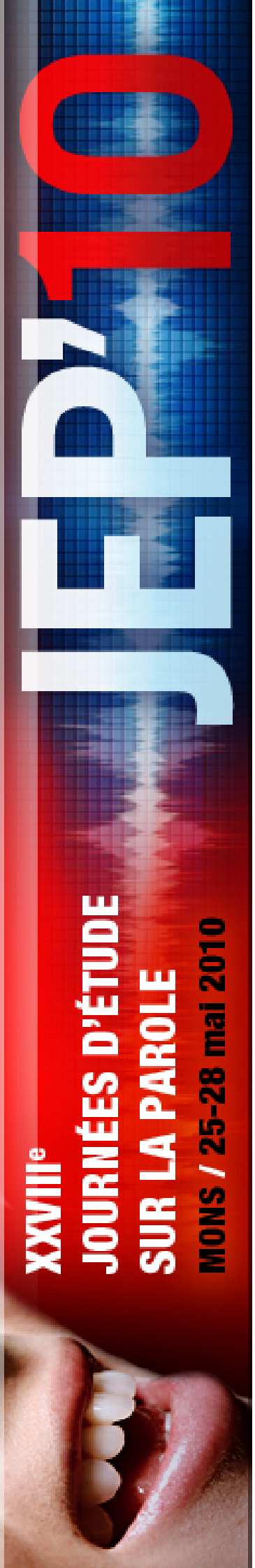


XXVIIIèmes Journées d'Etude sur la Parole JEP 2010

Actes

Mons, 25-28 mai 2010

UMONS Université de Mons

XXVIII^{èmes}
Journées d'Etude sur la Parole
JEP 2010

Université de Mons
Mons, Belgique

25 – 28 Mai 2010

Organisées par :

le Laboratoire des Sciences de la Parole (Académie Universitaire Wallonie-Bruxelles) : Laboratoire de phonétique (UMONS), Laboratoire de phonologie expérimentale (ULB), et Laboratoire Théorie des Circuits et Traitement du Signal (UMONS) ;

Sous l'égide de :

l'Association Francophone de la Communication Parlée (AFCP), un 'Special Interest Group' régional de l'International Speech Communication Association (ISCA) ;

Avec le soutien de :

l'Extension de l'Université de Mons, le Fonds National de la Recherche Scientifique, le Conseil de Recherche de l'Université de Mons-Hainaut, la Commission Recherche de l'Académie Universitaire Wallonie-Bruxelles, l'Ecole Doctorale Thématique Neurosciences, Multitel, Acapela.

Comité d'organisation

Véronique DELVAUX
Didier DEMOLIN
Stéphane DUPONT
Thierry DUTOIT
Bernard HARMEGNIES
Kathy HUET
Myriam PICCALUGA

Comité d'organisation local

Ludovic ABRASSART
Kanittarat BOOTTAWONG
Sarah BROHE
Nicolas CHAMPAGNE
Sandrine CLAIRET
Fernanda CONSONI
Anne COSYN
Gilles DELMEE
Véronique DELVAUX
Stéphanie DEMARTIN
Didier DEMOLIN
Virginie DUFRANCATEL
Stéphane DUPONT
Thierry DUTOIT
Matthieu DUVINAGE
Eric ERCULISSE
Bernard HARMEGNIES
Kathy HUET
Audrey LECLERCQ
Benjamin PICART
Myriam PICCALUGA
Flavien ROELANDT
Florence VERBANCK
Déborah WAIMBERG

Comité de programme

Présidente :

Cécile FOUGERON	LPP *(Paris)*

Gilles ADDA	LIMSI *(Paris)*
Melissa BARKAT-DEFRADAS	Praxiling *(Montpellier)*
Laurent BESACIER	LIG *(Grenoble)*
Véronique DELVAUX	UMONS-Labo phonétique *(Mons)*
Didier DEMOLIN	ULB-Labo phonologie expérimentale *(Bruxelles)*
Stéphane DUPONT	AUWB-UMONS-TCTS Lab *(Mons)*
Thierry DUTOIT	AUWB-UMONS-TCTS Lab *(Mons)*
Isabelle FERRANE	IRIT *(Toulouse)*
Corinne FREDOUILLE	LIA *(Avignon)*
Alain GHIO	LPL *(Aix-en-Provence)*
Guillaume GRAVIER	IRISA *(Rennes)*
Bernard HARMEGNIES	UMONS-Labo phonétique *(Mons)*
Kathy HUET	UMONS-Labo phonétique *(Mons)*
Irina ILLINA	LORIA *(Nancy)*
Bruno JACOB	LIUM *(Le Mans)*
Fabrice LEFEVRE	LIA *(Avignon)*
Christine MEUNIER	LPL *(Aix-en-Provence)*
François PELLEGRINO	DDL *(Lyon)*
Pascal PERRIER	Gipsa-Lab *(Grenoble)*
Myriam PICCALUGA	UMONS-Labo phonétique *(Mons)*
Solange ROSSATO	LIG *(Grenoble)*
Rudolph SOCK	IPS *(Strasbourg)*
Ioana VASILESCU	LIMSI *(Paris)*

Comité scientifique

Les membres du comité de programme sont particulièrement reconnaissants aux personnes dont les noms suivent pour l'aide qu'ils ont apportée lors de l'évaluation des communications soumises aux JEP 2010 :

Martine ADDA-DECKER	LIMSI
Alexandre ALLAUZEN	LIMSI
Régine ANDRÉ OBRECHT	IRIT
Véronique AUBERGÉ	Gipsa-Lab
Pierre BADIN	Gipsa-Lab
Odile BAGOU	Université de Neuchatel
Gérard BAILLY	Gipsa-Lab
Claude BARRAS	LIMSI
Denis BEAUTEMPS	Gipsa-Lab
Mathieu BEN	INRIA Rennes
Roxane BERTRAND	LPL
Frédéric BETTENS	UMONS
Brigitte BIGI	LPL
Frédéric BIMBOT	IRISA/CNRS
Anne BONNEAU	LORIA
Fethi BOUGARES	LIUM
Philippe BOULA DE MAREÜIL	LIMSI
Christian CAVÉ	LPL
Christophe CERISARA	LORIA
Vincent COLOTTE	LORIA
Lise CREVIER BUCHMAN	LPP
Christophe D'ALESSANDRO	LIMSI
Elisabeth DELAIS-ROUSSARIE	CNRS-Paris7
Marion DOHEN	Gipsa-Lab
Thomas DRUGMAN	UMONS
Thomas DUBUISSON	UMONS
Richard DUFOUR	LIUM
Sophie DUFOUR	LPL
Stéphane DUPONT	UMONS
Frédéric ELESEI	Gipsa-Lab
Robert ESPESSER	LPL
Yannick ESTÈVE	LIUM
Jérôme FARINAS	IRIT
Emmanuel FERRAGNE	Equipe CLILLAC-ARP, Université Paris 7, France
Dominique FOHR	LORIA
Cédric GENDROT	LPP
Alain GHIO	LPL
Laurent GIRIN	Gipsa-Lab
Pierre HALLÉ	LPP
Daniel HIRST	LPL
Irina ILLINA	LORIA

Denis JOUVET	LORIA
Lionel KOENIG	IRIT
Sacha KRSTULOVIC	Toshiba
Muriel LALAIN	Gipsa-Lab
David LANGLOIS	LORIA
Yves LAPRIE	LORIA
Antoine LAURENT	LIUM
Viet-Bac LE	LIMSI
Audrey LECLERC	UMONS
Benjamin LECOUTEUX	LIA
Hélène LOEVENBRUCK	Gipsa-Lab
Shinji MAEDA	ENST
Alain MARCHAL	LPL
Egidio MARSICO	DDL, CNRS - Université de Lyon, France
Hélène MAYNARD	LIMSI
Julien MEYER	Area de Linguistica, CCH, Museu Goeldi, Brésil
Alexis MICHAUD	Lacito-CNRS
Alexis MOINET	UMONS
Noel NGUYEN	LPL
Claire PILLOT LOISEAU	LPP
Julien PINQUIER	IRIT
Serge PINTO	LPL
Michel PITERMANN	LPL
Cristel PORTES	LPL
Christian RAYMOND	IRISA/INSA Rennes
Annie RIALLAND	LPP
Rachid RIDOUANE	LPP
Albert RILLIARD	LIMSI
Sophie ROSSET	LIMSI
Jean-Luc ROUAS	LIMSI
JeanLuc SCHWARTZ	Gipsa-Lab
Christine SÉNAC	IRIT
Natalie SNOEREN	LIMSI
Marion TELLIER	LPL
Jerome URBAIN	UMONS
Nathalie VALLÉE	Gipsa-Lab
Ioana VASILESCU	LIMSI
Béatrice VAXELAIRE	IPS
Anne VILAIN	Gipsa-Lab

Le mot du Comité d'organisation

Au moment où le point final s'imprime au bas de cette page, trois années ont passé depuis que l'idée, alors un peu folle, d'attirer à Mons les Journées d'étude sur la Parole a subitement surgi dans mon « arrière-cuisine » (le petit local attenant au laboratoire de phonétique où mes complices les plus assidues - Kathy, Myriam et Véronique - et moi-même tirons d'ordinaire des plans sur les comètes).

De rêve à quatre têtes, la perspective est vite devenue projet, d'abord informellement évoqué au CA de l'AFCP sur fond de grisaille parisienne, puis présenté officiellement en bonne et due forme, sous le soleil avignonnais.

Chemin faisant, des comparses « parole » de Belgique francophone (Didier Demolin, du *Laboratoire de phonologie expérimentale* de l'Université Libre de Bruxelles -ULB- et Thierry Dutoit, du *Laboratoire de Théorie des Circuits et Traitement du Signal* de la Faculté Polytechnique de Mons -FPMs-) nous avaient rejoints, et leurs laboratoires, associés au mien (*Laboratoire de phonétique* de l'Université de Mons-Hainaut -UMH-) allaient ainsi former le tripode où appuyer les XXVIII^{èmes} JEP.

Les trois laboratoires appartenaient alors à trois universités différentes, et avaient décidé, déjà, de se réunir dans le cadre du *Laboratoire des sciences de la parole* de l'Académie Universitaire Wallonie-Bruxelles (AUWB : un regroupement confédéral de ces trois universités, créé dans le cadre du Décret de 2004 réorganisant l'enseignement universitaire de Belgique francophone). Tout naturellement, il a ainsi été décidé de confier officiellement à cette structure ternaire l'organisation des JEP. Deux des trois universités partenaires (l'UMH et la FPMs) ayant fusionné en octobre 2009 pour créer une nouvelle institution (L'Université de Mons - UMONS -), c'est ainsi dans les murs séculaires d'une université nouvelle que se matérialise le Laboratoire des sciences de la parole de l'AUWB pour héberger les travaux des XXVIII^{èmes} JEP.

Les membres du Comité d'organisation des JEP se font les interprètes de la communauté Parole de Belgique francophone pour remercier le CA de l'AFCP d'avoir choisi de souligner le caractère international des JEP en leur faisant élire domicile un jet de pierre par-delà les confins septentrionaux de l'Hexagone. C'est là, pour les chercheurs francophones non français, une marque de reconnaissance et un témoignage de soutien dont nous savons le prix. C'est aussi pour tous les scientifiques qui travaillent en Francophonie le signe qu'une communication scientifique internationale de haut niveau peut se faire jour en français.

Mais la Belgique n'est sûrement pas *terra incognita* pour les JEP. Max Wajskop aimait à le rappeler, c'est en février 1968, au cours d'un colloque ayant fait converger à Bruxelles des membres des universités de Londres, d'Aix et de Grenoble, ainsi que du CNRS, que le projet - non trivial à l'époque - de réunir en un même lieu des scientifiques non seulement de pays différents, mais surtout d'origines disciplinaires variées, avait émergé : deux ans après il se concrétiserait à Grenoble par l'organisation des premières Journées d'Etude sur la Parole.

Depuis, les JEP sont revenues par trois fois déjà en Belgique (toujours à l'ULB) : en 1973, 1984 et 1992. Y aurait-il une composante BF (Belge Fréquence) dans la périodicité des JEP qui les ramènerait outre-Quiévrain sur une base quasi-décennale ? Aux générations futures de répondre à la question !

Mais pour l'heure, je voudrais adresser mes remerciements les plus vifs aux institutions, aux entreprises et aux personnes qui ont donné des moyens financiers et matériels, des lieux, du temps et de l'intelligence pour que les JEP montoises se concrétisent. Sans l'appui du Conseil de recherche de l'Université de Mons-Hainaut, de la Commission Recherche de l'Académie Universitaire Wallonie-Bruxelles, de l'Ecole Doctorale Thématique Neurosciences et du Fonds National de la Recherche Scientifique, sans le concours de Multitel, et Acapela, sans le soutien indéfectible de Véronique Delvaux, Kathy Huet, Myriam Piccaluga, Stéphane Dupont, Thierry Dutoit et Didier Demolin, sans l'aide précieuse de tous les membres du comité d'organisation ainsi que des services généraux de l'UMONS, sans l'efficacité de Fanny Lallemand et des autres membres de l'Extension UMONS, sans l'expertise et la sagesse des membres du comité scientifique, sans la disponibilité du comité de programme, sans la gentillesse et la clairvoyance attentive des président-e-s de l'AFCP - successivement Pascal Perrier et Cécile Fougeron - , je ne pourrais pas, au nom du Comité d'organisation, avoir l'honneur doublé du plaisir de souhaiter la bienvenue à Mons aux participants des XXVIII^{èmes} JEP.

Pour le Comité d'organisation,

Bernard Harmegnies

Le mot de la Présidente de l'AFCP

Chers collègues,

L'arrivée des JEP est toujours un temps fort dans la vie de la communauté Parole francophone. Cette année, l'AFCP se réjouit particulièrement de voir l'organisation des JEP en Belgique. D'une part, parce que la convivialité y est réputée, mais surtout parce que cette localisation réaffirme le caractère international de cette conférence francophone et de notre communauté scientifique.

De quoi cette édition sera faite ? Sur 136 communications soumises, 100 ont été retenues. Elles émanent de nombreux laboratoires situés dans différents pays comme la France, la Belgique, la Suisse, la Pologne, la Tunisie, l'Algérie, le Canada, le Maroc ou encore le Brésil. Comme par le passé, les thématiques abordées couvrent les multiples facettes de nos recherches sur la communication parlée, allant de la reconnaissance automatique de la parole à la description de pathologies de la parole, en passant par la description de contrastes phonologiques, l'apprentissage d'une langue seconde, ou encore la synthèse. Cette diversité thématique témoigne des liens forts entre les différentes disciplines dont l'objet d'étude est la parole. Il est donc à gager que ces JEP seront, comme à leur habitude, un carrefour de rencontres et d'échanges fructueux entre les différents acteurs de la recherche sur la parole. En particulier, pour témoigner de la synergie croissante entre sciences humaines et sciences de l'ingénieur à laquelle l'AFCP aspire, une session 'Convergence' a été organisée autour de communications ayant trait à des questions de recherche mariant ces deux communautés.

Les JEP se veulent aussi l'occasion d'échanges particuliers entre jeunes chercheurs et chercheurs confirmés. Dans ce sens, et dans le cadre de sa politique de soutien aux jeunes chercheurs, l'AFCP a octroyé de nombreuses bourses de transports à des jeunes chercheurs, et, pour la 3ème édition consécutive, se réjouit de pouvoir inviter cinq jeunes chercheurs appartenant à des laboratoires situés hors de France (Brésil, Tunisie, Maroc) à participer à ces rencontres.

L'organisation d'une rencontre comme les JEP n'est pas une tâche facile, avec laquelle il faut jongler pendant une longue période en accommodant les autres contraintes professionnelles et personnelles. Je tiens à remercier sincèrement tous les organisateurs, et particulièrement Véronique Delvaux, Kathy Huet, Bernard Harmegnies, Myriam Piccaluga, Stéphane Dupont, Thierry Dutoit, et Didier Demolin pour leur investissement dans cette aventure.

Au nom du comité de programme, je remercie aussi vivement les 105 relecteurs pour le temps consacré aux évaluations des articles soumis et le sérieux de leur travail. Pour rappel, les communications aux JEP sont sélectionnées sur la base d'un article complet (4 pages). Chaque soumission est évaluée par deux relecteurs. Le comité de programme, constitué des membres du CA de l'AFCP et de quelques membres du comité d'organisation, se réunit pendant deux jours pour sélectionner les communications. Les soumissions et leurs évaluations sont examinées, certaines sont relues par un 3ème lecteur, et la sélection est effectuée. Les communications sélectionnées sont alors groupées en grands thèmes, répondant aux thématiques des soumissions et aux thématiques souhaitées par les organisateurs. Sur cette base, les sessions thématiques de la conférence sont définies, et pour chaque session, cinq communications orales sont sélectionnées. Les autres communications, qui seront présentées sous forme de posters, ne sont pas regroupées thématiquement. En effet, comme pour les deux dernières éditions des JEP, nous avons souhaité des sessions poster couvrant un large spectre d'intérêts. Il est donc à noter qu'aux JEP, la sélection entre

communication orale et affichée s'effectue principalement sur la base d'un choix thématique pour les sessions orales et ne renvoie donc pas à un critère de qualité.
Pour conclure, je tiens à remercier les participants à ces XXVIIIèmes Journées d'Etude sur la Parole qui par leur dynamisme sont le moteur de notre communauté scientifique si sympathique.

Je vous souhaite à tous une conférence enrichissante et stimulante.

Cécile Fougeron
Présidente de l'AFCP
Présidente du Comité de Programme des XXVIIIèmes JEP

L'Association Francophone de la Communication Parlée (AFCP) est une structure d'animation et de réflexion de la communauté francophone travaillant sur la parole.
http://www.afcp-parole.org/

Table des matières

Résumés des conférences invitées

Narly GOLESTANI

Functional Brain Mapping Laboratory, Université de Genève

« Les bases cérébrales de l'apprentissage du langage et de l'expertise »

Guy CHÉRON

Laboratoire de Neurophysiologie et de Biomécanique du Mouvement, Université Libre de Bruxelles/ Service d'Electrophysiologie, Université de Mons

« Les oscillations cérébrales dans la perception active : une base sensori-motrice pour l'étude du langage »

Christophe D'ALESSANDRO

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur/CNRS

« La parole comme mouvement : glossolalies chironomiques »

James E. FLEGE

Division of Speech and Hearing Sciences, University of Alabama at Birmingham

« 'Age' effects on second language acquisition »

« Les bases cérébrales de l'apprentissage du langage et de l'expertise »

Narly A. GOLESTANI

Functional Brain Mapping Laboratory, Université de Genève

We have previously shown that both functional and structural brain differences underlie individual differences in foreign speech sound learning. This leads to the question of the relative influences of pre-existing, possibly 'innate' brain structural differences between individuals which might predict domain-specific language capacities, and of experience-dependent plasticity following systematic differences in learning. I will present recent evidence for both experience-dependent structural plasticity in experts, and aspects of brain anatomy that likely pre-date expertise training. Our results suggest that both pre-existing, possibly innate factors and environment influences play a role in brain structure and in specific language- related skills, with different relative contributions in different brain areas.

« Les oscillations cérébrales dans la perception active : une base sensori-motrice pour l'étude du langage »

Guy CHÉRON

Laboratoire de Neurophysiologie et de Biomécanique du Mouvement, Université Libre de Bruxelles/ Service d'Electrophysiologie, Université de Mons

L'étude des oscillations cérébrales mesurées par l'électroencéphalographie (EEG) dynamique ouvre un champ d'accès aux mécanismes neuronaux à la base du langage. En considérant que l'action et la perception sont fonctionnellement liées dans le cerveau, cette conférence présentera les nouveaux outils de l'EEG qui permettent aujourd'hui de dissocier les différentes étapes préalables à une perception consciente ainsi que l'élaboration de la réponse motrice. Une tentative d'intégrer la neuroanatomie fonctionnelle du langage aux données récentes de la neurophysiologie de la perception active sera envisagée.

« La parole comme mouvement : glossolalies chironomiques »

Christophe D'ALESSANDRO

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur / CNRS, France

La parole relève d'une part de la linguistique et d'autre par d'une sémiologie non linguistique, que l'on peut nommer « é-motive », ou incarnée. Les aspects prosodiques, au sens large de non (infra et supra) segmentaux, semblent plus particulièrement liés à l'émotion. Aborder la prosodie dans le cadre de la sémiologie du geste est une tradition qui remonte à la rhétorique antique, passe par la célèbre définition « d'ars bene movendi » chez Augustin, et réapparaît chez Fonagy (« la vive voix »), ou de Certeau (« Utopies vocales : glossolalies »), en particulier sous l'angle (en miroir) du corps dans la parole.

Une façon d'aborder de manière expérimentale cette « autre scène » de la parole est l'utilisation de systèmes de synthèse ou d'analyse/synthèse temps réels à contrôle gestuel. Le « mouvement » de la parole peut ainsi être externalisé comme « geste » (de la main, des bras, des pieds…) : c'est un nouveau point de vue (au sens de lieu d'observation et de manière d'étudier).

Deux aspects de ce programme de recherche sont présentés : la stylisation chironomique de l'intonation, et la glossolalie chironomique.

1- La stylisation prosodique est étudiée à l'aide d'un système chironomique. Le sujet trace sur une tablette graphique l'intonation (de phrases pré-enregistrées) à l'aide de gestes semblables à l'écriture ou au dessin. Les résultats montrent que, sans entrainement intensif, les sujets sont tout à fait capables de reproduire l'intonation de phrases cibles. Afin d'évaluer la précision de stylisation chironomique, on demande aussi aux sujets de reproduire la prosodie par une imitation vocale. Il apparaît que l'intonation dessinée est aussi proche (parfois plus, parfois moins) de l'original que l'intonation vocale. Dans un test perceptif de discrimination, l'intonation chironomique ne peut dans un grand nombre de cas pas être distinguée de l'intonation naturelle. Ainsi les mouvements intonatifs sont d'une certaine façon analogues aux mouvements expressifs de la main. Les travaux futurs porteront sur la stylisation chironomique du rythme.

2- Plusieurs nouveaux systèmes de synthèse vocale ont été développés. Il s'agit de modèles source/filtres temps réel à contrôle gestuel, pilotés par gants numérique (theremin vocal), clavier et pédales (orgue LF), joystick et tablette graphique (Calliphonie), méta instrument (Meta-calm), bras haptique (phonétogramme haptique). Ces systèmes permettent de synthétiser par le geste des énoncés de parole très expressifs, mais dépourvus de sens linguistique, des glossolalies chironomiques. Ce type de vocalisation est actuellement étudié pour des projets musicaux et pour l'analyse par la synthèse des attitudes en parole. Les travaux en cours portent également sur l'expression audio visuelle, avec une tête parlante, et sur la synthèse de consonnes.

Ces travaux sont le fruit de recherches, passées et en cours, menées au LIMSI avec Albert Rilliard, Sylvain Le Beux (maintenant Univ. McGill, Montréal), Lionel Feugère, Mehdi Ammi, Boris Doval (maintenant UPMC, Paris). Ils ont profité de fructueux échanges avec Thierry Dutoit et Nicolas D'Alessandro, en particulier dans le cadre du workshop Enterface'05, à Mons.

« "Age" effects on second language acquisition »

James E. FLEGE

Division of Speech and Hearing Sciences, University of Alabama at Birmingham

A consistent finding of L2 research is that early learners (age of arrival: 2-10 years) demonstrate greater L2 proficiency than late learners (AOA 15-23 years). Such effects might arise from neural maturation, cognitive changes, differences how L1-L2 systems interact, or input differences. No one hypothesis can account for all the data, suggesting that age effects arise from multiple factors that co-vary with age in ways not yet understood. Until potential causal variables are examined directly and confounded variables have been controlled, we can only speculate about the true bases for age-related differences in L2 proficiency.

Articles

Lecture Labiale, Surdité, et Langage Parlé Complété

Mario Aparicio[1], Philippe Peigneux[2], Brigitte Charlier[1], Charlotte Neyrat[1] et Jacqueline Leybaert[1]

[1] : Laboratoire Cognition Langage Développement, http://lcld.ulb.ac.be/ ;
[2] : UR2NF : Unité de Recherches en Neuropsychologie et Neuroimagerie fonctionnelle
Université Libre de Bruxelles

ABSTRACT

It has been shown that deaf subjects outperform hearing in speechreading. However, little is known about the reasons of this difference. In the present study, we measure speechreading performance in two groups of deaf and one group of hearing participants. One group of deaf was exposed early and intensively to Cued Speech (CS) whereas the other group of deaf has been exposed to oral language without CS. Results show that only the CS deaf group clearly outperformed the hearing group. We discuss the possibility that the early and accurate use of CS influences performance in speechreading

Keywords: speechreading, Cued Speech, deafness

INTRODUCTION

La Lecture Labiale (LL) est habituellement le seul moyen que le sourd possède pour décoder le message oral de son entourage. Cela peut être critique dans certaines situations, comme le milieu du travail. Par conséquent, une bonne capacité de LL s'avère importante pour procurer à la personne sourde une meilleure compréhension du message oral.

Quelques études ont montré que les personnes atteintes d'une surdité pré-linguale sévère ou profonde et qui utilisent la langue orale de manière quotidienne atteignent de meilleures performances d'identification des éléments linguistiques en LL [2], [7]. Toutefois, ces études font également état d'une forte variabilité parmi les sourds. Les causes des différences en LL entre entendants et sourds et de la variabilité existante dans la population sourde sont peu connues.

Les sourds exposés à la langue orale ont reçu des stimulations linguistiques différentes. Quelques sourds ont appris le Langage Parlé Complété (LPC) pour communiquer. Le LPC est la version française du Cued Speech, inventé afin d'aider des enfants sourds à résoudre l'ambigüité inhérente à la LL [4]. Dans le LPC, le locuteur utilise différentes positions de la main (clés) pour compléter l'information linguistique de la langue orale fournie par la lecture labiale (voir Figure 1) Le LPC est un mode de communication, capable de transmettre au récepteur toute l'information phonologique, sémantique et syntaxique de la langue orale en absence de la modalité auditive [6].

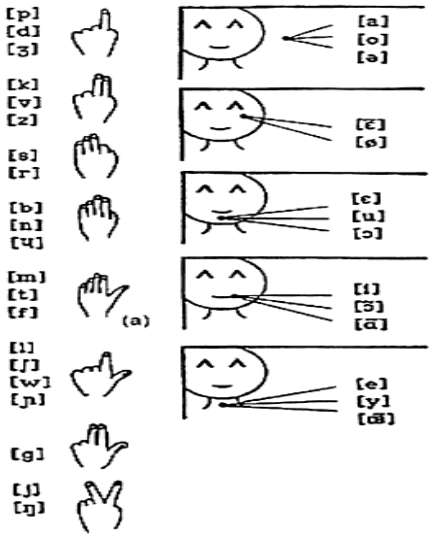


Figure 1 *Le LPC utilisé dans la langue française. Les clés sont déterminées par huit formes de la main et cinq positions. Tout le répertoire phonologique du Français est fourni*

Les études montrent que les sujets exposés au LPC de manière précoce (i.e. à la maison avec leurs parents) perçoivent mieux le message oral que les sourds qui n'ont pas été exposés au LPC [3] ; [5].

Le but de cette étude est d'investiguer si les sourds LPC-précoces sont des meilleurs labio-lecteurs que les sourds qui communiquent oralement sans utiliser LPC. Cela suggérerait que l'acquisition de bonnes représentations phonologiques à travers la modalité visuelle améliore l'intelligibilité du message oral fourni par la LL.

Pour examiner cela, nous avons créé un test de LL en Français. Les items utilisés sont des phrases, qui répondent plus que les mots isolés à une situation naturelle de communication. La tâche consiste à identifier l'image cible (parmi quatre images) qui correspondait a la phrase précédemment perçue en LL. Nous avons choisi cette tâche car l'identification des mots dans des phrases isolées (sans aucun contexte) est souvent très faible (autour de 5 %) [9]. Lorsque ces phrases sont contextualisées à l'aide des images, la performance peut augmenter jusqu'à 45-50 % [10]. En outre, ce type de réponse est mieux adapté à la population sourde car il lui permet d'exprimer son choix d'une manière non verbale.

METHODE

Sujets

36 participants adultes ont participé à cette étude. Parmi ces 36 participants, 12 étaient entendants, sans trouble linguistique et 24 participants étaient des sourds congénitaux avec une surdité sévère ou profonde (i.e. perte d'audition plus grande que 70 dB). Tous les participants sourds utilisaient la langue orale quotidiennement. Parmi les 24 participants sourds, 12 (le groupe de sourds-LPC) avaient reçu une stimulation en LPC riche et précoce (i.e. à la maison avec les parents) tandis que les 12 autres participants sourds avaient reçu une stimulation orale mais sans LPC (le groupe de sourds-non LPC). La majorité des sourds des deux groupes communiquaient aussi en Langue de Signes quotidiennement. L'âge moyen des sujets était de 25.2 ans (étendue : 21-37) pour le groupe d'entendants, de 25.0 ans pour le groupe de sourds-LPC (étendue : 18-33) et de 31.6 ans pour le groupe de sourds-non LPC (étendue : 18-49).

Materiel

Test de Lecture Labiale

Ce test est composé de 30 phrases qui étaient désignées pour tester les capacités de perception de la parole chez des sujets sourds et entendants, indépendamment de leur difference de compétence linguistique [8]. Par conséquent, les phrases étaient construites avec un materiel lingustique qui était accessible à tous les sujets. D'un côté, les mots utilisés dans les phrases avaient tous une haute frequence lexicale. D'un autre côté, les phrases possédaient toujours une même structure syntactique simple : sujet, verbe, complément. Chaque phrase était suivie d'une planche avec quatre images qui contenait la cible et trois distracteurs qui avaient la même structure syntaxique que la cible mais qui diffèrent sémantiquement et phonologiquement (voir figure 2 pour un exemple). La cible a été contrebalancée dans les quatre positions de la planche.

Les sujets regardaient l'interlocutrice prononcer la phrase et devaient choisir ensuite une de quatre images présentées simultanément: la cible et trois distracteurs.

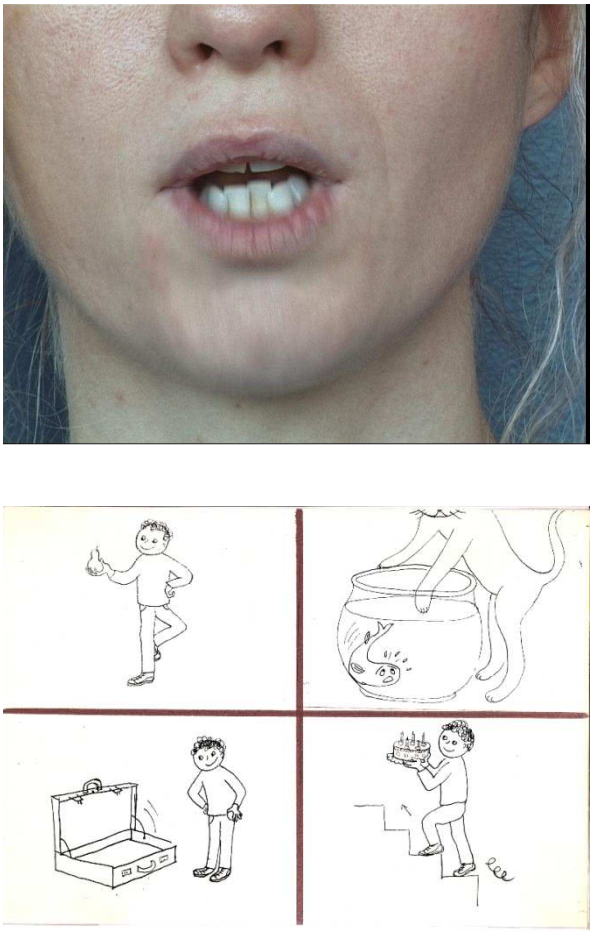


Figure 2 : *Un photogramme de la phrase « Il monte le gâteau » (en haut) et la planche des images correspondantes (en bas). Les trois distracteurs étaient les phrases : « Le garçon porte la poire », « le poisson a peur », « La valise était ouverte ».*

Les vidéos et les images étaient présentées sous MatLab à l'aide d'un ordinateur portable. Les items étaient prononcés par une locutrice de langue maternelle française. La vidéo montrait son visage depuis la base du nez jusqu'à la gorge (voir figure 2). Le sujet, assis devant l'écran de l'ordinateur, devait pointer avec le doigt l'image de son choix. Les réponses étaient enregistrées manuellement par l'expérimentateur.

RESULTATS

Tous les sujets avaient des performances significativement différentes du hasard (25 %) ($p < 0.05$)

La performance moyenne des entendants et des sourds (LPC et non LPC réunis) dans le test de Lecture Labiale est de 51.94 % (écart-type = 19.67) et de

71.39 % (écart-type = 17.8) respectivement (voir Figure 3)

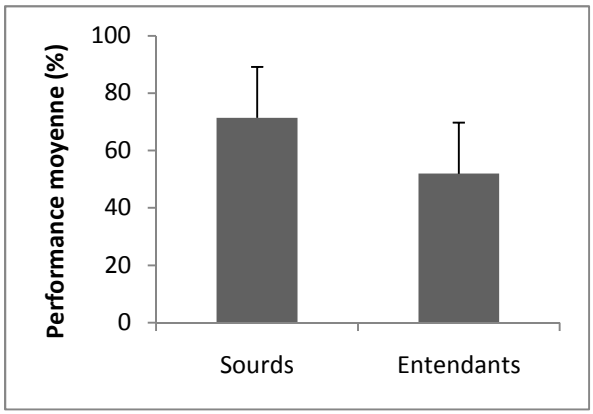


Figure 3 :*Pourcentage moyen de réponses correctes (et écarts types) pour les sourds et les entendants*

La performance des sourds est significativement plus élevée que celle des entendants ($t(34) = 2.99$; $p < 0.01$)

La comparaison des sourds en fonction du moyen de communication révèle une meilleure performance pour le groupe LPC (80.83 %, écart-type = 9.2) que pour le groupe non-LPC (61.94 %, écart-type = 18.99) (voir Figure 4)

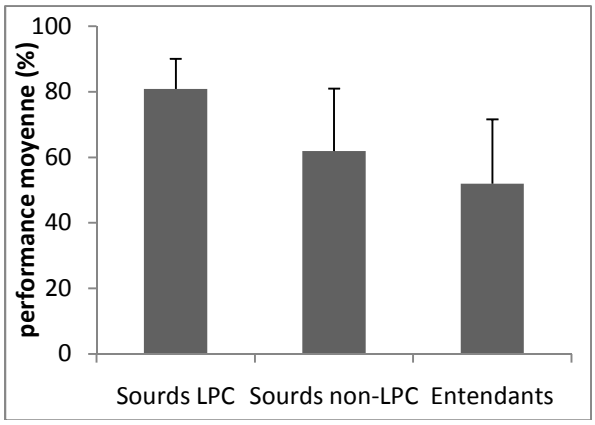


Figure 4 : *Pourcentage moyen de réponses correctes (et écarts-types) des sourds LPC, des sourds non LPC et des entendants*

Une analyse de variance sur le pourcentage moyen de RC montre une différence significative entre les 3 groupes, $F(2,33) = 7.04$; $p<0.01$. Un test post-hoc (Bonferroni-Dunn) montre une différence significative entre les sourds LPC et les entendants d'une part, ($p < 0.01$) et les sourds LPC et non-LPC ($p<0.05$) d'autre part. Il n'existe pas de différence statistiquement significative entre les sourds non-LPC et les entendants ($p > 0.2$).

DISCUSSION

Nous avons utilisé un test de LL en Français qui est adapté à la population de personnes sourdes. Nous avons comparé pour la première fois, deux populations de sourds différant par leur moyen de communication : LPC et LS. .

Nos résultats montrent une meilleure performance chez les sourds (LPC et non-LPC confondus) par rapport aux entendants (voir Figure 1). Cette différence de performance en LL entre sourds et entendants est conforme à d'autres études [2], [7].

Lorsque les sourds sont distingués en fonction du moyen de communication, le groupe des sourds-LPC témoigne d'une meilleure performance que celui des entendants et également que celui des sourds non-LPC. Les sourds non-LPC ont également une performance plus élevée que celle du groupe d'entendants, mais la différence n'est pas significative. Ces résultats suggèrent que la différence observée dans la Figure 1 est principalement due aux performances élevées du groupe de sourds LPC.

Dans les études précédentes, les causes de la différence en LL entre sourds et entendants ont été peu documentées. La distinction entre les sourds qui utilisent un code manuel pour la langue orale et ceux qui ne l'utilisent pas n'a pas été prise en compte [2].

Pour expliquer la différence entre les sourds LPC et les autres deux groupes, nous pourrions penser à l'existence d'une différence quantitative d'exposition à la LL. En effet, il est vraisemblable que le groupe de sourds LPC ait été davantage exposé à la LL que le groupe de sourds non-LPC et que le groupe d'entendants. Toutefois, le groupe de sourds non-LPC utilise également l'oral, et donc la LL, pour communiquer avec les entendants.

Une suggestion complémentaire est que la performance en LL ait été influencée par la bonne qualité des représentations phonologiques chez les sujets sourds LPC. En effet, les sourds qui ont été stimulés de manière précoce en LPC, possèdent un meilleur traitement phonologique qui améliore sa perception du message oral [1]. Ce meilleur traitement phonologique pourrait augmenter leur capacité de détection linguistique dans des conditions où l'information est incomplète (comme c'est le cas pour la LL). Toutefois, les participants entendants, qui ont des performances faibles en LL, possèdent également de bonnes représentations phonologiques.

Les sourds LPC se distinguent des sourds non-LPC par la qualité des représentations phonologiques de la langue orale et se distinguent des entendants par la modalité employée dans la perception du message (visuel seul vs audio-visuel). Par conséquent, nous suggérerons que ces deux éléments ensemble (i.e. la qualité des représentations phonologiques ET la modalité visuelle exclusive de perception de la langue)

sont importantes pour expliquer les différences de performance en LL.

Du point de vue méthodologique, notre test s'avère sensible à révéler des différences dans l'efficience en LL. Dans l'avenir, nous comptons l'utiliser pour investiguer les capacités de LL d'autres populations au développement atypique du langage (enfants dyslexiques, enfants avec troubles spécifiques du langage).

REFERENCES

[1] Alegria, J., Charlier, B. L., & Mattys, S. (1999). The role of Lip-reading and Cued Speech in the Processing of Phonological Information in French-educated Deaf Children. *European Journal of Cognitive Psychology, 11*(4), 451-472.

[2] Bernstein, L. E., Auer, E. T., Jr., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *Neuroreport, 13*(3), 311-315.

[3] Charlier, B. L., & Leybaert, J. (2000). The rhyming skills of deaf children educated with phonetically augmented speechreading. *The Quarterly Journal of Experimental Psychology A, 53*(2), 349-375.

[4] Cornett, O. (1967). Cued speech. *American Annals of the Deaf, 112*, 3-13.

[5] Leybaert, J. (2000). Phonology acquired through the eyes and spelling in deaf children. *Journal of Experimental Child Psychology, 75*(4), 291-318

[6] Leybaert, J., Alegria, J., Hage, C., & Charlier, B. (1998). The effect of exposure to phonetically augmented lipspeech in the prelingual deaf. In R. Campbell, B. Dodd & B. Charlier (Eds.), *Hearing by eye II. Advances in the psychology of speechreading and auditory visual speech* (pp. 283-301): Psychology Press.

[7] Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., & Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clinical Linguistics & Phonetics, 20*(7-8), 621-630.

[8] Perier, O., Charlier, B., Hage, C. & Alegria, J. (1990). Evaluations of the effects of prolonged Cued Speech Pratice upon the reception of spoken language. *Cued Speech Journal, IV, 47-59*

[9] Samuelsson, S., & Rönnberg, J. (1991). Script activation in lipreading. *Scandinavian journal of Psychology, 32*(2), 124-143.

[10] Samuelsson, S., & Rönnberg, J. (1993). Implicit and explicit use of scripted constraints in lip-reading. *European Journal of Cognitive Psychology, 5*(2), 201-233.

Relations temporelles entre parole et gestualité co-verbale en français spontané

Gaëlle Ferré

Laboratoire de Linguistique de Nantes (LLING)
Université de Nantes
Chemin de la Censive du Tertre, BP 81227
44312 Nantes cedex 3
Gaelle.Ferre@univ-nantes.fr

ABSTRACT

Several studies have described the links between gesture and speech in terms of timing, most of them concentrating on the production of hand gestures during speech or during pauses (Beattie & Aboudan [1]; Nobe [16]). Other studies have focused on the anticipation or delay of gestures regarding their co-occurrence with speech (Schegloff [18]; McNeill [15]; Chui [6]; Kida & Faraco [9]; Leonard and Cummins [13]) and we would like to take part in the debate in the present paper. We studied the timing relationships between iconic gestures and their lexical affiliates (Kipp, Neff et al. [11]) in a corpus of French conversational speech involving 6 speakers and annotated both in Praat (Boersma & Weenink [4]) and Anvil (Kipp [10]).

Keywords: Multimodality, co-verbal gestures, timing relationships, lexical affiliates

1. INTRODUCTION

Parmi les études toujours plus nombreuses en multimodalité qui s'intéressent à la gestualité co-verbale — dont le rôle communicationnel a été montré par McNeill [15] entre autres — un certain nombre s'est attaché à décrire les relations temporelles qui existent entre le geste et la parole. L'un des intérêts de ce type de recherche est de pouvoir comprendre les systèmes multimodaux et de pouvoir ainsi alimenter le développement d'avatars. Ainsi, par exemple, Beattie & Aboudan [1] et Nobe [16] se sont penchés sur la co-occurrence des gestes manuels et des pauses silencieuses ou du temps d'articulation. D'autres études (Schegloff [18], McNeill [15], Leonard & Cummins [13] sur l'anglais; Chui [6] sur le chinois; Kida & Faraco [9] sur le français) se sont concentrées sur l'anticipation ou le retard de la gestualité co-verbale par rapport à la parole. C'est sur ce point que portera le présent article, car avec le développement des corpus vidéos annotés, une plus grande précision peut être atteinte. Ainsi, nous avons travaillé sur le corpus CID (Bertrand, Blache et al. [2], Blache, Bertrand et al. [3]) et analysé les relations temporelles entre les gestes iconiques (décrits dans la section 2.2) et la parole. Pour ce faire, nous avons mis en relation les groupes intonatifs (IP) avec les phrases gestuelles (cf. section 2.2), et les affiliés lexicaux (cf. section 2.3) avec la phase de réalisation du geste (Gstroke, cf. section 2.2), car ces unités nous ont semblées comparables. Les résultats montrent une très nette anticipation de la gestualité par rapport à la parole. Ils montrent aussi que si une unité gestuelle peut être décomposée en plusieurs items comme l'unité intonative peut se décomposer en mots, les unités gestuelles sont également plus longues que les unités verbales.

2. CORPUS ET DONNÉES

Pour cette étude, nous avons travaillé sur une sous-partie du corpus vidéo CID (décrit dans Bertrand, Blache et al. [2]), soit 45 minutes de parole interactionnelle (3 dyades de 15 minutes chacune) impliquant 6 locuteurs. L'annotation et l'exploitation du corpus font l'objet actuellement d'un projet financé par l'ANR (ANR BLAN08-2_349062).

2.1. Transcription du corpus

Nous avons travaillé sur une transcription et un alignement semi-automatique du corpus dans Praat, corrigés manuellement. Les groupes intonatifs (Intonational Phrases, Selkirk [19]) ont également été annotés dans Praat. Nous avons en effet pensé que cette unité était beaucoup plus appropriée au découpage de l'oral que des unités comme la phrase syntaxique ou la proposition qui présentent certains inconvénients et ne correspondent pas toujours au découpage exprimé par les locuteurs : par exemple, il n'est pas rare qu'une conjonction soit insérée en fin de groupe intonatif et suivie d'une pause silencieuse. Si syntaxiquement, la conjonction fait partie du groupe syntaxique situé à sa droite, intonativement, elle est rattachée au groupe syntaxique gauche. Ceci a un impact pragmatique puisque cette stratégie permet au locuteur de conserver la parole (Ferré [7]). Le groupe intonatif nous semble pour cette raison plus approprié pour rendre compte du découpage de l'oral et peut plus facilement être mis en relation avec des unités gestuelles que nous allons décrire dans la section 2.2. Loehr [14] a d'ailleurs montré qu'il existe un lien entre groupes intonatifs et gestualité co-verbale.

Ces annotations sur la parole ont ensuite été importées dans Anvil (logiciel d'annotation des fichiers vidéos, Kipp [10]) afin de pouvoir comparer les données verbales et les données gestuelles. Anvil présente également l'avantage d'imposer une structuration hiérarchique des données de type XML ce qui a un impact sur l'annotation des gestes présentée ci-dessous.

2.2. Annotations gestuelles

L'ensemble des gestes manuels des 6 locuteurs a été transcrit manuellement sur les 45 minutes de corpus (l'annotation de 3 heures de corpus est actuellement en cours). Outre la configuration de la main, le type de mouvement, etc, qui ne nous sont pas directement utiles ici, nous avons annoté le type de geste (d'après la typologie de McNeill [15], décrite plus bas) dans ce qui constitue la Phrase gestuelle (Kendon [8]), c'est-à-dire le geste dans sa globalité, depuis la mise en place des articulateurs (bras, mains, doigts) jusqu'au repos final ou jusqu'au début du geste suivant lorsque deux gestes sont enchaînés sans retrait des articulateurs (en ayant cependant à l'esprit que l'annotation gestuelle est moins précise que l'annotation de la parole puisqu'elle est basée sur un enregistrement comptant 24 images/seconde).

Toujours selon Kendon [8], la Phrase gestuelle se décompose en différentes phases que sont la préparation (mise en place des articulateurs), la réalisation du geste (stroke), une éventuelle tenue du geste, et la rétraction. Seule la phase de réalisation est nécessaire pour former une phrase gestuelle, les autres phases étant facultatives. Ces différentes phases ont également été annotées sur 45 minutes d'enregistrement.

La typologie des gestes manuels employée pour l'annotation se compose des types de geste suivants : les iconiques représentent une caractéristique physique d'un objet de discours ou miment des actions, les métaphoriques représentent des idées abstraites, les déictiques pointent vers un référent (spatial ou énonciatif), les emblèmes sont des gestes conventionnels, les battements des gestes de scansion du discours, et enfin, les adaptateurs des gestes d'auto-contact.

Parmi ces gestes, nous avons retenu les gestes iconiques uniquement, plus nombreux que les autres, soit 107 occurrences (nous avons écarté 18 gestes iconiques pour lesquels il n'était pas possible de déterminer un affilié lexical).

2.3. Affiliés lexicaux

En effet, s'il s'agit de mettre en relation les gestes manuels co-verbaux et la parole, il faut pouvoir être certain de mettre en relation des unités de nature comparable, d'où la notion d'affiliation lexicale sur laquelle repose l'article de Schegloff [18] et définie par Kipp et al. [11] comme : « The word or words deemed to correspond most closely to a gesture in meaning ». Si l'on considère les gestes iconiques, il apparaît que dans 85.6% des occurrences, il est possible de déterminer un affilié lexical dans une relation de redondance par rapport à la parole et correspondant à un mot comme dans les Figures 1 et 2.

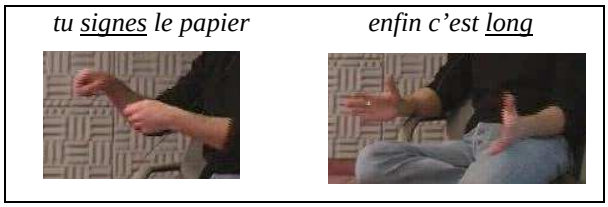


Figure 1 : Gestes iconiques correspondant aux affiliés lexicaux « signes » et « long ».

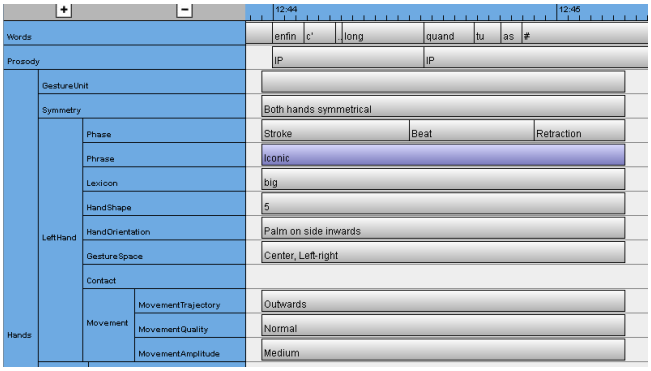


Figure 2 : Annotation Anvil correspondant à l'affilié lexical « long » dans « enfin c'est long quand tu as # tout le le la période de travail et tout ».

En ce qui concerne les autres catégories de gestes, soit elles contiennent très peu d'occurrences comme le cas des emblèmes, soit il est impossible de déterminer un affilié lexical précis comme dans le cas de nombreux métaphoriques qui apportent une modalité à tout l'énoncé comme dans la Figure 3. La relation sémantique entre geste et parole est alors implicite, or, pour pouvoir effectuer une comparaison en termes de temporalité, il faut pouvoir se baser sur une relation sémantique explicite entre geste et parole. C'est pour cette raison que nous avons choisi pour cette étude de ne retenir que les gestes iconiques, choix qui était également celui de Chui [6], alors que Kida & Faraco [9] et Loehr [14] ont travaillé sur différentes catégories de geste.

Figure 3 : Geste métaphorique produit sur « on n'en avait pas reparlé », qui permet d'apporter une modalité à l'énoncé mais pour lequel il est difficile de déterminer un affilié lexical précis.

3. RÉSULTATS

En ce qui concerne les résultats de cette étude, la première remarque que l'on peut faire est que les unités gestuelles (au niveau lexical, « Gstroke », et au niveau phrastique,

« Gphrase ») sont plus longues que les unités verbales correspondantes (mot et IP), même si l'écart entre unités phrastiques est moins important que l'écart entre unités lexicales.

En termes de relations temporelles (cf. pourcentages et écart moyen dans la Table 1), si l'on se place au niveau des unités lexicales en comparant le début et la fin du geste lui-même (« stroke ») au début et à la fin des tokens qui constituent les affiliés lexicaux, on constate qu'une large majorité de gestes (81.3%) commencent largement avant la production de l'affilié lexical en parole, et une proportion plus importante de gestes se terminent après la production de l'affilié lexical (61.7%) qu'avant celle-ci. Un Test T apparié montre que l'anticipation de la phase de réalisation sur l'affilié lexical en parole est significative (t=-7.85 ; p=1.73E-12). En revanche, on ne peut pas dire que le geste se termine avant ou après la parole de manière significative au niveau lexical (t=1.14 ; p=0.12). Ces statistiques ne s'expliquent donc pas uniquement par la durée plus importante du geste par rapport à la parole puisque le décalage temporel moyen est plus important dans le cas de l'anticipation du geste.

En ce qui concerne la relation temporelle entre les unités phrastiques (phrase gestuelle vs. IP), la tendance est la même, à savoir une anticipation de la gestualité sur la parole (60.75% des phrases gestuelles commencent avant les IP), mais le pourcentage est moins élevé que pour les unités lexicales. Egalement, 64.5% des phrases gestuelles se terminent après les IP avec une différence temporelle moyenne plus importante pour les gestes qui anticipent sur la parole. Pour le temps de début comme pour le temps de fin, le Test T apparié montre que le geste commence significativement avant la parole au niveau phrastique (t=-2.92 ; p=0.002) et se termine après la parole ((t=2.90 ; p=0.002). On notera toutefois qu'au niveau phrastique, la différence est moins importante qu'au niveau lexical. Si dans tous les cas rencontrés, il y avait nécessairement chevauchement entre la production des phrases gestuelles et la production des IP, lorsque le verbal anticipe sur le geste, le décalage temporel est plus important encore, de l'ordre d'une demi-seconde. Il faudrait donc regarder si dans ces cas précis, il n'y aurait pas d'hésitation au niveau de la production verbale.

Enfin, en ce qui concerne la comparaison entre la phrase gestuelle et l'affilié lexical, il est apparu que sur 107 iconiques, nous n'avons trouvé que 8 cas où la phrase gestuelle dans sa globalité était terminée avant la production de l'affilié lexical (nombre de ces cas contenaient des marques d'hésitation) alors que dans tous les autres cas, la phrase gestuelle et l'affilié lexical sont co-occurrents. Quant à la relation temporelle, 99% des phrases gestuelles commencent avant la production de l'affilié lexical (avec une différence hautement significative : t=-13.02 ; p=4.85E-24) et 85% d'entre elles se terminent après la production de l'affilié lexical (également de manière très significative : t=6.79 ; p=3.21E-10).

Table 1 : Pourcentage de gestes qui commencent ou finissent avant/après le verbal

	% de gestes qui commencent		**% de gestes qui se terminent**	
	avant la parole	**après la parole**	**avant la parole**	**après la parole**
Gstroke/ Affilié	**81,3**	**18,7**	**38,3**	**61,7**
Différence moyenne	0,566 s	0,14 s	0,44 s	0,391 s
Gphrase/IP	**60,75**	**39,25**	**35,5**	**64,5**
Différence moyenne	0,271 s	0,412 s	0,413 s	0,191 s
Gphrase /Affilié	**99**	**1**	**15**	**85**
Différence moyenne	0,76 s	0,098 s	0,578 s	0,804 s

4. DISCUSSION

Dans cette étude, nous avons présenté les résultats d'une des premières études portant spécifiquement sur le geste réalisées à partir du corpus CID. En effet, les récentes annotations gestuelles sur ce corpus nous ont permis de tester les relations temporelles entre gestualité co-verbale et parole dans le cas des gestes iconiques. Le choix de la catégorie gestuelle est justifié par la possibilité de déterminer pour ce type de geste un affilié lexical explicitement mentionné par les locuteurs.

Les résultats présentés dans ce travail – portant sur 107 gestes iconiques produits par 6 locuteurs pendant 45 minutes de français spontané – montrent que les relations temporelles qui existent entre la gestualité co-verbale et la parole, vont clairement dans le sens d'une anticipation du geste sur la parole. Mais si l'on considère les différentes études réalisées dans ce domaine, celles-ci affichent des résultats opposés. En effet, pour le chinois, Chui ([6]:878) a trouvé une plus grande proportion de gestes synchronisés avec la parole que de gestes anticipant la parole (60.1% vs 35.6%), avec des résultats semblables pour Loehr [14] sur l'anglais, toutes catégories de gestes confondues. En revanche, Schegloff [18], qui a travaillé sur les gestes déictiques en anglais, constate que les réalisations gestuelles (« strokes ») sont produites généralement de manière anticipée par rapport à leur affilié lexical. Leonard & Cummins [13], dans une récente étude, trouvent également une anticipation du geste sur la parole dans cette langue. Leur travail concernait plus précisément l'alignement des battements, décomposés en leurs différentes phases, avec l'affilié lexical. Ils ont montré – sur un corpus très réduit – que la phase de réalisation du battement anticipait sur l'onset de la voyelle dans l'affilié lexical correspondant. Ils ont aussi montré que le geste s'achève après la parole, comme dans notre corpus. Bourguet & Ando [5], sur les gestes déictiques en japonais, insistent plutôt sur la variabilité des relations temporelles entre geste et voix, et montrent qu'en fonction du type de déictique produit, le geste peut anticiper la

parole ou au contraire être produit après la parole. Enfin Kranstedt et al. [12], également sur les déictiques en anglais, montrent que le geste est produit avec un retard par rapport à la parole.

Devant une telle variabilité des résultats obtenus dans les différentes études, il convient de s'interroger sur les raisons de cette variabilité. Nous nous sommes tournés vers la réalisation du geste et notamment son amplitude qui pourrait agir sur les relations temporelles entre geste et parole. La tendance observée est une anticipation plus grande pour les gestes de grande amplitude et moins grande pour les gestes de petite amplitude. Mais ces observations de moyennes ne sont pas statistiquement significatives. Rochet-Capellan, et al. [17] remarquent que sur les déictiques produits en français et en portugais, la parole et la gestualité co-verbale tendent à l'isochronie, avec un décalage du geste afin que son apogée corresponde à la syllabe accentuée de l'affilié lexical. Nous n'avons pas pu vérifier cette tendance sur notre corpus, mais il est possible que la variabilité observée dans les études tienne à la nature de la relation geste / parole. En effet, la relation étudiée ici sur les gestes iconiques était une relation de redondance alors que les gestes déictiques dans d'autres travaux sont dans une relation de complémentarité avec la parole. On pourrait penser que dans le cas de la redondance, la synchronisation entre geste et parole est moins nécessaire que dans le cas de la complémentarité, et tend à être moins précise. Enfin, le type de corpus annoté (conversationnel vs. expérimental) pourrait également avoir un impact sur les relations temporelles entre gestes et parole.

BIBLIOGRAPHIE

[1] G. Beattie and R. Aboudan. Gestures, pauses and speech - an experimental investigation of the effects of changing social-context on their precise temporal relationships. *Semiotica*, 99:3-4, 1994.

[2] R. Bertrand, P. Blache, et al. Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *TAL*, 49(3): 105-133, 2008.

[3] P. Blache, R. Bertrand, et al. Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project. In M. Kipp et al. (eds.), *Multimodal Corpora.* Berlin, Heidelberg, Springer-Verlag, 38-53, 2009.

[4] P. Boersma and D. Weenink. *Praat: doing phonetics by computer (Version 5.1.05)* [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/

[5] M.-L. Bourguet and A. Ando. Synchronization of Speech and Hand Gestures during Multimodal Human-Computer Interaction. In Karat, C.-M., et al. (eds.), *Human Factors in Computing Systems, CHI 98.* Los Angeles, CA, ACM Press, 241-242, 1998.

[6] K. Chui. Temporal Patterning of Speech and Iconic Gestures in Conversational Discourse. *Journal of Pragmatics*, 37:871-887, 2005.

[7] G. Ferré. Les pauses démarcatives déplacées en anglais spontané : marquage prosodique et kinésique. *Lidil*, 26:155-169, 2002.

[8] A. Kendon. Gesture and speech: two aspects of the process of utterance. In M.R. Key (ed.), *Nonverbal Communication and Language*. The Hague, Mouton, 207-227, 1980.

[9] T. Kida and M. Faraco. Prédication gestuelle. *Faits de Langues*, 31-32:217-226, 2008.

[10] M. Kipp. Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*:1367-1370, 2001.

[11] M. Kipp, M. Neff, et al. An annotation Scheme for Conversational Gestures: How to Economically Capture Timing and Form. *Language Resources and Evaluation*, 41:325-339, 2007.

[12] A. Kranstedt, P. Kühnlein and I. Wachsmuth. Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach. In A. C. Volpe (ed.) *Gesture-Based Communication in Human-Computer Interaction.* Berlin, Heidelberg, Springer-Verlag, 112-123, 2004.

[13] T. Leonard and F. Cummins. Temporal Alignment of Gesture and Speech. In *Proceedings of Gespin*, Poznan, Pologne. [CD-Rom], 2009.

[14] D. Loehr. *Gesture and Intonation.* PhD Thesis. Georgetown University, 2004.

[15] D. McNeill. *Hand and Mind : What Gestures Reveal about Thought.* Chicago, London, The University of Chicago Press, 1992.

[16] S. Nobe. Where do *most* spontaneous representational gestures actually occur with respect to speech? In D. McNeill (ed.), *Language and Gesture*. Cambridge, CUP, 186-198, 2000.

[17] A. Rochet-Capellan, C. Vilain et al. Does the Number of Syllables Affect the Finger Pointing Movement in a Pointing-naming Task? *8th International Seminar on Speech Production.* Strasbourg, 257-260, 2008.

[18] E. A. Schegloff. On Some Gestures' Relation to Talk. In J. M. Atkinson & J. Heritage (eds.), *Structures of Social Action*. Cambridge, CUP, 266-298, 1984.

[19] E. Selkirk. On Prosodic Structure and its Relation to Syntactic Structure. In T. Fretheim (ed.), *Nordic Prosody II*. Trondheim, Tapir, 111-140, 1978.

Production conjointe de gestes bracchio-manuels et de focalisation prosodique : coordination temporelle et effets de la production d'un geste sur les corrélats acoustiques/articulatoires de la focalisation

Benjamin Roustan et Marion Dohen

Département Parole et Cognition – GIPSA-lab – UMR5216 CNRS, Université de Grenoble
961, rue de la Houille Blanche 38402 Saint-Martin-d'Hères, France

ABSTRACT

Speech, and prosody in particular, is tightly linked to manual gestures . This study investigates the coordination of prosodic contrastive focus and different manual gestures (pointing, beat and control gestures). We used motion capture on ten speakers to explore this issue. The results show that prosodic focus "attracts" the manual gesture whichever its type. The temporal alignment between speech and the manual gesture is the strictest for pointing. It is realized between the pointing apex and an articulatory vocalic target whatever the position of focus within the utterance. Moreover, the results show that the production of a manual gesture, whichever its type, does not affect the acoustic and articulatory correlates of prosodic focus.

1. Introduction

Gestes manuels et parole sont liés dans la production de la parole en interaction (*e.g.* 1, 2). Plusieurs études [3–9] ont notamment mis en évidence les liens entre indices prosodiques et gestes manuels au niveau de la coordination gestes manuels/parole.

La deixis est la capacité d'attirer l'attention, de désigner. Dans l'espace, elle peut-être réalisée par le pointage manuel [10]. Dans la parole, la focalisation peut jouer ce rôle. En particulier, la focalisation contrastive prosodique est utilisée pour mettre en avant (*i.e.* désigner) un mot ou un groupe de mots au niveau informationnel. La deixis peut ainsi être réalisée de façon multimodale. Plusieurs études se sont interessées à la coordination temporelle des réalisations unimodales de la deixis. De Ruiter [11] a montré que la variation de la position de l' accent lexical dans un mot simple n'a aucun effet sur la coordination temporelle parole/pointage. Par contre, il a montré que la variation de la position de la focalisation prosodique au sein d'un syntagme nominal (adjectif+nom) avait un effet sur l'instant d'initialisation du geste de pointer. Rochet-Capellan *et al.* [12] ont exploré la coordination parole/pointage dans une tâche combinée (pointer+parole) de dénomination d'une cible (non-mot de 2 syllabes) et en faisant varier la position de l'accentuation lexicale. Ils ont trouvé que la position de l'accent lexical avait une influence sur la coordination temporelle parole/pointage en ce sens que le geste était déplacé pour que la syllabe accentuée soit toujours contenue dans la partie du geste qui désigne (index maintenu pointé vers la cible). Il semblerait donc que les différentes modalités de la deixis puissent être liées dans leur réalisation. Ce lien doit pourtant encore être précisé. Soulignons de plus que les études citées ci-dessus se sont principalement intéressé à des productions vocales simples (non-mots, mots ou syntagmes nominaux isolés).

Plusieurs études (*e.g.* 1, 5) ont de plus mis en évidence le lien potentiel entre prosodie et gestes de battements (*beat gestures* ou *batons* : oscillation verticale de la main de haut en bas).

Outre la coordination temporelle gestes manuels/parole, se pose également la question de l'influence de la production de gestes manuels sur la parole. Krahmer et Swerts [13] ont montré que la production d'un "battement visuel" (geste manuel, mouvement de sourcil ou hochement de tête) avait un effet significatif sur la durée des productions vocales concomitantes et sur le formant F_2. Cet effet était par ailleurs proche des corrélats acoustiques d'une accentuation emphatique. Ce résultat suggère que la production d'un geste manuel aurait un effet sur la production de la parole concomitante. L'objectif de cette étude est d'analyser la coordination entre la production de la focalisation contrastive prosodique en français et plusieurs types de gestes manuels de natures différentes en utilisant des phrases complètes et des mesures de capture de mouvement. Les questions posées sont les suivantes : 1. La focalisation prosodique et les gestes manuels sont-ils coordonnés temporellement et si oui, comment ? 2. Cette coordination éventuelle est-elle influencée par le type de geste et notamment son lien fonctionnel avec la parole ? 3. La production d'un geste manuel a-t-elle un effet sur les corrélats acoustiques/articulatoires de la focalisation prosodique ?

2. Méthodologie

2.1. Protocole expérimental

Corpus Le corpus était constitué de 4 phrases de structure syntaxique sujet (S) – verbe (V) – objet (O) en français (ex : Mumu tient le bébé). La structure syllabique était la suivante : S = 2 syl (nom propre) ; V = 1 syl (verbe d'action au présent) ; O = 1+2 syl (article+nom commun). Les mots cibles (S et O) commençaient tous par une consonne bilabiale.

Conditions expérimentales Nous nous sommes intéressés à deux conditions de focalisation : focali-

sation sur le sujet (FS ; exemple : $MUMU_F$ tient le bébé) et sur l'objet (FO). Deux conditions de parole ont été étudiées : parole seule et parole+geste manuel. Trois types de gestes manuels ont été analysés : pointage (geste manuel déictique communicatif), battement (geste manuel non déictique communicatif), et contrôle (appui sur un bouton ; geste manuel non déictique non communicatif).

Tâches Une tâche de correction (voir 14) a été utilisée pour induire la production de focalisation prosodique dans un contexte naturel de dialogue. Les participants entendaient un prompt audio dans lequel deux locuteurs discutaient et devaient corriger la phrase du deuxième locuteur en fonction de ce qu'avait dit le premier. Il était simplement demandé aux participants d'effectuer la tâche de correction dans la condition demandée. Deux images relatives au dialogue entendu étaient affichées sur un écran en face du participant qui, en condition de pointage, pointait vers l'image adéquate en même temps qu'il effectuait la correction. En condition parole+geste manuel, la seule indication donnée aux participants était de produire le geste manuel en même temps qu'ils parlaient.

Protocole L'expérience était divisée en quatre blocs (un par condition : parole seule, parole+geste manuel avec trois types de gestes). Avant chaque bloc, les participants s'entraînaient brièvement à la tâche avec des phrases différentes de celles utilisées pendant les phases de test. L'ordre des blocs et l'ordre des stimuli à l'interieur d'un bloc étaient aléatoires et différents pour chaque participant. Chaque bloc comportait 16 essais (4 phrases, 2 types de focalisation, 2 répétitions).

2.2. Participants

Dix adultes droitiers et de langue maternelle française ont participé à l'expérience (8 femmes et 2 hommes ; âge moyen : $30,2$)

2.3. Dispositif Experimental

Les participants étaient assis sur une chaise devant un écran. Une position de repos était marquée sur une table située à leur droite. Il leur était demandé de placer leur index sur ce repère en phase de repos : de partir de cette position pour faire un geste puis d'y revenir une fois le geste terminé. Les mouvements de leurs lèvres et de leur main droite étaient enregistrés grâce à un système 3D de suivi du mouvement (Optotrak). Quatre diodes étaient placées sur leurs lèvres (une à chaque commissure, une au milieu de la lèvre supérieure et une au milieu de la lèvre inférieure). Trois autres diodes étaient placées sur leur main droite (2 sur l'index : une sur l'ongle et une sur la première phalange et une sur le dos de la main). Les productions vocales des locuteurs étaient enregistrées grâce à un microphone.

2.4. Mesures

Toutes les productions acoustiques ont été validées pour vérifier que les participants avaient bien produit la focalisation sur l'élément souhaité. Cette validation acoustique a été réalisée en vérifiant que les corrélats acoustiques de la focalisation prosodique (*e.g.* 14) avaient bien été produits. Les erreurs de production ont été exclues de l'analyse. Les frontières acoustiques des syllabes ont été étiquetées à l'aide du logiciel Praat [15]. Les maxima de fréquence fondamentale (F_0) et d'intensité (Int) correspondant à l'élément focalisé (S ou O) ont été également étiquetés. La durée (Dur) de l'élément focalisé a été calculée. L'ouverture des lèvres et la protrusion de la lèvre supérieure ont été extraites des données de suivi du mouvement (Optotrak). Les cibles vocaliques (CV_1, CV_2) correspondant à chacune des voyelles des syllabes de l'élément focalisé (2 voyelles) ont ainsi pu être annotées (maxima d'ouverture des lèvres ou de protrusion). Concernant le mouvement du doigt, l'apex (P_A) et le début du geste de retour (l'index repart de sa position d'apex pour revenir vers la position de repos ; P_R) ont été annotés. Pour le geste de pointage, l'apex correspond à la position la plus étendue de l'index vers la cible. Pour le geste de battement, l'apex a été identifié comme étant le point vertical le plus bas du mouvement. Pour le geste de contrôle, l'apex a été identifié comme étant le moment où l'index appuie sur le bouton. Les instants de réalisation de chacun de ces évènements (maxima de F_0 et d'intensité, P_A et P_R) ont été normalisés sur la durée totale de l'énoncé afin d'éliminer la variabilité due aux différences segmentales des énoncés ou celle liée aux temps de réponse variables.

Toutes les variables dépendantes ont été testées en utilisant des ANOVAs à mesures répétées avec deux facteurs intra-participants : type de focalisation (2 niveaux : FS et OF) et condition gestuelle (pour les variables gestuelles *i.e.* P_A et P_R : 3 niveaux : pointage, battement et contrôle ; pour les variables acoustiques et articulatoires *i.e.* F_0, Int, Dur, CV_1 et CV_2 : 4 niveaux : parole seule + 3 types de gestes).

3. Résultats

3.1. Timings : Coordination temporelle parole/geste

Résultats généraux La Table 1 donne les résultats des analyses statistiques sur les instants d'occurrence de P_A, P_R, F_0, Int, CV_1, CV_2 (resp. t_{P_A}, t_{P_R}, t_{F_0}, t_{Int}, t_{CV_1}, t_{CV_2}).

Table 1: Résultats des ANOVAs sur les instants d'occurrence des évènements acoustiques, articulatoires et gestuels

	Condition Focalisation	**Condition Geste**
t_{P_A}	$F(1,9) = 114.4$, $p < .001$	$F(2,18) = 24.3$, $p < .001$
t_{P_R}	$F(1,9) = 99.5$, $p < .001$	$F(2,18) = 0.6$, $p = .55$
t_{F_0}	$F(1,18) = 1571.6$, $p < .001$	$F(3,27) = 1.6$, $p = .21$
t_{Int}	$F(1,18) = 2478.6$, $p < .001$	$F(3,27) = 5.6$, $p = .01$
t_{VT_1}	$F(1,18) = 3746.1$, $p < .001$	$F(3,27) = 1.1$, $p = .36$
t_{VT_2}	$F(1,18) = 2655.7$, $p < .001$	$F(3,27) = 2.2$, $p = .11$

Étude du geste manuel — Le *type de focalisation* a un effet significatif sur les instants de réalisation de P_A. Le geste a tendance à être réalisé plus tard au sein de l'énoncé quand c'est l'objet qui est focalisé (FO) : on peut dire que la focalisation attire le geste. La *condition gestuelle* a également un effet significatif sur l'instant de réalisation de P_A mais pas sur celui

de P_R. Ceci suggère que les différents types de gestes ne sont pas réalisés de la même façon (en tout cas en ce qui concerne leur apex).

Étude de la parole — Le *type de focalisation* a un effet significatif sur toutes les variables acoustiques et articulatoires. Ceci correspond au fait que les corrélats acoustiques et articulatoires ont été mesurés sur S pour FS et sur O pour FO. Ils arrivent donc forcément plus tard en condition FO. La *condition gestuelle* n'a d'effet significatif sur aucune des variables considérées. La production d'un geste n'a donc aucun effet sur l'organisation temporelle interne de l'énoncé.

Alignements temporels On peut dire que deux points sont alignés dans le temps si la différence entre leurs instants de réalisation est proche de 0. Dans le but d'étudier l'alignement potentiel des gestes manuels avec la parole, nous avons calculé, pour chaque énoncé, les différences entre les instants de réalisation des évènements gestuels (t_{P_A} et t_{P_R}) et les instants de réalisation des évènements acoustiques (t_{F_0} et t_{Int}) et articulatoires (t_{CV_1} et t_{CV_2}). Nous avons ensuite calculé la moyenne de ces différences sur tous les énoncés pour chaque participant. La Figure 1 donne les résultats sur le calcul des moyennes et écart-types sur tous les participants (si une boîte est proche de zéro, les deux évènements sont proches dans le temps).

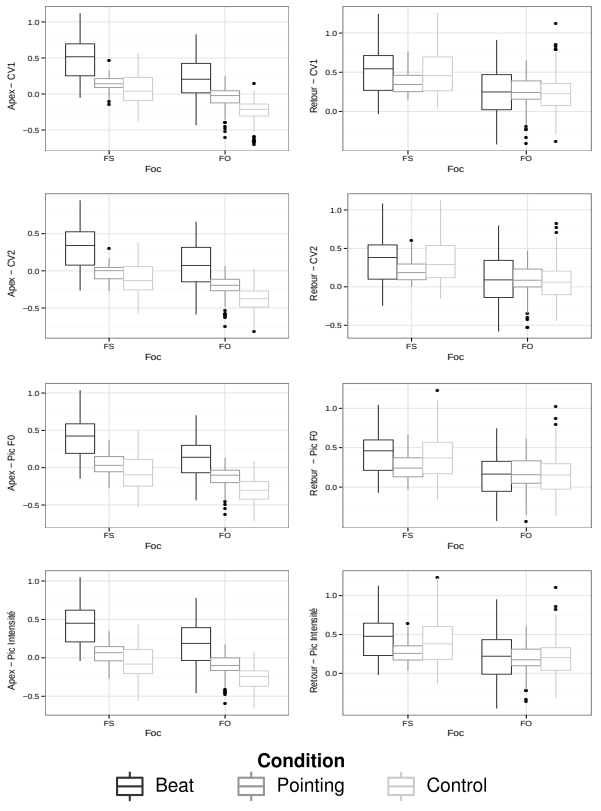


Figure 1: *Alignements temporels entre les évènements gestuels (P_A et P_R) et les évènements acoustiques (F_0, Int) et articulatoires (CV_1, CV_2) pour toutes les conditions de focalisation (FS et FO) et tous les types de geste (pointage, battement, contrôle). Données temporelles normalisées (c.f. section 2.4)*

La Table 2 donne les résultats des ANOVAs sur les alignements temporels. La *condition de focalisation* a un effet significatif sur toutes les différences temporelles. Quel que soit le type de geste, la coordination temporelle entre la parole et les gestes manuels au sein de l'élément focalisé est différente pour FS et FO. La *condition gestuelle* a également un effet significatif sur toutes les différences temporelles pour P_A mais pas pour P_R. Les différents gestes manuels sont donc coordonnés à la focalisation prosodique de façons différentes : P_R se trouve à peu près au même endroit relativement aux corrélats acoustiques et articulatoires mais pas P_A.

Table 2: Résultats des ANOVAs sur les différences temporelles

	Condition Focalisation	**Condition Geste**
t_{P_A} - t_{F_0}	$F(1,9) = 47.8,\ p < .001$	$F(2,18) = 25.1,\ p < .001$
t_{P_A} - t_{Int}	$F(1,9) = 55.3,\ p < .001$	$F(2,18) = 25.0,\ p < .001$
t_{P_A} - t_{VT_1}	$F(1,9) = 57.5,\ p < .001$	$F(2,18) = 24.4,\ p < .001$
t_{P_A} - t_{VT_2}	$F(1,9) = 55.6,\ p < .001$	$F(2,18) = 24.4,\ p < .001$
t_{P_R} - t_{F_0}	$F(1,9) = 32.8,\ p < .001$	$F(2,18) = 0.69,\ p = .51$
t_{P_R} - t_{Int}	$F(1,9) = 32.9,\ p < .001$	$F(2,18) = 0.95,\ p = .40$
t_{P_R} - t_{VT_1}	$F(1,9) = 38.6,\ p < .001$	$F(2,18) = 0.61,\ p = .55$
t_{P_R} - t_{VT_2}	$F(1,9) = 36.3,\ p < .001$	$F(2,18) = 0.64,\ p = .53$

Des tests t (Welch) ont ensuite été menés pour comparer les instants normalisés d'occurrence des évènements gestuels (P_A et P_R) aux évènements acoustiques (F_0 et Int) et articulatoires (CV_1 et CV_2). Ces comparaisons ont été effectuées pour chaque type de geste séparément (puisqu'il y a un effet de la condition gestuelle) et pour chaque type de focalisation séparément (puisqu'il y a un effet du type de focalisation).

Geste de pointage — Les tests ont montré que t_{P_A} n'était pas significativement différent de t_{F_0} (t(9)=1 ; p=0,3) et de t_{Int} (t(9)=1,5 ; p=0,2) pour FS mais pas pour FO (t_{F_0} : t(9)=-3,11 ; p=0,008 ; t_{Int} : t(9)=-2,7 ; p=0,02). Pour FS, t_{P_A} n'est pas significativement différent de t_{CV_2} (t(9)=-0,8 ; p=0,4) et pour FO, t_{P_A} n'est pas significativement différent de t_{CV_1} (t(9)=-1,5 ; p=0,2). Il apparaît donc que pour le pointage, il y a alignement entre l'apex du geste et une cible articulatoire de l'élément focalisé.

Geste de battement — Pour FS, il n'y a aucun alignement entre P_A et un des corrélats acoustiques ou articulatoires de la focalisation. Pour FO, P_A semble être aligné avec les maxima de F_0 et d'intensité (F_0 : t(9)=1,4 ; p=0,2 ; Int : t(9)=-0,8 ; p=0,4) ainsi qu'avec CV_2 (t(9)=0,7 ; p=0,5).

Geste de contrôle — Pour FS, P_A semble être aligné avec les maxima de F_0 et d'intensité (F_0 : t(9)=-1,2 ; p=0,3 ; Int : t(9)=-0,8 ; p=0,4). Ceci n'est cependant pas le cas pour FO. Pour FS, P_A est aligné avec CV_1 (t(9)=0,8 ; p=0,4). Pour FO, P_R est aligné avec CV_2 (t(9)=1,2 ; p=0,3). De façon très intéressante, la Figure 1 montre aussi que les alignements sont plus précis pour le geste de pointage (voir les écart-types). Ceci est d'autant plus vrai si on regarde les données concernant l'apex.

3.2. Réalisations acoustiques et articulatoires de la focalisation : effets de la production d'un geste manuel et du type de geste

Nous avons analysé les amplitudes des corrélats acoustiques (durée de l'élément focalisé et maxima de F_0

et d'intensité) et articulatoires (CV_1 et CV_2). Le *type de focalisation* a un effet significatif sur toutes les variables. Des analyses post-hoc révèlent que les amplitudes de tous les corrélats acoustiques et articulatoires de la focalisation sont plus faibles en FO qu'en FS ce qui n'est pas surprenant[14]. De façon plus intéressante, la *condition gestuelle* n'a d'effet significatif sur aucune des variables : la production d'un geste n'affecte ni les corrélats acoustiques ni les corrélats articulatoires de la focalisation prosodique.

Table 3: Résultats des ANOVAs sur les corrélats acoustiques et articulatoires de la focalisation

	Condition Focalisation	**Condition Geste**
Dur	$F(1,9) = 13.5$, $p = .05$	$F(3,9) = .4$, $p = .7$
F_0	$F(1,9) = 17$, $p < .01$	$F(3,9) = 2.7$, $p = .1$
Int	$F(1,9) = 76.2$, $p < .001$	$F(3,9) = .5$, $p = .2$
CV_1	$F(1,9) = 13.5$, $p < .01$	$F(3,9) = 2.3$, $p = .2$
CV_2	$F(1,9) = 59.6$, $p < .001$	$F(3,9) = 3.4$, $p = .08$

4. Conclusions et discussion

Les résultats de cette étude montrent que la focalisation et les gestes manuels sont coordonnés en ce sens que la focalisation "attire" le geste manuel. Les apex des gestes sont en effet soit inclus dans l'élément focalisé soit très proches. Ceci était prévisible pour le geste de pointage puisque les pointages manuel et vocal avaient le même objet qui était désigné au niveau vocal par la focalisation et au niveau gestuel par le pointage. Nous retrouvons donc bien là, dans la lignée des résultats de Rochet-Capellan et collègues [12], que la partie du geste manuel qui montre (index étendu vers la cible) et la partie de la parole qui montre (focalisation) se chevauchent. On trouve de plus que la coordination se fait essentiellement en alignant l'apex du geste avec une cible plutôt articulatoire qu'acoustique (toujours de façon cohérente avec Rochet-Capellan *et al.* [12]). L'attraction du geste manuel par la focalisation était également prévisible pour le geste de battement puisque plusieurs études ont déjà montré que le geste de battement était lié à l'emphase dans le discours [5]. Par contre, nous ne nous attendions pas à un tel résultat pour le geste de contrôle pour lequel nous pensions qu'il n'y aurait aucune coordination particulière entre le geste manuel et la focalisation. Il est possible que le geste de contrôle ait été mal choisi en ce sens qu'il est peut-être trop proche d'un geste de pointage (extension du bras et de l'index nécessaires pour aller appuyer sur le bouton). L'analyse des écart-types montre cependant que la coordination est beaucoup plus stricte pour le pointage ce qui va dans le sens de nos prévisions. Il apparaît donc que le lien fonctionnel entre geste et parole a une grande influence sur leur coordination. C'est en effet pour le geste de pointage que ce lien est le plus fort et la coordination la moins stricte est observée pour le geste de contrôle qui est un geste non-communicatif.

Les résultats montrent aussi que la production d'un geste manuel — quelque soit son type — n'a aucun effet sur les corrélats acoustiques et articulatoires de la focalisation prosodique. Il n'y a aucune différence entre la condition parole seule et les conditions parole+geste et aucune différence non plus pour les différents types de gestes. Ces résultats ne sont pas dans la continuité de ceux de Krahmer & Swerts [13] qui avaient trouvé que la production d'un geste de battement avait apparemment pour effet d'augmenter l'activité musculaire liée à l'articulation. En fait, il est possible que les résultats de ces auteurs soient un artefact de leur protocole expérimental. Les participants devaient en effet parfois produire un geste de battement sur un mot différent du mot accentué (condition d'incongruence). Ce type de production n'est pas naturel et il est possible que les locuteurs aient tout simplement eu tendance à produire un accent aussi sur le mot concomitant à leur geste donnant ainsi l'impression d'une augmentation de certains paramètres acoustiques en fait liés à la production pure et simple d'un accent qui n'existait pas dans la condition sans geste.

Notons enfin que pendant le déroulement des expériences, nous avons pu constater que bien que les gestes de battement soient produits très fréquemment dans la communication parlée, il est très difficile de les faire produire à un locuteur "sur commande". Il serait ainsi crucial de tenter d'effectuer ces mesures dans des conditions plus naturelles.

Références

[1] D. Mcneill, *Hand and Mind : What Gestures Reveal about Thought.* University Of Chicago Press, 1992.

[2] A. Kendon, *Gesture : Visible Action as Utterance.* Cambridge University Press, October 2004.

[3] D. Bolinger, "Intonation and gesture," *American Speech*, vol. 58, pp. 156–174, 1983.

[4] S. Nobe, "Representational gestures, cognitive rhythms, and acoustic aspects of speech : A network/threshold model of gesture production," Ph.D. dissertation, The Faculty of the Division of the Social Sciences, 1996.

[5] E. Mcclave, "Pitch and manual gestures," *Journal of Psycholinguistic Research*, vol. 27, no. 1, pp. 69–89, 1998.

[6] J. Boyer, A. Di Cristo, and I. Guaïtella, "Rôle de la voix et des gestes dans la focalisation," in *Oralité et gestualité. Interaction et comportements multimodaux dans la communication*, C. Cavé, I. Guaïtella, and S. Santi, Eds. L'Harmattan, 2001, pp. 459–463.

[7] L. Pietrosemoli, E. Mora, and M. A. Blondet, "Synchronisation des mouvements des mains et de la ligne de fréquence fondamentale en espagnol parlé," in *Oralité et gestualité. Interaction et comportements multimodaux dans la communication*, C. Cavé, I. Guaïtella, and S. Santi, Eds. L'Harmattan, 2001, pp. 492–495.

[8] D. P. Loehr, "Gesture and intonation," Ph.D. dissertation, Faculty of the Graduate School of Arts and Sciences of Georgetown University, 2004.

[9] S. Duncan, "Gesture and speech prosody in relation to structural and affective dimensions of natural discourse," in *GESPIN - Gesture & Speech in Interaction*, 2009.

[10] S. Kita and A. Özyürek, "What does cross-linguistic variation in semantic coordination of speech and gesture reveal ? : Evidence for an interface representation of spatial thinking and speaking," *Journal of Memory and Language*, vol. 48, no. 1, pp. 16–32, January 2003.

[11] J. P. de Ruiter, "Gesture and speech production," Ph.D. dissertation, Catholic University of Nijmegen, Netherlands, 1998.

[12] A. Rochet-Capellan, R. Laboissière, A. Galvan, and J.-L. Schwartz, "The speech focus position effect on jaw-finger coordination in a pointing task," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 1507–1521, December 2008.

[13] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence : Acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, 2007.

[14] M. Dohen and H. Lœvenbruck, "Identification des corrélats visibles de la focalisation contrastive en français," in *Proceedings of XXVe Journées d'Etudes sur la Parole*, April 19–22 2004, pp. 185–188.

[15] P. Boersma and D. Weenink, "Praat : doing phonetics by computer," 1995-2009. [Online]. Available : www.praat.org

Interactions audio-tactiles et perception de la parole : Comparaisons entre sujets aveugles et voyants

Christian Cavé[1], Marc Sato[2], Lucie Ménard[3], Annie Brasseur[3]

[1] Laboratoire Parole & Langage, CNRS & Aix-Marseille Université, France
[2] GIPSA-Lab, Département Parole & Cognition, CNRS & Grenoble Universités
[3] Département de Linguistique, Université du Québec à Montréal, Canada

Email : christian.cave@lpl-aix.fr

ABSTRACT

The present study investigated whether tactile information obtained manually by touching the speaker's face modulates the decoding of speech. Audio-tactile perception was compared to audio-only perception, and a group of congenitally blind adults was compared to a group of sighted adults. Participants performed a phonemic decision task in three conditions: audio-only, congruent audio-tactile, and incongruent audio-tactile. For the auditory modality, the phoneme sequences were presented either with a background white noise or without noise. The results showed that tactile information relevant to recovering speech gestures improved auditory speech perception in cases of degraded acoustic information. Moreover, the same audio-tactile interactions were found for the blind and sighted listeners, despite possible differences in these groups' sensory skills.

Keywords: speech perception; multimodality; audio-tactile interactions; blindness.

1. INTRODUCTION

On sait depuis longtemps que des informations visuelles modifient le traitement de la parole par exemple en améliorant l'intelligibilité de parole présentée dans le bruit si le visage du locuteur est visible [1-2]. L'effet McGurk [3] est une autre manifestation de l'influence d'informations visuelles sur le décodage de la parole qui a été étudiée pour de nombreuses langues et divers types de sujets, normo entendants, malentendants, avec troubles de parole, etc. (voir [4] pour une synthèse).

On sait aussi depuis longtemps, par la méthode Tadoma [5] utilisée par des sujets sourds et aveugles, que des informations tactiles (perception haptique), obtenues en plaçant une main sur le visage du locuteur, permettent de récupérer les gestes de production de la parole (voir la partie « Méthode » pour détails). Des utilisateurs entraînés peuvent ainsi arriver à un niveau de communication quasi-normal.

Quelques études ont montré que des interactions audio-tactiles pour la parole existaient chez les sujets normaux non entraînés. Ainsi Fowler et Dekle [6] ont présenté à des sujets naïfs non entraînés des syllabes acoustiques couplées avec le contact de la main sur le visage du locuteur. Ils mettent ainsi en évidence une forte interaction entre les modalités auditive et tactile : l'information tactile influence le décodage de la syllabe auditive et réciproquement, la syllabe auditive le décodage de la syllabe perçue tactilement. De plus, la présentation de paires de syllabes auditive-tactile non cohérentes, produit chez certains sujets un percept illusoire (fusion ou combinaison) de type McGurk. Il faut toutefois noter qu'il y a de fortes différences interindividuelles et que la plupart des sujets ne perçoivent pas de réponses illusoires. Plus récemment, Gick et collègues [7] ont montré que l'information tactile manuelle améliorait la perception de la parole tant visuelle qu'auditive chez des sujets non entraînés. Dans leur étude, les participants devaient percevoir des disyllabes présentés visuellement ou auditivement dans un bruit blanc continu, avec la main placée sur le visage du locuteur comme dans la méthode Tadoma. La comparaison entre la condition bimodale et les conditions unimodales montrent que l'information tactile améliore la perception de la parole d'environ 10%. Pris ensemble, ces résultats soulèvent d'importantes questions sur les relations entre la modalité auditive et les autres modalités sensorielles et sur un possible couplage fonctionnel entre systèmes de perception et de production de la parole [8-10].

Pour cette étude, nous avons utilisé la méthode Tadoma pour évaluer l'interaction entre information tactile et information auditive –cohérente ou non cohérente– lors de la perception de la parole en comparant une situation audio-tactile avec une situation audio seul ou tactile seul. Pour cela, il a été demandé aux sujets de réaliser une tâche d'identification phonémique dans trois conditions de présentation: 'audio seul', 'audio-tactile cohérent' ou 'audio-tactile non cohérent'. De plus, une condition 'tactile seul' avait pour but de déterminer dans quelle mesure la seule information tactile permettait d'identifier l'item articulé silencieusement. De façon à évaluer si le niveau de sensibilité tactile peut jouer sur la récupération des gestes de production de la parole et moduler les interactions inter modales, nous avons comparé deux groupes de dix sujets aveugles congénitaux et de dix sujets à vision normale.

Compte tenu des données ci-dessus, nous faisons l'hypothèse qu'une information tactile cohérente devrait améliorer la perception de la parole dans le bruit. Pour la condition audio-tactile, la perception d'un geste articulatoire non cohérent pourrait diminuer

l'intelligibilité de la parole ou même modifier l'expérience auditive du sujet comme lors de l'effet McGurk. La mise en évidence de percepts illusoires de type McGurk lors d'une stimulation audio-tactile non cohérente fournirait de plus un argument en faveur d'une véritable intégration bimodale audio-tactile.

2. MÉTHODE

Participants

Un groupe de dix sujets aveugles congénitaux (moyenne d'âge : 41 ans) et un groupe de dix sujets ayant une vision normale (moyenne d'âge : 27 ans) ont participé à l'étude. Les participants aveugles avaient une cécité de niveau 3, 4 ou 5 selon l'échelle de l'Organisation mondiale de la santé. Tous les participants étaient droitiers, n'avaient pas de troubles d'audition ou de langage connus, avaient le français québécois comme langue maternelle et n'avaient jamais participé à une expérience de même type. Tous ont signé un consentement et ont été payés pour leur participation.

Stimuli

Une locutrice dont la langue maternelle est le français québécois a été enregistrée (44,1 kHz, 16bits) et filmée (30 trames/s ; 720 x 480 pixels) de face sur un fond gris clair. Elle a produit, à plusieurs reprises, les séquences phonémiques /aba/ et /aga/ en maintenant une intensité normale, une intonation neutre et en fermant complètement les lèvres entre chaque réalisation. Une réalisation articulatoire unique de chaque item a été sélectionnée pour produire les 14 stimuli utilisés pour l'expérimentation. Ces stimuli permettent de présenter simultanément la piste audio aux sujets et la piste vidéo à l'expérimentateur (voir Procédure). Pour chaque item, il y avait 7 conditions de présentation :

- tactile seul
- audio seul avec ou sans bruit blanc
- audio-tactile cohérent avec ou sans bruit blanc
- audio-tactile non cohérent avec ou sans bruit blanc

Procédure expérimentale

La procédure expérimentale est inspirée de celle de Fowler et Dekle (1991). Les participants étaient assis face à l'expérimentatrice (AB). Dans la modalité tactile, leur main droite était placée sur son visage de façon à ce que le pouce soit vertical et frôle les lèvres et que les autres doigts soient placés horizontalement sur la mâchoire. Cette position permettait de capter les mouvements des lèvres et de la mandibule lors de la production silencieuse des items /aba/ et /aga/. Pour éviter que les sujets voyants ne regardent l'expérimentatrice, ils portaient un bandeau sur les yeux pendant toute la durée de l'expérience. Pour la modalité auditive, les items /aba/ ou /aga/ étaient présentées par écouteurs mixés ou non avec un bruit blanc à un niveau acoustique confortable.

L'expérimentatrice était assise face au sujet et à un écran d'ordinateur. À chaque essai, l'écran lui indiquait la séquence à prononcer et affichait les mouvements visuels des articulateurs à produire, ce qui lui permettait d'articuler silencieusement en synchronie avec l'item acoustique présenté au sujet. Avant l'expérimentation, elle était entraînée à articuler silencieusement en synchronie avec les mouvements affichés sur son écran. À aucun moment, pendant le déroulement d'une session, elle n'était au courant de la syllabe présentée de manière acoustique au sujet.

Avant l'expérience, les participants étaient informés qu'il leur serait présenté des items soit auditivement par écouteurs, soit tactilement par contact entre leur main et le visage de l'expérimentatrice, soit par les deux modalités en même temps. Leur tâche était de dire, à chaque essai, s'ils avaient entendu /aba/, /ada/, /aga/ ou n'importe quelle combinaison syllabique (par exemple, /agba/, /abga/ ou /agda/ par exemple). Ils ne recevaient aucune information particulière sur la façon d'interpréter l'information tactile, ni aucune autre consigne. L'expérimentatrice démarrait chaque essai en appuyant sur une touche du clavier tout en ayant la bouche fermée en position neutre. Avant la présentation de l'item auditif au sujet, les instructions relatives à l'articulation à réaliser (i.e., « aba », « aga » ou « ### ») s'affichaient sur son écran pendant 500 ms. L'item acoustique pour le sujet et le visage articulant ou non pour l'expérimentatrice étaient alors présentés simultanément. Après chaque présentation, un autre expérimentateur notait la réponse du sujet. De plus, la totalité de la session expérimentale était filmée pour vérification ultérieure si besoin. Chacun des 14 stimuli était présenté 6 fois dans un ordre aléatoire. La procédure expérimentale étant assez contraignante tant pour l'expérimentatrice que pour les participants, ils étaient autorisés à faire de courtes pauses à n'importe quel moment pendant la session expérimentale qui durait environ 20 minutes.

Analyse des résultats

L'ensemble des réponses recueillies a été analysé pour chaque participant et chaque condition. Exceptée pour la condition 'tactile seul', les données ont été traitées par une analyse de variance (Anova) avec pour variable inter-sujets le groupe (voyants, non-voyants) et pour variables intra-sujets la séquence phonémique présentée de manière acoustique (/aba/, /aga/), la production articulatoire silencieuse (/aba/, /aga/, pas de production) et la présence ou non de bruit blanc. Par convention, dans les conditions audio-tactiles non cohérentes, les termes 'réponses correctes' correspondent aux réponses basées sur la modalité auditive. Du fait de l'absence de bruit blanc associé, les données issues de la condition 'tactile seul' ont été traitées séparément des autres conditions par une analyse de variance (ANOVA) avec pour variable inter-sujets le groupe (voyants, non-voyants) et pour variable intra-sujets la syllabe articulée silencieusement (/aba/, /aga/). Pour ces analyses, le niveau de significativité a été fixé à $p < .05$, un test de Mauchly a été effectué de manière à vérifier l'hypothèse de sphéricité des données, enfin des tests de Newman-Keuls ont été utilisés pour les analyses post-hoc.

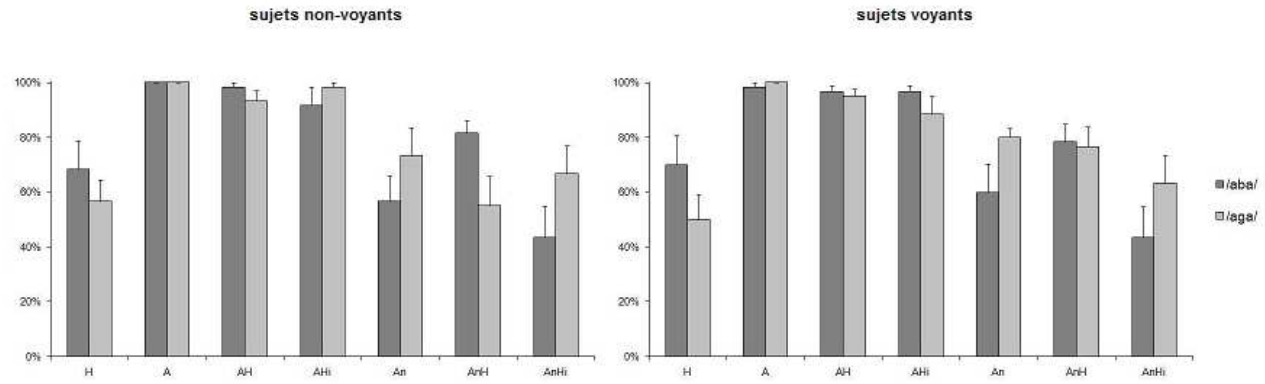


Figure 1. Pourcentage de réponses correctes des participants non-voyants et voyants en fonction des différentes conditions expérimentales (H : tactile seul ; A : audio seul, AH : audio-tactile cohérent, AHi : audio-tactile non cohérent, n : présence de bruit blanc).

3. RÉSULTATS

Le pourcentage de réponses correctes est indiqué dans la Figure 1 pour les deux groupes de sujets.

Condition tactile seul

La syllabe articulée a un effet significatif ($F(1,18) = 5,63$; $p=.03$) sur les réponses. On observe plus de réponses correctes pour /aba/ que pour /aga/ aussi bien pour les aveugles (68% vs 57%) que pour les voyants (70% vs 50%) et il n'y a pas de différence significative entre les deux groupes de sujets ($F(1,18) < 1$).

Condition audio et audio-tactile

Il y a un effet très significatif du bruit ($F(1,18) = 52,07$; $p < .0001$). On observe un plus grand nombre de réponses correctes pour les items présentés sans bruit (96%) que pour les items bruités (65%).

L'item articulé a aussi un effet significatif ($F(2,36) =5,57$; $p < .008$) ; l'item /aba/ donne lieu à un plus grand nombre de réponses correctes que l'item /aga/ (82% vs 76% ; $p < .02$).

En l'absence de bruit, aussi bien /aba/ que /aga/ sont identifiés de façon quasi parfaite (globalement 99% vs 98%) quelle que soit la condition et il n'y a pas de différence significative entre les conditions.

Par contre, en présence de bruit, les conditions de présentation ont un effet significatif (toutes les comparaisons significatives à au moins $p <.03$) sur le nombre de réponses correctes (voir tableau 1). Ainsi, pour la syllabe /aba/ perçue de manière acoustique en présence de bruit, on observe un nombre de réponses correctes supérieur dans le cas de la condition audio-tactile cohérent par rapport à la condition audio seul (80% vs. 58%, $p < .003$) et la condition audio-tactile non cohérent (80% vs. 43%, $p < .0001$), et dans le cas de la condition audio seul par rapport à la condition audio tactile incohérent (58% vs. 43%, $p < .008$). Pour l'item /aga/, on observe un nombre supérieur de réponses correctes pour les conditions audio tactile cohérent (72% vs. 59%, $p < .03$) et audio seul (77% vs. 59%, $p < .007$) par rapport à la condition audio tactile non cohérent.

Il est important de noter qu'il n'y a pas d'effet principal du groupe ni d'interaction entre la variable groupe et les autres variables.

Tableau 1. Pourcentage de réponses correctes en présence de bruit en fonction des conditions de présentation

	/aba/	/aga/
Audio-tactile cohérent	80%	72%
Audio seul	58%	77%
Audio-tactile non cohérent	43%	59%

4. DISCUSSION

Nos résultats indiquent clairement que l'information tactile provenant des gestes articulatoires telle qu'elle a été traitée par les participants peut moduler la perception de la parole dans des conditions acoustiques dégradées. En effet, on constate que le contact manuel avec le visage de la locutrice associé à une entrée auditive non cohérente diminue l'intelligibilité de la parole non seulement par rapport à la modalité audio seul mais aussi par rapport à la modalité tactile seul, révélant ainsi une certaine interaction entre les informations tactiles et les informations auditives. De même, une augmentation de l'intelligibilité est observée lors de la condition 'audio-tactile cohérent' par rapport aux modalités audio seul et tactile seul.

Toutefois nos résultats ne mettent pas en évidence d'illusions de type McGurk induites par nos stimulus audio-tactiles non cohérents. En effet, le pourcentage de réponses /ada/ ainsi que le pourcentage de réponses de type combinaison n'étaient pas significativement différents entre les conditions audio seul et audio-tactile. Notons cependant que la procédure utilisée ne permet pas

de contrôler l'information tactile perçue par les participants et qu'il en est de même pour les travaux cités [6], [7]. Ce fait pourrait jouer un rôle dans les différences inter- et intra-individuelles constatées. Toutefois, il ne peut en rien expliquer l'augmentation de l'intelligibilité dans la situation audio-tactile cohérent avec bruit.

Nos résultats ne révèlent aucune différence perceptive entre sujets voyants et sujets aveugles. Ce, notamment pour les situations unimodales tactile ou auditive, bien qu'une sensibilité tactile ou auditive supérieure par rapport aux sujets voyants ait été décrite chez des aveugles congénitaux [11-13]. Toutefois, on sait que même des sujets sourds-aveugles entrainés à la méthode Tadoma n'ont pas de meilleurs scores que des sujets standards non entraînés pour l'identification de syllabes sans signification [14, 15]. Il n'est donc pas surprenant de ne pas trouver de différences entre nos deux groupes de sujets pour les situations bimodales, qu'elles soient cohérentes ou non cohérentes.

Plusieurs sujets ont assez rapidement détecté que, dans certains essais, les informations tactiles et auditives n'étaient pas cohérentes. Il semble bien que cela ne se produise pas pour les non cohérences audiovisuelles donnant lieu à l'effet McGurk. Cela pourrait indiquer que les phénomènes d'intégration entre modalité tactile et modalité auditive ne sont pas de même nature ou du moins n'ont pas la même force que ceux intervenant entre l'audition et la vision.

En dépit de l'absence globale d'effet de type McGurk pour la condition audio-tactile non cohérente, certains sujets ont donné des réponses de type fusion /ada/ et de type combinaison telles que /abda/, /abga/, /adga/, ce qui est classique, ou même /adada/, /ababa/, /ania/ et /abia/. Alors que dans notre étude, plusieurs sujets n'ont donné aucune réponse de type fusion ou combinaison, le sujet le plus productif en a donné 32%. Ces observations vont dans le sens de celles de Fowler et Dekle [6] qui rapportent l'existence de percepts illusoires de type McGurk pour des stimulations audio-tactiles non cohérentes, en précisant toutefois qu'il existe de très fortes différences interindividuelles.

5. CONCLUSION

En conclusion, ces résultats démontrent : 1) que lors du processus de traitement de la parole, une information tactile sur les mouvements des articulateurs peut influencer le décodage de l'information auditive présentée simultanément et 2) que les interactions audio-tactiles ne sont pas différentes chez des participants non entraînés en dépit des possibles différences de capacités sensorielles entre sujets voyants et sujets aveugles.

BIBLIOGRAPHIE

[1] Sumby, W.H. and Pollack, I. Visual contribution to speech intelligibility in noise. Journal of Acoustical Society of America, 26: 212-215, 1954.

[2] MacLeod, A. and Summerfield, Q. Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology, 21: 131-141, 1987.

[3] McGurk, H. and MacDonald, J. Hearing lips and seeing voices. Nature, 264: 746-748, 1976.

[4] Colin, C. and Radeau, M. Les illusions McGurk dans la parole : 25 ans de recherches. L'Année Psychologique, 103(3) : 497-542, 2003.

[5] Alcorn, S. The Tadoma method. Volta Rev., 34: 195–198, 1932.

[6] Fowler, C. and Dekle, D. Listening with eye and hand: Crossmodal contributions to speech perception. J. Exp. Psychol. Hum. Percept. Perform., 17: 816–828, 1991.

[7] Gick, B., Jóhannsdóttir, K.M., Gibraiel, D. and Mühlbauer, M. Tactile enhancement of auditory and visual speech perception in untrained perceivers, Journal of Acoustical Society of America, 123: 72–76, 2008.

[8] Schwartz, J.-L., Sato, M. and Fadiga, L. The common language of speech perception and action: a neurocognitive perspective. Revue Française de Linguistique Appliquée, 13(2): 9-22, 2008.

[9] Schwartz, J.-L., Ménard, L., Basirat, A. and Sato, M. The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. Journal of Neurolinguistics, sous presse.

[10] Sato, M., Tremblay, P. and Gracco, V. A mediating role of the premotor cortex in phoneme segmentation. Brain and Language, 111(1): 1-7, 2009.

[11] Röder, B. and Neville, H.J. Developmental functional plasticity. In Grafman J, Robertson I (Eds), Plasticity and rehabilitation. Handbook of Neuropsychology, vol 9. Elsevier, Amsterdam, pp. 231–270, 2003.

[12] Gougoux, F., Lepore, F., Lassonde, M., Voss, P., Zatorre, R. J., and Belin, P. Pitch discrimination in the early blind. Nature, 430: 309, 2004.

[13] Ménard, L., Dupont, S. Baum, S.R. and Aubin, J. Production and perception of French vowels by congenitally blind adults and sighted adults. Journal of the Acoustical Society of America, 126: 1406-1414, 2009.

[14] Norton, S. J., Schultz, M. C., Reed, C. M., Braida, L. D., Durlach, N. I., Rabinowitz, W. M. & Chomsky, C. (1977). Analytic study of the Tadoma method: Background and preliminary results. J. Speech Hear. Res., 20: 574–595.

[15] Reed, C. M., Rubin, S. I., Braida, L. D. & Durlach, N. I. (1978). Analytic study of the Tadoma method: discrimination ability of untrained observers. J Speech Hear Res, 21: 625-637.

Perception interculturelle des attitudes audio-visuelles vietnamiennes

Dang-Khoa Mac [1,2], Véronique Aubergé [1], Albert Rilliard [3], Eric Castelli [1,2]

[1] Laboratoire d'Informatique de Grenoble, CNRS, France
[2] Centre de recherche international MICA, CNRS-UMI 2954, Hanoi, Vietnam
[3] LIMSI-CNRS, Orsay, France
dang-khoa.mac@imag.fr

ABSTRACT

Social affective expression is a main part of face-to-face interaction and it is highly linked to the language through the culture. This paper presents a cross-cultural study on Audio-Visual prosodic attitudes in Vietnamese, an under-resourced tonal language. Based on an audio-visual corpus of 16 attitudes, perception experiments were carried out with Vietnamese and French participants. The result analysis shows the relative contribution of audio, visual, and audio-visual information in attitude perception. It also shows how native and non-native listeners recognize and confuse the attitudes, thus allows us to investigate the cultural specificities and cross-cultural common attitudes in Vietnamese.

Keywords: Audio-visual prosody, Social affects, Cross-cultural perception, Vietnamese

1. INTRODUCTION

La parole est le vecteur privilégié de la communication humaine qui, au delà des informations strictement linguistiques, construit le sens de la communication en la situant par son contexte, ses acteurs, par les connaissances exprimées par le locuteur sur ses états émotionnels, attitudinaux (affects sociaux), intentionnels, mentaux et cognitifs. Les affects sont exprimés dans la parole et leurs expressions sont contrôlées à différents niveaux cognitifs [1], depuis un contrôle involontaire des émotions (communiquées éventuellement pendant le temps langagier) jusqu'à un contrôle volontaire — dont relève les attitudes [2]. Ainsi les expressions attitudinales et émotionnelles sont différentes par la nature de leur contrôle (volontaire ou involontaire) et en conséquence par leur ancrage temporel, qui participe ou non à la construction des unités linguistiques [2]. Une même étiquette de valeur, par exemple la surprise, peut relever fonctionnellement et cognitivement de deux processus distincts : une émotion (événement non inhérent à la construction langagière) ou une attitude (fait de langage, pragmatique ou linguistique, tel n'est pas, à notre sens, le débat), et sera distinguable en tant que telle, non pas par sa forme acoustique mais par l'organisation temporelle de cette forme. Selon ce point de vue, les attitudes sont des codes, construits et acquis dans la dynamique du langage et de ses réalités socioculturelles. Par là même, elles véhiculent les choix et les valeurs du langage dans sa culture. Elles expriment, autour d'invariants linguistiques et sociologiques, sur les spécificités socioculturelles de la langue : les expressions inconnues (de valeurs et de formes) et les faux-amis qui apparaissent en croisant les langues (en apprentissage par exemple [3, 10]).

Le vietnamien est une langue peu dotée. Peu d'études se sont intéressées aux affects dans la parole [7]. La présente étude s'inspire de travaux sur des langues très éloignées linguistiquement et culturellement (français [9], anglais [8], japonais [3,4,10]). Le vietnamien est une langue tonale, qui utilise phonologiquement des variations de fréquence fondamentale et de qualité de voix (avec des changements dus à la source glottique). Cet usage de la qualité de voix rend le vietnamien intéressant, les usages les plus communs de la qualité de voix étant en général non linguistiques. En raison du contraste linguistique entre français et vietnamien (morpho-syntaxes, prosodies démarcatives, phonologie non-tonale vs tonale), de la distance géo-culturelle, le français a été choisi comme langue de croisement pour cette étude interculturelle de l'affect social vietnamien. Cette étude a été menée dans une réalité pseudo-écologique d'interactions, c'est-à-dire en modalité audio-visuelle, afin de comprendre [4] la contribution relative des informations auditives, visuelles et audio-visuelles dans la perception d'une attitude.

Après la présentation du corpus d'attitudes, l'expérience perceptive est décrite pour des sujets vietnamiens puis français. Les résultats montrent le rôle des modalités Audio (A) et Vidéo (V) et leur intégration. Ils soulignent en particulier des différences d'identification et de confusions entre sujets natifs et non-natifs.

2. CORPUS

2.1. Sélection des attitudes

Les attitudes prosodiques ont été étudiées avec une même approche pour différentes langues [3, 7, 8, 9]. Dans ces études le choix des valeurs d'attitudes repose sur des usages d'enseignement, à l'origine des catégories empiriques issues des propositions de Fónagy pour le français. Sur la base des attitudes proposées pour le japonais [3], nous avons défini 16 attitudes , c'est-à-dire 16 étiquettes associées à une définition très précise pour le vietnamien (table 1).

Table 1 : 16 attitudes vietnamiens, avec leurs abréviations

Déclaration	DEC	Irritation	IRR
Question-simple	INT	Ironie sarcastique	SAR
Exclamation de surprise neutre	EXo	Mépris	MEP
Exclamation de surprise positif	EXp	Politesse	POL
Exclamation de surprise negatif	EXn	Admiration	ADM
Evidence	EVI	Maternel	MAT
Doute-incrédulité	DOU	Séduction	SED
Autorité	AUT	Familier	FAM

2.2. Corpus

125 phrases squelettes sans signification affective intrinsèque ont été établies de manière à équilibrer les structures syntaxiques de base (contrôle de l'effet de prosodie syntaxique) et de contrôler les effets croisés des événements tonals et de coarticulation : le corpus contient 8 phrases de 1 syllabe, qui correspondent aux variations des 8 tons vietnamiens et 72 phrases de deux syllabes, qui correspondent à toutes les combinaisons de tons (notons que la plupart des mots vietnamiens sont mono- ou bi-syllabiques [5,7]). Le reste du corpus est basé sur 45 phrases à partir de 3 à 8 syllabes et variés dans leur structure syntaxique: mono-mot, groupe nominal isolé, groupe verbal isolé ou structure simple « sujet-verbe-objet », courante en vietnamien.

Un locuteur, originaire de Hanoi (prononciation standard du vietnamien), a été choisi pour enregistrer le corpus, après une longue phase d'entraînement. Le corpus a été enregistré dans une chambre sourde avec une caméra numérique DV et un électroglottographe pour mesurer directement les vibrations des cordes vocales (impliquées dans la qualité de voix tonale et attendues comme fondamentales dans les réalisations de certaines attitudes). Pour contrôler les performances du locuteur, un spécialiste des attitudes et un locuteur natif vietnamien ont observé le processus d'enregistrement de l'extérieur de la salle, grâce à un système vidéo. Le locuteur a prononcé les 125 énoncés dans les 16 attitudes. Le corpus complet contient donc 2000 stimuli. Il correspond à plus de 90 minutes de signal audio-visuel, après traitement.

3. EXPERIENCES PERCEPTIVES

Le test de perception était destiné à évaluer la contribution relative des facteurs suivants:

- les 16 valeurs et expressions d'attitudes
- la longueur de la phrase (en nombre de syllabes)
- les trois modalités (A, V, AV)
- l'ordre de présentation des modalités (A premier ou V premier)

Pour examiner l'influence de la longueur des phrases, trois phrases d'une, deux et cinq syllabes ont été choisies dans le corpus. Afin de limiter la complexité de ce test, l'influence du ton n'a pas été étudiée dans cette expérience qui contient uniquement des énoncés sans variation tonale (l'influence du ton sera étudiée ultérieurement).Toutes les syllabes sont basées sur le ton 1 (le ton plat). Ces phrases sont présentées dans les 16 attitudes et dans les trois modalités (A, V, AV). Le test de perception est constitué de 3 * 16 * 3 = 144 stimuli.

Quarante auditeurs ont participé à cette expérience: 20 Vietnamiens (10 hommes et 10 femmes d'un âge moyen de 25 ans), qui parlent le même dialecte que le locuteur, et 20 Français (10 hommes et 10 femmes avec un âge moyen de 35 ans) qui n'ont aucune expérience de la langue ni de la culture vietnamienne. Ces participants vietnamiens et français ont été séparés en deux groupes. Le premier groupe écoute les stimuli en audio seul d'abord, puis regarde les stimuli en vidéo seul, et enfin les stimuli audio-visuels. Le deuxième groupe a commencé avec les stimuli en vidéo seul, ensuite en audio seul et termine avec les stimuli audio-visuels. Les tests de perception ont été effectués dans une pièce calme. L'interface donnait l'étiquette et l'explication des 16 attitudes dans la langue maternelle de l'auditeur. Tous les sujets ont écouté (et/ou regardé) chaque stimulus une seule fois. Après chaque stimulus, on leur a demandé d'indiquer l'attitude supposée parmi les 16 attitudes et d'indiquer une intensité allant de « à peine perceptible » (codé comme 1) à « très forte » (codé 100). Un score de 0 a été attribué aux 15 autres attitudes non sélectionnées.

4. ANALYSE DES RESULTATS

4.1. Effets des facteurs

Afin de mesurer l'effet des facteurs ci-dessus, deux ANOVA à mesures répétées ont été calculées (une pour chaque groupe d'auditeurs). L'intensité moyenne des bonnes réponses a été choisie comme variable dépendante de l'analyse de variance. Les facteurs inter-sujets sont les 16 attitudes, la longueur (3 niveaux), et les modalités (3 niveaux). Le tableau 2 montre les résultats d'analyses de la variance.

Table 2 : Résultats d'ANOVA sur l'intensité moyenne. Des effets significatifs au niveau de 1% sont en gras. Att: attitude; Mod: Modalité; Ord: ordre de présentation des modalités; Len: longueur de la phrase

		Vietnamien		Français	
	df	F	p	F	p
Att	**15**	**47.804**	**0.000**	**33.100**	**.000**
Mod	**2**	**45.373**	**0.000**	**74.767**	**.000**
Ord	**1**	.022	0.882	.001	.975
Len	2	3.735	0.024	1.655	.191
Att*Mod	**30**	**6.096**	**0.000**	**9.104**	**.000**
Att*Ord	15	1.527	0.087	2.971	.000
Att*Len	**30**	**3.542**	**0.000**	**3.007**	**.000**
Mod*Ord	2	0.749	0.473	4.955	.007
Mod*Len	4	1.822	0.122	**6.061**	**.000**
Ord*Len	2	.238	0.788	.564	.569
Att*Mod*Ord	30	1.175	0.235	.872	.666
Att*Mod*Len	**60**	**2.104**	**0.000**	**1.721**	**.001**
Att*Ord*Len	30	0.806	0.763	1.138	.277
Mod*Ord*Len	4	0.547	0.701	.913	.455
Att*Mod*Ord*Len	60	0.644	0.985	1.122	.244

Les résultats de perception pour les 16 attitudes dans chaque modalité sont présentés dans la figure 1. Globalement, la plupart des attitudes sont reconnues au-dessus le niveau du hasard et les auditeurs natifs ont des scores supérieurs à ceux des français, sauf dans le cas de l'attitude « Admiration ».

Pour les groupes de sujets vietnamiens et français, l'attitude et la modalité ont un effet significatif sur la perception. Au contraire, la longueur de phrase et l'ordre de présentation ne montrent pas d'influence. L'interaction

entre attitude et modalité, entre la longueur de la phrase et l'ordre de présentation (pour les auditeurs français), et entre l'attitude, la longueur de phrase et les modalités, ont des effets significatifs sur la perception des attitudes.

La modalité a un effet important sur la perception de l'attitude. Comme prévu, pour la plupart des attitudes, le score moyen en modalité audio-visuelle est meilleur que celui en audio seul ou visuel seul. Pour les auditeurs vietnamiens, les informations audio joue un rôle important dans le cas des attitudes Déclaration, Evidence, Autorité et Familier. Pour les auditeurs français, l'audio est essentiellement informatif pour Autorité et Irritation (dans lesquels on peut déjà supposer acoustiquement le rôle de la qualité de voix). Les informations vidéo sont importantes dans le cas de la Surprise positive, du Mépris et de la Politesse pour les vietnamiens et les français.

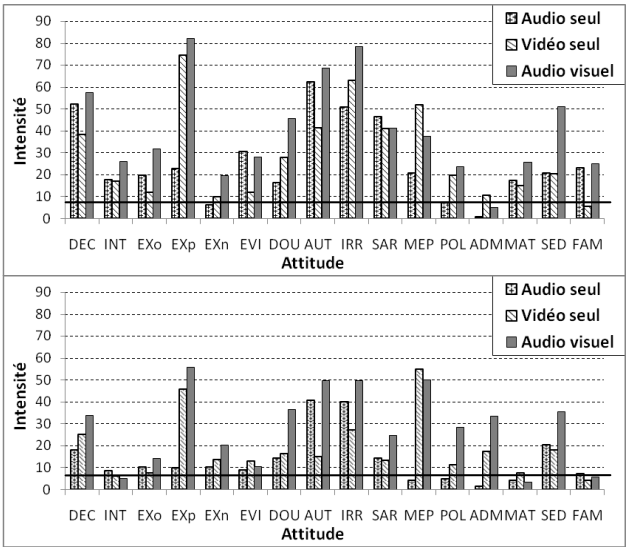


Figure 1 : Note moyenne d'intensité pour chaque attitude dans chaque modalité pour les auditeurs vietnamiens (en haut) et les français (en bas)

4.2. Classification des attitudes

Pour les groupes vietnamiens et français, les matrices de confusions sont analysées grâce à une méthode de classification hiérarchique, qui permet de regrouper les stimuli perçus selon la proximité des distributions des réponses pour chacune des attitudes proposées. La figure 2 présente la classification hiérarchique pour les auditeurs vietnamiens et français dans chacune des 3 modalités. En utilisant un seuil de 75 (proche de la moitié de la distance maximale obtenue), les 16 attitudes peuvent être répartis en groupes qui varient en fonction de la modalité considérée.

Dans la condition audio seul, le regroupement montre cinq groupes principaux pour les auditeurs vietnamiens et 3 groupes principaux pour les auditeurs français (ce qui confirme la plus forte confusion faite par des auditeurs non-natifs). Les auditeurs vietnamiens ont tendance à percevoir de manière similaire la Déclaration et la Politesse. Les auditeurs français confondent Autorité avec Irritation, et Séduction avec Familier.

Avec l'information de vidéo seule, les résultats pour les auditeurs vietnamiens donnent 5 clusters, plus EXp reconnu sans confusion. Les résultats pour les auditeurs français montrent 4 clusters, avec deux attitudes reconnues sans confusion. Les éléments de chaque groupe dans la condition de vidéo seule sont différents de ceux obtenus en audio seul. Pour les deux groupes vietnamiens et français, les attitudes SAR et MEP ont été regroupées dans un cluster d'expressions « impolies ». Contrairement au cas de l'audio seule. Avec les informations visuelles, les auditeurs vietnamiens confondent Autorité avec Irritation.

Comme prévu, la modalité AV permet une meilleure identification. Les vietnamiens regroupent les expressions en 5 clusters principaux et 5 attitudes sans confusion (EXp, POL, MAT, AUT et IRR). Le regroupement des auditeurs français donne 4 groupes et 4 attitudes bien reconnues (EXp, ADM, AUT et IRR). Pour les deux groupes vietnamiens et français, un cluster regroupant EXn avec DOU et un autre regroupant SED et EXo sont observés. SAR a été regroupée avec MEP avec les auditeurs vietnamiens. Dans le cas des auditeurs français, MAT a été ajouté à ce groupe. DEC et INT se trouvaient dans un groupe de deux attitudes pour les auditeurs vietnamiens, mais pour les auditeurs français, on y trouve 5 attitudes (POL, DEC, EXo, INT, EVI).

5. DISCUSSION

Selon les résultats, bien que le score d'intensité moyenne obtenue par les auditeurs français soit inférieur à celui des auditeurs vietnamiens, il est cohérent avec le résultat des auditeurs vietnamiens. Pour les deux groupes d'auditeurs, certaines attitudes sont bien reconnues: DEC, EXp, DOU, AUT, TRI et SED. Ceci suppose que les concepts et les réalisations prosodiques AV de ces attitudes sont similaires dans les deux langues et culture. Nous pourrons donc les considérer comme des affects sociaux interculturels entre le vietnamien et le français.

Par ailleurs, il y a des attitudes bien reconnues par les auditeurs natifs, mais ne sont pas reconnues par les francophones : INT, MAT et FAM. Les réalisations prosodiques vietnamiennes de ces concepts ne sont pas partagées avec les Français et elles ont besoin d'être acquises par des apprenants de langue étrangère. Des conclusions semblables ont déjà été discutées pour certaines attitudes en japonais, qui ne sont pas reconnues par les français ou les anglais [10].

Un cas intéressant est l'expression de l'admiration, qui est mal reconnue par les auditeurs natifs, mais mieux reconnue par les non-natifs (dans les modalités visuelle et audio-visuelle). On peut peut-être relier ce résultat au fait que pour les Vietnamiens, cette attitude ne peut pas se produire sans une cohérence morpho-lexicale [7]. C'est-à-dire qu'il y a impossibilité à séparer énoncé et prosodie. Donc une prosodie d'admiration sur un énoncé « neutre » n'est pas écologique, ce qui n'est pas le cas en français [11]. Toutefois, des investigations supplémentaires doivent être effectuées afin de vérifier cette hypothèse.

6. CONCLUSIONS

Ce travail vise à l'évaluation interculturelle des affects sociaux, en modalité audiovisuelle, pour le vietnamien. La réalisation des attitudes par un locuteur vietnamien entrainé a été assez bien évaluée par des auditeurs natifs et non-natifs. Les résultats expérimentaux montrent les facteurs influant sur la perception des attitudes : la modalité de présentation et l'expression de l'attitude elle-même. Ces résultats nous permettent également d'étudier les spécificités culturelles et la perception interculturelle des attitude vietnamiennes et de poser aussi des questions intéressantes pour des recherches futures (notre application visée est la synthèse vocale), par exemple dans le domaine de l'enseignement des langues étrangères.

Toutefois, les résultats doivent encore être validés par une analyse des paramètres prosodiques, afin de déterminer quels paramètres acoustiques et visuels conduisent à la perception de ces affects sociaux. D'autres expériences de perception avec des variations tonales sont en cours, afin d'explorer l'effet d'un tel système tonal sur la perception des attitudes pour les auditeurs natifs, mais aussi pour des locuteurs étrangers sans pratique d'une langue tonale: sont-ils capables de séparer l'information tonale locale de l'information globale d'attitude, sans confondre des deux niveaux informatifs ?

BIBLIOGRAPHIE

[1] Scherer, K.R., & Ellgring, H. "Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns?", Emotion, 7(1), 158-171, 2007.

[2] Aubergé, V., "A Gestalt Morphology of Prosody Directed by Functions: the Example of a Step by Step Model Developed at ICP", in *Speech Prosody*, 151-154, France, 2002.

[3] Shochi, T., Aubergé, V., and Rilliard, A., "How prosodic attitudes can be false friends: Japanese vs. French social affects", in *Speech Prosody*, 692-696, Dresden, 2006.

[4] Barkhuysen, P., Krahmer, E. & Swerts, M. "Cross-modal perception of emotional speech", in *ICPhS*, 2133-2136, Saarbruecken, Germany, 2007.

[5] Do T.D., Tran T.H. & Boulakia G., "Intonation in Vietnamese", in Intonation systems: A survey of 22 languages, D. Hirst and A. Di Cristo, Eds.: Cambridge University Press, 395-416, 1998.

[6] Shochi, T., Erickson, D., Rilliard, A., and Aubergé, V., "Recognition of Japanese attitudes in Audio-Visual speech", in *Speech Prosody*, 689-692, Campinas, Bresil, 2008.

[7] Le T.X., "Etude contrastive de l'intonation expressive en français et en vietnamien", PhD thesis of Linguistic and Phonetic, Université Paris 3, 1989.

[8] Diaféria, M.-L., "Les Attitudes de l'Anglais : Premiers Indices Prosodiques", Master thesis, INP Grenoble, France 2002.

[9] Morlec, Y., Bailly, G., & Aubergé, V., "Generating the prosody of attitudes", in *ETRW Workshop on Prosody*, 251-254, Athens, Greece, 1997.

[10] Shochi, T., Aubergé, V. & Rilliard, A. "Cross-Listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes", in *ICPhS*, 2097-2100, Saarbrucken, Germany, 2007.

[11] Rilliard, A., Shochi, T., Martin, J.C., Erickson, D. and Aubergé, V. "Multimodal Indices To Japanese And French Prosodically Expressed Social Affects", Language and Speech 52(2&3), 223-243, 2009.

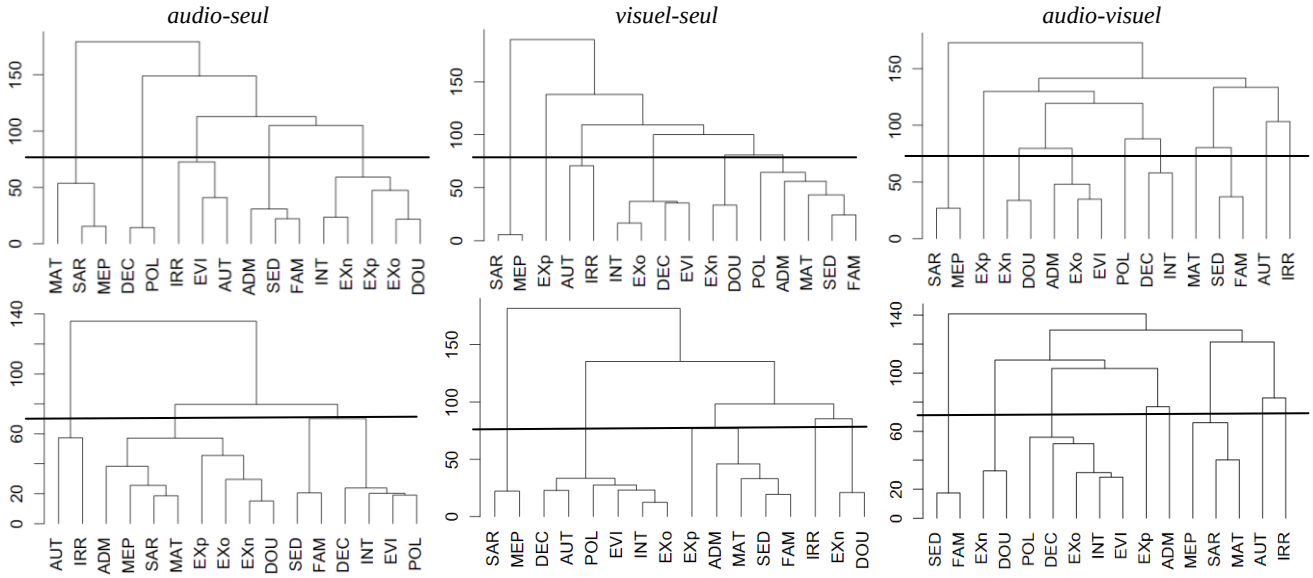


Figure 2 : Dendrogrammes des 16 attitudes pour les auditeurs vietnamiens (en haut) et les auditeurs français (en bas)

SPÉCIFICITÉS DE L'ACQUISITION DES CONSONNES EN FRANÇAIS ET EN DREHU : INFLUENCE DE LA LANGUE AMBIANTE

Julia Monnin[1,2] et Hélène Lœvenbruck[1]

[1]EA CNEP, Université de la Nouvelle-Calédonie, Nouméa, France; [2]Département Parole et Cognition, GIPSA-lab, CNRS UMR 5216, Université de Grenoble, France;
monninjulia@yahoo.fr; helene.loevenbruck@gipsa-lab.grenoble-inp.fr
http://www.gipsa-lab.inpg.fr

ABSTRACT

This study extends a cross-linguistic collaboration on phonological development, which aims to compare production of word-initial obstruents across sets of languages which have comparable consonants that differ in overall frequency or in the frequency with which they occur in analogous sound sequences. By comparing across languages, the influence of language-specific distributional patterns on consonant mastery can be disentangled from the effects of more general phonetic constraints on development. We compared production of French and Drehu word-initial obstruents to frequencies in the input, and did word and non-word repetition experiments with French and Drehu-acquiring two-year-old to five-year-old children. Results show that frequencies in the input are correlated to production of word initial obstruents especially in youngest children. Both groups of monolingual and bilingual children have similar scores but they show different patterns for particular obstruents that differ in the input frequencies.

Keywords: Phonological development, lexical frequency, obstruents, French, universal, language-specific, Drehu language.

1. INTRODUCTION

Des tendances universelles du développement phonologique chez l'enfant ont été mises en évidence dès 1941 [1]. Il a en effet été montré que certains sons dits « difficiles » sont acquis plus tardivement que d'autres qui apparaissent dès le babillage dans toutes les communautés linguistiques. Ainsi les fricatives et les affriquées sont maîtrisées plus tard que les occlusives [2]. Les constrictions dorsales sont maîtrisées plus tardivement quand elles sont suivies de /i/ que de /u/ [3]. Des tendances universelles sur les associations intrasyllabiques produites dans le babillage ont aussi été repérées et expliquées par des oscillations mandibulaires associées à des mouvements secondaires de la langue, du velum et des lèvres [3]. Or, s'il semble bien exister des contraintes biomécaniques universelles qui façonnent les premières productions des enfants, des influences de la langue ambiante ont également été observées. Certains sons « difficiles » sont en fait acquis plus ou moins tôt selon les langues. Ainsi, /v/ semble maîtrisé plus tôt en finnois, estonien et bulgare qu'en anglais. Ingram [3] suggère que la fréquence de /v/ en anglais est plus faible, ce qui explique les différences de calendrier des acquisitions. Certaines tendances de l'acquisition phonologique s'expliqueraient ainsi par des différences de phonèmes et de leurs fréquences selon les langues (cf. aussi [5]). Il existe donc une influence de la langue ambiante sur les premières productions des enfants. Il a même été montré que l'intonation du cri des nouveau-nés est déjà influencée par la langue maternelle [6].

Cette étude s'inscrit dans le cadre du projet Paidologos (http://www.ling.osu.edu/~edwards/) qui examine les influences des fréquences des consonnes à l'initiale des mots sur le développement phonologique, dans plusieurs langues. Il s'agit de comparer les productions de consonnes d'attaque dans différentes langues présentant des distributions fréquentielles peu similaires, chez des enfants de 2 à 5 ans. L'étude présentée ici étend ce projet au français et au drehu. La langue drehu est la langue majoritaire sur l'île de Lifou (Nouvelle Calédonie). Elle compte environ 15 000 locuteurs. Elle possède un riche répertoire consonantique. La situation bilingue drehu et français reste très présente parmi les enfants locuteurs du drehu. Toutefois, la langue drehu est dominante en contexte familial et scolaire sur Lifou [7].

Dans un premier temps, nous avons comparé les données fréquentielles pour ces deux langues avec les productions de consonnes à l'initiale des mots chez les enfants. Ensuite, les compétences des enfants locuteurs du drehu et des enfants monolingues français ont été comparées en répétition de non-mots et de mots.

2. CORPUS

Des enfants monolingues du français et des enfants locuteurs du drehu en situation bilingue drehu-français à dominance drehu (désormais dits « bilingues », par souci de concision) ont été enregistrés en répétition de mots et de non-mots. Huit groupes d'âge de 20 enfants ont été constitués en français, en respectant des écarts de 6 mois entre chaque groupe, de 2 ans jusqu'à 5 ans. En drehu, trois groupes d'âge ont été retenus, allant de 3 à 5 ans, chaque groupe étant constitué de 15 à 20 enfants. Les groupes d'enfants ont chacun répété des mots dans leur langue maternelle (en français pour les enfants dits monolingues, en drehu pour les enfants dits bilingues) et des non-mots typiques du français. Les mots et non-mots commençaient tous par une des consonnes étudiées dans le projet. Les enfants étaient installés devant un ordinateur

et voyaient défiler les images couplées aux mots ou non-mots prononcés. Les mots et non-mots avaient été préalablement enregistrés par deux locutrices natives dans un style de parole adressé à l'enfant. La consigne était de répéter les sons entendus. Ensuite, une locutrice native du français (première auteure) et un locuteur natif du drehu ont transcrit les données. En français, pour les mots, les consonnes étudiées étaient /t/, /d/, /k/, /g/, /s/, /ʃ/, /tʃ/, /z/, et /n/, suivies de différentes voyelles : /A/ (/a/, /ɑ/ et /ɑ̃)/, /E/ (/e/, /ɛ/, /ɛ̃/ et /œ̃/), /O/ (/o/, /ɔ/ et /ɔ̃/), /8/ (/ø/ et /œ/), /i/ et /y/ pour les mots. Les mots choisis étaient soit « faciles » avec une structure syllabique de type CV, CVC ou CV.CV, soit « difficiles » avec une structure syllabique de 3 syllabes. Ils étaient fréquents, dans la mesure du possible. Selon la consonne observée (par exemple /tʃ/ en français) certains mots étaient en réalité peu fréquents. En drehu, pour les mots, les consonnes considérées étaient /t/, /d/, /k/, /g/, /tʃ/, /dʒ/, /θ/, /ð/, /x/, /s/ et /z/, suivies des voyelles /a/, /i/, /u/. Les non-mots ont été construits pour être typiques du français et de façon à présenter soit une structure syllabique « facile » (CV ou CVC ou CV.CV), soit une structure syllabique « difficile » (CV.CVC.CV ou CVC.CV.CV). De plus, ces non-mots respectaient les fréquences moyennes des séquences de syllabes en français. Les consonnes retenues appartenaient aux systèmes phonologiques du français et du drehu, les fréquences des consonnes dans ces deux langues n'étant pas identiques et permettant de comparer les productions sur des phonèmes rares en français mais fréquents en drehu, ou réciproquement. Onze consonnes ont ainsi été testées : /d/, /dz/, /dʒ/, /g/, /k/, /kw/, /t/, /tʃ/, /tw/, /z/, /ʒ/.

Pour les deux langues, les transcriptions ont uniquement concerné l'initiale des mots et la voyelle suivante. Les transcriptions étaient à la fois de type phonologique (réussite ou échec de la répétition de la consonne et de la voyelle initiales) et phonétique (transcription respectant un certain codage de la production erronée de l'enfant).

3. PRODUCTIONS DE CONSONNES EN FRANÇAIS ET EN DREHU SELON LES DONNEES FREQUENTIELLES DE L'INPUT

3.1. Méthodologie

Les données obtenues en répétition de mots ont été comparées aux mesures fréquentielles de la langue ambiante de l'enfant. Certaines études suggèrent que les données fréquentielles sur les phonèmes peuvent varier selon le registre étudié. Les fréquences calculées sur des corpus de parole adressée à l'adulte peuvent ainsi différer de celles de la parole adressée à l'enfant. Nous avons donc utilisé des fréquences obtenues à partir d'enregistrements de parents s'adressant à l'enfant de 2 ans, en français, et en drehu [8]. Les coefficients de corrélations ont été calculés sur 11 paires de données, correspondant, pour chaque consonne, à un score de productions correctes et une fréquence de l'*input*.

3.2. Résultats

La figure 1 corrèle les pourcentages de répétition correcte des consonnes initiales des mots en français et les fréquences de l'*input*, pour les enfants les plus jeunes de notre étude (20 enfants de 2 ans à 2 ans 5 mois). Ces deux variables sont très fortement corrélées (corrélation de Pearson significative : 0,81). Notons que pour les enfants de 3 ans à 3 ans 5 mois, la corrélation passe à 0,77, et à 0,60 pour les enfants de 3 ans 6 mois à 3 ans 11 mois.

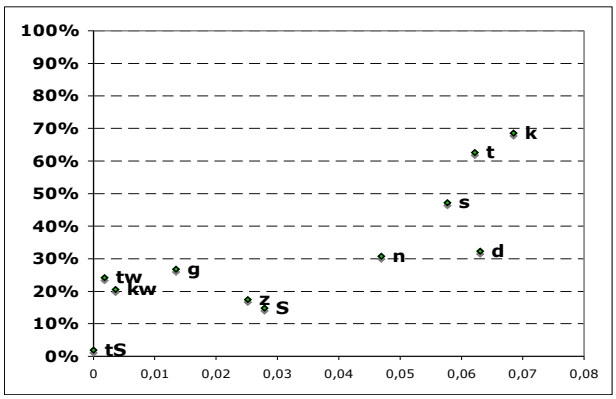


Figure 1 : Corrélation entre productions correctes et fréquences de l'*input* entre 2 ans et 2 ans 5mois en français (S : /ʃ/, tS : /tʃ/).

La figure 2 montre qu'à 5 ans (20 enfants âgés de 5 ans ½ à 5 ans 11 mois), les fréquences de l'input n'expliquent plus aussi fortement les résultats obtenus en répétition de mots en français (corrélation de 0,51, non significative).

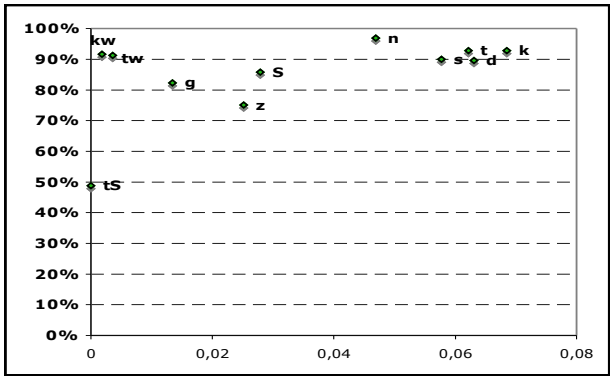


Figure 2 : Corrélation entre productions correctes et fréquences de *l'input* français entre 5 ans ½ et 5 ans 11 mois.

La figure 3 corrèle les pourcentages de répétition correcte des consonnes des mots en drehu et les fréquences de l'input, pour 16 enfants bilingues drehu-français âgés de 3 ans à 3 ans 11 mois. La corrélation est forte à 0,67. La figure 4 montre qu'à l'âge de 5 ans, comme en français, les productions correctes sont moins fortement corrélées aux fréquences (corrélation de 0,63 significative). Ainsi on observe qu'à l'âge de 2 et 3 ans, les productions des enfants sont très influencées par la langue ambiante ; plus tard, lorsqu'ils maîtrisent mieux la langue, l'influence de la langue est moins sensible. Les enfants français atteignent alors des scores supérieurs à 70% pour la plupart des consonnes, et les enfants drehu à 60%.

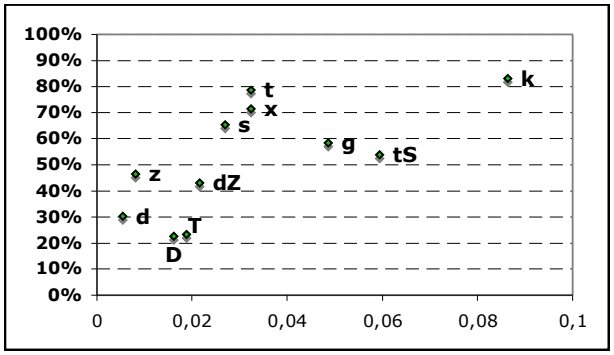


Figure 3 : Productions correctes de consonnes initiales des mots en drehu à 3 ans en fonction des fréquences de l'*input* (tS : /tʃ/, dZ = /dʒ/, T = /θ/, et D = /ð/).

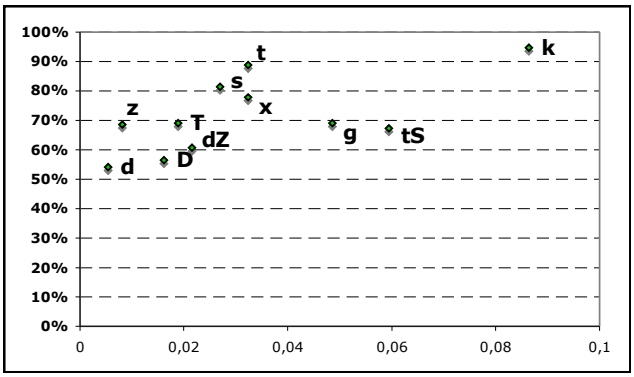


Figure 4 : Productions correctes de consonnes initiales des mots en drehu à 5 ans en fonction des fréquences de l'*input*.

4. COMPARAISON DU FRANÇAIS ET DU DREHU EN REPETITION DE NON-MOTS

Les enfants français et drehu ont également passé une épreuve de répétition de non-mots typiques du français. Ces données permettent de comparer deux groupes d'enfants, français et drehu, sur des répétitions identiques.

4.1. Méthodologie

Afin de pouvoir comparer les deux groupes d'enfants, les enfants francophones de 3 ans (les 20 plus jeunes et les 20 plus âgés) ont été regroupés, de même pour les enfants de 4 ans et de 5 ans. Les enfants français de 2 ans n'ont pas été retenus pour cette étude, aucun enfant bilingue de 2 ans n'ayant pu être enregistré. Nous prenons donc en considération les productions de 164 enfants (116 monolingues français et 48 bilingues drehu-français).

4.2. Résultats

Les enfants monolingues du français ou bilingues drehu-français réussissent l'épreuve de répétition de non-mots de façon similaire (figure 5). Une ANOVA montre que l'effet de la langue n'est pas significatif ($F(1, 158) = 3,01$, $p = 0,085$), alors que celui de l'âge l'est ($F(2,316) = 29,88$ $p< 0,001$). Les productions correctes des enfants augmentent progressivement de 3 ans à 5 ans et atteignent plus de 50% à 5 ans pour les 11 consonnes testées.

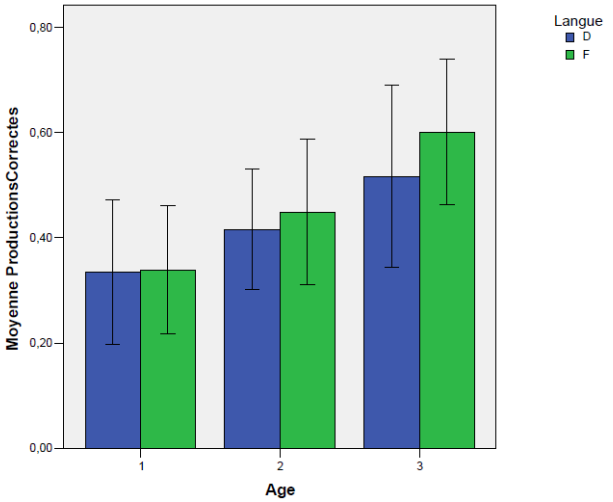


Figure 5 : Répétition de non-mots français par des enfants bilingues drehu-français (D, gauche) et monolingues du français (F, droite) à 3 ans (1), 4 ans (2) et 5 ans (3) : moyennes des répétitions correctes +- écarts-type.

La figure 6 présente les résultats obtenus en répétition de non-mots pour les enfants français et les bilingues drehu-français pour les phonèmes initiaux /k/, /g/, /t/ et /d/.

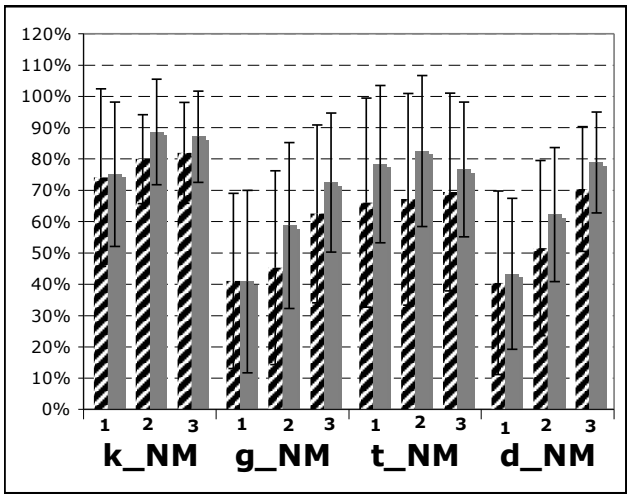


Figure 6 : Répétition correcte des consonnes initiales /k/, /g/, /t/ et /d/ des non-mots (NM) exprimée en pourcentage avec un écart-type. 1 : enfants de 3 ans ; 2 : enfants de 4 ans ; 3 : enfants de 5 ans. En hachuré : bilingues ; en gris : monolingues.

Il est intéressant de comparer ces consonnes car des tendances différentes ont été obtenues pour les dentales et les dorsales selon les langues. Dans certaines, /t/ est mieux produit que /k/ (anglais), dans d'autres, c'est l'inverse (japonais). Les résultats des analyses statistiques sur nos données sur les non-mots montrent que l'effet de la langue n'est pas significatif pour ces phonèmes, sauf pour /t/ qui est mieux prononcé par les monolingues français ($F(1,158)=6,38$ $p=0,012$). Or /t/ est plus fréquent dans l'*input* en français qu'en drehu. On note de plus que /k/, souvent considéré comme acquis plus tard [1], atteint plus de 70% de réussite même à un âge précoce. Ce résultat se rapproche de ce qui a été observé pour le japonais [9] et peut s'expliquer par le fait que la consonne /k/ est très fréquente en français comme en drehu.

La figure 7 fournit les résultats en répétition de non-mots pour les consonnes /tʃ/ et /dʒ/ fréquentes en drehu et rares en français, et /ʒ/ et /z/ assez fréquentes en français et rares en drehu. L'effet de la langue est significatif pour /tʃ/ ($F(1,158)=4,86$ $p=0,029$) et /dʒ/ ($F(1,158)=4,75$ $p=0,031$) pour les non-mots : les bilingues ont de meilleurs scores. La fréquence d'un phonème dans une langue semble bien jouer sur sa production par l'enfant.

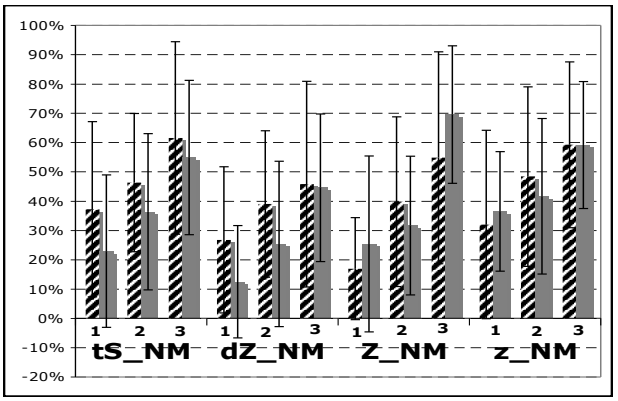


Figure 7 : % de répétitions correctes des consonnes initiales /tʃ/, /dʒ/, /ʒ/ et /z/ des non-mots (NM), avec 1 écart-type. 1 : 3 ans ; 2 : 4 ans ; 3 : 5 ans. Hachuré : bilingues ; en gris : monolingues.

5. COMPARAISON DU FRANÇAIS ET DU DREHU EN REPETITION DE MOTS

Les résultats obtenus pour certains phonèmes communs aux deux listes de mots (/k/, /g/, /t/ et /d/) sont comparés sur la figure 8. Les mots commençant par /d/ sont moins bien répétés par les enfants bilingues, la différence est significative ($F=53,65$, $p<0,001$). La consonne /t/ est aussi mieux répétée chez les français ($F=5,49$, $p = 0,02$). Les autres consonnes ont un taux de répétition correcte identique, l'effet de la langue n'est pas noté pour /k/ et /g/.

6. DISCUSSION ET CONCLUSION

Les capacités phonologiques initiales chez le jeune enfant (2 à 3 ans) semblent fortement corrélées aux données fréquentielles de l'*input*. Pour notre groupe d'enfants monolingues du français, cette corrélation est très forte, elle l'est moins mais reste forte cependant pour notre groupe d'enfants bilingues drehu-français. De façon attendue, ces corrélations diminuent lorsque les enfants sont plus âgés (5 ans) dans les deux groupes considérés. Il est établi que les enfants bilingues présentent parfois un délai à l'acquisition du lexique et que les capacités phonologiques sont en lien avec l'accroissement du lexique. Cependant, dans notre épreuve de répétition de non-mots, les enfants issus des deux groupes ne présentent pas de différences dans les pourcentages de répétition correcte des consonnes initiales des non-mots. Les profils des enfants monolingues et bilingues présentent pourtant des spécificités lorsque l'on compare plus précisément les répétitions de certains phonèmes. Les phonèmes peu fréquents en français et plus fréquents en drehu (/tʃ/, /dʒ/) sont mieux répétés par les bilingues. Réciproquement, /d/, plus fréquent en français, est mieux produit par les monolingues, surtout dans les mots (la différence n'est pas significative dans les non-mots). Ainsi les enfants bilingues drehu-français sont très influencés par l'*input* drehu, même s'ils reçoivent aussi un *input* français. Nos résultats montrent donc que les productions des jeunes enfants sont très influencées par la langue ambiante.

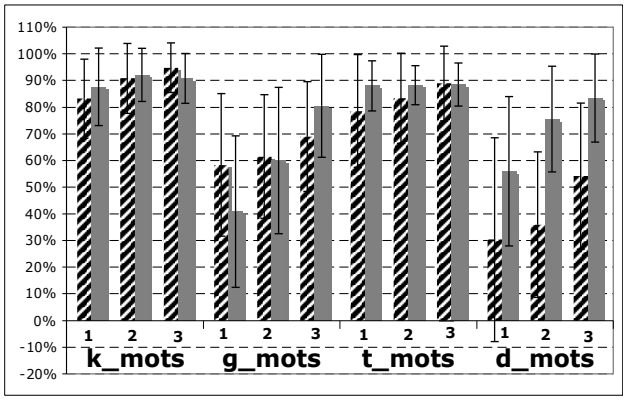


Figure 8 : Répétitions correctes (%) des consonnes initiales /k/, /g/, /t/ et /d/ des mots, avec 1 écart-type. 1 : enfants de 3 ans ; 2 : 4 ans ; 3 : 5 ans. En hachuré : bilingues ; en gris : monolingues.

REMERCIEMENTS

Nous remercions Fabrice Wacalie et Numa Henesewene pour les transcriptions phonétiques des corpus en drehu et Mary Beckman et Anne Vilain pour leurs commentaires.

BIBLIOGRAPHIE

[1] Jakobson, R. (1941/1968). A. R. Keiler, Child language, aphasia, and phonological universals. The Hague: Mouton.

[2] Hua, Z. & Dodd, B. (2000). The phonological acquisition of Putonghua (Modern Standard Chinese). *J. of Child Language*, 27, 3-42.

[3] Davis, B.L., MacNeilage, P.F. & Matyear, C. Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica*, 2002, 59, 75-107.

[4] Ingram, D. (1988). The acquisition of word-initial [v]. *Language and Speech*, 31, 77-85.

[5] de Boysson-Bardies B., Vihman M., Roug-Hellichius L., Durand C., Landberg I. & Arao F. (1992). Material evidence of infant selection from the target language. A cross-linguistic phonetic study. *Phonological development: models, research, implications.* C. A. Ferguson, L. Menn, C. Stoel-Gammon (eds.), York Press. 369-391.

[6] Mampe B., Friederici A. D., Christophe A. & Wermke K. (2009). Newborns' Cry Melody Is Shaped by Their Native Language. *Current Biology*, 19, 1-4.

[7] Moyse-Faure, C., (1983). *Le drehu, Langue de Lifou (Iles Loyauté)*. Paris-SELAF; 17-31.

[8] Monnin J. & Lœvenbruck H. (2008). Influence des fréquences lexicales des langues française et drehu sur l'acquisition des consonnes initiales de mots. *Actes des XXVIIèmes JEP*, Avignon.

[9] Beckman, M. E., Yoneyama, K., & Edwards, J. (2003). Language-specific and language universal aspects of lingual obstruent productions in Japanese-acquiring children. *J. of the Phonetic Society of Japan*, 7, 18-28.

Effets du discours adressé à l'enfant sur l'acquisition de la liaison : étude d'un corpus dense d'une fillette de 40 mois

CHABANAL Damien

lrl
Université Blaise Pascal
Damien.chabanal@univ-bpclermont.fr

ABSTRACT

The study of dense corpuses like (Behrens [1], Tomasello & Stahl [2]) aims at showing precisely the influence that the parents' way of speaking has on the language acquisition of their children. This study will focus on a 40-month-old French speaking girl whose words have been recorded for one hour and a half , every single day for 8 days. This study points out that there is a close link between the production of the compulsory and optional liaisons and also the production of optional « adjective + noun–type » liaison. Besides, it shows the effect of the frequency of the word1-word2 structure on the liaison acquisition for a child.
Key words : Acquisition, liaison, psycholinguistics, usage based theory, frequency

1. INTRODUCTION

Les recherches récentes sur les effets des discours adressés à l'enfant (Nardy [3] ; Chenu et Jisa [4]) tendent à contredire la thèse de la pauvreté de l'input dans les premières étapes de l'acquisition du langage. En effet, apparaissent dans ces différentes études, l'idée que précocement, l'enfant apprend par mémorisation des formes les plus fréquemment produites par son entourage en déduisant par la suite un certain nombre de règles ou de principes analogiques. Concernant l'acquisition de la liaison, les recherches effectuées à partir de la théorie basée sur l'usage (Chevrot, Chabanal, Dugua [5] ; Chabanal [6]) ont proposé un scénario développemental d'acquisition autour de trois étapes. Lors de la première, il est postulé que les liaisons s'apprendraient contexte après contexte suivant la fréquence des liaisons rencontrées. En clair, plus une liaison serait présente dans l'input de l'enfant, plus elle aurait des chances d'être mémorisé et reproduite plus rapidement. Le recueil de corpus dense, constitué grâce au projet ANR « phonlex », a notamment pour objectif de vérifier directement les effets du discours adressé à l'enfant sur l'acquisition de la liaison. Après avoir brièvement rappelé les débats qui animent la question de son acquisition, nous évoquerons les méthodes de recueil et les intérêts des corpus denses. Enfin, nous analyserons l'impact du discours des parents de Prune, fillette de 40 mois, enregistré durant une semaine à raison d'un heure trente d'enregistrement par jour, sur l'émergence de ses liaisons.

Nous essaierons de mesurer cet impact à travers deux comparaisons. D'une part en comparant les taux de liaisons justes par contexte de liaisons obligatoires et facultatives produites par Prune et ses parents. D'autre part en faisant le lien entre la fréquence des collocations mots1-mots2 du contexte adjectif + nom dans l'input et la production de ces contextes chez Prune. En d'autres termes, nous nous demanderons comme Alexander [7] pour les adultes, si la réalisation des liaisons facultatives est plus importante dans les collocations les plus fréquentes.

2. LA LIAISON ET SON ACQUISITION

La liaison est l'action de prononcer devant un mot à initiale vocalique ou commençant par un "h muet" (non aspiré) la consonne finale du mot précédent (telle qu'elle est écrite ou avec une modification) qui normalement est absente du langage oral. La liaison se produit pour une classe de mots lorsque cette consonne finale est en position pré-vocalique. Elle ne sera pas prononcée en fin de phrase ni en position pré-consonantique. Les consonnes de liaisons sont limitées. D'après Boé & Tubach [8], /n/, /z/ et /t/ représentent 99.7% des liaisons produites, le reste (0.3%) correspondant aux consonnes /p, /R/ et /d/. Les grammairiens classent habituellement les contextes de liaison en trois types : les liaisons obligatoires (désormais LO), les liaisons facultatives (désormais LF), les liaisons interdites/ fautives. Nous considérons qu'une liaison est obligatoire quand elle est réalisée par 100% des locuteurs. Cette approche empirique propose un autre classement que celui réalisé par les grammaires normatives. Celles-ci considèrent par exemple le contexte "adjectif + nom" comme obligatoire alors que dans l'usage les locuteurs ne marquent pas toujours ce type de liaison. De jong [9], à partir d'un corpus de locuteurs français, propose un classement sur lequel nous nous fonderons. Cet auteur distingue seulement quatre contextes où la liaison est catégoriquement réalisée : entre un déterminant et un nom (des enfants, un ami, etc.), entre un pronom personnel et un verbe (ils arrivent, nous allons, etc.), entre un verbe et un pronom personnel (vient-il, prenez en, etc.), dans des expressions figées (tout à coup, tout à l'heure,etc.). Durand, Lyche & Tarrier [10] et Durand & Lych [11], à partir du corpus recueilli dans le cadre du projet *Phonologie du Français Contemporain* font des remarques analogues.

Sur le plan psycholinguistique, la manière dont s'acquière la liaison est le lieu d'un débat dynamique. Deux approches différentes sont proposées. Une, que nous défendons, propose une conception exemplariste et basée sur l'usage (Chevrot, Chabanal, Dugua [5]). Dans ce cadre, l'acquisition de la liaison est lexicale, contexte après contexte, et c'est à partir d'un certain stock lexical que se dessine chez l'enfant la compréhension du phénomène de liaison. Pour l'autre courant, (Wauquier-Gravelines & Braud [12]) l'enfant construit une représentation phonologique abstraite de la liaison à partir de contraintes phonologiques universelles en cours de paramétrisation.

3. IMPACTS DU DISCOURS ADRESSE A L'ENFANT

Pour évoquer cette question récente (Tomasello [13]) et riche en questionnements, nous proposons de faire un point sur les modalités du recueil d'un corpus dense, son traitement et les axes de recherche qu'il engage.

3.1. Recueil d'un corpus dense

L'observation naturelle des interactions parents-enfants est le premier élément de définition de ce type de corpus. On peut y adjoindre un cahier de notes où nous notons « à la volée » des contextes produits hors de la période d'enregistrement. Le moment choisi pour enregistrer ces interactions est variable, il peut s'agir du repas, du bain ou lors de jeux. La question de la quantité de données suffisantes par jour et sur la totalité du corpus constitue l'autre donnée fondamentale pour la constitution d'un corpus dense. Le temps d'enregistrement doit être calculé à partir de l'émergence plus ou moins fréquente des formes que l'on veut étudier. Tomasello & Stahl [2] précisent par exemple que la fréquence d'utilisation de copules ou de pronoms en anglais est telle que l'enregistrement d'une heure par jour d'interactions parents-enfant sur deux semaines est suffisant. Ces mêmes auteurs inventorient les différents procédés pour évaluer la quantité suffisante de temps d'enregistrement recommandé pour avoir suffisamment de données à traiter. En ce qui concerne le corpus de Prune (40 mois), pour un enregistrement d'une heure trente par jour sur une semaine, nous avons en moyenne par jour, 25 contextes de liaison chez la mère, 123 contextes chez le père et 139 chez la fillette. Ces éléments sont suffisants pour travailler sur des études quantitatives.

3.2. Le traitement d'un corpus dense : nature et quantité d'input

Sur le plan du traitement des corpus denses, ces derniers, comme le précise Tomasello [13], peuvent nous aider à étudier l'influence de deux aspects de l'environnement langagier : l'un se manifestant au travers de la nature du discours entendu (*children hear differents things*), l'autre au travers de la quantité du discours entendu (*in differents quantities*).

Pour ce qui est de la nature de l'input, il apparaît que ce dernier est caractérisé par des énoncés courts et bien formés, une vitesse d'élocution plus lente, des pauses plus longues et de nombreuses répétitions. Cameron-Faulkner, Lieven & Tomasello [14], après avoir observé 12 dyades mère-enfant entre 1;9 et 2;6, ont noté ceci : (1) 20% des énoncés produits par les mères consistent en des fragments de phrases ne comportant pas de sujet ou de verbe ; (2) 31% des énoncés correspondent à des questions ; (3) 18% des énoncés possèdent un sujet et un verbe ; (4) 6% des énoncés sont des phrases complexes constituées de plusieurs propositions ; (5) 51 % des énoncés commencent par 52 mots ou séquences très fréquents comme *it's …*, *look at …*, *can you …*, *what …* et chacune de ces séquences est entendue plus de 40 fois par jour par la moitié des 12 enfants. Du point de vue de la quantité d'input, il a été prouvé que la fréquence d'usage de certaines formes linguistiques par les mères corrèle avec la fréquence d'usage et donc d'acquisition de ces mêmes formes par leur enfant. Chenu et Jisa [4] constatent un effet de la fréquence des verbes produits par la mère sur leurs acquisitions par de jeunes enfants français. Dans le cadre de l'ANR « phonlex », nous travaillons sur ces différents axes. En effet, nous observons les corpus denses de quatre fillettes qui ont respectivement 28, 32, 36 et 40 mois. Nous les enregistrons par deux fois à huit mois d'intervalles, de façon à avoir des renseignements sur les types d'effets de l'input à différents âges.

4. METHODOLOGIE

Le corpus sur lequel se fonde cette étude contient des interactions entre Prune, fillette âgée de 40 mois, et ses parents (père : maître de conférences, mère : orthophoniste) vivant à Clermont-Ferrand (Puy-de-Dôme). Au final, 2309 contextes de liaison ont été recueillis et analysés. Les contextes étudiés sont classés de la façon suivante : Pour les liaisons obligatoires : A : après un déterminant (les ours…), B : après un pronom clitique pré- verbal (j'en ai…), C : dans une expression figée (tout à coup), D : entre verbe et pronom clitique pré- verbal (prenez-en…) Pour les liaisons facultatives : E : après un adjectif pré-nominal (petit-ours), F : après un nom pluriel (des enfants idiots), G : après une forme du verbe avoir (ils ont un…), H : après une forme du verbe être (ils sont ici), I : après une forme d'un autre verbe (il vient aussi), J : après mot invariable (quand un…).

5. RESULTATS

5.1. Production de liaisons justes chez Prune et chez ses parents

Nous postulons ici qu'il existe une corrélation positive significative entre les productions des liaisons justes de LO et de LF chez Prune et ses parents.

Tableau 1 : Pourcentages de liaisons obligatoires et facultatives réalisées justes par Prune et ses parents

		Prune	**Parents**
LO	A. Après déterminant	96.1% (253/263)	98.7% (232/235)
	B. Après pronom clitique préverbal	99.6% (302/303)	100% (263/263)
	C. Dans expression figée	100% (21/21)	100% (19/19)
	D. Entre verbe et clitique post-verbal	100% (25/25)	100% (22/22)
Total liaisons obligatoires réalisées justes		**98,2% (601/612)**	**99,4% (539/536)**
LF	E. Après adjectif pré-nominal	69% (58/84)	80% (60/75)
	F. Après nom pluriel	0% (0/27)	5% (2/40)
	G. Après une forme du verbe avoir	0% (0/24)	0% (0/32)
	H. Après une forme du verbe être	8.2% (6/73)	10.1% (13/128)
	I. Après une forme d'un autre verbe	0% (0/177)	1,8% (4/212)
	J. Après mot invariable	39% (46/118)	23% (34/147)
Total liaisons facultatives réalisées justes		**21,8% (110/503)**	**17,8% (113/634)**

Nous observons une corrélation positive et significative entre les pourcentages de réalisations justes de Prune et de ses parents (Corrélation de Spearman: Rho = 0.978 ; p= 0.01). Autrement dit, les pourcentages de réalisation justes sont forts de manière équivalente, à la fois chez Prune et chez ses parents. Cependant, il conviendrait de comparer les résultats de Prune avec d'autres locuteurs adultes pour prétendre à authentifier ces corrélations.

5.2. Production de Liaisons facultatives après adjectifs chez Prune et chez ses parents

Nous avons choisi les contextes de LF les plus présents chez Prune (adjectif + nom) et nous avons testé une corrélation entre les productions de ces mêmes contextes chez Prune et ses parents.

Tableau 2 : pourcentages de liaisons adjectif + nom réalisées justes par Prune et ses parents

Contextes LF les plus réalisées par Prune	**% de réalisations justes prune**	**% de réalisations justes parents**
Grand +	93% (14/15)	80% (4/5)
Petit +	91% (21/23)	77% (7/9)
Petits +	84 % (16/19)	72,7% (8/11)
Grands +	62% (5/8)	25% (1/4)
Gros +	53% (8/15)	53 %(8/15)

Parmi les contextes en E, nous constatons que les contextes les plus réalisés justes par Prune sont également ceux les plus réalisées justes par ses parents. Il existe donc un effet d'input de ces contextes sur la production de Prune. La corrélation de Spearman ressort également positive (rho = 0 ;97, p = 0,05).

5.3. Liaisons facultatives après adjectifs : Effets de la collocation mot1-mot2

Nous voulons ici savoir si le taux de réalisation varie en fonction de la présence plus ou moins forte des collocations mots 1-mots2 du type adjectif + nom dans l'input de Prune. Pour y répondre, nous avons calculé le rapport du nombre d'occurrences de mot1 au nombre de mot2 différents après le mot1 réalisées par les parents. Plus ce rapport est élevé et plus un mot1 apparaît dans des contextes mots1-mots2 fréquents. En conséquence, Prune aurait plus de chances de les mémoriser et ces taux de réalisations justes seraient plus élevés.

Le rapport parents (cf tableau ci-dessous) équivaut au rapport du nombre d'occurrences de mots1 par rapport au nombre de mots2 différents réalisées pour les parents.

Tableau n°3 : liaisons du types adjectif + nom : effet de la fréquence du mot1 et des collocations mot1-mot2 dans l'input.

Contextes LF les plus réalisées par Prune	valeur	Ordre	Rapport parents	ordre
Grand +	93% (14/15)	5	4,6 (14/3)	5
Petit +	91% (21/23)	4	3,5 (21/6)	4
Petits +	84% (16/19)	3	2,6 (8/5)	3
Grands +	62% (5/8)	2	2,5 (5/2)	2
Gros +	53% (8/15)	1	1,8 (9/5)	1

L'ordre selon les pourcentages de liaisons justes de Prune est équivalent à l'ordre du rapport parents. Ce fait signifierait donc que plus Prune entend des mot1-mot2 fréquents, plus elle réalise juste ces contextes. Le calcul d'un coefficient de corrélation par rang confirme cette tendance (Rho de Spearman = 1,000, p= 0,01).

6. CONCLUSION

Comme en témoigne nos résultats, l'enfant de 40 mois semble sensible à l'input, mémorisant davantage les collocations les plus fréquemment réalisées par son entourage. Nous avons pu montrer deux types de corrélation positive entre la production des liaisons obligatoires et facultatives des parents et de Prune ainsi qu'entre la production des contextes de liaisons facultatives (adjectif + nom) émergeant le plus souvent chez Prune. D'autre part, nous avons révélé, à travers l'effet des collocations mot1-mot2, deux éléments d'importance en faveur de l'étude des corpus denses. Premièrement, l'enfant de 40 mois est sensible à la fréquence des contextes lexicaux dans l'input. Deuxièmement, les études précédentes sur la liaison et l'effet de l'input avaient corrélé les productions des enfants avec des corpus adultes (frantext…) pris en dehors des interactions avec les productions des jeunes sujets. Grâce à l'étude des corpus denses, nous pouvons davantage démontrer le lien réel entre le discours de l'enfant et le discours des parents adressé à ce même enfant.

7. BIBLIOGRAPHIE

[1] H. Behrens. The input-output relationship in first language acquisition. *Language and Cognitive Processes*, 21, 2-24, 2006.

[2] M. Tomasello and D. Stahl. Sampling children's spontaneous speech : how much is enough ?, *J. Child Lang*. 31, 101-121. Cambridge University Press, 2003.

[3] A. Nardy. *Acquisition des variables sociolinguistiques entre 2 et 6 ans : facteurs sociologiques et influences des interactions au sein du réseau social*. Thèse de doctorat, Université Stendhal, Grenoble 3, 2008.

[4] F. Chenu and H. Jisa. Impact du discours adressé à l'enfant sur l'acquisition des verbes en français, *lidil* 31, 85-100, 2005.

[5] J.P. Chevrot, J.-P, D. Chabanal and C, Dugua.. Pour un modèle de l'acquisition des liaisons basé sur l'usage: trois études de cas. *Journal of French Language Studies*, 17, (1), 103-128, 2007.

[6] D. Chabanal. *Un aspect de l'acquisition du français oral: la variation socio-phonétique chez l'enfant francophone*. Thèse de doctorat, Université Paul Valéry, Montpellier, 2003.

[7] J. Alexander. *Frequency, prosody, and French liaison: testing Bybee's hypothesis*. BA Dissertation, Boston University, Boston, 2004.

[8] L.-J. Boë and J.-P. Tubach. (). *"De A à Zut": dictionnaire phonétique du français parlé*. Grenoble: Ellug, 1992.

[9] D. De Jong, (). La sociophonologie de la liaison orléanaise. In Lyche, C. (Ed.), *French Generative Phonology: retrospective and perspectives*, 95-130, Salford, 1994.

[10] J. Durand, C, Lyche and J.-M. Tarrier. *Quelles liaisons dans PFC ?* Consulté le 17 juillet 2008. http://www.projet-pfc.net/, 2007.

[11] J. Durand and C, Lyche. French liaison in the light of corpus data. *Journal of French Language Studies*, *18,* (1), 33-66, 2008.

[12] S. Wauquier-Gravelines and V. Braud. (). Proto-déterminant et acquisition de la liaison obligatoire en français. In Chevrot J.-P., Fayol, M., Laks, B (Eds.), , *Nouvelles approches de la liaison, Langages 158* 53-65, Paris : Larousse, 2005.

[13] M. Tomasello. *Constructing a language: a usage-based theory of language acquisition*. Cambridge: Harvard University Press, 2003.

[14]T.Cameron-Faulkner, E, Lieven and M. Tomasello. A construction based analysis of child directed speech. *Cognitive Science*, *27,* (6), 843-873, 2003.

Influence des méthodes d'enseignement de la lecture sur les fonctions cognitives de l'enfant

Trappeniers, J. & Lefebvre, L

Laboratoire de sciences cognitives
Université de Mons, 7000 Mons, Belgique
julie.trappeniers@umons.ac.be
http://w3.umh.ac.be/~scoglab/index.html

ABSTRACT

In the literature, many researchers conclude that reading ability requires multiple cognitive functions like attention, anticipation, segmentation, executive functions, categorization and memory. On the basis of idea that the reading implies complex and multiple cognitive competences which interact, the objective of our research was to determine if the reading methods even imply or develop these cognitive functions and more particularly the attentional and executive functions, differently. The results show that mixed methods develop more flexibility than others methods. This result appears interesting to us because perhaps that will allow that, in the future, the difficulties encountered by the child in his reading training are allotted to the interaction child-method.
Keywords: reading methods, attention, executive functions, development, neuropsychological tests.

1. INTRODUCTION

La littérature distingue généralement trois méthodes d'enseignement de la lecture : (i) les méthodes synthétiques basées sur l'introduction rapide des correspondances entre graphèmes et phonèmes ; (ii) les méthodes globales basées sur l'introduction rapide du sens et sur la formation d'un stock lexical important ; (iii) les méthodes mixtes fondées sur les deux approches précédemment décrites. Depuis des décennies et aujourd'hui encore, de nombreux auteurs évaluent l'influence des ces méthodes sur l'apprentissage même de la lecture en faisant abstraction du système cognitif de l'enfant. Or, l'acte de lire requiert de nombreuses fonctions cognitives dont le développement peut varier grandement en fonction de l'approche considérée. Dès lors, connaître l'implication des méthodes d'enseignement de la lecture sur le développement cognitif permettrait peut-être aux enseignants d'orienter les élèves dans l'enseignement qui leur est le plus profitable.

2. LECTURE ET FONCTIONS COGNITIVES ASSOCIÉES

Selon Le Ny [1], l'activité de lecture est influencée par un ensemble de compétences cognitives. Parmi elles, Charpentier [2] évoque des capacités de repérages visuels d'indices, d'imagerie et de représentation mentale, d'anticipation, de segmentation, de catégorisation, d'induction et d'inférence, d'habiletés analytiques, d'ordination temporelle et spatiale, de compétence d'abstraction et de généralisation. A cela il faut ajouter la mémoire (Lefebvre [3] et Stoll, cité par Van Grunderbeeck [4]) ainsi que les fonctions attentionnelles et exécutives (Van Der Sluis [5]), fonctions sur lesquelles nous nous attardons dans la présente étude.

D'un point de vue neuropsychologique, les fonctions attentionnelles et exécutives sont des fonctions de haut niveau qui commandent et déterminent les autres fonctions cognitives : schématiquement, les fonctions attentionnelles sélectionnent les informations à traiter et les fonctions exécutives exécutent les traitements appropriés (Mazeau [6]). De par leurs rôles respectifs, les fonctions attentionnelles sont régulièrement associées aux fonctions exécutives avec lesquelles elles forment un construit multidimensionnel (Metz-lutz et al. [7]).

L'implication des fonctions attentionnelles dans l'acte de lire a clairement été illustrée par de nombreux auteurs. Parmi elles, on retrouve plus particulièrement l'attention soutenue, l'attention sélective ainsi que l'empan attentionnel. A ce sujet, Ross (cité par Fijalkow [8]) considère l'attention sélective comme le mécanisme cognitif qui différencie le mieux les enfants en difficultés d'apprentissage de la lecture des autres enfants. Quant à l'attention soutenue, elle serait très sollicitée par les processus de compréhension de l'écrit. En effet, l'écrit nous prive de signes supra-segmentaires (intonations) ou extra-linguistiques (mimiques, gestes…). Par conséquent, l'accès aux éléments syntaxiques et textuels est plus prégnant mais aussi plus coûteux en attention soutenue (Mazeau [6]).

Selon Boulc'h et al. [9], plusieurs travaux démontrent l'implication des fonctions exécutives dans des tâches complexes de mémorisation mais peu étudient son rôle lors de l'activité de lecture. Les seules études existantes confirment l'implication des fonctions exécutives et plus précisément de l'inhibition et de la flexibilité. L'inhibition serait requise dans l'identification des mots (Brosnan et al. [10]) et la flexibilité permettrait de différencier les normolecteurs des enfants en difficultés d'apprentissage de la lecture (Boulc'h et al. [9]).

2.1. L'impact des méthodes d'enseignement de la lecture sur les fonctions cognitives

A ce jour, beaucoup d'auteurs ont évalué l'implication des fonctions cognitives dans l'acte de lire. Cependant, très peu se sont intéressés à déterminer l'influence des méthodes de lecture sur ces fonctions cognitives. Seule l'étude de Lefebvre [3] peut être mentionnée. Cette étude avait pour objectif principal de déterminer avec plus de précisions quelles fonctions sont réellement impliquées, voire développées lorsque l'enfant est soumis à un certain type de méthode de lecture. Les analyses ont porté sur une population de quatre-vingts enfants âgés de 6-7 ans suivant un enseignement fondamental ordinaire, répartis selon le type de méthode auquel ils étaient soumis. Les résultats obtenus sont les suivants : (i) les méthodes de lecture ne semblent pas développer les mêmes compétences cognitives ; (ii) la méthode synthétique semble développer une organisation visuo-spatiale particulière, appliquée à des formes abstraites ; (iii) la méthode globale semble intervenir partiellement dans le développement du raisonnement inductif ; (iv) il n'y a pas d'apports significatifs concernant la méthode mixte. Selon l'auteur, ces résultats devraient nous permettre de mieux comprendre les difficultés auxquelles certains enfants doivent faire face durant leur apprentissage de la lecture. Ainsi, selon lui, si les méthodes n'impliquent ni ne développent les mêmes compétences cognitives, il est raisonnable de penser qu'un enfant, selon ses capacités initiales, soit plus ou moins réceptif au type d'enseignement qui lui est proposé.

3. MÉTHODE

Dans la continuité de l'étude de Lefebvre [3], nous souhaitions évaluer l'impact des méthodes d'enseignement de la lecture sur l'attention et le fonctionnement exécutif, fonctions essentielles dans tout apprentissage.

Notre hypothèse générale postule la présence de différences significatives aux tests évaluant les fonctions attentionnelles et exécutives en fonction de la méthode à laquelle les enfants sont soumis. Ainsi, nous pensions que l'utilisation précoce du déchiffrage, l'introduction rapide de sens dans les activités de lecture ou encore la combinaison des deux pouvait avoir une influence plus ou moins importante sur les fonctions attentionnelles et/ou exécutives.

Parallèlement à cette hypothèse, nous nous sommes posés deux questions : 1/ sachant que la lecture exige une alternance entre des phases d'apprentissage aussi bien visuelles (découverte de la forme des lettres, découverte des mots) qu'auditives (apprentissage des sons), les méthodes développeraient-elles davantage l'une ou l'autre modalité sensorielle (visuelle vs auditive) des fonctions cognitives investiguées ? ; 2/ étant donné la proximité des fonctions cognitives investiguées, est-il possible d'obtenir une mesure « pure » de chacune d'elle ?

3.1. Sujets

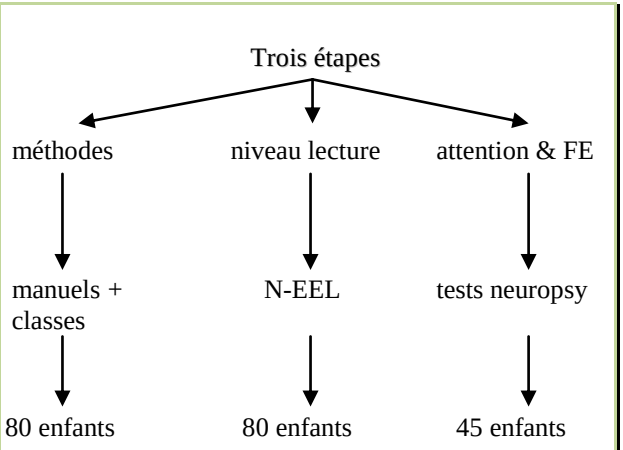
Nos premières observations et analyses ont porté sur quatre-vingts enfants issus de quatre classes de première année primaire de l'enseignement ordinaire. Ensuite, une fois diverses variables contrôlées telles que : le niveau en lecture, le niveau socio-culturel et l'absence de troubles dysexécutifs, nous avons poursuivi notre recherche avec 45 enfants, dont 23 de sexe masculin et 22 de sexe féminin.

3.2. Matériel

Il semblait nécessaire, avant tout, de connaître les compétences de nos sujets en lecture. En effet, le niveau d'efficience en lecture peut grandement influencer la manière dont les fonctions attentionnelles et exécutives seront sollicitées. Ainsi, un faible lecteur devra solliciter davantage ces fonctions pour lire un mot qu'un bon lecteur. Pour cette raison, tout enfant présentant soit un retard important en lecture, soit un trouble avéré de la lecture tel que la dyslexie s'est vu écarté de l'échantillon. Toutefois, ne disposant pas d'outil permettant d'évaluer la lecture chez des enfants de 6 ans, l'évaluation des compétences en lecture a été réalisée à l'aide de la N-EEL (Nouvelles Epreuves pour l'Examen du Langage ; Chevrie-Muller & Plaza [11]). Cette batterie n'est pas un test de lecture en soi, cependant elle permet d'évaluer certaines compétences langagières requises par la lecture. Ensuite, nous avons sélectionné plusieurs batteries ou subtests permettant d'évaluer les fonctions attentionnelles et exécutives chez l'enfant : la Nepsy (bilan Neuropsychologique de l'enfant ; Korkman, Kirk & Kamp [12]), le subtest « symboles » du Wisc IV (Wechsler [13]), la Tea-ch (Manly, Robertson, Anderson & Mimmo-Smith [14]) ainsi que le Trail Making test (Spreen & Strauss [15]).

3.3. Procédure

Notre recherche s'est déroulée en trois étapes principales (voir figure 1). Dans une première étape, nous avons déterminé les types de méthodes d'enseignement de la lecture proposés aux enfants. Pour ce faire, nous avons observé les activités de lecture au sein même des quatre classes et emprunté les manuels utilisés par les enseignants. À l'issue de ces observations, nous avons pu constater la présence des trois méthodes d'enseignement de la lecture précédemment décrites. Dans une deuxième étape, nous avons évalué le niveau en lecture de chaque enfant. Enfin, dans une troisième étape, les enfants ayant un niveau similaire en lecture ont été soumis aux épreuves évaluant les fonctions attentionnelles et exécutives. La durée de passation moyenne pour ces épreuves s'élevait à 1h30 par enfant, répartie en trois séances afin que l'enfant participant à la recherche ne cumule pas un retard trop important sur la matière abordée en classe.



***légende** : FE = Fonctions exécutives ; tests neuropsy = tests neuropsychologiques*

Figure 1 : procédure

4. RÉSULTATS

Les tests paramétriques suivants ont été utilisés lorsque la répartition de nos données le permettait : analyse de variance simple, test t de Student, coefficient r de Bravais Pearson. Dans le cas où la répartition des données empêchait l'utilisation des tests paramétriques, les tests non paramétriques suivants ont été choisis : test Kruskal-Wallis, test Mann-Whitney, test Wilcoxon, coefficient rho de Spearman.

4.1. analyses inter-méthodes

Les résultats statistiques, indiquent qu'aucun groupe ne présente de compétences supérieures concernant les fonctions attentionnelles, l'inhibition et la planification (voir table 1). Cependant, les enfants soumis à la méthode mixte ont une meilleure flexibilité cognitive que les enfants confrontés à la méthode synthétique (Z : -2.375 ; p = .018).

Table 1 : résultats statistiques des analyses inter-méthodes

	synth/glob	glob/mixte	synth/mixte
flexibilité	.485 (NS)	.062 (NS)	.018*
attent. sout.	.287 (NS)	.150 (NS)	.676 (NS)
attent. div.	.786 (NS)	.480 (NS)	.654 (NS)
attent. sél.	.735 (NS)	.497 (NS)	.322 (NS)
planification	.799 (NS)	.187 (NS)	.328 (NS)
inhibition.	.373 (NS)	.599 (NS)	.089 (NS)

***légende** : synth = synthétique ; glob = globale ; attent. = attention ; sout = soutenue ; div = divisée ; sél = sélective*

4.2. Analyses intra-méthodes

Les analyses statistiques indiquent que selon les méthodes, le type de modalité sensorielle étudié influencera ou non les résultats aux fonctions cognitives considérées.

1/ Les enfants soumis aux méthodes synthétiques présentent une meilleure inhibition auditive (Z = -2.412 ; p = .016) que visuelle.

2/ Les enfants confrontés aux méthodes globales ont de meilleures compétences en inhibition auditive (Z = -2.703 ; p = .007) et en flexibilité auditive (t = 2.581 ; p = .014).

3/ Les enfants soumis aux méthodes mixtes ont de meilleures aptitudes en inhibition auditive (Z = -2.157 ; p = .031) et en flexibilité visuelle (Z = -2.449 ; p = .014).

4.3. Analyse des corrélations entre les outils

Les résultats statistiques indiquent la présence de 13 corrélations significatives entre les outils neuropsychologiques administrés. Ces résultats démontrent soit une relation forte entre les fonctions cognitives investiguées, soit une faible validité des outils utilisés.

5. CONCLUSIONS ET PERSPECTIVES

L'objectif de notre recherche était de déterminer si les méthodes d'enseignement de la lecture impliquaient différemment les fonctions attentionnelles et exécutives. De manière générale, les résultats indiquent qu'aucune méthode n'implique davantage ces fonctions cognitives. Toutefois, nous observons également que les méthodes mixtes sollicitent davantage la flexibilité (visuelle et auditive) que les méthodes synthétiques. L'explication la plus appropriée pour justifier ce résultat significatif concerne le développement de la composante de flexibilité. Cette dernière serait, comparativement aux autres composantes exécutives, celle qui se développe le plus lentement (Blaye & Lemaire [16]). Alors que certaines composantes telles que la mémoire de travail ou l'inhibition atteignent un niveau adulte rapidement au cours de l'enfance, la flexibilité poursuit son développement jusqu'à l'adolescence, voire l'âge adulte. De par sa lenteur de développement, nous pensons que la flexibilité serait susceptible d'être plus sensible à divers facteurs parmi lesquels nous pourrions retrouver les méthodes d'enseignement de la lecture. Le fait qu'une approche mixte développe davantage la flexibilité visuelle n'est guère étonnant car, par définition, l'approche mixte requiert que l'enfant passe régulièrement d'un mode de traitement de l'écrit à un autre afin de rendre la lecture la plus efficiente possible.

Afin d'expliquer l'absence d'influence significative des méthodes sur les fonctions attentionnelles, des éléments théoriques peuvent être proposés. Ainsi, nous savons que l'activité de lecture nécessite notamment de bonnes capacités attentionnelles, exécutives et mnésiques. S'il est vrai que toutes ces fonctions sont

primordiales, elles seules ne permettent cependant pas l'efficience en lecture. En effet, apprendre à lire nécessite beaucoup d'aptitudes élémentaires telles que la connaissance des caractéristiques des lettres, la conversion des lettres en sons en vue de prononcer un mot ainsi que la compréhension de chacun des mots. Selon la théorie proposée par LaBerge et Samuels (cités par Reed [17]), la capacité à pouvoir acquérir des aptitudes complexes telles que la lecture dépend notamment de la capacité de traitement automatique. Selon ces auteurs, certaines aptitudes dites élémentaires peuvent être affectées sans nécessiter de ressources car, dans le cas contraire, la demande totale issue de toutes les aptitudes élémentaires serait trop importante pour que l'individu puisse réaliser la tâche. Une des aptitudes élémentaires présente essentiellement au début de l'apprentissage de la lecture, du moins pour certaines approches, est l'identification des caractéristiques d'une lettre. Ces caractéristiques doivent ensuite être combinées pour former une lettre, ce qui nécessite de l'attention. Cependant, après une pratique suffisante de la reconnaissance des lettres, les ressources attentionnelles peuvent être mobilisées pour d'autres aptitudes élémentaires. Si nous transférons ces éléments théoriques à notre étude, nous constatons que pour chaque méthode d'enseignement de la lecture, traitements automatiques et traitements élaborés se côtoient. Ainsi, aucune méthode ne semble solliciter davantage de traitements élaborés, ce qui expliquerait l'absence de différences significatives entre les performances de nos groupes aux épreuves évaluant les fonctions attentionnelles.

L'absence de résultats significatifs peut également se justifier par des éléments plus méthodologiques. Ainsi, pour chacune des fonctions cognitives investiguées, nous pouvons remettre en cause les tests utilisés, soit pour leur qualités psychométriques, soit parce qu'ils mesurent davantage de fonctions cognitives que prévu. Ainsi, les tests psychométriques ne sont jamais purs ; ils mettent en jeu une diversité de processus cognitifs. Ceci peut s'expliquer par la rapidité avec laquelle les nouveaux instruments se multiplient ainsi que par l'utilisation encore trop fréquente chez les enfants de tests conçus pour les adultes.

Pour conclure, nous pensons qu'à l'avenir, il pourrait être intéressant d'investiguer à nouveau la relation entre les méthodes d'enseignement de la lecture et le fonctionnement cognitif, non sans une certaine prudence à l'égard des épreuves neuropsychologiques.

BIBLIOGRAPHIE

[1] J.F. Le Ny. Lecture et pédagogie. In *Colloque de Tours*, pages 94-103, 1972.

[2] J. Charpentier. *Apprentissage de la lecture et développement de la pensée logique*. Presses universitaires de France, Paris, France, 1992.

[3] L. Lefebvre. Compétences cognitives et méthodes de lecture. *Psychologie & Education*, 4 : 71-91, 2006.

[4] N. Van Grunderbeeck. *Les difficultés en lecture, diagnostic et pistes d'intervention*. De Boeck, Bruxelles, Belgique, 1994.

[5] S. Van der Sluis, P.F. De Jong and A. Van der Leij. Executive functionning in children, and its relations with reasoning, reading and arithmetic. *Intelligence*, 35 : 427-449, 2006.

[6] M. Mazeau. *Conduite du bilan neuropsychologique chez l'enfant*. Masson, Paris, France, 2003.

[7] M-N Metz-Lutz, E. Demont, C. Seegmuller, M. De Agostini and N. Bruneau. *Développement cognitif et troubles des apprentissages : évaluer, comprendre, rééduquer et prendre en charge*. Solal, Marseille, France, 2004.

[8] J. Fijalkow. *Mauvais lecteurs : pourquoi ?* Presses Universitaires de France, Paris, France, 1986.

[9] L. Boulc'h, C. Gaux and C. Boujon. Implication des fonctions exécutives dans le décodage en lecture : étude comparative entre normolecteurs et faibles lecteurs de CE2. *Psychologie française*, 52 : 71-87, 2006.

[10] M. Brosnan, J. Demetre, S. Hamil, K. Robson, H. Sheperd and G. Cody. Executive functioning in adults and children with developmental dyslexia. *Neuropsychologia*, 40 : 2144–2155, 2002.

[11] C. Chevrie-Muller and M. Plaza. *Nouvelles épreuves pour l'examen du langage*. Editions du Centre de Psychologie Appliquée, Paris, 2001.

[12] M. Korkman, U Kirk and S. Kemp. *Bilan neuropsychologique de l'enfant : NEPSY*. Editions du Centre de Psychologie Appliquée, Paris, 2003.

[13] D. Wechsler. *Echelle d'intelligence de Wechsler pour enfants et adolescents*. Editions du Centre de Psychologie Appliquée, Paris, 2005.

[14] T. Manly, I.H. Robertson, V. Anderson and I. Mimmo-Smith. *Test d'évaluation de l'attention chez l'enfant.* Editions du Centre de Psychologie Appliquée, Paris, 2006.

[15] O. Spreen and E. Strauss. *A compendium of neuropsychological tests. Administration, norms and commentary.* Oxford University Press, Oxford, UK, 1998.

[16] A. Blaye and P. Lemaire. *Psychologie du développement cognitif de l'enfant*. De Boeck, Bruxelles, Belgique, 2007.

[17] S.K. Reed. Cognition. *Théories et applications*. De Boeck Université, Bruxelles, Belgique, 1999.

Corrélats neurocognitifs de la perception de la focalisation prosodique contrastive en français

Marcela Perrone[1], Marion Dohen[2], Hélène Loevenbruck[2], Marc Sato[2], Cédric Pichat[1], Gaëtan Yvert[1], Monica Baciu[1]

[1] Laboratoire de Psychologie et NeuroCognition, UMR CNRS 5105, UPMF, Grenoble, France
[2] Département Parole et Cognition, GIPSA-lab UMR CNRS 5216, Grenoble Universités, France

ABSTRACT

The present event-related functional magnetic brain imaging (fMRI) study deals with the perception of prosodic contrastive focus in French. Twenty-two right-handed French speakers participated in the experiment. The two conditions consisted in the auditory judgement of two kinds of utterances: with contrastive prosodic focus (Focus condition, Task) and without (Neutral condition, Control). The participants had to judge whether the utterances contained focus or not. Our results suggest that both hemispheres participate in the auditory perception of contrastive prosodic focus, but with a left-dominant contribution for morpho-syntactic processes and thematic role monitoring.

Keywords: prosody, prosodic contrastive focus, left hemisphere specialization.

1. INTRODUCTION

Les résultats de précédentes études sur la perception de la prosodie et ses corrélats neuronaux ne permettent pas de conclure sur une possible latéralisation hémisphérique. En effet, les résultats de certains travaux confortent la conception traditionnelle de la prosodie comme subordonnée à l'hémisphère droit– HD – [1] d'autres études suggèrent que le traitement de la prosodie n'est pas limité à ce seul hémisphère [2]. Très peu d'études en neuroimagerie se sont intéressées de manière spécifique à la focalisation prosodique. La focalisation contrastive est employée pour mettre en relief un mot (ou groupe de mots) par opposition à un autre dans l'énoncé. En français, elle peut être réalisée grâce à la prosodie (intonation, rythme, phrasé) : « $\text{Nico}_{\mathbf{F}}$ cassait le vélo » (voir Dohen & Loevenbruck, [3] sur les corrélats acoustiques de la focalisation prosodique). Parmi les rares études qui se sont penchées sur les corrélats neuroanatomiques fonctionnels de la perception de la focalisation prosodique, celle de Wildgruber et al. [4] compare deux types de prosodie : affective et linguistique. Pour explorer la prosodie linguistique, ces auteurs ont utilisé une tâche de détection indirecte de la focalisation (*trouver la réponse la plus appropriée à une question précise*). Pour explorer la prosodie affective, ils ont utilisé une tâche d'évaluation de l'expressivité émotionnelle. Lorsqu'ils ont comparé ces deux conditions (prosodie linguistique vs. affective), ils ont observé des activations latéralisées dans l'HG du cortex frontal inférieur (aire de Broca). Tong et al. [5] ont examiné le traitement de la focalisation prosodique dans le but de différencier le traitement du mode intonatif (énoncés interrogatifs vs. affirmatifs) de celui de la focalisation contrastive. Des participants anglais et chinois ont été examinés. Pour le traitement de la focalisation contrastive, les deux groupes de participants ont montré des activations bilatérales du sillon intra-pariétal (BA 40/7) et des activations prédominantes dans l'HD du cortex dorsolatéral préfrontal (BA 9/46). Pour le groupe des Chinois, ils ont de plus observé des activations prédominantes dans l'HG du gyrus supramarginal et du cortex temporal moyen dans sa partie postérieure (BA 21/20/37). L'étude en IRMf (Imagerie par Résonance Magnétique fonctionnelle) présentée ici a pour but de préciser, par localisation anatomo-fonctionnelle, les mécanismes cérébraux et cognitifs mis en œuvre lors de la perception de la focalisation contrastive prosodique en français.

2. MÉTHODES

2.1. Participants

Vingt-deux volontaires adultes (11 femmes, moyenne d'âge = 27,45 ans ± 3,48) ont participé à l'expérience. Tous les participants, étaient droitiers et de langue maternelle française.

2.2. Stimuli

L'expérience comportait deux conditions : une condition dans laquelle un des éléments de la phrase présentée était focalisé (Condition Focalisé, F) et une autre condition dans laquelle aucun élément de la phrase n'était focalisé (Condition Neutre, N). Le corpus utilisé comportait 24 phrases en français (durée moyenne de 2 secondes). Toutes les phrases avaient la même structure syntaxique et syllabique: Sujet (S: prénom de 2 syllabes) - Verbe (V: verbe à l'imparfait de 2 syllabes) - Objet (O: article défini de 1 syllabe + nom commun de 2 syllabes), tel que dans l'exemple suivant: «Nico cassait le vélo». Tous les constituants de la phrase avaient une structure syllabique de type CVCV. D'autre part, nous avons contrôlé la fréquence

lexicale des noms et verbes utilisés (http://www.lexique.org/). Par ailleurs, pour la condition F, les phrases ont été enregistrées une fois en focalisation sujet (FS : $Nico_F$ cassait le vélo) et une fois en focalisation objet (FO : Nico cassait le $vélo_F$). Au total, 72 énoncés ont ainsi été enregistrés dans une chambre sourde par une locutrice experte de langue maternelle française (24 phrases, trois types de focalisation : N, FS et FO).

2.3. Tâche

Il était demandé aux sujets de déterminer si la phrase entendue comportait une focalisation prosodique ou non. Leurs réponses manuelles étaient enregistrées afin d'évaluer leurs performances.

2.4. Paradigme IRMf

Les stimuli ont été présentés à l'aide du logiciel E-Prime (E-Prime Psychology Software Tools Inc, Pittsburgh, USA) sur un ordinateur PC. Les sujets entendaient les stimuli auditifs grâce à un casque audio compatible IRM. Au total, 96 stimuli ont été présentés dans un ordre aléatoire (i.e., F : 48 phrases incluant - 24 FS et 24 FO -, N: 24 phrases présentées deux fois chacune). Un paradigme événementiel pseudo-aléatoire (optimisé selon la méthode développée par [6]) à été utilisé pour la présentation des stimuli dans chacune des conditions. La session fonctionnelle était composée de 48 événements par condition prosodique. Par ailleurs, 30 événements « nuls » ont également été inclus afin d'établir une mesure de référence appropriée [6]. Ces derniers consistaient en la présentation d'une croix de fixation au centre de l'écran. L'intervalle inter-stimuli était de 4 secondes.

2.5. Acquissions IRM

Les données IRM ont été acquises en utilisant un scanner corps entier 3T (Bruker MedSpec S300). Les scans fonctionnels ont été réalisés en utilisant une séquence EPI en écho de gradient pondéré en T2* (39 coupes axiales adjacentes parallèles au plan bi-commissural, mode entrelacé, taille de voxel 3×3×3.5 mm^3, TR = 3 s, TE = 40 ms, angle de bascule = 77°). Une carte de champ a été acquise pour mesurer les inhomogénéités locales du champ magnétique B0. Enfin, un volume anatomique de haute résolution a été acquis en utilisant une séquence pondérée en T1 (champ de vue = 256×224×176 mm ; résolution : 1.333×1.750×1.375 mm^3).

2.6. Analyse des données IRMf

L'analyse des données a été réalisée en utilisant le modèle linéaire général (MLG) sous SPM5 (www.fil.ion.ucl.ac.uk/spm). Les prétraitements suivants ont d'abord été réalisés : *a)* corrections liées au décalage temporel lors de l'acquisition des volumes fonctionnels, *b)* corrections liées aux mouvements des sujets et liées à la distorsion du champ EPI (application des cartes de champ individuelles) , *c)* normalisation spatiale permettant d'établir la correspondance spatiale entre le volume anatomique du sujet et un modèle de référence, d*)* lissage de chaque volume fonctionnel par un filtre gaussien passe-bas. Un filtrage passe-haut a été appliqué pour supprimer le bruit induit par les basses fréquences et la dérive du signal pour chaque voxel. Après ces étapes de prétraitement, nous avons réalisé une analyse statistique sur les images fonctionnelles. Les deux conditions d'intérêts (F vs. N) ont été modélisées comme deux régresseurs convolués à une fonction canonique estimée de la réponse hémodynamique (HRF). L'activité de chaque voxel pour chacune des conditions et pour chacun des participants a été estimée grâce au MLG. Au niveau individuel, nous avons effectué la comparaison F vs. N afin d'évaluer les différentes régions impliquées de manière spécifique dans la détection de la focalisation prosodique. Les contrastes résultants de l'analyse individuelle ont été utilisés pour réaliser une analyse de groupe à effet aléatoire en utilisant un test t. Les groupes de voxels activés (15 voxels adjacents) ont été identifiés sur la base de l'intensité de la réponse individuelle ($p < 0.001$, non corrigé, $T = 3,53$). Les régions activées pour chaque condition ont été identifiées en fonction de leurs coordonnées Talairach. Enfin, nous avons défini un ensemble de régions d'intérêt (ROI) issues de l'activation obtenues par le contraste F vs N et en tenant compte des résultats obtenus d'études antérieures [5, 7]. Plus précisément, nous avons retenu tous les voxels activés inclus à l'intérieur d'une sphère de 5 mm de diamètre autour de chaque pic d'activation, et ce dans les deux hémisphères cérébraux. Les pics d'activation sont présentés dans la Table 2. Pour construire les ROI nous avons utilisé le logiciel Marsbar (http://marsbar.sourceforge.net/). L'intensité de variation du signal MR (MR%, estimations de paramètres) de chaque ROI a ensuite été extraite. Les valeurs des MR% pour chaque ROI et pour chaque participant, ont été incluses dans une analyse de variance (ANOVA) avec le facteur hémisphère comme variable intra-sujet (HD vs HG), afin d'identifier une prédominance hémisphérique éventuelle dans ces différentes régions d'intérêt pour le traitement de la focalisation prosodique.

3. RÉSULTATS

3.1. Résultats comportementaux

Les réponses comportementales enregistrées pendant l'expérience montrent que les sujets ont réussi à détecter la focalisation correctement : pourcentages de réponses correctes : F (M = 92.99%, ET = 6,73%) et N (M = 97,72%, ET = 3,85%).

Région	H	BA	k	x	y	z	T
Cortex prémoteur	G	6	108	-3	34	37	8.14
Gyrus frontal inférieur	G	47	80	-50	21	2	6.55
	D	47	67	53	29	-4	5.32
Gyrus frontal inférieur	D	44	47	50	13	19	5.66
Gyrus temporal moyen	D	21	352	62	-52	6	6.11
Gyrus temporal supérieur	G	22	78	-59	-32	5	4.91
Gyrus supramarginal (pariétal)	G	40	90	-50	-53	41	5.70
	D	40	27	50	-32	50	4.29
Gyrus pariétal supérieur	G	7	55	-42	-67	52	5.16
Insula (partie antérieure)	G	13	18	-33	18	2	4.85

Table 1 : Régions activées lors de l'analyse à effet aléatoire ($p > 0.001$, non corrigé) pour le contraste F (focalisé) > N (neutre). H = hémisphère, D = droite, G = gauche, k = nombre de voxels dans chaque cluster; BA = aire de Brodmann.

3.2. Résultats en IRMf

3.2.1 Principaux contrastes entre les conditions

Le contraste F vs. N (c.f. Table 1 et Figure 1, panel A) révèle des activations bilatérale des cortex frontaux, temporaux et pariétaux. Concernant le cortex frontal, nous avons observé des activations du gyrus prémoteur gauche (BA 6) et des gyri frontal inférieur gauche (BA 47) et droit (BA 44, 47). Concernant le cortex temporal, nous avons observé des activations du gyrus temporal supérieur gauche (BA 22), des gyri temporal supérieur et moyen droit (BA 21). Enfin, nous avons également observé des activations bilatérales du gyrus supramarginal (BA 40) et du lobule pariétal supérieur gauche (BA 7).

Région d'intérêt	BA	x	y	z	**F**	***p***
Cortex prémoteur	6	± 3	33	37	0.49	0.49
Gyrus frontal inférieur	44	± 50	12	18	1.33	0.26
Gyrus frontal inférieur	47	± 53	29	-4	6.04	**0.02**
Insula	13	± 33	18	2	5.32	**0.03**
Gyrus temporal supérieur	22	± 59	-31	4	2.9	0.1
Gyrus temporal moyen	21	± 62	-52	5	0.21	0.64
Gyrus supramarginal	40	± 50	-53	41	6.97	**0.01**

Table 2 : Coordonnées des pics d'activation pour les Régions d'Intérêt et valeurs statistiques (F et p) de la différence entre les hémisphères cérébraux pour chaque ROI.

3.2.2 Paramètres estimés (% de variation du signal RM) dans les ROIs

Les résultats de l'analyse réalisée pour chacune des ROI (Table 2 et Figure 1, panneau B) montrent une prédominance de l'HG pour le traitement de la focalisation prosodique dans le gyrus frontal inférieur (BA 47), le gyrus supramarginal (BA 40) et l'insula antérieur (BA 13). Sur les ROI temporaux étudiées, aucune prédominance hémisphérique n'était significative.

4. DISCUSSION ET CONCLUSION

Nos résultats montrent que le traitement de la focalisation contrastive implique de manière prédominante l'HG dans les ROI désignées, suggérant ainsi que le traitement de la prosodie n'est pas strictement latéralisé à droite comme l'ont proposé plusieurs études.

Parmi les régions d'intérêt observées comme étant activées de façon prédominante à gauche, nous avons trouvé des activations du gyrus frontal inférieur (BA 47, LIFG), du gyrus supramarginal (BA 40) et de l'insula antérieur (BA 13, aINS). Le LIFG est classiquement impliqué dans les traitements phonologiques et sémantiques [8], mais également dans des tâches faisant appel au jugement syntaxique. D'après [9], l'attribution de rôle thématique implique un traitement lexico-sémantique et morpho-syntaxique qui recruterait à la fois la partie antérieure (BA 45/47) et la partie postérieure (BA 44/45) du LIFG (i.e., aire de Broca). Le LIFG a également été désigné par [7] comme étant impliqué dans cette fonction de suivi du rôle thématique lors de la production de la parole. Cette dernière interprétation est renforcée par nos résultats en perception. L'INS gauche est généralement considérée comme impliquée dans les processus de mémoire de travail verbal [10]. Elle pourrait donc intervenir dans la détection de la focalisation via une répétition interne de la phrase cible. Le lobule pariétal supérieur gauche a été montré comme impliqué dans la production de différentes modalités du pointage, dont le pointage avec la voix réalisé par la focalisation prosodique, le pointage manuel et le pointage oculaire. [11]. Il a été suggéré que lors de la production de focalisation prosodique, les locuteurs utiliseraient des représentations multisensorielles, tout comme s'ils produisaient des pointages manuel ou oculaire. Ces représentations pourraient se former dans les régions associatives pariétales. Nos résultats actuels suggèrent que la perception de la focalisation prosodique nécessite également l'association de représentations multimodales pour pouvoir détecter des indices de la focalisation. Il est enfin à noter l'importante activation observée dans le gyrus temporal moyen de l'HD, un résultat qui suggère que le traitement du contour mélodique se réaliserait dans l'HD de manière prédominante, alors que les processus de décision linguistique seraient réalisés de manière prédominante par l'HG.

En résumé, nos résultats suggèrent que les deux hémisphères participent à la perception auditive de la prosodie, avec une contribution prédominante de l'HG pour les processus de nature morpho-syntaxique et les processus d'attribution du rôle thématique.

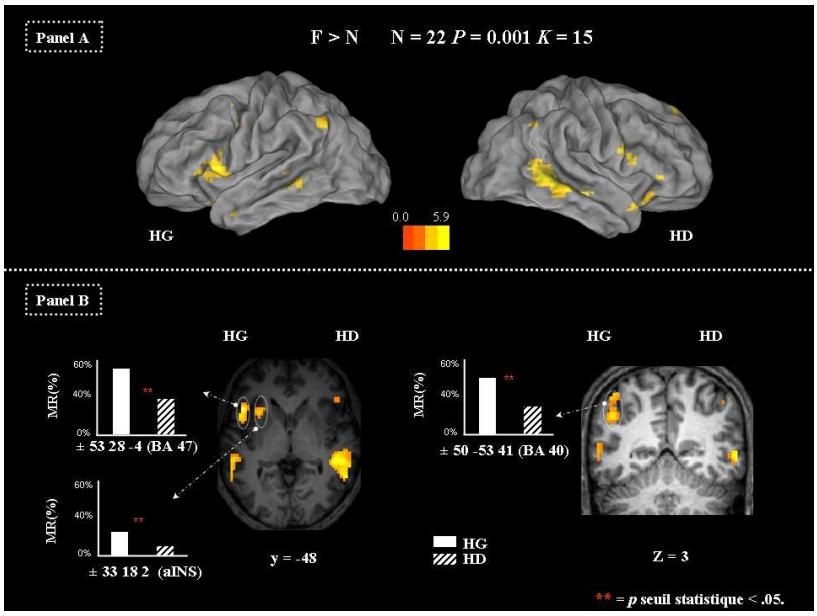


Figure 1 : Activations obtenues lors du contraste (F)ocalisé vs. (N)eutre. Panneau A : activations projetées sur des gabarits anatomiques en 3D. HD = hémisphère droit, HG = hémisphère gauche. Panneau B : Analyse dans les ROI : les graphiques montrent les différences significatives d'activités entre l'HG et l'HD pour chacune des ROI. Nous avons obtenu une prédominance gauche significative dans le gyrus frontal inférieur (BA 47), le gyrus supramarginal (BA 40) et dans l'insula antérieure (aINS,).

5. REMERCIEMENTS

Ce projet est soutenu par le Cluster 11 "Vieillissement, handicap, neurosciences" de la région Rhône-Alpes et la FFRE (Fondation Française pour la Recherche sur l'Epilepsie).

6. BIBLIOGRAPHIE

[1] N. Dronkers, S. Pinker, and A. Damasio. "*Language and the aphasias*", in E.R. Kandel, J.H. Schwartz & T.M. Jessel [Eds], Principles of neural science, 1169-1187, Mc Graw-Hill, 2000.

[2] C. Astésano, M. Besson and K. Alter. "Brain potentials during semantic and prosodic processing in French", *Cognitive Brain Research*, 18: 172-184, 2004.

[3] Dohen M. & Lœvenbruck H., 2004. Pre-focal Rephrasing, Focal Enhancement and Post-focal Deaccentuation in French. In Proc. 8th Intl. Conf. on Spoken Language Processing, volume 1, pages 785-788, 2004.

[4] D. Wildgruber, I. Hertrich, A. Riecker, M. Erb, S. Anders, W. Grodd and H. Ackermann. "Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation", *Cerebral Cortex*, 14: 1384-1389, 2004.

[5] Y. Tong, J. Gandour, T. Talavage, D. Wong, M. Dzemidzic, Y. Xu, X. Li and M. Lowe. "Neural circuitry underlying sentence-level linguistic prosody", *NeuroImage*, 28: 417-428, 2005.

[6] KJ. Friston, E. Zarahn, O. Josephs, R. Henson and A. Dale. "Stochastic Designs in Event-Related fMRI", *NeuroImage*, 10(5): 607-619, 1999.

[7] Loevenbruck, M. Baciu, C. Segebarth and C. Abry. "The left inferior frontal gyrus under focus: an fMRI study of the production of deixis via syntactic extraction and prosodic focus", *Journal of Neurolinguistics*, 18(3): 237-258, 2005.

[8] . Fiez. "Phonology, semantics, and the role of the left inferior prefrontal cortex", *Humand Brain Mapping*, 5: 79-83, 1997.

[9] A. Friederici. "Towards a neural basis of auditory sentence processing", *Trends in Cognitive Sciences*, 6 (2): 78-84, 2002.

[10] M. Sato, M. Baciu, H. Lœvenbruck, J-L. Schwartz, M-A. Cathiard, C. Segebarth and C. Abry. "Multistable representation of speech forms: An fMRI study of verbal transformations", *NeuroImag*e, 23(3): 1143-1151, 2004.

[11] H. Loevenbruck, M. Dohen and C. Vilain. "*Pointing is 'special"*, in S. Fuchs, H. Lœvenbruck D. Pape & P. Perrier [Eds], Some Aspects of Speech and the Brain, 211-258, Peter Lang, 2009.

Reconnaissance du Locuteur basée sur des Signatures Glottiques

Thomas Drugman, Thierry Dutoit

TCTS Lab - Faculté Polytechnique - Université de Mons
31, Boulevard Dolez - 7000 Mons - Belgique

ABSTRACT

The great majority of current speaker recognition systems are based on features related to the vocal tract. However some studies have shown that the glottal flow conveys relevant information about the speaker identity. This paper proposes the use of some glottal signatures in speaker recognition. These signatures are extracted from a speaker-dependent dataset of pitch-synchronous residual frames. Experiments of speaker identification are led on both TIMIT and YOHO databases. It is shown that the proposed approach outperforms other state-of-the-art methods based on glottal features.
Keywords : Speaker Recognition, Glottal Analysis, Residual Signal, Voiceprint

1. INTRODUCTION

Développer un système de reconnaissance du locuteur efficace implique une bonne connaissance de ce qui définit l'individualité d'un locuteur. Bien que des informations de haut niveau (comme par exemple l'usage de mots) puissent être envisagées, des attributs acoustiques de bas niveau sont généralement utilisés [11]. Ces derniers sont, la plupart du temps, extraits du spectre d'amplitude du signal de parole. Ils visent à paramétriser la contribution du conduit vocal, qui est une caractéristique importante de l'identité du locuteur. D'un autre côté, très peu de travaux ont étudié la possibilité d'utiliser en reconnaissance du locuteur des attributs émanant de la source glottique. Pourtant des différences significatives dans les formes d'onde glottiques ont été observées entre différents types de locuteurs [6].

Principalement, deux signaux véhiculent de l'information quant au comportement de la glotte : le flux glottique et le signal résidu. Le flux glottique est le débit d'air expulsé dans la trachée et passant à travers les cordes vocales. Son estimation directement à partir du signal de parole est un problème typique de séparation aveugle, puisqu'aucune des contributions glottique et du conduit vocal ne sont observables. Il est donc requis d'adopter un processus d'estimation incorporant une connaissance précise du mécanisme de production. De cette façon, le flux glottique peut être estimé, par exemple, par une analyse spectrale sur la phase fermée de la glotte. Par cette technique, Plumpe et al. [10] ont extrait un ensemble d'attributs temporels paramétrisant le flux glottique ainsi estimé. Dans un canevas similaire, Gudnason et al. [5] ont caractérisé le flux glottique par des coefficients de cepstre réel. Ces deux approches ont abouti à une amélioration, en termes d'identification du locuteur, en combinant ces paramètres glottiques à des attributs extraits du spectre d'amplitude de la parole (tels que les coefficients LP ou MFCC). D'un autre côté, le signal résidu désigne le signal obtenu par filtrage inverse, après avoir enlevé la contribution de l'enveloppe spectrale. Le signal résidu qui en résulte véhicule de l'information pertinente quant à l'excitation et, contrairement au flux glottique, a l'avantage d'être obtenu facilement. Dans [12], Thevenaz et al. ont suggéré d'utiliser, en vérification du locuteur, des coefficients LPC du signal résidu. Plus récemment, Murty et al. [8] ont mis en évidence, en reconnaissance du locuteur, la complémentarité de la phase résiduelle avec les MFCCs conventionnels. Dans cette dernière étude, l'information contenue dans la phase résiduelle a été extraite via des réseaux de neurones.

Le but de cet article est d'étudier la potentialité d'utiliser des *signatures glottiques* en reconnaissance du locuteur. La recherche d'un invariant dans le signal de parole, caractérisant univoquement une personne (comme pour les empreintes digitales), a toujours attiré la communauté scientifique [7]. Comme ceci semble utopique dû à la nature inhérente du mécanisme de phonation, nous préférons ici le terme de "*signature vocale*" pour désigner un signal contenant une information pertinente quant à l'identité du locuteur. Cet article est structuré comme suit. En Section 2, nous détaillons la façon d'extraire ces signatures vocales à partir du signal de parole et de les inclure dans un système de reconnaisance du locuteur. La Section 3 présente des résultats d'identification du locuteur menés sur les bases de données TIMIT et YOHO. Finalement, la Section 4 conclut cet article.

2. SIGNATURES GLOTTIQUES

2.1. Signatures Glottiques utilisées dans cette Etude

Les *signatures glottiques* utilisées dans cette étude proviennent du Modèle Déterministe plus Stochastique (DSM) du signal résidu que nous avons proposé dans [3] pour la synthèse paramétrique de parole. Ce modèle émane d'une analyse menée sur un ensemble de trames de résidu normalisées et pitch-synchrones. La Figure 1 présente le diagramme utilisé pour obtenir cet ensemble particulier de données à partir d'une collection d'enregistrements d'un locuteur donné. Tout d'abord, une analyse de prédiction linéaire (LP) classique, capturant l'enveloppe spectrale, est réalisée sur les signaux de parole. Les résidus sont ensuite obtenus par filtrage inverse. Les instants de fermeture glottique (GCIs) sont

alors identifiés en localisant les discontuinités les plus marquées dans le signal résidu, comme expliqué dans [2]. En parallèle, le pitch est estimé via la librairie Snack Sound Toolkit [9], disponible publiquement. Les trames de résidu pitch-synchrones sont ensuite isolées par un fenêtrage de Blackman centré sur un GCI et long de 2 périodes de pitch. Les trames résultantes sont finalement normalisées en prosodie, c-à-d à la fois en pitch et énergie. Cette opération de normalisation en pitch est réalisée par décimation/interpolation sur un nombre fixé d'échantillons (de façon à ce que les trames de résidu aient toutes la même longueur).

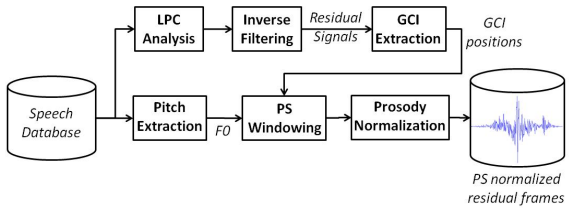


FIGURE 1: Diagramme permettant d'obtenir, pour un locuteur donné, un set de trames de résidu normalisées et pitch-synchrones.

Une fois que le set de trames de résidu est disponible, certaines caractéristiques dépendantes du locuteur et liées au modèle DSM sont extraites sur celui-ci. D'après ce modèle [3], le signal résidu voisé $r(t)$ est composé d'une structure déterministe basses-fréquences $r_d(t)$ et d'une composante stochastique hautes-fréquences $r_s(t)$, supposée modéliser principalement les turbulences présentes dans le débit d'air glottique. Le spectre est donc divisé en deux bandes délimitées par la *fréquence maximale de voisement* F_m (fixée à $4kHz$ au sein de cette étude). Le signal résidu synthétisé est alors obtenu comme décrit en Figure 2. La partie déterministe est modélisée par une forme d'onde unique dépendante du locuteur et appelée *premier résidu propre*. Cette forme d'onde est définie comme le premier vecteur propre obtenu par application d'une Analyse en Composantes Principales (PCA) sur le set de trames de résidu. Quant à la composante stochastique, elle est modélisée par un bruit Gaussien hautes-fréquences modulé temporellement par une enveloppe d'énergie pitch-synchrone. Cette *enveloppe d'énergie* est extraite du set de données précédent en moyennant l'enveloppe de Hilbert du contenu hautes-fréquences des trames de résidu.

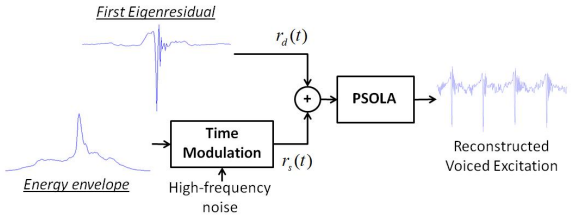


FIGURE 2: Reconstruction de l'excitation voisée selon le Modèle Déterministe plus Stochastique (DSM) du signal résidu. Les 2 signatures glottiques utilisées dans ce travail sont le premier résidu propre et l'enveloppe d'énergie.

En conclusion, le modèle DSM du signal résidu fait usage de deux formes d'onde dépendantes du locuteur, ci-après nommées *signatures glottiques* : le premier résidu propre (ou *résidu propre* tout court) et l'enveloppe d'énergie. La Figure 3 illustre la forme de l'enveloppe d'énergie pour deux locuteurs masculins. Des différences dans les formes d'onde suggèrent que les signatures glottiques proposées ont le potentiel pour être utilisées en reconnaissance automatique du locuteur.

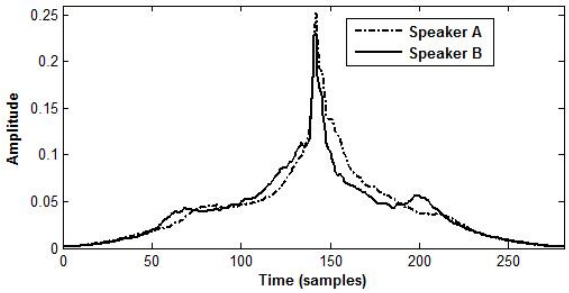


FIGURE 3: Formes d'onde de l'enveloppe d'énergie pour deux locuteurs masculins différents.

2.2. *Intégration des signatures Glottiques en Identification du Locuteur*

Afin d'être incorporées dans un système d'identification du locuteur, les signatures glottiques sont estimées à la fois sur le test d'entraînement et de test. Une *matrice de confusion* $C(i,j)$ entre le locuteur i et le locuteur j est ensuite calculée. Dans ce travail, le carré de l'erreur temporelle relative (RTSE) a été choisi comme mesure entre deux formes d'onde différentes. Si $v_{k,l,training}$ et $v_{k,l,test}$ désignent la k^{ieme} signature glottique (dans notre cas, $k = 1, 2$ respectivement pour le résidu propre et l'enveloppe d'énergie) pour le locuteur l, estimée respectivement sur les sets d'entraînement et de test, la matrice de confusion $C_k(i,j)$ en utilisant uniquement la k^{ieme} signature glottique est définie comme :

$$C_k(i,j) = \sqrt{\frac{\sum_{n=0}^{N-1} \left(v_{k,i,test}(n) - v_{k,j,training}(n)\right)^2}{\sum_{n=0}^{N-1} v_{k,j,training}(n)^2}} \quad (1)$$

où N est le nombre d'échantillons pour la normalisation en pitch. La matrice de confusion $C(i,j)$ est finalement obtenue comme :

$$C(i,j) = C_1(i,j) \cdot C_2(i,j) \quad (2)$$

Notez que plusieurs opérations pour combiner les deux matrices sont possibles. D'après nos expériences, la multiplication a donné les meilleurs résultats, bien que les différences de performance observées étaient relativement faibles.

Finalement, l'identification d'un locuteur i est réalisée en cherchant la plus petite valeur dans la i^{ieme} ligne de la matrice de confusion $C(i,j)$. Le locuteur est alors identifié correctement si la position du minimum est i. En d'autres mots, quand des enregistrements sont présentés au système, le locuteur identifié est celui dont les signatures glottiques sont les plus proches (au sens Euclidien) des signatures glottiques extraites sur ces enregistrements.

3. EXPÉRIENCES

Les expériences décrites dans cette Section ont été menées sur les bases de données TIMIT et YOHO. La base de données TIMIT [4] comporte 10 enregistrements prononcés par 630 locuteurs (438 hommes et 192 femmes) échantillonés à 16 kHz. Quant à la base de données YOHO [1], elle contient de la parole de 138 locuteurs (108 hommes et 30 femmes) échantillonée à 8 kHz. Ces enregistrements ont été collectés dans un environnement réel de bureau lors de 4 sessions sur une période de 3 mois. Pour chaque session, 24 phrases ont été prononcées par locuteur. Dans nos expériences, les données ont été séparées pour chaque locuteur (et chaque session pour YOHO) en 2 parts égales pour l'entraînement et le test. Ceci est fait de manière à garantir que, pour chaque étape, suffisament de trames de résidu soient disponibles pour estimer de façon fiable les signatures glottiques.

3.1. Résultats sur la base de données TIMIT

Pour donner une première idée sur le potentiel d'utiliser les signatures glottiques en reconnaissance du locuteur, la Figure 4 montre les distributions de $C_1(i,j)$ respectivement quand $i = j$ et quand $i \neq j$. En d'autres mots, ce graphique montre les histogrammes de la RTSE (voir Equation 1), en échelle logarithmique, entre les résidus propres estimés respectivement pour le même locuteur et pour des locuteurs différents. Il peut être clairement observé que la mesure d'erreur est bien plus grande (environ 15x en moyenne) quand la signature glottique n'appartient pas au locuteur considéré. Cependant, un faible recouvrement des distributions est noté, ce qui peut mener à certaines erreurs d'identification du locuteur.

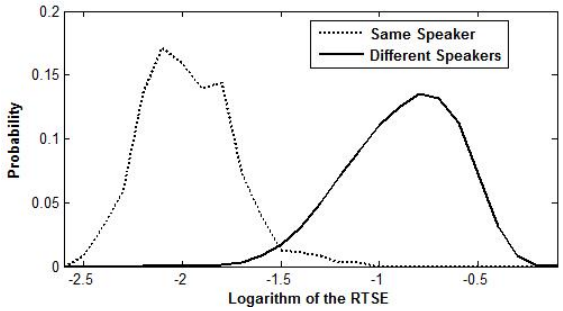


FIGURE 4: Distributions du carré de l'erreur temporelle relative (RTSE) entre les résidus propres estimés respectivement pour le même locuteur et pour des locuteurs différents.

La Figure 5 illustre l'évolution du taux d'identification avec le nombre de locuteurs considérés dans la base de données. Pour cela, l'identification a été réalisée en utilisant une seule des deux signatures glottiques, ou en utilisant leur combinaison comme suggéré par l'Equation 2. Comme attendu, la performance se dégrade quand le nombre de locuteurs augmente, puisque le risque de confusion devient plus important. Cependant cette dégradation est relativement lente dans tous les cas. Une autre observation importante est le clair avantage de combiner les informations des deux signatures glottiques. En effet, ceci mène à une amélioration de 7.78% comparé à l'utilisation unique du résidu propre.

Le Tableau 1 résume les résultats obtenus sur la base de

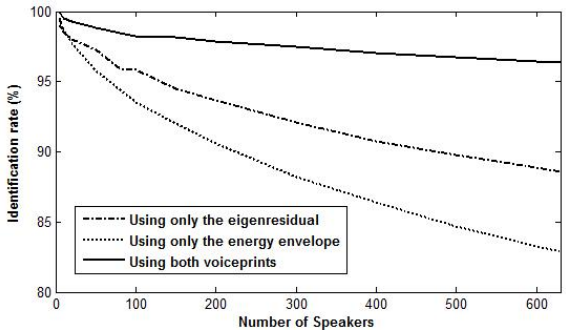


FIGURE 5: Evolution du taux d'identification avec le nombre de locuteurs pour la base de données TIMIT.

données TIMIT. Les taux d'identification pour 168 locuteurs sont aussi donnés pour des motifs de comparaison. En effet, dans [10] Plumpe et al. ont extrait un ensemble de 12 paramètres temporels caractérisant le flux glottique estimé par une analyse sur la phase fermée de la glotte. En utilisant ces attributs, ils ont rapporté un taux de mauvaise classification de 28.64% sur un sous-ensemble de 168 locuteurs. Sur le même sous-ensemble, Gudnason et al. ont rapporté dans [5] un taux de mauvaise classification de 5.06% en utilisant des coefficients du cepstre de la source vocale. Ces résultats peuvent être comparés aux 1.98% que nous avons obtenus en utilisant les deux signatures glottiques. Finalement, notons que Gudnason et al. [5], en utilisant leurs attributs glottiques, ont aussi obtenu un taux de mauvaise classification de 12.95% sur la totalité de la base de données TIMIT (630 locuteurs). Avec les signatures glottiques proposées, un taux de mauvaise classification de 3.65% est atteint.

	168 locuteurs	630 locuteurs
Résidu propre	5.88	11.43
Enveloppe d'énergie	8.76	17.14
Avec les 2 signatures	1.98	3.65

TABLE 1: Taux de mauvaise classification ($\%$) sur la base de données TIMIT obtenus en utilisant une seule des deux signatures glottiques, ou leur combinaison.

3.2. Résultats sur la base de données YOHO

Comparé à la base de données TIMIT, le corpus YOHO diffère en deux principaux aspects : 1) les enregistrements sont maintenant échantillonnés à 8 kHz, 2) les enregistrements ont été collectés en plusieurs sessions sur une période de 3 mois. Le premier point implique pour notre système que les GCIs sont plus difficiles à localiser, et de surcroît que les signatures glottiques vont perdre leurs détails hautes-fréquences (qui peut contenir de l'information pertinente pour distinguer des locuteurs). Concernant le second aspect, on peut s'attendre à une plus grande variabilité intra-locuteur lorsque les sessions d'entraînement et de test sont espacées sur une longue période de temps. Les résultats que nous avons obtenus sur le corpus YOHO en utilisant les 2 signatures vocales sont présentés en Figure 6. Ces résultats sont détaillés selon la période séparant les enregistrements d'entraînement et de test. De plus, les pourcentages des cas pour lesquels le locuteur correct est reconnu en seconde ou troisième position (au lieu d'être en première position) sont également donnés. De ce graphe il peut être remarqué que le système marche

parfaitement quand les enregistrements proviennent de la même session. Au contraire, quand le test est fait dans une session ultérieure, l'identification chute brutalement jusqu'à 70%. Cette chute est essentiellement imputable à la discordance entre les conditions d'entraînement et de test. Il peut être observé que le taux d'identification décroît ensuite d'environ 5% pour toute session ultérieure. Comme attendu, ceci résulte de la plus grande variabilité du locuteur quand l'intervalle de temps entre sessions augmente. Notons aussi que, quand les conditions d'entraînement et de test diffèrent, entre 12% et 16% des locuteurs sont identifiés en seconde ou troisième position. On peut s'attendre à ce que la combinaison des signatures glottiques proposées avec des attributs basés sur le spectre de magnitude de la parole enlève l'essentiel de cette ambiguïté. Finalement, dans un but de comparaison, Gudnason et al. ont rapporté dans [5] un taux de mauvaise identification de 36.3% en utilisant les coefficients cepstraux de la source vocale (avec des enregistrements de test répartis sur les 4 sessions). En moyennant nos résultats sur la totalité des sessions, nous avons trouvé un taux de mauvaise classification de 29.3% en utilisant les 2 signatures glottiques.

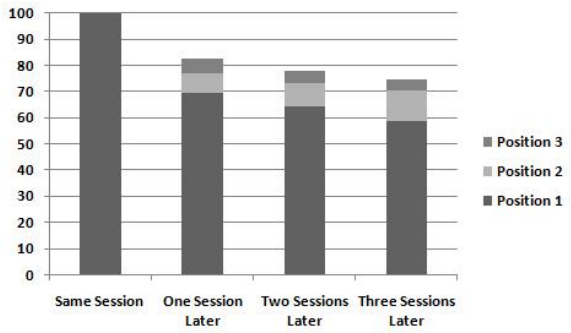


FIGURE 6: Taux d'identification (%) pour la base de données YOHO quand les sessions d'entraînement et de test peuvent être séparées sur une longue période. La proportion de locuteurs pour lesquels les signatures glottiques sont reconnues en seconde ou troisième position est également indiquée.

4. CONCLUSION

Cet article a étudié la potentialité d'utiliser des signatures glottiques en reconnaissance du locuteur. Ces signatures vocales ont été dérivées d'une analyse, pour un locuteur donné, d'un set de trames de résidu pitch-synchrones et normalisées en prosodie. Des résultats d'identification du locuteur ont été rapporté sur les bases de données TIMIT et YOHO. Dans ces expériences, les signatures glottiques proposées ont donné de meilleurs résultats que d'autres études similaires basées sur des attributs glottiques. Cependant, il a été montré que la performance est dégradée quand les sessions d'entraînement et de test sont espacées dans le temps.

Plusieurs améliorations pourraient être apportées à l'approche actuelle. En effet, les résultats ont été obtenus en utilisant *uniquement* les signatures glottiques proposées. D'après l'évidence d'une complémentarité entre les MFCCs et les caractéristiques basées sur l'excitation ([8], [10], [5]), il est raisonnable de penser qu'incorporer les signatures vocales proposées dans un système de reconnaissance du locuteur mènerait à une amélioration appréciable. Deuxièmement, l'application d'une compensation de canal pourrait réduire la discordance entre les sessions d'entraînement et de test. En effet, différentes conditions d'enregistrement imposent différentes caractéristiques au signal de parole. Parmi celles-ci, les différences en réponse de phase peuvent affecter sensiblement l'estimation des signatures glottiques (puisque l'information du résidu est essentiellement contenue dans sa phase). Ces deux possibles améliorations sont l'objet d'un travail en cours.

5. REMERCIEMENTS

Thomas Drugman est supporté par le Fonds National de la Recherche Scientifique (FNRS).

RÉFÉRENCES

[1] J. Campbell. Testing with the yoho cd-rom voice verification corpus. In *Proc. ICASSP*, pages 341–344, 1995.

[2] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[3] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. Interspeech*, 2009.

[4] W. Fisher, G. Doddington, and K. Goudie-Marshall. The darpa speech recognition research database : Specifications and status. In *Proc. DARPA Workshop on Speech Recognition*, pages 93–99, 1986.

[5] J. Gudnason and M. Brookes. Voice source cepstrum coefficients for speaker identification. In *Proc. ICASSP*, pages 4821–4824, 2008.

[6] I. Karlsson. Glottal waveform parameters for different speaker types. In *STL-QPSR*, volume 29, pages 61–67, 1988.

[7] L.G. Kersta. Voiceprint identification. In *Nature 196*, pages 1253–1257, 1962.

[8] S. Murty and B. Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. In *IEEE Signal Processing Letters*, volume 13, pages 52–55, 2006.

[9] [Online]. The snack sound toolkit. In *http ://www.speech.kth.se/snack/*.

[10] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. In *IEEE Trans. on Speech and Audio Processing*, volume 7, pages 569–586, 1999.

[11] D.A. Reynolds. An overview of automatic speaker recognition technology. In *Proc. ICASSP*, volume 4, pages 4072–4075, 2002.

[12] P. Thevenaz and H. Hugli. Usefulness of the lpc-residue in text-independent speaker verification. In *Speech Communication*, volume 17, pages 145–157, 1995.

Corrélation entre les différences entre les taux de reconnaissance de la parole sur deux ensembles de test et celles des distributions de probabilité des vecteurs acoustiques de ces mêmes ensembles

Cong-Thanh Do[1,2], *Dominique Pastor*[1,2] *et André Goalic*[1,2]

[1]Institut TELECOM, TELECOM Bretagne ; UMR CNRS 3192 Lab-STICC
Technopôle Brest-Iroise, CS 83818, 29238 Brest Cedex 3, France
[2]Université européenne de Bretagne, 35000 Rennes, France
Courriel : {thanh.do, dominique.pastor, andre.goalic}@telecom-bretagne.eu

ABSTRACT

A strong correlation was revealed between the Kullback-Leibler divergence (KLD), calculated between the probability distributions of the feature vectors extracted from two testing sets of speech signals, and the difference between the automatic speech recognition (ASR) word accuracies (WAs) computed on these two testing sets by an hidden Markov model-based ASR system. This strong correlation suggests that the variation of the difference between the ASR WAs computed on two testing sets could be predicted based on the variation of the KLD between the two probability distributions of the feature vectors extracted from these sets.

Keywords: Correlation, Kullback-Leibler divergence, hidden Markov model-based automatic speech recognition, speech feature vector.

1. Introduction

Les systèmes de reconnaissance automatique de la parole (RAP) pour les grands vocabulaires utilisent les principales composantes de la Fig. 1 [17]. Le décodeur recherche la séquence des mots $W = w_1, w_2, \ldots, w_K$ encodés dans les vecteurs acoustiques $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$ et extraits du signal de parole d'entrée. Le décodeur essaie de déterminer la séquence des mots les plus probables, $\hat{W}$, en maximisant la probabilité $P(W|\mathbf{O})$

$$\hat{W} = \arg\max_{W} P(W|\mathbf{O}) \tag{1}$$

La probabilité $P(W|\mathbf{O})$, étant difficile à modéliser directement [7], la règle de Bayes est utilisée pour transformer (1) en

$$\hat{W} = \arg\max_{W} P(W|\mathbf{O}) = \arg\max_{W} \frac{p(\mathbf{O}|W)P(W)}{p(\mathbf{O})} \tag{2}$$

où les vecteurs acoustiques $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$ sont considérés comme des variables aléatoires continues. Trouver la séquence des mots $\hat{W}$, maximisant $P(W|\mathbf{O})$, équivaut donc à trouver la séquence des mots $\hat{W}$ qui maximise le produit :

$$\hat{W} = \arg\max_{W} \left(p(\mathbf{O}|W)P(W)\right) \tag{3}$$

où la vraisemblance $p(\mathbf{O}|W)$ est déterminée par un modèle acoustique et $P(W)$ est déterminé par un modèle de langage. Dans le système RAP, basé sur les modèles de Markov cachés (MMCs), les modèles acoustiques sont les MMCs λ_i qui représentent les mots w_i. Les paramètres des MMCs sont obtenus à partir d'apprentissage d'un grand nombre de signaux de parole [14]. Dans le processus de décodage, la vraisemblance $p(\mathbf{O}_i|\lambda_i)$ peut être facilement estimée lorsque la séquence d'états correspondante $Q = q_1, q_2, \ldots, q_C$ est connue. $\mathbf{O}_i$ est une sous-séquence de longueur C de $\mathbf{O}$. La séquence optimale d'états $\hat{Q} = \hat{q}_1, \hat{q}_2, \ldots, \hat{q}_C$, qui maximise la probabilité $P(\hat{Q}|\mathbf{O}_i, \lambda_i)$, peut être efficacement trouvée par l'algorithme Viterbi [13]. On a

$$p(\mathbf{O}_i|\hat{Q}, \lambda_i) = b_{\hat{q}_1}(\mathbf{o}_{c_1}).b_{\hat{q}_2}(\mathbf{o}_{c_2}) \ldots b_{\hat{q}_C}(\mathbf{o}_{c_C}) \tag{4}$$

où $\{c_1, \ldots, c_C\}$ sont les indexes des vecteurs de caractéristiques de $\mathbf{O}_i$ et $b_{\hat{q}_t}(\mathbf{o}_{c_t}), t = 1, \ldots, C$ sont les vraisemblances d'observation. Une distribution de probabilité d'observation $b_q(.)$ est souvent modélisée par un modèle de mélange de gaussiens multivariés (MMGM) [13]. Les vecteurs acoustiques $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$ sont supposés être statistiquement indépendants pour que l'équation (4) soit correcte.

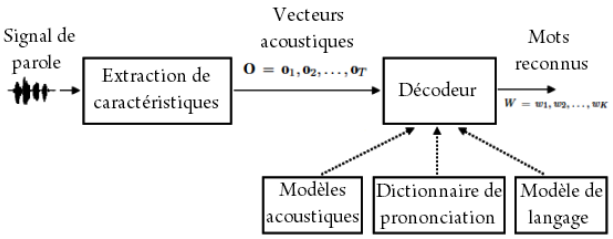


Fig. 1: Principales composantes d'un système de reconnaissance automatique de la parole (RAP) à grand vocabulaire [17].

Connaissant la séquence des états $\hat{Q}$, la vraisemblance $p(\mathbf{O}_i|\hat{Q}, \lambda_i)$ dans (4) dépend principalement des vraisemblances d'observations $b_{\hat{q}_t}(\mathbf{o}_{c_t}), t = 1, \ldots, C$. Ceci suggère que la distribution de probabilité des vecteurs acoustiques $p(\mathbf{o}_{c_t})$ devrait avoir une contribution significative aux vraisemblances d'observations $b_{\hat{q}_t}(\mathbf{o}_{c_t})$, et donc à la performance du sytème de reconnaissance [17, 12]. A cet égard, nous cherchons à établir une relation entre les différences de performance du système RAP et celles apparaissant entre les distributions de probabilité des vecteurs acoutiques extraits des ensembles de signaux de test. Nous avons expérimentalement trouvé que la différence entre les performances du système RAP, évaluées sur deux ensembles de signaux de test, est fortement corrélée avec celle apparaissant entre les distributions de probabilité des vecteurs acoustiques extraits de ces ensembles de signaux de test. Comme dans les MMCs, les distributions de probabilité des vecteurs acoustiques $p(\mathbf{o}_t)$ sont modélisées par un MMGM dont les paramètres sont estimés seulement sur les ensembles de signaux de tests. La différence entre les distributions de probabilités est caractérisée par la divergence Kullback-Leibler (KLD) [10].

2. Formulation du problème

2.1. Différence des distributions de probabilité

Soient $\mathbf{o}_x$ et $\mathbf{o}_y$ les vecteurs acoustiques de dimension L qui sont extraits d'une trame de deux ensembles de signaux de test, Ω_x and Ω_y, respectivement. Les distributions de probabilité $p_x(\mathbf{o})$ et $p_y(\mathbf{o})$ de $\mathbf{o}_x$ et $\mathbf{o}_y$, respectivement, peuvent-être modélisées par un MMGM de la façon suivante :

$$p_x(\mathbf{o}) = \sum_{m=1}^{M} \lambda_{x,m} \mathcal{N}(\mathbf{o}; \mu_{x,m}, \mathbf{\Sigma}_{x,m}) \tag{5}$$

$$p_y(\mathbf{o}) = \sum_{m=1}^{M} \lambda_{y,m} \mathcal{N}(\mathbf{o}; \mu_{y,m}, \mathbf{\Sigma}_{y,m}) \tag{6}$$

où

$$\mathcal{N}(\mathbf{o};\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{L/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{o}-\mu)^T\mathbf{\Sigma^{-1}}(\mathbf{o}-\mu)\right\}$$

est la densité gaussienne multivariée avec comme vecteur de moyenne μ de dimension L et la matrice de covariance $\mathbf{\Sigma}$ de dimension L^*L. Dans (5) et (6), M est le nombre de composantes gaussiennes, $\lambda_{x,m}$ et $\lambda_{y,m}$ sont les poids des m-ième composantes gaussiennes de $p_x(\mathbf{o})$ et $p_y(\mathbf{o})$, respectivement, et $\mathbf{o}$ est un vecteur généralisé des variables aléatoires. Les poids $\lambda_{x,m}$ et $\lambda_{y,m}$ qui satisfont les conditions

$$0 \leq \lambda_{x,m}, \lambda_{y,m} \leq 1; \sum_{m=1}^{M} \lambda_{x,m} = \sum_{m=1}^{M} \lambda_{y,m} = 1 \qquad (7)$$

peuvent être vus comme les probabilités à priori des composantes gaussiennes $\mathcal{N}(\mathbf{o};\mu_{x,m},\mathbf{\Sigma}_{x,m})$ et $\mathcal{N}(\mathbf{o};\mu_{y,m},\mathbf{\Sigma}_{y,m})$, respectivement.

La différence entre $p_x(\mathbf{o})$ et $p_y(\mathbf{o})$ est évaluée en utilisant la divergence Kullback-Leibler (KLD) [10]. Si $p_x(\mathbf{o})$ est utilisée comme distribution de probabilité de référence, la KLD entre $p_x(\mathbf{o})$ et $p_y(\mathbf{o})$ peut être écrite comme suit :

$$D\left(p_x(\mathbf{o})||p_y(\mathbf{o})\right) = \int p_x(\mathbf{o}) \log \frac{p_x(\mathbf{o})}{p_y(\mathbf{o})} d\mathbf{o} \qquad (8)$$

Le calcul de (8) n'est pas réalisable analytiquement car $p_x(\mathbf{o})$ et $p_y(\mathbf{o})$ sont, dans ce cas, les MMGMs. Parmi les méthodes disponibles [8] pour le calcul de (8), la simulation Monte Carlo est la méthode qui peut estimer $D\left(p_x(\mathbf{o})||p_y(\mathbf{o})\right)$ pour les grandes valeurs de L avec une précision arbitraire [8]. Nous utilisons donc la méthode de simulation Monte Carlo pour l'estimation de (8). L'idée principale de l'utilisation de la simulation Monte Carlo pour l'estimation de la KLD est d'appliquer la technique de l'échantillonnage d'importance [5] pour réécrire (8) comme

$$D\left(p_x(\mathbf{o})||p_y(\mathbf{o})\right) = E_{p_x(\mathbf{o})}\left[\log \frac{p_x(\mathbf{o})}{p_y(\mathbf{o})}\right] \qquad (9)$$

L'espérance dans (9) peut être approximativement estimée en générant N vecteurs de variables aléatoire indépendants identiquement distribués (i.i.d) $\{\mathbf{o}_i\}_{i=1\ldots N}$ suivant la distribution de probabilité $p(\mathbf{o}_x)$ et puis on calcule la moyenne empirique de $\log\left[p_x(\mathbf{o}_i)/p_y(\mathbf{o}_i)\right], i = 1,\ldots,N$ car

$$\frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{p_x(\mathbf{o}_i)}{p_y(\mathbf{o}_i)}\right] \longrightarrow E_{p_x(\mathbf{o})}\left[\log \frac{p_x(\mathbf{o})}{p_y(\mathbf{o})}\right] \qquad (10)$$

lorsque $N \to \infty$. Pour générer un vecteur de variables aléatoires $\mathbf{o}_i$ suivant la distribution de probabilité $p_x(\mathbf{o})$ qui est un MMGM, nous tirons d'abord un indicateur $\delta \in \{1,\ldots,M\}$ selon les probabilités à priori $\{\lambda_{x,m}\}_{m=1,\ldots,M}$. Puis, nous générons le vecteur des variables aléatoires $\mathbf{o}_i$ à partir de la composante gaussienne multivariée correspondante $\mathcal{N}(\mathbf{o};\mu_{x,\delta},\mathbf{\Sigma}_{x,\delta})$.

2.2. Différence de la performance de RAP

Etant donnée deux ensembles de signaux de parole, Ω_x et Ω_y, pour les tests avec un système de RAP basé sur les MMCs indépendant du locuteur. On cherche à caractériser la variation de la différence entre les taux de reconnaissance en termes de précisions sur les mots (word accuracies [18]), a_x et a_y, évaluées sur Ω_x et Ω_y, respectivement. Quand une de deux précisions sur les mots est fixe, la variation de l'autre précision sur les mots peut être prédite en se basant sur la variation de la différence entre deux précisions sur les mots. Dans ce travail, nous cherchons à établir expérimentalement une relation entre la différence de la précisions sur les mots, $\bar{a} = a_x - a_y$, et la KLD $D(p_x(\mathbf{o})||p_y(\mathbf{o}))$.

Dans ce papier, le système de RAP basé sur les MMCs indépendant du locuteur est établi d'apprentissage sur la base de données d'apprentissage de TI-digits [11] utilisant le logiciel de reconnaissance de la parole HTK [18]. Le système utilise également un modèle de langage de type bigram [18] et ses modèles acoustiques sont les triphone MMCs dépendantes du contexte, comportant trois états. Les distributions de probabilité d'observation $b_q(.)$ des MMCs sont modélisées par les MMGMs consistant en $M = 16$ composantes gaussiennes multivariées. Chaque vecteur acoustique, $\mathbf{o}_x$ et $\mathbf{o}_y$, consiste en 13 coefficients cepstraux en échelle de fréquence Mel (MFCCs) [2] qui sont extraits de chaque trame de parole fenêtré par une fenêtre Hamming et de longueur 25 ms en utilisant le logiciel de reconnaissance de la parole HTK. Le chevauchement entre deux trames successives est 15 ms. Les coefficients delta et accélération sont annexés aux MFCCs statiques pour produire les vecteurs acoustiques de taille $L = 39$. En utilisant cette configuration à la fois dans les conditions d'apprentissage et de test, les caractéristiques du signal de parole peuvent être considérées comme des variables aléatoires stationnaires. En outre, les matrices de covariance, $\Sigma_{x,m}$ and $\Sigma_{y,m}, m = 1,\ldots,M$, sont toutes diagonales car les caractéristiques du signal de parole sont supposées être des variables aléatoires statistiquement indépendantes.

3. Résultats expérimentaux

Un ensemble Ω_x se compose de 250 phrases de parole originale non-bruitée est utilisé comme l'ensemble référentiel. Les signaux de parole dans l'ensemble sont choisis dans la base de données de test de la base de données TI-digits [11]. Ces 250 phrases, qui sont des séquences de chiffres parlées, ont été choisies de sorte qu'elles ont été prononcés par des locuteurs adultes et enfants des deux sexes. La longueur des phrases dans l'ensemble varient de la longueur minimale (séquence de chiffres isolés) à la longueur maximale (séquence de sept chiffres) des séquences de la base TI-digits. Les données ont été recueillies dans un environnement non-bruité et numérisé à 20 kHz. Dans cette étude, les données ont été sous-échantillonnées à 8 kHz. Un test de reconnaissance est pris sur cet ensemble de test utilisant le système de RAP basé sur les MMCs et la précision sur les mots référentielle obtenue est $a_x = 99.76\%$. Les autres ensembles de signaux de test Ω_y sont dérivés de cet ensemble référentiel de différentes façons. Les tests de reconnaissance sont ensuite effectués sur ces ensembles de test pour estimer les précisions sur les mots a_y.

3.1. Expérimentations avec parole bruitée

Dans cette section, nous utilisons des signaux de parole contaminés par du bruit blanc pour les ensembles de signaux de test. Ces signaux de parole bruitée ont été produits en ajoutant artificiellement le bruit blanc de la base de données NOISEX-92 [16], à différents rapports signaux-sur-bruit (SNR), aux signaux de parole originale non-bruitée de l'ensemble référentiel Ω_x. Neuf ensembles de signaux de parole bruitée $\{\Omega_{y,i}\}_{i=1,\ldots,9}$ ont été produits correspondants aux neuf niveaux de SNR $\{-5, 0, 5, 10, 15, 20, 25, 40, 50\}$ dB qui ont été utilisés dans la phase de contamination des signaux de parole non-bruitée. En suite, utilisant le système de RAP basé sur les MMCs et indépendant du locuteur mentionnés ci-dessus, les tests de reconnaissance sont effectués sur ces ensembles de test de signaux de parole bruitée et les précisions sur les mots $\{a_{y,i}\}_{i=1,\ldots,9}$ sont obtenues. Les différences de précisions sur les mots, $\{\bar{a}_i = a_x - a_{y,i}\}_{i=1,\ldots,9}$, sont indiquées dans le Tab. 1.

Comme dans l'apprentissage du système de RAP basé sur les MMCs, les vecteurs acoustiques se composent de 13 MFCCs concaténés avec des coefficients delta et accélération. Ces vecteurs acoustiques de tailles 39, $\mathbf{o}_x$ et $\mathbf{o}_{y,i}, i = 1,\ldots,9$, sont extraits de l'ensemble référentiel de test et des ensembles de test des signaux de parole bruitée, respectivement. Un total de 56000 vecteurs acoustiques sont extraits de chaque ensemble de test. Comme mentionné ci-dessus, les distributions de probabilité $p_x(\mathbf{o})$ et $p_{y,i}(\mathbf{o}), i = 1,\ldots,9$ sont modélisées par les MMGMs qui se composent de $M = 16$ composantes gaussiennes multivariées. Leurs paramètres $\{\lambda_{x,m},\lambda_{y,i,m},\mu_{x,m},\mu_{y,i,m},\Sigma_{x,m},\Sigma_{y,i,m}\}_{m=1,\ldots,M}^{i=1,\ldots,9}$ sont estimés des vecteurs acoustiques observés en utilisant l'algorithme expectation-maximisation (EM) [3]. Les matrices de covariance

Tab. 1: Les différences de taux de reconnaissance en termes de précision sur les mots (en %) $\{\bar{a}_i\}_{i=1,\ldots,9}$ et les KLDs $\{D\left(p_x(\mathbf{o})||p_{y,i}(\mathbf{o})\right)\}_{i=1,\ldots,9}$, estimées de l'ensemble référentiel et des ensembles de test des signaux de parole bruitée. Le système de RAP basé sur les MMCs est obtenu par l'apprentissage sur la base de données d'apprentissage de signaux de parole non-bruitée de TI-digits [11]. La précision sur les mots évaluée sur l'ensemble référentiel est $a_x = 99.76\%$.

Différences	SNR (en dB)								
	-5	0	5	10	15	20	25	40	50
$\bar{a}_i = a_x - a_{y,i}$	64.44	61.09	44.58	24.91	14.56	7.43	4.51	0.98	0
$D\left(p_x(\mathbf{o})||p_{y,i}(\mathbf{o})\right)$	172.23	112.65	71.43	47.54	30.26	21.50	13.98	3.99	2.31

$\{\Sigma_{x,m}, \Sigma_{y,i,m}\}_{m=1,\ldots,M}^{i=1,\ldots,9}$ sont restreintes aux diagonales.

Les KLDs $D\left(p_x(\mathbf{o})||p_{y,i}(\mathbf{o})\right), i = 1,\ldots,9$ entre $p_x(\mathbf{o})$ et $p_{y,i}(\mathbf{o}), i = 1,\ldots,9$, respectivement, sont estimées en utilisant la méthode de simulation Monte Carlo [5] avec $N = 100000$ échantillons aléatoirement générés. Ces KLDs sont également indiquées dans le Tab. 1. Comme d'habitude, la dépendance linéaire entre les différences de précisions sur les mots $\{\bar{a}_i\}_{i=1,\ldots,9}$ et les KLDs $D\left(p_x(\mathbf{o})||p_{y,i}(\mathbf{o})\right), i = 1,\ldots,9$ est mesurée par le coefficient de corrélation de Pearson (Pearson product-moment correlation coefficient PMCC) [15] $r = 0.85$. Soient $\widetilde{a}_j = \bar{a}_j / max\{\bar{a}_i\}_{i=1,\ldots,9}, j = 1,\ldots,9$ les différences normalisées de précision sur les mots. De même, soient $\widetilde{D}_j = D\left(p_x(\mathbf{o})||p_{y,j}(\mathbf{o})\right) / max\{D\left(p_x(\mathbf{o})||p_{y,i}(\mathbf{o})\right)\}_{i=1,\ldots,9}, j = 1,\ldots,9$ les KLDs normalisées. Les précisions sur les mots normalisées $\{\widetilde{a}_j\}_{j=1,\ldots,9}$ et les KLDs normalisées sont illustrées ensemble dans la Fig. 2.

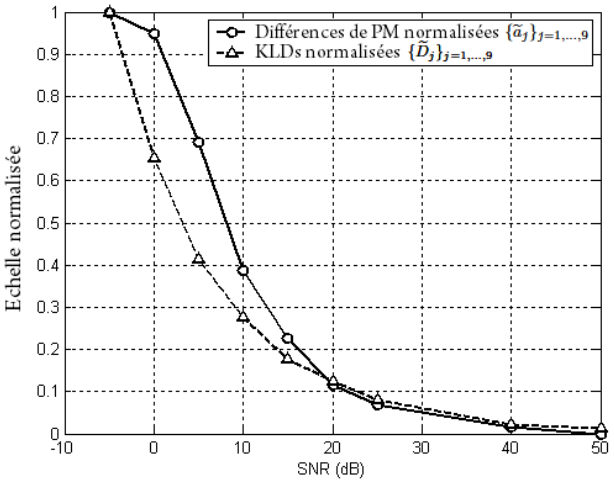


Fig. 2: Les différences de taux de reconnaissance en termes de précision sur les mots (PM) normalisées $\{\widetilde{a}_j\}_{j=1,\ldots,9}$ et les KLDs normalisées $\{\widetilde{D}_j\}_{j=1,\ldots,9}$ sont estimées sur l'ensemble référentiel et sur des ensembles de test des signaux de parole bruitée. Le PMCC entre $\widetilde{a}_j$ et $\widetilde{D}_j$ est $r = 0.85$.

Le PMCC $r = 0.85$ et les courbes dans Fig. 2 indiquent qu'il y a une corrélation forte entre les différences de précision sur les mots et les KLDs évaluées entre les distributions de probabilité des vecteurs acoustiques.

3.2. Expérimentations de la parole avec réduction spectrale

Nous effectuons des expérimentations avec un autre type de signal de parole avec réduction spectrale (spectrally reduced speech SRS) [4]. Cette fois, les ensembles de test se composent des signaux SRS synthétisés à partir de signaux originaux de parole non-bruitée de l'ensemble référentiel Ω_x. L'algorithme pour la synthèse de la SRS est décrit dans [4]. Les deux paramètres les plus importants pour la synthèse du signal SRS sont le nombre de sous-bandes de fréquence et la bande passante des enveloppes temporelles des sous-bandes [4]. Dans cette expérimentation, le nombre de sous-bandes de fréquence sont $\{4, 8, 16, 24, 32\}$ et les bandes passantes des enveloppes temporelles des sous-bandes sont $\{50, 500\}$ Hz. Il y a donc un total de dix ensembles de test de signaux SRS. Nous divisons ces ensembles en deux groupes. Le premier groupe comprend cinq ensembles de test $\{\mathbf{\Omega}_{y,i}\}_{i=1,\ldots,5}$ de signaux SRS ayant 50 Hz de bande passante des enveloppes temporelles des sous-bandes et le second comprend cinq ensembles de test $\{\mathbf{\Omega}'_{y,i}\}_{i=1,\ldots,5}$ de signaux SRS ayant 500 Hz de bande passante des enveloppes temporelles des sous-bandes. Les tests de reconnaissance sont ensuite effectués sur les ensembles de test $\{\mathbf{\Omega}_{y,i}\}_{i=1,\ldots,5}$ et $\{\mathbf{\Omega}'_{y,i}\}_{i=1,\ldots,5}$ pour obtenir les précisions sur les mots $\{a_{y,i}\}_{i=1,\ldots,5}$ et $\{a'_{y,i}\}_{i=1,\ldots,5}$, respectivement. Les différences de précision sur les mots $\{\bar{a}_i = a_x - a_{y,i}\}_{i=1,\ldots,5}$ et $\{\bar{a}'_i = a_x - a'_{y,i}\}_{i=1,\ldots,5}$ sont indiquées dans le Tab. 2.

Soient $\{p_{y,i}(\mathbf{o})\}_{i=1,\ldots,5}$ et $\{p'_{y,i}(\mathbf{o})\}_{i=1,\ldots,5}$ les distributions de probabilité des vecteurs acoustiques extraits des ensembles de test $\{\mathbf{\Omega}_{y,i}\}_{i=1,\ldots,5}$ et $\{\mathbf{\Omega}'_{y,i}\}_{i=1,\ldots,5}$, respectivement. Ces distributions de probabilité, modélisées par des MMGMs de 16 composantes gaussiennes multivariées, sont également estimées en utilisant l'algorithme EM [3]. Les KLDs $\{D(p_x(\mathbf{o})||p_{y,i}(\mathbf{o}))\}_{i=1,\ldots,5}$ et $\{D'(p_x(\mathbf{o})||p'_{y,i}(\mathbf{o}))\}_{i=1,\ldots,5}$ évaluées entre $p_x(\mathbf{o})$ et $\{p_{y,i}(\mathbf{o})\}_{i=1,\ldots,5}$, $\{p'_{y,i}(\mathbf{o})\}_{i=1,\ldots,5}$, respectivement, sont également indiquées dans le Tab. 2. Le PMCC entre $\bar{a}_i$ et $D(p_x(\mathbf{o})||p_{y,i}(\mathbf{o}))$ est $r = 0.8$. De même, le PMCC entre $\bar{a}'_i$ et $D'(p_x(\mathbf{o})||p'_{y,i}(\mathbf{o}))$ est $r' = 0.77$. Comme dans la section précédente, les différences de précision sur les mots normalisées, $\{\widetilde{a}_j\}_{j=1,\ldots,5}$ et $\{\widetilde{a}'_j\}_{j=1,\ldots,5}$, et les KLDs normalisées, $\{\widetilde{D}_j\}_{j=1,\ldots,5}$ et $\{\widetilde{D}'_j\}_{j=1,\ldots,5}$, sont illustrées dans la Fig. 3. Encore, les PMCCs, $r = 0.8$ et $r' = 0.77$, et les courbes dans la Fig. 3 indiquent qu'il y a des corrélations fortes entre les différences de précision sur les mots et les KLDs.

Tab. 2: Les différences de taux de reconnaissance en termes de précision sur les mots (en %) $\{\bar{a}_i\}_{i=1,\ldots,5}$, $\{a'_i\}_{i=1,\ldots,5}$ et les KLDs $\{D(p_x(\mathbf{o})||p_{y,i}(\mathbf{o}))\}_{i=1,\ldots,5}$, $\{D'(p_x(\mathbf{o})||p'_{y,i}(\mathbf{o}))\}_{i=1,\ldots,5}$, estimées de l'ensemble référentiel et des ensembles de test de signaux SRS. Le système de RAP basé sur les MMCs est obtenu par l'apprentissage sur la base de données d'apprentissage de signaux de parole non-bruitée de TI-digits [11]. La précision sur les mots évaluée sur l'ensemble référentiel est $a_x = 99.76\%$.

Différences	Nombre de sous-bandes de fréquence				
	4	8	16	24	32
$\bar{a}_i = a_x - a_{y,i}$	64.19	15.53	-0.06	0.06	0.19
$D(p_x(\mathbf{o})||p_{y,i}(\mathbf{o}))$	71.69	22.27	6.37	3.73	3.07
$\bar{a}'_i = a_x - a'_{y,i}$	60.42	2.20	0.19	0.06	0.13
$D'(p_x(\mathbf{o})||p'_{y,i}(\mathbf{o}))$	38.82	14.56	7.01	3.45	2.52

4. Conclusion

Dans ce papier, nous avons réalisé les expérimentations dont les résultats indiquent qu'il y a des corrélations fortes entre des différences de précision sur les mots et les KLDs entre des distributions de probabilité des vecteurs acoustiques. Les distributions de probabilité sont estimées seulement à partir des observations extraites des signaux dans des ensembles de test. L'idée de ces expérimentations vient du fait que la distribution de probabilité des vecteurs acoustiques devrait avoir une contribution significative sur les probabilités d'observation dans la RAP basée sur les MMCs et donc, sur la performance de la RAP [17, 12]. Cette corrélation forte suggère que la variation

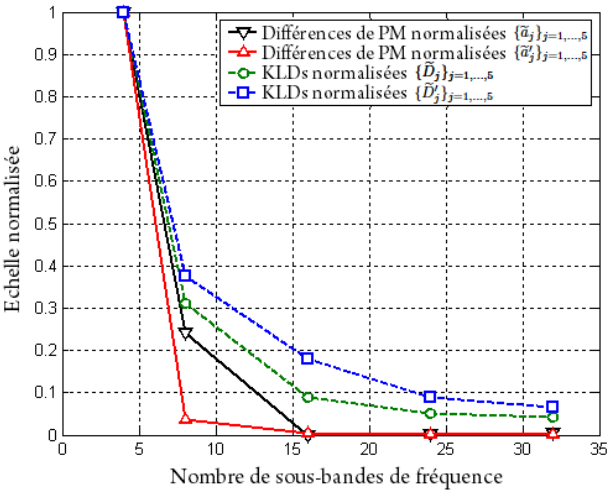


Fig. 3: Les différences de taux de reconnaissance en termes de précision sur les mots (PM) normalisées $\{\widetilde{a}_j\}_{j=1,...,5}$, $\{\widetilde{a}'_j\}_{j=1,...,5}$ et les KLDs normalisées $\{\widetilde{D}_j\}_{j=1,...,5}$, $\{\widetilde{D}'_j\}_{j=1,...,5}$, sont estimées de l'ensemble référentiel et des ensembles de signaux SRS. Les PMCCs entre $\widetilde{a}_j$, $\widetilde{a}'_j$ et $\widetilde{D}_j$, $\widetilde{D}'_j$ sont $r = 0.8$ et $r' = 0.77$, respectivement.

de la différence entre les précisions sur les mots évaluées sur deux ensembles de test peut être prédite en se basant sur la variation de la KLD entre les distributions de probabilité des vecteurs acoustiques extraits de ces ensembles de test. Cette corrélation serait également susceptible une base pour l'optimisation de précision sur les mots de la RAP évaluée sur un ensemble de signaux de test en se basant sur le principe de minimisation d'entropie croisée de Kullback-Leibler [6]. En minimisant la KLD évaluée entre deux distributions de probabilité des vecteurs acoustiques, on peut espérer que le taux de reconnaissance évalué sur un ensemble de test convergera autant que possible vers le taux de reconnaissance évalué sur l'ensemble référentiel. Le principe de minimisation d'entropie croisée de Kullback-Leibler a également été utilisé dans les autres applications tels que la synthèse de la parole à partir du texte [9] et la conception du texte pour la RAP [1].

Références

[1] X. Cui and A. Alwan. Efficient adaptation text design based on the Kullback-Leibler measure. In *Proc. IEEE ICASSP 2002, May 13 - 17, Orlando, USA*, volume 1, pages 613–616, May. 2002.

[2] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences. *IEEE Trans. Acoustics, Speech, Signal Processing*, 28(4) :357–366, Aug. 1980.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1) :1–38, 1977.

[4] C.-T. Do, D. Pastor, and A. Goalic. On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR. *IEEE Trans. Audio, Speech and Lang. Process.*, 2010 (in press http ://dx.doi.org/10.1109/TASL.2009.2032945).

[5] A. Doucet and X. Wang. Monte Carlo methods for signal processing. *IEEE Sig. Process. Mag.*, 22(6) :152–170, Nov. 2005.

[6] S.-C. Fang, J. R. Rajasekera, and H.-S. J. Tsao. *Entropy optimization and mathematical programming*. Kluwer Academic Publishers, Norwell, Massachusetts 02061 USA, 1997.

[7] M. Gales and S. Young. The application of hidden Markov models in speech recognition. *Foundations and trends in signal processing*, 1(3) :195–304, 2007.

[8] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *Proc. IEEE ICASSP 2007, April 15 - 20, Hawaii, USA*, volume 4, pages 317–320, Apr. 2007.

[9] A. Krul, G. Damnati, F. Yvon, and T. Moudenc. Corpus design based on the Kullback-Leibler divergence for text-to-speech synthesis application. In *Proc. ISCA Interspeech 2006, September 17 - 21, Pittsburgh, USA*, volume 1, pages 2030–2033, Sept. 2006.

[10] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1) :79–86, 1951.

[11] R. Leonard. A database for speaker-independent digit recognition. In *Proc. IEEE ICASSP 1984, March 19 - 21, San Diego, USA*, volume 9, pages 328–331, Mar. 1984.

[12] A. Nadas. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(4) :814–817, Aug. 1983.

[13] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2) :257–286, Feb. 1989.

[14] L. R. Rabiner. The power of speech. *Science*, 301(5639) :1494–1495, Sep. 2003.

[15] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1) :59–66, Feb. 1988.

[16] A. Varga and H.J.M Steeneken. Assessment for automatic speech recognition : II. NOISEX-92 : A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3) :247–251, Jul. 1993.

[17] S. Young. HMMs and related speech recognition technologies. *Springer handbook of speech processing (J. Benesty, M. M. Sondhi, and Y. Huang, Eds.)* Springer, pages 539–557, 2007.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book (for HTK version 3.4)*. Cambridge University Engineering Department, Cambridge, UK, 2006.

Approche multi-variable pour une reconnaissance de la parole distribuée robuste

Djamel Addou[1], Sid-Ahmed Selouani[2], Malika Boudraa[1], Bachir Boudraa[1]

[1]Lab. Communication parlée & Traitement de signal, USTHB Université des Sciences & Technologies, Alger, Algérie
[2]Lab. Larhis, Université de Moncton, Campus Shippang, New Brunswick, Canada
daddou@usthb.dz, selouani@umcs.ca, mboudraa@usthb.dz, bboudraa@usthb.dz

ABSTRACT

This paper describes a noise-robust Distributed Speech Recognition (DSR) front-end using a combination of conventional Mel-cepstral Coefficients (MFCCs) and Line Spectral Frequencies (LSFs). These features are adequately transformed and reduced in a multi-stream scheme using Karhunen-Loève Transform (KLT). We investigate the performance of this new front-end in terms of recognition accuracy in adverse conditions as well as in terms of dimensionality reduction. Our results showed that for highly noisy speech, using the proposed transformation scheme leads to a significant improvement in recognition accuracy on Aurora 2 task.

Keywords: Distributed speech recognition, MFCC, LSF, KLT, Multi-stream paradigm.

1. INTRODUCTION

Les systèmes de reconnaissance de parole, actuels, restent limités en matière de robustesse et doivent donc améliorer leur capacité à s'adapter à différents environnements bruités et changeants, comme par exemple les surfaces commerciales, les halls de gares ou d'aéroports. L'extraction des paramètres appropriés reste une question clé pour la reconnaissance de la parole. Au cours des dernières années, les MFCCs sont devenus les paramètres standards et sont actuellement employés dans les systèmes de reconnaissance de la parole distribuée (DSR). Pour de tels systèmes, il est crucial d'utiliser des paramètres robustes au bruit afin de maintenir une bonne performance lorsque le rapport signal sur bruit (RSB) diminue.

Dans des travaux précédents [1], nous avons présenté un paradigme multi-variable pour un système DSR où nous avons fusionné différentes sources d'informations qui se complétaient aux MFCCs. Nos expériences ont prouvé qu'une telle approche multi-variable, intégrant quelques paramètres basés sur un modèle simulant la cochlée ainsi que des indices acoustiques reflétant les résonnances spectrales (formants), conduit à une amélioration du taux de reconnaissance. Ceci a montré que les MFCCs, en dépit de leur popularité, perdent de l'information appropriée au processus de reconnaissance en milieu fortement bruité.

Dans ce papier nous étudions l'impact de l'utilisation des paramètres LSFs (fréquences de raies spectrales) sur la robustesse d'un système DSR en milieu bruité. Ces LSFs sont intégrés dans une approche multi-variable et employés comme nouveau *front-end* dans ce système. Il est important de noter que les LSFs ont l'avantage d'être utilisés dans les systèmes de codage de la parole. De nombreux travaux ont été publiés dans le but de proposer des systèmes de reconnaissance robustes dans les télécommunications mobiles [2].

D'autre part, nous visons à optimiser l'utilisation des flux de paramètres en réduisant la dimension des vecteurs acoustiques tout en améliorant la robustesse du système. Une façon efficace d'effectuer cette réduction est d'utiliser la transformée de Karhunen-Loève (KLT) [3]. C'est une technique de décomposition en sous-espaces également utilisée en rehaussement des signaux bruités [4]. Ainsi en intégrant KLT dans notre approche, nous réalisons deux objectifs : réduction optimale de paramètres et amélioration de la robustesse, en éliminant les composantes principales bruités (généralement celles d'ordres supérieurs). Pour l'évaluation de notre système LSF-KLT, nous avons effectué nos expériences sur la base de données AURORA.

Ce travail présente une solution complémentaire pour l'extraction des paramètres acoustiques adaptés à un environnement bruité. Le système de reconnaissance qui a servi de base à ce travail est présenté dans la section 2. L'extraction et le traitement des paramètres acoustiques sont décrits dans la section 3. La section 4 est dédiée à la validation expérimentale et à l'analyse de ses résultats. Une conclusion sur le travail présenté termine cet article.

2. LE STANDARD ETSI AURORA

Le système de reconnaissance ayant servi de base à ce travail utilise une modélisation statistique des paramètres du signal de parole à base de mélanges de gaussiennes. Le standard ETSI Aurora [5] a été crée à l'origine pour la reconnaissance de la parole sur des architectures distribuées. Le terminal a pour charge d'extraire les paramètres cepstraux et de les transmettre après compression (Figure 1). Le flux compressé est ensuite reçu par un serveur distant pour effectuer la reconnaissance. Les dégradations dues au codage bas débit de la voix ou au codage canal sont ainsi évitées.

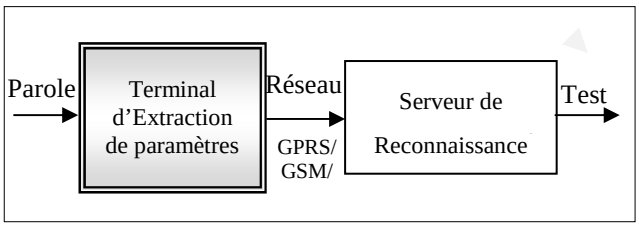


Figure 1: Synoptique d'un système DSR

Le standard Aurora utilise 12 paramètres cepstraux statiques et un paramètre d'énergie calculés toutes les 10 ms. Un étage de quantification est appliqué aux paramètres cepstraux. Le vecteur de paramètres est étendu en ajoutant les dérivées premières et secondes des cepstres.

3. TRAITEMENT AMONT DES PARAMÈTRES

Comme dans la plupart des systèmes de reconnaissance, nous utilisons une analyse cepstrale pour extraire les paramètres du signal de parole. Les paramètres cepstraux sont ensuite combinés avec d'autres paramètres: les fréquences de raies spectrales en utilisant une approche statistique multidimensionnelle, pour former le nouveau vecteur acoustique. Enfin, on réduit la dimension des vecteurs ainsi obtenus de façon optimale en effectuant la transformée de Karhunen-Loève qui a l'avantage supplémentaire de réduire l'effet du bruit qui pourrait entacher le signal de parole.

3.1 ANALYSE ACOUSTIQUE

Le signal échantillonné à 8 KHz est pré-accentué par un filtre FIR de premier ordre de la forme H (z)= $1+0.97z^{-1}$. Les coefficients LSFs et les coefficients MFCCs sont calculés toutes les 10 ms sur une largeur d'une fenêtre de Hamming de 25 ms. L'énergie de la trame est ajoutée aux coefficients statiques. Ensuite, les premières et secondes dérivées et les variations de leurs énergies correspondantes, sont calculées et ajoutées au vecteur statique.

Deux raisons principales ont motivé notre choix pour considérer les LSFs dans les communications mobiles bruitées. La première est liée au fait que les régions des LSFs du spectre peuvent rester au-dessus du niveau du bruit même dans le cas où le RSB est très faible, alors que les régions d'énergie inférieure tendront à être masquées par l'énergie du bruit. La deuxième raison est liée au fait que les LSFs sont employés couramment dans des codeurs conventionnels de la parole. Ceci évite l'incorporation de nouveaux paramètres qui peuvent exiger des modifications importantes et coûteuses aux dispositifs et aux codecs actuels

3.2 ANALYSE STATISTIQUE

Dans l'approche multi-variable acoustique, les différents paramètres obtenus à partir de différentes sources sont concaténés pour former le vecteur multi-variable utilisé pour l'apprentissage des HMM (Hiedden Model Markov). Considérons S sources d'informations qui fournissent des vecteurs synchrones O_{st} d'observation, où $s = 1,\ldots, S$ indique la source d'informations et t l'indice de temps. La dimension des vecteurs d'observation peut varier d'une source à une autre. Chaque séquence de vecteurs d'observation fournit des informations au sujet d'une séquence d'états cachés. Dans un système multi-flux (streams), au lieu de produire S séquences d'états pour S séquences d'observation, seule une séquence d'état est générée. En réalité, ceci est fait en introduisant une nouvelle fonction de distribution des états. La fonction de distribution de l'état *j* est définie comme suit :

$$b_j(O_t) = \prod_{s=1}^{S} \left[b_{js}(O_{st}) \right]^{\gamma_{js}}, \qquad (1)$$

où les distributions des vecteurs multiples d'observation sont fusionnées pour former une distribution à sortie unique pour l'état *j*. L'exposant γ spécifie la contribution de chaque flux à la distribution globale en mesurant sa distribution correspondante.

Dans les HMMs, les modèles de mixtures à mélanges de gaussiennes sont utilisés pour représenter la distribution d'émission des états. La probabilité du vecteur O_t à chaque instant *t* dans l'état *j* peut être déterminée par l'expression suivante :

$$b_j(O_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M} C_{jsm} N(O_{st}; \mu_{jsm}; \phi_{jsm}) \right]^{\gamma_{js}} \qquad (2)$$

M est le nombre de composants du mélange et c_{jsm} est le $m^{ème}$ poids de la gaussienne de l'état *j* pour la source *S*. *N* indique le multi-variable gaussien avec μ_{jsm} le vecteur moyen et Φ_{jsm} la matrice de covariance.

Dans ce travail, les HMMs correspondant au nouveau *front-end*, seront composés de trois flux (S=3) ; à savoir, Les MFCCs, leurs dérivées premières et enfin les LSFs.

3.3 ANALYSE DIMENSIONNELLE

KLT est une technique d'analyse statistique multi-variable. Son principe est basé sur le calcul des vecteurs propres de la matrice de covariance, puis leur agencement selon l'ordre décroissant des valeurs propres correspondantes. On établit enfin la matrice *W* de projection selon les *N* plus grands vecteurs propres (c.-à-d., les *N* directions de plus grandes variances) [3]. L'espace généré par les vecteurs correspondant aux valeurs propres d'ordres inférieurs est supposé être très faiblement influencé par le bruit. La projection des vecteurs bruités dans ce sous-espace est le principe des méthodes de décomposition en sous-espaces qui permettent de réaliser le rehaussement [4]. Chaque vecteur *x* de paramètres est alors prétraité pour produire le vecteur optimal y selon l'expression :

$$y = W(x - \mu), \qquad (3)$$

où μ représente le vecteur Moyenne des paramètres. Nous employons la matrice *W* pour extraire les paramètres. Dans nos expériences, nous avons opté pour une KLT à classe indépendante (CI-KLT) dans laquelle la matrice de transformation est globale et est déterminée pour toutes les classes, contrairement au cas d'une KLT à classe dépendante (CD-KLT) [6,7] où nous utilisons une matrice de transformation pour chaque modèle acoustique.

4. EVALUATION EXPERIMENTALE

Cette section présente la base de données et les protocoles utilisés pour la validation expérimentale de l'approche de traitement acoustique proposée.

4.1 Base de données Aurora

Nos expériences ont été effectuées sur la base de données Aurora-2. C'est une base de haute qualité originale de TI Digits, établie par un échantillonnage de 8kHz et un filtrage G712 (disponible sur le terminal GSM)[8] avec un bruit additif contrôlé pour couvrir une gamme de rapports signa/bruit (SNR) (clean, 20, 15, 10, 5, 0,-5dB). Dans cette base, il y a 8 différents bruits dans 3 ensembles de test : A, B et C. Les expériences présentées dans cet article sont réalisées sur les tests A et B, avec deux sortes de bruits différents : réception, voiture pour le test A et restaurant, aéroport pour B. Au total, Chaque test contient 2x7x1001=14014 élocutions.

4.2 Description du système de base

Le système de base est défini par un vecteur de dimension 39, qui se compose de 12 coefficients cepstraux (sans le coefficient C_0) avec le log-énergie en plus des composants deltas et accélérations correspondants. Ce vecteur noté MFCC-E-D-A est considéré comme *front-end* conventionnel par le standard DSR de l'ETSI [5]. Les phases d'apprentissage et de reconnaissance des HMMs sont réalisées par HTK [9]. On divise le vecteur d'observation en flux multiples avec des pondérations égales. Pour le cas du vecteur de base, on utilise 3 flux : le premier concerne les coefficients statiques et l'énergie, le second et le troisième sont réservés, respectivement, aux coefficients delta et accélération avec leur composante d'énergie.

4.3 Méthodologie expérimentale

Dans la méthode proposée, le filtre LPC et l'algorithme UIT [10] ont été utilisés pour extraire les coefficients LSFs. Aussi, pour le cas du *front-end* proposé dans le cadre de l'approche multi variable, les 12 coefficients MFCCs et leurs dérivées premières, sans la composante d'énergie, constituent le premier et second flux. Les 10 coefficients LSFs sont pris comme troisième flux. Ces flux multiples ont des pondérations égales. Ce nouveau *front-end* sera noté par MFCC_D_LSF (34), où 34 indique sa dimension. Les LSFs ajoutés produisent un ensemble multidimensionnel de paramètres, et remplaceront les composantes accélérations et énergies du front-end conventionnel.

Afin d'évaluer l'impact des coefficients de pondération γ_{js} de l'équation (1), nous avons effectué une autre expérience sur le même vecteur mais avec l'utilisation de pondérations différents satisfaisant à la contrainte suivante [11] :

$$0 \le \gamma_{js} \le 1, \quad \sum_{s=1}^{S} \gamma_{js} = 1. \qquad (4)$$

Le vecteur constitué sera noté MFCC_D.8_LSF.2 (34). Il indique une pondération de 0,8 pour les deux premiers flux et de 0.2 pour le troisième correspondant aux LSFs. D'autre part, nous visons à optimiser l'utilisation de ces flux de paramètres en réduisant la dimension des vecteurs acoustiques tout en améliorant la robustesse du système. Pour cela, on applique une KLT sur l'ensemble des flux constituant le vecteur. On notera, par exemple, MFCC_D.8_LSF.2 (KLT_22) pour indiquer que KLT est appliquée sur le vecteur de dimension 34, à l'ensemble des trois flux à pondérations différentes (0.8 pour les MFCCs et 0.2 pour les LSFs), où l'on retient les 22 premières composantes.

4.4 Analyse des résultats

Les résultats donnés dans la table 1 montrent que l'utilisation du *front-end LSF* de dimension 34 conduit à une amélioration significative du taux de reconnaissance par rapport au front-end basic. Nous remarquons que notre approche est d'autant meilleure lorsque le RSB diminue. La substitution des composantes accélérations et énergies par les LSFs dans le vecteur de base conduit à une amélioration de taux de reconnaissance et à une dimension de vecteur conséquente. De plus, pour une pondération différente des flux utilisés, on constate une nette amélioration à partir de 10dB. À ce niveau de rapport signal sur bruit (10 dB), 20% de contribution des LSFs par rapport aux MFCCs permet d'améliorer le taux de reconnaissance de manière significative (jusqu'à 25%). D'autre part, nous remarquons que KLT appliquée sur le nouveau vecteur pondéré, réduit et optimise à son tour l'espace original de paramètres. KLT a ainsi conduit à une meilleure performance avec moins de paramètres. En conditions défavorables, la décomposition par KLT en sous-espaces donne de meilleurs résultats comparativement au *front-end* conventionnel d'ETSI. Expérimentalement, l'optimisation a été obtenue pour une dimension N= 24 en utilisant la CI-KLT pour différents types de bruit avec des RSB variant de -5 dB à 20dB (Figure 2).

Figure 2 : Taux de reconnaissance moyen réalisés avec différents types de bruits et des valeurs de RSB allant de 20 à -5 dB pour différents nombres de composantes KLT.

5. Conclusion

Nous avons effectué une analyse complémentaire du codec basique de l'ETSI utilisé en reconnaissance distribuée de

Table 1.Taux de reconnaissance (en %) du système DSR de base et ceux utilisant l'approche multi-variable sur les répertoires Test A et B de la base de données Aurora. Les meilleurs taux sont donnés en gras

Type de bruit		Vecteur multi-variable	20db	15db	10db	5db	0db	-5db
Test A	Réception	MFCC-E-D-A (39)	90,15	73,76	49,43	26,81	9,28	1,57
		MFCC-D-LSF (34)	82,92	75,48	61.01	37,94	18,32	9,52
		MFCC-D.8-LSF.2 (34)	91.32	86.03	**75.51**	**53.72**	**25.01**	**11.52**
		MFCC-D.8-LSF.2 (KLT_22)	**93.74**	**89.06**	75.01	50.57	23.55	11.19
	Voiture	MFCC-E-D-A (39)	**97,41**	**90,04**	67,01	34,09	14,46	9,39
		MFCC-D-LSF (34)	88,46	77,87	53,18	24,63	15,57	10,23
		MFCC-D.8-LSF.2 (34)	86.46	78.65	58.75	29.88	17.95	11.96
		MFCC-D.8-LSF.2 (KLT_22)	95.85	89.29	**68.12**	**34.63**	**19.21**	**12.73**
Test B	Restaurant	MFCC-E-D-A (39)	89,99	76,24	54,77	31,01	10,96	3,47
		MFCC-D-LSF (34)	81,89	75,93	62,02	38,62	18,71	9,64
		MFCC-D.8-LSF.2 (34)	81.95	76.24	65.86	41.97	22.29	11.18
		MFCC-D.8-LSF.2 (KLT_22)	**93.00**	**88.92**	**76.27**	**49.62**	**22.81**	**11.82**
	Aéroport	MFCC-E-D-A (39)	90,64	77,01	53,86	30,33	14,41	8,23
		MFCC-D-LSF (34)	74,77	66,03	51,01	33,46	17,63	9,07
		MFCC-D.8-LSF.2 (34)	81.84	74.44	62.24	42.29	23.89	12.38
		MFCC-D.8-LSF.2 (KLT_22)	**92.25**	**87.56**	**76.71**	**51.27**	**26.84**	**13.39**

la parole ETSI DSR basic. Les résultats obtenus montrent que les LSFs améliorent la performance de la DSR, comparativement à celle de la DSR front-end de base de l'ETSI utilisant les MFCCs seuls. Cette amélioration est d'autant plus importante lorsqu'on effectue un prétraitement par KLT, particulièrement pour des milieux fortement bruités. D'autre part, bien que l'extraction des nouveaux paramètres ajoute une certaine complexité au processus d'analyse, l'utilisation d'un vecteur de dimension 24 au lieu de 39 diminuera le temps de calcul et la capacité de stockage pour le processus principal effectué sur le serveur. Ce travail est en cours de continuation pour évaluer l'apport de ces nouveaux paramètres dans un environnement bruité avec un *front end* robuste au bruit tel que l'ETSI *Advanced* ou *Extended front-end*, utilisant les Mel transformés des LSFs (MLSF), en complémentarité avec les MFCCs.

BIBLIOGRAPHIE

[1] D. Addou, S-A. Selouani, K. Kifaya, M. Boudraa, B. Boudraa. A noise-robust front-end for distributed speech recognition in mobile communications. *International Journal of Speech Technology*, pages. 167-173, 2009.

[2] Tan Zheng-Hua. Automatic Speech Recognition on Mobile Devices and over Communication Networks. *Lindberg Børge (Eds.) Springer-Verlag*, XX, 404 p. 115, 2008.

[3] I. T. Jolliffe. Principal Component Analysis. *Second Edition. Springer*, 2002.

[4] Yi Hu, P.C. Loizou. A subspace approach for enhancing speech corrupted by colored noise, *Signal Processing Letters IEEE*, Volume 9, Issue 7, pages 204 – 206, 2002.

[5] ETSI. Speech processing, transmission and quality aspects; distributed speech recognition; front-end feature extraction algorithm; *Technical report of ETSI* ES 201 108, 2003.

[6] A. Sharma, K.K.Paliwal, G.C. Onwubolu**.** Class-dependent PCA, MDC and LDA: A combined classifier for pattern classification. *Pattern Recognition*, Volume 39, Issue 7, pages. 1215-1229, 2006.

[7] K. Fukunaga. Introduction to Statistical Pattern Recognition. *Second Edition. Academic Press*, 1990.

[8] ITU Recommendation G.712, Transmission performance characteristics of pulse code modulation channels, Nov. 1996.

[9] S.J. Young et al. HTK version 3.4: Reference and User Manual. Cambridge University, 2006.

[10] ITU-T Recommendation G.723.1. Dual rate speech coder for multimedia communications transmitting at 5.3 kbit/s, 1996.

[11] R. Rose and P. Momayez, Integration of multiple features sets for reducing ambiguity in automatic speech recognition. *Proc. IEEE-ICASSP*, pages 325-328, 2007.

Expériences et recommandations pour la structuration des données sonores, physiologiques et cliniques dans le cas des dysfonctionnements de la parole

Alain Ghio[1], Gilles Pouchoulin[1,2], Lise Crevier[3], Cécile Fougeron[3], Corinne Fredouille[2], Antoine Giovanni[1], Danièle Robert[1], Antonia Simon[3], Bernard Teston[1], François Viallet[1]

[1]Laboratoire Parole et Langage, CNRS, Université Aix-Marseille, France
[2] Université d'Avignon, CERI/LIA, Avignon, France
[3] Lab. de Phonétique et Phonologie, UMR 7018 CNRS-Paris3/Sorbonne Nouvelle, Paris, France
alain.ghio@lpl-aix.fr, corinne.fredouille@univ-avignon.fr, cecile.fougeron@univ-paris3.fr

ABSTRACT

Research in clinical phonetics needs a structured organisation of various data. In such a domain, speech signal is useless if it is unlinked to the clinical state of the speaker or to the speech production context. In this paper, we present a model of database which is designed for the storage and accessibility of various speech disorder data. Information contained in this model can be considered as recommendations based on our own experience in recording data in a clinical context.

Keywords: clinical phonetics, voice/speech disorders, database

1. INTRODUCTION

Depuis une quinzaine d'années, l'étude des dysfonctionnements de la voix et de la parole est sortie du simple cadre de la recherche clinique et intéresse les laboratoires de recherche issus des sciences du langage ou du traitement automatique de la parole. Par l'observation des dysfonctionnements, les chercheurs non cliniciens confrontent les résultats de leur recherche établis sur des corpus de parole "normale" à des situations de dysfonctionnement. Ces recherches permettent un enrichissement des connaissances générales sur la parole et bénéficient de la complémentarité des expertises des communautés de cliniciens et des autres scientifiques. A ce titre, on peut citer les travaux de Hardcastle et Gibbon [1] où des techniques d'Electropalatographie conçues au départ pour les études de phonétique, ont été utilisées pour des locuteurs souffrant de troubles de l'articulation, initiative ayant donné lieu à un grand nombre de publications ainsi qu'à un réseau de santé CleftNet.

2. PARTICULARITÉS DES DONNÉES DE « PAROLE PATHOLOGIQUE »

L'expression "parole pathologique" est un raccourci pour désigner la parole produite par des locuteurs atteints de dysfonctionnements de la voix et de la parole. L'étude multidisciplinaire de la "parole pathologique" nécessite trois exigences : (1) des signaux de très bonne qualité pour ne pas imputer des distorsions, bruits… aux dysfonctionnements; (2) un matériau linguistique (énoncé produit par les locuteurs) suffisamment riche ; (3) des renseignements cliniques sur les locuteurs suffisamment précis (pathologie, thérapies, contexte clinique d'enregistrement). Actuellement, les études sur le dysfonctionnement de la voix et de la parole souffrent cruellement d'une dispersion et d'une hétérogénéité des données. Souvent, les analyses portent sur quelques locuteurs enregistrés pour les besoins ponctuels d'une étude. L'enregistrement des signaux et le stockage sont souvent effectués par du personnel non expérimenté à certains aspects techniques de la prise et du formatage de données. A cela s'ajoute la perte fréquente des métadonnées (ex: pathologie, durée de la maladie, âge, contexte d'enregistrement…) : ces informations sont utilisées comme critères d'inclusion dans une étude particulière puis ne sont plus conservées de façon pérenne et sont dissociées des données sonores.

Pourtant, toute généralisation à une population clinique particulière nécessite l'observation d'un grand nombre de patients du fait de la très forte variation interindividuelle rencontrée (différentes évolutions de la maladie, stratégies de compensation individuelles, sévérité et spécificité des atteintes variées, …). Certaines pathologies étant rares, la disponibilité des patients dans un contexte clinique n'étant pas toujours possible, l'acquisition de données de parole pathologique n'est pas chose aisée. Pour ces raisons, il est important de mutualiser les enregistrements existants. Or pour être utilisables, ces enregistrements doivent répondre aux exigences mentionnées ci-dessus, ce qui est rarement le cas en pratique. Si les problèmes de prise de son ou autres signaux physiologiques sont en passe de devenir anecdotiques grâce à la diffusion de matériels de qualité et à la meilleure formation des personnels en charge des enregistrements, si le stockage des signaux de parole ne constitue plus actuellement un obstacle, si le recours à du matériau linguistique suffisant se généralise, le maillon faible reste la normalisation et la structuration des données sur les locuteurs et leurs productions langagières. Concrètement, si les données sonores sont souvent accessibles, parfois au prix d'un important effort de numérisation quand le support de stockage est sous forme de bandes, de cassettes…, elles ne présentent au final aucun intérêt si les liens entre les enregistrements et les caractéristiques cliniques du locuteur sont rompus. Nous touchons là aux différences fondamentales entre recueil de données vs base de données, différences que nous aborderons plus loin. L'objectif de ce document est de présenter différentes actions de terrain et de proposer des recommandations pour la structuration des données sonores, physiologiques et cliniques.

3. CORPUS DE PAROLE PATHOLOGIQUE

3.1. Patients dysphoniques du service ORL du CHU Timone, Marseille

Depuis plus de quinze ans, sous l'impulsion d'Antoine Giovanni, le service ORL du CHU de la Timone à Marseille enregistre régulièrement des patients dysphoniques qui se présentent à la consultation. Ces personnes sont à la fois enregistrées avec l'appareillage EVA [2] mais aussi sur des cassettes D.A.T. Pour des raisons de service, l'information sur les patients est stockée sur un registre sous la forme de cahiers sur lesquels sont notés l'identité du locuteur, sa pathologie, la date de l'examen, le contexte pré/post opératoire… Un important travail de numérisation, d'indexation et de saisie de renseignements a permis de disposer à présent d'une base de données de 1530 patients dysphoniques produisant des /a/ tenus, de la lecture de texte, de l'improvisation chantée…, pour un total de 1961 sessions d'enregistrements (certains locuteurs sont enregistrés plusieurs fois). Cette base est constituée de 504 hommes et 1026 femmes. Les pathologies principales sont : 314 nodules, 228 paralysies laryngées, 202 polypes, 138 œdèmes de Reinke, 139 dysphonies dysfonctionnelles, 81 kystes, 37 sulcus, 25 vergetures. A notre connaissance, cette base de données dépasse le contenu de la référence internationale du Massachusetts Eye and Ear Infirmary (MEEI), qui inclut 700 patients [3]. Nous prévoyons de compléter cette base en ajoutant les données physiologiques récoltées avec EVA sur ces mêmes patients et de l'enrichir avec les locuteurs récemment enregistrés. L'ajout de résultats d'évaluation perceptive (GRBAS) ou instrumentale est à l'étude, ce qui permettrait d'extraire facilement, par exemple, toutes les locutrices avec nodules en situation pré-opératoire avec un grade de sévérité de dysphonie entre 1 et 2.

3.2. Patients dysarthriques du service de neurologie du CH du Pays d'Aix

Depuis une dizaine d'années, sous l'impulsion de François Viallet, le service de neurologie du CH du Pays d'Aix enregistre régulièrement des patients dysarthriques qui se présentent à la consultation. Ces personnes sont enregistrées avec l'appareillage EVA et les données cliniques sont saisies avec un tableur. Actuellement, nous disposons des enregistrements sonores et aérodynamiques de 990 patients et 160 sujets contrôles appairés en âge (âge moyen patients = 67,7 ans ; contrôles = 62 ans). La population de patients est composée pour la majorité de malades de Parkinson (601) et syndromes parkinsoniens (98). L'originalité de ce travail réside dans (1) la présence de signaux complémentaires au signal sonore (intensité SPL, débit d'air oral, pression sous-glottique estimée…), ce qui représente 75000 fichiers pour toute la base ; (2) la multiplication des contextes d'enregistrement des 601 malades de Parkinson (avec/sans médication, avec/sans stimulateur sous-thalamique…), ce qui représente 1616 sessions d'enregistrements ; (3) le recueil de renseignements très précis sur le locuteur (date et lieu de naissance, langue maternelle…), et les conditions cliniques (date d'apparition de la maladie, localisation des symptômes, dosage médicamenteux d'usage et au moment de l'enregistrement, caractéristiques des éventuels stimulateurs électro physiologiques, scores des examens cliniques de type UPDRS, …) Une telle précision est importante pour des études cliniques (ex : effet des thérapies sur la production de parole) mais aussi au niveau linguistique (recherche de caractéristiques phonético-acoustiques spécifiques à un groupe homogène de locuteurs dysarthriques).

3.3. Patients neurologiques CCM

Entre 1965 et 1997, Claude Chevrie-Muller (CCM) et son équipe du Laboratoire d'étude de la voix et de la parole (INSERM U3), ont enregistré des patients atteints de troubles neurologiques, qui lui étaient adressés dans un but d'évaluation des dysfonctionnements de la parole et de la voix. Ces enregistrements représentent plus de 1000 patients avec des pathologies diverses couvrant dysphonie et dysarthrie mais aussi anarthrie, aphasie, bégaiement et des pathologies psychiatriques. Les enregistrements ont été réalisés suivant un protocole standardisé (qui a un peu évolué dans le temps) d'environ 15 minutes de parole, incluant la production de listes de mots, de phrases, de séries automatiques, de voyelles tenues et de syllabes, une description d'image, la lecture d'un texte et de la parole spontanée. Les signaux audio et électroglottographiques étaient enregistrés en chambre sourde sur des bandes Revox avec indexation sur un cahier. Un dossier médical complet sur papier est associé à chaque enregistrement, comprenant les informations sur l'état civil ainsi que sur la nature de la pathologie, les symptômes et les traitements associés. Un important travail de numérisation des bandes Revox et des dossiers médicaux est actuellement en cours. Cette base est enrichie par des enregistrements réguliers de patients à l'Hôpital Européen Georges Pompidou par Lise Crevier-Buchman et son équipe, suivant le même protocole mais sur un support numérique de type D.A.T.

4. ORGANISATION EN BASE DE DONNÉES

L'intérêt majeur de construire une base de données mutualisant des ressources est de pérenniser ces informations, de permettre un échange et un enrichissement graduel via une plateforme web pour un groupe de travail. Pour cela, un modèle de BD a été conçu à partir d'une analyse fonctionnelle réalisée en milieu clinique, pour être ensuite affiné sur la disponibilité de données empiriques issues de différents corpus de parole pathologique comme ceux présentés dans la section 3.

Si les concepts autour des bases de données (BD) sont familiers aux informaticiens, il n'en est pas de même pour les non spécialistes. Il est fréquent de lire qu'un recueil d'enregistrements sonores constitue une base de données : on assimile ainsi corpus/collection de données avec BD. Cette dernière se distingue de la première par une structuration et une organisation cohérente régie par un

modèle, partageable par un groupe de personnes et mémorisable sur un support informatique permettant de sélectionner facilement les données répondant à des critères précis. Cela nous amène à aborder la notion de Système de Gestion de Bases de Données (SGBD) qui doit supporter les concepts définis au niveau du modèle de données, assurer le respect des règles de cohérence définies sur les données, rendre transparent le partage des données entre différents utilisateurs tout en assurant la confidentialité sélective d'une partie des données, pouvoir répondre à des requêtes avec un niveau de performances adapté, fournir différents langages d'accès selon le profil de l'utilisateur. Nous avons opté pour un modèle relationnel, considéré comme le plus simple et le plus élégant des modèles de BD. Sa simplicité provient de l'organisation tabulaire des données, atomiste et minimaliste, rendant intuitive l'architecture des données, les éléments de chaque table étant ensuite liés par des relations. Comme le montre la figure 1, la base de données est composée d'une cinquantaine de tables décrivant les locuteurs avec des informations civiles (date et lieu de naissance, lieu de résidence…), des informations sociolinguistiques (langue maternelle, professions…), médicales (prescriptions thérapeutiques, recueil des symptômes, diagnostics établis) et sur les sessions d'enregistrements. Pour standardiser certaines de ces informations, un ensemble de listes (Figure 1, tables "lst_") permet de saisir de façon commune aux différents centres fournisseurs de données, des renseignements normalisés tels que professions, langues, pays/régions, symptômes, thérapies, diagnostics, facteurs de risque, localisation des pathologies, contextes expérimentaux, méthodes d'évaluations… L'intérêt de ces listes fermées est d'éviter la multiplication d'appellations pour un même terme (ex : MDP, maladie de Parkinson, Parkinson, Parkinsonien => [52] maladie de Parkinson).

5. RECOMMANDATIONS ET RÉALISATIONS TECHNIQUES

Contrairement aux bases de données orales de type patrimonial, dialogal, conversationnel, comme celles disponibles au CRDO [4], la parole pathologique nécessite la collecte et la mémorisation précise des informations sur les locuteurs et le contexte d'enregistrement. Par conséquent, il est recommandé de saisir le maximum de renseignements possible sur les aspects :
(1) sociolinguistiques (ex : certaines dysarthries se manifestent par une élision des /r/, phénomène qui peut s'apparenter à un accent « créole » ; seules des informations sur le lieu de naissance ou de résidence du locuteur peuvent permettre de faire la part des choses) ;
(2) médicaux (quelques commentaires sur l'état du patient peuvent permettre de faire des choix d'inclusion/exclusion pour certaines études) ;
(3) symptomatiques (ex : date d'apparition et localisation des symptômes) ;
(4) contextuels (ex : « le patient sort d'une bronchite, porte un appareil dentaire, a pris ses médicaments 4h auparavant ») ;

De même, toute forme d'évaluation (ex : UPDRS pour les malades de Parkinson, GRBAS pour les dysphoniques) constituent une source d'information à conserver précieusement.

Si l'utilisation d'un SGBD est recommandée pour la traçabilité et l'exploitation des métadonnées, la standardisation du protocole d'acquisition de données sonores ou physiologiques apparaît difficilement conciliable avec le contexte clinique. En effet, un protocole complet incluant la production de voyelles tenues, d'efforts vocaux, de phrases, de répétitions, de textes lus, de parole spontanée… est difficilement réalisable compte tenu de la batterie complète d'examens que subit le patient et de la fatigabilité engendrée par de trop longs efforts. Il est donc préférable d'adapter les taches d'élocution à l'état dysfonctionnel du locuteur. Par exemple, l'étude de la nasalité est particulièrement intéressante dans le cas de dysarthries paralytiques du fait de l'immobilité du voile du palais, mais n'est pas centrale dans la maladie de Parkinson pour laquelle des exercices phonatoires peuvent être préférés du fait de l'hypophonie.

A propos de la réalisation technique, la base de données de parole pathologique est développée dans l'environnement PHP/MySQL sur un serveur Apache avec module SSL (Secure Sockets Layer) pour le cryptage des communications. Concernant la sécurité et l'accès aux données, il est recommandé de gérer des privilèges/rôles accordés aux utilisateurs et de crypter les données confidentielles. Ces recommandations nécessaires ne dispensent pas une déclaration à la Commission Nationale de l'Informatique et des Libertés (CNIL) pour le recueil, la saisie des informations cliniques et l'analyse statistique des données recueillies au cours de la recherche [5]. L'anonymisation des données ainsi que l'établissement de consentements éclairés auprès des locuteurs sont des aspects juridiques à ne pas négliger.

Remerciements : Ce travail a été financé par un PHRC CH du Pays d'Aix et est financé actuellement par l'ANR-08-BLAN-0125 « DESPHO-APADY ».

BIBLIOGRAPHIE

[1] W. Hardcastle, F. Gibbon, Electropalatography and its clinical applications. In *Instrumental Clinical Phonetics*. 149-195, 1997

[2] A.Giovanni, N.Estublier, D.Robert, B.Teston, M. Zanaret, M.Cannoni, Evaluation vocale objective des dysphonies par la mesure simultanée des paramètres acoustiques et aérodynamiques à l'aide de l'appareillage EVA, *Ann. Otolaryngol. Chir. Cervicofac.*, 112, p. 85-90, 1995

[3] Massachusetts Eye and Ear Infirmary. Elemetrics Disordered Voice Database. Voice and Speech Lab, Boston, MA, Kay Elemetrics Corp, 1994

[4] B. Bel, P. Blache, Le Centre de Ressources pour la Description de l'Oral (CRDO). *Travaux interdisciplinaires du LPL d'Aix-en-Prov. (TIPA)*, vol. 25. 2006, p. 13-18.

[5] CNIL, Méthodologie de référence pour les traitements de données personnelles opérés dans le cadre des recherches biomédicales, MR-001, 2006, www.cnil.fr

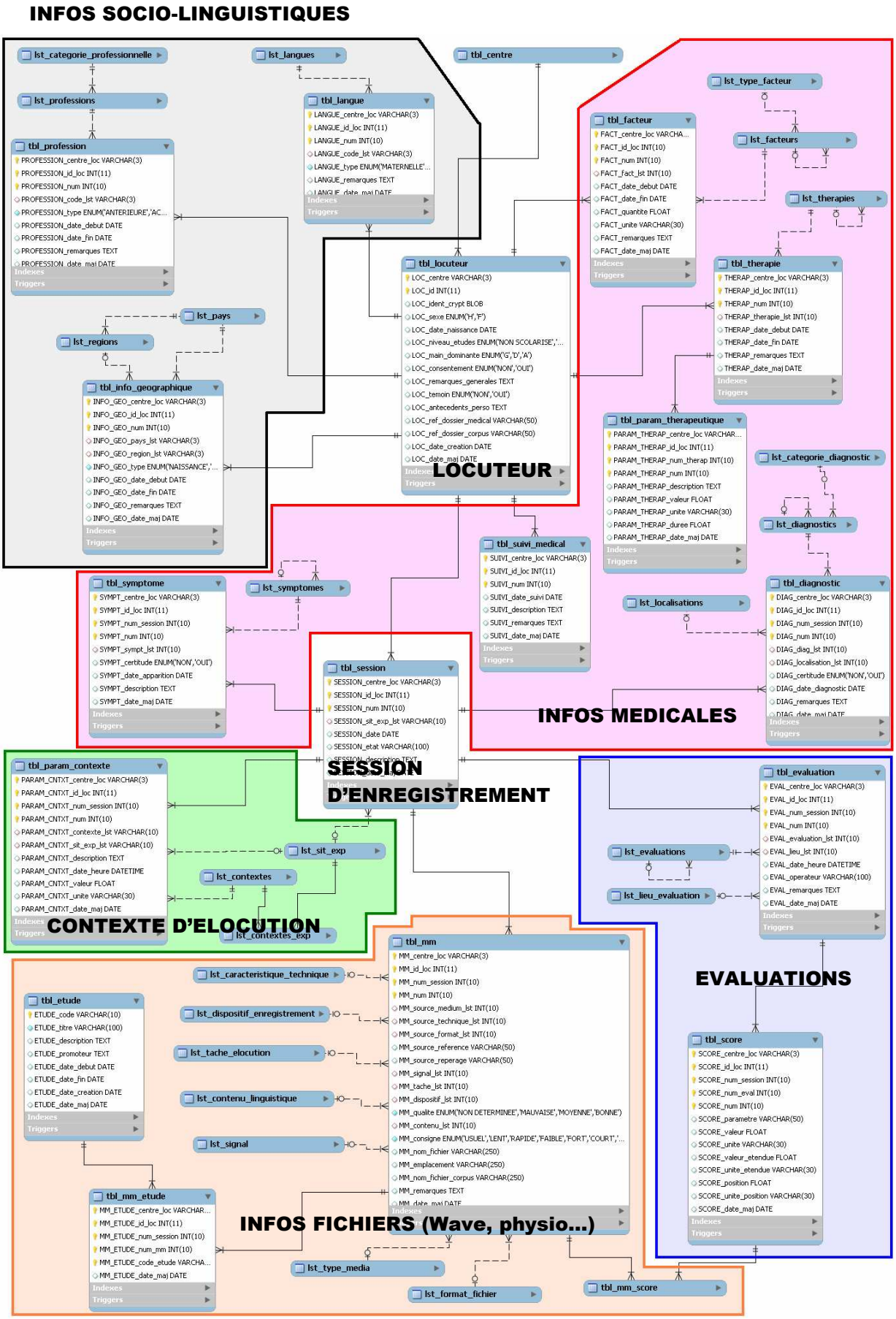


Figure 1 : Modèle conceptuel généralisable d'une base de données de « parole pathologique »

La base de données AVLaughterCycle

Jérôme Urbain[1], *Elisabetta Bevacqua*[2], *Thierry Dutoit*[1], *Alexis Moinet*[1], *Radoslaw Niewiadomski*[2], *Catherine Pelachaud*[2], *Benjamin Picart*[1], *Joëlle Tilmanne*[1] *et Johannes Wagner*[3]

[1]TCTS Lab, Faculté Polytechnique, Université de Mons, Boulevard Dolez 31, 7000 Mons, Belgium
[2]CNRS - LTCI UMR 5141, Institut TELECOM - TELECOM ParisTech,
37/39, rue Dareau - 75014 Paris, France
[3]Institut für Informatik, Universität Augsburg, Universitätsstr. 6a, 86159 Augsburg, Germany

ABSTRACT

In this paper, the AVLaughterCycle database is presented. The database consists in the audio and facial motion capture recordings of 24 subjects watching a funny video. It is the first database of laughter combining these 2 modalities. There are around 1000 laughter episodes in the database, covering a large variety of shapes. The database annotation protocol is meant to not only build a large laughter class grouping all the utterances, but distinguish different kinds of laughter. The database is publicly available for research purposes. It can be used for many laughter studies. For example, in the AVLaughterCycle project it was used to endow a virtual agent with the capability of appropriately joining its conversational partner's laughter.
Keywords : laughter, corpus, facial motion capture

1. Introduction

Le rire est un facteur essentiel des relations humaines. Il a la faculté non seulement de transmettre nos émotions mais aussi, par ses vertus communicatives, d'en susciter chez nos interlocuteurs. Différentes disciplines scientifiques (psychologie, médecine, étude du langage, etc.) s'intéressent au rire afin d'identifier ses causes et mécanismes de production, de décrire ses caractéristiques (par exemple, Bachorowski et al. [1]) ou encore de mesurer ses effets. Les progrès en traitement automatique de la parole suscitent un intérêt croissant pour des systèmes capables d'automatiquement détecter le rire ou de le synthétiser de manière naturelle. Si dans un premier temps seul le signal acoustique était étudié (Truong et van Leeuwen [11], Sundaram et Narayanan [10]), Petridis et Pantic [8] y ont récemment ajouté les mouvements faciaux pour détecter le rire. L'émergence des agents virtuels pousse également à combiner les deux modalités (audio et vidéo) afin de doter les avatars de la faculté de rire.

Pour ces travaux de traitement automatique du rire, les données sont cruciales. Mais le rire, comme tous les signaux émotionnels, est difficile à enregistrer car il n'y a aucune garantie qu'un rire acté soit similaire à un rire spontané. Il convient donc, idéalement, d'utiliser des rires naturels, produits par des locuteurs qui ne sont pas conscients qu'ils sont enregistrés. Peu de bases de données ont été enregistrées de cette manière. Celles qui s'en approchent le plus, quoique les sujets savaient qu'ils étaient enregistrés, sont les bases de données ICSI Meeting Corpus (Janin et al. [4]), uniquement audio, obtenue en plaçant des micros lors de réunions techniques, et AMI Meeting Corpus (Carletta [2]), audiovisuelle, consistant en des interview de citoyens racontant des situations émotionnelles. Dans aucun cas le rire n'était l'objectif principal de l'enregistrement des données.

C'est dans ce contexte que s'inscrit cet article. Il décrit une base de données audiovisuelle de rires, appelée AVLaughterCycle, enregistrée dans le cadre du projet eNTERFACE'09 du même nom. Le but du projet était de faire rire un agent virtuel, Greta (Niewiadomski [7]), de manière naturelle en synchronisant les signaux audio et visuels de rires humains. Le signal acoustique était simplement rejoué, alors que les mouvements faciaux ont dû être adaptés à l'animation et la morphologie de Greta. L'application AVLaughterCycle nécessitait une représentation précise des mouvements faciaux lors du rire, données difficiles à obtenir à partir de la base de données AMI. Nous avons donc décidé d'enregistrer une nouvelle base de données, focalisée sur le rire et combinant le signal acoustique à une capture de mouvements faciaux robuste. Les sujets étaient conscients de l'enregistrement : ils étaient laissés seuls dans une pièce éclairée, face à des caméras et avec des marqueurs placés sur le visage. Nous avons utilisé une méthode d'induction pour amener les sujets à l'état souhaité (le rire).

L'article est organisé comme suit : les outils utilisés pour enregistrer les données sont présentés à la Section 2. Le protocole d'enregistrement est exposé à la Section 3. L'annotation est décrite à la Section 4. La Section 5 est dédiée au contenu de la base de données. Enfin, l'application AVLaughterCycle est résumée à la Section 6, avant la conclusion (Section 7).

2. Outils d'enregistrement

La base de données à été enregistrée à l'aide d'une webcam (25 images par seconde, RGB 24 bits, 640x480 pixels) et d'un micro-casque pour l'enregistrement audio (16kHz, PCM 16 bits) et la diffusion de stimulus sonore. De plus, deux outils particuliers ont été utilisés pour construire la base de données : le logiciel "Smart Sensor Integration" (Wagner et al. [13]) et des systèmes de capture de mouvements faciaux. Ces outils sont présentés ci-dessous.

2.1. Smart-Sensor Integration (SSI)

Ce logiciel permet la lecture synchrone de plusieurs signaux d'entrée (dans notre cas, audio et vidéo) et

propose une interface graphique permettant de soumettre les utilisateurs à une série de stimulus audiovisuels (texte, images, vidéos, etc.) sous forme de pages HTML afin de susciter des réactions. Le logiciel permet également l'annotation des données. Les signaux enregistrés peuvent être analysés automatiquement (en temps réel ou non) à l'aide d'algorithmes définis dans une librairie éditable. Via la librairie Torch3D, le logiciel propose également plusieurs types de classificateurs ($HMMs$, $GMMs$, $kNNs$, etc.) pour entraîner des modèles sur les données annotées. Ces modèles peuvent ensuite être utilisés pour classer de nouvelles données, en temps réel ou non.

2.2. Capture des mouvements faciaux

Comme précédemment expliqué, nous souhaitions disposer de mesures précises des mouvements faciaux du rire. Nous nous sommes orientés vers des systèmes de capture de mouvements utilisant des marqueurs placés sur les sujets, plus robustes que les systèmes n'en utilisant pas. Deux logiciels commerciaux ont été successivement utilisés : ZignTrack et OptiTrack.

ZignTrack [6] est un logiciel bon marché nécessitant 22 marqueurs (autocollants ou points de maquillage) et n'utilisant qu'une seule caméra (notre webcam). Les mouvements ne sont donc pas réellement enregistrés en 3D mais en 2D, à partir de laquelle la 3D est regénérée en utilisant un modèle fixe de morphologie faciale. Cela pose des problèmes de reconstruction du visage en 3D lors de rotations de la tête. De plus, le tracking des marqueurs n'est pas assez robuste pour permettre une extraction automatique des mouvements faciaux du rire : de nombreuses corrections manuelles sont nécessaires.

OptiTrack [5] est un logiciel professionnel utilisant 6 caméras infrarouges disposées de manière semi-sphérique. Au moins 23 réflecteurs infrarouges doivent être collés sur le visage des sujets, en plus d'un bandeau muni de 4 réflecteurs placé sur la tête et servant à mesurer les rotations de la tête. Grâce aux 6 caméras, les mouvements sont captés directement en 3D, sans passer par un modèle morphologique. Le tracking réalisé à l'aide de ce système est très robuste.

3. Protocole

Les sujets étaient des volontaires parmi les chercheurs participant au Workshop eNTERFACE'09 à Gênes (Italie). Au total, 24 sujets ont été enregistrés individuellement. Ils étaient originaires de 11 pays : Belgique, Canada, Corée du Sud, Etats-Unis, France, Grèce, Inde, Italie, Kazakhstan, Royaume-Uni et Turquie. 8 participants (3 femmes, 5 hommes) ont été enregistrés avec une capture de mouvements faciaux via le système ZignTrack. Les 16 autres (6 femmes, 10 hommes) ont été enregistrés avec OptiTrack. L'âge moyen des sujets était de 29 ans (écart-type : 7.3 ans).

A cause des capteurs nécessaires à la capture robuste des mouvements faciaux, il nous était impossible d'enregistrer les sujets à leur insu. Pour obtenir des rires spontanés alors que les sujets se savaient analysés, nous avons décidé de leur montrer une vidéo humoristique d'une dizaine de minutes comprenant une série de clips trouvés sur Internet. Les clips se succèdent sans pause : le rire provoqué par un clip peut donc être influencé (écourté ou renforcé) par le clip suivant. Pour éviter des problèmes de sauvegarde, la vidéo a du être séparée en 3 sessions distinctes d'environ 3 minutes lorsque le système OptiTrack était utilisé.

Avant la séance, les capteurs étaient placés sur le visage du sujet. Celui-ci était alors invité à mettre le micro-casque et s'asseoir devant un écran muni de la webcam. Lorsqu'OptiTrack était utilisé, les 6 caméras infrarouges étaient ajoutées autour du sujet. Une page d'instructions lui était alors présentée, avec les consignes suivantes : le sujet devait se détendre et réagir librement à la vidéo, en faisant toutefois attention à garder sa tête vers l'écran et à ne rien placer entre la webcam et son visage tout au long de l'expérience. Une fois les instructions assimilées, le sujet était laissé seul dans la pièce et l'expérience démarrait. A la fin de la vidéo, une page invitait le sujet à produire un rire acté avant de terminer l'enregistrement. Tous les participants ont donné un accord écrit, signé au terme de l'expérience, d'utiliser leurs données à des fins non commerciales. La base de données est disponible à l'adresse *http://tcts.fpms.ac.be/~urbain*. Elle contient les enregistrements audio et les annotations. Les enregistrements vidéo (webcam) et les vidéos de stimulus peuvent être obtenus sur demande. La capture des mouvements faciaux (25FPS avec ZignTrack, 100FPS avec OptiTrack) y sera ajoutée prochainement.

4. Annotation

La base de données est annotée par une personne à l'aide du logiciel *SSI*. Un protocole d'annotation hiérarchique a été mis au point afin de distinguer 6 classes principales (silence, parole, respiration, applaudissement, rire et "bruit", qui regroupe les sons n'appartenant pas aux 5 autres classes), tout en donnant la possibilité d'ajouter des précisions à l'intérieur d'une de ces classes, en particulier la classe *Rire* qui est le centre d'intérêt de cette base de données. Dans cette catégorie, des précisions peuvent notamment être apportées sur :

- la structure du rire : en spécifiant s'il contient une seule syllabe ou plusieurs et s'il y a plusieurs sections distinctes séparées par des inhalations audibles, appelées "*bouts*".
- le type de son(s) rencontré(s) : voyelle, nasal, chuchotement, grognement, fredonnement, hoquet, etc.

Chaque segment est assigné à une seule classe principale. En cas d'ambiguïté, le segment est annoté "à écarter" afin d'éviter de détériorer les modèles entraînés sur les classes principales (par exemple lorsqu'un téléphone sonne au milieu d'un rire). Il n'y a pas de restriction sur le nombre de précisions apportées : les sous-classes relatives à la structure du rire sont mutuellement exclusives mais peuvent être combinées à toutes les sous-classes de contenu acoustique, car le type de son peut varier au sein d'un épisode de rire.

L'annotation se fait principalement à l'aide du signal audio. Néanmoins, l'enregistrement vidéo est consulté pour affiner les positions des début et fin des segments de rire, ainsi que pour annoter les rires (quasiment) inaudibles. Un rire se termine fréquemment par une

forte inhalation (Chafe [3]), parfois plusieurs secondes après les exhalations principales. Lorsqu'une telle inhalation, manifestement provoquée par le corps du rire qui la précède, est présente, la fin du segment de rire est placée à la suite de cette inhalation.

5. Contenu

La table 1 présente le nombre d'occurrences des classes principales, hormis la classe silence qui est la classe par défaut. La table 2 présente les occurrences des sous-classes de rire. Les rires actés ne sont pas pris en compte. La base de données contient un millier de rires. La plupart des rires est constituée d'un seul "*bout*", contenant lui-même plusieurs syllabes. Les sons du type "voyelle" sont les plus fréquents, mais concernent moins de la moitié des rires. De nombreux rires ont un contenu nasal, assimilable à de la respiration ou à un fredonnement. Les sujets ne disposaient pas d'interlocuteur au cours de l'expérience, ce qui explique la faible fréquence de la classe "Parole" et, en conséquence, la quasi-inexistence de "speech-laughs" (lorsque le sujet parle et rit en même temps, ce qui module le signal de parole (Chafe [3])).

Table 1: Occurrences des classes principales

Classe principale	Occurrences
Rire	1039
Bruit	267
Parole	186
Applaudissement	93
Respiration	41
A écarter	31

Table 2: Occurrences des sous-classes de rire

Catégorie	Sous-classe de rire	Nombre
Structure	Monosyllabique	185
	Un "bout"	697
	Plusieurs "bouts"	157
Acoustique	Voyelle	453
	Nasal	284
	Respiration	245
	Fredonnement	171
	Hoquet	96
	Grognement	18
	Speech-laugh	20
	Silencieux	95

La durée moyenne d'un rire est de 3.5s (écart-type : 5.3s), incluant l'éventuelle inhalation finale. La figure 1 montre un histogramme de la durée des rires et sa fonction de distribution cumulative. Il y apparaît clairement que la plupart des rires sont assez courts (83% des rires durent moins de 5s). Il ne faut néanmoins pas négliger les rires plus longs, qui représentent 51.4% de la durée totale des rires. Le plus long rire dure 82s.

Le nombre et la distribution des rires est extrêmement variable d'un sujet à l'autre : certains sujets rient très peu, d'autres énormément, certains ont tendance à rire brièvement, d'autres produisent de nombreux longs rires, etc. Les causes potentielles de ces réactions variables au stimulus présenté sont nombreuses : différences culturelles, sensibilités diverses à différents

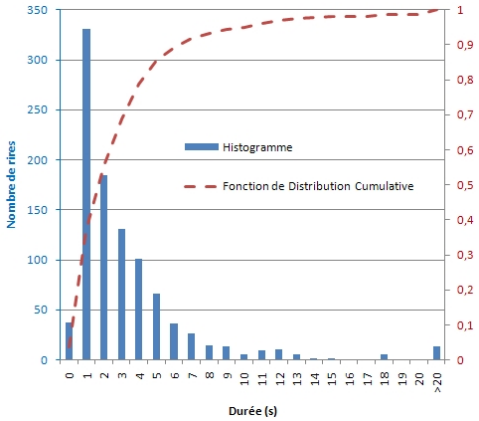


Figure 1: Histogramme et fonction de distribution cumulative de la durée des rires

types d'humour, influence de l'humeur des sujets au moment de l'enregistrement, etc. Des analyses supplémentaires seraient nécessaires pour déterminer les facteurs principaux expliquant les différents comportements constatés.

6. Application AVLaughterCycle

Le but du projet AVLaughterCycle était de construire un système capable d'enregistrer le rire de l'utilisateur et d'y répondre par un rire adéquat. C'est l'agent virtuel Greta qui joue le rire sélectionné. L'architecture de l'application est illustrée à la Figure 2.

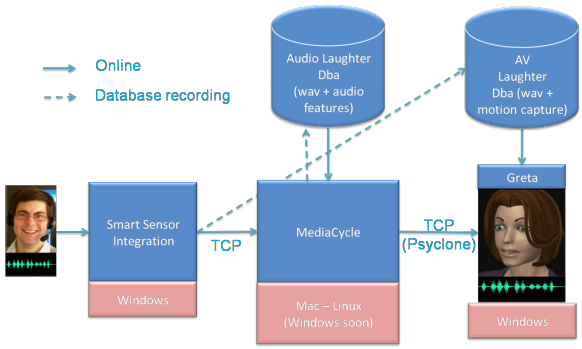


Figure 2: Architecture de l'application AVLaughterCycle

L'utilisateur rit dans le microphone. Le signal audio est analysé en temps réel par le logiciel *SSI* qui en extrait des caractéristiques spectrales pour chaque trame de 340 échantillons, avec un décalage de 85 échantillons entre 2 trames successives. Sur base du rapport signal à bruit, *SSI* segmente le signal audio. Dans le système actuel, il n'y a pas de réelle détection du rire mais une détection d'activité vocale. L'hypothèse est faite que le signal envoyé par l'utilisateur est un rire. Une fois un rire segmenté, *SSI* envoie les moyennes et écart-types de ses caractéristiques spectrales (MFCCs, Spectral Flatness, Loudness, etc.) au deuxième module, appelé MediaCycle.

MediaCycle (Siebert et al. [9]) est un logiciel permettant d'organiser une base de données multimédia et d'y naviguer de manière efficace. La base de données est organisée en fonction des similarités entre les objets. Les similarités sont estimées par la distance Euclidienne entre les vecteurs de caractéristiques des ob-

jets. Pour le projet AVLaughterCycle, seules les caractéristiques spectrales du signal audio ont été utilisées. Pour chaque rire de la base de données, MediaCycle gardait en mémoire les moyennes et écart-types des caractéristiques spectrales, normalisées. Lorsque *SSI* envoie un vecteur de caractéristiques à MediaCycle, celui-ci le compare avec les rires de la base de données. Le rire dont le vecteur de caractéristiques est le plus proche de celui du rire d'entrée est sélectionné. Ce système peut servir à naviguer dans la base de données en l'interrogeant par du rire.

Une fois le rire à jouer sélectionné, sa référence est envoyée à Greta afin qu'elle le joue instantanément : le son n'est pas modifié mais l'animation faciale de Greta est pilotée par les mouvements faciaux du rire, adaptés au visage de Greta (Urbain et al. [12]).

Le projet AVLaughterCycle a débouché sur la réalisation de cette chaîne complète de traitement en quasi temps réel : *SSI* analyse le signal audio en continu et, dès qu'un segment (supposé être du rire) est détecté, le rire le plus similaire est transmis à Greta qui le joue (en plus de son comportement "normal", ses mouvements des bras, etc., qui sont réalisés indépendamment du rire). Le système est opérationnel et les tests qualitatifs qui ont été effectués étaient prometteurs même s'ils ont mis en évidence la nécessité d'introduire des caractéristiques décrivant la structure et le rythme du rire. De plus amples informations sur le projet AVLaughterCycle peuvent être trouvées sur *http ://www.numediart.org/projects/07-4-avlaughtercycle/*. Les résultats de MediaCycle sont actuellement soumises à des évaluations objective (déterminer si MediaCycle est capable de grouper des rires du même type ou du même locuteur) et subjective (mesurer si MediaCycle est en accord avec perceptions de similarité humaines).

7. Conclusion

La base de données AVLaughterCycle, première base de données de rires comprenant à la fois le signal audio et un tracking précis des expressions faciales, a été présentée. Cette base de données est centrée sur le rire et est annotée en fonction, pour donner des indications sur la structure et le contenu de chaque rire. La base de données est disponible gratuitement pour des utilisations non commerciales. Elle contient environ un millier de rires, de types et longueurs variables. Ses applications potentielles couvrent les domaines suivants : l'analyse, la reconnaissance, la modélisation et la synthèse du signal audio du rire ou des mouvements faciaux ; l'étude simultanée de ces deux modalités et leur synchronisation, etc.

Remerciements

Le projet AVLaughterCycle a été partiellement financé par le projet Européen CALLAS (IP6, contrat n° 034800) et par le Ministère de la Région Wallonne en Belgique, via le Programme de Recherche Numédiart (contrat n° 716631). Joëlle Tilmanne dispose d'une bourse doctorale octroyée par le Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.) en Belgique.

Références

[1] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110 :1581–1597, 2007.

[2] J. Carletta. Unleashing the killer corpus : experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2) :181–190, 2007.

[3] W. Chafe. *The Importance of not being earnest. The feeling behind laughter and humor.*, volume 3 of *Consciousness & Emotion Book Series*. John Benjamins Publishing Company, Amsterdam, The Nederlands, 2007.

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong-Kong, April 2003.

[5] Natural Point, Inc. Optitrack. http ://www.naturalpoint.com/optitrack/.

[6] Zign Creations. Zign track. http ://www.zigncreations.com/zigntrack.html.

[7] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud. Greta : an interactive expressive ECA system. In C. Sierra, C. Castelfranchi, K. S. Decker, and J. S. Sichman, editors, *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 2*, pages 1399–1400. IFAAMAS, 2009.

[8] S. Petridis and M. Pantic. Is this joke really funny ? judging the mirth by audiovisual laughter analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1444–1447, New York, USA, June 2009.

[9] X. Siebert, S. Dupont, P. Fortemps, and D. Tardieu. MediaCycle : Browsing and performing with sound and image libraries. In T. Dutoit and B. Macq, editors, *QPSR of the numediart research program*, volume 2, pages 19–22. numediart, 2009.

[10] S. Sundaram and S. Narayanan. Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America*, 121(1) :527–535, January 2007.

[11] K. P. Truong and D. A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49 :144–158, 2007.

[12] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. AVLaughterCycle : An audiovisual laughing machine. In T. Dutoit and B. Macq, editors, *QPSR of the numediart research program*, volume 2, pages 97–104, Sept. 2009.

[13] J. Wagner, E. André, and F. Jung. Smart sensor integration : A framework for multimodal emotion recognition in real-time. In *Affective Computing and Intelligent Interaction*, 2009.

Effet du type de bruit sur le démasquage binaural chez l'adulte dyslexique

M. Dole[1], M. Hoen[2], F. Meunier[1]

1. Laboratoire Dynamique du Langage, UMR 5596, Lyon, France
2. Stem Cell and Brain Institute, U846, Bron, France
marjorie.dole@univ-lyon2.fr

ABSTRACT

The purpose of our study was to determine how binaural cues are used by dyslexic listener during speech-in-noise comprehension. We therefore designed a study testing binaural unmasking depending on the nature of the interfering noise. 15 adult dyslexics were tested using auditory stimuli made up of words presented in different types of noise (stationary noise, fluctuating noise and cocktail party noise) and in different stimulation configurations (dichotic, monaural and hybrid monaural/dichotic configuration).

Our results demonstrated important masking effects in the monaural condition. We observed a binaural unmasking ability in dyslexic participants, but only in the case of high-level informational masking.

Keywords: speech-in-noise, binaural unmasking, dyslexia, informational masking, energetic masking.

1. INTRODUCTION

La dyslexie développementale est définie comme un trouble de l'apprentissage affectant l'acquisition des compétences en lecture et en écriture. Plusieurs hypothèses ont été avancées pour expliquer les causes de ce trouble, la plus courante étant celle d'un déficit dans l'utilisation et/ou la représentation de l'information phonologique. Ces troubles phonologiques pourraient reposer sur un déficit de perception de la parole. Cependant, peu d'études ont réussi à montrer un tel déficit et les résultats sont plutôt ambigus [1-3], peut être parce que de nombreuses études ont cherché à mettre en évidence ce déficit dans des conditions d'écoute silencieuse. Or, les troubles caractérisant les dyslexiques sont souvent difficiles à observer en condition non-aversive, alors qu'ils sont magnifiés dans des situations plus difficiles (stress, bruit concurrent, etc).

Depuis quelques années, plusieurs études se sont ainsi intéressées à la perception de la parole dans le bruit chez le dyslexique. Ces études ont montré chez des enfants ayant des troubles développementaux du langage un déficit de perception de la parole dans le bruit mais pas dans le silence [4-5]. Ces données sont corroborées par des données neurophysiologiques montrant des réponses neuronales altérées dans le bruit mais pas dans le silence [6].

Du point de vue cortical, de nombreuses études expliquent l'origine du trouble dyslexique par un dysfonctionnement des aires temporales postérieures gauche associées au traitement phonologique [7]. La dyslexie semble aussi associée à des asymétries fonctionnelles moins marquées voire absentes; des résultats en MEG ont par exemple montré que la N1m enregistrée en réponse à la syllabe /ba/ était localisée plus postérieurement dans le cortex auditif droit chez les sujets dyslexiques adultes, produisant un pattern plus symétrique que chez les sujets normo-lecteurs [8]. Des enregistrements de potentiels évoqués auditifs en réponse aux syllabes /ba/ et /pa/ montrent une altération de la latéralisation à gauche chez des sujets adultes ayant un historique de dyslexie [9]. Au niveau anatomique, des anomalies péri-sylviennes ont été observées, telles qu'une moindre asymétrie du *planum temporale* corrélée aux résultats lors d'une tâche d'écoute dichotique [10], ou une moindre asymétrie des régions pariétales inférieures [11].

Ces anomalies de l'asymétrie corticale semblent aller de pair avec, au niveau périphérique, une réduction de l'asymétrie du système auditif efférent. Il s'agit d'une voie neuronale partant du cortex auditif primaire pour exercer un rétrocontrôle au niveau cochléaire et qui semble par ailleurs être impliquée dans la compréhension de la parole dans le bruit [12]. En 2007, Veuillet et al. ont montré une réduction de l'asymétrie de ce système chez des enfants dyslexiques, corrélée à une altération de la perception du voisement [13].

Ainsi il est envisageable que le déficit comportemental de perception de la parole dans le bruit observé chez les dyslexiques puisse être mis en rapport avec des anomalies structurales et fonctionnelles de l'asymétrie du système auditif. Une explication possible serait que ce défaut d'asymétrie observé chez les sujets dyslexiques soit un frein à l'utilisation d'indices acoustiques binauraux comme les différences interaurales, indices cruciaux pour la compréhension de la parole bruitée. Afin de tester la capacité d'utilisation de ces indices chez le sujet dyslexique, nous proposons une expérience de compréhension de la parole dans le bruit comparant des configurations d'écoute favorisant l'utilisation d'indices binauraux (démasquage binaural) ou au contraire l'abolissant (écoute monaurale). Ceci nous permettra d'évaluer les capacités de démasquage binaural chez le sujet dyslexique adulte, en fonction de la nature du bruit utilisé.

Le démasquage binaural est défini comme l'amélioration de l'intelligibilité d'un mot cible présenté avec du bruit dans les deux oreilles par rapport à une situation d'écoute monaurale où cible et bruit sont présentés dans une seule oreille. Il est notamment dû à l'utilisation d'indices binauraux qui aident à séparer les deux flux concurrents selon des critères spatiaux. Des résultats précédemment obtenus chez le sujet normo-lecteur ont montré que cet

effet de démasquage est plus important dans le cas d'un masquage informationnel de haut niveau (bruit de parole) que d'un masquage informationnel de bas niveau (bruit large spectre fluctuant) ou d'un masquage énergétique (bruit large spectre stationnaire) [14].

Le but de notre étude est donc d'évaluer de quelle manière le démasquage binaural est modulé chez le sujet dyslexique, en étudiant son amplitude dans différentes conditions de masquage.

2. METHODE

2.1. Participants et procédures

15 participants diagnostiqués dyslexiques ont participé à l'expérience. Tous sont de langue maternelle française, de 18 à 44 ans, droitiers et sans troubles auditifs. Leurs seuils auditifs ont été vérifiés à l'aide d'une audiométrie tonale ; sont exclues de l'expérimentation les personnes ayant des seuils auditifs supérieurs à 20dB dans le silence dans une gamme de fréquence de 125 à 8000 Hz.

Leur tâche consiste à écouter des stimuli auditifs, présentés à l'aide d'un casque audio (Beyerdynamic DT48, 200 Ω). Les stimuli sont composés de mots présentés dans du bruit. Il est demandé aux sujets de répéter le mot qu'ils ont entendu.

2.2. Stimuli

Mots cibles

126 mots bisyllabiques ont été sélectionnés dans une gamme de fréquence d'occurrence moyenne (de 0.23 à 338.19, moyenne : 16.81, D.S. : 43.74) selon la base de donnée Lexique 2 [15]. Ils sont prononcés par une voix féminine et enregistrés dans une pièce insonorisée.

Bruits

Trois types de bruits ont été utilisés : un bruit de paroles (Cocktail), un bruit fluctuant (Fluctuating Noise, FN) et un bruit stationnaire (Broadband Noise, BBN).

Le bruit de Cocktail a été construit en mélangeant 4 voix (2 voix féminines, 2 masculines) ; chaque locuteur a été enregistré dans une pièce insonorisée, lisant des extraits de journaux français. Chaque extrait individuel a été traité selon le protocole suivant : 1) suppression des silences de plus de 1 seconde ; 2) suppression des phrases contenant des erreurs de prononciation, une prosodie exagérée ou des noms propres ; 3) réduction du bruit optimisée pour les signaux de parole ; 4) normalisation à 80dBA ; 5) mixage de chaque source afin d'obtenir un cocktail à 4 voix.

Le bruit fluctuant (FN) contient les mêmes caractéristiques spectro-temporelles que le bruit Cocktail. Pour ceci, nous avons extrait de notre bruit Cocktail l'enveloppe temporelle en dessous de 60Hz, puis nous avons calculé l'énergie spectrale du signal d'origine et en avons extrait la distribution des phases (Transformée de Fourier). Les phases ont ensuite été redistribuées de façon aléatoire, puis réinjectées dans l'enveloppe temporelle du signal d'origine. Enfin, l'énergie globale du signal obtenu a été ajustée à celle du signal original. Le bruit résultant posséde donc la même énergie spectrale et la même enveloppe temporelle que le signal original, mais sans informations linguistiques. Le bruit stationnaire (BBN) a été généré de la même manière, mais en éliminant l'enveloppe temporelle du signal d'origine. Il contient donc uniquement l'information spectrale du signal d'origine.

Stimuli et listes de mots

Les stimuli ont été générés en mixant chaque mot avec une séquence de 4 secondes de bruit, le mot étant toujours inséré à 2,5 secondes.

3 configurations de présentation ont été testées : 1) une configuration dichotique (S_N), avec le mot cible présenté dans une oreille et le bruit présenté dans l'oreille controlatérale ; 2) une configuration monaurale (SN_Si), avec le bruit et le mot présentés dans l'oreille cible ; 3) une configuration hybride monaurale/dichotique (SN_N) avec le mot diffusé dans l'oreille cible et le bruit diffusé dans les deux oreilles, cette dernière configuration donnant lieu à un effet de démasquage binaural en comparaison avec la condition monaurale. Pour chaque configuration, les 3 bruits ont été testés, ce qui donne 9 conditions de stimulation, avec 14 mots par condition.

Table 1. Conditions utilisées dans cette expérience

Condition	Oreille cible	Oreille controlatérale	Bruit
S_N	Mot	Bruit	BBN / FN / Cocktail
SN_Si	Mot / Bruit	Silence	BBN / FN / Cocktail
SN_N	Mot / Bruit	Bruit	BBN / FN / Cocktail

Les stimuli ont été présentés à une intensité de 65dB, avec un rapport signal/bruit de 0 dB dans l'oreille ispilatérale et une intensité dans l'oreille contralatérale de 20dB inférieure à l'oreille ipsilatérale.

9 listes de stimuli ont été créées, afin qu'à travers les listes, chaque mot soit présenté dans chaque bruit et avec chaque configuration de présentation. Dans chaque liste, les fréquences d'occurrence des mots sont contrebalancées entre les différentes conditions. Chaque sujet écoute deux listes de 63 mots, l'une étant présentée dans l'oreille gauche, l'autre dans l'oreille droite. A travers les sujets, l'ordre de présentation des deux listes est contrebalancé.

3. RESULTATS

Les pourcentages de mots correctement restitués (taux d'identification) ont été utilisés comme variable dépendante dans une ANOVA à mesures répétées, avec l'oreille de présentation du mot (Oreille), le bruit (Bruit) et la configuration de présentation (Configuration) comme facteurs intra-sujets.

Cette analyse a révélé un effet significatif de la Configuration (F(2,28)=110.95, p<.001, Figure 1). Selon les comparaisons planifiées, nous obtenons des performances meilleures pour la condition S_N que pour la condition SN_Si (F(1,14)=254.867, p<.001). L'ajout d'un bruit controlatéral améliore les performances dans la condition SN_N par rapport à la condition SN_Si (F(1,14)=17.05, p<.005). Les performances pour les conditions S_N et SN_N sont également différentes (F(1,14)=181.167, p< .001).

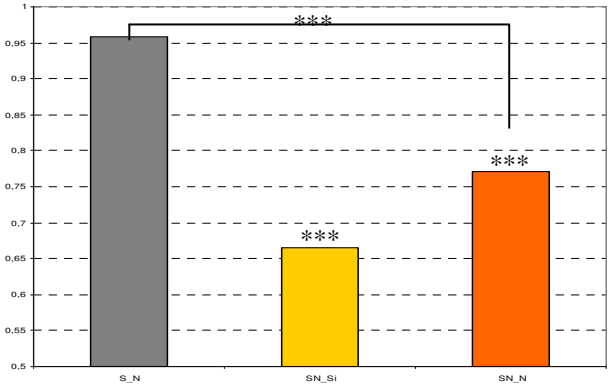


Figure 1 : Taux d'intelligibilité en fonction de la configuration de présentation.

L'ANOVA n'a pas révélé d'effets principaux de l'Oreille (F<1, n.s.) ni du Bruit (F (2,28) = 2.094, n.s.). En revanche, l'interaction Configuration * Bruit (Figure 2) est significative (F(4,56 =4.953, p<.005), montrant que dans les conditions S_N et SN_N les bruits ne diffèrent pas entre eux alors que dans la condition SN_Si le bruit Cocktail crée plus de difficultés de compréhension que le BBN (F(1,14)=5.676, p<.05) et le FN (F(1,14)=14.411, p<.005). L'effet de démasquage binaural est significatif lorsque le bruit utilisé est du Cocktail, avec de meilleures performances en configuration SN_N par rapport à la configuration SN_Si (F(1,14)=24.807, p<.001) ; cet effet n'est pas présent pour les bruits BBN (F(1,14)=1.887, n.s.) et FN (F<1, n.s.).

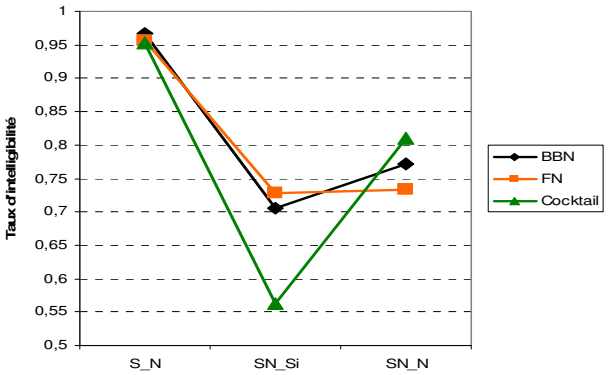


Figure 2 : Taux d'identification correcte en fonction du bruit et de la configuration de présentation

L'amplitude du démasquage binaural a été évaluée en soustrayant les performances obtenues dans la condition SN_Si de celles obtenues dans la condition SN_N (Figure 3). Une ANOVA à 1 facteur a été effectuée, avec le Bruit comme facteur intra-sujets.

Nous avons ainsi observé un effet significatif de ce facteur (F(2,28)=6.839, p<.005). Les comparaisons planifiées ont révélé un démasquage pour le bruit Cocktail significativement plus important que celui obtenu pour le bruit BBN (F(1,14)=5.44, p<.05) et pour le bruit FN (F(1,14)=14.484, p<.005), tandis que l'effet de démasquage binaural obtenu pour les bruit BBN et FN n'est pas significativement différent (F<1, n.s.).

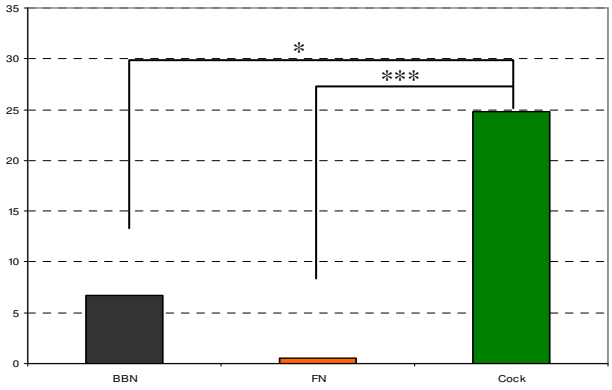


Figure 3 : Amplitude du démasquage binaural en fonction du type de bruit.

4. DISCUSSION

Nous avons cherché ici à étudier l'influence du type de bruit sur le démasquage binaural chez les dyslexiques adultes. Nos résultats ont montré qu'en condition dichotique, les sujets obtiennent 96% de bonnes réponses. En effet, dans cette condition, la séparation entre les deux sources est maximale, et les sujets n'ont donc aucune difficulté à les discriminer. En condition monaurale, nous avons observé un effet de masquage important du mot cible par le bruit, avec des performances chutant à 66%. En effet, dans cette condition, la cible et le bruit concurrent sont situés du même côté, et les sujets ne disposent d'aucun indice spatial pour les aider à discriminer les deux flux. Cet effet est particulièrement important dans le cadre d'un masquage informationnel de haut niveau représenté par la condition Cocktail (56% de réponses correctes). Nous pouvons donc en déduire que les informations linguistiques présentes dans le bruit Cocktail, comme les indices phonologiques et lexicaux, créent davantage d'interférences que des indices de bas niveau comme l'information spectrale (BBN) ou temporelle (FN).

Nous avons également obtenu un effet de démasquage binaural, avec des performances globales s'améliorant dans la condition SN_N (77% de réponses correctes). Les comparaisons planifiées nous ont cependant montré que cet effet de démasquage binaural n'est présent que dans le cas d'un masquage informationnel de haut niveau. Nos sujets semblent ne montrer aucun effet de démasquage pour les bruits BBN et SN. Ceci semble suggérer que nos sujets dyslexiques utilisent les indices binauraux pour compenser la présence du bruit environnant, mais seulement lorsque ces indices sont de nature linguistique.

Lors d'une expérience précédente [14], nous avions réalisé la même expérience avec 76 sujets normo-lecteurs

(NL). La différence résidait dans la tâche demandée aux participants, qui devaient noter eux-mêmes leurs réponses sur une feuille de papier. Bien que la constitution d'un groupe contrôle effectuant une tâche similaire soit nécessaire, nous avons comparé les performances de nos deux groupes à l'aide d'une ANOVA à mesures répétées, avec la Population comme facteur inter-sujets, et le Bruit et la Configuration comme facteurs intra-sujets. Cette analyse révèle un effet significatif de la population : d'un point de vue global, les dyslexiques ont des performances inférieures à celles des NL ($F(1,91)=5.834$, $p<.05$). Nous n'obtenons aucune différence entre les deux groupes en configuration monaurale, quel que soit le bruit utilisé. Par contre nous observons des performances inférieures chez les dyslexiques dans la condition SN_N pour le bruit BBN ($F(1,91)=7.489$, $p<.01$), et cette différence est proche du seuil de significativité pour le bruit FN ($F(1,91)=2.886$, $p=0.093$). Ces résultats semblent suggérer chez le sujet dyslexique l'existence d'un déficit du démasquage binaural lorsque la parole est masquée par un bruit physique, et non lorsque le bruit masquant est constitué de paroles.

Ce déficit pourrait en partie expliquer les troubles de la compréhension de la parole dans le bruit, bien que d'autres études soient nécessaires pour expliquer l'absence de déficit du démasquage binaural en situation de parole concurrente. Il est possible que ce résultat puisse s'expliquer par un effet compensatoire, notre groupe de sujets dyslexiques étant principalement constitué d'étudiants de niveau universitaire.

5. CONCLUSION

Cette étude avait pour but d'étudier les troubles de la compréhension de la parole dans le bruit chez l'adulte dyslexique. Nos résultats ont permis de mettre en avant un déficit de l'utilisation des indices binauraux lorsque la parole est masquée par un bruit physique. Nous n'avons cependant pas montré de tel déficit pour un masquage informationnel de haut niveau.

6. REMERCIEMENTS

Nous remercions la Région Rhône-Alpes (cluster Handicap, Vieillissement, Neurosciences) pour son Allocation Doctorale de Recherche, ainsi que l'European Research Council pour son financement attribué à Fanny Meunier (projet SpiN).

BIBLIOGRAPHIE

[1] W. Serniclaes, S. Van Heghe, P. Mousty, R. Carré and L. Sprenger-Charolles. Allophonic mode of speech perception in dyslexia. *Journal of Experimental Child Psychology*, 87:336-361, 2004.

[2] F.R. Manis, C. McBride-Chang, M.S. Seidenberg, P. Keating and al. Are speech perception deficits associated with developmental dyslexia? *Journal of Experimental Child Psychology,* 66: 211-235, 1997.

[3] P.L. Cornellissen, P.C. Hansen, L. Bradley and J.F. Stein. Analysis of perceptual confusions between nine sets of consonant vowel sounds in normal and dyslexic adults. *Cognition*, 59: 275-306, 1996.

[4] A.R. Bradlow, N. Kraus and E. Hayes. Speaking clearly for children with learning disabilities: sentence perception in noise. *Journal of Speech Language and Hearing Research*, 46:80-97, 2003.

[5] J.C. Ziegler, C. Pech-Georgel, F. George and C. Lorenzi. Speech-perception-in-noise deficits in dyslexia. *Developmental Science*, 12:732-745, 2009.

[6] B. Wible, T. Nicol and N. Kraus. Abnormal neural encoding of repeated speech stimuli in noise in children with learning problems. *Clinical Neurophysiology*, 113: 485-494, 2002.

[7] M. Habib. The neurological basis of dyslexia. An overview and working hypothesis. *Brain*, 123: 2373-2399, 2000.

[8] S. Heim, C. Eulitz and T. Elbert. Altered hemispheric asymmetry of auditory N100m in adults with developmental dyslexia. *Neuroreport*, 14: 501-504, 2003.

[9] K. Giraud, A. Trébuchon-DaFonseca, J.F. Démonet, M. Habib and C. Liégeois-Chauvel. Asymmetry of voice onset time-processing in adult developmental dyslexics. In *Clinical Neurophysiology*, 119: 1652-1663, 2008.

[10] K. Hugdahl, E. Heiervang, L. Ersland, A. Lundervold, H. Steinmetz and A.I. Smievoll. Significant relation between MR measures of planum temporale area and dichotic processing of syllables in dyslexic children. *Neuropsychologia*, 41: 666-675, 2003.

[11] F. Robichon, O. Levrier, P. Farnarier and M. Habib. Developmental dyslexia: atypical cortical asymmetries and functional significance. *European Journal of Neurology*, 7: 35-46, 2000.

[12] A.L. Giraud, S. Garnier, C. Micheyl, G. Lina, A. Chays and S. Chéry-Croze. Auditory efferents involoved in speech-in-noise intelligibility. *Neuroreport*, 8: 1779-1783, 1997.

[13] E. Veuillet, A. Magnan, J. Ecalle, H. Thai-Van and L. Collet. Auditory processing disorder in children with reading disabilities: effect of audiovisual training. *Brain*, 130, 2915-2928, 2007.

[14] M. Dole, M. Hoen and F. Meunier. Effect of contralateral noise on energetic and informational masking on speech-in-speech intelligibility. In *Proc.Interspeech*, volume 10, pages 136-139, 2009.

[15] B. New, C. Pallier, M. Brysbaert and L. Ferrand. Lexique 2: a new French lexical database, *Behavior Research Methods, Instruments & Computer*, 36: 516-524, 2004.

Production de l'enchaînement et de la liaison enchaînée en français: données psycholinguistiques et acoustiques

Odile Bagou et Marina Laganaro

Groupe NPL, NeuroPsychoLinguistique, FLSH, Université de Neuchâtel, Suisse

Odile.Bagou@unine.ch, Marina.Laganaro@unine.ch

ABSTRACT

This study investigates how sequences involving sandhi phenomena are produced by French speakers. Using a psycholinguistic production paradigm, we first investigated whether the encoding of *enchaînement* and *liaison enchaînée* involves a processing cost. Second, we examined how sequences were produced by measuring the durational properties of the critical V_1CV_2 sequences. Psycholinguistic results indicate that the encoding of sequences involving potential resyllabification engages an additional processing time compared to word-initial consonants. Durational data revealed that word-initial consonants are lengthened compared to their resyllabified counter-parts. Taken together, these data suggest that word-initial and resyllabified consonants are not encoded in the same way.

Index Terms: speech production, psycholinguistic, phonology, sandhi.

1. INTRODUCTION

La production d'une séquence de deux mots consécutifs peut engendrer des phénomènes de sandhi externes lorsque le premier mot (M1) se termine par une consonne et que le deuxième (M2) commence par une voyelle. Deux principaux phénomènes de sandhi ont été décrits en français. Dans l'enchaînement, la consonne sous-jacente finale de M1 est (re)-syllabée en position initiale de M2 dans la forme de surface (e.g. "sept" /sɛt/ + "amis" /ami/ → /sɛ.t#a.mi/[1]). Dans le cadre de la phonologie non linéaire auto-segmentale [1], ces consonnes sont dites fixes puisqu'elles sont ancrées au squelette syllabique du 1er mot de la séquence, mais leur position syllabique est flottante puisqu'elles perdent leur statut de coda pour devenir attaque de la syllabe suivante. Dans la liaison enchaînée, une consonne latente apparait à la frontière des deux mots dans la forme de surface seulement lorsque M2 commence par une voyelle (e.g. "mon" /mɔ̃/ + "ami" /ami/ → /mɔ̃.n#a.mi/[1]; mais /mɔ̃/+ /kopɛ̃/→ /mɔ̃.#ko.pɛ̃/). L'apparition de ces consonnes de liaison est donc régie par le contexte subséquent. Selon la phonologie non linéaire, ces consonnes n'ont pas d'ancrage dans le squelette syllabique contrairement aux consonnes d'enchaînement, ce qui leur confère un statut de consonne flottante par essence. Par ailleurs, comme les consonnes d'enchaînement, la production de ces consonnes modifie la structure syllabique de la forme de surface. Ainsi, on peut parler de flottement simple dans le cas de l'enchaînement et de flottement double dans le cas de la liaison enchaînée.

D'un point de vue psycholinguistique, la pertinence cognitive des statuts de flottements simple et double doit être établie. Ainsi, il s'agit de savoir si la réorganisation syllabique induite par les phénomènes de sandhi externes influence le traitement des séquences de mots. A notre connaissance, les études antérieures ont adressé la question du décodage de telles séquences [2,3], mais la question de l'encodage n'a jamais été abordée avec des paradigmes psycholinguistiques de production de la parole.

Le principal objectif des travaux menés antérieurement était de savoir si la réorganisation de la chaîne de parole découlant de tels phénomènes affectait l'identification du deuxième mot. En effet, lorsqu'il y a re-syllabation, les frontières syllabiques de la forme de surface ne sont pas alignées avec les frontières lexicales sous-jacentes. Si, comme le suggèrent les modèles d'identification lexicale, les syllabes [4] ou les débuts de syllabe [5] déclenchent l'accès au lexique mental, il est raisonnable de penser que ce non alignement retarde ou entrave la reconnaissance de M_2. Toutefois, les résultats indiquent que l'identification de M_2 est facilitée [2,3]. Les consonnes d'enchaînement et de liaison ne seraient donc pas traitées comme des consonnes initiales, ce qui confirmerait la pertinence cognitive de leur statut phonologique de consonnes flottantes.

Les études perceptives tendent également à valider la pertinence cognitive du flottement double, i.e. du statut phonologique des consonnes de liaison, mais cette conclusion ne serait pas généralisable aux consonnes d'enchaînement. Les résultats aux tâches de détection de phonèmes indiquent que les consonnes de liaison sont plus difficiles à détecter que les consonnes initiales et d'enchaînement [6]. Si certains auteurs ont vu en ces résultats la preuve de la pertinence du statut phonologique spécifique des consonnes de liaison, il est néanmoins possible que les consonnes de liaison soient structurellement similaires mais phonétiquement différentes des consonnes initiales comme le suggère la phonologie articulatoire [7].

Toutefois, les études acoustiques montrent que, quel que soit le phénomène de sandhi impliqué, les

[1] « . » : frontière syllabique « # »: frontière lexicale

consonnes resyllabées [e.g. 2, 3, 8, 9, 10, 11] sont moins saillantes que les consonnes initiales de mot. Les consonnes d'enchaînement devraient donc être également plus difficiles à détecter que les consonnes initiales, ce qui n'est pas le cas. En résumé, les études phonétiques suggèrent que la moindre saillance acoustique des consonnes re-syllabées pourrait expliquer la facilité d'identification de M_2 observée pour les deux phénomènes de sandhi. Cependant, elles n'expliquent pas pourquoi les consonnes d'enchaînement ne se distinguent pas perceptivement des consonnes initiales. Par ailleurs, la phonologie permet d'expliquer les résultats perceptifs et la facilité d'identification lexicale de M_2 dans la liaison, mais ne suffit pas à rendre compte de la facilité d'accès à M_2 dans les séquences enchaînées.

Les phénomènes de sandhi externes défient également les modèles psycholinguistiques de production de la parole. Selon le modèle le plus abouti actuellement [12], la structure syllabique ne serait pas représentée dans le lexique. L'accès à la forme phonologique des mots consisterait donc à encoder sa composition segmentale et la syllabation aurait lieu subséquemment sur le mot phonologique, avant l'adressage des gestes articulatoires [13, 14]. La chaîne de parole étant syllabée sur la base d'une séquence phonémique dont les positions syllabiques ne sont pas préalablement définies, les séquences donnant lieu à des phénomènes de sandhi externes seraient syllabées lors de l'encodage du mot phonologique comme le sont les autres segments. Ces séquences ne sont donc pas considérées comme « resyllabées » [12]. De plus, deux syllabes identiques au niveau segmental ne devraient pas se distinguer phonétiquement même si le contexte de production diffère. Ainsi, la syllabe /ta/ serait identique dans « sept amis » (/sɛ.ta.mi/) et dans « détaler » (/de.ta.le/), une prédiction qui va à l'encontre des résultats phonétiques mentionnés précédemment.

L'objectif de cette étude vise à répondre aux deux questions suivantes : (1) l'encodage phonologique de phénomènes de sandhi induit-il un coût de production comparé à l'encodage de séquences syllabiquement alignées ? (2) les consonnes resyllabées se distinguent-elles phonétiquement des initiales de mot ? Dans une étude préliminaire [14], nous avons montré que la production de séquences induisant une liaison enchaînée générait un coût d'encodage comparé aux deux autres conditions de frontière (alignement syllabique et enchaînement). Par ailleurs, les mesures acoustiques conduites sur les séquences critiques V_1CV_2 de 9 locuteurs indiquaient que la durée des segments critiques différait selon la condition de frontière, notamment pour les consonnes occlusives sourdes /t/ qui étaient significativement allongées à l'initiale de mot [14]. Toutefois, les différences observées entre enchaînement et liaison pourraient être liées aux propriétés segmentales de M_1 plutôt qu'aux phénomènes phonologiques. En effet, les consonnes critiques étaient strictement identiques d'une condition à l'autre, mais ni le premier segment de M_1 ni sa voyelle finale (V_1) n'avaient été contrôlés systématiquement. La présente étude vise donc à répliquer les résultats précédents en évitant ces biais potentiels.

2. EXPÉRIENCE

2.1. Méthode

Participants

17 étudiants de l'université de Neuchâtel ont participé à l'expérience.

Matériel

14 noms (M_2) à initiale consonantique (contrôles, tigrons) et 14 débutant à initiale vocalique (tests, igloo) appariés en fréquence lexicale et comparables quant à leurs propriétés phonologiques (longueur, segment initial et structure syllabique) ont été sélectionnés dans la base de données Lexique. 6 adjectifs (M_1) leur ont été associés afin de construire des séquences de deux mots (adjectif+nom) selon 3 conditions de frontière (Table 1). Pour l'enchaînement et la liaison enchaînée, les adjectifs induisaient une re-syllabation de la chaîne de parole dans les contextes V-initial. Pour éviter tout sandhi externe dans les contextes V-initial, deux adjectifs à finale vocalique ont été sélectionnés dans la condition d'alignement syllabique. 144 séquences de même nature étaient utilisées comme distracteurs pour éviter un biais éventuel dû à la répétition fréquente des adjectifs.

Table 1 : Conditions de frontière selon le statut du phonème final de M_1

		Noms	
		Contrôles-C-ini	***Tests-V-ini***
Statut du phonème final de M1	***C-Enchaînement*** *(dix-sept / trente)*	trente tigrons /tʀɑ̃t.#ti.gʀɔ̃/	**trente igloos** /tʀɑ̃.t#i.glu/
	C-Liaison enchaînée *(grand / parfait)*	**grand tigron** /gʀɑ.#ti.gʀɔ̃/	**grand igloo** /gʀɑ̃.t#i.glu/
	V-Alignem. Syll. *(demi / sacré)*	demi tigron /də.mi.#ti.gʀɔ̃/	demi igloo /də.mi.#i.glu/

Les séquences en gras ont été utilisées dans l'analyse acoustique

Procédure

Les participants étaient testés individuellement dans une chambre sourde et devaient lire les séquences le plus rapidement possible. Celles-ci étaient présentées au milieu de l'écran pendant 1500 ms en ordre semi-aléatoire. Les latences de production (TR) étaient mesurées entre l'apparition de la cible à l'écran et le début de la production orale mesurée par une clé vocale. Les productions étaient enregistrées pour une vérification ultérieure et pour les mesures acoustiques subséquentes. Les TR ont été recueillis sur les adjectifs isolés suivant les mêmes procédures.

2.2. Résultats

Les TR recueillis sur les adjectifs isolés ne diffèrent pas significativement d'une condition à une autre (F_1(2,32)=1.16, *MSE* = 1114; p=.3; F_2(2,3)=.4, *MSE*= 373; p=.7; *minF'*<1). Toutefois, les analyses ont été menées sur des différences de TR calculées en soustrayant la latence de production de chaque adjectif isolé à la latence de production de la séquence totale. Ainsi, nous évitons que des différences, même minimes, influencent les résultats.

Latences de production (Figure 1)

Les TR compris entre 250 ms et 1000 ms ont été inclus dans les analyses (96 % des données).

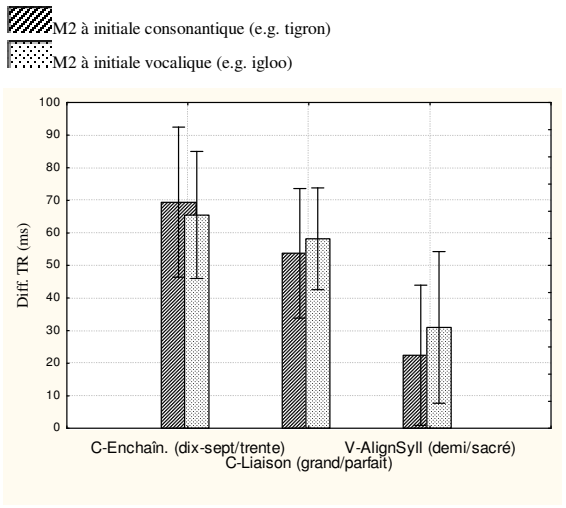


Figure 1 : Latences de production (TR) pour les 3 conditions selon le statut du phonème initial de M_2

Des ANOVAs à mesures répétées ont été menées avec le statut du phonème final de M_1 comme facteur intra-participants (F_1) et intra-items (F_2) et la nature de M_2 (Contrôle-Cini vs. Test-Vini) comme facteur intra-participants et inter-items. Seul un effet significatif du phonème final de M_1 est révélé (F_1(2,32)=7.1, *MSE*=2101.8, p=.003; F_2(2,52)=48.3, *MSE*=252.3, p<.00001; *minF'*(4,40)= 6.19, p<.01). Des comparaisons planifiées (Fisher) indiquent que les séquences syllabiquement alignées sont plus rapidement produites que les séquences impliquant une consonne d'enchaînement (p=.0009) ou de liaison (p=.01). En revanche, aucune différence n'est observable entre l'enchaînement et la liaison (p=.3).

Analyses acoustiques (Figure 2)

Des analyses de durée ont été conduites sur les productions des 17 locuteurs. Les séquences critiques (en gras dans le tableau 1) V_1CV_2 ont été extraites et la durée de chaque segment a été mesurée sous Praat.

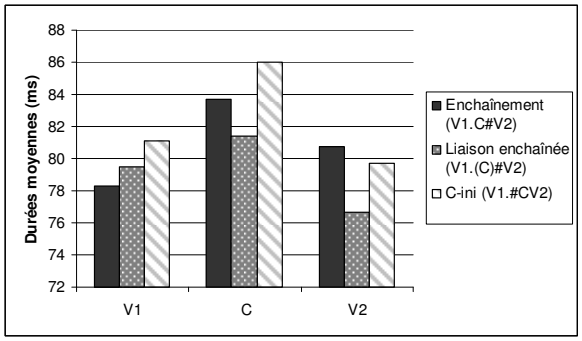


Figure 2 : Durées moyennes de V1, C et V2 selon la condition de frontière

Des ANOVAs à mesures répétées et des comparaisons multiples (Test de Fisher) ont été menées avec la condition de frontière comme facteur intra-participant et intra-items. Elles révèlent que les caractéristiques acoustiques de la consonne C (F_1(2,32)=5.44; *MSE* = .000013; p=.009; F_2(2,26)=13.5, *MSE* = .00001, p=.00009; *minF'*(4,58)=3.877, p < .01) et de la voyelle subséquente V2 (F_1(2,32)=6.3, *MSE* = .00001; p=.004; F_2(2,26)=5.7, *MSE* = .00002, p=.008; *minF'*(4,47)=3, p < .025) dépendent de la condition de frontière. Plus précisément, C est significativement plus longue en position initiale de mot qu'en consonne de liaison (p=.002) et marginalement plus longue qu'en consonne d'enchaînement (p=.08). De plus, V_2 est significativement plus courte dans le cas de la liaison que dans les deux autres conditions (Liaison enchaînée vs. C-ini: p=.01; Liaison vs. Enchaînement: p=.002). Pour résumer, la syllabe CV est significativement plus courte dans la liaison enchaînée que dans les deux autres conditions de frontière.

3. DISCUSSION

Le premier objectif de ce papier était de savoir si l'encodage de séquences induisant des phénomènes de sandhi externes génère un coût de traitement comparé à l'encodage de séquences syllabiquement alignées. Le second objectif était de savoir si les séquences critiques V_1CV_2 différent phonétiquement d'une condition de frontière à l'autre. L'analyse des latences de production (TR) confirme que l'encodage de séquences impliquant une consonne d'enchaînement ou de liaison engendre un coût de traitement comparé à l'encodage de séquences alignées. Cependant, puisque des TR plus longs ont également été observés lorsque M_2 débute par une consonne, le contexte phonétique ne module pas l'effet de la condition de frontière. Puisque l'effet de la condition de frontière n'est pas significatif lorsque l'adjectif est produit en isolation, ce coût de production résulte bien de l'encodage de l'adjectif précédant un M_2 dans une séquence de mots induisant potentiellement un phénomène de sandhi externe.

Concernant la liaison enchaînée, ces résultats suggèrent que l'encodage mais aussi le non-encodage des segments flottants induisent un coût de traitement, ce qui validerait la pertinence cognitive du statut de double flottement attribué aux consonnes de liaison [1]. A la lumière des modèles psycholinguistiques de

production, deux alternatives pourraient expliquer ce coût cognitif. D'une part, si l'on considère que deux formes phonologiques (avec/sans consonne de liaison) sont stockées dans le lexique mental, le coût serait du à la recherche de la forme appropriée au contexte phonologique. D'autre part, si l'on postule l'existence d'une seule forme phonologique stockée avec une unité cachée, le coût serait dû à l'activation/non activation de ce segment additionnel. Nos données ne permettent pas de départager ces deux hypothèses alternatives et des travaux ultérieurs sont donc nécessaires.

Concernant l'enchaînement, des TR plus longs ont également été observés quel que soit le contexte. Puisque, dans ce cas, il n'y a pas d'information segmentale supplémentaire à encoder, ce résultat plaide en faveur d'un coût de re-syllabation dans les contextes à initiale vocalique (/tʀɑ̃tiglu/). Dans les contextes à initiale consonantique, en revanche, le coût supplémentaire observé pourrait être dû à l'encodage du segment supplémentaire présent dans ces séquences (/tʀɑ̃ttigʀɔ̃/).

Les analyses acoustiques infirment également les prédictions faites par les modèles s'opposant à l'existence d'un processus de re-syllabation [12]. En effet, les contrastes entre des séquences identiques d'un point de vue segmental mais différant quant à leur condition de frontière sous-jacente ne sont pas neutralisés. En effet, ni la consonne sous-jacente finale de M_1 dans l'enchaînement ni la consonne latente ajoutée dans la liaison enchaînée n'ont adopté les caractéristiques acoustiques d'une consonne initiale dans la forme de surface. De plus, les caractéristiques de la voyelle subséquente V_2 sont aussi modulées par la condition de frontière. Ainsi, les gestes articulatoires diffèrent selon qu'il s'agit de produire une syllabe CV resyllabée ou initiale de mot.

En conclusion, ces données psycholinguistiques et phonétiques suggèrent que les syllabes sont générées à une étape précoce de l'encodage et que les consonnes de liaison et d'enchaînement sont ancrées au 1er mot de la séquence. Même si les traces de l'appartenance syllabique sous-jacente de ces consonnes résistent à la re-syllabation, ces données confirment néanmoins l'existence d'un processus de resyllabation.

4. REMERCIEMENTS

Cette recherche a été financée par le FNRS (no. PP01-118969). Nous remercions également Violaine Michel pour la passation des expériences et la correction manuelle des latences de production.

BIBLIOGRAPHIE

[1] Encrevé, P. *La liaison avec et sans enchaînement*. Seuil, Paris, 1988.

[2] Gaskell, M., Spinelli, E., and Meunier, F. Perception of resyllabification in French. *Memory and Cognition*, 30, 798-810, 2002.

[3] Spinelli, E., McQueen, J., and Cutler, A. Processing resyllabified words in French. *Journal of Memory and Language*, 48, 233-254, 2003.

[4] Mehler, J., Dommergues, J.Y., Frauenfelder, U.H. and Segui, J. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20, 298-305, 1981.

[5] Content, A., Kearns, R., and Frauenfelder, U. Boundaries versus onsets in syllabic segmentation. *Journal of Memory and Language*, 45, 177-199, 2001.

[6] Nguyen, N., Wauquier-Gravelines, S., Lancia, L., & Tuller, B. Detection of liaison consonants in speech processing in French: Experimental data and theoretical implications. In P. Prieto, J. Mascaró & M.J. Solé [Eds], *Segmental and Prosodic Issues in Romance Phonology*, 3-23, John Benjamins, Amsterdam, 2007.

[7] Browman, C.P. and Goldstein, L. Towards an articulatory phonology. In C. Ewen and J. Anderson [eds.], *Phonology Yearbook 3,* Cambridge: Cambridge University Press, pp. 219-252, 1986.

[8] Delattre, P. Le mot est-il une entité phonétique en français? *Le Français Moderne*, 8, 47-56, 1940.

[9] Dejean de la Bâtie, B. Word boundary ambiguity in spoken French, PhD thesis, Monash University, Victoria, Australia, 1993.

[10] Fougeron, C., Bagou, O., Content, A., Stefanuto, M. and Frauenfelder, U.H. Looking for acoustic cues of resyllabification in French. In Solé, M.J., Recasens D. and Romero, J. [Eds], *Proceedings of the 15th International Congress of Phonetic Sciences*, 2257-2260, 2003.

[11] Fougeron, C. Word boundaries and contrast neutralization in the case of enchaînement in French. In Cole, J. & Hualde J.I. [Eds], *Papers in Laboratory Phonology IX: Change in Phonology, Berlin: Mouton de Gruyter,* 609-642, 2007.

[12] Levelt, W. J.M., Roelofs, A. and Meyer, A.S. A theory of lexical access in speech production. *Brain and Behavioral Sciences,* 22 (1), 1-38, 1999.

[13] Laganaro M. and Alario, X. On the locus of syllable frequency effect. *Journal of Memory and Language*, 55, 178-196, 2006.

[14] Bagou, O., Michel, V. and Laganaro, M. On the production of sandhi phenomena in French: psycholinguistic and acoustic data. *In Proceedings of interspeech*, 2009.

C-PROM :
Un corpus de français parlé annoté pour l'étude des proéminences

M. Avanzi[1, 2], *A.C. Simon*[3], *J.-P. Goldman*[3, 4] *& A. Auchlin*[4]

[1]Chaire de linguistique française, Université de Neuchâtel; [2]MoDyCo, Université Paris Ouest Nanterre ;
[3]Institut Langage & Communication / Valibel Discours & Variation, Université catholique de Louvain;
[4]Département de linguistique, Université de Genève
mathieu.avanzi@unine.ch, anne-catherine.simon@uclouvain.be,
jeanphilippegoldman@gmail.com, antoine.auchlin@unige.ch

ABSTRACT

This paper presents C-PROM, an annotated corpus for studying prominence in French. The corpus includes 3 regional varieties of French (Belgian, Swiss and metropolitan French) and 7 speaking styles (from oral reading to spontaneous conversations); it has a total duration of 70 minutes and was annotated by two phoneticians. The two experts followed a strict annotation protocol, taking into account the previous mistakes encountered by prior research on prominence detection in French and elements of the methodology followed by scholars working on other languages. We conclude by discussing the average consistency between both annotators. The results obtained are quite encouraging, since the F-measure between the two annotators reaches 82.8%, and the kappa-score 0.77.
Keywords: corpus, spontaneous French, prominence, discourse-genre.

1. TRAVAUX ANTÉRIEURS

Les premiers travaux sur la détection de proéminences en français ont vu le jour dans le cadre du projet Phonologie du Français Contemporain (PFC) [1]. Dans le cadre de ce projet, les principes posés pour le protocole d'annotation de la prosodie étaient à l'origine les suivants [2]. La procédure de codage devait (i) être réalisée en dehors de tout cadre théorique spécifique, (ii) reposer sur des bases impressionnistes uniquement, (iii) être reproductible par des annotateurs non experts, (iv) permettre des études transversales à tous les domaines de la prosodie (accentuation, intonation, rythme, etc.).

Dans l'optique de mettre au point une première esquisse de protocole, une expérience pilote a été conduite par [3]. L'auteur a demandé à sept experts phonologues d'annoter les proéminences syllabiques dans un extrait de parole spontanée d'une durée de 3 minutes, sans aucune autre instruction. L'auteur s'attendait à un consensus assez élevé, partant de l'idée que les proéminences correspondaient à des syllabes accentuées, et que les règles de l'accentuation du français étaient bien connues des participants. De façon surprenante, il est apparu que le taux d'accord inter-juges n'était pas aussi bon que ce qui était attendu, puisque sur les 165 syllabes susceptibles de porter un accent primaire dans le corpus, la proportion de syllabes annotées comme proéminentes variait de 19% à 49%. Ces résultats, guère encourageants, ont confirmé les points de vue les plus pessimistes, à savoir que l'annotation des proéminences « s'apparentait plus à un art qu'à une pratique scientifique » [4].

Reprenant ces données, [5] ont cherché à évaluer le caractère robuste de deux paramètres acoustiques (la F0 et la durée) pour une détection automatique des proéminences, dans l'optique de suggérer des mesures objectives pour évaluer les performances des annotateurs. De cette seconde expérience, les auteurs ont conclu que les variations de F0 étaient corrélées avec le taux d'accord inter-juge (plus les variations relatives de F0 sont importantes, plus l'accord entre les codeurs est grand). Par contre ce n'est pas le cas avec les mesures relatives de durée : en effet, ils ont observé qu'au-delà un certain seuil de durée (approximativement entre 175 et 200 ms), la proportion d'accord inter-juge s'inverse et le taux décroît. Cela est dû au fait qu'au-delà d'une certaine durée syllabique, les allongements ne sont plus perçus comme des marqueurs de proéminence mais comme des signaux d'hésitation. Cela signifie en outre que des humains, même experts, ne partagent pas la même définition de la notion de proéminence.

Plusieurs enseignements ont été tirés de ces études pilotes par les initiateurs du codage de la prosodie dans les corpus PFC. Premièrement, il est apparu que le faible taux d'accord provenait vraisemblablement du manque de clarté et de précision dans la consigne et que, par conséquent, si l'on désirait avoir un meilleur taux de consensus entre les annotateurs, il fallait que la notion de proéminence soit correctement définie et qu'elle ne soit pas confondue avec celle d'accent (qui est une notion phonologique impliquant un savoir linguistique spécifique). Ensuite, il est ressorti que le codage perceptif des proéminences devait être borné à un intervalle temporel défini, pour éviter que différentes sous-parties du corpus subissent un codage différent (plus long est l'extrait écouté, moins élevé est la proportion de syllabes perçues comme proéminentes, et inversement). Finalement, l'étude des corrélats acoustiques des proéminences a montré que si la F0 était un indice pertinent pour la détection automatique,

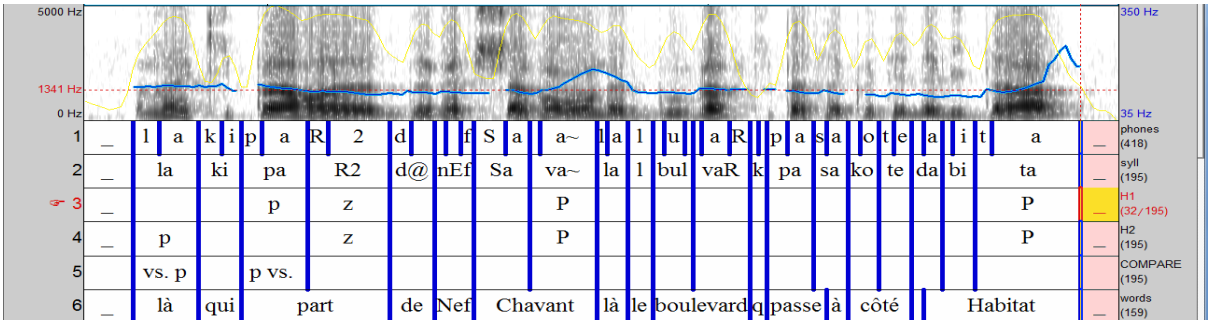


Figure 1 : Copie d'écran Praat, transcription de l'énoncé : « là qui part de Nef Chavant là le boulevard qui passé à côté d'Habitat » [iti-1]. Les couches d'annotations sont, du haut vers le bas, celle des phonèmes, celle des syllabes (toutes deux en SAMPA), celles des annotations manuelles H1 et H2 (tires « delivery »), la tire COMPARE et celle des mots graphiques.

les allongements liés à une hésitation devaient être balisés, d'une part pour ne pas être interprétés différemment selon les codeurs, et d'autre part pour ne pas interférer dans les mesures de durée relative dans une perspective d'automatisation de la détection.

2. DESCRIPTION DU CORPUS

Ces premiers travaux conduits par les participants du groupe Prosodie-PFC ont connu des prolongements dans diverses études menées par un consortium informel de linguistes français, suisses et belges et ont *in fine* abouti à l'élaboration d'un corpus multi-styles et multi-locuteurs annoté pour les proéminences. Ce corpus est présenté dans les sections suivantes. L'annotation perceptive a également permis de développer des algorithmes de détection automatique [6] qui ne sont pas décrits ici. Le corpus est composé d'enregistrements échantillonnés en sept genres avec, du plus formel au moins formel : Journaux radiophoniques (JPA) ; Lectures d'un texte (LEC), Discours politique (POL), Conférences (CNF), Interviews radiophoniques (INT), Descriptions d'itinéraires (ITI) et Récits de vie (NAR). Mis à part les genres ITI et INT, tous les genres de discours sont représentés par 3 enregistrements de 3 minutes environ, de locuteurs francophones natifs de Belgique (BE), de Suisse (CH) et de France métropolitaine (FR) (voir tableau 3 *infra*). Cet ensemble d'enregistrements, d'une durée totale de 70 minutes, a été automatiquement segmenté en phonèmes, syllabes et mots graphiques avec le script EasyAlign [7], implémenté sous Praat [8]. Les transcriptions ont été vérifiées manuellement, et corrigées le cas échéant.

3. ANNOTATION PERCEPTIVE DES PROÉMINENCES

3.1. Protocole

Deux annotateurs (H1 et H2, deux des auteurs de cet article) ont codé la totalité du corpus en suivant un protocole strict. Ce protocole d'annotation du corpus C-PROM tire parti des difficultés rencontrées lors des premières études menées dans le cadre de PFC, et tient compte de certaines des recommandations faites par les superviseurs du corpus de néerlandais parlé [9]. En pratique, chaque annotateur part d'une couche (« tire ») d'annotation vide, dupliquée à partir de la tire syllabe (tires H1 et H2, *cf.* figure 1), et annote les intervalles avec trois types de symboles. La première classe de symboles concerne la perception des proéminences à proprement parler. Deux symboles (p et P) peuvent être utilisés pour annoter les syllabes perçues comme faiblement ou fortement proéminentes. La distinction entre « p » et « P » a une valeur heuristique : elle force le codeur à développer une écoute plus fine et permet d'éviter un marqueur d'indécision du type « ? ». Si aucune proéminence n'est perçue, l'intervalle est laissé vide. Au cours de la phase de comparaison entre annotateurs, les deux symboles « p » et « P » sont confondus en une seule classe. La deuxième catégorie de symboles, caractérisant les marques de travail de formulation (tire « delivery »), est utilisée pour isoler des syllabes qui ont des propriétés spécifiques, susceptibles d'interférer avec la perception et la détection automatique des proéminences. On introduit ainsi dans l'annotation une distinction entre les proéminences supposément liées au système accentuel du français et des phénomènes liés au travail de formulation. Le symbole « z » est utilisé pour noter les syllabes allongées associées à une hésitation (leur longueur gênant les mesures de durée relative, comme l'ont montré [5]). Quant aux schwas post-toniques (@) et aux appendices ($), ils ont été identifiés spécifiquement pour se donner la possibilité d'étudier leurs statuts phonologiques spécifiques : les segments post-focaux sont décrits comme ayant un profil mélodique plat et peu modulé, en théorie; d'autre part, la capacité des schwas en position finale de mot à jouer le rôle de noyau accentuel est controversée. La troisième catégorie de symboles regroupe les « silences » et certains phénomènes para- ou non verbaux. Elle contient les pauses silencieuses repérées automatiquement, les prises de souffle audibles ainsi que les intervalles inexploitables, *i.e.* les segments qui n'ont pu être transcrits (bruits, rire, toux,

chevauchements), susceptibles d'interférer avec le traitement automatique.

3.2. Tâche d'annotation

Dans un premier temps, les deux annotateurs se sont entraînés sur un enregistrement d'indication d'itinéraire d'une minute environ (iti-5), puis ont annoté le reste du corpus chacun de son côté, genre de discours par genre de discours (l'annotation s'est déroulée sur plusieurs mois, de l'automne 2007 à l'automne 2008). En pratique, chaque annotateur écoute des segments de trois à cinq secondes, pas plus de trois fois (l'hypothèse étant qu'une sur-écoute entraînerait un surcodage). Comme l'annotation repose sur la perception de saillances syllabiques et non sur l'analyse visuelle des variations acoustiques (montées de F0 p. ex.), l'affichage du signal de parole est réservé aux cas retors (lors de la phase ultérieure de comparaison). Au fur et à mesure qu'un genre est codé, une tire de comparaison est automatiquement générée, afin de mettre en lumière les cas de désaccord entre les deux codeurs (voir tire COMPARE de la figure 1). Ces désaccords ont ensuite été discutés au cours de sessions communes ; on se reportera à [9] pour une analyse systématique des cas de désaccords, et les règles qui ont émergé pour les arbitrer.

3.3. Evaluation du taux d'accord inter-annotateurs

La tire COMPARE a été utilisée pour estimer le taux d'accord inter-annotateurs. Le Tableau 1 donne le détail de ce taux pour chaque enregistrement du corpus, sur la base du décompte des intervalles impliquant un conflit entre p, P, et un autre symbole.

Tableau 1: Taux d'accord inter-annotateurs : nom du fichier (voir Table 2) ; n00/n11 : syllabes annotées 0 (non proéminent) et P (proéminent) les deux annotateurs ; n10/n01 : syllabes annotées P par un transcripteur, 0 par l'autre; A : accord brut ; R : rappel ; P : précision ; F : F-mesure ; K : Kappa-score.

Fichiers	n00	n11	n10	n01	A	R	P	F	K
jpa-be	924	259	58	29	93	89,9	81,7	85,6	0,81
jpa-ch	587	172	29	50	91	77,5	85,6	81,3	0,75
jpa-fr	668	197	25	46	92	81,1	88,7	84,7	0,80
lec-be	495	87	22	5	96	94,6	79,8	86,6	0,84
lec-ch	390	156	26	33	90	82,5	85,7	84,1	0,77
lec-fr	469	120	32	2	95	98,4	78,9	87,6	0,84
pol-be	225	139	21	26	89	84,2	86,9	85,5	0,76
pol-ch	735	162	20	97	88	62,5	89	73,5	0,66
pol-fr	519	143	13	71	89	66,8	91,7	77,3	0,70
cnf-be	727	182	26	84	89	68,4	87,5	76,8	0,70
cnf-ch	586	186	40	59	89	75,9	82,3	79	0,71
cnf-fr	776	239	39	37	93	86,6	86	86,3	0,82
int-be	722	224	48	44	91	83,6	82,4	83	0,77
int-fr	899	262	56	72	90	78,4	82,4	80,4	0,74
iti-01	124	26	9	2	93	92,9	74,3	82,5	0,78
iti-02	84	41	1	1	98	97,6	97,6	97,6	0,96
iti-03	261	83	22	15	90	84,7	79	81,8	0,75
iti-04	547	154	15	21	95	88	91,1	89,5	0,86
iti-06	281	87	8	7	96	92,6	91,6	92,1	0,89
iti-07	94	24	7	1	94	96	77,4	85,7	0,82
nar-be	582	190	31	36	92	84,1	86	85	0,80
nar-ch	573	172	16	65	90	72,6	91,5	80,9	0,74
nar-fr	468	149	25	40	90	78,8	85,6	82,1	0,76
Total	**11736**	**3454**	**589**	**843**	**91**	**80,4**	**85,4**	**82,8**	**0,77**

Il ressort du Tableau 1 que les taux de syllabes proéminentes de H1 et H2 s'élèvent à 24,3% et 29,3%, respectivement (à comparer à la fourchette 19%-49% de [3]). Le taux d'accord inter-juges a été quantifié à l'aide du coefficient Kappa, et évalué à 0,77 pour l'ensemble du corpus. Quant à la F-mesure, elle indique un taux d'accord de 82,8% (rappel : 80,4 et précision 85,4). A la suite de [9], nous pensons que si le taux d'accord est si bon, c'est parce que les codeurs ont suivi un protocole strict, sur lequel ils se sont entraînés auparavant.

3.4. Annotation de référence

Une annotation consensuelle a émergé de la discussion des désaccords relevés dans la tire COMPARE, et elle a été considérée comme annotation de référence pour l'analyse des syllabes proéminentes. Le Tableau 3 donne le détail de la composition du corpus C-PROM, qui comprend 28 locuteurs (12 femmes, 16 hommes) et 17.778 syllabes. Parmi celles-ci, 805 (4.5%) sont notées spécifiquement dans la tire delivery, 4570 sont annotées proéminentes (25.7%) et 12.403 (69.7%) ne sont ni proéminentes ni notées « delivery ».

4. CONCLUSION

Le but de cet article était de présenter C-PROM, un corpus annoté pour l'étude des proéminences en français. Nous avons essayé de synthétiser en quelques lignes les principaux travaux à l'origine de la mise au point de ce corpus de référence et de présenter la méthodologie suivie pour sa constitution. La méthodologie proposée doit être testée sur de plus larges enregistrements avec davantage d'annotateurs. Cependant cette première expérience s'est révélée encourageante, le consensus entre les deux codeurs atteignant un taux satisfaisant. Des parties du corpus ont déjà été utilisées pour entraîner des systèmes de détection automatique [6] [9] [11], et pour des études de discrimination automatique des genres de discours sur des critères prosodiques [12] **Erreur ! Source du renvoi introuvable.**. Aussi espérons-nous que ce corpus permettra de rendre comparables les prochaines études relatives à la problématique de la proéminence, chose impossible auparavant en raison du manque de données communes.

Le corpus est téléchargeable à l'adresse suivante : http://sites.google.com/site/corpusprom

5. REMERCIEMENTS

Cette recherché a été soutenue par le Fonds National de la Recherche scientifique Suisse, (subsides n° PBNEP1-127788, Université de Neuchâtel) et par le Programme Wist2 Convention n°616422, financé par la Région wallonne (Belgique), Projet *EXPRESSIVE*.

BIBLIOGRAPHIE

[1] Durand, J., Laks, B. & C. Lyche. "La phonologie du français contemporain: usages, variétés et structure", *in* Pusch, C. & W. Raible (eds.), Romance Corpus Linguistics, Tübigen, Gunter Narr Verlag, 93-106, 2002.

[2] Lacheret-Dujour, A., Lyche, Ch. & M. Morel, "Pour une transcription prosodique normalisée au sein du projet PFC (phonologie du français contemporain): champ d'action et limites", Actes des 25^{s} JEP, Fès, Maroc, 2004.

[3] Poiré, P. "La perception des proéminences et le codage prosodique", Bulletin PFC, 6, 69-79, 2006.

[4] Martin, Ph. 2006. « La transcription des proéminences accentuelles : mission impossible ? », Bulletin PFC, 6, 81-87.

[5] Morel M., Lacheret-Dujour, A. Lyche, Ch., Morel, M. & F. Poiré. "Vous avez dit proéminence?", Actes des 26èmes journées d'étude sur la parole, Dinar, Maroc, 2006.

[6] Obin, N., Goldman, J.-P., Avanzi, M. & Lacheret-Dujour, A. "Comparaison de trois outils de détection semi-automatique des proéminences dans les corpus de français parlé", Actes des 22èmes JEP, Avignon, 2008.

[7] Goldman, J.-P. "EasyAlign: a semi-automatic phonetic alignment tool under Praat", http://latlcui.unige.ch/phonetique, 2008.

[8] Boersma, P. & Weenink, D. Praat: doing phonetics by computer (Version 5.2). www.praat.org, 2010

[9] Buhmann, J., Caspers, J., van Heuven, V, Hoekstra, H., Martens, J.-P. & M. Swerts, "Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non Expert Transcribers in the Spoken Dutch Corpus", LREC Processing, 779-785, 2002.

[10] Avanzi, M., Goldman, J.-P. Lacheret-Dujour, A. Simon, A.-C. & A. Auchlin, "Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé", Cahiers of French Language Studies, vol. 13, no. 2, pp. 2–30, 2007.

[11] Goldman, J.-P.; Avanzi, M.; Lacheret-Dujour, A.; Simon, A.-C.; Auchlin, A. A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French. In Proceedings of Interspeech'07, Antwerp, Belgium, August 27-31. 98-101, 2007.

[12] Obin, N., Lacheret-Dujour, A., Veaux, C., Rodet, X & A.C. Simon, "A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features", Interspeech, Brisbane, 2008.

[13] Simon A.C., Auchlin A., Avanzi M., Goldman J.-Ph., "Les phonostyles: une description prosodique des styles de parole en français", in Les voix des Français. En parlant, en écrivant, Abecassi, M. & G. Ledegen (eds), Berne, Peter Lang, 2010, sous presse.

Table 2. Contenu détaillé du corpus C-PROM. Avec, de gauche à droite : genre de discours, cote du sous-corpus, sexe du locuteur, durée, nb. de syllabes, nb. de syllabes non proéminentes, nb. de syllabes non proéminentes, nb. de syllabes associées à un marqueur delivery, nb. de syllabes associées à une hésitation, associées à une séquence postfocale et un schwa post-tonique. Les sous-totaux sont en gras. Les totaux pour l'ensemble du corpus figurent dans la ligne grisée tout en bas du tableau.

	Genres	Fichiers	Loc.	Durée (sec.)	Total syllables	Non-prom syllables	Prom syllables	Delivery syllables			
								total	Z	$	@
+formel	Journaux radiophoniques	JPA-BE	M	253	1315	963	312	40	24	0	16
		JPA-CH	F	180	879	610	242	27	12	0	15
		JPA-FR	M	188	971	683	256	32	19	0	13
		total	**2M/1F**	**621**	**3165**	**2256**	**810**	**99**	**55**	**0**	**44**
	Lectures	LEC-BE	M	114	606	492	111	3	0	0	3
		LEC-CH	M	137	606	403	196	7	0	0	7
		LEC-FR	M	150	618	462	153	3	0	0	3
		total	**3M**	**401**	**1830**	**1357**	**460**	**13**	**0**	**0**	**13**
	Discours politiques	POL-BE	M	188	420	246	160	14	0	0	14
		POL-CH	F	230	1011	753	257	1	0	0	1
		POL-FR	M	217	743	533	209	1	1	0	0
		total	**2M/1F**	**635**	**2174**	**1532**	**626**	**16**	**1**	**0**	**15**
	Conférences	CNF-BE	F	244	1066	776	250	40	35	0	5
		CNF-CH	M	219	950	627	260	63	55	7	1
		CNF-FR	F	224	1117	798	301	18	12	1	5
		total	**1M/2F**	**687**	**3133**	**2201**	**811**	**121**	**102**	**8**	**11**
	Interviews radiophoniques	INT-BE	2F	296	1189	769	317	103	56	32	15
		INT-FR	2M	331	1402	996	346	60	21	30	9
		total	**2M/2F**	**627**	**2591**	**1765**	**663**	**163**	**77**	**62**	**24**
	Itinéraires	ITI-01	M	50	172	117	36	19	15	2	2
		ITI-02	2M	47	142	82	42	18	17	1	0
		ITI-03	F	100	419	270	104	45	30	11	4
		ITI-04	2F	204	790	538	197	55	50	1	4
		ITI-05	M	28	128	92	27	9	8	0	1
		ITI-06	1M/1F	128	431	298	106	27	17	8	2
		ITI-07	M	33	140	98	30	12	10	2	0
		total	**6M/3F**	**590**	**2222**	**1495**	**542**	**185**	**147**	**25**	**13**
- formel	Récits de vie	NAR-BE	F	206	939	634	238	67	55	12	0
		NAR-CH	F	218	949	632	228	89	75	8	6
		NAR-FR	F	198	775	531	192	52	46	2	4
		total	**3F**	**622**	**2663**	**1797**	**658**	**208**	**176**	**22**	**10**
	7 genres	**24**	**16M/12F**	**4183**	**17778**	**12403**	**4570**	**805**	**558**	**117**	**130**

Indices phonétiques et contraintes phonologiques : caractérisation du syntagme intermédiaire en français

Amandine Michelas & Mariapaola d'Imperio

Université Aix-Marseille I et Laboratoire Parole et langage, CNRS, Aix-en-Provence, France

michelas@lpl-aix.fr, mariapaola.dimperio@lpl-aix.fr

ABSTRACT

The two experiments reported here support an analysis based on constraints which reflect the syntax-prosody interface. This analysis proves that the Intermediate Phrase (ip) exists in French. The ip is ranked higher than the Accentual Phrase and smaller than the Intonation Phrase in the prosodic hierarchy and it is not restricted to specific syntactic marked constructions as it was previously proposed. We predict that the interaction of (i) a syntactic constraint (ALIGN-XP,R, ip, R) aligns the right edge of a maximal projection with the right edge of an ip with (ii) a phonological constraint (MIN-BIN) stating that non final ip consists of minimally two APs, conspires to place an ip-boundary in French. An interesting interplay of duration and pitch height is responsible for signaling the boundary.

Keywords: prosodic phrasing, Intermediate Phrase, preboundary lengthening, downstep, pitch reset, French.

1. INTRODUCTION

Les deux expériences présentées ici ont été réalisées dans le cadre de la Théorie de l'Optimalité (OT) [4]. Dans cette théorie, la grammaire est envisagée comme un ensemble de contraintes, qui sont supposées universelles, tandis que la hiérarchisation de ces contraintes est propre à chaque langue. Les deux contraintes testées dans cet article sont des versions modifiées de deux contraintes proposées dans le cadre de l'OT. La première contrainte ALIGN,XP,R est une contrainte d'alignement proposée par [9], qui prédit des points d'ancrage où structure prosodique et structure syntaxique coïncident. Cette contrainte exige que la frontière d'une projection syntaxique majeure (XP) coïncide avec la frontière d'un constituant prosodique. La deuxième contrainte pertinente pour notre étude est une contrainte phonologique de binarité qui a attrait à la taille des constituants prosodiques. Ce type de contrainte (appelée MAX-BIN dans [7] et *Binary Minimum (Map)* ou *Binary Maximum (Map)* dans [8]) oblige un constituant de rang donné dans la hiérarchie à contenir soit au minimum, soit au maximum, un nombre défini de constituants prosodiques de niveau inférieur.

Les études qui se sont intéressées à la structure prosodique du français s'accordent généralement sur l'existence de deux niveaux de structure prosodique nommés IP (*Intonational Phrase*) et AP (*Accentual Phrase*) dans le modèle postulé par Jun & Fougeron, [3]. Certains travaux [1] ont déjà tenté de modéliser les groupements prosodiques du français dans le cadre de l'OT et ont notamment proposé un contrainte d'alignement particulière pour la rupture syntaxique entre un syntagme nominal (SN) un syntagme verbal (SV). Selon ces travaux, il semblerait que la frontière droite d'un SN soit alignée avec la frontière droite d'un syntagme intonatif (IP). Cependant, de récentes études [2], [5] ont mis évidence l'existence d'un niveau de structuration intermédiaire (ip) en français qui ne serait pas restreint à des structures spécifiques tel que cela avait été proposé précédemment [3]. Notre hypothèse est que l'interaction entre une contrainte d'alignement ALIGN-XP,R,ip,R (« aligne la frontière droite d'une projection syntaxique majeur avec la frontière droite d'un ip ») et une contrainte de binarité, que nous appelons MIN-BIN, serait responsable du placement d'une frontière d'ip à l'intérieur d'un syntagme intonatif (IP).

2. EXPERIENCE 1

Nous allons dans un premier temps tester la contrainte ALIGN-XP,R,ip,R. Notre hypothèse est que la durée segmentale ainsi que les valeurs de f_0 seront plus importantes lorsque la frontière d'AP correspond à une disjonction syntaxique majeure (AP/XP).

2.1. Méthode

20 phrases de type SVO contenant des syllabes cibles ont été construites. Chaque syllabe cible pouvait apparaître dans quatre contextes différents : 1- à l'intérieur d'un mot prosodique (noté PW, *Prosodic Word*), 2- en frontière droite d'AP qui n'est pas alignée avec la frontière droite d'une XP (AP), 3- en frontière droite d'AP qui est alignée avec la frontière droite d'une XP (AP/XP), 4- en frontière d'IP (Table 1).

1- Intérieur d'un PW	Les gre**NA**diers]AP de Marrakesh]AP/XP ne pousent pas bien vers chez nous.]IP
2- En frontière d'AP	Le sau**NA**]AP de Paolo]AP/XXP deviendra incontournable.]IP
3- En frontière d'AP/XP	Le sau**NA**]AP deviendra]AP/XP incontournable.]IP
4- En frontière d'IP	Le sau**NA**]IP d'après ce qu'on m'a dit]IP n'est pas très loin.]IP

Table 1 : Contextes d'apparition des syllabes cibles du corpus.

Deux locuteurs de langue maternelle française ont lu les vingt phrases du corpus à quatre reprises à deux débits d'élocution (normale/rapide) pour un total de 320 phrases expérimentales.

2.2. Résultats

Un modèle mixte (effets fixes : débit/frontière/locuteur ; effet aléatoire : consonne précédente) a été réalisé pour la durée des voyelles. L'effet du locuteur n'était pas significatif [t=2.45 p=0.11]. Conformément à nos hypothèses, les résultats montrent qu'à débit d'élocution normal, les voyelles en frontière d'AP/XP étaient significativement plus longues que les voyelles en frontière d'AP [t=2.64, p<0.05] (Figure 1). De plus la durée des voyelles en final d'AP/XP était significativement différente de celle des voyelles en final d'IP [t=3.83, p<0.05].

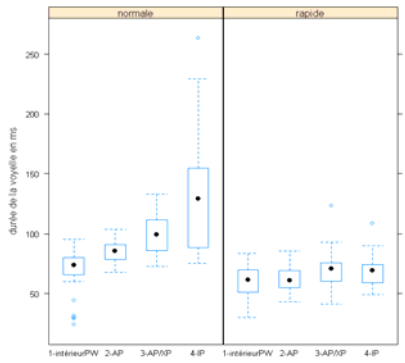


Figure 1 : Durée des voyelles cibles en fonction du type de frontière à débit d'élocution normal et rapide.

Concernant les valeurs de f_0, deux modèles mixtes (effets fixes : débit/frontière ; effet aléatoire : consonne précédente) ont été conduits séparément pour chaque locuteur. Contrairement à ce qui était attendu, chez les deux locuteurs et à vitesse de parole normale, les résultats montrent que les valeurs de f_0 associées aux syllabes finales d'AP/XP ne sont pas significativement plus hautes que les valeurs de f_0 associées aux syllabes finales d'AP (Figure 2).

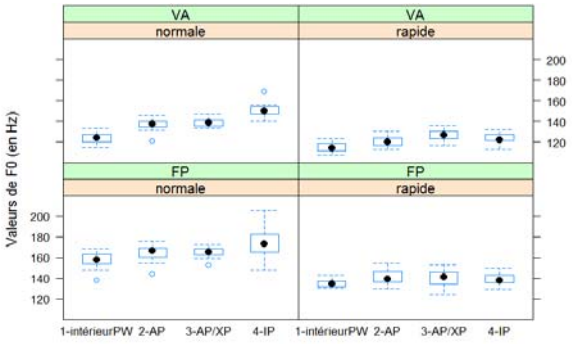


Figure 2 : Valeurs de f_0 associées aux syllabes cibles en fonction du type de frontière à débit d'élocution normal et rapide pour les deux locuteurs VA et FP.

2.3. Discussion

A débit d'élocution normal, les résultats mettent en évidence un allongement significatif associé à la frontière d'AP lorsque cette frontière est associée à une rupture syntaxique majeur. Par contre, les résultats relatifs aux valeurs de f_0 montrent que les syllabes en position finale d'AP/XP ne sont pas significativement différentes de celles situées en finale d'AP alors qu'elles sont significativement plus basses que celles situées en finale d'IP. Ces résultats nous permettent d'affirmer que (i) la rupture entre un SN et un SV ne semble pas être alignée avec une frontière droite d'IP contrairement à ce qui avait été proposé précédemment [1] (les valeurs de f_0 observées en position finale d'AP/XP sont significativement inférieures à celles observées en finale d'IP) et (ii) alors que l'alignement entre structure syntaxique et structure prosodique semble renforcer l'allongement présent avant la frontière, la contrainte ALIGN-XP,R,ip,R ne semble pas suffisante pour rendre compte de l'émergence de la frontière d'ip en français (les valeurs de F0 observées en finale d'AP/XP ne sont pas significativement supérieures à celles observées en finale d'AP). fait que le SN sujet contenait toujours un seul AP dans nos données nous a conduit à supposer l'interaction d'une contrainte de type phonologique avec la contrainte d'alignement. C'est ce que nous testons dans l'expérience suivante.

3. EXPERIENCE 2

Dans cette expérience nous testons l'interaction d'une contrainte de binarité MIN-BIN stipulant que l'ip est constitué au minimum de 2 APs en français, avec la contrainte d'alignement ALIGN-XP,R,ip,R. Nous prédisons que l'interaction de ces deux contraintes conspire à placer une frontière d'ip contenant minimum 2 APs en correspondance avec la frontière droite d'une rupture syntaxique majeur telle que la frontière entre un SN et un SV.

3.1. Méthode

Nous avons étudié la durée et les valeurs de f_0 de syllabes cibles contenues dans des phrases dont le SN sujet était composé soit de 2 APs (Figure 3a) soit de 3 APs (Figure 3b).

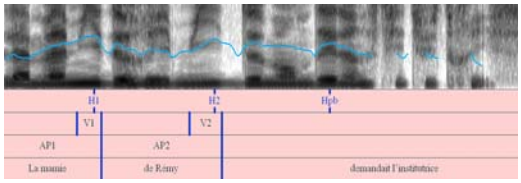


Figure 3a: Courbe de f_0 pour la phrase « La mamie de Rémy demandait l'institutrice » où le SN sujet est composé de 2 APs.

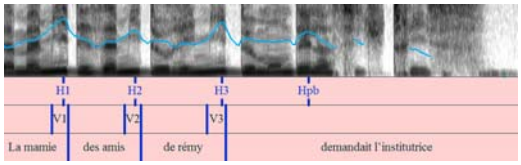


Figure 3b: Courbe de f_0 pour la phrase « La mamie des amis de Rémy demandait l'institutrice » où le SN sujet est composé de 3 APs.

Le corpus a été lu par deux locuteurs de langue maternelle française à quatre reprises à deux vitesses de parole (normale/rapide) pour un total de 128 phrases. Le découpage en APs des SN sujets de chaque phrase a été vérifié par un expert. A vitesse de parole normale, tous les SN sujets ont été découpés soit en 2 APs soit en 3 APs (frontières d'AP marquées par une montée de la f_0 ainsi que d'un allongement de la dernière syllabe du syntagme [3]). A vitesse de parole rapide, toutes les frontières d'APs n'étaient pas accompagnées d'un allongement. Celles-ci ont toutefois été considérées comme frontières d'AP en raison de la montée significative de la f_0 associée à l'accent primaire (LH*) qui affectait la dernière syllabe du syntagme et non le début du syntagme comme l'aurait fait une montée initiale (LHi) associée à l'accent secondaire (cf. [3,5] pour une discussion à ce sujet).

Dans un premier temps, nous avons comparé les valeurs de durée et de f_0 de V2 (voyelle finale du 2nd AP à l'intérieur du NP sujet) dans la condition 2 APs et dans la condition 3 APs. Les valeurs de durée ont été évaluées de manière relative à V1 (voyelle finale du 1er AP à l'intérieur du NP sujet). De la même manière, les valeurs de f_0 ont été calculées de manière relative à H1 (premier pic de f_0). Ainsi les rapport de V2/V1 et de H2/H1 ont été calculés. Ces mesures ont été réalisées de manière à normaliser la variabilité des durées et des variations de registre observées chez un même locuteur. Nos hypothèses étaient que les indices prosodiques (durée + valeur de f_0) associées à V2 seraient plus importants dans la condition 2 APs que dans la condition 3 APs en raison de la frontière d'ip auquel est associé V2 dans la condition 2 APs. Dans un deuxième temps, nous nous sommes intéressés au premier pic de f_0 (LH*) après la frontière d'ip dans la condition 2 APs et dans la condition 3 APs (noté Hpb dans les figures 4a et 4b). Dans d'autres langues que le français [6], il a été montré que le registre est réinitialisé après une frontière prosodique ce qui signifie un retour à des valeurs de f_0 semblable à ce que l'on trouve en début d'énoncé. Nous nous attendons donc à observer des valeurs de f_0 semblables à celles associées au premier pic de f_0 de l'énoncé (H1) en frontière d'ip. De plus des phénomènes de réinitialisation partielle (retour à des valeurs de f_0 qui ne sont pas aussi hautes que celles observées en début d'énoncé) ont été observés dans des langues germaniques [10] afin de marquer le début d'un constituant prosodique qui n'est pas le premier de l'énoncé. Notre hypothèse était que le début du second ip de l'énoncé « demandait l'institutrice » subirait un phénomène de réinitialisation partielle semblable à ce qui a été observé dans ces langues. Nous nous attendions donc à ce que le premier pic de f_0 après la frontière d'ip (Hpb) ne soit pas aussi haut que celui associé à la frontière d'ip. Notre hypothèse était que Hpb serait plus bas que le ton H associé à l'ip que nous nommons H- (correspondant à H2 dans la condition 2 APs et à H3 dans la condition 3 APs).

3.2. Résultats

Deux modèles mixtes (effets fixes : nombre d'APs/débit d'élocution/type de consonne/locuteur ; effet aléatoire : consonne précédente) ont été réalisés séparément pour les ratios de durée (V2/V1) et les ratios de f_0 (H2/H1). L'effet du locuteur n'était pas significatif [H2/H1 : t=2.16, p=0.2467 ; V2/V1 : t=2.607, p=0.1107].

Conformément à nos hypothèses et quelle que soit le débit d'élocution, le rapport de la durée de V2 sur la durée de V1 était significativement plus élevé dans la condition 2 APs que dans la condition 3 APs (débit normal : t=-8.487, p<0.05 ; débit rapide : t=-3.250, p<0.05) et le rapport de la hauteur de H2 sur la hauteur de H1 était significativement plus élevé dans la condition 2 APs que dans la condition 3 APs (débit normal : t=-3.67, p<0.05 ; débit rapide : t=-9.43, p<0.05) (Figure 4).

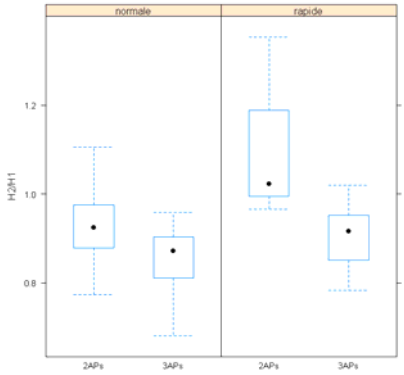


Figure 4 : Ratios de f_0 de H2 par H1 dans la condition 2 APs et 3 APs à débit d'élocution normal et rapide.

Concernant les valeurs de f_0 de Hpb, un modèle mixte nous a permis de mettre en évidence le fait que le rapport de la hauteur de Hpb sur la durée de H1 est significativement plus élevé que le rapport de la durée de H- sur la durée de H1 quelque soit le débit d'élocution [débit normal : t=-5.26, p<0.05 ;débit rapide : t=-9.30, p<0.05]. L'effet du locuteur n'était pas significatif [t=2.11, p=0.9118].

3.3. Discussion

Les syllabes finales d'ip ont été produites avec des valeurs de durée et de f_0 supérieures à celles en final d'APs situés en position médiane dans le SN sujet. Les résultats montrent également que les valeurs de f_0 des syllabes associées à la frontière d'ip sont à peu près équivalentes au premier LH* de l'énoncé (H1 dans la condition 2 APs et 3 APs). La frontière d'ip n'est donc pas signalée par des valeurs de f_0 plus importantes que celles que l'on peut trouver en frontière d'AP mais plutôt par un retour à la ligne de référence marquée par le premier LH* de l'énoncé (réinitialisation du registre). Ainsi dans la figure 3b, H2 est

significativement plus bas que H1 en raison du downstep des tons H* à l'intérieur de l'ip et H3 est plus haut que H2 et à peu près équivalent à H1 en raison d'un retour au registre initial avant la frontière d'ip. De plus nous pouvons observer en français un phénomène de réinitialisation partielle au début du second ip (sur Hpb). En effet, les valeurs de Hpb étaient significativement inférieures à celles observées en frontière d'ip.

4. DISCUSSION GENERALE

Les résultats obtenus nous permettent d'affirmer que les deux contraintes ALIGN XP,R,ip,R et MIN-BIN sont pertinentes pour rendre compte des placements des frontières de l'ip en français. Ces résultats montrent également que la rupture ente un syntagme nominal (SN) et un syntagme verbal (SV) est alignée avec la frontière droite d'un ip (syntagme intermédiaire) et non avec la frontière droite d'un IP (syntagme intonatif) comme cela avait été précédemment proposé [1]. La frontière droite de l'ip est marquée : (i) par un allongement significatif de la syllabe finale (ii) par un contour intonatif montant dû à la présence d'un ton H-qui est responsable d'un retour à la ligne de registre de référence marquée par le premier H* de l'énoncé. Nos résultats mettent également en évidence un phénomène de réinitialisation partielle qui survient au début du second ip de l'énoncé. Ces résultats vont dans le sens d'un emboîtement des downsteps comme cela a été proposé pour certaines langues germaniques telles que l'allemand [10] : un premier niveau de downstep relatif aux H* successifs à l'intérieur de l'ip et un deuxième niveau de downstep relatif aux ips à l'intérieur de l'IP en français.

Nos résultats laissent également entrevoir une compensation très intéressante des indices marquant les frontières prosodiques en fonction du débit d'élocution observé. Nos données montrent que la durée segmentale ne semble pas être un indice pertinent dans le marquage des frontières prosodiques à débit d'élocution rapide puisque nous n'avons noté aucune différence significative entre la durée des syllabes en finales d'AP, d'ip et d'IP dans la première expérience. Par contre, les indices relatifs à la f_0 semblent venir compenser la durée segmentale dans le marquage des frontières prosodiques du français puisque nous avons noté des variations de f_0 plus importantes à débit d'élocution rapide qu'à débit d'élocution normal (expérience 2). Ces résultats montrent que la prise en compte des différents indices est indispensable dans la définition des frontières prosodiques.

5. CONCLUSIONS

Nous proposons de définir l'ip en français, en tant que constituant prosodique hiérarchiquement inférieur à l'AP et supérieur à l'IP, à l'aide de deux contraintes formulées dans le cadre de l'OT : ALIGN, XP, R, ip, R et MIN-BIN. La frontière droite de ce constituant est marquée (i) par un allongement significatif de la dernière syllabe du constituant (ii) par un ton H- responsable d'un blocage du downstep des ton H* à l'intérieur de l'ip et (iii) par un phénomène de réinitialisation partielle qui affecte le début du second ip de l'énoncé.

BIBLIOGRAPHIE

[1] E. Delais-Roussarie. Pour Une approche parallèle de la structure prosodique: étude de l'organisation prosodique et rythmique de la phrase française. Thèse de Doctorat, Université de Toulouse Le Mirail, 1995.

[2] M. D'Imperio & A. Michelas. Embedded register levels and prosodic phrasing in French. *Proceedings of the Speech Prosody 2010 Conference*, Chicago, Etats-Unis, 4 pages, à paraître.

[3] S.A. Jun & C. Fougeron. A phonological Model of French Intonation. *Probus*, 14 :147-172, 2000.

[4] J. McCarthy & A. Prince. Generalized Alignment. In G. Booij and J. Van Merle [Ed] *Yearbook of Morphology 1993*, Dordrecht: Kluwer, pages 79-153.

[5] A. Michelas & M. D'Imperio. Durational Cues and Prosodic Phrasing in French. *Proceedings of the Speech Prosody 2010 Conference*, Chicago, Etats-Unis, 4 pages, à paraître.

[6] J. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. Ph.D thesis, MIT, 1980

[7] P. Prieto. The Intonational Phonology of Catalan. In S.A. Jun [Ed], *Prosodic Typology*, volume 2, Oxford University Press, Oxford, in press.

[8] F. Sandalo & H. Truckenbrodt. Some Notes on Phonological Phrasing in Brazilian Portuguese. *MIT Working Papers in Linguistics*, 42:285-310, 2002.

[9] E. Selkirk. The Interaction of Constraints on Prosodic Phrasing. In M. Horne [Ed], *Prosody:Theory and Experiment, Kluwer Academic Press*, pages 231-261, 2000.

[10] H. Truckenbrodt & C. Féry. Sisterhood and Tonal Scaling, *Studia Linguistica*, 59(2/3): 223-243, 2005.

Antériorisation/aperture des voyelles /ɔ/~/o/ en français du Nord et du Sud

Philippe Boula de Mareüil, Martine Adda-Decker, Cécile Woehrling

LIMSI-CNRS
BP 133 — F-91403 Orsay CEDEX
mareuil@limsi.fr, madda@limsi.fr, woehrling@vecsysresearch.com
http://www.limsi.fr

ABSTRACT

The considerable amount of data we now have at our disposal and advances in the area of speech processing enable us to quantify well-known and lesser-known phonetic tendencies. In this study, a large corpus of Northern and Southern French is investigated. The realisation of back mid vowels is examined through two approaches using automatic phoneme alignment, based on formant measurements and pronunciation variants. Two features are addressed: /ɔ/ fronting in the North and /o/ opening within some words (e.g. spelled with 'au' or 'ô') in the South. The results of the two approaches converge, showing that these phenomena contrast Northern and Southern French.

Keywords: corpus phonology, French varieties

1. INTRODUCTION

Dans cet article, nous nous proposons de montrer quelle peut être la contribution du traitement automatique de la parole à l'étude des voyelles moyennes /ɔ/~/o/ en français du Nord et du Sud. Des contraintes phonologiques spécifiques pèsent sur ces voyelles moyennes postérieures arrondies [15] : seul le timbre semi-fermé apparaît en syllabe finale ouverte (on n'oppose pas *pot* */pɔ/ à *peau* /po/ en français « standard » à tout le moins). Inversement, seul le timbre semi-ouvert apparaît avant /ʁ/.

Certains critères orthographiques entrent en ligne de compte : le digramme 'au' tend à être prononcé semi-fermé (/o/, comme le 'ô'), hormis devant 'r' où l'on a /ɔ/ ; et le timbre semi-ouvert /ɔ/ est considéré comme la forme sous-jacente du 'o' graphique ailleurs qu'en syllabe finale ouverte. Cependant, il existe de nombreuses exceptions où le phonème cible est /o/ : devant la consonne allongeante /z/, dans certains mots comme *fosse* ou *atome*, dans les préfixes *aéro-*, *micro-*, *psycho-*, etc. [16]. En outre, la voyelle racine (/o/) tend à être préservée par fidélité morphologique et sémantique à la base, dans des mots tels que *fossé*. Ajoutons que dans certains cas il y a désaccord entre les dictionnaires de prononciation quant à l'aperture de la voyelle finale ferme (ex. *synchrone*), voire antériorisation de /ɔ/ en [œ]. Dans un article célèbre, « C'est jeuli, le Mareuc ! » [14], Martinet analysait cette avancée du /ɔ/ en termes de rendement fonctionnel (relativement faible et sans grande incidence sur la compréhension, pour l'opposition /ɔ/~/œ/). Déjà pendant la Seconde Guerre mondiale, à partir des témoignages d'officiers recueillis dans un camp de prisonniers, l'auteur avait observé l'émergence de cette variante centralisée du /ɔ/ chez les locuteurs non méridionaux [13]. Il ouvrait ainsi des pistes pour des études empiriques et théoriques sur l'aménagement du système vocalique français.

Historiquement d'ailleurs, un mot latin comme *florire* a donné le français *fleurir* ; le verbe *florir* (d'où *florissant*) n'est qu'un archaïsme littéraire. On a d'autre part en synchronie les doublets *priorat~prieuré*, *senior~seigneur* (d'où *seigneurial* alors que l'adjectif dénominal de *directeur* est *directorial*) et des alternances morphologiques comme *mort~meurt*, parmi d'autres. Ce phénomène d'antériorisation a plus récemment été observé dans des travaux autour de l'harmonie vocalique [12][11][6]. Il serait aujourd'hui une marque de préciosité, alors qu'il avait une connotation populaire du XVII^e^ au XX^e^ siècle [1]. À notre connaissance, cependant, la prononciation du 'o' n'a pas été étudiée de façon systématique, en raison des difficultés pratiques à mener des enquêtes phonétiques sur le terrain.

De plus, ces observations, comme les règles phonologiques édictées plus haut, ont essentiellement été établies pour le français standard (parisien ou plus généralement du nord de la Loire). En français méridional, réputé pour ne pas faire la distinction *côte~cote*, des schibboleths comme *rose* ou *gauche* prononcés avec un [ɔ] ouvert sont pourtant bien connus [3][5].

La masse de travaux accumulés dans le cadre de projets récents, aussi bien que les instruments développés en traitement automatique de la parole, permettent aujourd'hui de regarder d'un œil nouveau ces phénomènes dont on n'a pas toujours conscience [17]. Nous avons pour quantifier ces différentes tendances déployé deux approches utilisant l'alignement automatique en phonèmes : (1) à base de formants, à partir d'un alignement tel qu'on peut le développer pour le français standard ; (2) à base de variantes de prononciations, où des alternances libres comme [ɔ]~[œ]~[o] sont autorisées.

Nous décrirons davantage la méthodologie après avoir introduit le corpus mobilisé pour cette étude. Nous présenterons ensuite les résultats en matière d'antériorisation du /ɔ/ et d'aperture du /o/ en français du Nord et du Sud, que nous discuterons pour finir.

2. CORPUS ET MÉTHODE

2.1. Corpus

Notre corpus est issu du projet « Phonologie du

Français Contemporain » (PFC) [4], qui a entrepris, dans le sillage de [13], de collecter des enregistrements (en lecture et en parole spontanée) couvrant un vaste territoire francophone, avec une dizaine de locuteurs bien ancrés géographiquement par point d'enquête. La partie analysée ici est constituée de douze points d'enquête : six dans la moitié nord de la France (Brécey, Brunoy, Dijon, Lyon-Villeurbanne, Roanne, Treize-Vents), un en Suisse romande (canton de Vaud) et cinq dans le sud de la France (Biarritz, Douzens, Lacaune, Marseille, Rodez). Malgré un substrat francoprovençal, la Suisse romande sera comptée comme Nord car sa variété de français est très peu perçue comme méridionale [17]. Aucun point d'enquête n'étant situé dans le département français du Nord, nous opposerons donc dans ce qui suit deux grandes variétés de français (Nord/Sud) sans nier que des divisions plus fines puissent être faites.

Le corpus traité représente plus d'une centaine de locuteurs : autant d'hommes que de femmes, de tranches d'âges équilibrées, de niveaux d'études et de professions variés, qui sont nés et ont passé la plus grande partie de leur vie en un même lieu. Totalisant plus de 30 heures d'enregistrement, transcrites orthographiquement, ces données comprennent 12 000 mots différents, représentant 15 000 occurrences de /ɔ/ et 9 000 occurrences de /o/ sous-jacents (dans des proportions Nord-Sud de 2/3-1/3). Pour chaque locuteur, nous avons à notre disposition — et utilisé dans cette étude — la lecture d'une liste d'une centaine de mots et d'un texte d'une vingtaine de phrases, ainsi que 10 minutes d'entretien guidé et de conversation libre, suivant un protocole labovien [10].

2.2. Alignement et extraction de formants

L'ensemble du corpus a été aligné en phonèmes par un système d'alignement automatique issu des travaux en reconnaissance de la parole menés au LIMSI [8]. L'alignement est fondé sur un principe identique, à la différence près que la suite de mots est ici connue. La transcription orthographique du signal de parole est utilisée pour générer des transcriptions phonétiques possibles à l'aide d'un dictionnaire de prononciations. Des modèles acoustiques sont utilisés pour comparer ces transcriptions phonétiques avec le signal de parole. La suite de phones la plus probable parmi les candidats est alors sélectionnée et alignée avec le signal acoustique.

À partir du signal ainsi segmenté en phones, les valeurs des formants extraites par un logiciel tel que PRAAT (http://www.praat.org) peuvent être moyennées pour chaque voyelle orale. Avant de calculer la moyenne, des filtres ont été appliqués — différents pour chaque voyelle, pour les hommes et pour les femmes — afin d'écarter les valeurs aberrantes, comme dans [9]. Dans cette première approche à base de formants, un dictionnaire de prononciation standard a été utilisé, dans lequel pour le mot *sauf*, par exemple, la forme canonique [sof] est donnée comme prononciation. Nous mesurerons d'éventuelles divergences de formants entre français du Nord et du Sud.

Dans la deuxième approche à base de variantes de prononciations, nous avons permis que chaque /ɔ/ et chaque /o/ sous-jacents soient alignés en [ɔ] (semi-ouvert), [œ] (antériorisé) ou [o] (semi-fermé). Ainsi a-t-on pour les mots *sol* et *sauf* les variantes [sɔl, sœl, sol] et [sɔf, sœf, sof] — nous notons entre barres obliques les prononciations standard et entre crochets leurs possibles réalisations régionales. Nous calculerons les taux de variantes alignées : cette approche est complémentaire de l'extraction de formants car elle manipule des classes symboliques qui sont intéressantes pour des interprétations catégorielles en phonologie.

Deux alignements ont donc été effectués, dans lesquels les mêmes modèles acoustiques, indépendants du contexte, avec mélanges de gaussiennes, ont été utilisés (512 gaussiennes par état, pour chaque phonème). Ces modèles acoustiques, appris sur de grandes quantités de données, correspondent à des formes relativement standard des phonèmes du français (non méridional).

3. RÉSULTATS

3.1. Analyse formantique

À l'aide du logiciel PRAAT, nous avons extrait les trois premiers formants (F1, F2, F3) toutes les 10 ms. Cependant, trop de valeurs de F3 (près du tiers) excédant 3000 Hz, nous avons considéré ces mesures insuffisamment fiables pour la qualité de nos données. En comparaison, nos filtres écartant les valeurs de F1 et F2 hors d'un intervalle raisonnable de ±500 Hz en moyenne n'ont rejeté que 5 % des voyelles orales. Nous nous en sommes donc tenus au plan F1/F2 pour tracer les triangles vocaliques du Nord et du Sud. Les triangles vocaliques correspondant aux femmes sont illustrés dans la Figure 1 — des différences Nord/Sud similaires étant manifestées par les hommes.

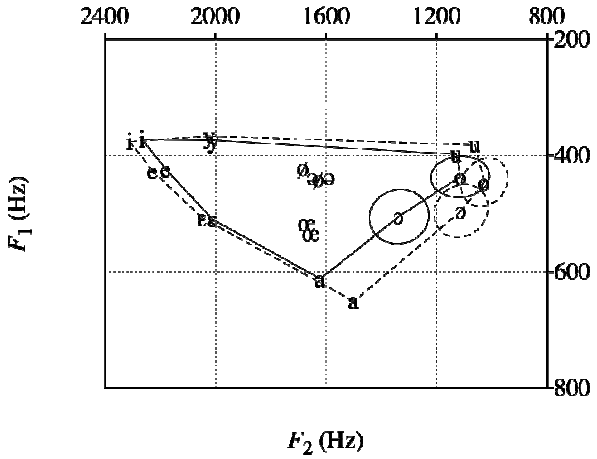


Figure 1. Triangles vocaliques des locutrices du Nord (en lignes pleines) et du Sud (en pointillés) analysées avec PRAAT. Les ellipses sont réglées à 10 % des occurrences autour du /ɔ/ et du /o/.

On peut remarquer que le triangle du Nord est inclus dans celui du Sud. Les voyelles d'arrière et le /a/ sont plus antérieurs chez les locutrices du Nord. Dans sa réalisation, le /ɔ/ est plus proche du /o/ dans le Sud que dans le Nord ; mais on note surtout que dans le Nord le /ɔ/ est plus fortement centralisé que les autres voyelles.

Pour cette voyelle /ɔ/, la différence de F2 entre locuteurs du Nord et du Sud est de l'ordre de 200 Hz chez les hommes comme chez les femmes.

Pour le /o/, on ne voit apparaître guère de différence Nord/Sud en matière de F1 (corrélat acoustique de l'aperture) : 380 Hz pour les locutrices du Nord, 400 Hz pour les locutrices du Sud. Les valeurs sont plus proches encore pour les hommes : 340 Hz dans le Nord, 350 Hz dans le Sud.

3.2. Antériorisation du /ɔ/

Nous nous sommes dès lors intéressés aux variantes de prononciations alignées pour le /ɔ/, en particulier. Les résultats de cette deuxième approche sont consignés dans la Table 1. Dans ce tableau, on peut observer 30 % d'antériorisations dans le Nord, contre seulement 10 % dans le Sud, où la fermeture en [o] est la prononciation majoritaire (à 52 %, en accord avec le rapprochement /ɔ/-/o/ qu'on pouvait observer dans les triangles vocaliques). Dans le Nord, c'est la variante [ɔ] qui est le plus souvent alignée, mais la variante [œ] est davantage sélectionnée que la variante [o].

Table 1 : taux de /ɔ/ alignés comme [ɔ], [œ] ou [o] en français du Nord et du Sud (%).

%/ɔ/	**[ɔ]**	**[œ]**	**[o]**
Nord	48	30	22
Sud	38	10	52

Des exemples de prononciation avec /ɔ/ majoritairement aligné en [œ] sont : *d'accord, personne, notre, votre, olympiques, officielles, connais, socialisme* (plus de 700 occurrences au total). Dans le premier mot (*d'accord*), le 'o' apparaissant avant une consonne vélaire a un F2 en moyenne de 1300 Hz pour les hommes et de 1400 Hz pour les femmes, soit 300 Hz au dessus d'un mot comme *bord(s)*. Ce cas est instructif dans la mesure où l'effet du /ʁ/ sur la voyelle précédente a été très controversé [1]. La séquence /œʁ/ est tout à fait possible en français. On note même que le phonème /œ/ apparaît le plus fréquemment avant /ʁ/ (notamment en raison de la productivité du suffixe agentif *-eur*) et que, pour Gadet [7] (p. 33), « les voyelles orales accentuées d'arrière connaissent une tendance à l'avancée, surtout devant [r] ». Nous observons pour notre part 15 % de cas d'antériorisation avec /ʁ/ en contexte gauche ou droit, ce qui est moins que pour d'autres phonèmes — 30 % (resp. 40 %) avec /s/ en contexte gauche (resp. droit), et même moins que la moyenne. Le /ɔ/ est surtout antériorisé lorsqu'en contexte gauche et droit figurent des consonnes d'avant, mais des cas d'antériorisation comme *d'accord* sont également possibles.

3.3. Aperture du /o/

Examinons à présent le cas du /o/ censé être semi-fermé en français de référence. Les résultats en termes de taux d'alignements sont rapportés dans la Table 2. Ils ne montrent pas de différences marquées entre français du Nord et du Sud, si ce n'est peut-être en ce qui concerne l'antériorisation en [œ], plus accusée dans le Nord. Dans le Sud, on a même davantage de /o/ alignés en [o], alors que les valeurs du F1 moyen sont légèrement plus élevées que dans le Nord (*cf.* § 3.1).

Table 2 : taux de /o/ alignés comme [ɔ], [œ] ou [o] en français du Nord et du Sud (résultats globaux).

%/o/	**[ɔ]**	**[œ]**	**[o]**
Nord	13	17	69
Sud	15	7	78

Il en va tout autrement si l'on restreint l'analyse à des mots comme *chose, cause, sauf, autre, pauvre, gauche, chaude, paume, rauque, fausses, côte, gnôle*, etc. Ces mots où le /o/ précède un /z/ et/ou est orthographié 'au' ou 'ô' sont au nombre de 600 dans le dictionnaire électronique ILPho [2]. Ils sont quelque 80 dans notre corpus, représentant 1500 occurrences. Le F1 du /o/ putatif, moyenné sur l'ensemble de ces mots pour les hommes et pour les femmes, est donné dans la Table 3.

Table 3 : F1 moyen du /o/ putatif dans quelque 80 mots tels que *autre*, en français du Nord et du Sud.

F1 /o/ (Hz)	Hommes	Femmes
Nord	360	410
Sud	410	480

Le F1 moyen est plus élevé dans le Sud que dans le Nord, avec une différence supérieure à 50 Hz, suggérant que le timbre est plutôt semi-ouvert ([ɔ]) : on atteint en effet, quand on ne les dépasse pas, les valeurs de F1 mesurées pour le /ɔ/ chez les locuteurs du Nord (410 Hz chez les hommes, 470 Hz chez les femmes). Les résultats de l'alignement avec variantes de prononciation, présentés dans la Table 4, corroborent très clairement ce fait : 50 % de /o/ sous-jacents sont alignés avec [ɔ] dans le Sud, alors que le pourcentage correspondant n'est que de 12 % dans le Nord.

Table 4 : taux de /o/ putatifs alignés comme [ɔ], [œ] ou [o] dans quelque 80 mots tels que *cause* en français du Nord et du Sud.

%/o/	**[ɔ]**	**[œ]**	**[o]**
Nord	12	16	72
Sud	50	7	43

Ces mots où le /o/ est majoritairement prononcé [o] dans le Nord et [ɔ] dans le Sud sont une bonne illustration de la nécessité de regarder en détail les résultats quand on analyse de grands corpus. Dans ces mots, on peut considérer un /ɔ/ sous-jacent dans le Sud, mais il se peut aussi que le phénomène soit lexical, à examiner au cas par cas. Ainsi, dans les points d'enquête méridionaux de PFC, le mot *autre(s)* n'est aligné avec [o] que dans 37 % des cas, tandis que le mot *chose(s)* l'est dans 51 % des cas — chacune de ces formes représentant plus de 100 occurrences.

Nous avons envisagé une possible influence du type de parole (lecture ou parole spontanée). Mais comme pour le /ɔ/ sous-jacent (§ 3.2), les différences en matière de taux d'alignement sont au maximum de 10 % entre lecture et parole spontanée. On n'observe pas non plus

de différences majeures entre les locuteurs les plus jeunes et les plus âgés. Les différences Nord/Sud qui émergent semblent donc relativement robustes.

4. CONCLUSION ET PERSPECTIVES

Les ressources considérables dont nous disposons aujourd'hui et les avancées dans le domaine du traitement automatique de la parole nous permettent de quantifier des phénomènes connus ou moins connus. Nous nous sommes concentrés dans cet article sur la réalisation de voyelles moyennes postérieures en français du Nord et du Sud. Deux approches ont été employées, à base de formants et de variantes de prononciations. Les résultats obtenus convergent pour une large part, mettant en évidence ou confirmant une tendance à l'antériorisation du /ɔ/ dans le Nord et à l'aperture du /o/ précédant un /z/ et/ou orthographié 'au' ou 'ô' dans le Sud.

Il y a à la fois la validation d'une approche à base d'alignement automatique qui peut être appliquée à d'autres traits de prononciation, d'autres dialectes et d'autres langues, et une comparaison à large échelle de variétés septentrionales et méridionales de français. Les résultats attendus pour la prononciation du digramme 'au' dans des mots comme *autre* suggèrent que la méthode à base d'alignement est appropriée, et permet d'éclairer de nouveaux phénomènes. En particulier, l'antériorisation du /ɔ/ est patente en français du Nord. Par sa fréquence au moins, elle pourrait devenir une variable discriminant le Nord du Sud au même titre (au moins) que les traditionnelles *choses jaunes ou rose*, prononcées avec un [ɔ] semi-ouvert. De telles formes sont moins nombreuses que des *d'accord* antériorisés dans notre corpus, mais font certainement davantage basculer la perception. Ceci pose la question du stéréotype, perçu de façon particulière dans la société [10] et de l'asymétrie entre français du Nord (représentant la norme) et français du Sud. Nos représentations et les discours (épi)linguistiques produisant en grande partie nos catégories de perception, il peut y avoir un décalage entre les réalités physiques (articulatoires ou acoustiques) et perceptives. Pour ce qui nous concerne ici, le mouvement d'avancée du 'o' fonctionnerait plutôt comme un indicateur au-dessous du seuil de la conscience [10].

Les résultats présentés ici sont donc à hiérarchiser et à mettre en relation avec la perception, qui n'est pas omnisciente mais sélective à certains événements saillants. Les résultats afférents à l'aperture du /o/ dans le Sud sont également à relier à la distinction entre syllabes accentuées/inaccentuées et à la prononciation du schwa final, pour interroger la « loi de position » [5]. Plus généralement, des données empiriques sur la pluralité des usages français sont d'un apport précieux pour la phonétique (descriptive ou didactique) et la phonologie (de corpus). Nous espérons que ces disciplines pourront dans le futur en tirer bénéfice.

5. REMERCIEMENTS

Ce travail a été partiellement financé par le programme Quæro de l'OSEO. Nous exprimons également notre profonde gratitude aux responsables du projet PFC (http:// www.projet-pfc.net).

6. BIBLIOGRAPHIE

[1] N. Armstrong & J. Low. C'est encœur plus jeuli, le Mareuc: some evidence for the spread of /ɔ/-fronting in French. *Transactions of the Philological Society*, 106(2) : 1–24, 2008.

[2] P. Boula de Mareüil, F. Yvon, C. d'Alessandro, V. Aubergé, J. Vaissière, A. Amelot. A French phonetic lexicon with variants for speech and language processing. *LREC*, Athènes, pages, 273–276, 2000.

[3] F. Carton, M. Rossi, D. Autesserre, P. Léon. *Les Accents des Français*, Paris, Hachette, 1983.

[4] J. Durand, B. Laks, C. Un corpus numérisé pour la phonologie du français. In G. Williams (ed.), *La linguistique de corpus*. Presses Universitaires de Rennes, Rennes, pages 205–217, 2005.

[5] J. Durand. Essai de panorama phonologique : les accents du Midi. In L. Baronian & F. Martineau (eds.), *Mélanges offerts à Yves-Charles Morin*, Presses de l'Université Laval, Québec, pages 123–170, 2008.

[6] Z. Fagyal N. Nguyen, P. Boula de Mareüil. From *dilation* to coarticulation: is there vowel harmony in French? *Studies in the Linguistic Sciences*, 32(1) : 1–21, 2002.

[7] F. Gadet. *Le français populaire*, Presses Universitaires de France, Paris, 1992.

[8] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, H. Schwenk. Where are we in transcribing French broadcast news? *Eurospeech*, Lisbonne, pages 1665–1668, 2005.

[9] C. Gendrot, M. Adda-Decker. Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German, *Eurospeech*, Lisbonne, pages 2453–2456, 2005.

[10] W. Labov. *Sociolinguistic patterns*, University of Pennsylvania Press, Philadelphia, 1972.

[11] M. Landick. The Mid-Vowels in Figures: Hard Facts. *French Review*, 68(1) : 88–102, 1995.

[12] I. Malderez. Vers la perte de l'opposition du lieu d'articulation des voyelles moyennes arrondies dans la parole des jeunes gens d'Île-de-France. *JEP*, Trégastel, pages 361–366, 1994.

[13] A. Martinet. *La prononciation du français contemporain*. Droz, Paris, 1945.

[14] A. Martinet. C'est jeuli, le Mareuc ! *Romance philology*, 11 : 345–355, 1958.

[15] D.C. Walker. *French Sound Structure*. University of Calgary Press, Calgary, 2001.

[16] H. Walter. *La dynamique des phonèmes dans le lexique français contemporain*. France-Expansion, Paris, 1976.

[17] C. Woehrling. *Accents régionaux en français : perception, analyse et modélisation à partir de grands corpus*. Thèse de doctorat de l'Université Paris-Sud XI, Orsay, 2009.

Hiérarchie prosodique et réalisation spectrale des voyelles

Cédric Gendrot et Kim Gerdes

Laboratoire de Phonétique et Phonologie, Université Paris3 Sorbonne Nouvelle, CNRS UMR7018
cgendrot@univ-paris3.fr ; kim@gerdes.fr

ABSTRACT

The link between the duration of vowels and their spectral realization has been validated for a long time by Lindblom (1963), i.e. the longer the vowels the more strengthened they are. Similarly, the relation between prosodic constituents of different levels (the prosodic hierarchy) and the duration of phonemes close to their boundaries has been demonstrated (in French, Fougeron 2001; Tabain 2003, 2005). In this study we aim at reproducing these results on non-controlled continuous speech. Procedures that try and detect automatically known prosodic categories are detailed. We show that the higher a vowel is in the prosodic structure of French, the more strengthened it is.

Keywords: prosody, formants, vowels, strengthening.

1. INTRODUCTION

Cette étude s'inscrit dans la cadre d'un projet de plus grande envergure s'attachant à décrire la variabilité des phonèmes en français. Grâce à l'utilisation de corpus de très grande taille segmentés automatiquement, nous avons pu étudier un nombre important de contextes afin de mieux quantifier leurs influences et leurs interactions. Nous avons observé lors d'études précédentes (Gendrot & Adda-Decker 2005) que la réalisation spectrale des voyelles était grandement influencée par leur durée : les voyelles les plus longues sont considérablement renforcées par rapport aux voyelles les plus courtes. Si l'on considère un espace acoustique formé par les 2 premiers formants de chaque voyelle, plus les voyelles sont longues et plus l'espace vocalique occupé par l'ensemble de ces voyelles est important (figure 1).

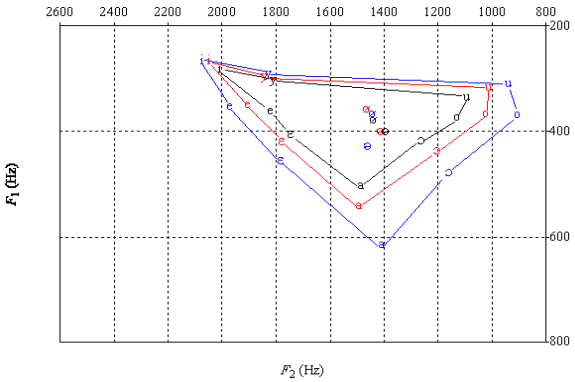


Figure 1: Valeurs moyennes de F1/F2 des voyelles orales du français en fonction de leur durée. Par ordre ascendant et de l'intérieur vers l'extérieur : noir [30-50ms], rouge [60-80], noir [90-110]).

Cette plus grande occupation de l'espace acoustique permet notamment aux voyelles de mieux se distinguer les unes des autres afin d'être plus facilement identifiables.

Les raisons de ces variations de durée sont multiples et des facteurs tels que le style et le débit du locuteur ont une grande influence. Des facteurs plus linguistiques tels que le contexte segmental, la position de la voyelle dans la syllabe, le mot, le syntagme ou l'énoncé ont également une forte influence. Les quatre unités précédemment mentionnées sont traitées comme des constituants prosodiques puisque représentés par des variations caractéristiques de durée et de f0 dans un but démarcatif (ils signalent des occurrences de frontières). Ces constituants prosodiques sont généralement considérés comme étant organisés au sein d'une hiérarchie prosodique : chaque constituant est imbriqué dans un constituant de plus haut niveau (Nespor et Vogel, 1986). Dans ce courant de recherche, il a été montré que plus le niveau du constituant prosodique est élevé dans la hiérarchie, plus les phonèmes aux frontières de ces constituants sont longs (pour le français, Fougeron 2001; Tabain 2003, 2005). Ces positions ont plus rarement été étudiées d'un point de vue spectral (et/ou articulatoire). Pour le français, Fougeron (2001) à partir de l'électropalatographie sur les consonnes (EPG) et Tabain (2003, 2005) grâce à l'articulographie sur les voyelles (EMMA) ont travaillé respectivement sur les positions initiales et finales. Leurs résultats montrent qu'un phonème en position strictement initiale ou finale de constituant sera influencé par sa position dans la hiérarchie prosodique. Plus le niveau sera élevé, plus le phonème sera renforcé.

Nous tentons ici de répliquer ces études sur de la parole continue à partir de mesures acoustiques. Les productions des locuteurs ne sont ainsi plus contrôlées ; si les données sont considérablement plus naturelles que celles de corpus lus, il est impossible de contrôler la catégorie prosodique produite par les locuteurs. Il est donc nécessaire de les détecter de façon indirecte. Notre objectif est de montrer que le niveau des constituants dans la hiérarchie prosodique a un effet sur la réalisation spectrale des voyelles. Quatre constituants prosodiques unanimement reconnus ont été choisis : la syllabe, le mot, le syntagme accentuel et le syntagme intonatif. Nous expliquerons quels choix ont été faits pour approcher ces constituants de façon automatique, et ceux-ci seront analysés à leurs frontières (en position initiale et finale).

2. MÉTHODE

2.1. Corpus et analyses

Les données de parole utilisées correspondent approximativement à 30 heures de parole radiophonique (environ 500 hommes et 300 femmes extraites du corpus ESTER (Galliano et al., 2005) et à partir de la segmentation effectuée par l'IRISA. Les procédures de transcription et de segmentation sont détaillées dans Bürki et al. (2008)

Des mesures des trois premiers formants (F1, F2 et F3) ont été effectuées automatiquement à l'aide de l'algorithme Burg implémenté dans Praat (Boersma et Weenink 2009). Des fourchettes de valeurs ont été établies afin de filtrer les valeurs fantaisistes (approx. 4% au final) dues à des erreurs de détection automatique. Plus de détails sur les mesures de formants et les précautions prises sont fournis dans Gendrot et Adda-Decker (2005). Comme mentionné dans l'introduction, les voyelles renforcées forment un espace acoustique beaucoup plus large, elles sont donc plus éloignées du centre acoustique. Nous émettons l'hypothèse que les voyelles aux frontières des constituants prosodiques de haut niveau (syntagme accentuel, puis syntagme intonatif) seront renforcées comparées aux voyelles en frontière de constituants prosodiques de plus bas niveau (le mot, puis la syllabe). Afin de quantifier cette variation, nous avons effectué par simple distance euclidienne une mesure d'éloignement du centre acoustique déterminé sur l'ensemble de nos données (F1 : 450Hz, F2 : 1540Hz) (Gendrot & Adda-Decker 2007; Bradlow 1996).

2.2. Sélection des catégories prosodiques

Quatre catégories ont été analysées dans cette étude, ces catégories tentent d'approcher les catégories prosodiques les plus acceptées : de la plus basse à la plus élevée, la syllabe, le mot, le syntagme accentuel et le syntagme intonatif. Nous décrivons ci-dessous les choix effectués pour détecter automatiquement chaque catégorie, en commençant par le mot. Les frontières de chaque catégorie seront analysées, c'est-à-dire aux positions initiales et finales. Les voyelles décrites comme initiales sont strictement initiales ("armée"), et non pas plus simplement dans la syllabe initiale. En effet, des résultats préliminaires ont montré que l'effet dû à la hiérarchie prosodique était essentiellement concentré sur la voyelle strictement initiale seulement. A l'inverse, pour les positions finales, les voyelles à la fois finales et en pénultième position ont été prises en compte car un effet semblable a été remarqué pour ces 2 positions.

Les frontières de mots ont été obtenues sur la base de la transcription manuelle, puis de l'alignement effectué automatiquement.

Les syllabes ont été déterminées à partir de la segmentation phonémique. Des règles de syllabation inspirées de Pallier (1994) ont été utilisées depuis le flux continu de phonèmes, i.e. les frontières de mots n'interviennent pas dans la syllabation. Par exemple, la séquence de mots "bon ami" est segmentée en 3 syllabes "bo", "na" et "mi" sauf s'ils sont séparés par une pause. Par manque d'occurrences, nous n'avons pu déterminer des positions initiales et finales pour le niveau syllabique ; par exemple les voyelles initiales de syllabe mais internes de mots sont très rares ("a**é**roport") et toutes les voyelles en syllabe interne de mot ont donc été analysées pour la catégorie "syllabe".

Le troisième niveau analysé ici est le syntagme accentuel (voir exemple en (1)). Un chunking syntaxique a été effectué sur la base d'un étiquetage grammatical automatique (Lefff : Clément et al. 2004) combiné à quelques règles de regroupement mises en place grâce au chunker du Natural Language Toolkit (http://nltk.org/index.php/Main_Page) :

o Les segments nominaux, prépositionnels et verbaux sont regroupés avec leur entourage le plus proche (clitiques, déterminants, prépositions, adjectifs, etc.)

o Puis, 3 règles de combinaisons sont appliquées :

1- Combinaison de tout segment terminant sur un auxiliaire ou modal avec le segment suivant.
2- Combinaison de tout segment verbal avec le segment suivant si la combinaison fait moins de 7 syllabes.
3- Combinaison de tout autre suite de segments qui fait moins de 7 syllabes.

Les segments découpés par cet algorithme peuvent avoir plus de 7 syllabes, si les règles précédentes le permettent, par exemple « avec qui j'ai pu m'entretenir », qui forme un groupe très naturel et difficilement à découper. La règle des 7 syllabes (Vaissière, 1971) pourra ultérieurement être supplantée par un calcul plus respectueux du débit de parole.

(1) combien de fois, la justice française a-t-elle accepté de se remettre en question, comme cela,

Par ce "chunking", nous tentons de nous approcher de la réalisation du syntagme accentuel. Nous sommes conscients que tous ces syntagmes ne seront pas "accentués", c'est-à-dire qu'ils ne seront pas tous caractérisés par un allongement final et/ou un contour mélodique montant. Cependant, utiliser des informations prosodiques pour s'en assurer aurait introduit un caractère circulaire dans notre étude puisque les voyelles les plus longues sont elles-mêmes renforcés. Notre méthode vise donc à évaluer la réalisation spectrale de syntagmes accentuels à un niveau syntaxique (sous-jacent), plutôt qu'en considérant des syntagmes accentuels d'après leurs caractéristiques prosodiques.

La quatrième catégorie analysée vise à s'approcher du syntagme intonatif et est détectée automatiquement sur la base des pauses (supérieures à 50ms) indiquées par l'alignement. En effet, il a été montré que la présence de pauses est un facteur important pour signaler la réalisation d'un syntagme intonatif (Jun & Fougeron 2000). Une détection de la forme du contour final de f0 (montant/descendant) a été effectuée dans le but de ne prendre en compte que les syntagmes ayant un contour montant. Cette méthode permet ainsi de les distinguer d'une position finale d'énoncé ayant un contour descendant. Aucune précaution de cet ordre n'a pu être effectuée pour les positions initiales. Nous sommes encore une fois conscients que la détection automatique de cette catégorie peut engendrer un certain nombre de détections erronées. Une explication semblable à celle évoquée pour les syntagmes accentuels sera proposée ici : l'objectif de cette étude est d'obtenir quatre catégories prosodiques, aussi proches que possible de celles mentionnées dans la littérature, sans se baser sur des caractéristiques prosodiques d'allongement. Une analyse supplémentaire (non détaillée ici) a montré des valeurs de durée et de f0 augmentant progressivement pour nos 4 catégories de la hiérarchie prosodique, confirmant ainsi la fiabilité de ces catégorisations.

3. RÉSULTATS

3.1. Positions initiales

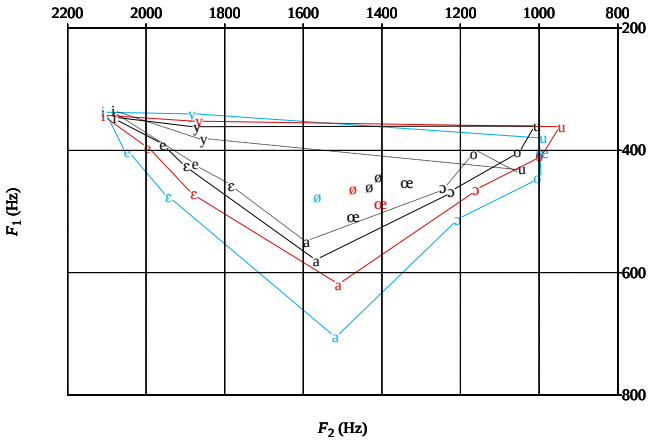


Figure 2 : Comparaison des 4 catégories prosodiques en position initiale. De l'intérieur vers l'extérieur : syllabe, mot, syntagme accentuel et syntagme intonatif.

Comme le montre la figure 2, les voyelles occupent un espace acoustique de plus en plus important en remontant la hiérarchie prosodique. Les mesures de dispersion présentées par la figure 3 révèlent que les valeurs augmentent globalement avec le niveau de la hiérarchie prosodique, c'est-à-dire que les voyelles s'éloignent du centre de l'espace vocalique, étant ainsi plus distinctes les unes des autres perceptivement.

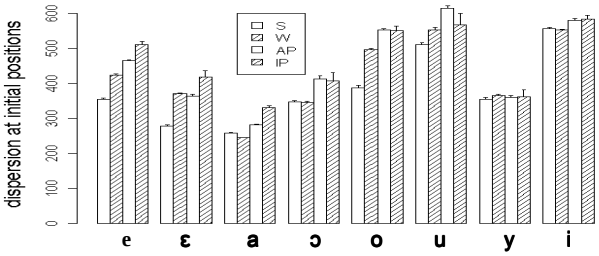


Figure 3 : Dispersion en position initiale en fonction de la hiérarchie prosodique (S: syllabe, W: mot, AP: syntagme accentuel, et IP: syntagme intonatif.

Des ANOVAS à 1 facteur (catégorie) ont été effectuées pour déterminer les différences significatives pour chaque voyelle. Seul /e/ affiche des différences significatives pour chaque niveau. Les voyelles /ɛ/, /a/, /o/ et /u/ montrent une dispersion significative entre trois niveaux. Pour /ɔ/ et /i/, seuls deux niveaux peuvent être distingués significativement, dans les deux cas, les niveaux syllabe/mot vs. syntagme accentuel/intonatif. Seul /y/ ne révèle aucune tendance en fonction de la hiérarchie prosodique.

3.2. Positions finales

Comme observé en 3.1., les voyelles occupent un espace acoustique de plus en plus important en remontant la hiérarchie prosodique pour les positions finales.

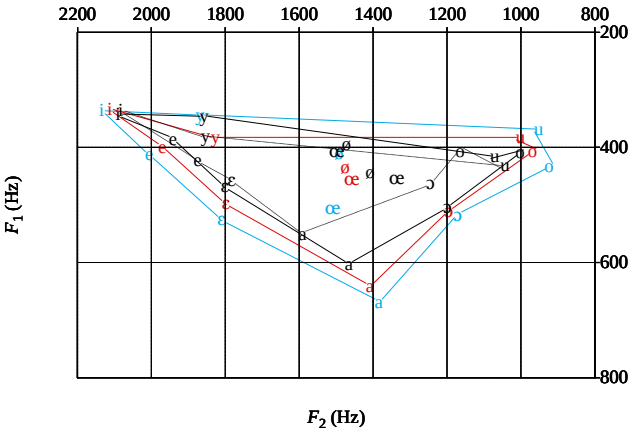


Figure 4 : Comparaison des 4 catégories prosodiques en position finale. De l'intérieur vers l'extérieur : syllabe, mot, syntagme accentuel et syntagme intonatif.

Les mesures de dispersion présentées par la figure 5 révèlent également des valeurs qui augmentent globalement avec le niveau de la hiérarchie prosodique. Les voyelles /e/, /a/, /o/ and /i/ montrent des différences significatives pour tous les niveaux, malgré une amplitude moins importante pour /i/. Les voyelles /ɛ/, /ɔ/ et /u/ révèlent une dispersion significativement plus importante pour trois niveaux. Pour /y/ encore, les variations sont soit inattendues, soit non significatives.

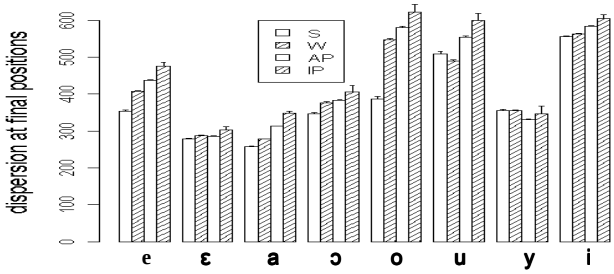


Figure 5 : Dispersion en position finale en fonction de la hiérarchie prosodique (S: syllabe, W: mot, AP: syntagme accentuel, et IP: syntagme intonatif.

4. DISCUSSION ET CONCLUSION

Comme nous en avions fait l'hypothèse, nous observons une hiérarchie prosodique sur la base de mesures spectrales à partir des catégories prosodiques que nous avons déterminées. A la différence des travaux effectués sur de la parole lue, les catégories sont déterminées à postériori sur le continuum de parole, et cette procédure nécessitera à l'avenir une réflexion méthodologique : combien de catégories significativement différentes pourraient ainsi se distinguer statistiquement ? Comme signalé par les études précédentes dans ce domaine (Fougeron 2001; Tabain 2003, pour le français), tous les niveaux ne peuvent pas être distingués de façon systématique. Par exemple nous avons noté que /i/ et surtout /y/ étaient moins variables sur l'axe F1/F2. Gendrot et al. (2008) avaient précédemment observé ce même résultat en fonction des variations de durée, en indiquant que les variations spectrales étaient plus importantes sur un plan F3/F4. Une analyse des valeurs de F3 et F4 a pu montrer que /y/ est caractérisé par un

rapprochement de F2 et F3 à mesure qu'il s'élève dans la hiérarchie prosodique. /i/ est quant à lui caractérisé par un rapprochement de F3 et F4. Ces rapprochements sont ainsi plus pertinents que les variations chaotiques observées sur F1 et F2 et servent à accroître les caractéristiques focales de ces voyelles (Schwartz et al.,1997).

D'autres voyelles telles que /ɔ/, /o/ et /u/ ont révélé des variations inattendues, parfois non significatives, parfois dans des directions ne correspondant pas aux hypothèses. Ces résultats peuvent être en partie expliqués par une quantité assez restreinte de ces voyelles dans ces positions phonotactiques. Par exemple, nous avons remarqué que /ɔ/ a des variations atypiques dans les positions finales strictes avec des valeurs de F2 particulièrement élevées ; or /ɔ/ apparaît en syllabe fermée ('or', 'corps', 'corpus', etc.) selon les règles phonologiques du français. Finalement, nous pouvons également mentionner qu'une augmentation de f0 comme celle signalée en 2.2 aboutit à une augmentation des valeurs de F1, ce qui pourrait également expliquer pourquoi les voyelles fermées varient peu sur F1 en fonction des positions prosodiques, contrairement à ce que nous avions observé sur la figure 1, i.e. les voyelles fermées renforcées ont traditionnellement des valeurs de F1 de plus en plus basses.

Il convient de se rappeler que les positions initiales chaque de catégorie prosodique correspondent également à la fin d'une catégorie de même niveau. Comme proposé par Byrd & Saltzman (2006), ces frontières sont des moments de ralentissement articulatoire ('pi-gesture') qui favorisent le renforcement. Il est donc normal d'observer les mêmes résultats et phénomènes de renforcement en position initiale et finale de tous nos niveaux prosodiques, et ce quelque soit la catégorie grammaticale impliquée (en effet, les positions initiales de syntagmes accentuels et intonatifs sont majoritairement occupées par des mots grammaticaux, généralement considérés comme hypoarticulés). Un autre résultat intéressant est le fait que pour les positions finales, les voyelles finales mais également pénultièmes ont été considérées pour l'analyse. Par contre, pour les positions finales, seules les voyelles strictement initiales ont été prises en compte. Comme suggéré par Byrd et al. (2006) pour l'anglais américain, l'empan semble plus important sur les positions finales que sur les positions initiales.

Pour finir, nous avons remarqué que les valeurs de durée et de f0 augmentaient parallèlement aux valeurs de dispersion, en remontant la hiérarchie prosodique. Il semblait dans un premier temps que ces paramètres étaient liés puisque la f0 et la durée sont connus pour marquer la présence de frontières. Des mesures de corrélations ont été effectuées mais se sont révélées faibles, ce qui pourrait suggérer que des phénomènes de compensation pourraient exister, au sein de stratégies entre locuteurs, entre des variations spectrales et des variations prosodiques pour marquer la présence de frontières.

BIBLIOGRAPHIE

[1] Boersma, P. & D. Weenink (2009). Praat: doing phonetics by computer (Version 5.1.22) [Computer program]. Retrieved September 15, 2009.

[2] Bradlow, A.R., G.M. Torreta & D.B. Pisoni (1996). Intelligibility of normal speech: global and finegrained acoustic-phonetic characteristics. *Speech Communication 20*, pp. 255-272.

[3] Byrd, D., J. Krivokapic & S. Lee (2006). How far, how long: On the temporal scope of phrase boundary effects. *J. Acoust. Soc. Am. 120*, pp. 1589-1599.

[4] Clément, L., B. Sagot & B. Lang (2004). Morphology based automatic acquisition of large-coverage lexica. In *proc. of LREC'04*, Lisboa, Portugal, pp. 1841–1844.

[5] Fougeron C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics 29:2*, pp. 109-135.

[6] Galliano, S., E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre & G. Gravier (2005). ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proc. of Interspeech*, Lisboa, Portugal, pp. 1149-1152.

[7] Gendrot, C. & M. Adda-Decker (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In *Proc. of Interspeech* – Lisboa, Portugal, pp. 2453-2456.

[8] Gendrot, C., M. Adda-Decker & J. Vaissière (2008). Les voyelles /i/ et /y/ du français : focalisation et variations formantiques. In *proc. of 26èmes Journée d'Etude de la Parole*, Avignon, France, pp.205-208.

[9] Lindblom B. (1963). Spectrogaphic study of vowel reduction. *J. Acoust. Soc. Am. 35*, pp.1773-1781.

[10] Nespor M. & I. Vogel (1986). *Prosodic phonology*, Dordrecht : Foris.

[11] Pallier, C. (1994). Rôle de la syllabe dans la perception de la parole : études attentionnelles. *PhD thesis*, EHESS, Paris.

[12] Schwartz J. L., L. J. Boë, N. Vallée & C. Abry (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics 25*, pp. 255-286.

[13] Stevens, K.N. (1997) Articulatory-acoustic-auditory relationships. In W.J. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences* , Blackwell: Oxford, pp. 462-506.

[14] Tabain M. (2003). Effects of prosodic boundary on /aC/ sequences: acoustic results. *J. Acoust. Soc. Am.. 113*, pp. 516-531.

[15] Tabain M. & Perrier (2005). Articulation and acoustics of /i/ in pre-boundary position in French. *Journal of Phonetics*, 33, pp. 77-100.

[16] Vaissière, J. (1971). Contribution à la synthèse par régles du français. Thèse de doctorat, Université de Grenoble.

L'effet du contenu narratif sur la focalisation dans les gestes iconiques d'enfants âgés de 9 à 11 ans

Djaber FANTAZI, Jean-Marc COLLETTA***
*Laboratoire LIDILEM, université de Grenoble, ** Laboratoire LIDILEM, IUFM et université de Grenoble
fantazi_djaber@yahoo.fr ; jean-marc.colletta@u-grenoble3.fr

ABSTRACT

This study shows how the narrative content influences the production of iconic gestures in a narrative task. We analyzed 24 oral narratives produced by middle-grade children aged 9-11 years. Half of the narratives was elicited after showing a short extract of a Tom & Jerry cartoon presenting a rapid flow of actions; the other half was elicited after showing a short extract of a Wallace & Gromit animated movie mainly presenting static situations. All narratives were analysed using *ELAN* as an annotation tool. The study results show the effect of the narrative content on iconic forms of gesture (*character's viewpoint* vs *observer's viewpoint* gestures) and add support to the thesis of speech as a multimodal process.

1. L'EFFET DU CONTENU NARRATIF SUR LA PRODUCTION DES GESTES ICONIQUES

Raconter est parmi les tâches discursives les plus complexes car d'une part il faut se rappeler les événements pertinents de l'histoire et les transformer en une suite de propositions, et d'autre part il faut maintenir la cohésion entre les propositions successives et, dans le même temps, introduire une progression dans les informations pour susciter et maintenir l'intérêt [7,8]. Ainsi, si rapporter des événements fait partie des pratiques langagières courantes pour l'adulte, c'est loin d'être une mince affaire pour l'enfant puisque, d'un point de vue acquisitionnel, celui-ci, même âgé de neuf-dix ans, n'a pas fini d'acquérir les structures prototypiques du récit (le schéma narratif et ses composantes) lui permettant de restituer le plus fidèlement possible les événements d'une histoire. En plus, d'un point de vue développemental, nous savons depuis les travaux de Piaget que les capacités de décontextualisation et de décentration qui sous-tendent les capacités discursives sont encore en devenir à cet âge. Cependant, les raisons développementales ne sont pas les seules en cause et d'autres éléments comme le mode de production peuvent influer sur la structuration du récit [18], ou son organisation syntaxique et gestuelle [2]. Si l'on se tourne du coté de la tâche narrative, celle-ci est considérée comme une variable puissante puisque elle conditionne, en quelque sorte, la production langagière de l'enfant comme d'ailleurs celle de l'adulte. Dans sa conception, la tâche narrative prédétermine l'organisation lexicale et syntaxique des énoncés [18, 9] et *in fine* la construction thématique [16] et référentielle [17]. Le travail de Vion et Colas [17] montre par exemple que la présentation de séquences d'images successivement ou en bloc influe quantitativement et qualitativement sur l'introduction et le maintien de la référence (objets de la narration). Mais, la plupart de ces études ne prennent en compte que la dimension verbale de la production langagière, sans se préoccuper des dimensions vocale et gestuelle. Or, compte tenu de la thèse du traitement multimodal du langage en production [10,12,11], on sait que les signifiants kinésiques sont intimement liés aux signifiants verbaux, et qu'ensemble, ils donnent naissance à un énoncé « total » [6] composé de gestes, paroles et vocalisations. En conséquence, si la tâche narrative influe sur la dimension verbale du langage oral, on doit s'attendre à ce qu'elle influe aussi sur la dimension gestuelle de celui-ci.

Une étude de Colas, Lo Giudice et Vion [1], portant sur la seule dimension gestuelle dans la communication référentielle selon que le destinataire est entendant ou sourd fait apparaître, en plus de l'ajustement gestuel au destinataire, un rôle important des caractéristiques du référent (description d'éléments géométriques) dans la production des gestes, notamment iconiques. D'autres études, privilégiant une entrée multimodale qui intègre paroles et gestualité, se révèlent intéressantes à cet égard, que ce soit dans l'étude des communautés sourdes [14, 15] ou chez les entendants [12]. L'analyse de McNeill [12], portant sur des récits enfantins montre que les gestes référentiels, notamment les gestes iconiques (gestes représentant des objets, personnages, lieux et actions, par opposition aux gestes de l'abstrait, qui représentent des idées), participent à la construction référentielle en donnant une indication sur la focalisation ou le point de vue adopté : le narrateur peut adopter le point de vue interne du personnage *en* mimant des actions impliquant un objet (saisir, porter, jeter, avaler un objet, etc.), et donc utiliser des gestes *C-VPT (character's viewpoint)* pour (se) représenter ces actions ; il peut aussi adopter le point de vue d'un observateur extérieur en prenant de la distance par rapport au personnage et exécute alors des gestes *O-VPT (observer's viewpoint).* Ce choix de codage a été également observé chez les sourds [14, 15]. L'identification du narrateur avec le personnage de l'histoire laisse penser que le contenu narratif imagé peut influencer la production gestuelle du langage et notamment de la gestualité iconique, puisque celle-ci, en vertu de la thèse défendue par McNeill et d'autres auteurs [13], est constituée de mouvements corporels exprimant des images mentales, c'est-à-dire des représentations d'objets concrets ou abstraits.

L'objectif de la présente étude est de disposer d'observations complémentaires sur la gestualité iconique dans les récits enfantins en contexte francophone. Plus précisément, il s'agit de voir jusqu'à quel niveau la focalisation (acteur *versus* observateur) dans la gestualité iconique est sensible au contenu narratif selon que celui-ci est axé sur un rappel d'actions et d'événements ou sur un rappel de situations et de descriptions. Nous posons l'hypothèse que l'enfant narrateur devant procéder à un rappel de récit riche en actions s'identifie au personnage

de l'histoire en reproduisant ses attitudes et en revanche, qu'il prend de la distance par rapport au personnage dans le rappel de récits riche en situations statiques en sollicitant une posturo-mimo-gestualité moins impliquée. Concrètement nous nous attendons à trouver plus de gestes *C-VPT* dans les récits d'actions, et inversement plus de gestes *O-VPT* dans les récits de situations.

Par ailleurs, si l'on se tourne du coté des liens entre les éléments verbaux et les éléments non verbaux, comme le rappellent Colletta et Millet [4], le choix de la focalisation dans les gestes iconiques n'est pas sans incidence sur l'organisation des éléments linguistiques telle que la syntaxe. D'après McNeill [12], il semble que le geste à point de vue interne accompagne plutôt des énoncés brefs et des constructions transitives, tandis que le geste à point de vue externe accompagne plutôt des énoncés longs et des constructions intransitives. A cet égard, il sera particulièrement intéressant d'observer si on retrouve ce résultat dans nos données.

2. METHODOLOGIE

Le corpus est constitué de 24 récits oraux produits par une population de 24 enfants de CM1/CM2 issus de familles de classe moyenne de l'agglomération grenobloise. Au préalable, chaque enfant a passé un test d'évaluation du langage oral *ELO*. Ce test a été administré dans le but de contrôler les capacités lexicales et syntaxiques de l'enfant et de disposer ainsi d'une mesure précise permettant d'écarter de l'étude les enfants dont le développement langagier est atypique. Ensuite, après avoir visionné un clip vidéo de moins de 3 mn (extrait d'un épisode de *Tom et Jerry* pour les uns, extrait d'une aventure de *Wallace et Gromit* pour les autres) présentant une mini-histoire complète avec début et fin, chaque enfant a produit un récit oral monologique en présence d'un interlocuteur.

Bien que ces deux mini-histoires partagent la même structure narrative (présentation, complication, résolution), elles s'opposent essentiellement par leur contenu narratif (récit d'actions *vs* récit de situations) ; on trouve dans le premier (RT&J) un enchaînement rapide d'actions et d'événements tandis que le second (RW&G) se présente comme un enchaînement de situations comportant de nombreux éléments de descriptions et appelant la formulation d'hypothèses de la part du narrateur.

Les récits filmés des 24 enfants ont été transcrits. Nous avons procédé à un codage des aspects syntaxique, narratif et gestuel de ces productions à l'aide du logiciel *ELAN* qui permet, grâce à l'alignement des sources vidéo et audio, un croisement fin des paroles et des mouvements corporels. Comme la mimo-gestualité représentationnelle s'étend à tous les signaux de la posturo-mimo-gestualité, nous avons analysé non seulement les gestes manuels, mais aussi les gestes céphaliques et les changements de postures. Les gestes produits durant le récit ont été catégorisés selon la grille fonctionnelle de Colletta et al. [3], qui a été élaborée à l'occasion du projet ANR 0178-08 « Multimodalité». Afin de disposer des valeurs et formes objectives des gestes, plusieurs annotateurs, dont un juge, tous préalablement formés à l'annotation gestuelle, ont été sollicités.

3. RESULTATS ET DISCUSSION

Sur un total de 216 gestes coverbaux dans RT&J, on dénombre 132 gestes représentationnels, dont 55 gestes iconiques codant des événements, des actions ou des déplacements des personnages. Parallèlement, sur un total de 126 gestes coverbaux dans RW&G, on dénombre 61 gestes représentationnels, dont 43 gestes iconiques. Dans nos données, environ un geste sur deux ou plus est un geste représentationnel, ce qui va dans le sens des résultats trouvés sur l'ensemble du corpus recueilli à l'occasion du projet ANR susmentionné [5]. S'agissant des gestes iconiques, entre un tiers et un quart des gestes produits à l'occasion de ces récits sont donc des gestes qu'on peut coder comme *C-VPT* ou O-*VPT*.

A partir de ce premier constat, il se dégage à l'évidence qu'il y a plus de gestes iconiques dans les récits de RT&J (55 gestes) que dans ceux de RW&G (43 gestes). Or, étant donné que la longueur des récits varie de façon importante (le nombre moyen de propositions des récits RT&J est plus petit que celui des récits RW&G, cf. tableau 1), nous sommes amenés à comparer la production gestuelle en tenant compte de la production verbale. Le moyen le plus communément utilisé est de diviser le nombre de gestes par le nombre de propositions pour obtenir un « taux gestuel ». Voir les résultats dans le tableau 1 ci-dessous :

Tableau 1 : Taux gestuel moyen dans les deux tâches narratives:

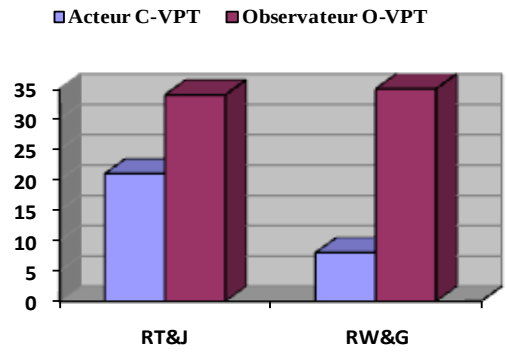
	RT&J	RW&G
Nombre moyen de propositions	42	50
Nombre moyen de gestes iconiques	4.58	3.58
Taux gestuel moyen	0.11	0.07

La comparaison des moyennes des taux gestuels montre qu'effectivement il y a plus de gestes iconiques dans le récit d'action RT&J que dans le récit de situation RW&G ($t = 3.31$, $p < .005$). Cette différence peut peut-être s'expliquer à la lumière des travaux de McNeill [12] : le contenu imagé riche en actions de RT&J a peut-être incité les sujets à produire davantage de gestes lors du rappel du récit.

À présent, nous savons que le taux gestuel dans RT&J est supérieur à celui de RW&G. Mais qu'en est-il de la répartition du point de vue exprimé gestuellement dans ces deux tâches narratives ? A la lueur de nos résultats, il apparaît que les gestes iconiques dont la focalisation est interne (le narrateur se positionne comme acteur) sont produits davantage dans le récit riche en actions (RT&J) que dans le récit présentant des situations (RW&G). Par contre, les gestes iconiques dont la focalisation est externe (le narrateur se positionne comme observateur) sont produits de façon quasi similaire dans les deux types de récits.

Le test χ^2 montre que les répartitions sont significativement différentes, i.e. qu'un lien existe entre la tâche narrative et la focalisation gestuelle ($\chi^2 = 4.38$,

p<.05), même si ce lien est visible davantage pour la focalisation interne *C-VPT* que pour la focalisation externe *O-VPT* ; comme l'illustre le graphe 1 ci-dessous :



Graphique 1 : Répartition de la focalisation des gestes iconiques selon la tâche narrative.

Dans notre étude, le contenu narratif de RT&J présente de très nombreuses actions (coudre, border, donner, avaler, porter, etc., quelque chose) ce qui explique la part importante des gestes *C-VPT*. Mais on y trouve aussi des localisations et des déplacements de personnages (chute, roulade, entrée, sortie, poursuite, etc.), autant d'actions non transitives qu'il faut aussi (se) représenter, d'où l'emploi important également de gestes *O-VPT*. En revanche, RW&G présente un contenu narratif fait de situations stables, avec peu d'actions, davantage d'éléments de décor à décrire (intérieur d'une maison, machine, etc.), incorporant une composante mystérieuse et appelant un questionnement, des hypothèses, et en conséquence, la part des gestes *C-VPT* y est moins importante.

Pour McNeill, il est plus aisé de mimer des actions impliquant un objet (saisir, porter, hisser, jeter un objet, le frapper, etc.) et donc d'utiliser un geste *C-VPT* pour représenter de telles actions, et par contre, l'utilisation de gestes *O-VPT* convient mieux pour représenter des actions n'impliquant pas d'objet tel, par exemple, le déplacement d'un personnage. De façon plus précise et au vu de la grille de catégorisation des coverbaux adoptée par Colletta, [2], les gestes iconiques rencontrés dans le corpus peuvent être rangés dans les catégories suivantes : les gestes mimétiques représentant les mouvements d'actions et les attitudes des personnages sont à l'évidence des *C-VPT* tandis que les gestes locatifs représentant le ou les référents dans l'espace, leur localisation et leurs déplacements, sont très majoritairement des *O-VPT*. Quant aux gestes illustratifs, ils peuvent servir l'un ou l'autre des deux points de vue selon que le référent dont ils dessinent la forme, les dimensions, la texture ou l'aspect est un référent animé ou non. Il semble que comme les gestes locatifs, les illustratifs représentant des référents inanimés codent davantage le point de vue externe (le narrateur se positionne comme observateur) tandis que les illustratifs représentant des référents animés, tout comme les gestes mimétiques, codent plutôt le point de vue interne (quand le narrateur se positionne comme acteur). Le tableau 2 présente la répartition de ces sous-catégories de gestes iconiques en fonction du contenu narratif et du point de vue.

Tableau 2 : Répartition des gestes iconiques selon la tâche narrative et le point de vue.

		Focalisation gestuelle			
		Acteur *C-VPT*		Observateur *O-VPT*	
		Mimétique	Illustratif		Locatif
Contenu narratif	Action RT&J	8	13	10	24
	Situation RW&G	4	4	21	14
		Référence « animée »		Référence « spatiale »	

Au regard de leur forme, ce qui distingue ces deux genres de focalisation, c'est notamment l'utilisation des ressources corporelles expressives : les gestes *C-VPT* mobilisent la posture, le mime en plus du geste manuel, et c'est souvent tout le corps qui se met en mouvement pour reproduire le référent. En contrepartie, la focalisation *O-VPT* emploie des gestes de moindre amplitude, sans implication du buste, de la posture ou des mimiques faciales. On y trouve certes de brefs mouvements céphaliques et/ou manuels mais beaucoup de gestes de pointage abstrait qui localisent le référent dans l'espace frontal du locuteur et représentent ses déplacements à l'intérieur de la scène représentée par cet espace.

Concernant la deuxième hypothèse et comme on peut s'y attendre, la focalisation (interne *vs* externe) va de pair avec les formes de constructions verbales. Après avoir groupé les deux tâches narratives (RT&J+RW&G), l'étude de liaison (test de χ^2) révèle que les gestes *C-VPT* accompagnent plutôt des constructions transitives (i.e. incluant des compléments COD ou COI) tandis que les gestes *O-VPT* accompagnent plutôt des constructions non transitives (i.e. sans complément) (χ^2=8.59, p<.01). Les résultats sont présentés dans le tableau 3.

Tableau 3 : Répartition de gestes iconiques selon la transitivité des constructions verbales.

	Focalisation gestuelle	
	Acteur *C-VPT*	Observateur *O-VPT*
Constructions transitives	22	30
Constructions non transitives	7	39

Ces résultats sont également en accord avec les observations de McNeill [12] et confirment l'incidence du contenu narratif tant sur les aspects linguistiques du récit que sur la gestualité coverbale. Dans notre étude, par exemple, nos jeunes locuteurs combinent volontiers construction transitive et geste *C-VPT* lorsqu'ils miment l'action d'un personnage sur un objet (faire un tricot, consulter l'heure, manger quelque chose, porter quelque-chose, etc.), et plus rarement construction

intransitive et mime d'action (*C-VPT*). L'effet est moins net lorsqu'on examine la relation entre gestualité *O-VPT* et type de construction syntaxique puisque la gestualité *O-VPT* accompagne à la fois construction intransitive (combinaison bien adaptée à l'évocation de localisations et de déplacements) et aussi construction transitive lorsqu'il s'agit, par exemple, d'évoquer des états mentaux (voir, s'interroger sur, soupçonner quelque chose, etc.).

5. CONCLUSION ET PERSPECTIVE

Au terme de ce travail, il ressort que le contenu sémiotique présenté dans une tâche langagière (ici narrative) influence non seulement la production linguistique du locuteur, comme cela a été montré dans de nombreuses recherches, mais aussi sa production gestuelle. Ce résultat plaide en faveur de la thèse du traitement multimodal de la parole en production puisqu'il met en relief de subtiles relations entre les codages linguistiques et les codages gestuels, codages liés à l'expression du point de vue. La production gestuelle accompagnant la parole est indissociable de la production verbale, et ce vaut aussi bien pour l'enfant que pour l'adulte. Au plan méthodologique, la grande sensibilité de ces codages à un paramètre aussi fin que celui du contenu narratif conduit à penser que, dans toute approche comparative, l'identité de la tâche pour tous les sujets est seule apte à offrir une garantie efficace.

Enfin, si dans cet article nous avons traité des liens entre la stratégie narrative (le point de vue), la gestualité iconique et l'organisation syntaxique, d'autres restent à étudier, notamment entre la gestualité iconique et l'organisation du récit, la cohésion narrative et la continuité référentielle, ou encore, la focalisation gestuelle et le rappel des événements centraux *vs* périphériques. Ces nouvelles pistes sont de nature à nous permettre de mieux cerner les relations entre développement discursif, développement gestuel et développement cognitif, et feront l'objet d'investigations complémentaires.

4. BIBLIOGRAPHIE

[1] A. Colas, N. Lo Giudice et M. Vion. Evolution de l'ajustement gestuel au destinataire (sourd/ entendant) en tâche de communication référentielle. In S., Santi et al. Eds. *Oralité et gestualité*. Communication multimodale, interaction. Paris, l'Harmattan. 259-265, 1999.

[2] J.-M. Colletta. *Le développement de la parole chez l'enfant âgé de 6 à 11 ans*. Belgique, Mardaga, 2004.

[3] J.-M. Colletta, R. Kunene, A. Venouil, V. Kaufmann, et J.-P. Simon. Multitrack annotation of child language and gestures, in M. Kipp (Ed.), *Multimodal Corpora*, LNAI, Springer. 2009.

[4] J.-M. Colletta et A. Millet. Introduction. *Lidil*, 26, 7-26. Grenoble: PUG. 2002.

[5] J.-M. Colletta, C. Pellenq et M. Guidetti. Age-related changes in co-speech gesture and narrative: Evidence from French children and adults. Accepté.

[6] J. Cosnier et A. Brossard. *La communication non verbale*. Neuchâtel-Paris, Delachaux & Niestlé. 1984.

[7] M. Fayol. *Le récit et sa construction*. Neuchâtel-Paris, Delachaux & Niestlé. 1985.

[8] M. Fayol. *Des idées au texte*. Paris, PUF. 1997.

[9] F. Gayraud, S. Gonnand, S. Kern et A. Viguié. L'effet de différentes tâches narratives sur la connexion dans des textes d'enfants francophones de 10ans. *Acquisition des langues*, Berder, 22-25 mars 1999.

[10] A. Kendon. Some relationships between body motion and speech. In A.W. Siegman et B. Pope, Eds, *Studies in dyadic communication*: 177-210, 1972.

[11] S. Kita. How representational gestures help speaking, In D. McNeill (Ed.), *Language and Gesture*. Cambridge, Cambridge University Press, pp: 162-185, 2000.

[12] D. McNeill. *Hand and mind*. What gestures reveal about thought. Chicago, University of Chicago Press. 1992.

[13] D. McNeill. Ed. *Language and gesture*. Cambridge, Cambridge University Press. 2000.

[14] A. Millet. Les dynamiques iconiques et corporelles en langue des signes française (LSF). *Lidil*, 26, 27-44. Grenoble: PUG. 2002.

[15] A. Millet et I. Estève. Contacts de langues et multimodalité chez des locuteurs sourds : concepts et outils méthodologiques pour l'analyse. *Journal of Language Contact*. Varia2, 111-133. 2009.

[16] M. Vion et A. Colas. Contrôle de la production d'informations nouvelles et anciennes par des enfants âgés de 4 à 11 ans : les constructions présentatives. *Bulletin d'autophonologie, V3*, 6, 671-686. 1987.

[17] M. Vion et A. Colas. L'introduction des référents dans le discours en français :contraintes cognitives et développement des compétences narratives. *L'année psychologique*, 98, 37-59. 1998.

[18] G. Weck (de). *La cohésion dans les textes d'enfants*. Etude du développement des processus anaphoriques. Neuchâtel, Delachaux et Niestlé. 1991.

Quantificateur vectoriel divisé à commutation SSVQ appliqué au codage des paramètres LPC du codeur MELP de 2.4 Kbits/s

Merouane BOUZID, Salah Eddine CHERAITIA, Moussa HIRECHE

Laboratoire Communication Parlée et Traitement du Signal (LCPTS)
Faculté d'Electronique et d'Informatique, Université *USTHB*, BP 32, El-Alia,
Bab-Ezzouar, ALGER, 16111, ALGERIE. Email: mbouzid@usthb.dz

ABSTRACT

In this paper, we present an optimized switched split vector quantization (SSVQ) scheme developed for low bit-rate encoding of the LPC parameters represented by the line spectral frequencies (LSF). It will be shown that the SSVQ provides better performances in terms of bit-rate, spectral distortion and computational complexity than the traditional split vector quantizer. We further applied the SSVQ system, called LSF-SSVQ encoder, to quantize the LSF parameters of the narrowband speech coder MELP of 2.4 Kbits/s, operating oven an ideal noiseless channel.

Keywords: Split vector quantizer, Switched SVQ, Speech coding, LSF parameters, MELP speech coder.

1. INTRODUCTION

Un des problèmes importants dans le codage de la parole à bas débit est la conception de quantificateurs efficaces pour le codage des coefficients de prédiction linéaire (LPC) du filtre de synthèse dont la fonction de transfert est donnée par $H(z) = 1/A(z)$, avec $A(z) = 1 + a_1 z^{-1} + \ldots + a_{10} z^{-10}$ [1]. Les 10 coefficients LPC $\{a_i\}_{i=1,2,\ldots,10}$ de ce filtre sont dérivés du signal d'entrée en utilisant une analyse par prédiction linéaire (LP) sur chaque trame du signal de parole. En pratique, ces coefficients ne sont pas appropriés pour une transmission directe vu qu'ils ont des propriétés de quantification médiocres. Ainsi, plusieurs représentations paramétriques équivalentes ont été formulées afin de les convertir en paramètres beaucoup plus appropriés à la quantification. Parmi les représentations qui se sont avérées plus efficaces, les fréquences de raies spectrales LSF (Line Spectral Frequencies) sont sans doute les plus utilisées [2].

Les paramètres LSF, qui sont liés aux zéros de polynômes dérivés de $A(z)$, présentent un certain nombre de propriétés intéressantes [1], [3]. Exploitant ces propriétés, divers schémas de codage basés sur la quantification scalaire et vectorielle ont été suggérés pour la quantification efficace des paramètres LSF. Cependant, plusieurs travaux ont démontrés que les schémas conçus à base de quantificateurs vectoriels (VQs) structurés, comme les VQs multi-étages MSVQ [4], les VQs divisés SVQ [3], les VQs codés par treillis TCVQ [5]…, peuvent réaliser une quantification de qualité transparente des paramètres LSF à des débits binaires nettement plus bas comparés à ceux conçus à base de quantificateurs scalaires (SQs).

Dans cet article, nous présentons un système de codage à base de quantification vectorielle divisée à commutation SSVQ (Switched Split Vector Quantization). Ce système, que nous avons appelé "Encodeur LSF-SSVQ", a été conçu pour le codage efficace à bas débit des paramètres LSF du codeur de parole normalisé MELP de 2.4 Kbits/s. Nous montrerons que l'encodeur LSF-SSVQ, incorporé dans le MELP, présente des performances comparables à celles de l'encodeur-LSF original du MELP avec en plus un gain en débit.

2. QUANTIFICATION VECTORIELLE DIVISEE À COMMUTATION

Le quantificateur vectoriel divisé à commutation SSVQ (Switched Split Vector Quantizer) est un schéma de codage hybride conçu à base d'un VQ à commutation combiné avec plusieurs quantificateurs vectoriels divisés SVQ (Split Vector Quantizer).

Rappelons qu'un SVQ de N parties (noté N-SVQ) est composé de N VQs classiques de tailles et de dimensions plus petites [3]. Son principe de base consiste à partitionner l'ensemble de tous les vecteurs x de dimension k de la base d'apprentissage en N sous-ensembles composés de sous-vecteurs de dimension k_i (avec $\sum_{i=1}^{N} k_i = k$). Ensuite, pour chaque partie, le dictionnaire VQ correspondant sera conçu en utilisant l'algorithme LBG-VQ conventionnel [6].

Comparé à un VQ conventionnel de dimension k, de débit R bits/échantillon (bpe) et de taille $L = 2^{Rk}$, un N-SVQ est constitué donc de N dictionnaires de tailles $L_i = 2^{R_i k_i}$ (où $L = \prod_{i=1}^{N} L_i$ et R_i est le débit partiel). Quantifier un vecteur-source d'entrée par un SVQ revient donc à décomposer ce vecteur en N sous-vecteurs de dimensions plus petites qui seront par la suite quantifiés séparément en utilisant les dictionnaires des parties correspondantes.

2.1. Principe de conception du SSVQ

Le principe de base du SSVQ consiste à diviser l'espace des vecteurs de la base d'apprentissage en plusieurs parties (division), où chaque partie est représentée par un quantificateur SVQ local approprié [7], [8].

La figure 1 présente le schéma bloc du principe de construction du dictionnaire SSVQ. La première étape consiste à appliquer l'algorithme LBG-VQ sur toute la base d'apprentissage afin de produire m représentants (vecteurs-code). L'ensemble de ces vecteurs-code est appelé dictionnaire VQ commutateur Y_m où m représente le nombre de direction de commutation. Ensuite, ce dictionnaire sera utilisé pour partager la base d'apprentissage en m classes suivant le critère du voisin le plus proche. C'est-à-dire que chaque vecteur de la base d'apprentissage est comparé avec les m représentants puis envoyé dans la direction de son plus proche voisin. A la fin de cette étape, on aura donc m classes correspondantes à m directions de commutation.

Dans la deuxième étape, chaque classe sera représentée par un N-SVQ local. Il s'agit donc de diviser les vecteurs de chaque classe en N sous-vecteurs puis appliquer l'algorithme LBG-VQ sur chaque ensemble de sous vecteurs afin de produire les N dictionnaires locaux correspondants. A la fin de la conception, on obtient alors (N m + 1) dictionnaires; le premier est celui du VQ commutateur, et les autres N m dictionnaires sont ceux des m N-SVQ locaux.

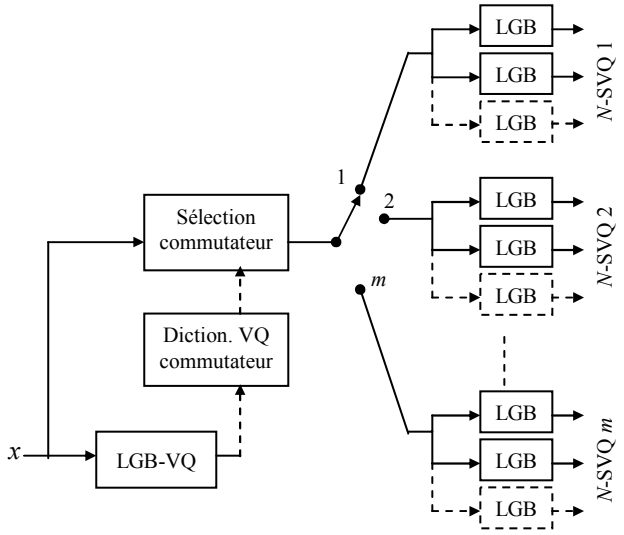


Figure 1 : Construction du dictionnaire SSVQ

2.2. Codage/décodage SSVQ

Le codage SSVQ d'un vecteur source d'entrée x passe par deux étapes. Au début, le vecteur x est commuté vers une des m directions possibles suivant le critère du voisin le plus proche. Par la suite, ce vecteur sera quantifié par le N-SVQ correspondant qui a été sélectionné par le VQ commutateur Y_m.

Ainsi, l'encodeur SSVQ fournit au décodeur un indice i composé de N + 1 indices concaténés. Le premier indice i_s (s = 1,…, m) est celui du vecteur-code le plus proche de x parmi les m vecteurs-code du dictionnaire Y_m. Les N indices i_n (n = 1,…, N) suivants seront fournit par le SVQ local correspondant à la direction i_s. Il s'agit des N indices des N sous-vecteurs de x codé séparément par le N-SVQ de la partie sélectionnée par l'indice i_s.

Le décodeur SSVQ, qui possède les mêmes dictionnaires que ceux de l'encodeur, reçoit l'indice envoyé $i = (i_s\ i_n)$ avec n = 1,…, N. Il utilise le premier indice i_s pour sélectionner la direction de commutation. Ensuite, il construit le vecteur décodé de x en utilisant les vecteurs-code d'indices i_n correspondants au N-SVQ de la partie i_s. Dans ce travail, on suppose que les transmissions s'effectuent à travers un canal idéal non-bruité.

3. CODAGE EFFICACE DES PARAMETRES LSF PAR LA METHODE SSVQ

En utilisant la technique de quantification SSVQ, un schéma d'encodage, appelé "Encodeur LSF-SSVQ", a été conçu pour réaliser une quantification efficace de qualité transparente des paramètres spectraux LSF.

Pour un codage de la parole en bande étroite téléphonique, l'ordre de prédiction est souvent fixé à 10. Ainsi, la taille d'un vecteur LSF à coder est de 10. Dans la conception des N-SVQ locaux de nos encodeurs LSF-SSVQ, le vecteur LSF de dimension 10 est divisé en deux parties avec la division (4 – 6). Concernant l'allocation des bits, ces derniers sont uniformément alloués aux sous-parties individuelles d'une même division [7], [8]; ceci bien sûr dans la mesure du possible.

Afin d'améliorer davantage les performances des encodeurs-LSF et obtenir une quantification transparente à des débits plus bas, une mesure de distance plus appropriée a été utilisée. Il s'agit de la mesure de distance euclidienne pondérée qui permet de mettre en évidence des parties particulières du spectre (c.-à-d., les régions formantiques, les basses fréquences,…) en pondérant les LSFs de ces parties plus que les autres. La mesure de distance pondérée, que nous avons utilisée dans ce travail, est donnée par la formule [5] :

$$d(f,\hat{f}) = \sum_{i=1}^{10} c_i w_i (f_i - \hat{f}_i)^2 \ , \qquad (1)$$

où f_i et $\hat{f}_i$ sont respectivement les $i^{\text{ème}}$ coefficients des vecteurs LSF original f et quantifié $\hat{f}$; c_i et w_i représentent respectivement les poids fixe et variable assignés au $i^{\text{ème}}$ coefficient du vecteur LSF.

Nous présentons, ci-dessous, les performances de nos encodeurs LSF-SSVQ opérant à différents débits de codage. Les performances sont évaluées ici par la distorsion spectrale SD (Spectral Distorsion) moyenne qui est souvent utilisée comme mesure objective de la performance d'encodage des paramètres LSF. Calculée discrètement sur une largeur de bande limitée, l'expression de la SD pour une trame i est donnée en décibels par la formule suivante [4], [5]:

$$SD_i = \sqrt{\frac{1}{n_1 - n_0} \sum_{n=n_0}^{n_1-1} \left[10 \log_{10} \frac{S(e^{j2\pi n/N})}{\hat{S}(e^{j2\pi n/N})} \right]^2}, \qquad (2)$$

où $S(e^{j2\pi n/N})$ et $\hat{S}(e^{j2\pi n/N})$ représentent respectivement les spectres de puissance original et quantifié du filtre de synthèse LPC de la $i^{ème}$ trame du signal de parole.

En général, une quantification de qualité transparente est obtenue si les trois conditions suivantes sont maintenues [3]: **1)-** la distorsion spectrale (SD) moyenne est d'environ 1 dB, **2)-** le pourcentage des trames externes (outliers frames) ayant une SD entre 2 et 4 dB est moins de 2% et **3)-** aucune trame "outliers" ne doit avoir une SD qui dépasse les 4 dB.

La base de données des vecteurs LSF utilisée dans ce travail a été construite à partir d'environ 85 minutes de parole prise de la base de données internationale TIMIT (f_e de 16 kHz) [9]. Les signaux de parole sont d'abord filtrés passe-bas à une fréquence de coupure de 3.4 KHz puis sous échantillonnés à 8 KHz. On a utilisé la même fonction d'analyse par prédiction linéaire (LP) du standard fédéral FS1016 de 4.8 Kbits/s [10] où une analyse LP d'ordre 10 par la méthode d'autocorrélation est effectuée sur chaque trame de 30 ms (pondéré par la fenêtre de Hamming). Une partie de la base (144984 vecteurs LSF) est utilisée pour l'apprentissage et l'autre partie, de 26560 vecteurs LSF (différente de la base d'apprentissage), est utilisée pour les tests.

Pour différents débits de codage b et de directions de commutation $m = 2^{bs}$, les performances d'un exemple d'encodeur LSF-SSVQ utilisant 2 parties (4 – 6) sont données dans la Table 1.

Table 1 : Performances de l'encodeur LSF-SSVQ de 2 parties, avec la division (4 – 6).

m	**Bits/trame** b (b_s+ b_1 + b_2)	**SD Moy. (dB)**	**SD Outliers (%)**	
			2-4 dB	**> 4 dB**
8	**26 (3 + 11 + 12)**	0.98	0.46	0.000
	25 (3 + 11 + 11)	1.06	0.97	0.000
	24 (3 + 10 + 11)	1.10	1.20	0.000
	23 (3 + 10 + 10)	1.18	2.49	0.000
	22 (3 + 9 + 10)	1.22	3.06	0.000
16	**26 (4 + 11 + 11)**	0.98	0.65	0.000
	25 (4 + 10 + 11)	1.03	0.80	0.000
	24 (4 + 10 + 10)	1.11	1.58	0.007
	23 (4 + 9 + 10)	1.15	2.00	0.003
	22 (4 + 9 + 9)	1.23	3.79	0.003
32	**26 (5 + 10+ 11)**	0.95	0.44	0.000
	25 (5 + 10 + 10)	1.03	0.81	0.000
	24 (5 + 9 + 10)	1.08	1.01	0.000
	23 (5 + 9 + 9)	1.16	1.97	0.000
	22 (5 + 8 + 9)	1.21	2.62	0.000

Ces résultats de simulation montrent que les performances des encodeurs LSF-SSVQ, en termes de SD moyenne et de trames "outliers", peuvent être toujours améliorées en augmentant le nombre de commutation m. En effet, un grand nombre de commutation correspond à un nombre de représentant plus grand (c.-à-d. plus de dictionnaires); ce qui explique cette amélioration. L'encodeur LSF-SSVQ à 2 parties (4 – 6) peut réaliser une quantification de qualité transparente à un débit de 23 bits/trame (bpt).

Dans la table 2, nous présentons une évaluation comparative des performances entre un SVQ conventionnel de 2 parties et un encodeur LSF-SSVQ de deux parties avec m = 32 directions.

Table 2 : Performances comparatives entre le 2-SVQ et le 2-SSVQ (m = 32), pour différents débits d'encodage

Débit Bits/trame	**SD Moyenne (dB)**		**SD Outliers 2-4 (%)**	
	2-SVQ	**2-SSVQ** (m = 32)	**2-SVQ**	**2-SSVQ** (m = 32)
26	1.05	0.95	0.82	0.44
25	1.03	1.03	0.95	0.81
24	1.17	1.08	2.11	1.01
23	1.21	1.16	2.63	1.97
22	1.29	1.21	5.05	2.62

Ces résultats montrent que le SSVQ apporte une nette amélioration aux performances d'encodage des paramètres LSF, surtout en termes de trames externe "Outliers". Ce gain apporté par le SSVQ est sans doute le résultat de l'exploitation de la corrélation, à travers toutes les dimensions des vecteurs LSF, par le VQ commutateur initial. En effet, ce dernier utilise toute la dimension des vecteurs sans aucune contrainte de structure à l'inverse du SVQ qui divise les vecteurs LSF dès le début.

D'autre part, il est connu que la complexité de calcul du SSVQ est inférieure à celle du SVQ [7]. Ce qui constitue un autre avantage du SSVQ sur le SVQ. Illustrant cet avantage par un simple exemple de calcul de la complexité de recherche qui désigne le nombre de distance à calculer pour coder un seul vecteur. Considérons un 2-SSVQ à 16 directions de commutation opérant à un débit de 22 bits/trame. En retranchant les 4 bits nécessaires pour coder l'indice de commutation, il restera donc 18 bits à allouer pour chaque SVQ local (9 bits pour chaque partie). Le nombre total de recherches requis par le 2-SSVQ pour quantifier un vecteur est donc égal à $2^4 + 2^9 + 2^9 = 1040$ recherches. Par contre, pour le même débit de 22 bpt, un 2-SVQ conventionnel requière $2^{11} + 2^{11} = 4096$ recherches pour quantifier un vecteur.

4. ENCODEUR LSF-SSVQ APPLIQUÉ AU CODAGE DES LSFS DU CODEUR MELP

Dans cette section nous présentons les performances de l'encodeur LSF-SSVQ (avec distance pondérée) appliqué au codage des paramètres LSF du codeur de parole normalisé MELP [11]. Le MELP (Mixed Excitation Linear Prediction) est un codeur de parole à excitation mixte de débit de 2.4 Kbits/s, développé par le DoD des USA. Suivant la norme du codeur MELP, ses paramètres LSF sont codés à l'origine par un MSVQ de 25 bits/trame.

La base de données utilisée dans les simulations suivante est composée de séquences de parole d'environ 113s extraites aléatoirement de la base de test TIMIT. L'évaluation objective de la qualité de la parole synthétisée par le codeur MELP a été faite suivant le modèle PESQ (*Perceptual Evaluation of Speech Quality*), normalisé à l'UIT-T sous le nom P.862 [12]. Le PESQ est un algorithme qui permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation, aboutissant à une corrélation très proche avec les évaluations subjectives. En effet, la note globale de qualité renvoyée par l'algorithme PESQ est hautement corrélée avec la note MOS (Mean Opinion Score).

La table 3 présente une évaluation comparative des performances du codeur de parole MELP global où ses paramètres LSF ont été codés séparément par les encodeurs suivants: le MSVQ original de 25 bits/trame, le LSF-SSVQ de 24 (7 + 8 + 9) bits/trame (division (4 – 6) avec m = 128) et le 2-SVQ de 24 (12 + 12) bits/trame avec la division (4 – 6).

Table 3 : Performances-PESQ du codeur global MELP

Codage-LSF du MELP	**Performance Encodeurs-LSF**			**Performance du MELP *PESQ***
	SD Moy. (dB)	**SD Outliers (%)**		
		2-4 dB	**> 4 dB**	
MSVQ Orig. (25 bpt)	1.018	1.230	0.000	3.160
LSF-SSVQ (24 bpt)	1.078	1.032	0.000	3.157
2-SVQ (24 bpt)	1.193	2.759	0.000	3.148

Ces résultats montrent que l'encodeur LSF-SSVQ de 24 bits/trame, incorporé dans le MELP, présente des performances comparables à celles du MSVQ original de 25 bits/trame. En effet, le LSF-SSVQ de 24 bpt a pu assurer une quantification de qualité transparente des LSFs-MELP avec en plus un gain de 1 bits/trame. On remarque aussi la supériorité du SSVQ sur le SVQ. D'autres part, les performances du codeur MELP en terme de PESQ sont très acceptables (PESQ supérieure à 3); assurant ainsi des communications de bonne qualité.

5. CON CLUSION

Dans ce travail, un système d'encodage basé sur la méthode SSVQ a été appliqué avec succès dans le codage efficace à bas débit des paramètres spectraux LSF du codeur de parole MELP de 2.4 Kbits/s. Comparé à l'encodeur des LSF conçu à base du SVQ conventionnel, l'encodeur LSF-SSVQ a permit de diminuer le débit d'environ 1-2 bits/trame, tout en maintenant des performances comparables.

Les performances de l'encodeur LSF-SSVQ en présence des erreurs de canal reste à être étudié. En outre une étude soigneuse de la robustesse de l'encodeur, en tenant compte des changements dans les conditions d'enregistrement des signaux parole, est aussi nécessaire.

BIBLIOGRAPHIE

[1] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*, Elsevier Science B.V., 1995.

[2] F. Itakura, Line spectrum representation of linear predictive coefficients of speech signals, *Journal of Acoust. Society America*, volume 57, page 535, 1975.

[3] K. K. Paliwal and B. S. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, *IEEE Transactions on Speech and Audio Processing*, Vol. 1, no.1, pages 3-14, 1993.

[4] W. F. Leblanc, B. Bhattacharya, S. A. Mahmoud and V. Cuperman, Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding, *IEEE Trans. Speech and Audio Processing*, volume 1, n° 4, pages 373-385, 1993.

[5] M. Bouzid, A. Djeradi et B. Boudraa, Optimized Trellis Coded Vector Quantization of LSF Parameters: Application to the 4.8 Kbps FS1016 Speech Coder, *Signal Processing Elsevier*, volume 85, issue 9, pages 1675-1694, 2005.

[6] Y. Linde, A. Buzo et R. M. Gray, An Algorithm for Vector Quantization Design, IEEE Transactions on Communications, volume COM-28, pages 84-95, 1980.

[7] S. Stephen et K. K. Paliwal, Efficient vector quantization of line spectral frequencies using the switched split vector quantiser, *in Proc. Int. Conf. Spoken language Processing*, Jeju, Korea, 2004.

[8] S. Stephen et K. K. Paliwal, A comparative study of LPC parameter representations and quantisation schemes for wideband speech coding, *Digital Signal Processing journal, Elsevier*, volume 17, pages 114-137, 2007.

[9] J. S. Garofolo et al., DARPA TIMIT Acoustic-phonetic Continuous Speech Database, *National Institute of Standards and Technology (NIST)*, Gaithersburg, October 1988.

[10] J. P. Campbell, T.E. Tremain, V. C. Welch, The Proposed Federal Standard 1016 4800 bps Voice Coder: CELP, *Speech Technology Magazine*, pages 58-64, 1990.

[11] A. McCree, K. Truong, E. B. George, T. P. Barnwell et V. Viswanathan, A 2.4 kbits/s MELP Coder Candidate for the New U.S. Federal Standard, *Proceedings of IEEE ICASSP-96*, pages 200-203, 1996.

[12] ITU-T, Recommendation P.862, Perceptual evaluation of speech quality assessment of narrowband telephone networks and speech codecs, February 2001.

Décodage guidé par un modèle cache sémantique

Benjamin Lecouteux, Pascal Nocera, Georges Linarès

LIA-CERI, université d'Avignon (France)

ABSTRACT

This paper proposes an improved semantic based cache model: our method boils down to using the first pass of the automatic speech recognition (ASR) system, associated to confidence scores and semantic fields, for driving the second pass. We use a Driven Decoding Algorithm (DDA), which allows us to combine ASR systems, by guiding the search algorithm of a primary system with an auxiliary system. We propose a strategy that uses DDA to drive a semantic cache, according to the confidence measures. The method works like an unsupervised language model adaptation. Results show, on 8 hours, that semantic-DDA yields significant improvements to the baseline: we obtain a 4% word error rate relative improvement without acoustic adaptation, and 1.9% after adaptation.

Keywords: Speech recognition, driven decoding, Latent Semantic Analysis, cache model

1. Introduction

Bien que les modèles n-gramme aient démontré leur efficacité dans le cadre des systèmes de reconnaissance automatique de la parole (SRAP), ils sont limités dans les modélisations à long terme ou sémantiques. Quelques travaux ont adressé ces aspects, principalement en modifiant les probabilités n-gramme en fonction de dépendances distantes ou de thèmes spécifiques :

- Les modèles caches introduits par [1], augmentent la probabilité des mots apparus récemment. L'hypothèse initiale étant que si un mot particulier est utilisé, ce dernier a de fortes chances de réapparaître.
- Les modèles triggers, présentés par [2, 3], où le problème de dépendance sur le long terme est résolu en interpolant des n-grammes avec des paires de mots "déclencheurs" (*triggers*) sélectionnés en fonction de leur information mutuelle. Cette approche est une généralisation des modèles cache.
- Les mélanges de modèles à base de thèmes: [4, 5, 6] proposent des techniques pour mixer des modèles en fonction de thèmes spécifiques. Les données d'apprentissage sont partitionnées en ensembles de thèmes qui sont utilisés pour adapter le modèle. [7] a introduit une méthode pour construire des modèles de langage en exploitant à la fois des contraintes locales et globales basées sur une analyse sémantique LSA (Latent Semantic Analysis). Cette approche propose d'estimer deux modèles, l'un à base de n-grammes, l'autre basé sur LSA. Ces modèles sont combinés, afin d'introduire l'information sémantique au sein du modèle n-gramme.
- La combinaison entre des mélanges de modèles et un modèle cache proposée par [8], combine ces deux approches pour capturer des dépendances éloignées au sein du langage.

Les modèles cache manquent de robustesse car les mots mis en avant dépendent de l'hypothèse courante : une erreur peut facilement se propager. De plus, ces modèles se basent uniquement sur le passé (l'historique de l'hypothèse courante). Le mélange de modèles à base de thèmes estime généralement des poids sur une première transcription ou sur l'hypothèse courante, sans prise en compte des erreurs potentielles. De plus, le principal problème de ces modèles est la sélection/détection des thèmes destinés à l'apprentissage.

Dans les travaux cités, des informations telles que les mesures de confiance produites par le SRAP ne sont pas utilisées. Elles sont généralement exploitées pour l'adaptation non supervisée des modèles acoustiques et rarement pour l'adaptation du modèle de langage : dans [9], les scores de confiance associés aux mots sont utilisés directement dans le graphe d'exploration, améliorant ainsi le décodage.

Notre objectif est d'exploiter toute l'information issue d'une première passe au cours de la seconde, afin de d'obtenir un modèle de langage adapté automatiquement. Nous proposons d'appliquer un modèle cache durant le processus de décodage, qui exploite les informations sémantiques et les mesures de confiance issues de la passe précédente.

Cet article présente une méthode intégrée permettant de diriger un SRAP avec son hypothèse précédente, associée à des mesures de confiance ainsi qu'à un modèle cache sémantique. Notre stratégie se focalise uniquement sur les mots mal reconnus, afin de réduire le bruit introduit par le modèle cache. Récemment, dans [10], nous avons proposé un algorithme qui permet d'introduire la sortie d'un SRAP auxiliaire au sein de l'algorithme de recherche. Nous présentons une extension de cet algorithme dédiée à l'adaptation non supervisée d'un SRAP. La première section présente l'ensemble du système. La seconde présente le protocole expérimental sur 8 heures extraites de la campagne ESTER [11]. La dernière section présente les expériences liées à notre décodage sémantiquement guidé. Finalement, nous concluons et suggérons quelques améliorations.

2. Approche intégrée : Le décodage guidé dédié à un cache sémantique

Le décodage guidé (Driven Decoding Algorithm, DDA) initial consiste à intégrer la sortie d'un

système auxiliaire dans l'algorithme de décodage d'un système primaire. Nous proposons de modifier l'algorithme pour obtenir un modèle cache amélioré. Les prochaines sous-sections présentent les différentes parties du système.

2.1. L'algorithme A* du système Speeral

Le SRAP du LIA est utilisé comme système primaire. Il est basé sur un algorithme de recherche A* opérant sur un treillis de phonèmes. Le processus de décodage repose sur une fonction d'estimation $F(h_n)$ qui évalue la probabilité de l'hypothèse h_n passant par le noeud n:

$$F(h_n) = g(h_n) + p(h_n), \quad (1)$$

Où $g(h_n)$ est la probabilité de l'hypothèse partielle au noeud n, qui résulte de l'exploration partielle du graphe. $p(h_n)$ est une sonde qui estime la probabilité restante entre le noeud n et le noeud final. Afin d'intégrer l'information issue du système auxiliaire, la partie linguistique de g dans (1) est modifiée en fonction de l'hypothèse auxiliaire, comme décrit ci-après.

2.2. Le décodage guidé

Le SRAP Speeral génère des hypothèses au fur et à mesure de l'exploration du treillis de phonèmes. La meilleure hypothèse à un temps t est étendue en fonction de la probabilité de l'hypothèse courante et du résultat de la sonde. Afin de combiner l'information issue de la transcription auxiliaire H_{aux} avec le processus de recherche, un point de synchronisation doit être trouvé pour chaque mot que le système évalue. Ces points sont trouvés en alignant dynamiquement la transcription fournie avec l'hypothèse courante; cette tâche est effectuée en minimisant la distance d'édition entre les deux hypothèses. Ce processus permet d'identifier dans la transcription auxiliaire H_{aux}, la meilleure sous-séquence qui correspond à l'hypothèse courante h_{cur}. Cette sous-séquence h_{aux} est utilisée pour une ré-estimation du score linguistique en fonction des probabilités *a posteriori* $\phi(w_i)$:

$$L(w_i|w_{i-2}, w_{i-1}) = P(w_i|w_{i-2}, w_{i-1})^{1-\beta}.\phi(w_i)^{\beta}$$
$$\beta = 0 \text{ si } w_i \text{ n'est pas trouvé dans } H_{aux} \quad (2)$$

Où $L(w_i|w_{i-2}, w_{i-1})$ est le score linguistique résultant, $P(w_i|w_{i-2}, w_{i-1})$ est la probabilité initiale du trigramme, β est un facteur d'échelle estimé empiriquement et $\phi(w_i)$ est le score de confiance du mot w_i.

2.3. Les mesures de confiance

Les mesures de confiance sont estimées en deux étapes. La première extrait des paramètres de bas niveau relatifs à l'acoustique et à la topologie du graphe, ainsi que des paramètres plus haut niveau liés à la linguistique. A partir de ces paramètres, un classifieur basé sur l'algorithme boosting assigne à chaque mot une probabilité d'être correct, comme détaillé dans [12]. Chaque mot de l'hypothèse est au final représenté par un vecteur de 23 paramètres, qui se regroupent en 3 classes :

- Les **paramètres acoustiques** tels que la log-vraisemblance acoustique du mot et la log-vraisemblance moyenne par trame.
- Les **paramètres linguistiques** sont basés sur les probabilités estimées par le modèle de langage 3-gramme utilisé dans le SRAP. Nous utilisons la probabilité 3-gramme, la perplexité du mot dans une fenêtre définie et la probabilité unigramme. Nous ajoutons un index représentant le repli actuel du mot au niveau du modèle de langage.
- **les paramètres liés au graphe** se basent sur l'analyse du mot dans le réseau de confusion. Nous utilisons le nombre de chemins alternatifs ainsi que la probabilité *a posteriori*. Nous incluons également des valeurs relatives à la distribution des probabilités *a posteriori* dans le réseau de confusion.

Le classifieur a été entraîné sur un corpus annoté en mots décodés correctement ou non. Le taux d'erreur confiance (CER) est de 19.5% sur le corpus de développement et 18.6% sur le corpus de test pour un seuil de 0.5. L'entropie normalisée croisée est quant à elle de 0.373 sur le corpus de développement et 0.282 sur le corpus de test. Un seuil de 0.85 a été choisi pour maximiser la confiance de décision dans le module sémantique : 55% des mots sont sélectionnés comme corrects avec seulement 2.7% d'erreurs.

2.4. Le module d'analyse sémantique

L'analyse sémantique latente (LSA) [7] est une technique permettant d'associer des mots qui sont corrélés sémantiquement à travers plusieurs documents. L'hypothèse formulée est que les mots co-occurrents dans un même document sont sémantiquement corrélés.

Dans notre système, une séquence de mot sémantiquement pertinente peut être considérée comme incohérente par le modèle de langage du SRAP en raison de la limite du modèle de langage n-gramme. Pour cette raison, nous ajoutons un estimateur de consistance sémantique qui permet de valider ou rejeter certaines hypothèses.

Dans nos expériences, le module LSA a été entraîné sur les données d'apprentissage du modèle de langage. Pour une meilleure couverture, le corpus a été lemmatisé et le vocabulaire réduit au lexique lemmatisé (environ 33K mots). De plus, une stop-liste a été appliquée pour filtrer les mots non porteurs de sens.

Lorsqu'un mot est présenté au module, ce dernier retourne les 100 meilleurs mots associés avec leurs scores de confiance LSA.

2.5. Sélection de mots pour LSA

Les mots sont sélectionnés en fonction de leur score de confiance, afin de ne pas introduire de bruit dans le module sémantique. Le seuil a été fixé à 0.85 où 55% du corpus est sélectionné tandis que le taux d'erreur de sélection est de 2.7%. Pour chaque mot sélectionné, 100 mots sont extraits avec le module LSA, générant au final un groupe de mots destiné à notre modèle cache.

2.6. Décodage guidé avec LSA

L'utilisation du décodage guidé est indispensable dans ce contexte, car un simple model cache LSA introduirait trop de bruit. Ainsi, le système est dirigé par ses hypothèses précédentes pour limiter les déviations des mots corrects. Le *trigger* LSA est appliqué uniquement sur les mots à faible confiance (< 0.5) : les mots corrects sont ainsi préservés. De plus, les mots associés à de faibles mesures de confiance sont sous-évalués, permettant au SRAP d'explorer des chemins alternatifs.

Le système final tel que détaillé dans la figure 1 fonctionne comme un modèle cache amélioré. Le DDA-LSA devient :

$$\phi(w_i) \geq 0.5 : \begin{cases} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\beta}.\phi(w_i)^{\beta} \\ \beta = 0 \text{ si } \phi(w_i) \text{ est trouvé dans } H_{aux} \end{cases}$$

$$\phi(cw_i) < 0.5 : \begin{cases} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\alpha}.\theta(w_i)^{\alpha} \\ \alpha = 0 \text{ si } \theta(w_i) \text{ non trouvé} \end{cases} \quad (3)$$

Où $L(w_i|w_{i-2}, w_{i-1})$ est le score linguistique résultant, $P(w_i|w_{i-2}, w_{i-1})$ est la probabilité initiale du trigramme, β et α sont des facteurs d'échelle calculés empiriquement, $\phi(w_i)$ est le score de confiance du mot w_i, cw_i est le mot aligné après l'historique (w_{i-2}, w_{i-1}) dans la transcription auxiliaire et $\theta(w_i)$ est le score LSA de w_i.

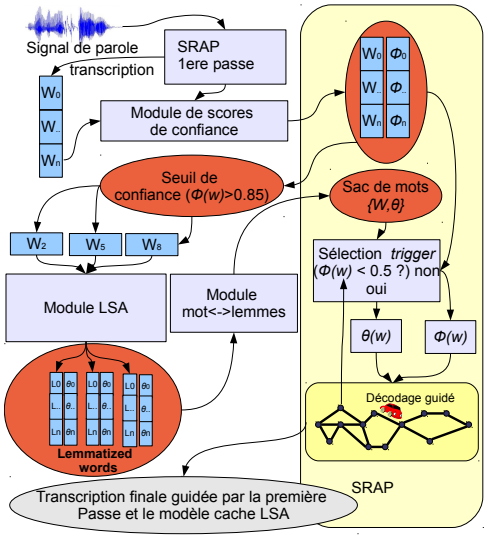


Figure 1: Principe du décodage guidé par LSA

2.7. *Le système de transcription du LIA*

Le système du LIA, Speeral, est un décodeur asynchrone opérant sur un treillis de phonèmes ; les modèles acoustiques utilisent des HMM et sont contextuels à base de tri-phones. Le modèle de langage est 3-gramme, estimé sur environ 200 millions de mots issus du journal *Le Monde* ainsi que du corpus ESTER (environ 1 million de mots). Le lexique est composé de 67000 mots. Dans ces expériences, une seule passe est effectuée en 3 fois le temps réel (3xRT).

2.8. *Le corpus ESTER*

Les expériences ont été effectuées sur les données issues de la campagne d'évaluation ESTER [11] : des émissions radio Marocaines et Françaises, des interviews, des locuteurs non natifs et des transcriptions à la volée. Les résultats sont issus de 8 heures extraites du test d'ESTER (sur 6 radios différentes). Les données d'apprentissage sont composées de 80 heures annotées manuellement, soit environ 1M mots ainsi que de 200M de mots issus du journal *Le Monde*.

3. Expériences

3.1. *Le SRAP guidé par sa 1ère passe*

Nous avons testé une seconde passe guidée par la première, associée à ses mesures de confiance. Le résultat n'a pas été surprenant : le SRAP converge sur la première passe et aucun changement significatif n'est observé. Ceci est du au fait qu'aucun chemin alternatif n'est proposé à l'algorithme d'exploration.

3.2. *Modèle cache sans contrainte*

Ces expériences testent l'utilisation d'un cache LSA sans contrainte. Pour chaque segment de parole, un ensemble de mots est associé via le module LSA. Nous avons implémenté un simple modèle cache : durant le processus d'exploration, chaque mot de l'hypothèse est recherché dans cet ensemble. Si le mot est trouvé, sa probabilité est augmentée en fonction de son score LSA. Les résultats sont peu intéressants, car une dégradation est observée, malgré l'information apportée par le LSA.

Cette dégradation est due à la quantité de mots dont la probabilité est modifiée : le bruit introduit impacte le résultat final. Cet aspect met en évidence la difficulté liée au réglage d'un modèle cache.

3.3. *Décodage guidé avec un cache LSA*

Ces expériences sont réalisées avec un système 3xRT, sans adaptation acoustique. Le SRAP utilise sa meilleure hypothèse précédente associée à ses scores de confiance. Le cache sémantique n'est appliqué qu'aux mots de faible confiance. Nous espérons guider le système avec les bons mots tout en réévaluant les moins bons avec le module LSA. Les résultats sont présentés dans le tableau 1. Le taux d'erreur mots (TEM) est réduit de 4% relatifs par rapport à la première passe : le cache LSA permet donc de se focaliser uniquement sur les erreurs. Cette stratégie réduit l'introduction de bruit.

Heure	P1 3xRT	P2 3xRT	P2-LSA DDA 3xRT
Classique 1h	21.4%	20.8%	20.9%
Culture 1h	34.0%	31.9%	33.3%
INTER 1h	22.7%	22.0%	21.6%
INFO 2h	25.8%	24.6%	25.0%
RFI 1h	28.6%	26.0%	27.1%
RTM 2h	35.4%	32.3%	33.6%

Table 1: DDA-LSA sans adaptation acoustique : *baseline* en première passe (P1 3xRT), *baseline* en seconde passe avec adaptation acoustique (P2 3xRT) et seconde passe avec DDA-LSA sans adaptation acoustique (P2-LSA DDA 3xRT)

3.4. *DDA-LSA avec adaptation mllr 3RT*

Ces expériences combinent l'adaptation acoustique par maximum de vraisemblance par régression linéaire (Maximum Likelihood Linear Regression, MLLR) pour la seconde passe (3xRT) avec le décodage guidé sémantique, afin de tester la complémentarité du DDA-LSA avec l'adaptation acoustique. Les résultats présentés dans le tableau 2 montrent un TEM réduit de 1.9% relatifs. Ceci montre la complémentarité avec le processus d'adaptation acoustique. Nous observons un meilleur gain sur la plus mauvaise heure (4.6% relatifs).

Heure	P2 3RT	P2-LSA DDA 3RT
Classique 1h	20.8 %	20,5 %
Culture 1h	31.9 %	31.8 %
INTER 1h	22.0 %	21.6 %
INFO 2h	24.6 %	24.5 %
RFI 1h	26.0 %	25.5 %
RTM 2h	32.3 %	30.8 %

Table 2: *baseline* en seconde passe 3xRT (P2 3xRT), décodage guidé par la sémantique avec adaptation acoustique en 3xRT (P2-LSA DDA 3xRT)

3.5. DDA-LSA avec adaptation mllr 10RT

Ces dernières expériences testent l'approche avec une exploration maximale du graphe de recherche au cours de la seconde passe en 10xRT. Les résultats sont présentés dans le tableau 3 : le TEM est réduit de seulement 1.1% relatifs ; la complémentarité entre l'adaptation acoustique et l'adaptation linguistique devient faible. Cependant, la plus mauvaise heure (RTM) présente une amélioration relative de 3.8%. Plus globalement, notre stratégie est intéressante dans le contexte d'un système 3xRT où les résultats convergent vers ceux d'un système 10xRT.

Heure	P2 10xRT	P2-LSA DDA 3RT	P2-LSA DDA 10RT
Classique 1h	20.2 %	20,5 %	20.0 %
Culture 1h	31.7 %	31.8 %	31.5 %
INTER 1h	21.6 %	21.6 %	21.6 %
INFO 2h	24.0 %	24.5 %	23.9 %
RFI 1h	25.4 %	25.5 %	25.3 %
RTM 2h	31.7 %	30.8 %	30.5 %

Table 3: *baselines* issues de la seconde passe après adaptation des modèles acoustiques en 10xRT (P2 10xRT), décodage guidé par la sémantique avec adaptation acoustique en 3xRT (P2-LSA DDA 3xRT) et décodage guidé par la sémantique avec adaptation acoustique en 10xRT (P2-LSA DDA 10xRT)

Le DDA-LSA améliore significativement le système 3xRT et plus particulièrement l'heure RTM. Avec le système 10xRT l'amélioration est faible, excepté sur l'heure RTM. Contrairement aux autres heures, RTM est une radio Marocaine, tandis que les données d'apprentissage du modèle de langage sont dérivées d'un journal Français. Il en résulte une couverture moins bonne pour RTM. Ceci explique les gains plus significatifs sur RTM et montre la contribution du modèle sémantique.

4. Conclusion

Nous avons proposé un décodage guidé par de l'information sémantique extraite de la première passe d'un SRAP. Notre stratégie s'est concentrée sur un modèle cache sémantique, qui s'applique en fonction du score de confiance de chaque mot.

Les expériences montrent que cet algorithme améliore le système initial et complète la phase d'adaptation acoustique. L'enrichissement de la première passe avec les scores de confiance est nécessaire pour orienter correctement l'algorithme de recherche, tandis que l'information sémantique permet de sélectionner des chemins alternatifs corrects quand les scores de confiance sont bas : cette stratégie s'assimile à une adaptation non-supervisée du modèle de langage. Un gain de 4% relatifs de TEM est obtenu sur la première passe sans adaptation acoustique, 1.9% sur le système 3xRT après adaptation acoustique et 1.1% sur le système 10xRT. La stratégie est plus intéressante dans le cadre d'un système 3xRT où le TEM est proche du système 10xRT, tout en réduisant les coûts de calcul. Malgré tout, les meilleurs résultats sont obtenus sur les heures les plus éloignées des données d'apprentissage.

Actuellement notre méthode est limitée à la granularité des segments. Nous souhaitons l'étendre à des ensembles de segments (discussion entre plusieurs locuteurs etc.). Nous envisageons également d'intégrer des données sémantiques externes telles que des résumés de documents audio ou des méta données (titres, locuteur, etc.).

Bibliographie

[1] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[2] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech and Language*, vol. 10, no. 3, pp. 187–228, 1996.

[3] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modeling," in *Computer Speech and Language, vol. 8, pp. 1-38,*, 1994.

[4] Yoshihiko Gotoh and Steve Renals, "Topic-based mixture language modelling," *Natural Language Engineering*, vol. 5, pp. 355–375, 1999.

[5] N. Singh-Miller and C. Collins, "Trigger-based language modeling using a loss-sensitive perceptron algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, 15–20 April 2007, vol. 4, pp. 25–28.

[6] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here ?," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, Aug. 2000.

[7] J.R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[8] P.R. Clarkson and A.J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, 1997, vol. 2, pp. 799–802 vol.2.

[9] P. Fetter, F. Dandurand, and P. Regel-Brietzmann, "Word graph rescoring using confidence measures," in *Fourth International Conference on Spoken Language ICSLP*, 1996, vol. 1, pp. 10–13 vol.1.

[10] B. Lecouteux, G. Linares, Y. Esteve, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 1549–1552.

[11] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase 2 evaluation campaign for the rich transcription of french broadcast news," in *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

[12] P. Moreno, B. Logan, and B. Raj, "A boosting approach for confidence scoring," in *Interspeech, Aalborg, Denmark*, 2001, pp. 2109–2112.

Liage et fusion audiovisuelle en perception de la parole : on peut « débrancher » lœffet McGurk par un contexte audiovisuel incohérent

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz

GIPSA-Lab ï DPC, ICP
UMR 5216 ï CNRS Uiversité de Grenoble
Olha.Nahorna, Jean-Luc.Schwartz, Frederic.Berthommier@gipsa-lab.grenoble-inp.fr
http://www.gipsa-lab.inpg.fr

ABSTRACT

The McGurk effect demonstrates the existence of a fusion process in audiovisual speech perception: the combination of the sound "ba" with the face of a speaker who pronounces "ga" is frequently perceived as "da". We assume that in the upstream of this phonetic fusion process, there is another early fusion process, which controls the combination of image and sound, and can block it in the case of audiovisual inconsistencies (conditional binding process), as in the case of a dubbed film. To test this early fusion hypothesis , we designed an experiment in which a consistent or inconsistent audiovisual context is placed before McGurk stimuli, and we show that the inconsistent contextual stimulus can remove the effect McGurk.

Keywords: McGurk effect, binding, multisensory fusion, audiovisual speech perception, audiovisual scene analysis.

1. INTRODUCTION

La perception visuelle fait partie intégrante de la perception de la parole chez les humains. Le célèbre effet McGurk [1] montre bien lôinfluence de lôinformation visuelle sur la parole perçue. Le montage du son « ba » avec un film de « ga » est perçu comme « da » chez de nombreux sujets.

Plusieurs architectures de fusion audio-visuelle ont été proposées dans la littérature [2]. Elles ont en commun de considérer des prises dôinformation auditive et visuelle indépendantes. Or, il y a déjà une quinzaine dôannées est apparue lôhypothèse de lôexistence de mécanismes précoces pour extraire lôinformation auditive et visuelle (voir [3]). Pour rendre compte de ce type de phénomène, Berthommier [4] a proposé un modèle dans lequel la fusion audio-visuelle est précédée dôun niveau primitif et pré-phonétique (Figure 1). Le rôle des interactions bas-niveau serait de renforcer la modulation dôamplitude des segments de la parole, sans distorsion des signaux phonétiques, spectrale ou temporelle. Ce niveau précoce permettrait de conditionner les mécanismes de fusion.

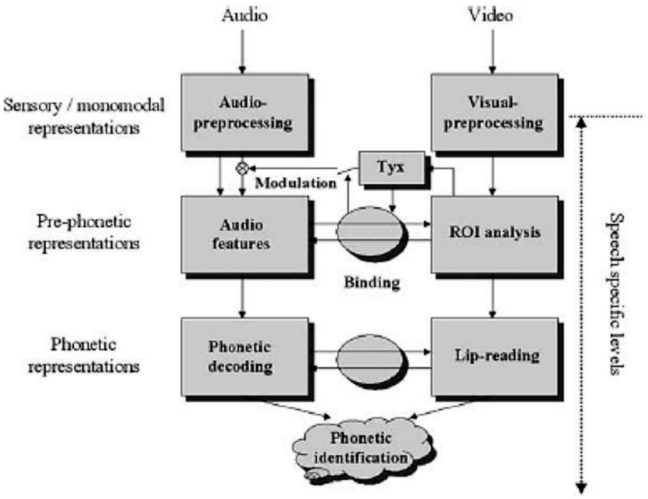


Figure 1 : Un modèle de fusion intégrant une interaction bas-niveau [4])

Ainsi, ce modèle postule deux niveaux dôinteraction audiovisuelle, un niveau précoce (détection) et un niveau tardif (fusion). La question expérimentale de ce travail est de savoir si le mécanisme de détection précoce fait partie dôun système plus large assurant un rôle de liage conditionnel. Ce système permettrait, au cas par cas, de lier les entrées auditives et visuelles en un même flux, ou au contraire de les séparer en deux flux différent. Si cêst le cas, on doit pouvoir construire des situations expérimentales où on peut « débrancher » le second niveau de fusion, comme cêst probablement le cas dans les filmes doublés, où il ne faut pas intégrer les entrées auditive et visuelle dans la reconnaissance, puisquêelles sont incongruentes et ne portent pas dôinformation phonétique cohérente.

Nous avons pris lœffet McGurk comme indicateur de la fusion. Nous allons donc essayer de construire un paradigme expérimental visant à supprimer ou modifier lœffet McGurk. Nous cherchons à déterminer si lœffet McGurk résiste à des variations du contexte préalable, qui permettrait de lier/délier les flux auditif et visuel. Nous supposons que par manipulation du contexte, on peut produire un « décrochage » du lien audiovisuel, conduisant à une diminution de la fusion audio-visuelle.

2. MÉTHODOLOGIE

Le paradigme expérimental, consiste à présenter à des sujets un flux de parole audiovisuelle et de leur demander de détecter en ligne la présentation de stimuli « ba » ou « da ». Nous présentons aux sujets deux types de stimuli cibles : un stimulus cohérent « ba » (audio « ba » + vidéo « ba »), dont on attend quôil soit correctement identifié « ba », et un stimulus « McGurk » (audio « ba » + vidéo « ga »), dont on attend quôil soit souvent perçu « da ».

Notre hypothèse est que lôeffet McGurk disparaît en fonction du contexte préalable. Pour cela nous construisons deux types de contexte : « cohérent » et « incohérent ». Dans le cas cohérent le contexte consiste en une séquence de syllabes, présentées en modalité audiovisuelle : le sujet voit donc le visage du locuteur qui prononce des syllabes synchronisées avec les syllabes audio que le sujet entend. Le contexte incohérent est constitué du même matériel audio, superposé avec la vision du même locuteur, qui prononce de la parole quelconque et non pas des syllabes. Les cibles auditives sont les mêmes dans les deux cas. Comme nous ne savons pas a priori combien de temps il faut présenter le contexte incohérent pour perturber lôeffet McGurk, nous utilisons des contextes de durées variables (Figure 2).

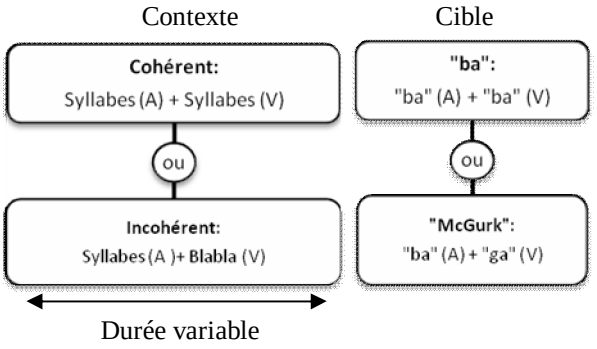


Figure 2 : Principe expérimental

2.1. Mise en oeuvre

Pour préparer lôexpérience nous avons enregistré des séquences avec des syllabes et de la parole quelconque. Le contexte acoustique est constitué de séquences aléatoires de syllabes françaises (syllabes CV, C étant une plosive ou une fricative, à lôexclusion des syllabes « ba », « da » et » ga », soit 13 syllabes possible) et les syllabes « ba » ou « ga » servent de cible. Nous avons utilisé des contextes préalables incohérents de 5, 10, 15, 20 syllabes.

Les stimuli cibles « ba » ne présentent pas dôintérêt direct dans cette expérience, puisque nous prédisons quôils devraient être identifiés correctement « ba » quelque soit le contexte. Seuls les stimuli McGurk nous intéressent, la prédiction étant quôils produisent moins de réponses de fusion « da » (et plus de réponses « auditives » « ba ») dans le cas de contexte incohérent. Mais pour être sûr que les sujets sont attentifs pendant toute lôexpérience et répondent correctement, on ne peut présenter uniquement des stimuli McGurk. Les données empiriques montrent que lôeffet McGurk apparait en moyenne dans 35-50% des cas, tandis que les stimuli « ba » produisent des réponses « ba » dans presque 100% des cas. Pour équilibrer dans notre expérience la fréquence attendue des réponses « ba » et « da », et pour optimiser le nombre de cibles «McGurk » qui concentrent notre intérêt, nous avons décidé présenter les stimuli dans les proportions : ¼ des stimuli « ba » et ¾ des stimuli « McGurk ».

Pour résumer, nous avons 3 facteurs principaux à contrôler dans la préparation de lôexpérience :

- Stimuli : ¾ de « McGurk » versus ¼ de « ba »
- Cohérence : contexte cohérent versus incohérent
- Durée : 5, 10, 15, 20 syllabes.

2.2. Préparation des matériaux expérimentaux

Enregistrement

Nous avons enregistré 80 séquences audiovisuelles de contextes de durée variée, se terminant toujours par la cible « ba » ou « ga » (prononcées par un locuteur français, JLS, avec les lèvres maquillées en bleu). Les 40 séquences destinées à produire le contexte audio pour toute lôexpérience, et le contexte vidéo pour le cas de contexte cohérent, sont produites par des arrangements aléatoires de 13 syllabes françaises : « pa », « ta », « va », « fa », « za », « sa », « ka », « ra », « la », « ja », « cha », « ma », « na ». 20 séquences se terminent par une syllabe « ba » et 20 par une syllabe « ga ». La longueur des séquences est 5, 10, 15, 20 syllabes, correspondant à des durées de lôordre de 3, 7, 10 et 13 s. Les séquences étaient présentées au locuteur sur un écran de contrôle. Le locuteur devait répéter les séquences proposées, en laissant à chaque fois un silence court entre deux syllabes consécutives, de façon à fournir des points de montage acoustique simples.

Les 40 séquences destinées à produire le contexte incohérent consistent en un flux de parole quelconque de durée 4, 7, 10, 13 secondes, se terminant dans la moitié des cas par une séquence « ba » et dans lôautre moitié par une séquence « da ». Le locuteur devait parler librement sur le sujet de son choix, et au bout dôune durée correspondant à la condition correspondante (4, 7, 10, 13 secondes), lôindication de la syllabe terminale apparaissait, indiquant au locuteur quôil devait conclure en prononçant cette syllabe.

Sélection et montage

Pour préparer les stimuli McGurk nous avons fait un montage audio en remplaçant le son « ga » par le son « ba », pris dans lôautre groupe des séquences avec « ba » à la fin (Fig. 3). Les données ont été ensuite normalisées en amplitude, et sélectionnées sur des critères dôamplitude de mouvement visuel, de manière à ce que les stimuli « cohérents » et « incohérents » ne diffèrent pas en terme de contenu audiovisuel des stimuli cible McGurk [5].

Nous avons ainsi préparé 16 stimuli originaux de chaque type (4 par durée de contexte, avec 4 durées de contexte),

et ce pour les 4 types définis par la cible (« ba » vs. McGurk) et le contexte (cohérent vs. incohérent). Nous les avons combinés dans lœexpérience avec les proportions : ¾ de « McGurk » versus ¼ de « ba ». Pour ce faire les stimuli de type McGurk ont été répétés 3 fois Au total nous avons donc présenté 128 stimuli (16 « ba » dans les 2 contextes, 48 McGurk dans les deux contextes) répartis en 4 blocs de 32 stimuli, répartis aléatoirement.

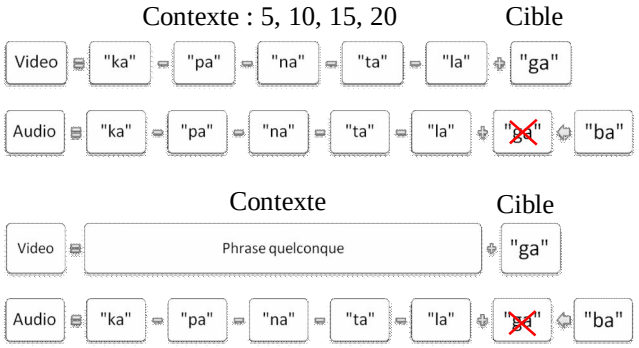


Figure 3 : Montage des stimuli "McGurk" à contexte cohérent (en haut) ou incohérent (en bas)

2.3. Passation du test et analyse des réponses

Protocole et sujets

Le protocole consistait, selon lœinstruction donnée aux sujets, à observer les films et, chaque fois que le sujet entendait le son "ba" ou "da", à appuyer immédiatement sur le bouton correspondant, indiqué par le présentateur au début de lœexpérience. Les boutons de réponse ont été spécifiés par des marques « ba » et « da ». Pour la moitié des sujets le bouton « ba » était à gauche et le bouton « da » était à droite, pour lœautre moitié les boutons ont été inversés. Les mêmes marques étaient également affichées sur l'écran, de chaque côté de lœécran au niveau des yeux. La durée dœexpérience était environ 25 minutes. Entre les blocs le sujet pouvait faire une pause de durée arbitraire. Les réponses des sujets, avec leur date précise, étaient enregistrées automatiquement au cours de lœexpérience, par le logiciel de test.

19 sujets français ont participé à lœexpérience avec vision et audition normale ou corrigée, soit 8 femmes et 11 hommes, entre 22 et 51 ans (17 droitiers et 2 gauchers).

Analyse des résultats

Pendant lœexpérience les stimuli sont fournis en ligne, et le sujet peut répondre à chaque instant, quœil y ait ou non la présence dœune cible perceptive « ba » ou « da ». Il peut donc se produire deux types dœerreurs : la présence dœune réponse « ba » ou « da » en lœabsence de cible (stimulus « ba » ou « McGurk ») ou lœabsence de réponse à une cible. Pour traiter correctement les réponses nous avons mis en place la méthodologie suivante. (1) Pour un stimulus on compte les réponses qui sont apparues après sa présentation, mais avant le stimulus suivant. (2) Des analyses nous avons été conduits à limiter la validité temporelle de réponse par un seuil, qui est égal à la durée dœune séquence minimale (3500ms). Sur lœhistogramme temporel de toutes les réponses (correctes et incorrectes), données par tous les sujets, nous avons pu observer que les plupart des réponses sont à lœintérieur de ce seuil. (3) Pour déterminer les réponses incorrectes, nous distinguons 2 types dœerreurs : « Fausses alarmes » et « Absence de réponse ». Toutes les réponses au-delà du seuil sont considérées comme « fausses alarmes ». Sœil nœy a pas de réponse dans lœintervalle entre le stimulus et le seuil, on compte une « Absence de réponse » pour ce stimulus. Sœil y a plusieurs réponses dans cet intervalle, on fait une vérification de lœidentité des réponses. Si elles sont identiques, nous ne prenons que lœune dœentre elles et la comptons comme une réponse normale, sinon nous les éliminons toutes, et considérons une « absence de réponse » pour le stimulus.

3. RESULTATS

Les résultats bruts sont présentés dans la Table 1. Il apparaît une tendance à obtenir plus dœabsence de réponse et moins de réponses multiples proportionnellement avec les stimuli McGurk quœavec les « ba », sans que le contexte ne joue fortement sur ces tendances (voir les deux colonnes de droite). Si lœon en vient à ce qui est le focus de notre étude, la proportion de réponses « ba » par rapport au nombre total de réponses (« ba » + « da »), on obtient les données de la Figure 4.

Une analyse de la variance à trois facteurs (stimulus, contexte, sujets) sur ces proportions (après transformation en asin(sqrt), pour assurer la gaussianité) montre que les 3 effets sont fortement significatifs. Lœeffet significatif du stimulus traduit lœeffet McGurk (moins de réponses « ba » pour les stimuli McGurk : $F(1,18)=61.77$, $p<0.0001$). Lœeffet sujet traduit les fortes différences interindividuelles classiques dans lœeffet McGurk ($F(18,18)=3.76$, $p<0.004$). Lœeffet contexte, traduisant la chute du nombre de réponses « ba » en contexte incohérent ($F(1,18)=35.67$, $p<0.0001$) est essentiellement produit par les stimuli McGurk, ainsi que le montre lœexistence dœune interaction entre stimulus et contexte ($F(1,18)=24.14$, $p<0.0001$).

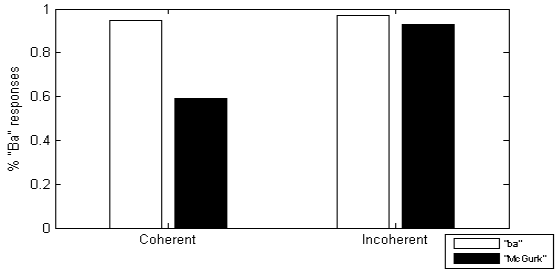


Figure 4 : Pourcentage de réponses « ba » rapportées à lœensemble des réponses (« ba »/(« ba » + « da »))

Table 1 : Résultats des tests.

Stimuli		Stimuli présentés	Réponse « ba »	Réponse « da »	Absence de réponses	Plusieurs réponses
Cohérent	« ba »	304	250	12	25	17
	« McGurk »	912	455	301	117	39
Incohérent	« ba »	304	260	7	21	16
	« McGurk »	912	724	53	110	25

Une seconde analyse de la variance, centrée sur les stimuli dôintérêt, les stimuli McGurk, à trois facteurs, sujet, contexte et durée, ne fait pas apparaître dôeffet durée global ($F(3,54)=2.07$, $p=0.1156$) mais un effet dôinteraction durée-contexte ($F(3,54)=2.85$, $p<0.05$) faiblement significatif. La Figure 5 montre que cet effet est dû au contexte cohérent, pour lequel un allongement du contexte augmente légèrement lôeffet McGurk, effet confirmée par une analyse par régression linéaire.

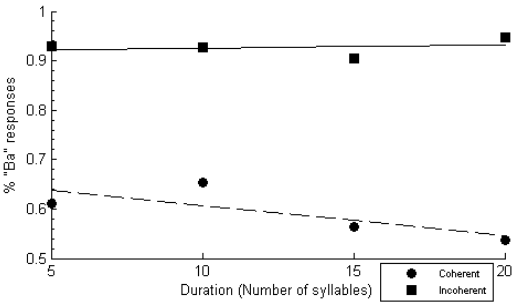


Figure 5 : La régression moyenne des réponses en fonction du contexte et sa durée (en contexte cohérent, r= -0,31 significativement différent de zéro, p.<0.0059).

4. DISCUSSION ET CONCLUSION

Les résultats obtenus montrent clairement que lôeffet McGurk dépend du contexte préalable. La suppression dôeffet McGurk signifie que lôon peut bloquer la fusion audio-visuelle. Dans le cas cohérent les flux auditif et visuel se combinent dans un même percept, avec des taux de McGurk classiques (60% de réponses « ba » contre 40% de réponses « da », ce qui correspond aux données dôautres études en français, voir [6]. Dans le cas incohérent lôeffet McGurk disparaît complètement et la réponse est essentiellement gérée par lôinformation auditive, avec des scores supérieurs à 90% de réponses « ba » pour des stimuli McGurk, presque identique à la réponse à un stimulus cohérent « ba ».

Après avoir longtemps considéré lôeffet McGurk comme automatique [1], des données récentes ont indiqué quôil était sous la dépendance de mécanismes attentionnels de divers types [7,8]. Néanmoins, côest la première fois, à notre connaissance, quôil est démontré sa sensibilité à des mécanismes ï vraisemblablement préattentionnels ï de liage conditionnel, qui se réfèrent à des travaux préalables que nous avons menés sur lôanalyse de scènes audiovisuelle [9]. Il nous faudra montrer par la suite comment fonctionnent ces mécanismes de liage, à quels mécanismes ils sont reliés à de quels paramètres expérimentaux ils dépendent, ainsi que dôen mettre à jour les corrélats neuronaux (voir par exemple les données récentes de Bernstein et col. [10] sur le rôle potentiel du Gyrus Supramarginal comme « hub », lieu de fusion de lôinformation provenant de flux différents, et de contrôle des différences et incohérences entre flux).

Remerciements ï Cette étude est financée par le projet ANR-08-BLAN-0167 MULTISTAP

BIBLIOGRAPHIE

[1] H. McGurk & J. MacDonald, "Hearing lips and seeing voices," *Nature*, 264, 746-748, 1976.

[2] J.-L. Schwartz, J. Robert-Ribes, & P. Escudier, "Ten years after Summerfield. a taxonomy of models of audiovisual fusion in speech perception," in *Hearing by Eye*, R. Campbell and et al., Eds. Hove, UK: Psychology Press, 1998, pp. 85-108.

[3] J.-L. Schwartz, F. Berthommier, & C. Savariaux, "Seeing to hear better: evidence for early audio-visual interaction in speech identification," *Cognition*, 93, B69-B78, 2004.

[4] F. Berthommier, "A phonetically neutral model of the low-level auio-visual interaction," *Speech Communication*, 44, 31-41, 2004.

[5] O. Nahorna, "Lôémergence des formes audiovisuelles dans le traitement multisensoriel de la parole : expériences et modélisation," Rapport de stage, Master IC2A-AST, Grenoble INP, 2009.

[6] M.A. Cathiard, J.L. Schwartz, & C. Abry, "Asking a naive question about the McGurk Effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]?", *Proc. AVSP-2001*, 138-142, 2001.

[7] K. Tiippana, M. Sams, and K.S. Andersen, "Visual attention modulates audiovisual speech," *Proc. AVSP-2001*, 167-171, 2001.

[8] A. Alsius, J. Navarra, R. Campbell & S. Soto-Faraco, ñAudiovisual Integration of Speech Falters under High Attention Demandsò, *Current Biology*, 15, 839-843, 2005.

[9] J. Barker, F. Berthommier, & J.L. Schwartz, ñ Is primitive coherence an aid to segment the scene? ò *Proc. AVSPô98*, Terrigal, Australia, 103ï 108, 1998.

[10] L.E. Bernstein, Lu Z.-L., & J. Jiang, "Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing," *Brain Research*, 1242:172-84, 2008.

Estimation du pitch utilisant le spectre du produit multi-échelle du signal de parole en présence de bruit blanc

Mohamed Anouar Ben Messaoud, Aïcha Bouzid et Noureddine Ellouze

Laboratoire signal, image et reconnaissance de formes (LSTS-ENIT)
Le Belvédère, B. P. 37, 1002, Tunis
anouar.benmessaoud@yahoo.fr, bouzidacha@yahoo.fr, N.Ellouze@enit.rnu.tn

ABSTRACT

In this work, we propose and describe an algorithm for estimating the fundamental frequency of a voiced and noisy speech signal. Our approach is based on the spectral analysis of the speech multi-scale product. The multi-scale product consists of making the product of the speech wavelet transform coefficients at three successive dyadic scales. The wavelet used is the quadratic spline function with a support of 0.8 ms. In each time frame, The fundamental frequency corresponds to the frequency that matches with the spectral ray localised on the power spectral density function following a defined strategy. We evaluate our approach using the Keele University database. Experimental results show the effectiveness of our method comparing other algorithms. Besides, the proposed approach is robust in the noisy environment.

Keywords: speech, wavelet transform, multi-scale product, spectral analysis, fundamental frequency

1. INTRODUCTION

Le signal de parole est aléatoire non stationnaire constitué de zones voisées, semi voisées, non voisées et de silence. Dans la parole voisée, l'excitation acoustique principale se produit à l'instant de fermeture des cordes vocales [1]. Celle-ci délimite la période du pitch donnant ainsi une estimation exacte de la fréquence fondamentale F_0.

L'estimation du paramètre F_0 est utile dans un grand nombre d'applications comme la synthèse vocale ou la reconnaissance, la transcription de musique polyphonique, etc… La variation de F_0 contribue à la prosodie et apporte de l'information concernant le message sonore. Elle permet de distinguer les langages tonals, d'exprimer des émotions, de distinguer les questions des phrases déclaratives et d'insister sur des parties d'une phrase. La localisation du pitch est aussi la base de la séparation du signal de parole harmonique d'autres composantes parole ou bruit [2].

De multiples approches qui peuvent être considérées assez robustes ont été proposées pour la détermination de pitch des signaux de parole dans un contexte non bruité [3].

Les algorithmes d'estimation de pitch peuvent être classés en trois classes: La première classe concerne les approches temporelles dont principalement la fonction d'autocorrélation [4], PRAAT [5], YIN [6]. La deuxième classe porte sur les approches fréquentielles comme le spectre d'amplitude [7]. La troisième classe concerne les approches temps-fréquence comme les transformées en ondelettes [8].

Dans ce papier, nous présentons une approche temps-fréquence simple et robuste pour l'estimation du pitch. La méthode consiste à identifier et localiser la raie spectrale qui permet de déterminer la fréquence du pitch et ce à partir de la fonction carré du spectre d'amplitude du produit multi-échelle du signal de parole.

Ce papier est organisé comme suit. La section 2 décrit le principe de produit multi-échelle. La section 3, présente l'approche que nous proposons pour l'estimation de la fréquence fondamentale. La section 4 décrit l'évaluation de notre approche sans bruit et en présence de bruit. Enfin, la section 5 conclut ce travail.

2. PRODUIT MULTI-ÉCHELLE

Utilisé sur le signal de parole, le produit multi-échelle (PM) a montré ses preuves quant à la détection des instants d'ouverture et de fermeture de la glotte, et l'estimation de la fréquence fondamentale et du quotient ouvert [9], [10].

Dans ce travail, nous allons utiliser le même outil pour l'estimation de la fréquence fondamentale moyenne sur des fenêtres de longueur fixe.

Le PM calcule le produit des coefficients de la transformée en ondelettes pour différentes échelles dyadiques successives selon l'équation 1:

$$p(n) = \prod_{j=1}^{n} W\ f(n, s_j). \qquad (1)$$

Le produit p (n) montre des pics aux transitions présentes dans le signal et présente de faibles valeurs ailleurs. Dans cette opération non linéaire, les pics des coefficients de la transformée en ondelettes sont renforcés par le PM. Bien que certaines échelles de lissage ne soient pas optimums, la combinaison non linéaire tend à rehausser les maxima en atténuant les faux pics. Le nombre impair des termes de p (n) permet de préserver le signe de la singularité. Trois échelles dyadiques consécutives suffisent généralement pour la détection des pics.

Nous avons choisi l'ondelette spline quadratique qui est la dérivée première d'une fonction de lissage spline cubique et donc à un seul moment nul. La particularité de cette ondelette est qu'elle est la mieux adaptée pour caractériser les maxima associés aux véritables discontinuités du signal acoustique selon [11]. La Figure 1 récapitule les étapes de PM.

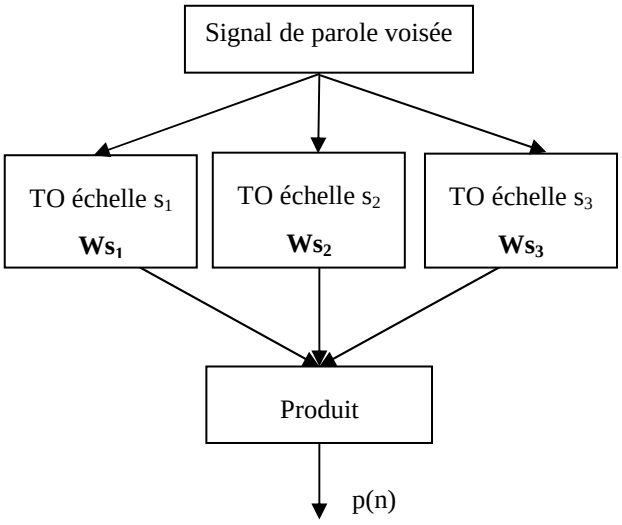


Figure 1 : Schéma block de calcul du produit multi-échelle du signal de parole.

Le PM appliqué au signal de parole voisé non bruité suivi de son produit multi-échelle est représenté dans la Figure 2. Le PM montre une structure de périodicité similaire au signal d'origine, mais avec une forme plus simplifiée qui rappelle celle de la dérivée du signal électroglottographique DEGG. Les oscillations dues à l'effet du système phonatoire (conduits nasal et vocal) sont atténuées, le PM fait ressortir les discontinuités présentes dans la structure périodique en zone voisée et atténue le reste. Il a permis de transformer le signal de parole en une forme moins complexe tout en gardant les éléments essentiels de la source.

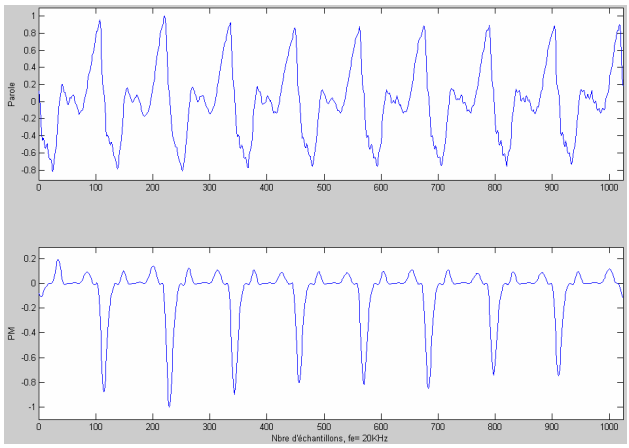


Figure 2 : Signal de parole non bruité suivi de son PM.

La Figure 3 montre le signal de parole voisé corrompu par un bruit blanc gaussien à un RSB de -5 dB suivi de son produit multi-échelle. Malgré la puissance du bruit, le PM réduit nettement le bruit générant ainsi un signal exploitable pour d'autres transformations.

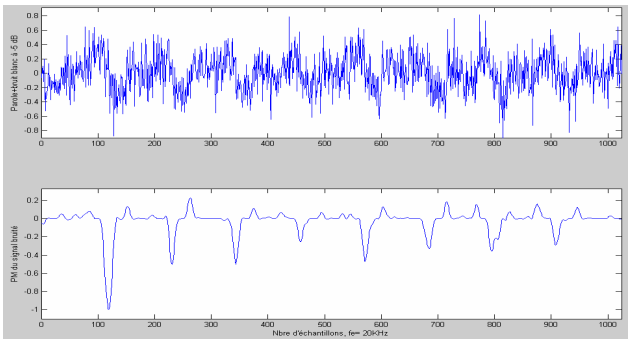


Figure 3 : Signal de parole corrompu par un bruit blanc gaussien de RSB = -5dB suivi de son PM.

3. ESTIMATION DU PITCH PAR ANALYSE SPECTRALE DU PM

L'approche que nous proposons est basée sur la sélection et la localisation de la raie spectrale correspondant à la fréquence fondamentale sur la fonction carrée du spectre d'amplitude du produit multi-échelle (DSPM) du signal de parole voisé.

La méthode peut être décomposée en trois étapes selon le schéma bloc donné par la Figure 4. La première étape consiste à opérer une analyse multi-échelle du signal. Nous calculons la transformée en ondelettes du signal à trois échelles dyadiques successives, ensuite nous effectuons le produit des coefficients en ondelettes générant le signal de la parole voisée. Nous avons utilisé l'ondelette spline quadratique de support 0.8 *ms* aux échelles ½, 1 et 2. La seconde étape consiste à calculer le carré du spectre d'amplitude du PM sur des fenêtres. La troisième étape consiste à détecter et localiser tous les pics de la fonction calculée et à déterminer la fréquence fondamentale F_0 sur chaque fenêtre.

La première étape est détaillée dans la section précédente. Pour la seconde étape, Le produit p(n, s_1, s_2, s_3) est pondéré par une fenêtre glissante $w[n]$:

$$p_w[n,i] = p[n]\,w[n - i\Delta n] \qquad (2)$$

i est l'indice de la fenêtre et Δn est le recouvrement. La fenêtre de Hanning est utilisée. La longueur de la fenêtre est choisie de telle sorte qu'elle assure la stationnarité du signal et une meilleure résolution spectrale. Nous avons choisi une fenêtre de 1024 échantillons avec un recouvrement de 512 points à une fréquence d'échantillonnage de 20 kHz. La transformée de Fourier est calculée par l'algorithme FFT sur N=4096 points.

La FFT est calculée selon l'équation suivante :

$$P_w^i[k] = \sum_{n=0}^{N-1} p_w[n,i]\, e^{-j\frac{2\pi}{N}nk} \qquad (3)$$

Ensuite, on calcule le carré du spectre d'amplitude du produit multi-échelle égal à $\left|P_w^i[k]\right|^2$.

Pour chaque tranche du PM du signal de parole, le carré du spectre d'amplitude permet de ressortir des pics spectraux qui correspondent à la fréquence fondamentale et leurs harmoniques.

La troisième étape de la méthode DSPM consiste à localiser tous les pics générés par l'analyse spectrale du PM, puis à rechercher pour chaque pic la série des harmoniques qui lui est associée. La série de pics la plus homogène et contenant le plus de pics est validée. La fréquence fondamentale recherchée correspond à la position du premier pic de la liste validée.

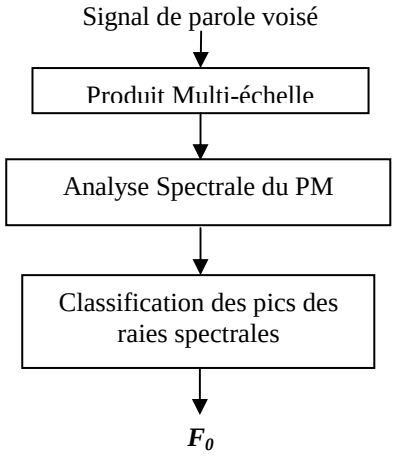


Figure 4 : Schéma block de l'approche proposée pour l'estimation de la fréquence fondamentale.

La Figure 5 illustre l'analyse spectrale du PM du signal de parole non bruité. Le PM est pondéré par la fenêtre de Hanning. Cette fonction donne un spectre de raies correspondant à la fréquence fondamentale et à ses harmoniques.

La Figure 6 donne l'analyse spectrale du produit multi-échelle du signal de parole corrompu par un bruit blanc gaussien de RSB = -5dB. On observe des raies claires sans bruit, la première raie représente la fréquence fondamentale et les autres correspondent aux harmoniques. On observe l'effet du produit à réduire le bruit sur le carré du spectre d'amplitude du PM.

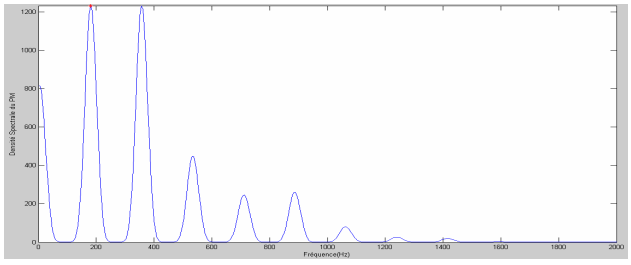


Figure 5 : Analyse spectrale du produit multi-échelle d'un signal de parole voisé non bruité. Le premier pic donne une estimation de la fréquence moyenne du pitch sur une fenêtre de 1024 points.

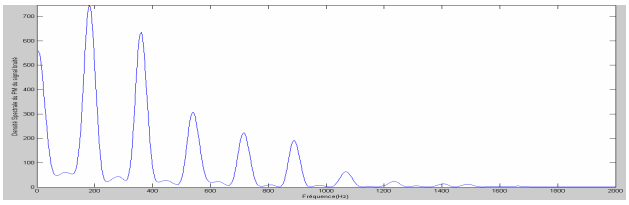


Figure 6 : Analyse spectrale du produit multi-échelle d'un signal de parole voisé en présence de bruit blanc gaussien (RSB=-5dB). Le premier pic donne une estimation de la fréquence moyenne du pitch sur une fenêtre de 1024 points.

4. EVALUATION DE LA MÉTHODE DSPM

La méthode que nous avons proposée est évaluée sur la base de son de l'université de Keele [12]. Elle est composée de 5 voix féminines et 5 voix masculines. Ces signaux de parole sont échantillonnés à la cadence de 20 kHz. L'algorithme proposé est évalué sur les zones voisées données par la base de sons utilisée.

4.1. Estimation de F0 dans un environnement non bruité

Pour l'évaluation des méthodes d'estimation du pitch, Rabiner [13] propose de calculer le taux d'erreurs grosses (GPE) défini comme le pourcentage des segments pour lesquels le F_0 estimé et le F_0 référence diffère par plus de 20%.

Pour se faire une idée de l'efficacité de la méthode proposée, nous avons comparé ses performances à celles d'autres méthodes. Comme le montre le Table 1, la méthode proposée DSPM surpasse les autres méthodes avec le taux le plus faible d'erreurs grossières. Ces performances nous motivent pour utiliser la méthode DSPM en présence de bruit.

Dans notre méthode DSPM, nous n'avons appliqué aucun post-traitement parce que l'analyse multi-échelle a réduit nettement les erreurs grossières et donc celles de dédoublement de pitch.

Table 1 : Taux des erreurs grossières pour l'estimation de F_0 en utilisant la base de Keele.

Méthodes	**GPE (%)**
DSPM	0.67
NMF [14]	0.9
MLS [15]	1.5
RAPT [4]	2.2
YIN [6]	2.35
PRAAT [5]	3.1
RCEPS [7]	3.95

4.2. Estimation de F0 dans un environnement bruité

Les taux des erreurs grossières pour l'estimation de F_0 en présence de bruit blanc gaussien par l'algorithme DSPM sont reportés dans le Table 2.

Le Table 2 montre la robustesse de notre algorithme DSPM pour l'estimation du pitch en présence de bruit. En effet, la méthode DSPM garde malgré la puissance du bruit de RSB -5dB, un taux d'erreurs grossières assez faible de l'ordre de 1.3% largement inférieur aux taux donnés par les méthodes RCEPS, PRAAT, YIN et NHMM.

Table 2 : Taux des erreurs grossières pour l'estimation de F_0 en présence de bruit blanc utilisant la base de Keele.

RSB (dB)	5					0					-5				
Méthodes	**DSPM**	NHMM [16]	YIN [6]	PRAAT [5]	RCEPS [7]	**DSPM**	NHMM [16]	YIN [6]	PRAAT [5]	RCEPS [7]	**DSPM**	NHMM [16]	YIN [6]	PRAAT [5]	RCEPS [7]
GPE (%)	0.93	1.2	3.9	4.6	5.3	1.14	1.28	5.1	6.1	6.6	1.32	1.43	5.9	6.2	7.1

5. CONCLUSION

Dans ce travail, nous avons conçu et développé un algorithme qui permet l'estimation du pitch du signal de parole sans bruit et en présence de bruit basée sur le calcul du carré du spectre d'amplitude du produit multi-échelle.

La méthode consiste à calculer le produit des coefficients de la transformée en ondelettes du signal de parole aux échelles ½, 1 et 2 avec l'ondelette spline quadratique de support 0.8 ms. Le signal produit obtenu est pondéré par la fenêtre de Hanning. Pour chaque fenêtre, nous calculons le carré du spectre d'amplitude du signal pondéré.

Les performances de la technique proposée sont évaluées en utilisant la base de Keele sans bruit et en présence de bruit blanc gaussien à différents RSB. La méthode montre son efficacité et sa robustesse en comparaison avec d'autres algorithmes.

BIBLIOGRAPHIE

[1] M. Brookes, P.A. Naylor and J. Gudnason. A quantitative assessment of group delay methods for identifying glottal closures in voiced speech. *IEEE Trans. Audio, Speech, and language Process*, 14:456-466, 2006.

[2] P.J.B. Jackson and C.H. Shadle. Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. *IEEE Trans. Acoust., Speech, Signal Process*, 9:713-726, 2001.

[3] W. Hess. *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer, Berlin, 1983.

[4] D. Talkin. A robust algorithm for pitch tracking (RAPT). In *Proc. Intl. Conf. on Speech Coding and Synthesis, Elsevier, Amsterdam*, pages 495-518, 1995.

[5] P. Boersma. PRAAT, a system for doing phonetics by computer. In *Proc. Intl. Conf. on Glot*, volume 5, pages 341-345, 2001.

[6] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Amer*, 111:1917–1930, 2002.

[7] A.M. Noll. Cepstrum pitch determination. *J. Accoust. Soc. Amer*, 41:293-309, 1967.

[8] T. Shimamura and H. Takagi. Noise-robust fundamental frequency extraction method based on exponentiated band-limited amplitude spectrum. In *Proc. Intl. Conf. on Symposium on Circuits and Systems*, pages 141-144, 2004.

[9] A. Bouzid and N. Ellouze. Open quotient measurements based on multiscale product of speech signal wavelet transform. *Research Letters in Signal Processing, Hindawi Publisher*, 2007:5 pages, 2007.

[10] M.A. Ben Messaoud, A. Bouzid and N. Ellouze. Spectral multi-scale product analysis for pitch estimation from noisy speech signal. *Advances in NonLinear Speech Processing, Springer*, 5933:95-102, 2010.

[11] B.M. Sadler and A. Swami. Analysis of multi-scale products for step detection and estimation. *IEEE Trans. Information. Theory*, 45:1043-1051, 1999.

[12] F. Plante, G. Meyer and W.A. Ainsworth. A pitch extraction reference database. In *Proc. Intl. Conf. on EUROSPEECH 95*, pages 837-840, 1995.

[13] L.R. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Trans. Acoust., Speech, Signal Process*, 25:24-33, 1977.

[14] F. Sha and L.K. Saul. Real time pitch determination of one or more voices by nonnegative matrix factorization. In *Proc. Intl. Conf. on Advances in Neural Information Processing Systems*, volume 17, pages 1233-1240, 2005.

[15] F. Sha, J.A. Burgoyne and L.K. Saul. Multiband statistical learning for F0 estimation in speech. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 661-664, 2004.

[16] D. Joho, M. Bennewitz and S. Behnke. Pitch estimation using models of voiced speech on three levels. In *Proc. Intl. Conf. on Acoust, Speech, Signal Process*, volume 4, pages 1077-1080, 2007.

Modéliser un locuteur : Influence des signaux d'apprentissage sur les performances d'un système de RAL

Juliette Kahn[1,2], Nicolas Audibert[1], Solange Rossato[2], Jean-François Bonastre[1]

[1]Université d'Avignon et des Pays de Vaucluse, CERI/LIA, France
[2]Universités de Grenoble, Laboratoire Informatique de Grenoble (LIG),France

ABSTRACT

Speaker verification systems (SVS) have shown significant progress and have reached a level of performance that support their use in practical applications. This context emphasizes the importance of a deeper analysis of the system's performance over basic error rate. In this paper, the effect of the training excerpt is investigated on ALIZE/SpkDet. The results show that the SVS performances are highly dependant on the voice samples used to train speaker models. Phonemic distribution differences between the excerpts are not sufficient to explain the range from 1% of EER to 33% on BREF 120.

Keywords: speaker verification, evaluation, phonetic information

1. INTRODUCTION

Les systèmes de Reconnaissance Automatique du Locuteur (RAL) sont évalués internationalement tous les ans depuis 1996 par le National Institute of Standards and Technology (NIST) [1]. Les performances des systèmes ont largement progressé ces dernières années. Leur niveau de performance (~1% d'EER pour les segments longs [2]) permet aujourd'hui d'envisager des applications concrètes.

La performance des systèmes est estimée à partir de deux types d'erreurs potentielles. Dans le cas d'un Faux Rejet (FR), le fichier test a bien été produit par le locuteur modélisé (test cible) mais le système considère l'hypothèse inverse. Dans le cas d'une Fausse Acceptation (FA), alors que l'auteur du fichier test est différent du locuteur cible (test imposteur), le système les considère comme identiques. Ces deux types d'erreurs sont liés par le seuil choisi pour prendre la décision.

Pour comparer les performances de différents systèmes, la courbe DET, le taux d'Egal Erreur (EER) et la fonction de coût de décision (DCF) sont le plus couramment utilisés [3]. La courbe DET représente l'évolution des deux types d'erreur en fonction du seuil. L'EER correspond au point où le taux de FA est égal au taux de FR, le DCF introduisant une fonction de coût. Toutes ces mesures sont calculées globalement sur une grand nombre de tests cible et imposteur.

Plusieurs études se sont intéressées aux facteurs potentiellement générateurs d'erreur pour ces systèmes de RAL. Doddington *et al.*[4] ont mis en évidence le fait que certains locuteurs entraînent plus ou moins des erreurs. D'autre part, Fauve *et al*. [9] ont fait varier les durées du signal de test et d'apprentissage et ont montré l'importance du durée du signal d'apprentissage. Outre la durée, certains phonèmes permettraient de mieux reconnaître le locuteur [11].

Etant donné que dans les évaluations NIST, un même locuteur peut être modélisé par plusieurs signaux, de contenu segmental sûrement différent, nous pouvons nous interroger sur la différence de performance des modèles, d'un même locuteur, construits à partir de ces différents signaux. Les expériences décrites dans cet article tentent de quantifier la variation de performance en fonction du choix du signal d'apprentissage pour un locuteur donné. Il s'agit ici de rendre compte de la précision du système avec des signaux de longueur constante dans un premier temps puis avec des signaux dans lesquels le nombre de trames sélectionnées par le système est constant. La variation due au type de locuteur n'est pas étudiée dans cet article, pour chaque série d'expériences les locuteurs restant les mêmes.

2. SYSTÈME UTILISÉ

Les expériences ont été réalisées à l'aide du système libre de droit ALIZE/SpkDet [5] qui montre de bonnes performances lors des campagnes NIST-SRE [2]. Il se fonde sur une approche UBM/GMM [6] et peut inclure les techniques du Factor Analysis [7] afin de modéliser la variabilité inter-session.

3. VARIATION DE PERFORMANCES EN FONCTION DES SIGNAUX D'APPRENTISSAGE

Notre objectif ici est de quantifier les variations de performance d'un système de RAL en fonction des fichiers apprentissage utilisés pour modéliser le locuteur. Nous effectuons une permutation apprentissage/test pour évaluer l'importance relative des signaux d'apprentissage et de test sur les performances d'un système.

3.1. Matériel et Méthode

Corpus NIST-08

Le corpus NIST-08 utilisé, est constitué d'enregistrements de parole téléphonique (conditions short 2-short 3) d'une durée de 2,5 min de la campagne NIST 2008. A l'origine,dans ce corpus, 221 hommes sont modélisés par 648 modèles différents.

Utilisation maximale de NIST-08 : M-08

Pour augmenter le nombre de modèles par locuteur, nous avons construit la base M-08 à partir des données NIST-08 par une procédure de *leave-one-out*. Chaque fichier d'apprentissage de NIST-08 ainsi que les fichiers

test ayant servi en tests cible dans NIST-08 ont été utilisés pour créer un modèle de locuteur différent. Afin d'analyser les variations inter-modèles pour un locuteur donné, nous avons exclu les 50 locuteurs représentés par moins de 3 modèles.

M-08 comprend alors 171 locuteurs représentés au total par 816 modèles. Chaque modèle est testé avec l'ensemble des fichiers sélectionnés précédemment, excepté celui ayant servi à la construction du modèle considéré, ce qui conduit à 661 416 tests imposteur et 3 624 tests cible.

Sélection des pires et des meilleurs modèles

Afin de mettre en évidence la variation inter-modèles, nous déterminons pour chaque locuteur le pire et le meilleur modèle. Le meilleur et le pire sont les modèles qui génèrent respectivement le moins d'erreurs (FA+FR) et le plus d'erreurs pour le seuil de l'EER de M-08. Nous obtenons alors 2 séries de modèles appelées *Modèles-Min* et *Modèles-Max* comportant respectivement les meilleurs et les pires modèles.

Nous comparons les performances de ces deux séries de modèles avec la série proposée par NIST-08.

Pour que chaque modèle du même locuteur soit comparé au même ensemble de fichiers test, les fichiers retenus dans Modèles *Min* et *Max* et utilisés en test dans NIST-08 n'ont pas été utilisés comme tests ici. Ces contraintes conduisent à mettre en place un protocole contenant 511 tests cible et 2 856 tests imposteur pour chacune des 3 conditions suivantes :

- *NIST-3 :* les fichiers d'apprentissage sont ceux définis dans le protocole NIST-08.
- *Min* : les fichiers d'apprentissage utilisés sont ceux de *Modèles-Min*
- *Max* : le fichier d'apprentissage utilisé sont ceux de *Modèles-Max.*

Permutation des fichiers d'apprentissage et de test

La question de la symétrie entre signaux test et signaux d'apprentissage est envisagée pour évaluer leur relative importance dans les performances obtenues. Si cette symétrie est avérée cela signifie que les facteurs d'erreur sont à chercher dans l'étude du couple signal d'apprentissage/signal de test. Si nous observons des différences de performance entre une série et sa série miroir cela pondère l'importance de ces signaux. Nous avons créé par permutation les séries « miroir » de *NIST-03*, *Min* et *Max* appelées respectivement *NIST-03-inv, Min-inv* et *Max-inv*. Dans ces 3 séries, les fichiers d'apprentissage de *NIST-03*, *Min* et *Max* sont utilisés comme tests et les fichiers test sont utilisés comme fichiers d'apprentissage.

3.2. Résultats

Influence du modèle du locuteur

La Figure 1 présente les courbes DET pour chacune des 3 séries pour lesquelles chaque locuteur est modélisé soit par le modèle proposé par NIST (série *NIST-3*), soit par le meilleur modèle (série *Min*), soit par le pire modèle (série *Max*). La série *Min* obtient un EER de 4,1% tandis que la série Max a un EER de 21,9%. La série NIST-3 obtient un EER de 12,1%. Ainsi, en fonction du modèle sélectionné pour représenter le locuteur, le taux d'EER peut varier du simple au quintuple.

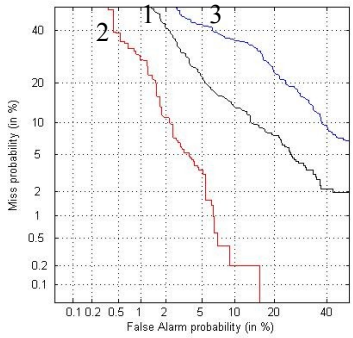


Figure 1: Courbes DET pour les 3 séries de tests *NIST-03* (1), *Min* (2) et *Max* (3). EER_{NIST}=12,1%, EER_{Min}=4,1%, EER_{Max}=21,9%

Permutation

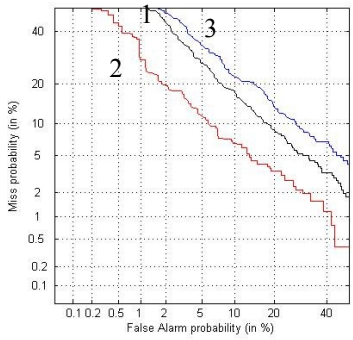


Figure 2: Courbes DET pour les 3 séries de tests *NIST-03-inv* (1), *Min-inv* (2) et *Max-inv* (3). $EER_{NIST-inv}$=13,5%, $EER_{Min-inv}$=7,4%, $EER_{Max-inv}$=17,0%

La Figure 2 présente les courbes DET pour chacune des 3 séries « miroir » *NIST-3-inv*, *Min-inv* et *Max-inv*. La série *Min-inv* obtient un EER de 7,4% tandis que la série Max a un EER de 17,0%. La série NIST-3 obtient un EER de 13,5%.

3.3. Discussion

Le choix du modèle permet de modifier considérablement les performances globales du système alors que ce sont les mêmes locuteurs qui sont modélisés dans chacune des séries *Min*, *Max* et *NIST-03*. Ici, pourtant, tous les signaux utilisés pour l'apprentissage des modèles sont enregistrés suivant le protocole établi par NIST, avec des conditions d'enregistrement similaires, même si la quantité d'information sélectionnée entre *Min* et *Max* est significativement différente ($t(170)=11.11$, $p<0.001$).

La permutation des fichiers test et apprentissage montre que, même si la tendance reste la même, les écarts de performance sont beaucoup plus importants en faisant

varier les signaux d'apprentissage qu'en faisant varier les signaux test. Ceci souligne la nécessité d'étudier en priorité le fichier d'apprentissage.

Il est dès lors nécessaire de comprendre ce qui distingue les fichiers d'apprentissage constituant les séries *Min* et *Max* et qui pourraient expliquer ces différences de performance. Pour cela nous étudions tout d'abord le contenu phonétique des différents signaux d'apprentissage.

4. Influence du contenu phonétique sur la vérification du locuteur

Pour tester si le contenu phonétique pourrait expliquer les différences de performance, nous utilisons le corpus BREF 120 en raison de sa qualité d'enregistrement et de la possibilité d'obtenir un alignement des signaux. Nous nous abstrayons ici de ce paramètre et testons l'influence du contenu phonétique avec un nombre de trames sélectionné constant.

4.1. Matériel et Méthode

Corpus BREF 120

Nous avons conservé les 64 femmes et 43 hommes français natifs du corpus BREF 120 [8]. Pour chaque locuteur, nous avons sélectionné aléatoirement des phrases parmi celles prononcées afin de créer 39 fichiers d'au moins 30 secondes de parole au sens du système de RAL (et non 30 secondes de durée de signal) afin d'obtenir un nombre de trames sélectionné constant. 18 ont servi à la création des modèles du locuteur, 21 autres fichiers ont été réservés comme fichiers test.

Tous les croisements sont effectués ce qui nous conduit à 17 766 tests pour chaque locuteur homme (378 tests cible et 17 388 tests imposteur) et 24 192 tests pour chaque locuteur femme (378 tests cible contre 23 814 tests imposteur). Au total, ont été réalisés 835 002 tests pour les hommes et 1 548 288 tests pour les femmes.

Paramétrage d'ALIZE/SpkDet

Au vu des conditions d'enregistrement du corpus BREF 120 (une seule session par locuteur et même matériel d'enregistrement), nous n'avons pas utilisé la technique du Factor Analysis sur ce corpus.

Sélection des pires et des meilleurs modèles

Les tests effectués nous ont permis de sélectionner pour chaque locuteur le pire et le meilleur modèle, qui engendrent respectivement le plus et le moins de FA et de FR pour le seuil de l'EER de l'ensemble des tests.

Pour les hommes, les séries *Min-Hommes* et *Max-Hommes* comportent chacune 43 modèles (un par locuteur). Les séries *Min-Femmes* et *Max-Femmes* comprennent 64 modèles. Nous avons également effectué des tirages aléatoires d'un modèle par locuteur comme base de comparaison. Pour les hommes comme pour les femmes, les signaux test sont exactement les mêmes quelque soit la série de modèles sélectionnés.

Analyse des fichiers d'apprentissage

Pour obtenir une transcription phonétique, un alignement forcé des fichiers sons a été réalisé de manière semi-automatique à l'aide du logiciel *Speeral* [10].

Nous avons regroupés en 10 catégories les phones présents dans chaque fichier d'apprentissage. Les trames sélectionnées sont classées dans les catégories suivantes : occlusives sourdes (OS), et sonores (OV), fricatives sourdes (FS) et sonores (FV), approximantes et latérales (A-L), consonnes nasales (CN), voyelles orales fermées (VOF), orales médianes et ouverte (VOMO), nasales (VN) et non-parole. La catégorie non-parole rassemble toutes les trames sélectionnées par le système de RAL mais qui n'appartiennent à aucune catégorie phonétique selon l'alignement. Le nombre de trames de chaque catégorie nous informe de la quantité d'information phonétique utilisée pour chaque modèle.

4.2. Résultats

Performance globale d'ALIZE sur BREF-120

Nous obtenons, lorsque tous les modèles sont testés, un EER de 8,8% pour les hommes et de 9,9% pour les femmes. Ces performances s'approchent de l'état de l'art des tests effectués lors des campagnes d'évaluation [9].

Pires et meilleurs modèles

Un EER de 1,0% est obtenu avec la série *Min-Hommes* tandis que l'EER de *Max-Hommes* s'élève à 33,0%. La même tendance est observée pour les femmes, avec un EER de 1,1% pour *Min-Femmes* contre 28,5% de EER pour *Max-Femmes*. Les 10 séries pour lesquelles les modèles de locuteur ont été choisis aléatoirement présentent un EER variant de 6,3% à 11,6% pour les hommes (m=9,0%, σ=1,4) et 8,8% à 11,5% pour les femmes (m=10,3%, σ=1,1). La Figure 3 présente les courbes DET correspondantes pour les hommes.

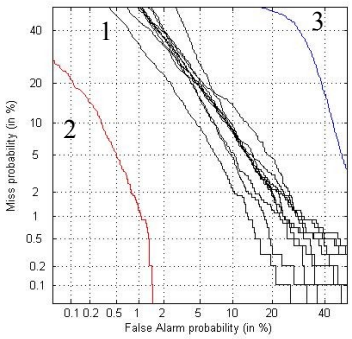


Figure 3: Courbes DET pour les séries de tests *Min-Hommes* (2) et *Max-Hommes* (3) et 10 séries tirées au hasard (1). 6,3%<EER_{Random}<11,6%, $EER_{Min\text{-}Hommes}$=1,0%, $EER_{Max\text{-}Hommes}$=33,0%

Les écarts de performance sont ici encore plus importants que pour NIST. Or, les locuteurs et le nombre de trames sélectionnées pour construire les modèles sont les mêmes. Ces variations étant comparables pour les hommes et pour les femmes. Un facteur explicatif pourrait être

l'information phonétique contenue dans les signaux *Min* et *Max*.

Analyse phonétique des fichiers d'apprentissage

La quantité d'information phonétique présente dans les signaux est ici simplement représentée par la répartition du nombres de trames sélectionnées par le système suivant les catégories phonétiques. Ces répartitions sont comparées pour les signaux des différentes séries.

Une MANOVA par genre a été réalisée. La variable dépendante est la répartition en nombre de trames dans chaque catégorie phonétique. Le facteur fixé est la performance. La répartition des distributions des deux groupes *Min* et *Max* ne montre aucune différence significative (femmes : F(10,17)=1,52;p=0,142, hommes : F(10,83)=1,51;p=0,150).

Des t-tests appariés entre meilleurs et pires modèles ont été réalisé pour chaque catégorie phonétique. Seules les consonnes nasales pour les femmes (p=0,025) et les fricatives sonores (p=0,037) sont en quantité significativement différente. La Figure 4 illustre la distribution des trames en fonction des différentes catégories phonétiques pour les femmes.

5. Discussion

Sur deux corpus différents, les performances d'un système de RAL montrent des variations très importantes en fonction du signal d'apprentissage. Une telle variation ne peut s'expliquer ni par la durée des signaux (section 3), ni par le nombre de trames sélectionnées par le système (section 4). Si certaines catégories phonétiques sont légèrement plus représentées entre nos séries *Min* et *Max*, cela ne suffit pas à expliquer de telles différences de performance. Plutôt que dans le nombre de trames par catégorie, qui n'est qu'une information très limitée, les différences entre les signaux de la série Min et série Max résident peut-être dans l'information spectrale fournie. La question des facteurs influençant la performance d'un système reste largement ouverte.

Ce travail a été en partie financé par le projet européen MOBIO (http://www.mobioproject.org) et par la Direction Générale de l'Armement.

Bibliographie

[1]A. Martin et M. Przybocki, "NIST speaker recognition evaluation chronicles", in *Odyssey*, 2004.

[2]D. Matrouf, J-F. Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLaren et F. Huenupan, "LIA GMM-SVM system description: NIST SRE," in *Odyssey NIST SRE*, Montreal (Canada), 2008.

[3]A, Martin, G. Doddington, T. Kamm, M. Ordowski, et M.A Przybocki, "The det curve in assessment of detection task performance," in Eurospeech 1997, 1997.

[4]G. Doddington, W. Liggett, A. Martin, M. Przybocki, et D. Reynolds, 1998, "Sheep, goats, lambs and wolves, an analysis of individual differences in speaker recognition performances in the NIST 1998 speaker recognition evaluation," *in ICSLP '98,* Sydney, 1998.

[5]J.F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans**,** 2007**"**ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition**"** *proceeding Interspeech 2007*.

[6]D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[7]P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel,"Factor analysis simplified," *in ICASSP, 2005*

[8]L.F. Lamel, J.L. Gauvain, M. EskEnazi, "BREF, a Large Vocabulary Spoken Corpus for French," *Proceedings of EUROSPEECH 91,* Genova, 24-26 September 1991, Vol.2, pp. 505-508.

[9]B. Fauve, N. Evans, J. Mason, 2008, "Improving the performance of text-independent short duration SVM- and GMM-based speaker verification" *in Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2008.

[10]G. Linares, P. Nocera, D. Massonie, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Lecture Notes in Computer Science*, 2007.

[11]I. Magrin-Chagnolleau, J-F. Bonastre, F. Bimbot, "Effect of Utterance duration and Phonetic content on speaker identification using second-order statistical methods", *EUROSPEECH 1995.*

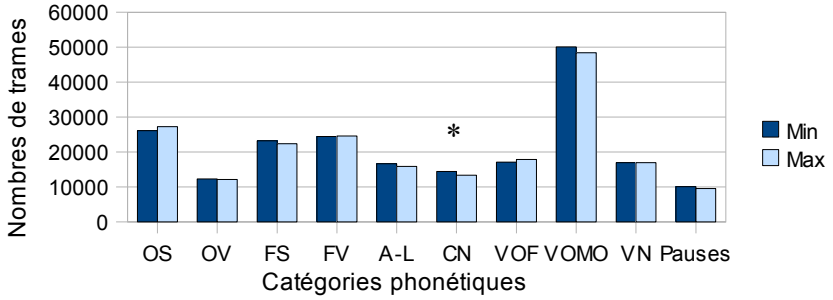


Figure 4: Distributions des catégories phonétiques pour les séries Min-Femmes et Max-Femmes

Utilisation conjointe de modèles locaux et globaux pour la caractérisation et la détection de segments de parole spontanée *

Richard Dufour [1] ,Yannick Estève [1], Paul Deléglise [1], Frédéric Béchet [2]

[1] LIUM - Université du Maine, Le Mans, France
prénom.nom@lium.univ-lemans.fr
[2] LIF - Université de la Méditerranée, Marseille, France
prénom.nom@lif.univ-mrs.fr

ABSTRACT

Processing spontaneous speech is one of the many challenges that Automatic Speech Recognition (ASR) systems have to deal with. The main evidences characterizing spontaneous speech are disfluencies (filled pause, repetition, repair and false start), and many studies have focused on the detection and the correction of these disfluencies. Ungrammaticality and language register are also important as well as prosodic patterns. In this study we define spontaneous speech as unprepared speech, in opposition to prepared speech where utterances contain well-formed sentences close to those that can be found in written documents. This paper proposes a set of acoustic and linguistic features that can be used for characterizing and detecting spontaneous speech segments from large audio databases. More, we introduce a strategy that takes advantage of a global classification process using a probabilistic model which significantly improves the spontaneous speech detection.

Keywords: spontaneous speech characterization, spontaneous speech detection, speech recognition

1. Introduction

Différents types de parole se retrouvent dans les émissions radiophoniques, notamment la parole spontanée. Ainsi, il serait intéressant de détecter les passages de parole spontanée afin de découper des documents en fonction de leur contenu et de leur structure. Il serait de plus utile de détecter au plus tôt les segments de parole spontanée dans le but d'adapter et de structurer les processus des Systèmes de Reconnaissance Automatique de la Parole (SRAP) à ce type de parole, comme présenté dans [6].

La parole spontanée apparaît dans les émissions radiophoniques sous différentes formes : débats, interviews, dialogues, etc. Les disfluences sont les principaux indices de ce type de parole (morphèmes spécifiques comme *euh*, répétitions, réparations et faux départs) et des études se sont focalisées sur la détection de ces disfluences [9, 8]. Toutes ces études montrent une chute importante des performances entre les résultats obtenus sur les transcriptions de référence et ceux obtenus sur les transcriptions automatiques. Cette différence peut s'expliquer par le bruit généré par les systèmes de RAP sur les segments de parole spontanée, avec un taux d'erreurs mots (WER) plus élevé que sur la parole préparée.

A ces disfluences s'ajoutent l'agrammaticalité et un registre de langage différent de celui retrouvé dans les textes écrits [4]. En fonction du locuteur, de l'état émotionnel et du contexte, le langage utilisé peut être très différent. Dans cette étude, nous définissons la parole spontanée comme de la parole non préparée, en opposition à la parole préparée où les discours contiennent des phrases bien formées, proches de celles pouvant être retrouvées dans des documents manuscrits. Nous avons déjà proposé dans un précédent travail une série de caractéristiques acoustiques et linguistiques pour caractériser la parole spontanée [6].

Dans cet article, nous présentons les évolutions de cette étude. En particulier, nous présentons une approche globale prenant en compte une modélisation probabiliste améliorant significativement la détection de la parole spontanée.

2. Caractérisation de la parole spontanée

2.1. Niveaux de spontanéité

La classification de la parole préparée/spontanée est très subjective. Idéalement, pour pouvoir étiqueter le niveau de spontanéité de chaque segment de parole d'un corpus audio, il faudrait demander à chaque locuteur d'annoter ses propres interventions. Cet étiquetage est irréaliste, mais il est cependant possible de définir un protocole d'annotation. Ce protocole est établi sur la perception d'un juge humain définissant un niveau de spontanéité à un segment donné. Notre approche a été d'étiqueter manuellement un corpus contenant des segments de parole avec une série de huit étiquettes, chacune correspondant à un niveau de spontanéité : le degré 1 pour la parole préparée, presque similaire à de la parole lue, jusqu'au degré 8 pour de la parole presque incompréhensible. Cette approche nous permet de choisir subjectivement la limite entre la parole préparée et spontanée. Dans nos expériences nous considérons trois classes : *parole préparée* correspondant au degré 1 ; *légèrement spontanée* correspondant aux degrés compris entre 2 et 4 ; *fortement spontanée* correspondant aux degrés 5 et plus.

Deux annotateurs humains ont annoté un corpus de

*Ces travaux sont financés par l'ANR sous le contrat numéro ANR-06-MDCA-006.

parole en écoutant des enregistrements audio. Le corpus a été divisé en segments au moyen d'un outil de segmentation automatique et de regroupement en locuteurs [5]. Aucune transcription n'a été fournie aux annotateurs. Afin d'évaluer l'accord inter-annotateur pour cette tâche d'annotation, nous avons calculé le coefficient Kappa [3] sur une heure d'émissions radiophoniques. Le coefficient obtenu était très haut : 0,852 — une valeur supérieure à 0,8 étant habituellement considérée comme excellente [7].

Le corpus obtenu suite à ce processus d'étiquetage a été produit à partir de 11 fichiers contenant des émissions radiophoniques françaises issues de 5 media différents (France Culture, France Inter, France Info, Radio Classique, RFI). Les fichiers ont été choisis de manière à contenir de la parole spontanée. La durée globale est de 11h37 pour un total de 11 821 segments (en enlevant les segments sans parole). Parmi tous ces segments, 3 670 ont été annotés avec la classe *parole préparée*, 4 107 avec la classe *légèrement spontanée* et 4 044 avec la classe *fortement spontanée*.

2.2. Description des caractéristiques

Dans ce travail, comme décrit dans [6], nous utilisons trois séries de caractéristiques : des caractéristiques acoustiques (prosodiques), des caractéristiques linguistiques, ainsi que des mesures de confiance. Nous les combinons afin de caractériser la classe de spontanéité de chaque segment : cette tâche est différente de la tâche de détection de parole disfluente, car les segments de parole spontanée ne contiennent pas obligatoirement des disfluences (e.g. forte variation dans le débit de parole).

Caractéristiques prosodiques Les caractéristiques prosodiques utilisées sont liées à la durée des voyelles ainsi qu'au débit phonémique.

Durée : en suivant le travail précédent décrivant le lien entre la prosodie et la parole spontanée [11], nous utilisons la durée des voyelles ainsi que la longueur d'une syllabe en fin de mot. Ce dernier a été proposé dans [2] et est associé au concept de *mélisme*. En sus de leurs durées moyennes, leur variance et leur écart type ont également été ajoutés comme caractéristiques.

Débit phonémique : des études précédentes [2] ont montré la corrélation entre les variations du débit de parole et l'état émotionnel d'un locuteur. Nous utilisons alors le débit de parole de chaque segment, afin de pouvoir observer son impact sur le niveau de spontanéité. Nous calculons le débit phonémique de deux manières : la moyenne et la variance de ce débit sur le segment global, en incluant dans un premier temps les pauses et les morphèmes spécifiques (*euh*, *ben...*), puis en les excluant.

Caractéristiques linguistiques Le concept des *disfluences* représente la principale caractéristique de la parole spontanée. Elles peuvent être catégorisées comme des pauses, des répétitions, des réparations ou des faux départs. De nombreuses études se sont focalisées sur leur description au niveau acoustique [11] et au niveau lexical [12]. Nous utilisons deux caractéristiques les représentant dans les segments de parole :

– morphèmes spécifiques : le lexique du système de RAP contient de nombreux symboles représentant les hésitations en français (e.g. euh, ben, hum). Leur nombre d'occurences est notre première caractéristique.
– répétition et faux départ : nous calculons le nombre de répétitions d'unigramme et bigramme dans le segment.

Comme défini par [4] sur les enregistrements d'émissions radiophoniques, l'agrammaticalité et un registre de langage différent de la parole préparée caractérisent également la parole spontanée. Pour montrer le lien entre le niveau de spontanéité, et le lexique et la syntaxe, nous appliquons un processus d'analyse syntaxique sur les segments de parole, incluant un étiquetage en partie du discours (POS) et un processus de découpage syntaxique. Les caractéristiques suivantes sont utilisées pour décrire les segments :

– paquets de n-grammes (de 1 à 3) sur les mots, étiquetage POS et découpage syntaxique en catégorie (groupe nominal et prépositionnel)
– taille moyenne des découpages syntaxiques des segments

De plus, comme présenté dans [1], nous utilisons le nombre de noms propres contenu dans un segment, un grand nombre de noms propres étant une caractéristique de la parole préparée.

Mesures de confiance Les mesures de confiance (MC) sont des scores exprimant la fiabilité des décisions de reconnaissance prises par un système de RAP. Ces scores sont utilisés pour caractériser la spontanéité des segments de parole, car, comme vu dans [6], les SRAP ont beaucoup plus de difficultés à transcrire des segments de parole spontanée que des segments de parole préparée.

3. Détection automatique des segments de parole spontanée

Afin d'extraire automatiquement les descripteurs acoustiques, linguistiques et les MC, nous utilisons le système de RAP du LIUM [5]. Ce système a été développé pour participer à la campagne d'évaluation ESTER 2 sur la transcription automatique d'émissions radiophoniques.

Les caractéristiques présentées précédemment sont utilisées lors d'une tâche de classification permettant de classifier notre corpus selon nos trois classes de spontanéité. L'outil de classification utilisé est *icsiboost*, un outil open source s'appuyant sur l'algorithme AdaBoost, comme l'outil *Boostexter* [10].

Chaque segment est catégorisé individuellement au cours du processus de classification, en prenant en compte la classe ayant le meilleur score.

4. Décision globale au moyen d'un modèle probabiliste

Notre précédente approche [6] prenait en compte les descripteurs extraits à l'intérieur d'un segment, sans cependant prendre en considérant les segments voisins. Afin d'améliorer notre approche, nous proposons de nous intéresser à la nature des segments entourant

le segment à classifier. Cette vision implique que la catégorisation de chaque segment dépend de la catégorisation des segments voisins à ce segment : le processus de décision devient un processus global. Nous avons choisi d'utiliser une approche statistique classique en utilisant une méthode du maximum de vraisemblance, qui prendra en compte les scores de classification au niveau du segment, ainsi qu'un score de probabilité sachant les segments voisins.

Soit s_i une classe du segment i, avec $s_i \in$ {"*fortement spontanée*", "*légèrement spontanée*", "*préparée*"}. Nous définissons $P(s_i|s_{i-1}, s_{i+1})$ comme la probabilité d'observer le segment i associé à la classe s_i, sachant que le segment précédent est associé à la classe s_{i-1} et que le segment suivant est associé à la classe s_{i+1}. Soit $c(s_i)$ la mesure de confiance fournie par la classifieur *icsiboost* par rapport au choix de la classe s_i pour le segment de parole i, en prenant en compte les valeurs des caractéristiques extraites de ce segment. S est une séquence de classes s_i associée à la séquence de tous les segments de parole i (simplement une classe par segment). Le processus de décision globale consiste à choisir la classe-séquence hypothèse $\bar{S}$ qui maximise le score global obtenu en combinant $c(s_i)$ et $P(s_i|s_{i-1}, s_{i+1})$ pour chaque segment de parole i détecté sur le fichier audio. La séquence $\bar{S}$ est calculée en utilisant la formule suivante :

$$\bar{S} = \underset{S}{\operatorname{argmax}}\ c(s_1)\times c(s_n)\times \prod_{i=2}^{n-1} \times c(s_i)\times P(s_i|s_{i-1}, s_{i+1}) \quad (1)$$

où n est le nombre de segments de parole détectés automatiquement dans le fichier audio.

5. Expériences

Le corpus expérimental (partie 2.1) est constitué de 11 fichiers audio d'enregistrements radiophoniques. Durant ces expériences, nous avons utilisé la méthode du *Leave one out* : 10 fichiers ont été utilisés pour l'apprentissage, 1 pour l'évaluation, et ce processus a été répété jusqu'à ce que tous les fichiers aient été évalués.

5.1. Performances du système de RAP

Le tableau 1 présente les résultats en terme de taux d'erreurs mots (WER) et d'entropie croisée normalisée (NCE) du système de RAP du LIUM sur les données expérimentales. Ces données n'ont pas été incluses dans les corpus d'apprentissage et de développement du décodeur. Le WER est la métrique classique d'évaluation des SRAP, alors que le NCE est utilisé habituellement pour évaluer les mesures de confiance fournies par un SRAP.

Nous constatons que les performances globales du système de RAP, avec un WER de 15 % et un NCE de 0,331, sont bonnes pour la transcription d'émissions radiophoniques françaises. Comme attendu, plus la parole est fluide, plus le WER est bas : de 10,1 % pour les segments de parole annotés en *parole préparée*, jusqu'à 28,5 % pour les segments *fortement spontanés*. Notons qu'il existe une corrélation entre l'annotation subjective du niveau de spontanéité et le WER obtenu par le décodeur.

Tab. 1: Performances du système de RAP en fonction de la classe de parole en termes de WER et de NCE, ainsi que du nombre de segments.

classe de parole	# segments	WER	NCE
parole préparée	3 670	10,1 %	0,358
légèrement spontanée	4 107	18,4 %	0,315
fortement spontanée	4 044	28,5 %	0,237
all	11 821	15,0 %	0,331

5.2. Détection et catégorisation automatique de la parole spontanée

Afin de mesurer le gain fourni pour chaque sorte de descripteurs et le gain fourni grâce à l'utilisation d'une décision globale, quatre conditions ont été évaluées : les caractéristiques linguistiques sur la transcription de référence *ling(ref)*, toutes les caractéristiques (acoustiques et linguistiques) sur la transcription automatique *all(rap)*, avec l'utilisation d'un modèle probabiliste global sur les résultats fournis par *ling(ref)* : *ling+global(ref)*, et enfin avec l'utilisation d'un modèle probabiliste global sur les résultats fournis par *all(rap)* : *all+global(rap)*. Nous pouvons ainsi comparer les résultats sur les transcriptions de référence avec les transcriptions obtenues automatiquement par le SRAP. Une comparaison plus détaillée de chaque caractéristique utilisée séparément (*acous vs. ling* sur *ref* et *rap*) est disponible dans [6].

Le tableau 2 présente les résultats sur la détection (précision et rappel) pour chaque classe de spontanéité. Nous constatons que les performances sur la détection des segments *légèrement spontanés* sont basses, ce qui n'est pas surprenant car les segments peuvent facilement être faussement classifiés en *parole préparée* d'un côté, et *fortement spontanée* d'un autre côté.

Tab. 2: Précision et rappel de la classification des segments de parole en fonction des 3 classes de spontanéité.

parole préparée				
Carac.	ling (ref)	all (rap)	ling+global (ref)	all+global (rap)
Préc.	56,0	57,8	61.6	62,1
Rapp.	64,1	61,7	66.5	64,2
légèrement spontanée				
Carac.	ling (ref)	all (rap)	ling+global (ref)	all+global (rap)
Préc.	43,8	45,5	46.9	49,2
Rapp.	37,7	40,5	42.8	44,2
fortement spontanée				
Carac.	ling (ref)	all (rap)	ling+global (ref)	all+global (rap)
Préc.	65,2	65,5	70.3	**69,3**
Rapp.	65,9	68,8	71.5	**74,6**

En examinant les résultats, nous pouvons voir que le modèle contextuel probabiliste appliqué à *all(rap)* et *ling(ref)* permet d'améliorer significativement les performances de classification, peu importe la classe de spontanéité ou la métrique utilisée.

En comparant l'utilisation des caractéristiques linguistiques provenant des transcriptions manuelles

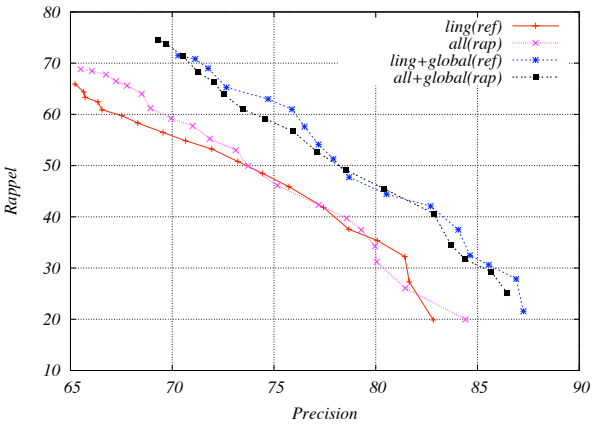


Fig. 1: Performances sur la détection des segments de parole *fortement spontanée* en fonction du seuil choisi sur le score de classification.

ling(ref), avec la fusion des caractéristiques linguistiques et acoustiques du système de RAP *all(rap)*, nous obtenons de meilleurs résultats quelque soit la classe de spontanéité, exception faite du rappel pour la parole *légèrement spontanée*. De plus, nous constatons que les résultats obtenus avec *all+global(rap)* sont légèrement meilleurs en général que ceux obtenus avec cette méthode globale sur les transcriptions manuelles *ling+global(ref)*.

En nous focalisant sur la détection de la classe de parole *fortement spontanée*, notre méthode de classification (locale et globale) permet d'atteindre 69,3 % en précision et 74,6 % en rappel. Nous avons également constaté que 83,5 % des erreurs de détection de la classe *fortement spontanée* sont dues à la confusion avec la classe *légèrement spontanée*.

En utilisant les scores $c(s_i)$ fournis par le classifieur et combinés avec les probabilités $P(s_i|s_{i-1}, s_{i+1})$ fournies par le modèle contextuel, il est possible de filtrer la proposition de classe en appliquant un seuil sur la valeur de $c(s_i) \times P(s_i|s_{i-1}, s_{i+1})$. La figure 1 présente les performances de détection de la classe de parole *fortement spontanée* obtenues en faisant varier ce seuil. Nous constatons que notre système peut être plus précis (la précision augmente) quand moins de décisions sont prises (rappel diminue). Cette possibilité de seuillage permet d'adapter l'utilisation de la méthode de classification en cherchant le meilleur compromis entre le rappel et la précision pour l'application souhaitée. La nouvelle approche globale *all+global(rap)* utilisant un modèle probabiliste surpasse nettement les résultats obtenus avec la précédente approche locale *all(rap)* : n'importe quel seuil donné avec cette nouvelle approche permet d'atteindre un meilleur rappel et une meilleure précision.

6. Conclusion

Nous proposons une série de caractéristiques acoustiques et linguistiques pour caractériser et détecter les segments de parole spontanée issus de données audio. Afin de mieux définir la notion de parole spontanée, une série de segments de parole représentant 11 heures de corpus d'émissions radiophoniques françaises a été étiquetée manuellement par niveau de spontanéité.

Les caractéristiques acoustiques et linguistiques extraites des sorties d'un système de RAP et combinées entre elles, permettent d'obtenir de meilleurs résultats sur le rappel et le précision que l'extraction seule de caractéristiques linguistiques issues des transcriptions manuelles. De plus, en utilisant un modèle contextuel probabiliste pour globaliser le processus de classification, nous obtenons une meilleure précision de 74,6 % et rappel de 69,3 % sur la détection de la parole *fortement spontanée*, où 83,5 % des erreurs de classification pour cette classe sont dues à la confusion avec la classe *légèrement spontanée*. En appliquant un seuil sur les scores obtenus pendant le processus de classification, la détection de la classe *fortement spontanée* peut atteindre 85 %, mais avec un rappel chutant à 25 %. Bien que la tâche de classification de segments de parole en fonction d'un niveau de spontanéité est difficile (même pour des annotateurs humains), beaucoup de progrès ont été faits dans la détection automatique depuis notre précédent travail [6], surtout avec l'ajout du processus de classification global.

Cette détection de parole spontanée fournit des informations très utiles qui pourront être utilisées dans différentes applications, comme par exemple en reconnaissance de la parole, en réalisant des traitements spécifiques sur la parole spontanée.

Références

[1] T. Bazillon, Y. Estève, and D. Luzzati. Manual vs assisted transcription of prepared and spontaneous speech. In *LREC 2008*, Marrakech, Morroco, 2008.

[2] G. Caelen-Haumont. Perlocutory Values and Functions of Melisms in Spontaneous Dialogue. *1st International Conference on Speech Prosody, SP*, pages 195–198, 2002.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 :37–46, 1960.

[4] P.B. de Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda, and P. Paroubek. A quantitative study of disfluencies in French broadcast interviews. In *Workshop Disfluency In Spontaneous Speech (DISS)*, Aix-en-Provence, France, 2005.

[5] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. Improvements to the LIUM French ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? In *Interspeech*, Brighton, Great Britain, 2009.

[6] R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linarès. Spontaneous speech characterization and detection in large audio database. In *SPECOM*, St Petersburg, Russia, 2009.

[7] B. Di Eugenio and M. Glass. The Kappa statistic : A second look. *Computational Linguistics*, 30(1) :95–101, 2004.

[8] M. Lease, Johnson M., and E. Charniak. Recognizing Disfluencies in Conversational Speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5) :1566–1573, 2006.

[9] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper. Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection. *Interspeech*, 2005.

[10] R. E. Schapire and Y. Singer. BoosTexter : A boosting-based system for text categorization. *Machine Learning*, 39 :135–168, 2000.

[11] E. Shriberg. Phonetic consequences of speech disfluency. *International Congress of Phonetic Sciences (ICPhS-99)*, pages 619–622, 1999.

[12] M.H. Siu and M. Ostendorf. Modeling disfluencies in conversational speech. *International Conference on Spoken Language Processing ICSLP*, 1, 1996.

Structures de frames sémantiques pour le dialogue Homme-Machine par processus de décision markoviens

Florian Pinault, Fabrice Lefèvre

CERI - Université d'Avignon
339, chemin des Meinajaries - 84911 AVIGNON Cedex 9, France
{florian.pinault, fabrice.lefevre, renato.demori}@univ-avignon.fr

ABSTRACT

Current Human-Machine dialogue managers are still not very robust against speech recognition errors. A common answer to this issue is by ensuring a complete control of the dialogue flow, leading to rigid and non-natural interactions. In order to allow user-driven dialogues, a richer semantic modeling is needed. Inspired from FrameNet, semantic frames trees are proposed in our work to model the dialogue states. Then the Partially Observable Markov Decision Process (POMDP) with summary space provides a general framework to address the incertainty issue from a statistical principled approach. In this paper, before field trials can be carried out, a first set of experiments with summary POMDP show good robustness to ASR errors when optimizing the dialogue policy.

Keywords: dialogue management, POMDP, semantic modeling, semantic frames

1. Introduction

Dans un sytème de dialogue homme machine, le module de compréhension (reconnaissance vocale et compréhension sémantique) fournit des informations sur ce qu'a dit l'utilisateur au gestionnaire de dialogue. Ce module maintient une estimation de l'état du dialogue (*croyance* ou *belief*) : il y intègre les nouvelles informations. Il applique ensuite une stratégie utilisant cette modélisation pour choisir l'action (la réponse) la plus appropriée. Il la transmet enfin au module de génération (génération de texte et synthèse vocale). L'optimalité d'une décision est définie en fonction de la réalisation finale du but recherché par l'utilisateur. Lors de l'optimisation de la stratégie, une planification est donc nécessaire afin d'anticiper le cours futur du dialogue. Cependant, les erreurs commises lors de l'étape de compréhension rendent cette tâche difficile. Ce papier présente un traitement statistique de l'incertitude liée aux erreurs de compréhension (modèle statistique de POMDP) appliqué à une modélisation de l'état de dialogue qui permet une réprésentation d'états de dialogue complexes (modèle sémantique de frames).

Issu du domaine de l'apprentissage automatique (*machine learning*) et du contrôle en robotique, les processus de décision de Markov partiellement observables (POMDP) ont été proposés [16] pour permettre une planification statistique, tout en modélisant l'incertitude sur l'état réel du dialogue. L'optimisation des POMDPs est restée longtemps difficile, mais des algorithmes de résolution approchée ont été récemment proposés. Ils sont de deux types : avec ou sans modèle, où le modèle en question est celui de la dynamique du dialogue (probabilités de transition d'état et d'observation), un modèle de l'état de dialogue est évidemment toujours nécessaire.

Les algorithmes sans modèle [17] peuvent fonctionner par interaction directe avec leur environnement (*on-line*). Ces algorithmes nécessitent généralement de grandes quantités de données annotées rarement disponibles. On peut générer de telles quantités de données grâce à des utilisateurs simulés [11], mais les limites du système obtenu sont alors celles du simulateur utilisé et de son réalisme.

Pour un algorithme avec modèle [15, 10] une possibilité, présentée dans cet article, est d'entrainer le modèle sur un corpus réaliste de dialogues entre un utilisateur et un magicien d'Oz (WoZ), puis d'optimiser et d'évaluer la stratégie dans ce modèle.

Dans la littérature, la modélisation de l'état de dialogue dépend du domaine de la tâche considérée. Pour un domaine généraliste (par ex. pour un test de Turing), des motifs reprenant les propres mots de l'utilisateur sont utilisés. Lorsque le domaine est restreint, l'état du dialogue est modélisé par un formulaire de requête à une base de données qui est rempli au cours du dialogue par les informations fournies par l'utilisateur (*slot-filling*). Une plus grande souplesse est nécessaire pour modéliser des états de dialogue complexes, par exemple lorsque le dialogue comporte des étapes de négociation. Dans cette logique l'approche suivie dans ce papier utilise des structures d'arbres de frames inspirées du paradigme FrameNet [1] telles que présentées dans [8].

La section 2 présente le modèle de représentation sémantique utilisé dans nos expériences jusqu'à la représentation des états de dialogue en arbre de frames. La section 3 offre une introduction au concept de POMDP et son application au dialogue par espaces résumés. Enfin, la section 4 montre les expériences réalisées qui confirment l'intérêt de cette approche pour rendre les systèmes plus robustes face aux erreurs de reconnaissance.

2. Modèle sémantique

La modélisation de l'état du dialogue par une structure d'arbre de frames sémantiques vise à apporter la souplesse qui manque aux systèmes par *slot-filling*. Les dialogues peuvent ainsi être plus fortement dirigés par l'utilisateur et non plus uniquement par la machine.

2.1. Arbres de frames sémantiques

La campagne d'évaluation EVALDA [4] a permis de réaliser un corpus de dialogues portant sur des informations touristiques. Le corpus MEDIA est constitué de 1257 dialogues (18831 tours de parole utilisateur) enregistrés par la technique du magicien d'Oz.

Transcriptions. A partir de l'enregistrement du signal acoustique, deux transcriptions exacte (manuelle) et automatique ont été effectuées. Le tableau 1 présente un extrait du corpus :

Tab. 1: *Transcriptions manuelles du corpus* MEDIA

woz :	Donnez le nom de la ville où vous souhaitez aller.
loc :	Je voudrais aller à Évreux.
woz :	À Dreux à quelle période souhaitez vous y séjourner ?
loc :	Euh c'est pas Dreux c'est Évreux.
woz :	À Évreux, à quelle période souhaitez vous...

Annotations en concepts. Le corpus comporte également des annotations sémantiques (manuelle et automatique) en concept-*mode*-valeur. Une étiquette *mode* est ajoutée à chaque concept, le *mode* d'un concept est positif (+), négatif (−), interrogatif (?) ou optionnel (∼). Le dictionnaire sémantique comporte 83 concepts. Le tableau 2 présente un extrait de l'annotation en concept du corpus MEDIA.

Tab. 2: *Annotations en concepts du corpus* MEDIA

transcription	concept	mode	valeur
euh c'est	null		
pas Dreux	localisation-ville	−	Dreux
c'est	null		
Évreux	localisation-ville	+	Évreux

Annotations en frames. Le corpus comporte enfin une annotation sémantique structurée par des arbres de frames sémantiques. Inspiré du paradigme de FrameNet [1], l'annotation en frame ajoute aux concepts des rôles sémantiques constituant les éléments des frames. L'ontologie MEDIA contient 21 frames et 86 éléments de frames. De plus, des liens entre les frames peuvent être établis : un élément de frame pointant sur une autre frame comme dans l'exemple du tableau 3.

Taux d'erreurs. La transcription automatique obtient un taux d'erreur mots sur le corpus de test MEDIA de 33.5%.

Les annotations automatiques en concepts construites à partir des mots ont été réalisées par réseaux Bayésien dynamiques (DBN). Le taux d'erreur concepts est de 21.3% sur les transcriptions exactes et de 43.4% sur les transcriptions automatiques [6].

Tab. 3: *Exemple simplifié d'annotation en frames*

transcription	Je voudrais réserver euh un hôtel plutôt euh vers Nice euh oui un hôtel à Nice pour deux nuits.
concepts	vouloir, + , ' ', réservation , + , ' ' hôtel , + , ' ' nom-ville , + , 'Nice' accept , + , ' ' hôtel , + , ' ' nom-ville , + , 'Nice' nombre-nuit , + , '2'
frames	F1 : VOULOIR(objet = F1) F2 : RESERVATION (établissement = F3, période = F4) F3 : HOTEL(ville = 'Nice' , nom = ' ') F4 : PERIODE (date-début = ' ', date-fin = ' ', durée = '2 jours')

Les annotations en frames ont été réalisées par un ensemble de règles logiques consistant à instancier des frames lorsque certains concepts ou mots sont présents, puis à fusionner ou relier les frames obtenues. L'approche statistique basée sur des DBN proposée par [8] permet d'estimer l'incertitude sur les frames sous la forme de liste scorée de $n - best$ et sera utilisée prochainement.

2.2. Mémoire de frames

Les graphes de frames sont agrégés à chaque tour de parole dans un même arbre sémantique, appelé mémoire de frame ($Frame\ Memory$ FM) selon des règles déterministes.

Tab. 4: *Exemples de règles d'agrégation de frames dans la mémoire de frames FM*

Aggrégation d'une nouvelle frame X	
Si $\exists f \in$ FM tq $nom(f) = nom(X)$	fusionner chaque sous-élément de X dans f
Sinon	insérer(X) dans FM
Aggrégation d'un nouveau élément x dans une frame f de FM	
Si $\exists e \in$ f tq $nom(e) = nom(x)$	remplacer e par x
Sinon	insérer(x) dans f

Ainsi pour le gestionnaire de dialogue à base de POMDP, l'état courant du dialogue s est la FM définie à partir des frames issues des transcriptions exactes et l'observation o à partir des frames observées issues des transcriptions automatiques.

3. Modèle statistique pour la gestion du dialogue.

L'approche statistique vise à rendre le gestionnaire de dialogue plus robuste face aux erreurs de reconnaissance vocale (ASR) et de compréhension (SLU). Nous définissons dans cette partie le modèle POMDP utilisé ainsi que la simplification (POMDP résumé) permettant de traiter la complexité de la résolution ainsi que le manque de données.

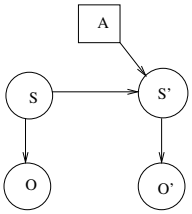
3.1. POMDP

Une introduction aux Processus de Décision de Markov Partiellement Observables (POMDPs) peut être trouvée dans [14] ou encore [16].

Définition. Un POMDP est défini par :

- une chaîne de Markov cachée (états s_t, observations o_t) conditionnée par les décisions (actions a_t) comme illustré par la figure 1,
- une fonction de récompense R pour définir les récompenses immédiates $r_t = R(s, a, s')$.

Fig. 1: Graphe d'influence d'un POMDP



Soit H l'historique observé : $H = (o_t, a_t, o_{t-1}, ...)$. Par souci de clarté, l'indice $_t$ ainsi que les paramètres $(H, o, s, a$ ou $b)$ seront parfois omis et $_{t+1}$ remplacé par un prime $'$.

Croyance. La croyance b est une distribution sur l'espace des états qui représente la probabilité d'être effectivement dans un état donné sachant l'historique observé du dialogue H. Elle est définie par l'équation :

$$\text{Pour tout état } \sigma, \qquad b(\sigma, H) = P(s = \sigma | H) \quad (1)$$

ou par la relation de récurrence :

$$b'(s') \propto P(o'|s', a) \sum_s P(s'|s, a) b(s) \quad (2)$$

b est une statistique de H suffisante pour l'état courant de dialogue s.

Stratégie. Une stratégie π donne une action a en fonction de l'historique observé : $a = \pi(H)$. Mais on se restreint généralement aux stratégies sur b : $a = \pi(b(H))$.

Optimalité. La fonction valeur $V^\pi(b)$ d'une stratégie π est définie en fonction de R par

$$V^\pi(.) = E(r_t + \gamma * V^\pi(b') \mid b = . \ , \ a = \pi(b)) \quad (3)$$

où γ est un facteur d'escompte proche de 1 et "." représente la variable de la fonction V^π.

Une stratégie π est dite optimale lorsque sa fonction valeur V^π est maximale.

Résolution. Des algorithmes de résolution exacte ou approchée sont disponibles, ainsi que différentes méthodes pour en réduire la complexité : POMDP hiérarchiques [2], POMDP factorisés [12] ou comme ci-dessous POMDP résumés [9]. D'autres représentations (*Predictive State Representations* PSR, *observable operator models* OOM) ont été également proposées ([13, 5]).

3.2. POMDP résumé

L'utilisation d'un modèle résumé permet d'une part de limiter la complexité de l'algorithme de recherche de stratégie optimale, d'autre part de palier la faible quantité de données disponibles pour apprendre les modèles.

À la date t, un état de dialogue s_t est une FM définie à partir des frames exactes (i.e. d'après des concepts annotés manuellement) tandis que l'observation o_t est une FM définie à partir des frames observées (i.e. d'après des concepts annotés automatiquement).

Ces FMs sont projetées par une fonction déterministe M depuis les espaces d'état/observation principaux virtuellement infinis dans les espaces résumés de taille réduite :

$$\tilde{s} = M(s) \qquad \text{et} \qquad \tilde{o} = M(o)$$

L'application de la fonction de résumé M implique une très forte réduction de la quantité d'information, visant à sélectionner uniquement l'information utile pour la sélection de l'action. L'espace d'états résumés contient six éléments. Dans cet article, M est définie empiriquement en s'appuyant sur des outils simples d'analyse de données (histogrammes, fréquences des motifs de frames, etc.), on appliquera ensuite des méthodes automatiques de regroupement en *clusters* comme dans [7].

4. Expériences

L'estimation complète des paramètres $P(\tilde{s}'|\tilde{s}, a)$ et $P(\tilde{o}|\tilde{s}, a)$ de transition et d'observation du modèle POMDP résumé est effectuée à partir du corpus résumé. Comme ce corpus n'est pas exhaustif : il ne contient pas tous les triplets $(\tilde{s}, a, \tilde{s}')$ et $(\tilde{s}, a, \tilde{o})$, un modèle de langage factorisé permet d'appliquer une méthode de repli généralisé avec l'outil SRILM [3].

La résolution du POMDP est effectuée par optimisation approchée (algorithme Perseus [15]).

La performance de la stratégie π obtenue est évaluée en simulant 10000 dialogues selon le modèle résumé : a généré par une machine suivant la stratégie π et $(\tilde{s}, \tilde{o})$ généré par un utilisateur simulé suivant la dynamique du modèle de POMDP résumé.

La fonction récompense R pénalise chaque tour de dialogue (-1) (afin de raccourcir la longueur du dialogue) et détermine si le dialogue est un succès ou un échec (+/-10) :

- $R(a = \texttt{Closing}, \tilde{s} = S_{Closing}) = +10$
- $R(a = \texttt{Closing}\ , \tilde{s} \neq S_{Closing}) = -10$

Pour évaluer la robustesse du système face aux erreurs de reconnaissance (ASR) et de compréhension (SLU), deux systèmes sont comparés : ASR+SLU et SLU. Les deux systèmes sont entrainés sur les mêmes données concernant l'état s et l'action a. En revanche, pour les observations o, le premier (ASR+SLU) utilise des données des transcriptions automatiques annotées automatiquement (fortement bruitées par les erreurs d'ASR et de SLU), tandis que le deuxième utilise des transcriptions manuelles annotées automatiquement (bruitées seulement par le SLU).

La figure 2 montre que le système avec ASR obtient des performances comparables à celui avec des trans-

Fig. 2: Récompenses moyennes des systèmes (ASR+SLU) et SLU selon le temps d'apprentissage (les deux courbes sont quasiment confondues).

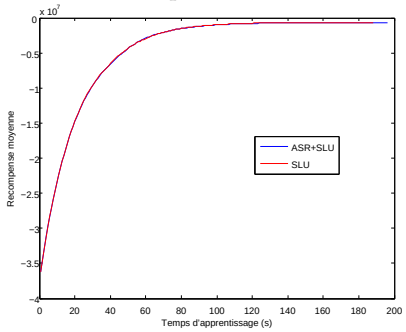
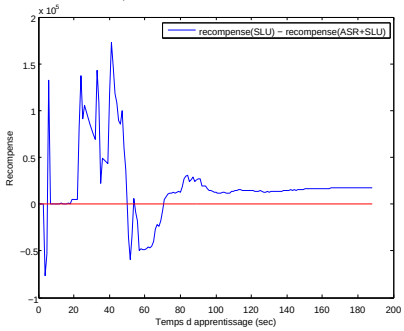


Fig. 3: Différence des récompenses moyennes : SLU − (ASR+SLU) selon le temps d'apprentissage.



criptions exactes. Cette robustesse apparente s'explique en fait principalement par le recours aux espaces résumés qui réduisent l'influence des erreurs de reconnaissance.

La figure 3 montre la différence de récompenses moyenne SLU - ASR+SLU. On constate que le système SLU moins bruité obtient une meilleure performance que le système ASR+SLU lorsque le temps d'apprentissage est suffisament long.

5. Conclusion

On a présenté dans cet article un gestionnaire de dialogue qui utilise des modèles statistiques pour mieux prendre en compte les erreurs de reconnaissance vocale et de compréhension. Par rapport à d'autres approches, la représentation choisie en graphes de frames sémantiques permet une modélisation de l'état de dialogue plus souple qu'un simple modèle de formulaire. L'utilisation de modèles résumés permet de projeter l'espace d'états initial, potentiellement infini, dans un espace d'états résumé restreint, et de réduire ainsi la complexité algorithmique de la résolution.

Une première évaluation des stratégies optimisées peut être alors réalisée en simulant des dialogues selon les modèles statistiques résumés. On observe que les erreurs de compréhension ont un impact limité sur les performances des stratégies en terme de récompense moyenne. Un prototype de système complet est en cours de construction, des évaluations plus complètes seront alors possibles dans l'espace d'états non résumé avec de véritables utilisateurs.

Références

[1] C.F. Baker, C.J. Fillmore, and J.B. Lowe. The Berkeley FrameNet project. *COLING-ACL*, 1998.

[2] A.G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13 :41–77, 2003.

[3] J.A. Bilmes and K. Kirchhoff. Factored language models and generalized parallel backoff. *HLT/NACCL*, 2003.

[4] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa. Semantic Annotation of the French Media Dialog Corpus. *Ninth European Conference on Speech Communication and Technology*, 2005.

[5] H. Jaeger. Discrete-time, discrete-valued observable operator models : A tutorial. Technical report, 1998.

[6] F. Lefèvre. Dynamic Bayesian Networks and Discriminative Classifiers for Multi-Stage Semantic Interpretation. *IEEE ICASSP*, 2007.

[7] F. Lefèvre and R. de Mori. Unsupervised state clustering for stochastic dialog management. *ASRU*, 2007.

[8] M.J. Meurs, F. Lefèvre, and R. De Mori. Spoken language interpretation : On the use of dynamic bayesian networks for semantic composition. In *ICASSP*, 2009.

[9] F. Pinault, F. Lefèvre, and R. De Mori. Feature-based summary spaces for stochastic dialogue modeling with hierarchical semantic frames. *INTERSPEECH*, 2009.

[10] J. Pineau, G.J. Gordon, and S. Thrun. Anytime point-based approximations for large pomdps. *JAIR*, 2006.

[11] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young. Agenda-based user simulation for bootstrapping a POMDP dialogue system. *ACL*, 2007.

[12] H.S. Sim, K.E. Kim, J.H. Kim, D.S. Chang, and M.W. Koo. Symbolic heuristic search value iteration for factored POMDPs. 2008.

[13] Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations : a new theory for modeling dynamical systems. *AUAI*, 2004.

[14] EJ Sondik. The optimal control of partially observable decision processes. *Ph. D. thesis, Stanford University, Stanford, California, USA*, 1971.

[15] M.T.J. Spaan and N. Vlassis. Perseus : Randomized point-based value iteration for POMDPs. *JAIR*, 2005.

[16] J. Williams, S. Young, J.D. Williams, and S. Young. Scaling pomdps for spoken dialog management. *IEEE Transactions on Audio Speech and Language Processing*, 15(7) :2116, 2007.

[17] S. Young, J. Schatzmann, K. Weilhammer, and H. Ye. The Hidden Information State Approach to Dialog Management. *ICASSP*, 2007.

Indices utiles à la cohésion lexicale pour la segmentation thématique de documents oraux*

Camille Guinaudeau - Guillaume Gravier - Pascale Sébillot

IRISA UMR 6074 & INRIA Rennes - Bretagne Atlantique
Campus de Beaulieu, F - 35042 Rennes Cedex, France
camille.guinaudeau@irisa.fr – guillaume.gravier@irisa.fr – pascale.sebillot@irisa.fr

ABSTRACT

The increasing quantity of TV material requires methods to help users navigate such data streams. Topic segmentation of TV broadcast is a first stage to structuring tasks. The goal of this article is to determine to what extent confidence measures and semantics can compensate errors in automatic transcripts for topic segmentation. To this end, we introduce confidence measure and semantic relations in a topic segmentation method. We show that our F1-measure is improved by +1.5 and +1.9 when integrating confidence measure and semantic relations respectively. Such improvement demonstrates that simple clues can conteract errors in automatic transcripts for topic segmentation.
Keywords: confidence measure, semantic relations, topic segmentation, TV broadcast

1. Introduction

L'augmentation du nombre de documents télévisuels disponibles rend indispensable la mise en place de méthodes de structuration de ces flux, structuration nécessitant une phase préalable de segmentation : du flux en émissions successives d'une part et des émissions en segments thématiques d'autre part. La segmentation thématique de documents oraux peut désormais s'effectuer par le biais des transcriptions automatiques de la bande sonore contenue dans les documents, les performances des systèmes de reconnaissance automatique de la parole (RAP) s'étant considérablement améliorées ces dernières années [5]. La plupart des travaux développés en ce sens appliquent généralement sur ces transcriptions des méthodes issues de la segmentation de documents textuels, très fréquemment fondées sur le critère de cohésion lexicale. Ainsi [4] et [7] proposent respectivement une méthode modélisant le problème de la segmentation par un modèle de Markov caché et une méthode consistant à rechercher la meilleure segmentation parmi toutes les segmentations possibles. Des marqueurs discursifs, obtenus lors d'une phase préalable d'apprentissage peuvent aussi servir à repérer des frontières thématiques [2, 3]. Christensen et *al.* [3] ont d'ailleurs établi que les erreurs de transcriptions n'avaient que peu d'effets sur les performances d'un algorithme de segmentation supervisée utilisant de tel marqueurs. Cependant, nous avons obtenu, lors de précédents travaux sur la segmentation d'émissions radiophoniques par une approche non supervisée se basant sur la cohésion lexicale, un gros écart de performances entre transcriptions manuelles et automatiques. Ceci laisse penser qu'il convient, pour pallier les erreurs de transcription, d'intégrer à la cohésion lexicale des indices propres aux documents oraux. Par exemple, [1] exploite la détection de locuteur afin de repérer le présentateur du journal télévisé, celui-ci introduisant de nouveaux reportages et donc les changements thématiques. Conjointement à la transcription, les auteurs de [6] utilisent quant à eux la prosodie. Cependant de tel indices sont globalement peu employés car leur extraction automatique est difficile. L'objectif de cet article est donc d'étudier si des indices plus facilement accessibles peuvent aider notre algorithme de segmentation à être plus robuste aux erreurs de transcription. Pour cela, nous testons si l'intégration des mesures de confiance, associées à chacun des mots transcrits par le système de RAP, et de relations sémantiques permet d'améliorer les performances de notre méthode de segmentation.

Dans cet article, nous présentons tout d'abord le critère de cohésion lexicale tel qu'il est utilisé pour la segmentation thématique de documents textuels, avant de décrire, en section 3, les modifications apportées à ce critère pour la segmentation de documents oraux. Les méthodes de segmentation employées ainsi que leurs résultats sont exposés dans les sections 4 et 5, avant la présentation de quelques perspectives.

2. Cohésion lexicale pour l'écrit

La notion de cohésion lexicale fait référence aux relations lexicales qui existent au sein d'un texte et lui donnent une certaine unité. Les méthodes de segmentation thématique utilisant cette cohésion se basent sur l'analyse de la distribution des mots au sein du texte : une rupture thématique est détectée lorsque le vocabulaire utilisé change de façon significative.

La valeur de la cohésion lexicale d'un segment S_i peut être vue comme la mesure de la capacité d'un modèle de langue Δ_i appris sur le segment S_i à prédire les mots contenus dans le segment. Deux étapes importantes sont nécessaires dans le calcul de la valeur de la cohésion lexicale d'un segment :
- le calcul du modèle de langue Δ_i du segment S_i,
- le calcul de la probabilité traduisant la capacité du modèle de langue Δ_i à prédire les mots de S_i.

Modèle de langue. Le modèle de langue Δ_i appris sur un segment S_i est un modèle de langue, sur l'en-

*Travaux partiellement financés par le projet Quaero.

semble des mots du vocabulaire du texte, pour ce segment. Le calcul du modèle de langue du segment S_i se formalise, pour un lissage donné, par

$$\Delta_i = \{P_i(u) = \frac{C_i(u)+1}{z_i},\ \forall\ u \in V_K\}\ , \qquad (1)$$

avec V_K le vocabulaire du texte de taille K et $C_i(u)$ le compte du mot u qui correspond à son nombre d'occurrences dans le segment S_i. La distribution de probabilité est lissée en incrémentant le compte de chacun des mots de 1. L'objectif de ce lissage est d'empêcher la concentration de la masse de probabilité sur les mots observés dans le segment – le nombre de mots observés dans le segment étant relativement petit au regard du nombre de mots dans le texte. On a donc $z_i = K + \sum_{j=1}^{n_i} Ci(w_j^i)$, avec n_i le nombre de mots du segment S_i et w_j^i le j^e mot de S_i.

Vraisemblance. La seconde étape du calcul de la valeur de la cohésion lexicale d'un segment consiste à calculer une probabilité traduisant à quel point le modèle de langage Δ_i permet de prédire les mots contenus dans le segment S_i, soit

$$\ln(P(S_i|\Delta_i)) = \sum_{j=1}^{n_i} \ln(\frac{C_i(w_j^i)+1}{z_i})\ . \qquad (2)$$

Intuitivement cette probabilité favorise les segments les plus cohérents lexicalement puisque sa valeur est plus importante lorsque les mots apparaissent plusieurs fois au sein du segment et qu'elle diminue si beaucoup de mots sont différents.

Le calcul de la cohésion lexicale tel que nous venons de le présenter accorde autant d'importance à chacun des mots, qu'ils soient ou non correctement transcrits. Or une erreur de transcription peut avoir un impact important sur la valeur de la cohésion lexicale.

3. Cohésion lexicale pour l'oral

Afin d'adapter le calcul de la cohésion lexicale aux documents oraux, en étant robuste aux erreurs de transcription, nous souhaitons y intégrer deux indices : des mesures de confiance et des relations sémantiques.

3.1. Utilisation des mesures de confiance

Les mesures de confiance, qui correspondent à la probabilité pour un mot d'être correctement transcrit, peuvent être prises en compte à deux niveaux lors du calcul de la cohésion lexicale : au moment du calcul du modèle de langue Δ_i ou lors du calcul de la probabilité des mots du segment S_i pour ce modèle de langue.

Dans le premier cas, on se propose de remplacer le compte $C_i(u)$ par la somme des confiances associées à chacune de ses occurrences, soit

$$C'_i(u) = \sum_{w_j^i = u} c(w_j^i)^{\lambda_2} \qquad (3)$$

où $c(w_j^i)$ correspond à la valeur de l'indice de confiance du j^e mot dans le segment S_i et où λ_2 est un paramètre permettant de faire diminuer le poids des mots dont la valeur de la mesure de confiance est faible.

Dans le second cas, la probabilité d'apparition de l'occurrence d'un mot dans un segment est multipliée par la mesure de confiance de l'occurrence de ce mot, ceci dans le but de diminuer l'importance du mot dans le calcul de la cohésion lexicale si sa mesure de confiance est faible

$$\ln(P(S_i|\Delta_i)) = \sum_{j=1}^{n_i} (c(w_j^i))^{\lambda_1} \ln(\frac{C_i(w_j^i)+1}{z_i})\ , \qquad (4)$$

avec λ_1 équivalent à λ_2.

Il est également possible de combiner les deux techniques d'intégration en remplaçant dans (4) C_i par C'_i.

3.2. Utilisation de la sémantique

Contrairement à un mot correctement transcrit, un mot mal transcrit a peu de chance d'être relié sémantiquement aux autres mots du segment. À partir de cette hypothèse, des relations sémantiques s'apparentant à des synonymes – nous considérons que deux mots sont sémantiquement liés s'ils apparaissent dans des contextes similaires – sont intégrées dans le calcul de la cohésion lexicale afin de diminuer le poids des mots mal transcrits dans ce calcul. Le compte C_i dans (1) est modifié de la manière suivante

$$C''_i(u) = C_i(u) + \beta \sum_{j=1, w_j^i \neq u}^{n_i} r(w_j^i, u)\ , \qquad (5)$$

avec $r(w_j^i, u)$ la proximité sémantique des mots w_j^i et u. Cette proximité correspond à la valeur de similarité des contextes de leurs occurrences.

Le paramètre β est utilisé pour pondérer l'importance des relations sémantiques lors du calcul du modèle de langue. Cette pondération est rendue nécessaire par le fait que les valeurs des relations sémantiques intégrées sont trop faibles par rapport au lissage et aux nombres d'occurrences.

4. Segmentation thématique

Nous présentons maintenant la méthode de segmentation thématique, basée sur le calcul de la cohésion lexicale, que nous utilisons, et la façon dont nous exploitons l'intégration des indices cités pour la segmentation de documents oraux.

4.1. Approche générale

Notre méthode de segmentation se base sur l'une des meilleures techniques de segmentation thématique existante, développée par Utiyama et Isahara [7], qui consiste à rechercher la segmentation qui produit les segments les plus cohérents d'un point de vue lexical, tout en respectant une distribution *a priori* de la longueur des segments. Le principe de cette technique est de trouver la segmentation la plus probable d'une séquence de l unités élémentaires (mots, phrases, *etc.*) $W = W_1^l$ parmi toutes les segmentations possibles, soit

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P[W|S]P[S]\ . \qquad (6)$$

En supposant que $P[S_1^m] = n^{-m}$, la probabilité d'un texte W pour une segmentation $S = S_1^m$ est donnée par

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}} \sum_{i=1}^{m} (\ln(P[S_i|\Delta_i]) - \alpha \ln(n))\ , \qquad (7)$$

avec n le nombre de mots du texte. La cohésion lexicale $\ln(P[S_i|\Delta_i])$ pour le segment S_i est calculée comme décrit en section 2. Le facteur α permet de contrôler la taille moyenne des segments retournés.

4.2. Approches proposées

Intégration des mesures de confiance. Afin d'adapter la méthode de Utiyama et Isahara à des documents oraux, nous définissons les 4 méthodes d'intégration suivantes :
- V^{λ_1} : l'intégration des mesures de confiance se fait lors du calcul de la vraisemblance (Éq. 4),
- M^{λ_2} : l'intégration des mesures de confiance se fait lors du calcul du modèle de langue (Éq. 3),
- $V^{\lambda_1} + M^{\lambda_2}$: l'intégration se fait lors du calcul de la vraisemblance et lors du calcul du modèle de langue,
- *Seuil* : seuls les mots dont la mesure de confiance est supérieure à un certain seuil sont pris en compte lors du calcul du modèle de langue (Éq. 1).

Cette dernière méthode est comparable à la méthode M^{λ_2}. En effet, plus la valeur du λ_2 augmente, plus la valeur de $c_i(w_j^i)^{\lambda_2}$ devient petite si $c_i(w_j^i)$ est faible. Ainsi les seuls mots pris en compte sont ceux dont la valeur de $c_i(w_j^i)$ est supérieure à un certain seuil.

Intégration de relations sémantiques. Les relations sémantiques intégrées, en modifiant le calcul du modèle de langue comme décrit dans (5), ont été apprises sur un corpus composé de textes contenant des articles *du Monde*, *de l'Humanité*, des transcriptions manuelles des campagnes *Ester 1* et *Ester 2*. Elles sont calculées en associant à chaque mot un vecteur composé des mots qui apparaissent dans ses voisinages, pondérés par leur fréquence d'apparition. La proximité sémantique de deux mots correspond à la valeur du cosinus de leurs vecteurs de contexte.

5. Résultats

Nos méthodes de segmentation prenant en compte les mesures de confiance ou les relations sémantiques ont été testées sur un corpus composé de 57 journaux télévisés (d'environ 1/2 heure chacun) diffusés en février et mars 2007 sur la chaîne de télévision France 2. Ces émissions ont été transcrites par un système de RAP, implémenté pour la transcription de journaux radiophoniques, atteignant un taux d'erreur d'environ 20% sur les journaux français du corpus *Ester 2*. Pour chacune des transcriptions, nous avons supprimé la partie précédant le lancement du premier reportage (titres) ainsi que celle suivant la fin du dernier, ces deux parties très spécifiques perturbant l'algorithme de segmentation. Cette extraction manuelle aurait pu être effectuée en utilisant des indices vidéo ou audio tels que la détection du bandeau de titre par exemple. Une segmentation de référence a été effectuée en considérant un changement de thème à chaque changement de reportage, bien que ce ne soit pas toujours le cas : ainsi, les premiers reportages traitent généralement du principal titre du journal et abordent donc tous le même thème. Nous obtenons ainsi un total de 1 180 frontières thématiques. L'évaluation de nos méthodes de segmentation se fait en considérant comme correcte une frontière éloignée de moins de 10 secondes d'une frontière de référence. Nous utilisons les métriques précision, rappel et F1-mesure pour chiffrer les résultats de nos algorithmes. Afin de confronter nos différentes méthodes, nous comparons leurs résultats pour une valeur de α optimale, c'est-à-dire conduisant à une segmentation dont la longueur moyenne des segments est la plus proche de celle de la segmentation de référence (96.1 secondes).

Tab. 1: F1-mesure pour la méthode *Seuil*

seuil	0	0.1	0.3	0.5	0.7	0.9
F1-mesure	58.9	58.9	58.7	59.1	57.9	55.0

5.1. Intégration de mesures de confiance

Nous présentons tout d'abord la méthode *Seuil*, la plus facile et la plus immédiate à mettre en œuvre, avant de décrire les résultats des méthodes nécessitant une modification de la technique de calcul du critère de cohésion lexicale.

Dans le tableau 1, nous pouvons constater que la méthode *Seuil* ne permet pas d'améliorer significativement les performances de l'algorithme de segmentation. En effet, un seuil égal à 0.5 conduit à une très faible amélioration de la F1-mesure (+0.2) par rapport à une segmentation sans prise en compte de la mesure de confiance (seuil = 0). De plus, nous remarquons que pour les seuils dont la valeur est supérieure à 0.7, la valeur de la F1-mesure diminue fortement car le nombre de mots pris en compte lors du calcul du modèle de langue est trop faible.

Dans le tableau 2 sont résumés les résultats obtenus en intégrant les mesures de confiance grâce aux méthodes modifiant la technique de calcul de la cohésion lexicale ; la première ligne correspond à une segmentation sans prise en compte des mesures de confiance. Nous pouvons constater que ces trois méthodes ont un comportement similaire. En effet, pour toutes ces techniques, l'intégration des mesures de confiance permet d'améliorer de façon statistiquement significative (t-test) la valeur de la F1-mesure lorsque le paramètre λ_k[1] est égal à 1 ou 2, mais dégrade la qualité de la segmentation lorsque que la valeur du λ_k est trop importante. Nous remarquons également que l'amélioration est plus importante lorsque les deux méthodes sont combinées, +1.5 contre +1.2 et +1 pour les méthodes V_{λ_1} et M_{λ_2} respectivement. Finalement, nous observons que la méthode V_{λ_1} entraîne une dégradation plus importante que M_{λ_2} lorsque le paramètre λ_k augmente – constatation valable également lorsque les deux techniques sont combinées. Ceci peut s'expliquer par le fait que la probabilité d'apparition d'un mot u dans le segment S_i correspond à une probabilité lissée, ce qui implique que, même si la valeur de $c_i(w_j^i)^{\lambda_2}$ est très petite, la probabilité d'apparition du mot dans M_{λ_2} n'est pas autant diminuée que dans V_{λ_1}.

En étudiant les courbes rappel/précision pour α variant de 0 à 1, pour chacune des 4 méthodes – courbes calculées pour une valeur de λ_k optimale – nous avons pu constater que, pour un rappel supérieur à 55%, l'intégration de mesures de confiance permettait d'améliorer à la fois le rappel et la précision pour toutes les mé-

[1] avec $k = 1$ ou $k = 2$.

Tab. 2: F1-mesure pour V^{λ_1}, M^{λ_2} et $V^{\lambda_1} + M^{\lambda_2}$.

λ_k	V^{λ_1}	M^{λ_2}	$V^{\lambda_1} + M^{\lambda_2}$ avec $\lambda_1 = 1$	$V^{\lambda_1} + M^{\lambda_2}$ avec $\lambda_1 = \lambda_2$
0	59.8	59.8	59.8	59.8
1	**61.0**	60.1	**61.3**	**61.3**
2	60.1	**60.8**	60.3	60.0
3	58.3	58.5	60.2	59.2
4	57.7	59.4	60.1	58.3
5	57.1	59.7	59.8	57.7
6	56.0	59.4	59.1	56.2
7	38.7	59.1	58.6	55.5
8	38.2	59.0	58.5	54.8
9	37.9	58.3	58.6	54.1
10	37.7	57.9	58.1	53.1

thodes. De plus, nous avons remarqué que l'intégration de mesures de confiance conduit d'une part à l'augmentation du nombre de frontières correctes détectées par notre algorithme mais également au déplacement de frontières préalablement reconnues, les rapprochant ainsi de frontières de référence.

L'intégration des mesures de confiance pourrait également être effectuée en combinant la méthode *Seuil* et celles modifiant la technique de calcul de la cohésion lexicale. Cependant, un premier test sur l'association de *Seuil* et M_{λ_2} n'a pas fourni de meilleurs résultats que M_{λ_2} seule (+0.2 contre +1). Nous avons donc choisi de ne pas tester plus avant les combinaisons.

De toutes ces constations, nous pouvons conclure que les méthodes modifiant le calcul de la cohésion lexicale pour intégrer les mesures de confiance offrent une plus forte amélioration de la qualité de la segmentation thématique que la méthode *Seuil*. De plus, cette amélioration est plus importante lorsque l'intégration des mesures de confiance est effectuée lors des deux étapes nécessaires au calcul de la cohésion lexicale.

5.2. Intégration de relations sémantiques

Lors de l'intégration de relations sémantiques dans le calcul de la cohésion lexicale, nous avons tout d'abord effectué différents tests en faisant varier le paramètre β afin de donner plus ou moins de poids aux relations sémantiques. Nous avons également modifié le nombre de relations sémantiques utilisées lors du calcul de la cohésion lexicale. En effet, nous avons pu constater qu'un nombre trop important bruitait énormément ce calcul. Nous avons donc choisi de limiter le nombre de relations associées à chaque mot. Avec un paramètre β optimal, la valeur de la F1-mesure est augmentée de +1.9 lorsque le nombre de relations associées à chaque mot est égal à 2, et de +0.7 lorsqu'il est égal à 3.

De plus, en observant en détails les résultats, nous remarquons que l'intégration de relations sémantiques permet à la fois de supprimer des frontières incorrectes et, comme pour les mesures de confiance, de rapprocher certaines frontières de frontières de référence.

L'intégration de relations sémantiques permet donc, elle aussi, d'améliorer les performances de l'algorithme de segmentation. Cependant, il est nécessaire de contraindre le nombre de relations intégrées et le poids qui leur est associé, un poids trop important ou un trop grand nombre de relations faisant diminuer considérablement la valeur du rappel.

6. Conclusion et perspectives

L'intégration d'indices supplémentaires que sont les mesures de confiance et les relations sémantiques semble rendre notre algorithme de segmentation plus robuste aux erreurs de transcription et améliore ses performances. En effet, la valeur de la F1-mesure est augmentée de +1.5 dans le cas de l'intégration des mesures de confiance et de +1.9 lors de l'utilisation de relations sémantiques lorsque nous utilisons la méthode de segmentation proposée dans [7]. De plus, des résultats obtenus sur une méthode basée sur le principe de fenêtres glissantes fournit des résultats similaires, ce qui semble indiquer que l'intégration des deux indices permet d'améliorer la qualité de la segmentation quelle que soit la méthode utilisée.

Afin de consolider ces résultats, nous souhaitons tester l'intégration des deux indices sur un autre corpus composé d'émissions télévisées différentes. En effet, même si un premier test sur 4 émissions de reportages donne des résultats encourageants – la F1-mesure est augmentée de +12 lorsque l'on utilise la méthode $V^{\lambda_1} + M^{\lambda_2}$ par rapport à une segmentation sans prise en compte des mesures de confiance – le corpus n'est pas suffisamment conséquent pour nous permettre actuellement d'en tirer des conclusions. De plus, dans le but d'améliorer les performances de notre algorithme de segmentation, nous souhaitons également combiner l'intégration des mesures de confiance et des relations sémantiques lors du calcul de la cohésion lexicale. Enfin, une perspective à plus long terme consiste à appliquer notre méthode de segmentation non pas sur la transcription finale mais sur les graphes de mots.

Références

[1] R. Amaral and I. Trancoso. Topic indexing of TV broadcast news programs. In *6th International Workshop on Computational Processing of the Portuguese Language*, 2003.

[2] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. In *Machine Learning*, 1999.

[3] H. Christensen, B. Kolluru, and Y. Gotoh et *al.* Maximum entropy segmentation of broadcast news. In *30th IEEE ICASSP*, 2005.

[4] P. Van Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron. Segmentation of automatically transcribed broadcast news text. In *DARPA Broadcast News Workshop*, 1999.

[5] M. Ostendorf, B. Favre, and R. Grishman et *al.* Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 2008.

[6] A. Stolcke, E. Shriberg, and D. Hakkani-Tür et *al.* Combining words and speech prosody for automatic topic segmentation. In *DARPA Broadcast News Workshop*, 1999.

[7] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *9th ACL*, 2001.

Identification du genre vidéo à la volée par combinaison de paramètres acoustiques

Mickael Rouvier, Georges Linarès et Driss Matrouf

LIA-CERI - Université d'Avignon

ABSTRACT

Video genre identification methods are frequently based on image or motion analysis, which are relatively time-consuming processes. Since such approaches are tractable by batch processing, as-soon-as-possible identification requires faster methods. We investigate the use of audio-only methods for on-the-fly video classification. We propose to use several acoustic feature streams and we evaluate various combination schemes at the frame or at the score level. Results are compared to those obtained by humans, according to the listening duration. Although the system based on model combination slightly outperforms the humans on very soon detection. The latter remain significantly more accurate on long sessions.
Keywords: Video genre identification, video classification, audio processing, speech processing

1. Introduction

Ces dernières années, la quantité de données multimédia accessibles sur Internet a considérablement augmenté, rendant le parcours et la recherche de données multimédia impossibles sans une indexation préalable. Dans la plupart des cas, les méta-données associées aux vidéos sont inexistantes, empêchant l'indexation sémantique des vidéos. Il existe plusieurs types de descripteurs, de plus ou moins haut niveau, qui peuvent être utilisés pour structurer des bases de données vidéos. Ils peuvent être liés aux contenus lignuistiques du document, aux types de contenu (musique, parole,...), aux sujets, etc... Nous utiliserons ici comme descripteur le genre, qui se réfère aux styles éditoriaux des vidéos. Une complète taxonomie du genre peut être trouvée dans [12], mais la plupart des systèmes d'identification essaient d'identifier les genres les plus couramment utilisés tel que la publicité, film, cartoon, actualité, musique.

Dans cet article, nous nous focaliserons sur un usage particulier pour lequel l'utilisateur a besoin d'identifier un flux vidéo à la volée, par exemple afin de supprimer les publicités ou de sélectionner un genre particulier. Deux principales contraintes doivent être satisfaites pour le scénario d'identification à la volée. Premièrement, les descripteurs globaux estimés sur la totalité d'une vidéo ne peuvent être utilisés, car l'identification doit être synchrone. Deuxièmement, l'identification à la volée doit se faire le plus rapidement possible ce qui disqualifie les méthodes issues du traitement de l'image.

Dans la littérature, de nombreuses personnes ont abordé la question de la classification vidéo. Les systèmes basés sur l'analyse de l'image obtiennent de bonnes performances mais sont liés à des paramètres coûteux à extraire. Les approches basées sur le texte utilisent des classifieurs sur les sous-titres ou flux teletext, mais ces applications sont limitées au cas où ces données sont disponibles.

L'identification basée sur l'audio a été, ces dernières années, développée dans deux directions. Les approches dites de haut niveau, qui consistent à détecter des événements ou à analyser la sortie d'un système de reconnaissance de la parole. Ces méthodes sont liées sur une connaissance *a priori* du contenu audio (par exemple les jingles) et la reconnaissance de la parole qui est un processus gourmand en terme de CPU et de mémoire. De plus, dans un contexte télévisé ou de données Web, le taux d'erreur mot est généralement élevé, ce qui complique l'analyse linguistique. Les approches de bas niveau utilisent des classifieurs statistiques (GMM, SVM,...) sur des paramètres cepstraux [9, 4, 14] ou dans le domaine temporel [7]. Différents types de paramètres et stratégies de classification ont été étudiés ces dernières années. La section suivante présente un aperçu de ces méthodes et discute de leur application pour la tâche d'identification de genre vidéo à la volée.

La suite de cet article est organisée comme suit : la section suivante présente les paramètres acoustiques utilisés par le système d'identification de genre vidéo. Dans le paragraphe 3, nous détaillerons la tâche d'identification et le corpus utilisé dans nos expériences. Le paragraphe 4 décrira notre combinaison au niveau des trames et au niveau des scores. Finalement, nous présenterons une expérience qui a pour but d'estimer les performances humaines sur la tâche de l'identification du genre vidéo à la volée. La dernière section conclut l'article et propose quelques futurs développements et expériences.

2. Paramètres audio pour l'identification de genres vidéo

Les paramètres dans les domaines cepstraux et temporels sont les plus couramment utilisés comme descripteurs acoustiques bas niveau pour la classification vidéo. De plus, les paramètres cepstraux ont été largement utilisés, non seulement dans le domaine de la parole, mais aussi dans des tâches de traite-

ment audio. Dans [9], les auteurs évaluent une paramétrisation MFCC (Mel-Frequency Cepstral Coefficient) et un Réseau de neurones pour l'identification de genre vidéo (IGV). Ce système permet d'obtenir un taux de classification correcte de 51% pour une détection de 5 genres (actualité, publicité, sport, film, cartoon). Dans [7], la pertinence du domaine temporel est étudié. Ces paramètres permettent de représenter la structure d'un document sur un axe temporel, typiquement en utilisant le taux de passage par zéro et la variance de l'énergie. Dans un article précédent [10], nous combinons ces paramètres de bas niveau à d'autres paramètres de plus haut niveau, tels que l'interactivité du locuteur et la qualité de la parole. Ces deux derniers types de paramètres reposent sur une analyse globale qui ne peut pas être utilisés pour de l'IGV à la volée car pour extraire ces paramètres la vidéo doit être traiter en entier.

Par conséquent, nous nous focaliserons uniquement sur l'analyse cepstrale à court terme du flux vidéo. Plusieurs systèmes de transcription de parole utilisent plusieurs paramètres acoustiques pour augmenter la précision des systèmes [15, 1]. Nous analyserons l'intégration de plusieurs paramètres acoustiques est leur complémentarité pour l'IGV à la volée. Dans le paragraphe 4, différents systèmes sont étudiés pour tirer pleinement parti de la complémentarité entre les 3 paramètres acoustiques PLP (Perceptual Linear Prediction), Rasta-PLP et MFCC.

3. Tâche et corpus

La tâche d'IGV consiste à identifier 7 genres vidéos (actualité, sport, cartoon, musique, documentaire, film et publicité). Le corpus est composé de 1680 vidéos indexées manuellement, avec une durée comprise entre 2 et 5 minutes. 1400 de ces vidéos sont utilisées pour le corpus d'entraînement de notre système, 280 composant le corpus de test. Les 7 genres sont représentés de manière uniforme (environ 200 vidéos par genre pour le corpus de train, 40 par genre pour le corpus de test). La langue présente sur les vidéos est systématiquement du français, bien que la catégorie musique contiennent quelques chansons en anglais.

4. Système proposé

4.1. Aperçu du système

Le système proposé est une architecture à 2-niveaux, où le premier niveau extrait les paramètres cepstraux et le second niveau combine les paramètres et les classifie en genres. Les trames acoustiques sont calculées toutes les 10ms dans une fenêtre de Hamming de 20ms de large. Les PLP et Rasta-PLP sont composés de vecteur de 12 coefficients plus l'énergie auxquels sont ajoutés les dérivées premières et secondes. Les MFCC sont composées de vecteurs de 14 coefficients plus l'énergie auxquels sont ajoutés les dérivés premières et secondes [3].

Dans le deuxième niveau, le classifieur est un modèle GMM-UBM (Gaussian Mixture Model - Universal Background Model) dépendant du genre avec une

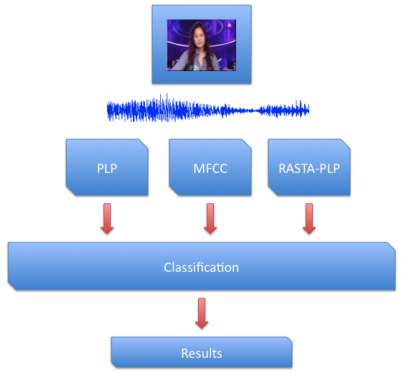


Figure 1: *Architecture du système d'identification de genre vidéo à la volée.*

réduction de la variabilité par une Analyse Factorielle (AF) [11]. De précédentes recherches dans le domaine de la reconnaissance de la parole ont montré l'intérêt de combiner des paramètres acoustiques complémentaires. Nous évaluerons et comparerons 2 méthodes de combinaisons qui opèrent respectivement à la trame et au score. Ces points sont développés dans le paragraphe 4.3.

4.2. Analyse factorielle pour la classification vidéo

Une approche de classification de genre audio consiste à utiliser des classifieurs tels que les GMM sur les paramètres cepstraux [8]. Une des principales difficultés de la classification de genre vidéo est due à la diversité des vidéos d'un même genre. Ce problème a été récemment contourné en appliquant une AF afin de réduire la variabilité intra-classe, sur des tâches variables telles que l'identification du locuteur [6] ou de la langue [13]. Nous avons avec succès appliqué cette technique à l'IGV dans un article [11]. Ici, l'AF est appliqué de la même façon au modèle GMM-UBM dépendant du genre, indépendamment des paramètres utilisés.

4.3. Combinaison des paramètres cepstraux

Le but de la combinaison acoustique est d'exploiter l'information complémentaire apportée par les différents paramètres acoustiques. Dans la littérature, 2 approches sont proposées pour combiner différents paramètres acoustiques :

- Combinaison au score : cette approche consiste à combiner les score issus des modèles du GMM dépendant du genre pour toutes les paramètrisations (PLP, MFCC et Rasta-PLP). Cela est réalisé en entraînant un Multi-Layer Perceptron (MLP) sur un groupe de vecteurs qui groupe toutes les sorties des GMM.
- Combinaison à la trame : dans cette technique, tous les paramètres acoustiques sont groupés dans un super-vecteur sur lequel le classifieur opère.

Nous testons ces deux différentes approches et comparons les résultats sur notre corpus de test.

Combinaison des scores Dans cette approche, une combinaison est appliquée au niveau du score. L'idée est d'estimer une probabilité *a priori* du genre en combinant les scores fournis par les GMM dépendant des paramètres. Cette combinaison est réalisée avec un MLP. Ce choix est motivé par des résultats de manière empirique.

Les probabilités postèrieures sont estimées sur chaque flux de paramètres. Les trois estimations sont ensuite groupées dans un vecteur de scores, lequel est utilisé comme entrée du méta-classifieur MLP.

Le MLP est un réseau de neurones à 3 couches avec respectivemment 21, 11 et 7 neurones. Les 21 neurones en entrée contiennent les scores issus depuis les 3 classifieurs acoustiques. Chacune des 7 sorties du MLP correspond à un genre.

Table 1: Taux de classification selon le type de paramètrisation acoustique sur les 7 genres.

	Rasta-PLP	PLP	MFCC
Results	85.81	85.46	85.46

On peut observer que les rappels pour les 3 paramètres acoustiques sont vraiment proches. Nous obtenons un taux d'identification d'environ 85% pour l'ensemble des paramètres acoustiques. Ces différents jeux de paramètres sont combinés afin d'évaluer leur complémentarité. Dans le Tableau 2, on voit qu'on obtient des résultats comparables à ceux de l'Oracle, qui sont obtenus en choisissant la meilleure hypothèse de classification, connaissant la cible du genre vidéo.

Table 2: Les performances Oracle et MLP des combinaisons de paramètres acoustiques. Les résultats démontrent une très bonne complémentarité entre les paramètres.

	MLP-based combination	Oracle
RPLP+PLP	0.91	0.92
PLP+MFCC	0.90	0.90
RPLP+MFCC	0.90	0.91
RPLP+MFCC+PLP	0.93	0.94

En combinant deux à deux les jeux de paramètres acoustiques pour n'importe quel jeu de paramètres (Rasta-PLP, MFCC ou PLP), nous observons une réduction du taux d'erreur relative d'environ 33% (de 85% à 90% en valeur absolue). En combinant les 3 paramètres acoustiques, nous observons une autre réduction du taux d'erreur relative de 30% (de 90% à 93% en valeur absolue). Cependant, les taux d'identification pour chaque type de paramètres acoustiques sont proches. Les résultats tendent à confirmer que ces 3 paramètres acoustiques sont complémentaires.

Combinaison des trames Nous proposons de combiner directement les différents paramètres acoustiques. Le vecteur acoustique sortant sera de taille 39 + 39 + 45 = 123, où 39 est la dimension des paramètres PLP et Rasta-PLP, et 45 est celle des paramètres MFCC. Cependant, le paramètre (supervector) peut contenir des redondances significatives, et inclure toutes les dimensions n'est pas toujours nécessaire pour améliorer l'identification du genre vidéo.

Afin de réduire la dimension du vecteur de paramètres, nous utilisons l'*Heteroscedastic Linear Discriminant Analysis* (HLDA). La HLDA est une technique qui a pour but d'estimer un sous-espace de représentation où les classes sont facilement séparables.

HLDA généralise la LDA en supprimant la restriction d'une matrice de covariance commune. HLDA a été recemment utilisée avec succès dans le domaine de la RAP [2]. La théorie de HLDA est décrite plus en détail dans [5].

Dans la Figure 2, nous avons essayé d'évaluer l'impact de la réduction des dimensions du super-vecteur. Ces résultats sont comparés à une classification classique sur le meilleur paramètre acoustique Rasta-PLP en utilisant une AF.

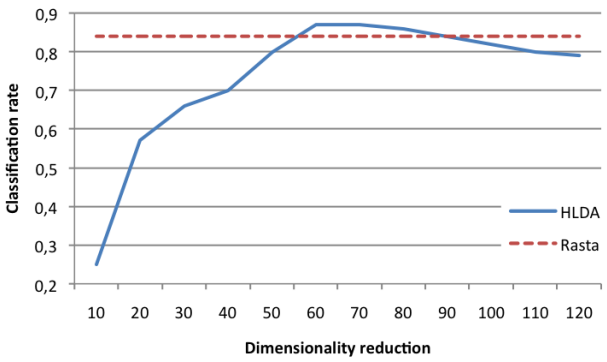


Figure 2: *Impact de la réduction de dimension sur l'HLDA.*

Finalement, le meilleur système (avec une réduction de dimension à 60 coefficients) obtient un taux d'identification de 87% contre 85% pour le Rasta-PLP. Le gain absolu est d'environ 2%, lequel est vraiment bas comparé au MLP où le gain absolu est de 8%. Les résultats montrent que le HLDA semble être moins précis que la combinaison MLP.

5. Comparaison avec les performances humaines

Dans les expériences précédentes, nous choisissons différentes façons de combiner les paramètres acoustiques sur la totalité de la vidéo. Ici, nous comparons les systèmes d'identification aux performance humaines. Les résultats sont reportés seulement pour les 2 systèmes hybrides (MLP et HLDA) et pour le meilleur système acoustique (basé sur la paramétrisation Rasta-PLP).

Les tests sont réalises sur un groupe de 14 personnes. Ce panel est constitué de 14 personnes ayant entre 14 et 53 ans. Chacune d'elles déterminent le genre des 28 vidéos choisies aléatoirement, soit un total de 392 réponses pour l'ensemble du panel. Toutes les 5 secondes, la vidéo est arrété et la personne doit essayer de donner une hypothèse d'identification du genre.

6. Résultats

Comme attendu, la combinaison des trois paramètres acoustiques surpasse la classification basée sur le meilleur paramètre acoustique. On peut observer que le comportement des systèmes de combinaisons acoustiques diffère significativement : bien que le MLP surpasse les performances des autres approches sur des longs segments (avec une durée supérieure à 15s), les performances de HLDA sont meilleures sur de courtes durées. Sur de très courtes durées d'identification, la combinaison basée sur le HLDA surpasse les humains : les premières 5 secondes, nous observons un rappel de 53%, 45%, 45% et 40% respectivement pour les méthodes HLDA, Humain, MLP et Rasta-PLP. Cependant, ce classement change rapidement : les combinaisons basées sur le MLP obtiennent de très bons résultats (79%) pour une durée de 40s.

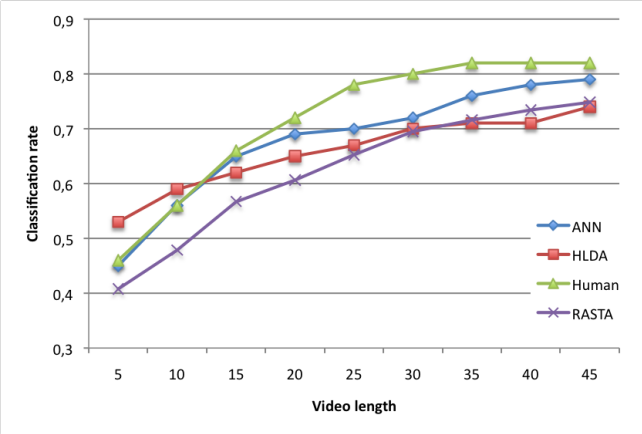


Figure 3: *Taux d'identification pour toutes les méthodes.*

7. Conclusion

Nous avons évalués les performances de méthode de combinaison des paramètres acoustiques pour l'IGV à la volée des genres vidéos. Différentes stratégies de combinaison ont été proposées, une combinaison directe (HLDA) ou une combinaison indirecte (MLP). Les résultats démontrent clairement l'intérêt de combiner les paramètres pour une classification synchrone. Les comparaisons avec les performances humaines suggèrent que les personnes ont besoin de quelques secondes pour correctement identifier le genre vidéo, probablement parce qu'ils utilisent des informations de plus haut niveau, reliées à la linguistique ou à la semantique.

Nos expériences sont conduites sur des vidéos où le genre est le même tout au long de la vidéo. Cependant, plusieurs applications opèrent sur des vidéos structurées, où non seulement la classification mais aussi la segmentation est requise. Nous envisageons de développer un système de segmentation en genre et d'intégrer des paramètres de bas niveau vidéo dans le processus d'identification.

Références

[1] Loic Barrault, Christophe Servan, Driss Matrouf, Georges Linarès, and Renato De Mori. Frame-based acoustic feature integration for speech understanding. In *ICASSP 2008*, 2008.

[2] Lukas Burget. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. In *ICSLP 2004*, 2004.

[3] H. Hermansky and N. Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4) :578–589, Oct 1994.

[4] R.S. Jasinschi and J. Louie. Automatic tv program genre classification based on audio patterns. In *Euromicro Conference, 2001*, 2001.

[5] Nagendra Kumar. *Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition.* PhD thesis, 1997. Adviser-Andreou, Andreas G.

[6] Driss Matrouf and al. A straightforward and efficient implementation of the factor analysis model for speaker veriïňĄcation. In *InterSpeech 2007*, 2007.

[7] Simon Moncriefi, Svetha Venkatesh, and Chitra Dorai. Horror film genre typing and scene labeling via audio analysis. In *Multimedia and Expo, 2003*, 2003.

[8] Matthew Roach and John Mason. Classification of video genre using audio. In *European Conference on Speech Communication and Technology*, 2001.

[9] Matthew Roach, Li-Qun Xu, and John Mason. Classification of non-edited broadcast video using holistic low-level features. In *IWDC'2002*, 2002.

[10] Mickael Rouvier, Georges Linarès, and Driss Matrouf. Robust audio-based classification of video genre. In *InterSpeech 2009*, 2009.

[11] Mickael Rouvier, Driss Matrouf, and Georges Linarès. Factor analysis for audio-based video genre classification. In *InterSpeech 2009*, 2009.

[12] Cees G. M. Snoek and Marcel Worring. Multimodal video indexing : A review of the state-of-the-art. *Multimedia Tools Appl.*, 25(1) :5–35, 2005.

[13] Florian Verdet, Driss Matrouf, Jean-FranÃğois Bonastre, and Jean Hennebert. Factor analysis and svm for language recognition. In *InterSpeech 2009*, 2009.

[14] Li-Qun Xu and Yongmin Li. Video classification using spatial-temporal features and pca. In *Multimedia and Expo, 2003. ICME '03*, 2003.

[15] András Zolnay, Daniil Kocharov, Ralf Schlüter, and Hermann Ney. Using multiple acoustic feature sets for speech recognition. *Speech Commun.*, 49(6) :514–525, 2007.

Les trajectoires formantiques respectant les lois de la physique contribuent-elles à une meilleure perception de la parole ?

Daniel Pape[1,3], Pascal Perrier[1], Susanne Fuchs[2], Sonia Kandel[4&5]

[1]ICP/GIPSA-lab, CNRS, Grenoble INP, Grenoble, France
[2]Phonetik/ZAS, Berlin, Allemagne
[3] IEETA, Université d'Aveiro, Portugal
[4]LPNC, CNRS, Université Pierre-Mendès France, Grenoble, France
[5]Institut Universitaire de France
Pascal.Perrier@gipsa-lab.grenoble-inp.fr
http://www.gipsa-lab.inpg.fr/page_pro.php?vid=168

ABSTRACT

Physical properties of speech articulators contribute to shape articulatory and formant trajectories. This study aims at evaluating the role of this shaping in speech perception. We conducted perception tests of synthetic stimuli generated with speech production models accounting for different degrees of physical complexity. Our results do not provide any support to the hypothesis that the degree of physical realism in the models influences the perception of naturalness. However for degraded speech (silent center speech), significant differences are observed.

Keywords: Perception-Action Interaction; Perception and Physics, Speech Goals

1. INTRODUCTION

La question du rôle des trajectoires formantiques dans la perception de la parole est ancienne et elle a été au cœur de nombreux débats. Deux théories s'opposent. Pour la première, ces trajectoires ne contiennent pas l'information pertinente phonétiquement. Cette théorie est en particulier défendue par ceux qui prônent l'existence d'une cible perceptive associée à chaque son élémentaire et le caractérisant ([6, 11] pour les voyelles, et [14] pour les consonnes). Dans ce cadre théorique, la trajectoire serait cependant exploitée perceptivement dans les cas où la cible perceptive n'est pas atteinte. Elle serait donc, non pas l'objet de la perception, mais le vecteur d'une information à partir de laquelle la cible perceptive pourrait être retrouvée [3, 4]. Pour la seconde théorie, les caractéristiques temporelles intrinsèques de la transition porteraient en elles-mêmes, et indépendamment de toute cible, l'information pertinente, celle du geste nécessaire à la production du son élémentaire. Ainsi, c'est la trajectoire formantique dans tous ses détails qui serait l'objet de la perception ([12, 13])

Dans ce contexte, les travaux de Cai et al [1] sont extrêmement intéressants. Ces auteurs ont étudié les mécanismes d'adaptation de locuteurs chinois (Mandarin) produisant la triphtongue /iau/ quand leur feedback auditif était perturbé en temps réel. Cette perturbation modifie la trajectoire formantique, tout en préservant les patrons formantiques atteints pour chacune des voyelles. Statistiquement, les sujets ont eu tendance à réagir à la perturbation en modifiant leurs stratégies articulatoires de façon à corriger l'effet de la perturbation sur la trajectoire formantique perçue. Ceci laisse penser que la trajectoire formantique entre les cibles pourrait constituer l'objectif majeur de la tâche motrice. Et dans une perspective d'interaction locuteur-auditeur, on peut conclure que la trajectoire dans sa totalité serait porteuse de l'information que le locuteur veut transmettre à l'auditeur. Ces travaux peuvent donc être interprétés comme un soutien aux hypothèses défendues par Strange et collègues [12, 13].

Cependant une autre explication est envisageable. Elle est suggérée par les travaux de Viviani et collègues (par exemple [18]) sur la perception visuelle et l'identification des mouvements de la main. Ces auteurs ont d'une part observé que, dans ces mouvements, la vitesse varie comme la racine cubique de l'inverse de la courbure de la trajectoire. Ils ont d'autre part constaté expérimentalement que cette loi était exploitée par le système de perception visuelle pour catégoriser les trajectoires. Ainsi, par exemple, un point lumineux parcourant sur un écran un cercle à vitesse constante (conformément à la courbure) est perçu comme un mouvement circulaire, alors que si ce point décrit ce même cercle à une vitesse variable, rapide-lente-rapide-lente, c'est un mouvement elliptique qui est perçu. Viviani et al. en ont conclu que la perception visuelle du mouvement chez l'homme est influencée par les connaissances que l'homme a des propriétés intrinsèques de ses mouvements. Il est alors possible d'interpréter les résultats de Caï et al., non pas comme la preuve du caractère central de la trajectoire dans la perception de la parole, mais comme la conséquence d'une perturbation non écologique qui ne respecterait par les règles physiques et les règles de contrôle moteur régissant les mouvements de la parole. Les stratégies de compensation ne viseraient alors pas à reproduire une trajectoire pertinente phonétiquement, mais à donner au mouvement ses propriétés naturelles.

Cet article présente la première étape d'une démarche d'évaluation du rôle potentiel, dans la perception de la parole, de l'impact des propriétés physiques des articulateurs de la parole sur les trajectoires formantiques. Nous y décrivons et analysons des tests

perceptifs de stimuli synthétiques générés avec des modèles intégrant plus ou moins la complexité physique de ces articulateurs. Après la présentation de la méthodologie, nous exposerons les résultats et nous conclurons dans le contexte du débat sur le rôle des trajectoires dans la perception de la parole.

2. METHODE

2.1. Sujets

Vingt-trois sujets (16 hommes, 7 femmes, âgé(e)s entre 25 et 50 ans) ont participé à l'expérience. Aucun n'était au courant des méthodes de synthèse utilisées. Tous sont de langue maternelle française. Aucun d'entre eux n'a fait part de problème d'élocution ou auditif.

2.2. Matériel et procédure

Génération des stimuli synthétiques

Des stimuli Voyelle1-Voyelle2-Voyelle1 ($V_1V_2V_1$) et Voyelle1-/g/-Voyelle1 (V_1CV_1) ont été synthétisés. Les voyelles ont été choisies parmi /i/, /e/, /ɛ/, /a/, ou /ɔ/ Ces stimuli ont été obtenus par synthèse articulatoire, à partir de formes sagittales du conduit vocal. Dans tous les cas, la synthèse acoustique a impliqué la génération de la fonction d'aire à partir de la forme sagittale [8], puis celle du signal acoustique par un modèle de type Kelly-Lochbaum (développé par B. Story [15,16]) excité par un modèle de la source vocale [17]. Dans tous les cas, la fréquence fondamentale a été maintenue constante et égale 110Hz. Les modèles de synthèse diffèrent par la façon dont les formes sagittales sont générées dans le temps.

Pour une première classe de stimuli, que nous appellerons ***Mod1,*** les formes sagittales ont été obtenues avec un modèle biomécanique bidimensionnel du conduit vocal [9]. Le modèle de contrôle est de type *cible* : on spécifie les commandes motrices de chaque son élémentaire et les mouvements entre ces sons sont la conséquence d'une évolution temporelle linéaire des commandes entre leurs valeurs cibles [7], selon un timing bien défini (tenue des cibles : 150ms ; transition entre cibles : 120ms). Différentes évaluations de ce modèle ont attesté de sa capacité à rendre compte de manière réaliste des caractéristiques cinématiques importantes des mouvements de la parole, amplitude du mouvement, valeur du pic de vitesse, forme du profil de vitesse [7], formes des trajectoires dans le plan sagittal [9], et relation vitesse-courbure [10].

Pour la synthèse des autres classes de stimuli, les formes de la langue effectivement atteintes pour chaque son élémentaire dans les stimuli de la classe ***Mod1*** ont été extraites, et ont servi de formes cibles. Pour chacune d'elles les instants auxquels elles sont atteintes et les durées de leurs tenues ont été mesurés. De nombreuses données expérimentales ont montré que pour /g/ la langue se déplace vers l'avant pendant la tenue consonantique tout en restant en contact avec le palais. C'est pourquoi deux formes cibles ont été extraites pour ce son, l'une au début de la phase de contact avec le palais (g_f) et l'autre à l'instant du relâchement consonantique (g_o). Ces formes cibles sont définies par la position de 17 points dans le plan sagittal. Tous les stimuli atteignent et maintiennent les formes cibles selon le timing mesuré dans les stimuli de la classe ***Mod1***, et le passage d'une forme cible à une autre se fait par le déplacement des 17 points selon des trajectoires rectilignes. Les classes de stimuli se différencient par le décours temporel des déplacements de ces points. Dans la classe ***Mod2,*** le déplacement se fait à vitesse constante. Dans la classe ***Mod3***, la vitesse est un arc de sinusoïde conforme aux caractéristiques d'un système du second ordre. Différentes données expérimentales ont en effet montré que ce type de profil de vitesse était couramment observé dans les mouvements de la parole. Enfin pour V_1CV_1 une quatrième classe de stimuli a été générée, ***Mod4,*** dans laquelle seule la forme g_o a été retenue pour /g/. Cette forme est maintenue pendant toute la durée de la tenue consonantique et les transitions entre formes-cibles ont le schéma temporel de la classe ***Mod2.*** Les stimuli ont donc été générés à partir de 3 modèles intégrant différentes complexités dans la représentation physique des articulateurs : ***Mod1*** correspond à la description la plus réaliste, suivi par ***Mod3*** puis par ***Mod2*** (et ***Mod 4*** pour V_1CV_1).

Pour ce travail nous avions deux objectifs majeurs : évaluer dans quelle mesure le degré de réalisme des modèles influence la qualité perçue de la synthèse ; étudier si dans des conditions de parole dégradée la perception est influencée par le réalisme des modèles. La parole dégradée a été générée en transformant les stimuli $V_1V_2V_1$ selon le paradigme des *centres silencieux* [13]. Le signal de parole y est dégradé du fait du remplacement de la partie stable de V_2 par du silence. Cette transformation a été réalisée manuellement à l'aide du logiciel PRAAT, en remplaçant par des 0 la portion de signal déterminée par les passages par zéro situés 10ms avant et 10 ms après la zone de stabilité du formant F2.

Tests perceptifs

Tous les tests ont été réalisés en chambre sourde au GIPSA-lab, à l'aide du logiciel gratuit Alvin [2]. Les sujets ont d'abord passé le test sur les stimuli à centre silencieux, puis les tests d'évaluation de la qualité de la synthèse, d'abord pour $V_1V_2V_1$, puis pour V_1CV_1. De courts tests d'entrainement ont été effectués en début de session.

Lors des tests sur les stimuli à centre silencieux, la tâche de l'auditeur a consisté à identifier V_2. Toutes les combinaisons $V_1V_2V_1$, y compris celles du type $V_1V_1V_1$, ont été évaluées. Ainsi 2 répétitions de 75 stimuli (3 modèles x $5V_1$ x $5V_2$) ont été présentées aux sujets. L'instruction était d'« *identifier la voyelle2 manquante aussi vite et aussi précisément que*

possible ». Les sujets répondaient en appuyant sur l'un des 5 boutons affichés à l'écran. Sur chaque bouton était représenté le symbole phonétique de la voyelle V_2, associé à un mot monosyllabique dont la prononciation contient cette voyelle. Ces boutons étaient situés sur un cercle centré sur un 6ème bouton sur lequel il fallait appuyer pour continuer l'expérience. Ainsi, à chaque nouvelle écoute, la souris était positionnée à égale distance des 5 boutons réponses. Ceci a permis une mesure fiable des temps de latence, en évitant une variabilité liée à la celle de la distance à parcourir avec la souris. La réécoute des stimuli n'était pas possible.

Pour évaluer les liens entre le réalisme du modèle et la qualité des stimuli synthétiques, des paires de stimuli générés par deux modèles ont été présentés. Les sujets devaient effectuer une tâche de discrimination et choisir « *aussi rapidement et précisément que possible lequel de ces deux stimuli est le plus naturel* ». Pour cela, ils devaient appuyer avec l'index gauche sur la touche 1 du clavier (stimulus 1) ou avec l'index droit sur la touche 2 (stimulus 2) du clavier numérique, les mains restant immobiles. Les tests ont été élaborés selon la procédure à choix forcé 2I-2AFC, où les deux stimuli sont présentés séquentiellement, séparés par une pause. Chaque stimulus dure 650ms et la pause est de 500ms. Les différentes classes de stimuli ont été combinées au sein des paires de manière aléatoire pour chaque séquence (exemple « ***Mod1***-Pause-***Mod2*** », « ***Mod2***-Pause-***Mod3*** », ou « ***Mod1***-Pause-***Mod3*** »). Pour chaque combinaison de 2 modèles, la moitié des stimuli, toutes séquences confondues, a été présentée dans un ordre, et la moitié dans l'ordre inverse. Cette procédure a été choisie pour garantir au mieux un traitement équivalent de tous les stimuli et rendre ainsi pertinente la mesure du temps de latence. Certaines paires contenaient aussi des stimuli identiques afin de tester la fiabilité des sujets (Test aveugle). L'évaluation perceptive des stimuli $V_1V_2V_1$ et V_1CV_1 a été faite en deux tests séparés. Pour $V_1V_2V_1$, 2 répétitions de 80 paires de stimuli [(3 couples de modèles + test aveugle) x 5V1 x 4V_2] ont été présentées. Pour V_1CV_1, /i/ n'a pas été prise en compte et 4 répétitions de 28 paires de stimuli [(6 couples de modèles + test aveugle) x 4V1], ont été présentées. La réécoute n'était pas possible.

2.3. Analyse statistique

L'objectif de l'analyse statistique des données est de voir si, tous sujets et tous stimuli confondus, il existe des différences entre les classes de stimuli. Nous avons utilisé pour cela le modèle linéaire généralisé avec effets mixtes disponible dans le logiciel R (2008). Nous avons choisi cette analyse plutôt qu'une ANOVA classique, car elle permet de traiter des classes qui n'ont pas le même nombre de données. Elle offre aussi la possibilité d'éliminer dans le traitement statistique la contribution des stimuli ($V_1V_2V_1$ et V_1CV_1) et des sujets dans la variance globale, en les considérant comme des facteurs aléatoires de variabilité. La classe des stimuli (***Mod1, Mod2, Mod3, Mod4)*** a été choisie comme facteur fixe.

3. RÉSULTATS

Réalisme du modèle et qualité de la synthèse

Les réponses aux paires de stimuli identiques ont été utilisées pour vérifier les performances de nos sujets. Théoriquement, pour de tels stimuli, les choix devraient se répartir équitablement entre les deux réponses. C'est pourquoi nous avons éliminé de l'analyse, pour l'ensemble des tests, quatre sujets qui présentaient une répartition des réponses plus déséquilibrée que 70%-30%. Pour les 19 sujets restants, que ce soit pour les séquences $V_1V_2V_1$ ou les séquences V_1CV_1, les réponses aux tests de qualité de la synthèse n'ont pas permis de mettre en évidence une différence significative entre les classes de stimuli. Ces tests ne permettent donc pas de démontrer qu'il existe une influence du réalisme de la modélisation physique sur le caractère plus ou moins naturel de la synthèse.

Perception des stimuli à centres silencieux.

Si on considère toutes les réponses correctes données pour les stimuli à centre silencieux par les 19 sujets sélectionnés, aucune différence entre les modèles ne peut être montrée. Cependant diverses études [5] ont montré que les temps de latence (TL) pouvaient être très informatifs des mécanismes perceptifs impliqués, notamment les TL courts. En effet, de TL longs peuvent être liés à l'implication de traitements cognitifs de haut niveau, dépassant le stade de la perception auditive proprement dite.

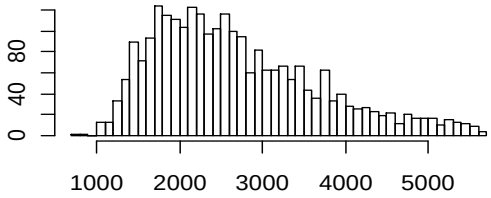


Figure 1 : Distribution des temps de latence (ms)

Pour notre test, la variabilité est grande, avec une moyenne, tous sujets et tous stimuli confondus, de 2885ms et un écart-type de 1375ms. La figure 1 donne la distribution des TL dans l'intervalle [-2σ +2σ].

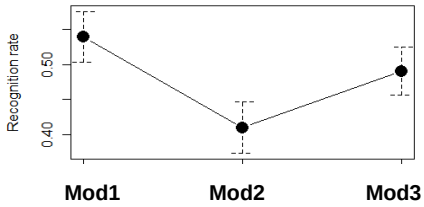


Figure 2 : Moyennes et erreurs-types (Int. Conf. :0.95) (nombre sujets : 19 ; nombre données =592)

Sur cette base, seules les réponses avec des TL dans l'intervalle [1000ms 2000ms] ont été prises en compte. Les moyennes et erreurs-types du pourcentage de bonnes réponses calculées par classe de stimuli sont

présentées fig. 2. Ce pourcentage est significativement meilleur (pMCMC=0.016, t=-2.45) pour les stimuli ***Mod1*** que ***Mod2***. Une tendance suggère un meilleur taux d'identification des stimuli ***Mod3*** que ***Mod2***, mais elle n'est pas significative (pMCMC=0.0874, t=1.678). Les différences entre les stimuli ***Mod1*** et ***Mod3*** ne sont pas significatives (pMCMC=0.4098, t=-0839).

4. CONCLUSIONS

Les tests évaluant le caractère naturel de la synthèse n'ont pas permis de faire apparaître un rôle quelconque du réalisme physique de la modélisation. Ce résultat ne va pas dans le sens de notre explication aux résultats expérimentaux de Cai et al. [1]. Cependant, il est possible que leurs stimuli perturbés soit encore moins écologiques que ceux qui ont été générés par le moins physique de nos modèles (***Mod2***). Nos résultats ne permettent pas non plus de dire que les auditeurs exploitent une connaissance du comportement physique des articulateurs pour évaluer le caractère naturel des gestes, contrairement à ce qu'ont montré Viviani et collègues pour la perception visuelle.

Les tests sur les stimuli à centre silencieux donnent des résultats significativement moins bons si les stimuli ont été générés par le modèle purement cinématique (***Mod2***) que s'ils l'ont été avec le modèle le plus réaliste (***Mod1***). Ainsi le réalisme physique semble aider à retrouver dans les transitions l'information phonétique manquante. Une information purement directionnelle (***Mod2***) sur la variation formantique est moins efficace qu'une description dynamique plus complexe. Ces résultats confirment le rôle des trajectoires dans la perception de la parole dégradée. Cependant pour les stimuli de la classe ***Mod1***, puisque le modèle de contrôle est de type « cible », les trajectoires sont intrinsèquement liées à la cible du mouvement. Nos résultats suggèrent donc que les auditeurs pourraient extraire des trajectoires l'information sur la physique des articulateurs afin de retrouver la cible manquante. Cela va dans le sens des hypothèses formulées par Loevenbruck & Perrier [3].

REMERCIEMENTS

A Brad Story (Univ. of Arizona) pour son synthétiseur acoustique. Ce travail est soutenu par l'Université Franco-Allemande (Sarrebruck) (Projet PILIOS) et le financement SFRH/BPD/48002/2008 du FCT Portugal.

BIBLIOGRAPHIE

[1] Cai, S., Boucek, M., Ghosh, S.S., Guenther, F. H. & Perkell, J.S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/. *Proc. of ISSP-2008.*, (pp. 65-68), Strasbourg, France.

[2] Hillenbrand J. & Gayvert R. (2005) Open Source Software for Experiment Design and Control, *J.S.L.H.R., 48,* 45-60.

[3] Lœvenbruck H. & Perrier P. (1996). How could undershot vowel targets be recovered? A dynamical approach based on the Equilibrium Point Hypothesis for the control of speech movements. *Proc. of ISSP-1996* (pp. 117-120), Autrans, France.

[4] Lindblom, B. & Studdert-Kennedy M. (1967) On the role of formant transitions in vowel recognition. *J Acoust Soc Am, 42*, 830-843.

[5] Miller, J.L. & Dexter, E.R. (1988). Effects of speaking rate and lexical status on phonetic perception. *J Exp Psychol Hum Percept Perform, 14*, 369–378.

[6] Nearey, T. (1977). *Phonetic feature systems for vowels*,Doctoral dissertation, University of Connecticut, Storrs, CT.

[7] Payan, Y. & Perrier, P. (1997). Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech Comm., 22 (2/3)*, 185-205.

[8] Perrier P., Boë L.J. & Sock R. (1992). Vocal Tract Area Function Estimation From Midsagittal Dimensions With CT Scans and a Vocal Tract Cast: Modeling the Transition With Two Sets of Coefficients. *J.S.H.R.,35*, 53-67

[9] Perrier, P., Payan, Y., Zandipour, M. & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: a modeling study. *J Acoust Soc Am, 114 (3)*, 1582-1599.

[10] Perrier, P. & Fuchs, S. (2008). Speed-curvature relations in speech production challenge the 1/3 power law. *J Neurophysiol, 100 (3), 1171-1183.*

[11] Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data, in E. E. David, Jr. & P. B. Denes, (Eds.), *Human Communication: A Unified View* (pp. 51–66).

[12] Strange, W., Edman, T.R., & Jenkins, J.J. (1979). Acoustic and phonological factors in vowel identification. *J Exp Psychol Hum Percept Perform, 5(4),* 643-656

[13] Strange, W., Jenkins, J.J. & Johnson, T.L. (1983). Dynamic specification of coarticulated wovels, *J Acoust Soc Am, 74(3)*, 695-705.

[14] Stevens, K. N. & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *J Acoust Soc Am, 64 (5)*, 1358-1368

[15] Story, B. H. (2004). Vowel acoustics for speaking and singing, *Acta Acustica 90(4)*, 629--640.

[16] Story, B.H. (2005). A parametric model of the vocal tract area function for vowel and consonant simulation, *J. Acoust. Soc. Am., 117(5)*, 3231-3254.

[17] Titze, I.R. (1984). Parameterization of the glottal area, glottal flow, and vocal fold contact area, *J. Acoust. Soc. Am., 75*, 570-580.

[18] Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions *J Exp Psychol Hum Percept Perform*, 18 (3), 603-623..

Corrélats neuroanatomiques des systèmes de perception et de production des voyelles du Français

Krystyna Grabski[1], Laurent Lamalle[2,3], Jean-Luc Schwartz[1], Coriandre Vilain[1], Nathalie Vallée[1], Irène Tropres[2,4], Monica Baciu[5], Jean-François Le Bas[2,6], Marc Sato[1]

[1]GIPSA-Lab, Département Parole et Cognition, UMR CNRS 5216 & Grenoble Universités; [2]IFR1 'RMN Biomédicale et Neurosciences', Unité IRM 3T, CHU de Grenoble; [3]INSERM; [4]Université Joseph Fourier; [5]Laboratoire de Psychologie et NeuroCognition, UMR CNRS 5105 & Université Pierre Mendès France; [6]CHU de Grenoble

krystyna.grabski@gipsa-lab.grenoble-inp.fr

ABSTRACT

Recent neurobiological studies argue for a tight connection between speech perception and production systems. By means of functional magnetic resonance imaging, we here investigated whether common sensorimotor brain areas might participate in both vowel perception and production. A conjunction analysis between speaking and listening conditions showed a shared neural network with common activations observed in the left inferior frontal gyrus and in the superior temporal gyrus bilaterally. Direct comparisons between the two tasks also revealed specific modulated responses within the superior temporal and supramarginal gyri. These results provide evidence for sensory-to-motor and motor-to-sensory feedback control mechanisms during vowel perception and production.

Keywords: vowel perception and production, internal model, sensorimotor interactions, fMRI.

1. INTRODUCTION

Bien que suggérée depuis plus d'un siècle par les théories neurologiques issues de l'études de patients aphasiques, l'hypothèse d'une séparation fonctionnelle stricte des régions temporales auditives et frontales motrices lors de la perception et de la production de la parole apparaît de plus en plus contestée. Des études récentes en neuroimagerie/neurophysiologie suggèrent en effet l'existence de mécanismes de contrôle et de comparaison en ligne des représentations articulatoires et auditives et appuient ainsi l'hypothèse d'un possible couplage fonctionnel des systèmes de perception et de production de la parole [1-2].

De fait, au-delà de l'implication des régions auditives temporales, l'activation de régions corticales impliquées dans la planification et l'exécution des gestes de parole (à savoir, le gyrus frontal inférieur gauche, le cortex prémoteur ventral et le cortex moteur primaire) et des aires proprioceptives liées aux mouvements de la bouche (dans le cortex somatosensoriel) ont été régulièrement observées lors de la perception auditive, visuelle ou audiovisuelle de la parole. Ces études suggèrent ainsi que la perception de la parole implique l'activation automatique et spécifique de représentations motrices liées aux gestes articulatoires perçus [3-5]. De même, la production de gestes de parole semble impliquer de telles interactions entre régions sensorielles et motrices. En effet, une modulation de l'activité des régions auditives temporales et somatosensorielles pariétales a été démontrée lors de la production ouverte et même silencieuse de parole. La modulation d'activité de ces régions sensorielles refléterait ainsi l'existence de mécanismes de contrôle et de comparaison en ligne entre les conséquences sensorielles prédites et le feedback sensoriel réel des gestes produits [6-8].

Face à ce possible couplage fonctionnel et à de précédentes études ayant permis d'obtenir une première caractérisation générale des régions neuronales associées à la production et perception de séquences de parole plus complexes sur le plan articulatoire et acoustique, cette étude en imagerie par résonance magnétique fonctionnelle (IRMf) a pour but, ce pour la première fois, une exploration des cartes sensorielles et motrices activées aussi bien lors de la production que lors de la perception des voyelles du Français chez les mêmes participants.

2. MÉTHODES

2.1 Participants

Quatorze volontaires droitiers de langue maternelle française ont participé à l'étude (dont 8 hommes; âge : 21-43 ans). Cette étude a reçu un avis favorable du Centre Hospitalier Universitaire de Grenoble, du Comité de Protection des Personnes pour la Recherche Biomédicale de Grenoble et de l'Agence Française de Sécurité Sanitaire des Produits de Santé.

2.2 Stimuli et Procédure

L'expérience impliquait deux tâches successives de perception et de production des voyelles. Les voyelles /i/, /y/, /u/, /e/, /ø/, /o/, /ɛ/, /œ/, /ɔ/, préalablement enregistrées dans une chambre sourde par chacun des participants pour la tâche de perception, ont été utilisées. Les voyelles ont été choisies de manière à effectuer des analyses relatives aux traits phonétiques, en regroupant les voyelles selon le degré d'ouverture de la mâchoire (voyelles mi-ouvertes, voyelles mi-fermées, voyelles fermées) ainsi que la position de la langue et

arrondissement des lèvres (antérieure-non arrondie, antérieure- arrondie postérieure-non arrondie). Il est à noter que les deux tâches impliquaient une perception ou une production de voyelles maintenues (d'une durée approximative de 600ms).

La tâche de perception consistait en l'écoute attentive de chacune des voyelles propres au participant. Lors de la tâche de production, une consigne visuelle indiquait pour chaque essai le type de voyelle à produire. Pour ces deux tâches, une voyelle était perçue ou produite toutes les 10s selon un ordre pseudo-aléatoire. Chaque tâche incluait trois blocs de 60 essais d'une durée de 10 minutes: soit 6 occurrences pour chacune des 9 voyelles et 6 essais consistant en une condition de repos. 360 scans fonctionnels ont ainsi été acquis (2 tâches x 3 blocs x 60 essais x 10s) pour une durée totale d'environ 75 minutes.

2.3 Matériel et acquisition des données IRM

A l'aide du logiciel Presentation (Neurobehavioral Systems, Albany, EU), les consignes visuelles ont été projetées au moyen d'un vidéo projecteur sur un écran situé derrière le participant et, par réflexion, sur un miroir placé au dessus de ses yeux. Un système casque-microphone compatible IRM a été utilisé pour la transmission des stimuli auditifs et l'enregistrement des participants. Lors de l'expérience, les participants portaient des bouchons d'oreille et un casque antibruit.

Les acquisitions des images anatomiques et fonctionnelles ont été réalisées sur un imageur corps entier 3T (Bruker Medspec S300) muni d'une antenne tête émission/réception à champ de vue large. Pour les scans fonctionnels, une séquence d'acquisition en écho de gradient pondérée en T2* a été utilisée. Pour chaque volume fonctionnel, quarante coupes axiales adjacentes ont été acquises en mode entrelacé (temps de répétition: 10s, temps d'acquisition : 2600ms, résolution: 3 mm^3). Entre les conditions de perception et de production, un volume anatomique de haute résolution (1 mm^3) pondérée en T1 a également été acquis.

Afin de minimiser de possibles artefacts de mouvement sur les images fonctionnelles, un paradigme d'acquisition de type 'sparse sampling' a été utilisé. Cette technique d'acquisition est basée sur le délai temporel existant entre l'activé neuronale liée à une tâche motrice ou à l'écoute d'un stimulus auditif et le délai de la réponse hémodynamique associée. Face au délai estimé dans de précédentes études du pic de la réponse hémodynamique lors de la production de mouvement orofaciaux ou de séquences de parole [7-11], l'intervalle de temps séparant la perception ou la production d'une voyelle et l'acquisition du volume fonctionnel correspondant variait aléatoirement entre chaque essai de 4s à 6s.

2.4 Prétraitements et analyses statistiques

Les données ont été analysées à l'aide du logiciel SPM5 (Statistical Parametric Mapping; Wellcome Department of Cognitive Neurology, Londres, RU). Pour chacun des participants, les images fonctionnelles ont été réalignées, normalisées dans l'espace commun du Montreal Neurological Institute (repère MNI) et lissées via un filtre gaussien passe-bas de 6 mm^3. Les réponses hémodynamiques correspondantes aux conditions expérimentales ont ensuite été estimées selon un modèle linéaire général, incluant la caractérisation d'une réponse à impulsion unique pour chaque scan fonctionnel et l'ajout de régresseurs de non-intérêt liés aux paramètres de mouvements. Enfin, un filtrage des basses fréquences *a priori* non-reliées aux conditions expérimentales (variations lentes d'origine physiologique) a été appliqué (fréquence de coupure de 1/128 Hz). Suite à l'estimation pour chaque participant des activations observées lors de la perception et de la production des voyelles par rapport à la condition de repos, une analyse de groupe "à effets aléatoires" a été réalisée via une ANOVA à mesures répétées. Différents contrastes ont été alors estimés de manière à déterminer les régions activées lors de la tâche de perception et lors de la tâche de production, ainsi que les activations communes ou spécifiques à ces deux tâches (voir Figure 1). L'ensemble de ces analyses ont été calculées selon un seuil statistique défini à $p < .01$ corrigé et une taille minimale des clusters de 100 voxels.

3. RÉSULTATS

Perception: Par rapport à la condition de repos, la perception des voyelles implique une activation bilatérale de régions du gyrus temporal supérieur (GTS). Ces régions, situées de part et d'autre des sillons latéral et temporal supérieur, incluent l'aire auditive primaire (gyrus de Heschl, HG), ainsi que des régions antérieures et postérieures (aire de Wernicke) associatives. Enfin, une activation de la partie operculaire du gyrus frontal inférieur gauche (aire de Broca) est également observée.

Production: La production des voyelles implique un ensemble de régions corticales et sous-corticales classiquement dévolues au contrôle moteur. Ces régions, activées de manière bilatérale, incluent le cortex moteur primaire orofacial, le cortex somatosensoriel adjacent et une partie du lobule pariétal inférieur, l'aire motrice supplémentaire s'étendant au gyrus cingulaire antérieur, le cortex prémoteur ventral et la partie operculaire du gyrus frontal inférieur, l'opercule rolandique et l'insula, les ganglions de la base et le thalamus. On observe également une activation du cunéus et du précunéus. Enfin, on retrouve les régions auditives temporales telles qu'observées en perception.

Activations communes: Les régions communes activées lors des deux tâches correspondent à l'ensemble des régions activées lors de la perception des voyelles.

Activations différentiées: A l'exception de l'aire de Broca, la comparaison de différences d'activations entre les deux tâches implique les régions motrices observées lors de la tâche de production (ces mêmes régions n'étant pas activées lors de la tâche de perception). En revanche, aucune différence significative n'est observée entre les deux tâches par rapport aux régions temporales auditives

citées précédemment et à l'aire de Broca. Enfin, cette comparaison fait également ressortir une déactivation bilatérale de certaines régions lors de la production des voyelles, ce aussi bien par rapport à la tâche de perception que par rapport à la condition de repos. Ces régions correspondent à la partie ventro-postérieure du gyrus supramarginal (GSM), à la partie dorso-postérieur du gyrus temporal supérieur, au gyrus cingulaire postérieur et à une partie du gyrus frontal supérieur droit.

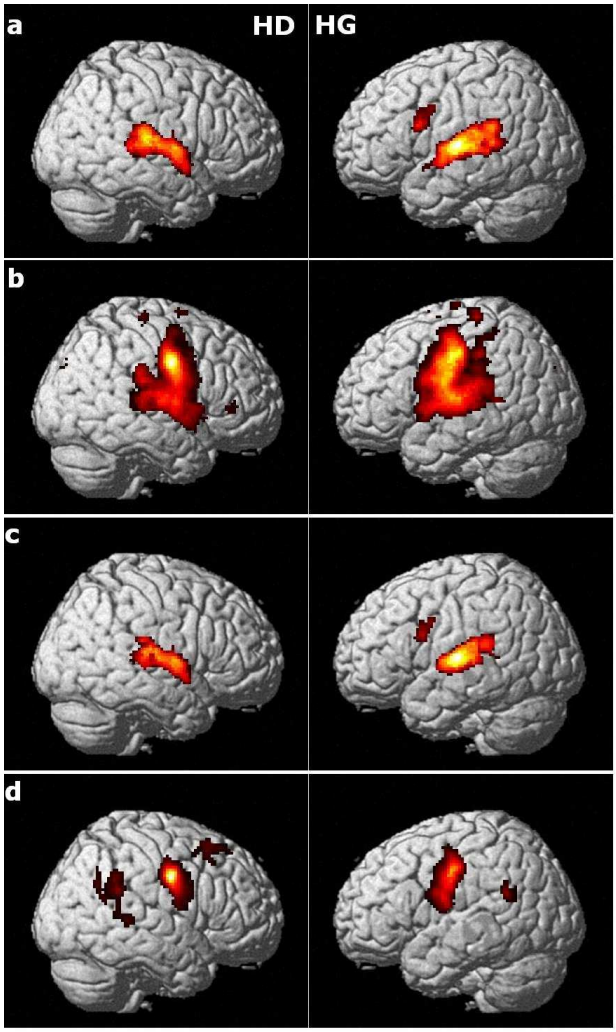


Figure 1: Projections surfaciques des activations observées lors de la perception (a) et de la production (b) des voyelles, ainsi que les activations communes (c) ou spécifiques (d) aux deux tâches (p < .01 corrigé, taille minimale des clusters de 100 voxels, HD/HG: hémisphère droit/gauche).

4. DISCUSSION

De manière attendue, des activations bilatérales du cortex temporal auditif ont été observées lors de la perception des voyelles. Ces régions sont traditionnellement considérées comme impliquées dans les traitements acoustiques (gyrus de Heschl [12-13]) et de décodage acoustico-phonétique (partie antérieure [14-15] ou postérieure [16] du GTS). De même, en plus des activations du cortex auditif, la plupart des régions activées lors de la tâche de production sont classiquement impliquées dans le contrôle moteur de la parole [6-11, 17]. Enfin, l'activation du cunéus en production pourrait refléter des processus visuels liés à la consigne tandis que le précuneus ferait appel à l'imagerie visuo-spatiale et au 'soi-acteur' lors de la réalisation d'une action [18].

Bien que les déactivations observées lors de la production des voyelles doivent être considérées avec précaution (les mécanismes sous-jacents étant encore mal compris et difficilement interprétables), de telles déactivations induites par la tâche pourraient refléter une réattribution de ressources de traitement des aires déactivées envers d'autres aires nécessaires pour la performance de la tâche et/ou l'existence d'interactions inhibitrices dans ces régions dans le but d'inhiber des pensées et comportements inappropriées dans le contexte de la tâche demandée [21]. De plus, la déactivation des régions associatives auditives et somatosensorielles (parties dorso-postérieure du GTS et ventro-postérieure du GSM) appuient indirectement l'hypothèse de mécanismes de contrôle permettant de distinguer entre les conséquences sensorielles attendues d'une production propre et les signaux sensoriels associés aux changements ou perturbations du monde extérieur [6-8, 22]. Il a été en effet observé une plus forte activation de ces régions lors d'une condition de repos par rapport à une condition de production de parole [22]. Inversement, ces régions sont plus fortement activées lors la production de parole avec modification en ligne du feedback acoustique par rapport à une production normale [8]. Il est dès lors possible de postuler une activité 'attentionnelle' minimale de ces régions en dehors de toute action par le sujet et, lors de la production de parole, une activation moindre du fait de projections inhibitrices par le système moteur (aire de Broca) ou, inversement, accrue en cas de conflit entre conséquences sensorielles attendues et effectives.

L'activation de l'aire de Broca, typiquement impliquée dans les processus de planification articulatoire [17], est également retrouvée dans de nombreuses études de perception de la parole, par exemple lors de tâches d'écoute 'passive' [5, 22-23] ou d'identification phonétique [19]. De plus, des études ont démontré l'engagement de cette région aussi bien lors de la perception que lors de la production de syllabes ou de phrases [5, 22-23]. Cette activation suggère ainsi la mise en œuvre de mécanismes de simulation motrice [1,2] lors de la perception des voyelles. Néanmoins, aucune activation du cortex prémoteur ventral gauche n'est en revanche ici observée. Si ce résultat s'oppose à ceux de précédentes études de perception 'passive' de parole [3-5], il est intéressant de constater que ces études consistaient en l'écoute de syllabes ou de phrases et que, pour certaines de ces études [3,4, 22-23], la présentation des stimuli était conjointe au bruit du scanner. Dès lors, l'absence d'activation du cortex ventral prémoteur gauche pourrait être liée à la perception de voyelles intelligibles, moins complexes du point de vue articulatoire que des syllabes et, de plus, propres à chaque participant. La perception de ces voyelles pourrait ainsi nécessiter une médiation moindre du système moteur dans des

mécanismes de simulation motrice interne, du fait de processus de désambiguïsation et de recodage phonétique de l'input acoustique simplifiés [20].

5. CONCLUSIONS & PERSPECTIVES

Cette étude constitue une première exploration des régions sensorielles et motrices activées lors de la production et lors de la perception des voyelles du Français. A notre connaissance, peu d'études ont examiné les substrats neuronaux liés aux systèmes de perception [15] et de production [11] vocalique et aucune de manière conjointe. Pris ensemble, nos résultats démontrent l'existence de mécanismes de contrôle et de comparaison en ligne des représentations articulatoires et auditives et, de là, appuient l'hypothèse d'un possible couplage fonctionnel des systèmes de perception et de production de la parole [1-2].

Par la suite, les réseaux de perception et de production de la parole seront étudiés plus en détail, ce en fonction de chaque voyelle et de leurs traits articulatoires respectifs. Pour cela, les résultats d'une étude IRMf démontrant une somatotopie motrice des articulateurs supralaryngés de la parole réalisée auprès des mêmes participants (voir Grabski, K. et al., "Somatotopie motrice des articulateurs supralaryngés de la parole", ce volume) seront utilisés en tant que localisateurs moteurs. Ces analyses ultérieures devraient permettre d'affiner les présents résultats en établissant une cartographie précise des activations motrices, somatosensorielles et auditives liées à la perception et production des voyelles du français et ainsi de mieux comprendre de quelle façon le cerveau code et exploite les informations phonétiques/phonologiques.

REMERCIEMENTS

Cette étude s'inscrit dans le cadre du BQR "Modyc: Modélisation dynamique de l'activité cérébrale" financé par l'Institut National Polytechnique de Grenoble.

BIBLIOGRAPHIE

[1] Schwartz, J.-L., Sato, M. and Fadiga, L. The common language of speech perception and action: a neurocognitive perspective. *Revue Française de Linguistique Appliquée*, 13(2): 9-22, 2008.

[2] Schwartz, J.-L., Ménard, L., Basirat, A. and Sato, M. The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, sous presse.

[3] Wilson, S.M. et al. Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.*, 7: 701-702, 2004.

[4] Pulvermüller F. et al. Motor cortex maps articulatory features of speech sounds. *PNAS*, 13: 7865-7870, 2006.

[5] Skipper, J.I., Van Wassenhove, V., Nusbaum, H.C. and Small, S.L. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17(10): 2387-2399, 2007.

[6] Guenther, F.H. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39: 350-365, 2006.

[7] Bohland J.W. and Guenther F.H. An fMRI investigation of syllable sequence production. *NeuroImage*, 32(2): 821-841, 2006.

[8] Tourville, J.A., Reilly, K.J. and Guenther F.H. Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, 39(3):1429-43, 2008.

[9] Gracco, V.L., Tremblay, P. And Pike, G.B. Imaging speech production using fMRI. *Neuroimage*, 26: 294-301, 2005.

[10] Özdemir, E., Norton, A. And Schlaug, G. Shared and distinct neural correlates of singing and speaking. *NeuroImage*, 33: 628–635, 2006.

[11] Sörös P. et al. Clustered functional MRI of overt speech production. *NeuroImage*, 32(1): 376-387, 2006.

[12] Binder, J.R. et al. Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10:512-528, 2000.

[13] Zatorre, R.J., Belin, P. and Penhune, V.B. Structure and function of auditory cortex: music and speech. *Trends Cognit. Sci.*, 6: 37–46, 2002.

[14] Scott, S.K. and Wise, R.J. The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, 92: 13-45, 2004.

[15] Obleser, J. Et al. Vowel Sound Extraction in Anterior Superior Temporal Cortex. *Human Brain Mapping*, 27: 562–571, 2006.

[16] Hickok, G. And Poeppel, D. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8: 393-402, 2007.

[17] Jürgens, U. Neural pathways underlying vocal control. *Neuroscience and Biobehavioral Reviews*, 26: 235-258, 2002.

[18] Cavanna, A.E. and Trimble, M.R. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564-583, 2006.

[19] Callan, D. and Kawato, M. Influence of brain regions involved with articulatory processinf of phoneme identification. In *Proc. Intl. Conf. on Phonetic Sciences*, pages 1873-1876, 1996

[20] Sato, M., Tremblay, P. and Gracco, V. A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language*, 111(1): 1-7, 2009.

[21] McKiernan, K.A. et al. A parametric manipulation of factors affecting task-induced deactivation in functional neuroimaging. *Journal of Cognitive Neuroscience*, 15(3):394-408, 2003.

[22] Jardi, R. et al. Self awareness and speech processing: An fMRI study. *NeuroImage*, 35: 1645–1653, 2007.

[23] Okada, K and Hickok, G. Left posterior auditory-related cortices participate both in speech perception and speech production: Neural overlap revealed by fMRI. *Brain and Language*, 98: 112-117, 2006.

Etude longitudinale des productions multimodales d'enfants français âgés de 18 mois à 3 ans et demi (41 mois).

Aurore Batista & Jean-Marc Colletta.

Laboratoire LIDILEM (Laboratoire de linguistique et didactique des langues étrangères et maternelles)
aurore.batista@u-grenoble3.fr, jean-marc.colletta@u-grenoble3.fr
Université de Grenoble - 1180, avenue centrale - 38040 Grenoble Cedex 9

1. ABSTRACT

The current paper deals with the effects of an experience which aims at improving the French underprivileged children's oral abilities.
We are going to investigate the success of this technique on (1) our subjects' capacities to produce multimodal or verbal combination to communicate and (2) the length of their statements.
We discovered that this experience has a notable effect on our subjects' global production: the size of their statements is longer.

Keywords:
Multimodality, child, development, language, performance, MLU.

2. INTRODUCTION

Le projet « P.A.R.L.E.R.[1] » a été mis en place en collaboration avec de nombreux partenaires[2] dans plusieurs crèches et écoles primaires de l'agglomération grenobloise (38) afin de favoriser les acquisitions scolaires enfantines.
Dans la présente recherche, nous allons cibler une partie précise du projet « P.A.R.L.E.R.» nommée « P.A.R.L.E.R. Bambin » concernant 70 enfants grenoblois placés en crèche et âgés de 18 mois à 3 ans et demi environ. Au sein de ce groupe d'enfant, les « parleurs tardifs » ont été identifiés [3] grâce à un questionnaire rempli par les parents. Ces enfants ont ensuite participé par petits groupes de trois, à des ateliers leur permettant de développer leurs compétences langagières et communicationnelles. Pour ce faire, une éducatrice professionnelle avait pour rôle de favoriser l'interaction au sein du groupe d'enfants ainsi qu'entre elle et les enfants, en utilisant des jeux et des livres d'images. Pour connaître l'effet de ces ateliers, des enfants ont été filmés dans une tâche de jeu avant puis après y avoir été soumis. À des fins de comparaisons, les enfants d'un groupe contrôle ont aussi été filmés aux mêmes périodes.

Le but de la présente étude est de comparer de manière globale[3] les productions des groupes « test » et « contrôle » afin de savoir si la méthode a porté ses fruits et ce, à quel niveau.
Divers chercheurs américains [5] et italiens [8] ont soulignés les éléments primordiaux du développement enfantin :
- A 16 mois, l'enfant italien utilise des mots mais communique majoritairement grâce à des gestes seuls. Cette tendance s'inverse à 20 mois : les mots seuls deviennent son moyen de communication privilégié [8].
- Pour les enfants italiens et américains, utiliser un mot seul (« stade un mot » [5]) ne semble être qu'une étape transitoire du développement verbal, ils deviennent très vite capables de produire des énoncés à deux (« stade deux mots » [5]) ou trois mots et plus par la suite.
- Quand ces enfants verbalisent un mot (à 20 mois), ce dernier est très souvent accompagné d'un geste déictique [8]. Ces combinaisons multimodales permettent d'abord à l'enfant italien d'introduire des mots déictiques (pointage + « *là* ») et des dénominations (pointage vers une balle + « *balle* »), puis des mots à valeur supplémentaire (pointage vers une balle + « *tombée* »).
Nous pouvons conclure de ces études que l'enfant communique en premier grâce à des énoncés[4] composés d'un seul élément (gestuel puis verbal) puis en utilisant des combinaisons multimodales (combinaisons geste-mot) ou verbales (association de deux mots) avant d'être capable de créer des énoncés à trois éléments et plus [1] [2]. Nous allons étudier les patterns de cette évolution chez les enfants français.
Nous supposons que le fait d'échanger, en groupe restreint, plusieurs fois par semaines, avec un adulte francophone peut influencer un paramètre de la progression langagière que nous avons décrite ci-dessus :
- la compétence de l'enfant à combiner des éléments verbaux ou multimodaux pour produire du

[1] **P**arler **A**pprendre **R**éfléchir **L**ire **E**nsemble pour **R**éussir

[2] L'Inspection Académique de l'Isère, Grenoble-Alpes Métropole, le Conseil Régional Rhône-Alpes, le Laboratoire des Sciences de l'Education de l'Université Pierre Mendès France de Grenoble et les municipalités d'Echirolles, de Fontaine et de Grenoble.

[3] Nous n'avons pas le temps ici de faire une analyse en profondeur mais pour plus d'informations à ce sujet voir [1] [1].

[4] Par « énoncé », nous entendons ici aussi bien l'énoncé verbal que « l'énoncé bimodal » formé d'un (ou plusieurs) mot(s) et d'un geste.

sens. Nous supputons que grâce à l'entraînement précoce du projet PARLER :

- les enfants composant le groupe testé seront plus vite capables de produire des énoncés composés de deux ou trois éléments et plus.
- la compétence de l'enfant à produire des énoncés composés de plus de trois éléments sera influencée.

3. OBJECTIFS

Pour commencer, nous allons analyser le développement de l'ensemble de nos sujets (groupes testé et contrôle confondus) en fonction du nombre d'éléments composant leurs énoncés [1]. Notre double objectif est de :

- Mieux appréhender le développement langagier général des enfants issus de milieux défavorisés,
- Observer si la méthode « P.A.R.L.E.R. Bambin » influe sur le développement de leur compétence à combiner des éléments communicationnels pour produire du sens.

Dans un premier temps, nous utiliserons donc un indice basé sur la longueur moyenne des énoncés ou LME [1] [4] [6] [7] pour étudier si la méthode influe sur la taille des verbalisations produites par l'enfant.

4. CORPUS

Soixante dix enfants[5] âgés de 18 à 41 mois (Voir Table 1) ont été filmés, deux fois en six mois, en situation de jeu triadique avec une maison de jeu Fisher Price et un adulte.

Table 1 : Répartition de l'effectif en fonction de l'âge, du groupe d'appartenance et de la période filmée.

	Groupe contrôle		Groupe testé	
	T0	**T1**	**T0**	**T1**
18-23	11		13	
24-29	12	11	18	13
30-35	10	12	6	18
36-41		10		6
Total	33		37	

L'enfant était libre de communiquer avec l'adulte et de jouer comme bon lui semblait. L'adulte, quand à lui, tenait le rôle d'un partenaire de jeu bienveillant dont la tâche était d'engager l'enfant à communiquer lorsque ce dernier restait muet en attirant son attention vers des éléments de la situation de jeu. Seuls 33 de ces enfants ont participé aux ateliers du projet « Parler Bambin », les 37 autres correspondent au groupe contrôle.

5. MÉTHODOLOGIE :

5.1. Transcription et annotation des énoncés :

Grâce au logiciel ELAN®[6], nous avons transcrit par groupes de souffle les verbalisations produites par l'adulte et l'enfant sur deux lignes séparées. Après, nous avons annoté les énoncés verbaux de l'enfant en fonction du nombre de mots le composant (un mot, deux mots ou trois mots et plus). Pour finir, les gestes seuls ainsi que les combinaisons geste-mot ont été transcrits sur une quatrième ligne.

5.2. Classement en fonction du nombre d'élément composant les énoncés :

Comme nous l'avons dit en introduction, nous avons d'abord étudié l'ensemble des productions langagières de nos sujets en nous intéressant au nombre d'élément qui compose leurs énoncés multimodaux [7] :

- Les 1EL (ou énoncés à un élément) prenant en compte les gestes communicationnels produits seuls (pointages) et les énoncés composés d'un mot.
- Les 2EL (ou énoncés à deux éléments) représentant les combinaisons multimodales (geste-mot) et les énoncés composés de deux mots.
- Les 3EL+ (ou énoncés à trois éléments et plus) englobant toutes les autres productions.

5.3. La longueur moyenne des énoncés verbaux ou LME :

La LME (« *Mean length of utterance* » [4]) a été calculée en nombre de mots suivant l'équation suivante : $\frac{\text{Nombre total de mots}}{\text{Nombre total d'énoncés verbaux}}$[8]

Le résultat de cette équation nous sert d'indice pour évaluer l'âge linguistique de l'enfant en fonction de ses performances verbales. Le classement en fonction du nombre d'éléments ne prend pas en compte le nombre de mots réellement prononcés par l'enfant. Par contre, la LME pallie ce manque car elle permet d'apprécier l'évolution des conduites linguistique de l'enfant de manière plus fine.

6. RÉSULTATS

Nous allons commencer par rappeler les résultats obtenus pour l'ensemble de nos sujets [1] avant de passer à une analyse opposant groupe contrôle et groupe testé selon le nombre d'éléments composants leurs énoncés puis en fonction de la LME.

[5] 85 enfants ont été filmés en tout mais seulement 70 à deux reprises car certains d'entre eux n'étaient pas présents lors de la seconde séance d'enregistrement ce qui explique les différences d'effectifs entre groupe contrôle et groupe testé.

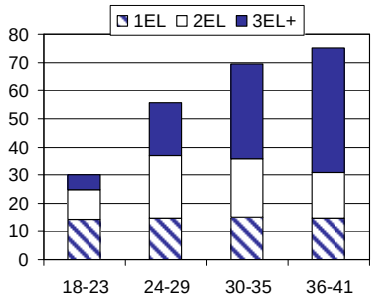
[6] http://www.lat-mpi.eu/tools/elan/

[7] Pour une analyse plus précise prenant en compte tous les autres types de gestes voir [1] [1].

[8] Dans notre cas, nombre total de groupes de souffles.

6.1. Productions globales, tous groupes de sujets confondus

Le Graphique 1, nous permet de constater que :
- globalement, la production verbale augmente avec l'âge (30 groupes de souffles comptabilisés dans la première classe d'âge contre 75 dans la dernière).
- nos sujets les plus jeunes produisent 50% d'1EL, le reste des productions se divise entre 2EL (35%) et 3EL+ (15%).
- chez les 24-29 mois, 40% des énoncés produits sont des 2EL, 30% des 1EL et les 30% restants des 3EL+.
- 50% des productions des 30-35 mois sont des 3EL+ et les 3EL représentent plus de 60% dans la dernière classe d'âge.

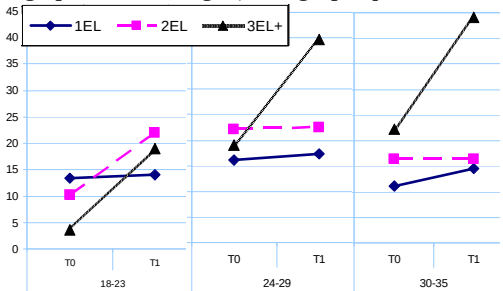


Graphique 1 : Moyennes des énoncés produits dans les quatre tranches d'âge.

Ceci nous montre qu'il y a, en effet, une progression des productions langagières enfantines allant du plus simple (1EL) au plus complexe (3EL+). Cependant, même si l'enfant produit une plus grande quantité de 1EL ou 2EL respectivement dans la première et la seconde classe d'âge, il est aussi capable d'employer des 3EL+ aux mêmes périodes. La compétence des sujets à produire des énoncés n'est donc pas réduite à un type d'énoncé.
Maintenant que nous avons évalué la production moyenne de tous nos sujets, nous allons comparer son évolution entre T0 et T1[9] afin de voir les effets des ateliers « P.A.R.L.E.R. ».

6.2. Comparaison entre T0 et T1.

Nous allons étudier l'évolution constatée classe d'âge par classe d'âge (Voir graphique ci-dessous).



[9] Soit T0, le temps du premier film avant la période de test et T1 le temps du second film après la période d'expérimentation pour le groupe testé.

Graphique 2 : Moyenne des énoncés produits à T0 et T1 dans chaque classe d'âge.

La comparaison des moyennes montre que l'évolution des énoncés produits entre T0 et T1 par les membres de :

<u>- la première classe d'âge :</u>

→ n'est pas significative pour les 1EL mais est significative pour les 2EL : t(24) = -3,575 ; p = ,002 ainsi que pour les 3EL+ : t(24) = -4,509 ; p = ,000

<u>- la seconde classe d'âge :</u>

→ n'est pas significative pour les 1EL ni pour les 2EL mais est significative pour les 3EL+ : t(30) = -4,28 ; p = ,000

<u>- la dernière classe d'âge :</u>

→ n'est pas significative pour les 1EL ni pour les 2EL mais est significative pour les 3EL+ : t(16) = -3,729 ; p = ,002

Parmi les productions de l'enfant, la part des énoncés brefs (les 1EL) n'augmente pas car leur acquisition est déjà faite. Par contre, les 2EL évoluent au cours de cette phase critique qui va de 18 à 23 mois. De plus, les 3EL+ évoluent dans toutes les classes d'âges. La capacité à produire des énoncés longs est donc en acquisition.
Si nous comparons les productions des six groupes contrôle et testé, qu'elles comportent 1EL, 2EL ou 3EL+, nous remarquons très peu de différences entre les deux groupes (Voir Table 2) et ces différences ne peuvent pas être attribuées à d'autres facteurs qu'aux variations interindividuelles influant sur nos moyennes.

Table 2 : Moyennes des productions multimodales à T0 et T1 pour chaque classe d'âge.

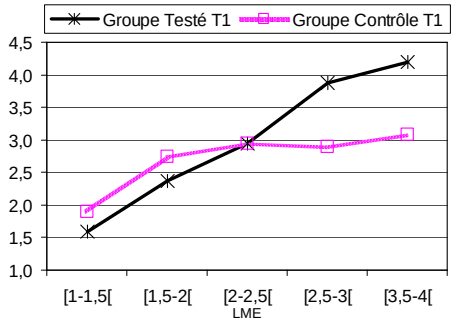
		18-23		24-29		30-35m	
		T0	T1	T0	T1	T0	T1
1EL	Groupe Contrôle	15,8	14,7	18,6	20,0	12,3	17,1
	Groupe Testé	11,3	13,5	14,5	15,6	9,3	10,3
2EL	Groupe Contrôle	10,2	22,1	25,0	25,3	14,1	16,8
	Groupe Testé	10,3	21,9	20,4	20,7	20,5	15,7
3EL+	Groupe Contrôle	3,7	15,6	17,8	42,2	20,4	43,6
	Groupe Testé	3,6	22,1	19,8	38,1	25,2	45,0

D'une part, si on s'en tient à ce résultat, les ateliers du programme « P.A.R.L.E.R. Bambin » n'ont pas d'influence sur les capacités de nos sujets à combiner des éléments. Pourtant, nous allons voir qu'ils ont un effet sur la LME.
Aussi, peut-être, faudrait-il à l'avenir détailler la catégorie des 3EL+ pour mieux prendre en compte le contenu des énoncés bimodaux longs. Concentrons-nous, à présent, sur l'analyse des productions verbales enfantines.

6.3. Comparaison entre groupe contrôle et groupe testé.

Nous proposons, ici, une analyse basée sur les performances purement verbales de nos sujets grâce

à la LME. Cet indice nous permettra tout d'abord, d'avoir une connaissance plus précise de la longueur des énoncés verbaux de nos sujets et de comparer les performances des sujets testés (à présent EGT[10]) avec celles du groupe contrôle (à présent EGC[11]). A T0, les LME des deux groupes sont presque identiques cependant nous notons des différences importantes à T1 (Voir Graphique 3).



Graphique 3 : Comparaison des LME entre groupe testé et groupe contrôle à T1.

Après l'expérimentation, nous ne remarquons aucun progrès entre les EGT et EGC ayant une LME inférieure à 2,5 (exclus). Par contre, nous notons une différence de LME entre les sujets des deux groupes dont l'indice est supérieur à 2,5. Nous pouvons interpréter cet écart comme l'effet du programme et confirmer que les ateliers portent leurs fruits sur le groupe testé en leur permettant de produire plus tôt des énoncés plus longs.

CONCLUSION :

En ce qui concerne les compétences globales des sujets (tous types confondus) à combiner des éléments (verbaux ou multimodaux) pour créer du sens, nos données révèlent une première phase d'acquisition : de 18 à 23 mois, au cours de laquelle les combinaisons 2EL augmentent de manière significative jusqu'à devenir les plus nombreuses à 24 mois. Ensuite, une seconde phase d'acquisition démarre au cours de laquelle les combinaisons à 3EL augmentent significativement. Par la suite, il faudra étudier plus précisément le contenu des ces énoncés verbaux et bimodaux comportant plus de trois éléments. En toute logique, étant donné que la LME augmente, le nombre d'éléments des énoncés bimodaux devrait également augmenter au fil de l'âge. Si tel est bien le cas, nous pourrions alors disposer d'un outil de mesure plus performant que la LME car non limité aux énoncés verbaux.

Notre analyse met en avant l'effet bénéfique du programme sur des sujets ayant déjà une bonne compétence langagière (LME supérieure à 2,5) mais il faudrait aussi considérer les gestes dans cet indice afin d'avoir une appréhension plus réelle des ressources communicationnelles mobilisées par l'enfant.

[10] Enfants du Groupe Testé
[11] Enfants du Groupe Contrôle

REMERCIEMENTS

Nous remercions Marie-Thérèse Le Normand de nous permis de travailler sur les données du programme « P.A.R.L.E.R. Bambin »

BIBLIOGRAPHIE

[1] Batista A. & Le Normand M.-T. (en cours). Etude des productions langagières d'enfants âgés de 17 à 41 mois et issus de quartiers défavorisés : Evaluation des capacités langagières des sujets selon l'âge. In *LIDIL*, 42. Grenoble : PUG.

[2] Batista A. & Colletta J.-M. (en cours). Analyse des productions multimodales d'enfants français âgés de dix sept à quarante et un mois en situation de jeu. In *Actes du colloque AcquisiLyon*, 3-4 décembre 2009.

[3] Bloyer, J. (2009), Parler bambin à l'espace petite enfance « 3 pom' ». In *Langages & réussite éducative : des pratiques innovantes.* Actes du colloque du 11 mars 2009, pp 13-14.

[4] Brown, R.W. (1973). *A first language: the early stages*. Cambridge, Mass. : Harvard University Press.

[5] Goldin-Meadow, S. (2003), *The resilience of language*, Psychology press, New York.

[6] Le Normand, M.-T., Parisse, C. & Cohen, H (2008). Lexical diversity and productivity in French preschoolers: Developmental, Gender and Sociocultural factors. In *Clinical Linguistics and Phonetics*, 22, pp 47-58.

[7] Le Normand, M.-T. (1991), Individual differences in the production of word classes in eight specific language-impaired preschoolers. In *J Commun Disord*, 24, Oct/Dec, pp 331-351.

[8] Volterra, V. & al. (2004). Gesture and the emergence and development of language. In M. Tomasello & D. Slobin (Eds.), *Beyond Nature-Nurture. Essays in Honor of Elizabeth Bates*, pp 3-40. London : Erlbaum.

Etude préliminaire de la perception précoce des voyelles labialisées par des auditeurs déficients visuels

Fabrice Hirsch[1], Henri Dreyfus[2], Rudolph Sock[1], Béatrice Vaxelaire[1], Camille Fauth[1], Fayssal Bouarourou[1], Marion Béchet[1]

[1] Institut de Phonétique de Strasbourg & E.A. 1339 - LiLpa, Composante Parole et Cognition
Université de Strasbourg, 22, rue René Descartes, 67084 Strasbourg
[2] Unité de Recherche INSERM U.44, 12, rue de Copenhague, 67000 Strasbourg
f.hirsch@unistra.fr

ABSTRACT

The present research deals with auditory effects of anticipatory lip movements in [V^1CV^2] sequences, where V^1 is vowel [i], C is the fricative [s] and V^2 a rounded vowel [V^{lab}]. The aim of the investigation is triple: first, we wanted to observe for potential anticipatory acoustic effects of vowel rounding, when a speaker pronounces sequences which include vowels [y], [ø], [u] or [o] in the V^2 position. Secondly, attention is paid to the moment when sighted listeners may start to perceive probable anticipatory effects of the rounded vowel. Thirdly, the core purpose of this work was to analyse the behaviour of blind subjects carrying out the same experimental tasks. Results show that vowels [y] and [u] are perceived earlier than [ø] and [o] by both control and blind subjects. When data for the control group and the group of blind subjects are compared, it is noticed that blind subjects perceive rounded vowels earlier than control sighted subjects.

Keywords: auditory perception, blindness, anticipation, rounded vowels

1. INTRODUCTION

Un grand nombre d'études ([1] par exemple) a montré que les mouvements propres à chaque son d'une séquence [CV] ne sont pas produits de manière successive. Au contraire, les études articulatoires menées sur ce sujet ont observé un phénomène de coarticulation, et plus précisément celui de la coarticulation anticipatoire. Ce phénomène de coarticulation anticipatoire que nous venons d'évoquer a des répercussions sur la perception de la parole. En effet, l'anticipation des gestes de la parole peut être perçue auditivement ([2] ; [3]) et visuellement ([4]), notamment pour des syllabes [CV], où la voyelle est labialisée. En d'autres termes, dans une séquence [su] par exemple, la voyelle [u] peut être perçue, par des auditeurs n'ayant aucun trouble auditif ou visuel, dans le [s], alors même que cette voyelle n'est pas encore rendue acoustiquement.

Si des auditeurs bien-voyants ont la capacité de percevoir précocement un son qui est sur le point d'être produit grâce à certains indices articulatoires et acoustiques, qu'en est-il d'auditeurs mal-voyants ou non-voyants ? La question mérite d'être posée, étant donné que certaines études ont montré des particularités au niveau de la perception des sons et du message linguistique chez les personnes atteintes de cécité. Moos *et al.* [5] ont, par exemple, observé que les auditeurs non-voyants sont capables de percevoir le sens d'une parole utra-rapide à une vitesse d'élocution incompréhensible pour des auditeurs sans trouble visuel. Il en est de même pour Menard *et al.* [6] qui ont notamment mis en avant le fait que les personnes souffrant de cécité présentaient des scores de discrimination plus élevés lorsqu'il s'agissait de catégoriser des voyelles par rapport à des auditeurs de contrôle. Ainsi, il semblerait que les auditeurs non-voyants ou malvoyants développeraient des facultés compensatrices en matière d'audition.

L'*objectif* de notre étude sera donc de vérifier deux *hypothèses* : premièrement, nous pensons que les auditeurs non-voyants percevraient la voyelle protruse avant les auditeurs sans trouble visuel, et de manière plus certaine (selon un seuil de confiance), dans des séquences [isV^{lab}], cela grâce aux réajustements auditifs compensatoires liés à la perturbation du canal visuel ; Secondement, nous posons le fait que la perception de la voyelle serait moins précoce à mesure que cette dernière soit moins labialisée, et cela aussi bien pour les auditeurs non-voyants que voyants. Ainsi, l'élément vocalique [y], par exemple, serait discerné avant le [ø].

2. MÉTHODE

2.1. Corpus, enregistrement et mesures acoustiques

Afin de vérifier nos hypothèses, des séquences [isV^{lab}] insérées dans une phrase porteuse du type « C'est [isV^{lab}] ça » ont été enregistrées dans la chambre insonorisée de l'Institut de Phonétique de Strasbourg. Il est à noter que la voyelle [V^{lab}] était soit [y], [ø], [u] ou [o]. Chaque phrase a été prononcée à trois reprises par un locuteur. Le choix de ce corpus permettait de tester l'effet de l'aperture et de la labialité sur la perception anticipée de la voyelle, et de vérifier si les résultats observés sont les mêmes selon que les éléments vocaliques soient antérieurs ou postérieurs.

Une fois les données acquises, des mesures qualitatives et quantitatives ont été effectuées. La structure formantique

(F1, F2 et F3) a ainsi été mesurée au milieu du [i] et au milieu de la voyelle labialisée. Parallèlement à cela, la fréquence de la limite inférieure du bruit de friction du [s] a été relevée, à partir de coupes spectrales, et cela toutes les 10 ms avec, comme point de départ, le début de la structure formantique stable de la voyelle protruse. Notons que la fréquence du bruit de friction a tendance à diminuer à mesure que l'on se rapproche d'un élément labialisé (Hirsch, [3] ; Calliope, [7]).

En outre, la durée du [i], de [V^{lab}] et du [s] a été quantifiée. De même, nous indiquerons la durée de l'intervalle compris entre le début de l'inflexion du bruit de friction (si elle a lieu) et le début de la structure formantique stable de la voyelle labialisée en termes de valeurs absolues et relatives, cela afin d'évaluer l'effet de l'élasticité temporelle du signal acoustique.

2.2. Tests de perception et auditeurs

Afin d'étudier le moment où les voyelles labialisées commencent à être perceptibles, quatre phrases ont été sélectionnées parmi le corpus présenté, comportant chacune une voyelle labialisée différente. Le paradigme du *gating* ou du dévoilement progressif du signal, a été utilisé puisque les séquences étudiées ont été tronquées toutes les 10 ms en partant de la structure formantique stable de la voyelle labialisée, et en allant vers le [i]. Un montage a été réalisé pour chacune des phrases proposées avec, entre chaque stimulus, un bip prévenant de l'imminence d'une séquence placé 1,5 sec. avant toute séquence. Les séquences tronquées étaient présentées en ordre aléatoire.

20 auditeurs ont été recrutés pour écouter les stimuli : 10 personnes sans trouble visuel ni auditif, et 10 personnes déficientes visuelles (8 non-voyants et 2 mal-voyants). Etant donné que certains sujets non-voyants n'étaient pas en mesure de cocher une réponse ou de l'écrire, les auditeurs avaient tous pour consigne d'écouter, dans un premier temps la séquence tronquée, puis de donner oralement la voyelle suivant le [s] à l'expérimentateur qui prenait les notes. Les auditeurs avaient le choix entre trois possibilités : [e], [a] (retenues comme « distracteurs ») ou [V^{lab}]. En outre, il était demandé d'allouer une note de confiance (nc) à la réponse donnée, note qui pouvait aller, dans une échelle subjective, de 0 à 5, selon que la voyelle ait été donnée au hasard (0) ou avec une assurance totale (5).

3. RÉSULTATS

3.1. Observations acoustiques

Deux tendances se dégagent lorsque l'on observe la limite inférieure du bruit de friction : pour [y] et [u], la fréquence de la limite inférieure du bruit de friction du [s] a tendance à diminuer très légèrement dès la fin du [i], avant qu'une inflexion plus prononcée ait lieu jusqu'à l'arrivée de la structure formantique de la voyelle labialisée. Dans le cas de la séquence [isy], la limite inférieure du bruit de friction s'élevait à 3864 Hz au premier relevé situé à proximité du [i] avant d'être évalué à 3473 Hz, 90 ms avant le début acoustique du [y] ; à cette date, une forte inflexion est observée, puisque la limite inférieure du bruit de friction diminue fortement pour atteindre 1337 Hz, à 10 ms de l'émergence de la structure formantique stable du [y]. Notons que la date de l'inflexion correspond à 51 % de l'intervalle obstruent [s].

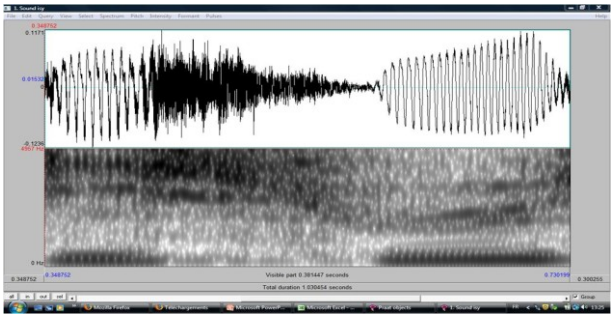


Figure 1 : Signal acoustique et spectrogramme d'une séquence [isy]. La fréquence de la limite inférieure du bruit de friction diminue en deux temps : une légère baisse d'abord avant qu'une inflexion brutale ne soit observée.

Notons que nous avons observé le même phénomène pour [isu]. En ce qui concerne les séquences comportant les voyelles [ø] et [o], la tendance est différente dans la mesure où aucune inflexion forte n'a pu être constatée. En effet, seule une légère baisse de la limite inférieure du bruit de friction est observée, celle-ci passant de 4027 Hz (à la date la plus proche du [i]) à 3277 Hz.

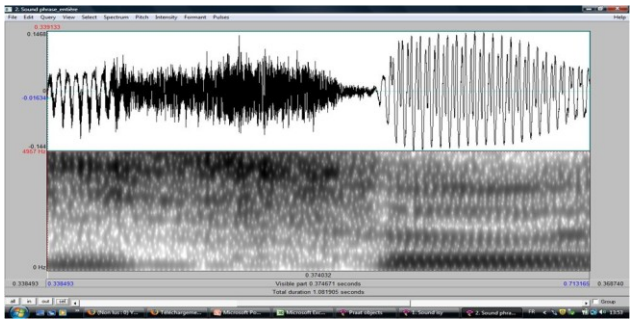


Figure 2 : Signal acoustique et spectrogramme d'une séquence [isø]. La fréquence de la limite inférieure du bruit de friction diminue légèrement mais aucune inflexion n'est visible.

3.2. Tests de perception : les résultats

Résultats pour les auditeurs de contrôle

Les résultats (Figure 3) révèlent que la voyelle [y] commence à être perçue 90 ms avant son début acoustique, et cela par 70% des sujets témoins (nc = 2). Il est intéressant de rappeler ici que le début de l'inflexion du bruit de friction du [s] a eu lieu 10 ms plus tôt. Notons aussi qu'à la date suivante, c'est-à-dire à 90 ms de la structure formantique stable de la voyelle arrondie, le pourcentage d'auditeurs qui ont donné une réponse correcte n'est plus que de 50% (nc = 2), ce qui signifie que le seuil du hasard a été franchi.

En outre, il a également été possible d'observer que le seuil de confiance a tendance à diminuer à mesure que l'on s'éloigne de la voyelle labialisée.

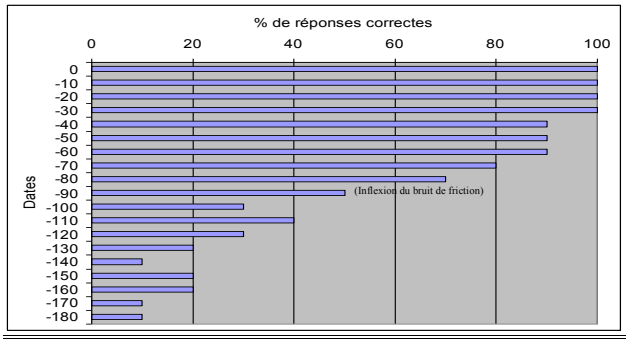


Figure 3 : Pourcentage de réponses correctes pour les auditeurs de contrôle (abscisse) en fonction des points de troncation (ordonnée). La voyelle [y] est perçue 80 ms avant son début acoustique.

Pour ce qui concerne les résultats obtenus pour la voyelle [ø] (Figure 4), ils ne permettent pas de parler de perception précoce, étant donné qu'elle n'a été identifiée, par 90 % des auditeurs, qu'à l'apparition de sa structure formantique clairement définie (nc = 3). 10 ms plus tard, le pourcentage de réponses correctes diminue déjà pour atteindre 50% (nc = 1).

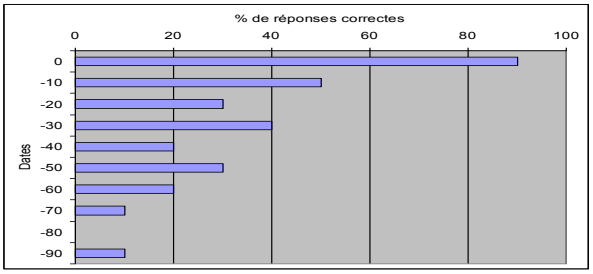


Figure 4 : Pourcentage de réponses correctes pour les auditeurs de contrôle (abscisse) en fonction des points de troncation (ordonnée). La voyelle [ø] n'est pas perçue avant son début acoustique.

Pour ce qui est de l'autre couple étudié, les séquences [isu] et [iso], le schéma reste le même que pour les voyelles antérieures, à savoir que la voyelle de petite aperture et très labialisée [u] est perçue de manière anticipée 80 ms avant son début acoustique (nc = 2), alors que le [o] est difficilement perceptible avant le début de sa structure formantique stable.

Comparaison des résultats pour les auditeurs de contrôle *vs.* les auditeurs déficients visuels

La comparaison des résultats des auditeurs de contrôle avec ceux des auditeurs ayant une déficience visuelle montre que les non-voyants et les mal-voyants tendent à percevoir la voyelle labialisée avant le groupe de contrôle. C'est le cas pour le [y] qui commence à être reconnu par 70 % des auditeurs atteints de déficience visuelle, à 110 ms de son début acoustique (nc = 3). 10 ms plus tard, la voyelle est encore perçue par 60% des auditeurs déficients visuels, score que nous estimons trop bas pour être robuste. Quant aux sujets de contrôle, rappelons qu'ils ne percevaient cette même voyelle que 80 ms avant son début acoustique.

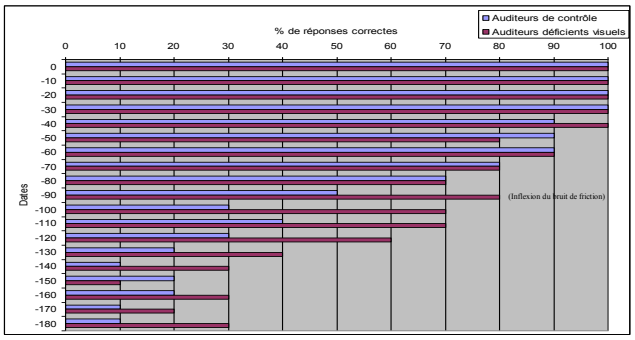


Figure 5 : Comparaison du pourcentage de réponses correctes (abscisse) chez les auditeurs de contrôle et les sujets déficients visuels en fonction des points de troncation (ordonnée) pour la voyelle [y]. Le groupe d'auditeurs non-voyants et mal-voyants perçoit la voyelle arrondie avant les auditeurs de contrôle.

Quant à la voyelle [ø], elle est identifiée par 70% des auditeurs 10 ms avant son début acoustique (nc = 3). Cela dit, nous notons une légère baisse du pourcentage de réponses correctes à 20 ms du [ø], étant donné que seuls 60% des sujets présentant un trouble visuel sévère ont identifié la voyelle correctement (nc = 3). A 30 ms du début de la voyelle labialisée, le nombre d'auditeurs qui ont indiqué une réponse juste remonte légèrement pour atteindre à nouveau 70% (nc = 2). Par la suite, le pourcentage de réponses correctes diminue pour atteindre des scores non-significatifs (50% des auditeurs déficients visuels ont identifié la voyelle à 40 ms du [ø], puis 30 % à la date suivante).

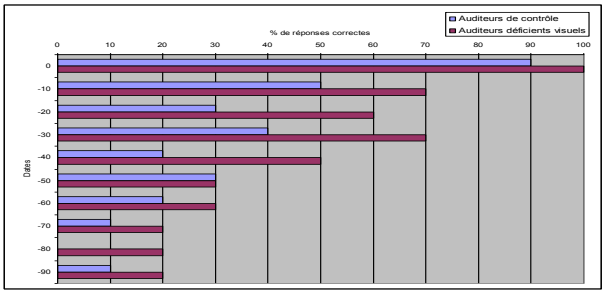


Figure 6 : Comparaison du pourcentage de réponses correctes (abscisse) chez les auditeurs de contrôle et les sujets déficients visuels en fonction des points de troncation (ordonnée) pour la voyelle [ø]. Le groupe d'auditeurs non-voyants et malvoyants perçoit la voyelle arrondie avant les auditeurs de contrôle.

Les mêmes tendances ont pu être observées pour les voyelles postérieures du corpus puisque le [u] est identifié plus tôt par les auditeurs non-voyants et malvoyants (à 100 ms avant le début acoustique de la voyelle *vs.* à 80 ms pour les auditeurs de contrôle), tout comme le [o] qui est reconnu 20 ms (nc = 2) avant l'apparition de la structure formantique, par le groupe présentant des

troubles sévères de la vue, alors que les sujets témoins n'avaient pas réussi à identifier cette même voyelle.

Quant aux notes de confiance, elles sont généralement plus élevées chez les auditeurs déficients visuels (2,8 en moyenne) qui semblent, de ce fait, plus sûrs de leur réponse que le groupe de contrôle (2,2 en moyenne). De manière générale, tous les sujets ont été sévères dans l'attribution de leurs notes de confiance, puisque les valeurs moyennes restent relativement bas, même pour des scores d'identification correcte élevés.

4. SYNTHÈSE ET DISCUSSION

Au commencement de ce travail, nous avions formulé deux *hypothèses* : nous pensions tout d'abord que la perception de voyelles labialisées serait moins précoce à mesure que celles-là soient moins protruse. Deuxièmement, nous supposions que des auditeurs non-voyants percevraient la voyelle protruse avant les auditeurs sans trouble visuel.

L'étude acoustique et perceptive des séquences [isV^{lab}], menée auprès des *sujets témoins*, révèle deux phénomènes, l'un concernant les voyelles très labialisées et de petite aperture, l'autre les voyelles moins labialisées de moyenne aperture. Pour les premières citées ([y] et [u]), la fréquence de la limite inférieure du [s] diminue d'abord légèrement, et cela dès la fin de la voyelle [i], avant qu'une inflexion forte ait lieu. La date à laquelle se trouve l'inflexion correspond, plus ou moins, au moment où les auditeurs de contrôle commencent à percevoir la voyelle labialisée. Notons que ce résultat peut être mis en parallèle avec ceux obtenus par Vaxelaire *et al.* [8] qui avaient remarqué que, pour une séquence comportant une fricative suivie d'un [y], le moment où les auditeurs perçoivent précocement la voyelle labialisée correspond, au niveau acoustique, au début de l'inflexion du bruit de friction qui coïncide, lui-même, avec le pic d'accélération du mouvement labial au niveau articulatoire. Par conséquent, les sujets témoins sembleraient sensibles à certains événements cinématiques qui leur permettraient d'identifier auditivement une voyelle avant son émergence acoustique. Pour le second groupe de voyelles ([ø] et [o]), la limite inférieure du bruit de friction diminue, de manière plus ou moins régulière, de la fin du [i] à la voyelle [V^{lab}], mais aucune inflexion n'est visible. Cela laisse supposer qu'une augmentation de l'aperture provoquerait une diminution de la protrusion, ce qui entrainerait une modification des paramètres cinématiques. En d'autres termes, étant donné que la protrusion est moins prononcée, il est envisageable que le pic d'accélération intervienne plus tardivement, ce qui expliquerait pourquoi il ne provoquerait pas d'inflexion au niveau de la limite inférieure du bruit de friction. En ce qui concerne les auditeurs *déficients visuels*, nous avons pu observer les mêmes tendances, les voyelles très labialisées étant perçues bien avant les voyelles labialisées de moyenne aperture. En comparant les résultats obtenus pour ces derniers avec ceux des auditeurs de contrôle, il a été constaté que le groupe atteint d'un trouble sévère de la vue perçoit plus tôt les éléments vocaliques labialisés par rapport aux sujets témoins. Ce résultat pourrait s'expliquer par le fait que les auditeurs déficients visuels développeraient une sensibilité auditive plus prononcée leur permettant de percevoir des changements plus fins au niveau de la limite inférieure du bruit de friction.

En perspective à cette étude, il semblerait intéressant de continuer cette recherche en recrutant davantage d'auditeurs dans chaque groupe. En outre, nous souhaiterions vérifier si les auditeurs témoins seraient capables de rattraper leur « retard » de perception lorsqu'on leur présente des informations audio-visuels ou lorsqu'on leur masque la vue. Dans tous les cas, nos données acoustiques et auditives sont à rapprocher des celles connexes, obtenues dans les domaines articulatoire ([3] ; [8]) et cinématique ([4]), cela pour pouvoir déceler les relations sensori-motrices impliquées dans la perception précoce de faits anticipatoires.

Remerciements : Programme MISHA (Maison Interuniversitaire des Sciences de l'Homme d'Alsace) « Perturbations et réajustements : parole normale *vs.* parole pathologique » 2008-2012 ; ANR-07-CORP-018-01, DOCVACIM, 2007-2011» ; Association des Aveugles d'Alsace et de Lorraine (AAAL) ; Coralie Vincent.

BIBLIOGRAPHIE

[1] C. Abry and T. Lallouache. Le MEM : un modèle d'anticipation paramétrable par locuteur : données sur l'arrondissement en français. *Bulletin de la communication parlée,* volume 3, pages 85-89, 1995.

[2] V. Ferbach-Hecker. *La perception auditive de l'anticipation des gestes vocaliques en français*. Thèse de doctorat Nouveau Régime soutenue à l'Université de Strasbourg, 2002.

[3] F. Hirsch, R. Sock, P.-Y. Connan and G. Brock. Auditory effects of anticipatory rounding in relation with vowel height in French. In *Proceedings of the International Phonetic Sciences*, pages 1445-1448, 2003.

[4] J.P. Roy. Visual perception of anticipatory rounding gestures in French. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 2949-2952, 2005.

[5] A. Moos, I. Hertrich, S. Dietrich, J. Trouvain and H. H. Ackermann. Perception of ultra-fast speech by a blind listener - Does he use his visual system?. In *Proceedings of the International Speech Production Seminar, Strasbourg*, pages 297-300, 2008.

[6] L. Menard, S. Dupont, S.R. Baum and J. Aubin. Production and perception of French vowels, by congenitally blind adults and sighted adults. In *Journal of acoustical society of America*, volume 126 (3), pages 1406-1414, 2009.

[7] Calliope. *La parole et son traitement automatique.* Editions Masson, 1992.

[8] B. Vaxelaire, R. Sock, F. Hirsch, V. Ferbach-Hecker J.P. Roy and F. Bouarourou. *The Anticipatory Perception Based on Events (APE) Hypothesis. C*ommunication présentée aux Vè Convegno Nazionale AISV, Zurich, 2009.

L'effet sur la parole du locuteur de sa représentation du statut linguistique de son interlocuteur. Un essai exploratoire de caractérisation phonique.

Florence Verbanck, Myriam Piccaluga, Bernard Harmegnies

Laboratoire des sciences de la parole de l'Académie Universitaire Wallonie-Bruxelles
UMONS, 18 place du parc, B-7000 Belgique
bernard.harmegnies@umons.ac.be

http://w3.umh.ac.be/~compa/

ABSTRACT

French Foreigner-directed Speech is studied for the first time, using a totally new material. Results show that speakers modify their output when modifying their speech. Results also suggest that transformations of the pitch would appear when representation of the listener is changing. Utilisations of µ-index will be discussed.
Keywords: Foreigner-directed Speech, Representation, Pitch, µ-index

1. INTRODUCTION

Partant du postulat que la plupart des locuteurs adaptent leur discours en fonction de la représentation qu'ils se font de leur interlocuteur (personne ou dispositif non humain) à qui ils s'adressent [1], nombre de recherches se sont penchées sur les modifications du discours dans différents contextes langagiers. Dans cet ordre d'idées, le discours adressé aux non-natifs, souvent dénommé *Foreigner-Talk* ou *Foreigner-directed speech*, est un champ de recherches encore jeune et très peu développé. En effet, depuis le premier article de Ferguson en 1971[2], le terme ne revient que de façon sporadique dans les bases de données. Le concept, d'abord utilisé par les sociologues et les sociolinguistes dans le cadre de recherches sur les phénomènes de créolisation, a peu à peu investi le domaine d'investigation de la didactique des langues. Cependant, souvent associé à d'autres phénomènes d'adaptation du locuteur à son interlocuteur, il n'a été traité, le plus souvent, qu'en comparaison avec le discours adressé aux jeunes enfants (Freed [3], Knoll [4]).

Il n'y a donc rien d'étonnant à ce que la notion demeure mal définie dans la littérature. Ainsi, le dépouillement de la cinquantaine d'articles peu ou prou liés au sujet révèle qu'une dizaine de termes sont utilisés, en anglais, pour référer au concept. Les définitions disponibles sont, quant à elles, au moins aussi nombreuses. Par ailleurs, notre analyse de la littérature n'a porté à notre connaissance aucune étude basée sur des sujets s'exprimant en français. En outre, dans toutes les expériences qui ont été réalisées sur le sujet, le passage entre le discours adressé à un natif et le discours adressé à un non-natif se fait abruptement ; il est donc difficile de déterminer finement quels sont les éléments impliqués dans la dynamique de la modification. De même, il est difficile d'identifier les facteurs déterminants du changement. Enfin, la plupart des études reposent sur les perceptions, toujours subjectives, de quelques juges. La présente étude exploratoire, qui vise à tester les principes d'un dispositif expérimental et la sensibilité de diverses variables phonétiques, se veut une contribution au développement, dans ce domaine, d'études objectivées reposant sur des variables acoustiques et se centrant sur des productions vocales en langue française.

2. DISPOSITIF

2.1. Les sujets

Quinze sujets, 7 femmes et 8 hommes, ont été enregistrés aux fins de l'expérience. Tous francophones natifs, ils disposaient également de connaissances de base en langue anglaise.

2.2. La tâche

La tâche du sujet consistait à faire reconstituer à l'identique par une autre personne une grille où étaient disposées des images. L'interlocuteur (un comparse) pouvait entendre et voir le locuteur, alors que ce dernier ne pouvait que le voir à travers une vitre isolante, observer sur un écran de contrôle le résultat de ses manipulations du dispositif, mais en aucun cas l'entendre. Le sujet ne pouvait communiquer avec l'interlocuteur que par injonctions verbales. Au départ, l'interlocuteur suivait à la lettre les injonctions du sujet, mais petit à petit, son efficacité s'émoussait, et il paraissait ne plus aussi bien comprendre les indications ; les erreurs commises étaient préméditées et conçues de sorte à conduire le sujet à se représenter l'interlocuteur comme non-francophone et, en tout état de cause, susceptible de commettre des confusions avec l'anglais. Par exemple, lorsqu'il fallait replacer l'image d'une glace, au lieu de l'aliment, un verre (*glass* en anglais) apparaissait sur la grille; de même, alors que le déplacement d'une photo de plume était attendu par le sujet, une image de prune (*plum* en anglais) était déplacée par l'interlocuteur.

2.3. Construction de la grille

Nous avons utilisé quatre niveaux de proximité entre termes français et termes anglais : Reconnaissance immédiate ($Igloo_{fr}$ *vs.* $Igloo_{en}$) ; Reconnaissance après plusieurs répétitions ($Vampire_{fr}$ *vs.* $Vampire_{en}$) ; Mauvaise

reconnaissance (*Glace*$_{fr}$ *vs. Glass*$_{en}$) ; Aucune reconnaissance possible (*Cuillère*$_{fr}$ *vs. Spoon*$_{en}$). Enfin, l'ordre de présentation des images avait été pensé de façon à ce que des mots de la troisième catégorie soient présents très rapidement dans l'expérience et permettent au sujet de s'orienter dans la voie escomptée.

2.4. Prises de mesure

Tous les enregistrements ont été effectués dans la chambre anéchoïque du Laboratoire de Phonétique de l'Université de Mons, au moyen d'un microphone Neumann U87P48, connecté à un codeur PCM 501ES Sony, dont les signaux de sortie étaient stockés sur un magnétoscope VHS Panasonic. Les analyses acoustiques ont été effectuées au moyen du logiciel Praat, sur base de cinq critères acoustiques: fréquence fondamentale en début, milieu et fin de syllabe et fréquences des premier et deuxième formants en milieu de syllabe.

3. RESULTATS

3.1. La représentation de l'interlocuteur

Chaque sujet a été mis en contact deux fois avec le dispositif. D'abord, au cours de la session pré-expérimentale, les locuteurs ont été confrontés à une première effectuation de la tâche comme indiqué ci-dessus (Cf. 2.2). Le but poursuivi était d'amener le sujet à se former insensiblement, sur la seule base des manipulations du dispositif par l'interlocuteur, une représentation de l'identité de celui-ci, éventuellement susceptible d'intégrer la dimension allophone. Pour la phase expérimentale, les sujets ont, dans un second temps, à nouveau été confrontés au même dispositif, mais cette fois en étant avertis que le temps mis par l'interlocuteur pour effectuer la tâche serait pris en considération dans son évaluation. A la faveur d'une enquête rétrospective pratiquée immédiatement après l'effectuation de la tâche expérimentale, les sujets ont indiqué s'ils avaient ou non eu l'impression de se trouver face à un allophone, et si oui, à partir de quel moment. La figure 1, qui résume ces informations, donne un aperçu de la dynamique du basculement de la représentation, dont on observe que, si elle intervient, elle se fait jour endéans les deux premiers tiers de la tâche, mais à des moments très variables de sujet à sujet. L'enquête a par ailleurs montré que les sujets ont maintenu la même représentation entre la fin de la phase pré-expérimentale et la phase expérimentale, et ce, que la représentation de l'interlocuteur soit celle d'un francophone ou d'un allophone; seule exception : le sujet 5, qui témoigne d'un basculement de représentation entre les deux phases.

Pour la suite de notre analyse, nous prendrons en considération deux groupes de sujets : d'une part, ceux qui sont entrés dans la phase expérimentale avec le sentiment de ne pas avoir affaire à un interlocuteur francophone (représentation allophone : sujets 1, 4, 5, 6, 8, 9, 10, 11, 12, 14, 15) et d'autre part, ceux qui, au contraire, n'ont jamais pensé se trouver face à un interlocuteur allophone (représentation francophone : sujets 2, 3, 7, 13).

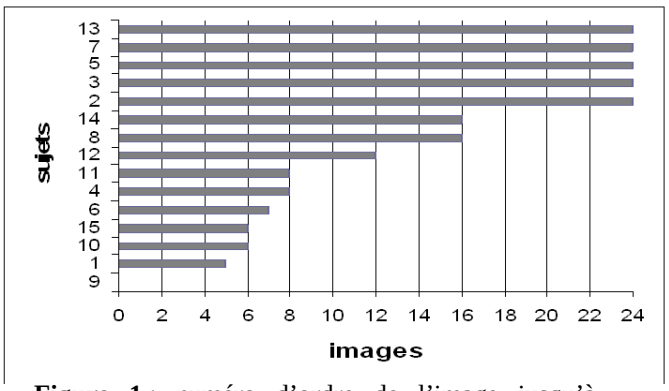


Figure 1 : numéro d'ordre de l'image jusqu'à laquelle les sujets ont pensé avoir affaire à un interlocuteur francophone

3.2. Une approche segmentale

Nous tentons ici de déterminer si les structures d'ensemble de l'espace vocalique ont été affectées par la représentation qu'avait le locuteur de l'identité de son interlocuteur. Plusieurs recherches ont en effet montré la sensibilité de l'organisation du système vocalique à diverses sources de variation, sur la base de la mise en œuvre d'outils métrologiques permettant une grande finesse d'analyse [5, 6]. La plupart du temps, cependant, ces démarches requièrent que les réalisations dans un type de condition puissent être pairées avec celles obtenues dans l'autre, les sujets étant placés tour à tour dans les diverses conditions à comparer (contrôle de la variabilité inter-sujets) et/ou les voyelles étant rendues disponibles en environnement similaire dans les diverses conditions (contrôle des effets coarticulatoires). Or, les groupes de réalisations que nous étudions ici sont par définition des groupes inter-sujets, et les productions recueillies au moyen de notre dispositif ne permettent pas le pairage inter-conditions d'un nombre de voyelles suffisant.

Nous avons dès lors décidé de recourir à une évaluation globale de l'extension du plan vocalique, pour chacune des deux conditions. A cet effet, nous avons sélectionné uniquement l'ensemble des voyelles [i], [a] et [u], qui peuvent être considérées comme les trois sommets du triangle vocalique. Les F_1 et F_2 des 514 voyelles ainsi isolées ont été évaluées sur la base de sonagrammes obtenus via Praat. Comme le montre la figure 2 (haut, gauche), le triangle vocalique obtenu pour l'ensemble des points-voyelles correspondant à chacun des groupes de sujets est légèrement plus étendu pour les sujets considérant leur interlocuteur comme non francophone. Afin d'obtenir une évaluation moins impressionniste, nous avons par ailleurs recouru à la technique introduite par Hirsch [7], consistant à déterminer la surface de chacun des triangles vocaliques. Nous avons, à cet effet, pour chacun des sujets, recouru à la formule ci-dessous (avec ***a***, l'aire et F, la fréquence formantique exprimée en Hertz).

$$a = \frac{1}{2}\left|\left(F_{2[u]} - F_{2[i]}\right)\left(F_{1[a]} - F_{1[i]}\right) - \left(F_{2[a]} - F_{2[i]}\right)\left(F_{1[u]} - F_{1[i]}\right)\right|$$

Comme l'indique la figure 2 (bas, gauche), ces calculs confirment l'observation qualitative, en montrant que les sujets considérant leur interlocuteur comme francophone tendent à se caractériser par des surfaces moins importantes (en moyenne, 86% de celles des autres locuteurs). Cette observation descriptive ne s'assortit cependant pas d'une confirmation inférentielle ($p > .50$ pour les tests de Wald-Wolfowitz, U de Mann-Whitney et W de Wilcoxon), ce qui n'est pas très étonnant vu la faible puissance des tests non-paramétriques utilisés et la disparité des effectifs d'ailleurs faibles (4 contre 11). Mais ce raisonnement ne doit pas scotomiser l'importante variabilité inter-sujets qui prévaut dans chacun des deux groupes. Un simple coup d'œil à la partie droite de la même figure, qui présente, pour chacun des deux groupes, les points moyennes pour chacune des voyelles et chacun des sujets, montre d'ailleurs combien la structure du plan F_1/F_2 est dépendante du sujet.

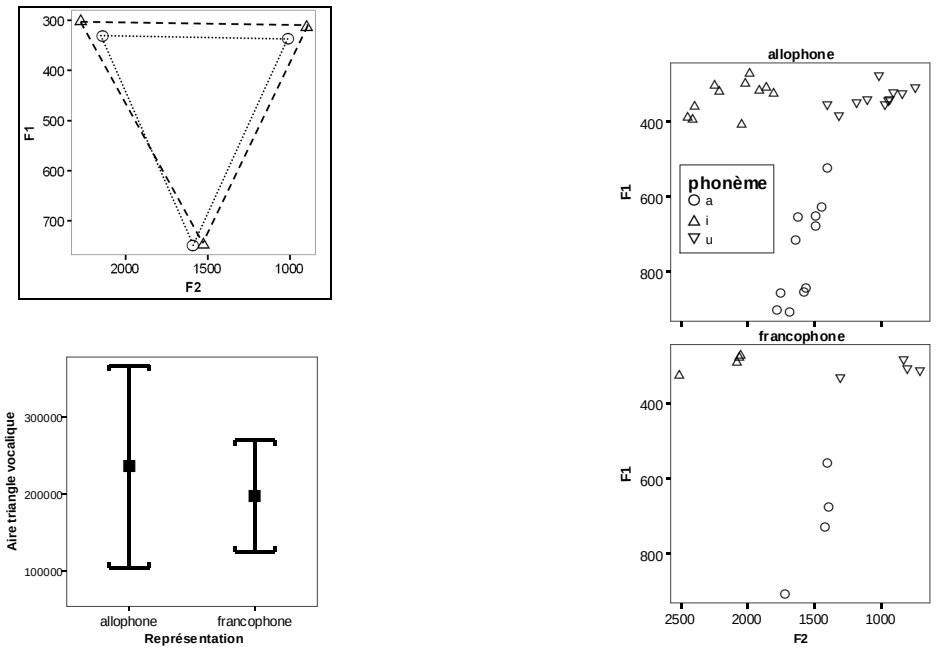


Figure 2 : (haut à gauche) triangles vocaliques, tous sujets confondus, pour les groupes à représentation allophone (pointillés) et francophone (tirets) ; (bas à gauche) moyennes et écarts types des aires des triangles vocaliques pour chaque sujet des 2 groupes ; (droite) dispersions des moyennes par sujet pour chaque voyelle, dans chaque groupe.

3.3. Une approche supra-segmentale

Des 3 corrélats acoustiques de la prosodie, seuls 2 nous sont accessibles (durée et mélodie), car les variations observées de l'intensité ne sont pas valides, la distance micro-lèvres n'étant pas contrôlée dans notre dispositif.

Les durées de toutes les voyelles prononcées par les 15 sujets ont été manuellement mesurées (2659 observations). Leur comparaison inter groupes ne révèle aucune différence significative (analyse de variance : $F = 1{,}523$, $dl=1$, $p = .217$; test U de Mann-Whitney : $z = 1.025$, $p = .305$).

Les fréquences fondamentales au centre de chacune des voyelles ont été estimées au moyen de l'algorithme de détection de pitch de Praat. Dans les cas douteux, une confirmation a été obtenue via le recours à un sonagramme en narrow. Pour la comparaison des informations relatives au pitch, nous devons tenir compte du fait que des locuteurs différents présentent naturellement des valeurs de f_0 moyen variables, eu égard à leur conformation physique et à leurs habitudes de production ; d'autre part, il convient d'être attentif au fait qu'un accroissement donné, en Hertz, peut correspondre, en fonction de la fréquence initiale, à un accroissement d'ampleur variable sur le plan mélodique. A la suite de Piérart et Harmegnies [8], nous avons dès lors, selon la formule ci-dessous, proposée par Zwicker et Feldtkeller [9], transformé les fréquences fondamentales exprimées en Hertz (*f*) en hauteurs harmoniques (Ha).

$$Ha = \frac{1}{Log(2)} * Log(\frac{f}{131})$$

Pour les raisons invoquées plus haut, une comparaison des valeurs de f_0 entre les deux groupes n'aurait évidemment aucun sens. Par contre, il peut être intéressant de noter que la variabilité de la f_0 mesurée au centre de la voyelle varie de sujet à sujet, comme l'indique la figure 3. La figure suggère en outre que la variabilité de la variance est supérieure dans le groupe à représentation allophone. Un test F de Snedecor, avec au numérateur la somme des carrés des écarts intra sujets du groupe à représentation allophone (rapportée au nombre d'observations diminué du nombre de sujets dans le groupe) et au dénominateur la

somme des carrés des écarts intra sujets du groupe à représentation francophone (rapportée au nombre d'observations diminué du nombre de sujets dans le groupe) confirme la significativité de la différence ($F = 1{,}347$, $dl_1 = 717$, $dl_2 = 351$, $p < .001$).

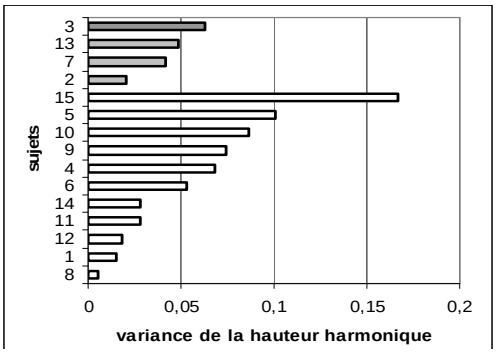


Figure 3 : variance de la hauteur harmonique en fonction du sujet, pour les sujets à représentation allophone (barres creuses) et à représentation francophone (barres pleines)

Enfin, nous avons investigué la variation micro-mélodique en évaluant, pour chaque voyelle prise en considération, outre sa fréquence fondamentale au centre (f_c), ses fréquences fondamentales en ses début (f_i) et fin (f_f) et en définissant, au départ de ces fréquences transformées en hauteurs harmoniques, un indice (μ_{Ha}) inspiré de l'indice μ [10], dont les valeurs sont d'autant plus supérieures à 1 que la fréquence centrale est élevée par rapport aux fréquences initiale et finale.

$$\mu_{Ha} = Log_2 \frac{f_c}{131} - \frac{1}{2}\left(Log_2 \frac{f_i}{131} + Log_2 \frac{f_f}{131} \right)$$

Il apparaît que l'indice μ_{Ha} est significativement plus grand (t de Student : $t = 4{,}438$, $dl = 1079$, $p < .001$; U de Mann-Witney : $z = 5.755$, $p < .001$) dans le groupe à représentation francophone que dans le groupe à représentation allophone.

4. CONCLUSIONS

Par une action sur le seul résultat comportemental, dans le chef d'un interlocuteur silencieux, des injonctions du locuteur, notre expérience nous a permis d'éprouver un dispositif susceptible de modifier, en la contrôlant, la représentation qu'a ce dernier de l'identité linguistique de la personne à qui il s'adresse. Cette modification, validée par enquête à posteriori, s'est avérée influer sur certains aspects acoustiques de la parole du locuteur (tendance au changement en termes d'hyper-hypo-articulation, suggérée par les rapports des triangles vocaliques, variation significative de la monotonie du pitch, confirmée par les différences de variances fréquentielles, variations micro-mélodiques significatives révélées par l'indice μ_{Ha}). Ces constats - à ce jour, originaux - engagent à développer la recherche centrée sur cette forme particulière d'adaptation du locuteur francophone à son interlocuteur, surtout devant les exigences actuelles de la mondialisation des échanges qui trop souvent laisse de côté le français. Plus profondément, la variabilité du temps mis par les sujets pour que bascule leur représentation de l'interlocuteur, de même que la très importante variabilité inter-individuelle des espaces formantiques, pose la question de la diversité non seulement de la sensibilité du locuteur aux caractéristiques de l'autre mais aussi de la nature des processus de faire-face convoqués. Les développements expérimentaux à venir devront viser, en conséquence, à mieux contrôler le glissement d'un état de représentation à un autre et aussi à affiner l'appréhension des variations intra-individuelles dans tous leurs aspects.

5. BIBLIOGRAPHIE

[1] B. Lindblom. Explaining Phonetic Variation : A sketch of the H&H Theory. In *Speech Production and Speech Modelling*, Kluwer Academic Publishers, 1989.

[2] C. A. Ferguson. Absence of copula and the notion of simplicity: A study of normal speech, baby talk, foreigner talk, and pidgins. In D. Hymes, *Pidginization and Creolization of Languages*. Cambridge University Press, Cambridge, UK, 1971.

[3] B. F. Freed.. Foreigner Talk, Baby Talk, Native Talk. *Int. J. of the Sociology of Language*, 28:19-39, 1981.

[4] M. Knoll, L. Scharrer, A. Costall. Are actresses better simulators than female students? The effects of simulation on prosodic modifications of infant- and foreigner-directed speech. *Speech Communication*, 51 (3):296-305, 2009.

[5] D. Poch-Olivé, B. Harmegnies. Vowel reduction in spontaneous speech in Spanish. In *Proc. of the ESCA Workshop on Phonetics and Phonology of Speaking Style*. Barcelona, 311-315, 1991.

[6] K. Huet, B. Harmegnies, D. Poch-Olivé. Une méthode statistique pour le contrôle des changements vocaliques sous l'effet du style de parole. Application à l'espagnol. In M. Matthey (éd.). *Le Changement linguistique, Evolution, variation, hétérogénéité*. TRANEL, 34/35:233-249, 2001.

[7] F. Hirsch. *Le bégaiement. Perturbation de l'organisation temporelle de la parole et conséquences spectrales*. Thèse doctorale, Université Marc Bloch, Strasbourg 2, 2007.

[8] B. Piérart, B. Harmegnies. Dysphasie simple de l'enfant et langage de la mère. *L'année psychologique*, 93:227-268, 1993.

[9] E. Zwicker, R. Feldtkeller. *Psychoacoustique; l'oreille, récepteur d'information*. Masson, Paris, 1981.

[10] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, D. Poch-Olive, Time- and Spectrum related variabilities in stressed speech under laboratory and real conditions. *Speech Communication*, 20 (1-2):111-129, 1996.

COREIL, un corpus pour l'étude de l'acquisition de la prosodie en Français et Anglais Langue Etrangère

Elisabeth DELAIS-ROUSSARIE et Hi-Yon YOO

CNRS - UMR 7110 / Laboratoire de Linguistique Formelle
Université Paris-Diderot
elisabeth.roussarie@wanadoo.fr
yoo@linguist.jussieu.fr

ABSTRACT

In the last decade, many research projects aimed at developing language corpora specifically designed to study L2 acquisition process (see among others [1] & [2]). In most of the cases, however, these studies focused on the acquisition of morpho-syntactic, lexical and pragmatic competence through the observation of learners' written productions. In this paper, we present a corpus that has been specifically designed to collect oral productions from learners of French and English as a foreign language. The data collection protocol has been thought in order i) to carry research on the acquisition of suprasegmental phenomena, ii) to compare the acquisition processes along several dimensions (L1 vs. L2, difference among the learners L1, etc.).

Keywords: oral corpus, L2 acquisition, prosody, interlanguage analysis.

1. INTRODUCTION

Depuis quelques années, de nombreuses recherches sur l'acquisition des langues étrangères se sont basées sur l'étude de larges corpus (cf., entre autres, [1], [2] ou [3]). Cela a permis de mieux évaluer les relations entre la langue première de l'apprenant, son niveau en langue étrangère (établi d'après le cadre commun de référence pour les langues ou CECR) et sa compétence grammaticale en langue étrangère. Ainsi, dans un projet comme English Profile (cf. [3]), les chercheurs ont essayé de déterminer, à partir d'études sur corpus, comment s'effectue la maîtrise de plusieurs faits morpho-syntaxiques de l'anglais. Mais, force est de constater que ces études se sont centrées sur l'acquisition des compétences syntaxiques et morphologiques, plutôt que phonologiques. Cela peut s'expliquer par la difficulté d'obtenir des corpus oraux permettant de mener de telles recherches (mentionnons cependant [4] et [5]). Afin de palier ce manque, nous avons élaboré un protocole et développé le corpus COREIL, un corpus d'apprenants conçu pour l'étude de l'acquisition des phénomènes suprasegmentaux en langue étrangère. L'objectif de ce papier est d'en présenter les différentes caractéristiques.

Dans la première partie, nous expliquerons quels présupposés théoriques ont été retenus pour constituer ce corpus. Dans une seconde partie, le protocole d'enregistrement sera décrit. Une attention particulière sera accordée à la sélection des locuteurs, à la construction des tâches demandées (lecture, mini-dialogues, etc.) et aux modalités d'utilisation du protocole. Pour finir, nous présenterons certains principes retenus pour la transcription et l'annotation des données collectées.

2. CADRE GENERAL ET PRESUPPOSES

2.1. Hypothèses sur l'acquisition de la phonologie en langue étrangère

Nombre de travaux consacrés à l'acquisition de la phonologie en langue étrangère reposent essentiellement sur l'idée qu'il existe un transfert important entre la langue première des apprenants et la langue cible. Des notions comme le crible phonologique ou la surdité phonologique en sont la preuve, tout comme, d'ailleurs, certaines pratiques didactiques reposant sur l'analyse contrastive. Des études sur corpus auraient le mérite de valider ou non ces hypothèses, d'autant que certains travaux ont montré que l'ordre d'acquisition de quelques phénomènes morpho-syntaxiques (gestion des temps verbaux, détermination nominale, etc.) se fait selon un ordre similaire pour les enfants en langue première et pour les apprenants en langue étrangère, et cela quelle que soit leur langue première. L'accès à des données rigoureusement collectées est un préalable à l'étude de l'acquisition de la prosodie en langue étrangère. Aussi le protocole de collecte de données de COREIL a été construit en suivant cette idée.

Jusqu'à maintenant, les travaux consacrés à l'acquisition de la phonologie en langue étrangère se sont surtout centrés sur l'acquisition des phénomènes segmentaux (cf. [6]). Mais des études et travaux sont nécessaires pour répondre à plusieurs types de questions concernant la prosodie :

- Existe-t-il une différence dans l'acquisition des phénomènes prosodiques entre la langue étrangère et la langue première ?
- Est-ce que le transfert joue un rôle fondamental dans l'acquisition de ces phénomènes ?
- Les réponses apportées aux questions précédentes sont-elles valables pour tous les domaines d'étude de la prosodie (accentuation, intonation, phrasing, rythme, etc.) ?
- Y a-t-il des différences entre l'acquisition des faits segmentaux et celle des faits suprasegmentaux ?

- Pour les phénomènes prosodiques comme le rythme – phénomènes qui impliquent des compétences phonologiques autant que phonétiques – y a-t-il un ordre strict dans le processus d'acquisition ? L'acquisition des faits phonétiques doit-elle devancer celle des faits phonologiques ou inversement ?

Le corpus COREIL a été conçu afin de répondre à ces questions, en évitant tout présupposé sur le poids du transfert dans l'acquisition de la phonologie en langue étrangère. De plus, le protocole d'enregistrement présenté dans la section 3 a été pensé afin de permettre des comparaisons selon plusieurs dimensions : entre apprenants de différentes langues premières, entre la langue première et la langue étrangère, etc.

2.2. Hypothèses et collecte des données

En plus des questions théoriques relatives à l'acquisition des langues étrangères, le corpus COREIL a été construit en suivant la méthodologie *AGILE* inspirée du développement logiciel (cf. [7]). D'une façon générale, la constitution d'un corpus est une activité longue et fastidieuse, puisqu'elle implique la collecte des données, leur mise en forme, leur annotation. L'étude ne peut intervenir qu'une fois ces tâches préparatoires accomplies. Mais, très vite, à l'épreuve des faits, il n'est pas rare que de nouvelles questions surgissent :

- les données collectées sont-elles assez représentatives pour l'objet d'étude envisagé ?
- des données complémentaires ne permettraient-elles pas de mieux répondre aux questions posées ?
- les annotations utilisés sont-elles satisfaisantes d'un point de vue théorique (peu de présupposés, aisance des requêtes, bonne adéquation entre transcripteurs, etc.) ?

Lorsqu'un chercheur a élaboré un corpus et mené un travail de recherche dessus, il n'est pas rare de l'entendre dire « *si c'était à refaire, je procéderais différemment* ». L'approche défendue par [7] est la suivante : la constitution et l'annotation des données doivent se faire parallèlement à leur étude, quitte à modifier le protocole et les schémas d'annotation dès que d'autres pistes de recherche voient le jour.

Pour construire le corpus COREIL, une telle approche est également retenue. Selon nous, elle offre de nombreux avantages. D'une part, elle permet de travailler sur certaines données du corpus, même si le processus de collecte des données n'est pas achevé. D'autre part, des données ou tâches supplémentaires peuvent aisément être intégrées au protocole et aux schémas d'annotation sans que l'ensemble soit à remettre en cause. La description détaillée des protocoles d'enregistrement et d'annotation va le montrer clairement.

3. PROTOCOLE D'ENREGISTREMENT

Le protocole d'enregistrement a été conçu de façon très modulaire, cela pour plusieurs raisons : i) les données sont produites par des locuteurs d'âge, de niveau et de langue première différents ; ii) l'acquisition peut être étudiée de façon longitudinale ou comparative ; iii) les résultats obtenus peuvent conduire à collecter de nouvelles données, à réviser les schémas d'annotation, etc.

3.1. La sélection des locuteurs

Le corpus COREIL regroupe à la fois des productions d'apprenants en langue étrangère, des productions d'enfants de moins de sept ans et d'adulte en langue cible (leur langue première). Pour étudier le poids du transfert, et également la différence entre l'acquisition de la phonologie (et plus précisément de la prosodie) en langue première et en langue étrangère, il est fondamental d'avoir des données de langue première produites par des natifs adultes ou enfants. Pour le moment, les données en langue première sont essentiellement limitées à des productions en français. Sont actuellement en cours d'enregistrement des productions d'une vingtaine d'enfants monolingues dont l'âge varie entre deux et cinq ans. Pour les adultes, une dizaine de personnes est enregistrée, mais, pour certaines tâches, des comparaisons peuvent être faites avec des productions extraites de corpus existants (PFC ou EUROM 1 [8])

Pour les apprenants de langue étrangère, deux paramètres ont été contrôlés, notamment pour le FLE : leur langue première et leur niveau de compétence linguistique.

Langue première des apprenants

Pour le FLE, les locuteurs retenus pour participer aux séances d'enregistrement ont comme langue première une des langues qui suit : anglais, espagnol, chinois et arabe. Pour sélectionner les apprenants d'anglais langue étrangère (ALE), aucune contrainte sur la langue première n'a été posée. Cela s'explique par le fait que la majorité d'entre eux a le français comme langue première, puisque les enregistrements sont effectués dans des institutions françaises (collèges, lycées, écoles, etc.).

Parmi les langues premières retenues, aucune restriction n'est faite sur les variétés parlées même si de grosses différences existent. Dans un premier stade, les variétés sont traitées indifféremment, mais elles sont notées dans le profil des apprenants. Si les études montrent qu'elles doivent être prises en compte puisqu'elles ont une influence sur l'acquisition de la prosodie en langue étrangère, elles le seront.

Age des apprenants de langue étrangère

Pour constituer le corpus, aucune contrainte sur l'âge des apprenants de FLE ou de ALE n'a été posée ; mais, d'une façon générale, seuls sont pris en compte les apprenants de plus de douze ans (ayant donc déjà achevé l'acquisition de leur langue première). Notons que les apprenants jeunes et scolarisés dans des structures comme les collèges ou lycées sont intéressants pour mener des suivis de cohorte. En effet, à ces âges, les élèves d'une même classe varient peu, et peuvent aisément être enregistrés à plusieurs reprises sur une année. Dans le cadre universitaire ou professionnel, cela est plus délicat.

Niveau de compétence en langue étrangère

Le niveau de compétence en langue étrangère des locuteurs a été pris en compte. Cette information est importante pour effectuer des études sur l'ordre d'acquisition, opérer des comparaisons en fonction des langues premières, etc. Pour ce corpus, une préférence a été donnée à des apprenants ayant un niveau de compétence pas trop avancé. Nous avons donc retenu des apprenants ayant des niveaux A2 et B1 d'après le CECR**.** Les apprenants débutants de niveau A1 ont été exclus dans la mesure où il leur serait difficile d'effectuer les différentes tâches demandées pour collecter les productions langagières.

Pour évaluer le niveau des étudiants, et être en mesure de s'en servir pour mener des recherches, deux méthodes ont été utilisées :

- les résultats obtenus aux tests de positionnement ou de niveau proposés lors d'inscription dans des cours de langue ont été pris en compte, d'autant qu'ils sont fréquemment formulés d'après le CECR.
- un questionnaire d'auto-évaluation a été proposé à chaque locuteur lors de la première séance d'enregistrement. Les questions posées tentent de déterminer ce qu'ils savent ou non faire dans la langue cible aussi bien à l'oral qu'à l'écrit, en production qu'en compréhension. Ces questions ont été construites en fonction du référentiel du CECR.

Profil des apprenants

Pour chaque apprenant, une fiche est établie afin de bien noter les informations nécessaires à l'établissement de son profil linguistique. Sont ainsi encodées les informations relatives à sa langue première (monolingue, bilingue, autres ; variété parlée, etc.), à son niveau de compétence en langue étrangère (niveau donné, niveau déterminé à partir de l'auto-évaluation, etc.). Sont également notés les langues étrangères maîtrisées par les apprenants enregistrés, les séjours effectués dans un pays où la langue cible est utilisée (date, durée, etc.). L'ensemble de ces informations est nécessaire pour étudier les processus d'acquisition et établir des comparaisons afin d'évaluer l'impact des langues premières dans l'acquisition des faits prosodiques en langue étrangère.

3.2. Les tâches demandées

Selon nous, il est important d'avoir accès à des données variées où les compétences communicatives et grammaticales des locuteurs ne sont pas nécessairement mobilisées de façon identique, d'où la modularité du protocole. Ce dernier propose en effet de nombreuses tâches qui sont regroupées selon trois critères essentiels : la typologie de la tâche, le niveau et l'âge requis pour l'effectuer. Cette manière de procéder à un double avantage : i) des tâches comparables, mais adaptées, peuvent être proposées à tous les locuteurs, même si leur âge ou leur niveau diffèrent, ii) en cas de suivi de cohorte, des tâches comparables mais différentes, peuvent être données à un même apprenant à chaque séance.

Cinq types de tâches ont été retenus. Parmi eux, trois grandes catégories peuvent être distinguées :

- la tâche de lecture oralisée où l'apprenant doit lire un texte donné. Il n'y a ni interaction ni liberté de production. Les textes retenus sont des extraits du corpus EUROM 1 et des extraits de textes de forme dialoguée. L'avantage de ces seconds textes est qu'ils font appel à l'utilisation d'intonation ou de prosodie particulière (usage d'exclamatives, ironie, etc.).
- les tâches monologuées, généralement élicitées à partir de matériaux variées : ces tâches sont au nombre de quatre. L'une d'entre elles consiste à faire un récit. Selon l'âge et le niveau des apprenants, deux possibilités : i) le locuteur peut avoir à répéter une histoire qui lui a été lue ou qu'il connaît déjà ; ii) il peut avoir à raconter l'histoire qui figure sur une série d'images qui lui est présentée (cf. figure 1).

Figure 1 : exemple d'images pour une tâche monologuée

  Dans deux autres tâches, on demande à l'apprenant de décrire un document visuel présenté. Dans le premier cas, il faut commenter ou décrire une image qui représente quelque chose de très statique comme une chambre, un tableau de peinture, une maison, etc. Dans le second, l'apprenant doit décrire une photo où des personnages sont en activités (sportifs, vendeurs sur un marché, etc.). La dernière tâche de cette catégorie consiste à effectuer une activité qui requiert l'utilisation de formes linguistiques particulières : donner une recette de cuisine, expliquer comment jouer à un jeu, etc.
- les tâches en situation d'interaction: ces tâches sont au nombre de deux. Dans l'une d'entre elles, l'apprenant doit poser une série de questions afin de connaître l'identité d'un personnage, ses goûts, son âge, etc. Dans l'autre, l'apprenant doit répondre à quelques questions qui lui sont posées par un investigateur.

3.3. Modalités expérimentales

Le protocole regroupe différentes tâches comparables, mais classées en fonction de l'âge ou du niveau des apprenants. Cette modularité permet de construire des séances d'enregistrement pour chaque locuteur en tenant compte de ses compétences, mais également des modalités expérimentales retenues. Pour une sous-partie du corpus, un suivi de cohorte a été entrepris. Il permet d'enregistrer un même apprenant à plusieurs reprises pendant un an, cela dans le but i) de voir comment se fait l'acquisition de tel ou tel fait prosodique, et ii) d'évaluer l'apport de certaines pratiques pédagogiques dans l'acquisition de la prosodie. Les enregistrements effectués

en une seule séance peuvent en revanche être utilisés pour mieux appréhender le poids de la langue première, etc.

Sur le plan matériel, l'ensemble des données a été enregistré avec le système développer à Munich dans le cadre du projet Wikispeech (cf. [9]).

4. ANNOTATION DES DONNEES

4.1. Présupposés théoriques

Le corpus COREIL se distingue de bien des corpus d'apprenants dans le mode d'annotation des données linguistiques. En effet, dans la plupart des cas, l'annotation linguistique de ce type de données consiste en un codage des erreurs (comparativement à la langue cible), codage effectué sans prise en compte du contexte de production (cf. [2] et [3]). Ces deux points sont, selon nous, contestables : d'une part, n'encoder que les erreurs, c'est oublier que la langue de l'apprenant fonctionne comme un système dont il importe de comprendre les règles ; d'autre part, en traitant les énoncés en dehors de leur contexte, il n'est pas toujours possible d'en évaluer, avec pertinence, l'acceptabilité. De plus, les annotations sont faites en minimisant les interprétations.

4.2. Transcription orthographique

L'ensemble du corpus est transcrit orthographiquement, la transcription étant alignée sur le signal au niveau de l'énoncé. Les conventions de transcription ont été élaborées en conformité avec les recommandations du TEI : recours à l'orthographe standard sans trucage orthographique, chiffres transcrits en toutes lettres, etc.

Les recommandations ont été adaptées afin de tenir compte de la spécificité du parler des apprenants. Lorsque ces derniers produisent une forme inexistante dans la langue cible, elle est transcrite selon les règles orthographiques de la langue cible (*I buyed a new car* pour *I bought a new car*). Lorsque l'apprenant produit une forme de façon incorrecte sur la plan phonétique ou phonologique, deux cas sont envisagés : i) si l'erreur est manifestement d'ordre phonétique ou phonologique, la transcription se fait selon l'orthographe standard (lorsque l'apprenant a dit [kɔn'ʃjus] au lieu de ['kɒnʃəs] pour *conscious* on a transcrit *conscious* ; de même, lorsqu'il a prononcé [mɔ̃teebu] pour [mɔ̃teeby], la transcription retenue a été *mon thé est bu*) ; ii) lorsque la prononciation erronée peut être liée à des difficultés d'ordre morpho-syntaxique, le recours à une multi-transcription proposant plusieurs choix possibles est obligatoire.

4.3. Annotation linguistique

Chaque annotation est faite sur une tire séparée, ce qui permet aisément d'ajouter des annotations. Dans l'immédiat, seul l'étiquetage grammatical a été effectué sur l'ensemble des données transcrites. Les programmes d'étiquetage intégrés au logiciel CLAN (Mor & Post) ont été utilisés. Ils permettent la création de dictionnaire et d'étiquette particulière. Cette fonctionnalité a été utilisée pour l'encodage des formes verbales, des locutions et des déterminants. Le choix des étiquettes s'est fait en respectant les présupposés retenus (cf. § 4.1). Un exemple d'étiquetage est donné sous (1) et (2).

(1) *APO: er what kind of courses do you have ?
%mor: fil|er pro:wh|what n|kind prep|of n|course-PL aux|dopro|you v|have

(2) *APO: euh je fais musique avec mon ordinateur et aussi euh je [x 2] j'écris euh j' écris des articles .
%mor:fil|euh pro:subj|je v:mdllex|faire-PRES-n|musique prep|avec det:poss|mon n|ordinateur&_ conj|et conj|aussi fil|euh pro:subj|je v|écrire-PRES fil|euh det|des&PL n|article.

Actuellement, une réflexion est menée sur la méthode qui pourrait être utilisée pour annoter les faits accentuels et intonatifs. La difficulté essentielle, propre à la transcription de la prosodie (cf., entre autres, [11]), est de trouver une façon d'encoder les phénomènes de façon symbolique et formelle, sans pour autant faire de présupposés lourds sur leur catégorisation et leur fonction dans la langue de l'apprenant.

5. CONCLUSION

Le corpus COREIL a été développé afin de travailler sur l'acquisition de la prosodie en langue étrangère, mais il est évident que les données collectées peuvent également être mises à disposition pour d'autres types d'étude linguistique. Le protocole retenu pour l'acquisition des données a été conçu afin de permettre une constitution évolutive du corpus. De plus, il permet de mener des études sur l'acquisition en croisant plusieurs paramètres : âge et niveaux des apprenants, rôle de la langue première, différences et/ou similitudes dans l'acquisition des langues premières et des langues étrangères, etc. De plus, l'annotation des données doit faciliter une réutilisation des données et éviter la mise en relation entre la langue cible et la langue des apprenants, cela dans le but d'étudier les systèmes des apprenants pour eux-mêmes.

BIBLIOGRAPHIE

[1] S. Granger. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL 37 (3):* 538-546, 2003.

[2] A. Lüdeling, S. Doolittle, H. Hirschmann, K. Schmidt & M. Walter. Das Lernerkorpus Falko. *Deutsch als Fremdsprache 45:* 67-73, 2008.

[3] Hawkins, J.A. & Buttery, P. Using learner language from corpora to profile levels of proficiency. In: *Studies in Language Testing*. Cambridge University Press, 2009.

[4] J.-T. Milde & U. Gut. A prosodic corpus of non-native speech. In *Proceedings of the Speech Prosody 2002*, pp. 503-506, 2002

[5] A. Tortel. ANGLISH: base de données comparative L1 & L2 de l'anglais lu, répété, parlé. *TIPA* 27 : 111-122, 2008.

[6] L. Rasier & P. Hiligsmann. Prosodic transfer from L1 to L2. Theoretical and methodological issues. *Nouveaux cahiers de linguistique française 28*, 2007.

[7] H. Voormann & U. Gut. Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory 4 (2)* : 235-251, 2008.

[8] D. Chan et alii. EUROM : a spoken language resource for the E.U. In *Proceedings Eurospeech 95* : 867-870, 1995.

[9] C. Draxler et K. Jänsch. SpeechRecorder – A Universal Platform Independent Multi-Channel Audio Recording Software. *LREC*. 2004.

[10] D. Hirst. Form and function in the representation of speech prosody. *Speech Communication 46* : 334-347, 200

Éléments de modélisation de l'effet de transformation verbale

Anahita Basirat[1,2] *et Jean-Luc Schwartz*[1]

[1]Gipsa-lab, CNRS UMR 5216, Grenoble INP, Université Joseph Fourier, Université Stendhal, Grenoble, France
[2]École Polytechnique Universitaire de Lille, Lille, France

ABSTRACT

The verbal transformation effect refers to perceptual alternations while listening to a speech sequence repeated rapidly. This effect is a rich source of information about the speech processing and speech perceptual binding mechanisms. In this work, we added some components to the TRACE model of speech perception in order to simulate our experimental results. Our key proposition concerns the role of the articulatory constraints in the verbal transformation effect.
Keywords: Multistable perception, Speech perception, Perceptual binding, TRACE model

1. Introduction

L'effet de Transformation Verbale (TV dans la suite) réfère aux changements perceptifs lorsqu'on écoute une répétition rapide d'une séquence verbale [18] : la mise en boucle du mot anglais *life* fournit la perception du mot *fly*. Ce phénomène est une source riche d'informations sur l'organisation perceptive de la parole et la nature de l'objet parole.

Dans le cadre de l'analyse de scènes auditives, Bregman propose des principes généraux d'organisation perceptive basés sur les mécanismes primitifs auditifs et les schémas [3]. Remez et al. montrent que l'organisation perceptive de la parole pourrait ne pas être seulement expliquée par ces principes [11]. En se référant à la théorie motrice, ils soulignent le rôle des représentations motrices dans l'organisation perceptive de la parole. Récemment, Sato et al. ont mis en évidence l'existence d'un biais d'asymétrie des transformations relevant de contraintes motrices liées à la cohérence entre gestes articulatoires des séquences répétées [15]. Ceci indique que des contraintes motrices peuvent participer à la construction des représentations mentales phonologiques. Nos études neuroanatomiques utilisant le paradigme de TV ont montré que les aires frontales et pariétales liées à la perception/production de la parole sont actives en lien avec les TV [13][2]. De plus, à travers l'effet de TV, nous avons étudié l'aspect multisensoriel de l'objet parole. Nous avons montré que les TV sont audio-visuelles et que la modalité visuelle intervient d'une manière active dans l'organisation des percepts [14]. En somme, ces études proposent que les mécanismes à la base des TV ne sont pas des mécanismes purement audiophonétiques mais plutôt perceptuo-moteurs et multisensoriels.

Nous proposons dans cet article quelques éléments de modélisation de l'effet de TV qui peuvent s'intégrer dans un modèle psycholinguistique tel que le modèle TRACE [9]. Nous avons choisi le modèle TRACE car, malgré les critiques, il peut expliquer un large spectre des données expérimentales. De plus, des travaux récents permettent d'améliorer et faciliter l'utilisation de ce modèle [10][17]. Dans la suite, nous présentons dans un premier temps deux modèles existants de l'effet de TV, fournissant des éléments de modélisation, mais faisant apparaître également le manque actuel d'un modèle psycholinguistique et computationnel convaincant. Dans un second temps, nous décrivons notre modèle en partant d'un rappel rapide du modèle TRACE et en indiquant comment nous l'avons modifié pour les simulations, dont nous présentons le principe et quelques résultats. Nous concluons cet article en proposant les perspectives de ce travail.

2. Modèles de l'effet de TV

Si l'effet de TV a été beaucoup étudié expérimentalement, aucun modèle computationnel réaliste n'a été proposé et confronté à l'ensemble de la phénoménologie. Nous présentons ci-dessous deux modèles partiels et complémentaires, qui indiquent clairement dans quelle direction il faut travailler.

2.1. Node Structure Theory, un modèle représentationnel incomplet

Ce modèle proposé par MacKay et al. explique l'effet de TV dans un cadre général de la perception/production de la parole [7]. Selon ce modèle, trois systèmes sont impliqués dans la perception de la parole : un système d'analyse acoustique, un système phonologique et un système phrastique (figure 1). Lorsqu'on écoute la mise en boucle du mot *base*, le mot le plus cohérent avec l'entrée (*base*) gagne et devient actif. Après une certaine durée, *base* devient saturé, ce qui entraîne l'activation du deuxième mot le plus cohérent avec l'entrée.

Ce modèle fournit une base intéressante pour modéliser l'effet de TV, en s'appuyant essentiellement sur des mécanismes de compétition avec saturation et fatigue des unités décisionnelles (« satiété »). Cependant, ce modèle n'a jamais été implémenté sous une forme computationnelle, ce qui conduit à une absence de précisions sur les mécanismes proposés. De plus, il ne peut expliquer le rôle de contraintes articula-

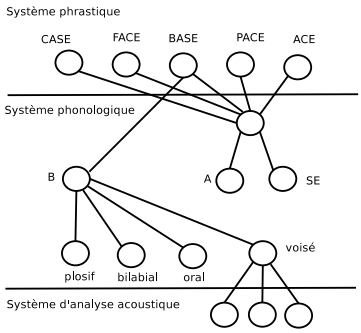


Fig. 1: *Node Structure Theory.* Entrée : mot *base.*

toires à la base des asymétries observées par notre groupe [15].

2.2. Le modèle de Ditzinger, un système dynamique sans représentations

Ditzinger et al. proposent un modèle de type système dynamique pour expliquer l'effet de TV [5]. Dans ce type de modèle, chaque percept est représenté comme un attracteur dans le système. Les attracteurs sont les états représentant les minima de l'énergie du système vers lesquels le système converge. Dans ce modèle (équation 1), la saturation de l'attention augmente l'énergie de la forme actuellement perçue, ce qui entraîne l'instabilité du percept et conduit à une transformation. Bien que ce modèle puisse rendre compte de la dynamique des transformations, il ne prend pas en compte la nature phonétique de ce phénomène, ce qui rend ce modèle très abstrait. Globalement, on peut dire que c'est un modèle mathématique non représentationnel, qui ne nous dit rien en réalité sur les processus de construction de l'identité phonémique dans la perception de la parole.

$$\dot{\xi}_k = \lambda_k - B\sum_{k'=1}^{M}\xi_{k'}^2\left(1-\alpha_{kk'}(1-\frac{2\xi_{k'}^4}{(\xi_k^2+\xi_{k'}^2)^2})\right) + B\xi_k^2 - C\sum_{k'=1}^{M}\xi_{k'}^2 \quad (1)$$

M : nombre de TV possibles
ξ_k : paramètre d'ordre k (ressemblance entre l'entrée et le percept k)
$\alpha_{kk'}$: biais (différence) entre les percepts k et k'
λ_k : saturation de l'attention au percept k

3. Eléments de modélisation des TV

On voit s'inscrire en creux de ces deux modèles existants la ligne que nous souhaitons suivre : celle du développement d'un modèle représentationnel computationnel, capable à la fois de prendre en compte les connaissances sur les processus de traitement et de représentation et notamment sur la nature perceptuo-motrice et multisensorielle de l'effet, et de fournir des prédictions quantitatives comparables aux données. Notre objectif est également de pouvoir associer les différents mécanismes proposés à des localisations neuroanatomiques en relation avec nos propres données. Le travail que nous présentons ici n'est encore que préliminaire. Il s'inscrit dans le cadre du développement du modèle TRACE, que nous rappelons rapidement.

3.1. Model TRACE

Le modèle TRACE est un modèle connexionniste à trois niveaux : niveau de trait phonétique, niveau phonémique et niveau lexical. Les unités dans chaque niveau sont représentées par des neurones artificiels. A l'intérieur des niveaux, les neurones s'inhibent les un les autres. Par contre, les connexions entre deux niveaux différents sont de type excitatrices. Ainsi, l'entrée du modèle excite les traits qui activent les phonèmes et les phonèmes à leur tour activent les mots. Un feedback est prévu du niveau lexical au niveau phonémique et du niveau phonémique au niveau de trait. Chaque unité est dupliquée au cours du passage du stimulus pour rendre compte du décours temporel.

Le modèle TRACE explique certaines observations expérimentales. Par exemple, l'effet Ganong existe dans TRACE. Cet effet concerne l'influence du statut lexical sur la catégorisation d'un phonème ambigu : un phonème ambigu entre [t] et [d] précédant /æsk/ est catégorisé comme [t] plutôt que [d] car *task*, contrairement à *dask*, est un mot. En exposant ce stimulus au modèle TRACE, le mot *task* s'active et renvoie cette activation au niveau du phonème. Les phonèmes constituant *task*, y compris [t], deviennent ainsi plus actifs, ce qui résout l'ambiguïté entre [t] et [d].

En ce qui concerne les TV, si on présente à l'entrée du modèle le mot *life* en boucle, la sortie sera constamment *life*. Autrement dit, l'effet de TV est absent dans le modèle. Nous présentons dans la suite quelques mécanismes qui pourraient d'une part faire émerger des TV dans le modèle TRACE, et d'autre part introduire des propriétés multisensorielles et perceptuo-motrices compatibles avec nos données.

3.2. Données expérimentales

Dans ce travail, nous nous sommes concentrés sur les stimuli de type /pata/ et /tapa/. Ces stimuli conduisent à des TV de resegmentation, qui sont le type le plus classique de TV, et qui se prêtent tout naturellement à des simulations dans le cadre de TRACE. Ces stimuli nous ont permis de démontrer à la fois les effets perceptuo-moteurs et audiovisuels. En ce qui concerne les premiers, il existe une tendance en faveur des séquences « Labiale-Coronale » (LC, comme /pata/) par rapport à celles de type CL (comme /tapa/) dans les langues du monde [8] et lors de l'acquisition du langage [4]. Cet effet, appelé l'effet LC, semble avoir une explication articulatoire : les séquences LC peuvent être réalisées en un seul cycle de mâchoire grâce à l'anticipation du geste de la consonne coronale, ce qui n'est pas possible pour les séquences CL [12]. Dans une expérience de TV, Sato et al. ont montré dans l'effet de TV qu'il existe une asymétrie perceptive entre ces deux formes : le percept /pata/ est un attracteur perceptif plus fort que /tapa/ [16]. Ils expliquent cette asymétrie par la cohésion articulatoire présentée ci-dessus.

En se basant sur ces études, nous proposons que le

liage perceptif en parole se fait par la cohésion articulatoire. Le liage préférentiel serait inséré dans le gabarit d'un geste d'ouverture de la mâchoire. Notre étude sur l'effet de TV audio-visuelle confirme cette proposition. Si l'on superpose au stimulus auditif /pata/ ou /tapa/ un geste visuel qui ne porte que sur une des deux syllabes (/paaa/ ou /taaa/), ce geste renforce la stabilité du percept commençant par cette syllabe (respectivement /pata/ ou /tapa/), ce que nous interprétons comme un effet de liage de l'information déclenché par le geste d'ouverture de la mâchoire, visible dans le stimulus visuel (effet de l'onset visuel) [14]. Nous proposons donc que les informations sur le geste de mâchoire peuvent être récupérées aussi bien à partir du signal auditif que visuel, et qu'elles font partie du processus de liage/segmentation/décision simulé dans la nouvelle version du modèle TRACE, que nous baptisons TRACE-VT (pour *Verbal Transformation*).

3.3. Implémentation de TRACE-VT

La figure 2 illustre les ajouts que nous proposons pour produire les TV au sein du modèle TRACE-VT. Ces ajouts sont de trois ordres.

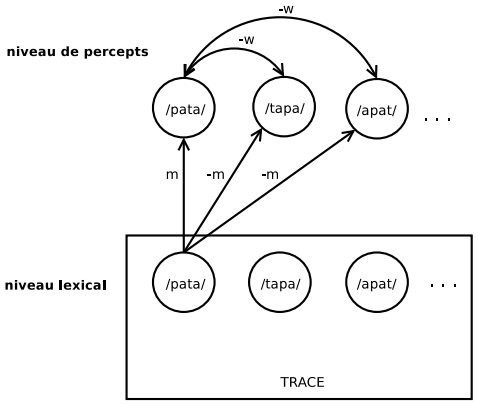


Fig. 2: Architecture de TRACE-VT.

Implémentation d'un processus de décision dynamique et stochastique, produisant des TV. Le niveau de décision est une machine de Boltzmann [6] où chaque percept est représenté par un neurone. La probabilité de chaque percept est calculée en fonction de l'énergie du système en équilibre lorsque l'unité représentant ce percept est active (équations 2 et 3). Ainsi, les percepts conduisant le système à une énergie plus basse sont plus probables. L'adaptation agit sur le seuil des unités : lorsqu'une unité est active, son seuil augmente, ce qui entraîne une augmentation de l'énergie et une baisse de sa probabilité. Cette baisse favorise l'activation d'une autre unité et une TV peut ainsi avoir lieu. La décision se fait sur une fenêtre de temps qu'on appelle fenêtre de liage. Après chaque décision, la fenêtre glisse à la fin du percept pour fournir une nouvelle entrée au modèle.

$$P\alpha = \frac{e^{-E_\alpha}}{\sum_{\text{Tous les états}} e^{-E_\beta}} \quad (2)$$

$$E_\alpha = -\sum_{i,j} y_i w_{i,j} y_j + \sum \theta_i y_i \quad (3)$$

$P\alpha$: probabilité de percevoir la forme phonétique α lorsque le système est en équilibre
$E\alpha$: énergie de l'état correspondant au percept α
y_i : sortie de l'unité i
$w_{i,j}$: poids de connexion de l'unité i vers l'unité j
θ_i : seuil de l'unité i

Introduction des biais articulatoires, liés au geste d'ouverture de la mâchoire, dans le processus de liage. La préférence de certains liages par rapport à d'autres en fonction du geste de la mâchoire (effet LC) est modélisée par une matrice qui peut amplifier ou non les sortie du niveau lexical de TRACE-VT (avant le niveau de décision). Cette matrice est calculée en fonction des traits reconnus. Par exemple, l'activation des unités qui commencent par /pa/ (dynamique importante du geste d'ouverture) est plus amplifiée que celle des unités commençant par /ta/ (présentant une dynamique moins importante du geste d'ouverture) et a fortiori encore plus que celles qui commencent par /ap/ ou /at/ (geste de fermeture). Cette préférence conduit d'abord à l'émergence de syllabes CV plutôt que VC, résultat classique et majeur des TV, en accord avec les théories de la sonorité. Elle généralise cette préférence à l'effet LC : dans cette implémentation, LC est à CL (ou « pata » est à « tapa ») ce que CV est à VC (ou « pa » à « ap »). Ainsi, la probabilité du percept /pata/ sera plus importante que celle de /tapa/, /apat/ ou /atap/.

Introduction d'une entrée visuelle Pour pouvoir intégrer à TRACE-VT nos données sur la nature multisensorielle du processus de transformation, et notamment notre résultat expérimental sur l'effet de l'onset visuel, nous avons ajouté à TRACE la possibilité de fournir des entrées visuelles. Ainsi, les biais articulatoires représentant le degré d'ouverture de la mâchoire sont plus importants lorsque l'entrée est présentée en modalité audio-visuelle que lors d'une présentation purement auditive.

On trouvera dans [1] toutes les informations détaillées sur les algorithmes et les paramètres numériques ayant permis d'effectuer les simulations.

3.4. Simulations

La figure 3 illustre des simulations de TRACE-VT. L'entrée auditive consiste en une mise en boucle des séquences /pata/ ou /tapa/ (150 répétitions). La modalité de présentation est auditive (A), audio-visuelle (AV), audio-visuelle /paaa/ (AV-pa) et audio-visuelle /taaa/ (AV-ta). Lors de la modalité AV-pa et AV-ta l'entrée visuelle est respectivement /paaa/ et /taaa/, /p/ et /t/ étant synchrones avec ceux du signal auditif. Les résultats illustrés sur la figure 3 consistent en 20 simulations par modalité et par type de stimulus. Les transformations possibles sont /pata/, /tapa/, /atap/, /apat/, /pa/ et /ta/. Les valeurs *delta* présentées sur cette figure correspondent à la stabilité globale du percept /pata/ moins celle du percept /tapa/ divisée par la durée de stimulus. Afin de pouvoir comparer l'apport des biais articulatoires sur la stabilité des percepts /pata/ et /tapa/, nous présentons également sur la figure 3 des simulations de TRACE-VT avec adaptation mais sans biais articulatoire. Nous pouvons constater que le mécanisme d'adaptation dans le processus de décision permet de produire effectivement des bascules mais il ne conduit

pas à la préférence pour certains percepts par rapport à d'autres (les valeurs de *delta* égalent à zéro). Ces résultats montrent que c'est en ajoutant les biais articulatoires à TRACE-VT que l'effet LC et l'effet de l'onset visuel peuvent émerger.

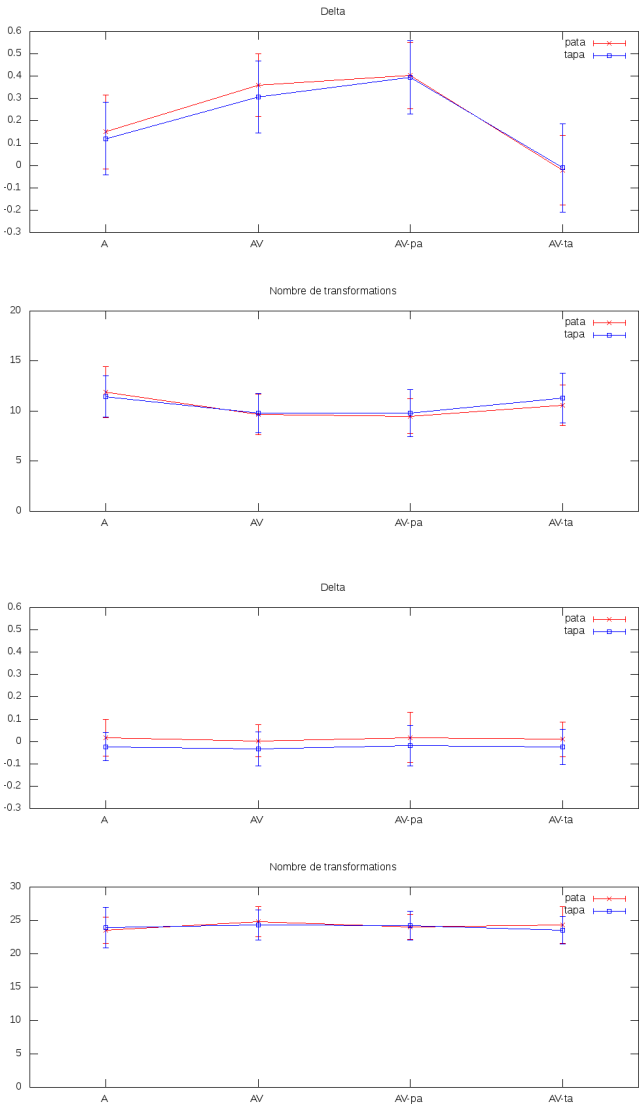


Fig. 3: Simulations : moyennes des valeurs de *delta* et du nombre de transformations verbales. En haut : TRACE-VT avec adaptation et avec biais articulatoires. En bas : TRACE-VT avec adaptation mais sans biais articulatoire.

4. Conclusion et perspectives

Le travail présenté dans cet article est un travail préliminaire pour modéliser la multistabilité perceptive en parole. Cette modélisation nous a permis de rassembler nos hypothèses sur les mécanismes sous-jacents à l'effet de TV et, plus généralement, ceux sur le liage perceptif en parole. Les perspectives de ce travail sont essentiellement de trois ordres. D'une part, des améliorations computationnelles devront être apportées pour ajouter au modèle des transformations autres que celles de type segmentation. Pour cela, l'ajout d'un module de traitement du signal auditif et visuel à l'entrée du modèle doit être envisagé. Le second enjeu est celui des données, expérimentales et de simulation. Les études expérimentales sur l'effet du geste de la mâchoire et sur la nature de la fenêtre de liage devront être poursuivies. Enfin, nous envisageons de préciser l'attribution neuroanatomique des différentes composantes de TRACE-VT, en relation avec les données de neuroimagerie.

C'est dans ce cadre unificateur, alliant données comportementales, neuroimagerie/neurophysiologie, et modélisation computationnelle, que nous parviendrons à mieux comprendre le fonctionnement de l'effet de TV, et à travers lui, les processus de construction de l'objet parole dans la cognition humaine.

Références

[1] A. Basirat. PhD thesis, Grenoble INP, soutenance prévue au printemps 2010.

[2] A. Basirat, M. Sato, J.L. Schwartz, P. Kahane, and J.P. Lachaux. *Neuroimage*, 42 :404–413, 2008.

[3] A.S. Bregman. *Auditory scene analysis : The perceptual organization of sound*. The MIT Press, 1990.

[4] B. Davis and P.F. MacNeilage. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 379–382, 2003.

[5] T. Ditzinger, B. Tuller, H. Haken, and J.A.S. Kelso. *Biological Cybernetics*, 77 :31–40, 1997.

[6] G.E. Hinton and T.J. Sejnowski. In *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, pages 282–317. 1986.

[7] DG MacKay, G. Wulf, C. Yin, and L. Abrams. *Journal of Memory and Language*, 32 :624–646, 1993.

[8] P.F. MacNeilage and B.L. Davis. *Science*, 288 :527, 2000.

[9] J. L. McClelland and J. L. Elman. *Cognit Psychology*, 18 :1–86, 1986.

[10] D. Mirman, J.L. McClelland, and L.L. Holt. *Psychonomic Bulletin and Review*, 13 :958, 2006.

[11] R.E. Remez, P.E. Rubin, S.M. Berns, J.S. Pardo, and J.M. Lang. *Psychological Review*, 101 :129–156, 1994.

[12] A. Rochet-Capellan and J.L. Schwartz. *The Journal of the Acoustical Society of America*, 121 :3740–3754, 2007.

[13] M. Sato, M. Baciu, H. Loevenbruck, J.L. Schwartz, M.A. Cathiard, C. Segebarth, and C. Abry. *NeuroImage*, 23 :1143–1151, 2004.

[14] M. Sato, A. Basirat, and J.L. Schwartz. *Perception & Psychophysics*, 69 :1360–1372, 2007.

[15] M. Sato, J.L. Schwartz, C. Abry, M.A. Cathiard, and H. Lœvenbruck. *Perception & Psychophysics*, 68 :458–474, 2006.

[16] M. Sato, N. Vallee, J.L. Schwartz, and I. Rousset. *Journal of Speech, Language and Hearing Research*, 50 :1466, 2007.

[17] T.J. Strauss, H.D. Harris, and J.S. Magnuson. *Behavior Research Methods*, 39 :19, 2007.

[18] M.R. Warren and R.L. Gregory. *American Journal of Psychology*, 71 :612–613, 1958.

Somatotopie motrice des articulateurs supralaryngés de la parole

Krystyna Grabski[1], Laurent Lamalle[2,3], Coriandre Vilain[1], Jean-Luc Schwartz[1], Nathalie Vallée[1], Irène Tropres[2,4], Monica Baciu[5], Jean-François Le Bas[2,6], Marc Sato[1]

[1]GIPSA-Lab, Département Parole et Cognition, UMR CNRS 5216 & Grenoble Universités; [2]IFR1 'RMN Biomédicale et Neurosciences', Unité IRM 3T, CHU de Grenoble; [3]INSERM; [4]Université Joseph Fourier; [5]Laboratoire de Psychologie et NeuroCognition, UMR CNRS 5105 & Université Pierre Mendès France; [6]CHU de Grenoble

krystyna.grabski@gipsa-lab.grenoble-inp.fr; marc.sato@gipsa-lab.grenoble-inpg.fr

ABSTRACT

In order to localize cerebral regions involved in articulatory control, ten participants were examined using functional magnetic resonance imaging while executing lip, tongue and jaw movements. Although the three motor tasks activated a set of common brain areas classically involved in motor control, distinct movement representation sites were nevertheless found in the primary motor cortex. These results support and extend previous brain imaging studies by demonstrating a sequential dorso-ventral somatotopic organization of lips, jaw and tongue in the motor cortex.

Keywords: supralaryngeal articulators, motor cortex, somatotopy, fMRI.

1. INTRODUCTION

Penfield et Boldrey [1] ont été les premiers à démontrer l'existence d'une organisation topographique musculaire au niveau du cortex moteur primaire. Lors de stimulations électriques appliquées à des régions distinctes du cortex moteur, ils ont en effet observé le déclenchement de mouvements musculaires spécifiques et ont ainsi établi une cartographie corticale motrice liée au contrôle de différentes parties du corps. Cette organisation corticale motrice, dite 'somatotopique', des parties corporelles ne reflète néanmoins pas la morphologie humaine, et l'étendue des zones motrices associées à chaque partie du corps correspond plutôt à la précision et la sensibilité du contrôle moteur des gestes associés à celle-ci. Une représentation spatiale différenciée des articulateurs de la parole a également été démontrée par Penfield et Rasmussen [2]. Ils ont en effet observé une organisation séquentielle dorso-ventrale des activations liées au contrôle des lèvres, de la mâchoire et de la langue au sein du cortex moteur primaire. Plus récemment, des études en imagerie par résonance magnétique fonctionnelle (IRMf) ont également démontré des activations différenciées au niveau du cortex sensorimoteur lors de différentes tâches motrices orofaciales. Ainsi, Lotze et al. [3], Hesselman et al. [4], Pulvermüller et al. [5] et Brown et al. [6] ont observé des activations sensorimotrices spatialement distinctes lors de la réalisation de mouvements labiaux et linguaux. De manière importante, les pics d'activations observés dans ces différentes études présentaient entre eux une très forte similarité spatiale, en accord avec une organisation dorso-ventrale lèvres-langue telle que suggérée par les travaux de Penfield et Rasmussen [2]. Enfin, Brown et al. [6] ont également observé une organisation somatotopique sensorimotrice dorso-ventrale larynx-lèvres-langue (voir également [7]).

Face à ces résultats et outre la mise en évidence d'un réseau neuro-anatomique fonctionnel commun associé au contrôle moteur des lèvres, de la mâchoire et de la langue, cette étude IRMf a pour but de tester une possible somatotopie dorso-ventrale lèvres-mâchoire-langue dans le cortex moteur.

2. MÉTHODES

2.1 Participants

Dix volontaires droitiers de langue maternelle française ont participé à l'étude (dont 8 hommes; âge : 21-43 ans). Cette étude a reçu un avis favorable du Centre Hospitalier Universitaire de Grenoble, du Comité de Protection des Personnes pour la recherche biomédicale de Grenoble et de l'Agence Française de Sécurité Sanitaire des Produits de Santé.

2.2 Procédure

L'expérience consistait en la réalisation distincte de trois tâches motrices, chacune à partir d'une position de base immobile (mâchoire fermée, lèvres et langue 'au repos') : protrusion des lèvres, mouvement retroflèxe de la langue et ouverture mandibulaire. Une tâche contrôle, sans aucune activité motrice, a également été testée. Les participants devaient produire l'une des 4 conditions toutes les 10 secondes selon un ordre pseudo-aléatoire (une même condition ne pouvant survenir plus de deux fois de suite). Chaque condition consistait en 18 essais. Pour chaque essai, précédée d'une croix de fixation durant 500ms, une consigne visuelle ('lèvres', 'langue', 'mâchoire' ou 'repos') indiquait au participant la condition à réaliser et la durée du mouvement (1s). 75 scans fonctionnels ont ainsi été acquis pour une durée totale de 13 minutes.

2.3 Matériel et acquisition des données IRM

A l'aide du logiciel Presentation (Neurobehavioral Systems, Albany, EU), les consignes visuelles ont été projetées au moyen d'un vidéo projecteur sur un écran

situé derrière le participant et, par réflexion, sur un miroir placé au dessus de ses yeux. Lors de l'expérience, les participants portaient des bouchons d'oreille et un casque antibruit.

Les acquisitions des images anatomiques et fonctionnelles ont été réalisées sur un imageur corps entier 3T (Brucker Medspec S300) muni d'une antenne tête émission/réception à champ de vue large. Pour les scans fonctionnels, une séquence d'acquisition en écho de gradient pondérée en T2* a été utilisée. Pour chaque volume fonctionnel, quarante coupes axiales adjacentes ont été acquises en mode entrelacé (temps de répétition: 10s, temps d'acquisition : 2600ms, résolution: 3 mm^3). Entre les conditions de perception et de production, un volume anatomique de haute résolution (1 mm^3) pondérée en T1 a également été acquis.

Afin de minimiser de possibles artefacts de mouvement sur les images fonctionnelles, un paradigme d'acquisition de type 'sparse sampling' a été utilisé. Cette technique d'acquisition est basée sur le délai temporel existant entre l'activité neuronale liée à une tâche motrice ou à l'écoute d'un stimulus auditif et le délai de la réponse hémodynamique associée. Face au délai estimé dans de précédentes études du pic de la réponse hémodynamique lors de la production de mouvement orofaciaux ou de séquences de parole [8-11], l'intervalle de temps séparant la perception ou la production d'une voyelle et l'acquisition du volume fonctionnel correspondant variait aléatoirement entre chaque essai de 4s à 6s.

2.4 Prétraitements et analyses statistiques

Les données ont été analysées à l'aide du logiciel SPM5 (Statistical Parametric Mapping; Wellcome Department of Cognitive Neurology, Londres, RU). Pour chacun des participants, les images fonctionnelles ont été réalignées, normalisées dans l'espace commun du Montreal Neurological Institute (repère MNI) et lissées via un filtre gaussien passe-bas de 6 mm3.

Les réponses hémodynamiques correspondantes aux conditions expérimentales ont ensuite été estimées selon un modèle linéaire général, incluant la caractérisation d'une réponse à impulsion unique pour chaque scan fonctionnel et l'ajout de régresseurs de non-intérêt liés aux paramètres de mouvements. Enfin, un filtrage des basses fréquences *a priori* non-reliées aux conditions expérimentales (variations lentes d'origine physiologique) a été appliqué (fréquence de coupure de 1/128 Hz).

Analyses statistiques individuelles: 3 contrastes ont été calculés pour déterminer les régions spécifiquement activées lors des 3 tâches motrices par rapport à la condition contrôle de repos (contrastes [langue-contrôle], [lèvres-contrôle], [mâchoire-contrôle]) selon un seuil statistique défini à $p < .005$ corrigé pour les comparaisons multiples (test de type "family-wise error", ou FWE) et une taille minimale des clusters de voxels activés supérieure à 25. Les cordonnées tridimensionnelles x,y,z des centres de gravité ('COG') des clusters activés dans les cortex moteurs gauche et droit ont été calculées au moyen du logiciel Anatomy [12]. Afin de situer spatialement les COGs, une analyse de variance (ANOVA) à mesures répétées à été calculée avec pour variable dépendante la cordonnée z (relative à la localisation dorso-ventrale du centre de gravité) et pour variables indépendantes l'hémisphère (cortex moteur gauche/droit) et le geste articulatoire (lèvres, langue, mâchoire).

Analyse statistique de groupe: A partir des contrastes individuels, une analyse de groupe 'à effets aléatoires' (consistant en une ANOVA à mesures répétées à un facteur (geste) de trois niveaux (lèvres, langue, mâchoire) et un facteur 'participant' implicite) a été réalisée. Trois nouveaux contrastes ont été calculés de manière à déterminer les régions spécifiquement activées lors des 3 tâches motrices par rapport à la condition contrôle de repos, selon un seuil statistique défini à $p < .005$ corrigé FWE et une taille minimale des clusters supérieure à 25 voxels.

3. RÉSULTATS

Analyse de groupe: Les résultats de l'analyse à effets aléatoires démontrent un ensemble d'activations cérébrales commun à la réalisation des trois tâches motrices (Figure 1 & Table 1). Ce réseau neural inclut des régions classiquement dévolues au contrôle moteur et observées lors de précédentes études liées à la réalisation de mouvements orofaciaux [3,4, 6,13]. Il comprend des

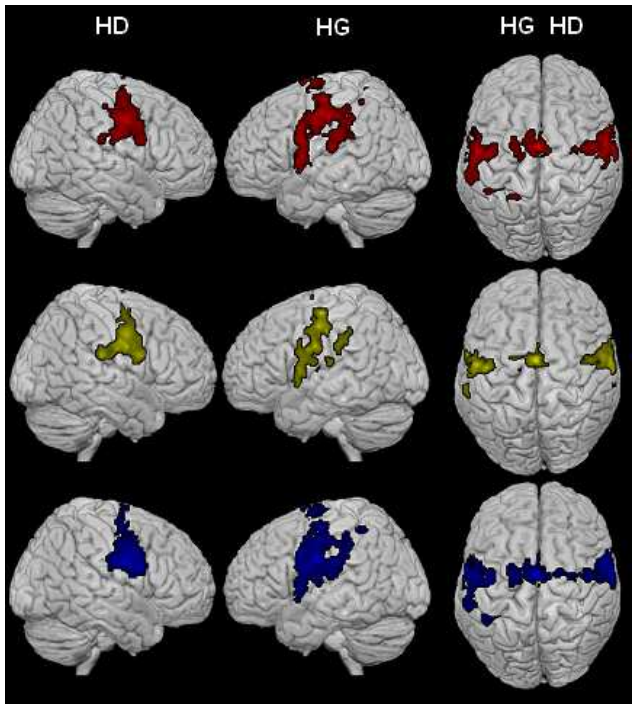


Figure 1: Projection surfacique des activations observées lors de la réalisation des mouvements des lèvres (haut), de la mâchoire (milieu) et de la langue (bas). Analyse de groupe, $p < .005$ corrigé, taille minimale des clusters supérieure à 25 voxels. HG : hémisphère gauche, HD : hémisphère droit.

Aires cérébrales	Lèvres						Mâchoire						Langue			
	Hém. Gauche				Hém. Droite		Hém. Gauche				Hém. Droite		Hém. Gauche				Hém. Droite	
	x	y	z	*t*	x	y	z	*t*	x	y	z	*t*	x	y	z	*t*	x	y	z	*t*
Aires frontales																				
motrice supplémentaire	-2;	-4;	60	*28,02*	12;	-4;	64	*10,87*	-2;	-4;	60	*9,75*					-2;	-4;	60	*32,91*	12;	2;	64	*11,69*
prémotrice	-54;	-10;	48	*22,44*	42;	-10;	50	*25,06*	-46;	-12;	60	*2,81*	42;	-10;	50	*22,57*	-62;	0;	28	*21,02*	60;	-4;	30	*29,5*
motrice primaire	-48;	-14;	42	*8,29*	56;	-4;	40	*20,86*	-48;	-14;	42 8	*5,67*	40;	-12;	40	*15,77*	-56;	-6,	30	*30,96*	54;	-12;	34	*14,32*
aire de Broca	-56;	8;	34	*19,42*	60;	8;	24	*16,96*					54;	2;	30	*15,12*	-56;	8;	34	*21,18*	60;	8;	24	*23,37*
Aires pariétales																				
somato-sensorielle primaire	-58;	-18;	48	*18,36*	52;	-24;	28	*15,21*	-58;	-16;	48	*15,63*	64;	-12;	28	*18,5*	-36;	-40;	50	*12,78*	64;	-12;	28	*20,59*
pariétale inférieure	-58;	-26;	30	*18,00*													-58;	-24;	28	*19,44*	40;	-34;	52	*18,46*
Aires auditive									-52;	8;	0	*15,62*								
Aires sous-corticales																				
Insula					40;	-2;	-4	*18,7*	-40;	-2;	-4	*17,55*	44;	-28;	22	*13,59*	-40;	0;	-2	*18,55*				
Cingulaire	-6;	8;	38	*12,86*	6;	12;	40	*13,84*	-4;	0;	50	*12,48*	6;	12;	40	*13,43*	-6;	8;	40	*11,62*	6;	12;	40	*13,84*
Putamen	-24;	2;	-4	*14,00*					-24;	2;	-4	*16,37*	26;	6;	-6	*14,8*					26;	4;	-6	*14,83*
Thalamus									-12;	-18;	4	*14,23*	12;	-20:	4	*12,89*				
Cervelet																	-16;	-62;	-20	*13,1*	16;	-58;	-22	*12,4*
Tronc cérébral													8;	-26;	-4	*11,84*				

Table 1: Coordonnées dans l'espace MNI (Montreal Neurological Institute) des pics d'activation observés pour les trois articulateurs.

activations du cortex moteur primaire orofacial (exécution de mouvements) et du cortex somatosensoriel adjacent (proprioception), de l'aire motrice supplémentaire dont l'activation s'étend au cortex cingulaire médian (initiation de patterns de mouvements), du cortex prémoteur ventral (planification motrice), du gyrus frontal inférieur, notamment la pars opercularis de la région de Broca (réalisation de mouvements complexes, notamment orofaciaux), de l'insula (coordination de gestes articulatoires), du putamen (sélection du mouvement particulier à réaliser), du thalamus (programmation motrice), du cervelet (coordination musculaire), et du lobule pariétal inférieur (site d'intégration d'informations somesthésiques et de coordination spatio-temporelle de mouvements).

Les pics d'activations observés dans les cortex moteur primaire gauche et droit suggèrent une organisation dorso-ventrale de type lèvres-mâchoire-langue [2] et, pour les lèvres et la langue, sont similaires à ceux observés lors de précédentes études IRMf [3-6]. Cependant, du fait de l'étendue des clusters activés lors des trois tâches motrices et leur recouvrement important, cette analyse de groupe ne permet pas d'établir une réelle somatotopie motrice des trois articulateurs.

Analyses individuelles: Pour tous les participants, on observe une organisation dorso-ventrale entre les COGs dans le cortex moteur relatifs aux lèvres, la mâchoire et la langue (Figure 2). Cette représentation somatotopique des 3 articulateurs sur un plan dorso-ventral est confirmée par les résultats de l'ANOVA montrant un effet geste significatif ($F(2,18) = 18.39$, $p < .00005$). Des analyses post-hoc (tests LSD de Fisher) démontrent une localisation plus dorsale des lèvres par rapport à la mâchoire ($p < .02$) et à la langue ($p < .0001$), et une localisation plus ventrale de la langue par rapport à la mâchoire ($p < .005$).

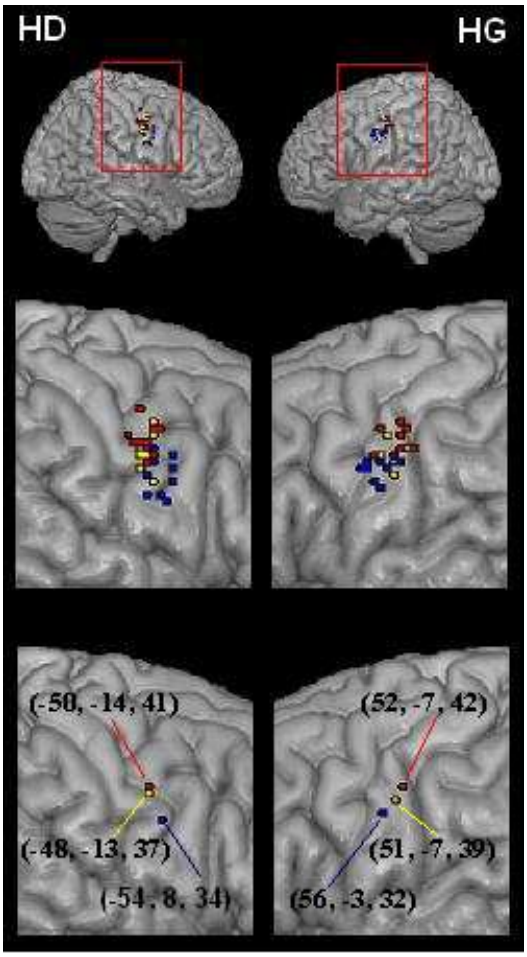


Figure 2: COGs individuels (en haut) et moyens (en bas) observés dans le cortex moteur lors de mouvements des lèvres, de la mâchoire et de la langue avec les coordonnées moyennes dans l'espace MNI des COGs moyens pour les trois articulateurs. HG/HD : hémisphère gauche/droit.

En absence de différences significatives de localisation des COGs entre les cortex moteur gauche et droit et d'interaction entre les variables 'geste' et 'hémisphère', ces résultats suggèrent ainsi une organisation motrice bilatérale somatotopique dorso-ventrale et montre une concordance relative entre les dix participants quant aux COGs liés à chacun des articulateurs.

4. DISCUSSION & PERSPECTIVES

En plus de la mise en évidence d'un réseau neuro-anatomique fonctionnel commun associé au contrôle moteur des lèvres, de la mâchoire et de la langue, cette étude a permis de montrer une somatotopie individuelle dorso-ventrale dans le cortex moteur liée au contrôle des lèvres, de la mâchoire et de la langue. Cette organisation topographique des trois articulateurs, cohérente avec les résultats obtenus par stimulations électriques de Penfield et Rasmussen [2] et avec de précédentes études IRMf des lèvres et la langue, est ainsi démontrée pour la première fois en IRMf. De plus, les résultats obtenus lors de cette étude, compte tenu d'un délai d'acquisition relativement limité, démontrent également l'intérêt de l'utilisation du paradigme d'acquisition de type 'sparse sampling', qui à notre connaissance n'avait jamais encore été testé dans la recherche d'un gradient somatotopique des représentations motrices. Ce paradigme d'acquisition pourrait ainsi s'avérer d'un grand intérêt pour les recherches et applications cliniques portant par exemple sur l'étude des mécanismes de plasticité motrice ou comme outil de localisation préalable dans le domaine de la neurochirurgie.

L'ajout de tâches motrices supplémentaires, notamment liées au contrôle du larynx, permettrait d'affiner nos connaissances sur l'organisation cérébrale des articulateurs de la parole, tout en permettant de valider les réseaux neuraux et l'organisation somatotopique motrice obtenus. Enfin, les activations observées seront utilisées en tant que localisateurs moteurs des différents articulateurs de la parole dans des études ultérieures. Cette étude s'inscrit, en effet, dans le cadre d'un projet plus vaste qui a pour but d'établir une cartographie précise des activations motrices, somatosensorielles et auditives liées à la perception et production des voyelles du français et, de tester de possibles interactions fonctionnelles entre les systèmes de perception et de production de la parole (voir Grabski, K. et al., "Corrélats neuroanatomiques des systèmes de perception et de production des voyelles du Français", ce volume).

REMERCIEMENTS

Cette étude s'inscrit dans le cadre du BQR "Modyc: Modélisation dynamique de l'activité cérébrale" financé par l'Institut National Polytechnique de Grenoble.

BIBLIOGRAPHIE

[1] Penfield, W. and Boldrey, E. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain, 60: 389-443, 1937.

[2] Penfield, W. and Rasmussen, T. The Cerebral Cortex of Man. New York: Macmillan, 1950.

[3] Lotze, M. et al. The representation of articulation in the primary sensorimotor cortex. Neuroreport, 11: 2985-2989, 2000.

[4] Hesselmann, V. et al. Discriminating the cortical representation sites of tongue and lip movement by functional MRI. Brain Topography, 16: 159-167, 2004.

[5] Pulvermüller, F. et al. Motor cortex maps articulatory features of speech sounds. PNAS, 103: 7865-7870, 2006.

[6] Brown, S. et al. Larynx area in the human motor cortex. Cerebral Cortex, 18: 837- 845, 2007.

[7] Loucks, T. et al. Human brain activation during phonation and exhalation: common volitional control for two upper airway functions. NeuroImage, 36: 131-143, 2007.

[8] Gracco, V.L., Tremblay P. And Pike, G.B. Imaging speech production using fMRI. NeuroImage, 26: 294-301, 2005.

[9] Sörös, P. et al. Clustered functional MRI of overt speech production. NeuroImage, 32(1): 376-387, 2006.

[10] Bohland J.W. and Guenther F.H. An fMRI investigation of syllable sequence production. *NeuroImage*, 32(2): 821-841, 2006.

[11] Özdemir, E., Norton, A. And Schlaug, G. Shared and distinct neural correlates of singing and speaking. *NeuroImage*, 33: 628–635, 2006.

[12] Eickhoff, S.B. et al. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage, 25: 1325-1335, 2005.

[13] Watanabe, J. et al. The human parietal cortex is involved in spatial processing of tongue movement - an fMRI study. NeuroImage, 21(4): 1289-1299, 2004.

Étude statistique de la durée pausale dans différents styles de parole

Jean-Philippe Goldman [1,4], Thomas François [2,5], Sophie Roekhaut [3,5], Anne Catherine Simon [4]

[1] Département de Linguistique, Université de Genève, Suisse ;
[2] Aspirant FNRS ; [3] TCTS Lab, Université de Mons, Belgique ;
[4] Institut Langage & Communication/Valibel - Discours & Variation, Université catholique de Louvain, Belgique ;
[5] Institut Langage & Communication/CENTAL , Université catholique de Louvain, Belgique ;
Jean-Philippe.Goldman@unige.ch, [Thomas.Francois, Anne-Catherine.Simon]@uclouvain.be, Sophie.Roekhaut@umons.ac.be

ABSTRACT

This paper presents a study on duration of silent pauses in spontaneous and read speech for several speaking styles in French. A univariate analysis shows that the observed multimodality is more exactly a mixture of logarithmic distributions. Then, various explanatory factors are considered, such as audible breathing, hesitation, speech rate and position regarding different level of syntactic boundaries, in order to build a predictive model of pause duration.
Keywords : pause duration, mixture model, spoken French, speaking styles

1. INTRODUCTION

Élément à part entière du langage oral, la pause silencieuse peut véhiculer bon nombre d'informations linguistiques et extra-linguistiques qui dépendent de sa longueur, de sa position dans l'énoncé ou encore de la présence ou non d'une prise de souffle. Grosjean [15, 14] a jeté les bases des travaux sur la pause en français en définissant les variables temporelles simples et complexes, et en tentant d'expliquer les variations de la durée des pauses (respiratoires et non respiratoires) en fonction du débit de parole et des frontières de constituants mineurs. Plusieurs études plus récentes [17, 8] font état d'une répartition multimodale (bimodale ou trimodale) de la durée des pauses, mais suggèrent chacune une valeur seuil différente [1]. Campione et Véronis [8] rappellent également deux éléments importants : la distribution de la durée des pauses s'apparente à une distribution lognormale et l'utilisation de seuils mal choisis peut fausser des conclusions.

Dans cet article, nous cherchons, sur la base de l'analyse de divers styles de parole en français, à mieux expliquer la variation de la durée des pauses à l'aide d'une série de variables parmi lesquelles la présence de respirations, la présence d'hésitations, le débit d'articulation, la position syntaxique. Cette approche analytique s'insère dans la perspective d'une meilleure prédiction de la durée et du type de pauses dans un système de synthèse de parole multi-styles. Dans un premier temps, nous présentons le corpus sur lequel l'étude a été réalisée, avant d'exposer les résultats des analyses statistiques. Celles-ci se limitent d'abord à une approche univariée de la distribution de la durée des pauses, qui tend à répliquer les résultats obtenus par Zellner [17] et Campione & Véronis [8]. Dans un second temps, elles visent à affiner le diagnostic et à expliquer la répartition des pauses en plusieurs catégories de durée à l'aide de différentes variables explicatives, et ce pour deux niveaux différents : pour l'ensemble des pauses du corpus et au niveau des styles de parole.

1. Le seuil est situé à 200ms chez Candea [9] et entre 180 et 250 ms chez Duez [11].

2. CORPUS D'ÉTUDE

Le matériel utilisé est une sous-partie du corpus C-PROM [2], un corpus multi-styles de français parlé, développé dans le but initial d'étudier les proéminences prosodiques syllabiques.

2.1. Description des enregistrements

Le sous-corpus fait environ 40 minutes, regroupe 11 locuteurs répartis en 4 styles de parole (JPA : journaux parlés, LEC : lecture, NAR : narration conversationnelle, CNF : conférence scientifique). Chaque style est représenté par des enregistrements de 3 locuteurs provenant de France, de Suisse et de Belgique, étant des professionnels des médias pour JPA, des enseignants universitaires pour CNF, et des particuliers pour LEC. Au total, le matériel d'étude, dont les principales caractéristiques sont décrites dans la Table 1, est composé de 806 pauses.

TABLE 1: Présentation du corpus par style de parole selon la durée (en secondes), le nombre de locuteurs, de syllabes, d'unités de rection, de séquences, de groupes rythmiques (ou "chunks)" et de pauses

style	loc (nb)	durée (sec)	syll (nb)	urc (nb)	seq (nb)	rg (nb)	pauses (nb)
CNF	3	687	3439	375	206	783	297
JPA	3	621	3383	452	150	746	206
LEC	2	250	1326	162	58	327	110
NAR	3	622	2881	814	569	1168	193
ALL	11	2180	11029	1803	983	3024	806

2.2. Description des annotations

Le corpus est entièrement annoté sous Praat [6], chaque type d'annotation faisant l'objet d'une ou plusieurs tires spécifiques :
– alignement phonétique (réalisé semi-automatiquement [13] et vérifié manuellement), syllabique et par mots graphiques ;
– détection par deux experts, selon un protocole strict, des syllabes perçues comme proéminentes et des phénomènes de production pouvant avoir une incidence sur la structuration prosodique (hésitations, interruptions, prises de souffle, etc.) [2] ;
– annotation manuelle et semi-automatique en unités syn-

taxiques de différents rangs, selon la grammaire de dépendance (voir [5] et suivant le protocole décrit dans [10]) :

– l'unité syntaxique maximale est l'unité de rection, c'est-à-dire un élément recteur (le plus souvent verbal) et tous les constituants qui en dépendent ;
– les unités de rection sont analysées en séquences fonctionnelles (par ex. séquence sujet, séquence verbe, séquence régie, insert, etc.) selon les critères exposés dans [4] ; les éléments non régis (éléments disloqués, adjoints, marqueurs de discours, connecteurs, etc.) sont annotés de manière spécifique en dehors des unités de rection et des séquences ;
– le dernier niveau d'analyse syntaxique contient des "chunks" [7] ou groupes rythmiques comprenant un mot grammatical suivi par un ou plusieurs mots lexicaux [2].

Un exemple d'annotations syntaxiques pour une phrase du corpus est donné à la Table 2. Une série de scripts permettent de créer des tables rassemblant des informations issues de la confrontation des différents niveaux d'annotation.

TABLE 2: Annotations syntaxiques : exemple

	beaucoup	d'	êtres	humains	ont	faim
catégories	DETCONST		NOM	ADJ	VERB	NOM
séquences	Séquence sujet				Séquence verbe	
groupes rythm.	RG1				RG2	
rection	urc					

3. ANALYSE STATISTIQUE UNIVARIÉE

Divers chercheurs [17, 8, 16] ont montré que la distribution de la durée des pauses est plutôt lognormale [3] que normale. Nos analyses ont effectivement conforté l'intérêt d'utiliser le logarithme de la durée plutôt que la durée, puisqu'on observe une queue de distribution alongée sur la droite. Toutefois, alors que Campione & Véronis [8, p. 200] rapportent une distribution lognormale pour l'ensemble de leurs données multilingues, nous n'observons pas cette lognormalité pour nos données : ni pour l'ensemble du corpus (Wde Shapiro-Wilk $= 0,971; p < 0.001$), ni pour chacun des styles pris séparément. La distribution apparaît plutôt comme multimodale.

Par conséquent, nous avons cherché à compléter les résultats de Campione & Véronis [8], lesquels rapportaient également une distribution multimodale pour le français, sans toutefois caractériser les diverses populations qui la constituent. Dans ce but, la distribution multimodale est analysée comme un mélange de lognormales :

$$f(x) = \sum_{i=1}^{N} \pi_i \Lambda_i(\mu_i, \sigma_i^2, x) \quad (1)$$

où π_i représente la probabilité qu'une pause soit générée par la distribution lognormale Λ_i, dont les paramètres sont μ_i, la moyenne et σ_i^2, la variance. Notons que $\sum_{i=1}^{N} \pi_i = 1$. Pour chaque distribution de pauses (pour tout le corpus et par style), nous avons appliqué un algorithme de clustering automatique par modèle probabiliste nommé

2. Ce découpage est réalisé automatiquement à partir de l'étiquetage semi-automatique en unités grammaticales réalisé par le module NLP du synthétiseur eLite [3].

3. Le logarithme employé par la suite est en base 10.

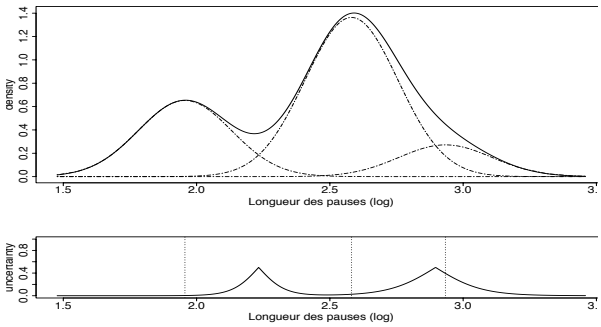


FIGURE 1: En haut : fonction de densité multimodale pour l'ensemble du corpus, superposée aux fonctions de densité des 3 composantes correspondant aux 3 types de pauses. En bas : fonction d'incertitude utilisée pour repérer les seuils.

mclust disponible pour R. Celui-ci opère en deux temps [12]. D'une part, il effectue un clustering hiérarchique qui permet de sélectionner le nombre N optimal de composantes par optimisation du critère BIC (Bayesian Information Criteria). Ce résultat constitue une bonne initialisation pour la seconde étape au cours de laquelle est estimée la valeur des paramètres de ces N composantes au moyen d'une procédure d'Expectation Maximization. La Table 3 reprend, par style et pour l'ensemble du corpus (ALL), les probabilités d'appartenance des données aux différentes distributions ($\pi(x)$), la durée moyenne des pauses pour ces différents groupes (μ ; dont la valeur présentée ici est retransformée en ms et correspond à la moyenne géométrique) et l'estimation d'un point de coupure (t ; en ms) entre les différentes composantes [4]. Ces seuils déterminent dès lors trois types de pauses : courtes, moyennes et longues.

TABLE 3: Estimation des paramètres pour le mélange lognormal : moyenne(ms) et probabilités d'appartenance(%) pour chaque type de pauses (courtes, moyennes et longues) et seuils entre les modes (t en ms) .

	COURTES			MOYENNES			LONGUES	
	μ (ms)	π (%)	t (ms)	μ (ms)	π (%)	t (ms)	μ (ms)	π (%)
CNF	88	30	159	363	70	_	_	_
JPA	94	39	233	332	33	417	481	28
LEC	84	23	162	516	77	_	_	_
NAR	117	23	222	505	77	_	_	_
ALL	90	29	171	381	60	790	858	12

Au niveau des divers styles, on remarque une très nette tendance à la bimodalité. Seul le style JPA présente un troisième mode (28,24% des données). Nous pouvons remarquer que ce troisième mode contient uniquement des pauses avec prises de souffle (voir 4). Nos résultats semblent donc se démarquer des observations de Campione & Véronis [8], pour qui la présence d'un troisième composant serait associé au discours spontané. Il est vrai que, pour l'ensemble des données, un troisième compo-

4. Les seuils correspondent aux points de l'échantillon qui sont des maximums locaux de la fonction d'incertitude du modèle (voir figure 1). L'incertitude se calcule pour une observation donnée comme 1 − la probabilité d'appartenance de cette observation à la composante la plus probable du modèle.

sant apparaît (voir Figure 1), mais sa présence reflète vraisemblablement le fait que ces données sont constituées de pauses issues de divers styles, et en des proportions diverses. On note également que la valeur du seuil entre les 2 modes principaux (pauses courtes et moyennes) se situe autour de 200ms comme dans les études citées mais surtout qu'elle varie en fonction du style de parole.

4. INFLUENCE DES VARIABLES EXPLICATIVES SUR LA DURÉE

Repérer une bimodalité de la distribution de la durée des pauses est une chose. Pouvoir l'expliquer en termes de différences de caractéristiques des deux composantes en est une autre. C'est pourquoi, nous avons isolé une série de variables susceptibles d'expliquer la variabilité de la durée des pauses : la présence d'une prise de souffle (*pds*), la présence d'une hésitation avant ou après la pause (*hesi*), le débit local (*debit*) [5] et la position de la pause relativement aux niveaux de l'annotation syntaxique, soit à une frontière de groupe rythmique (*rgr_pos*), de séquence (*seq_pos*) ou d'unité de rection (*urc_pos*) (voir la Table 4). Par ailleurs, afin de mieux isoler les caractéristiques propres à chaque composante, la variable *durée* a été discrétisée en fonction des seuils décrits dans la Table 3, afin d'obtenir un nombre de niveaux équivalent au nombre de composants. Ainsi, pour l'ensemble du corpus, on distingue les pauses courtes, moyennes et longues.

TABLE 4: Résumé des différentes caractéristiques des pauses par style : pourcentage de pauses avec prise de souffle, jouxtant une hésitation, en fin d'unité de rection, de séquence et de groupe rythmique, par rapport au nombre de pause total, et débit local.

	CNF	JPA	LEC	NAR	ALL
Pauses avec pds (%)	56	58	60	60	58
Pauses avec hesi (%)	24	6	0	34	19
Débit local (syll/sec)	5,17	5,7	6,26	5,21	5,47
Pauses en fin d'urc (%)	29	42	42	70	44
Pauses en fin de seq (%)	41	61	65	81	59
Pauses en fin de rgr (%)	74	81	86	91	81

Le facteur le plus significativement associé avec la durée des pauses est la présence d'une prise de souffle. C'est le cas pour l'ensemble du corpus ($\chi^2 = 448,32$; $p < 0.001$), comme pour chacun des styles. Cette significativité impressionnante du Khi-carré correspond bien à un effet important de *pds* (V de Cramer = 0.74), visible dans la table de contingence (Table 5). On remarque ainsi au niveau du corpus, et cela se confirme pour chacun des styles, que les pauses courtes ne comportent pas de prise de souffle. Il s'agit donc d'une caractéristique qui distingue radicalement le premier composant du mélange des autres. Notons que cette première conclusion avait déjà été soulignée par [16] pour l'anglais.

Par contre, le débit local ($Rs = -0.01$; $p = 0.78$) et la présence d'hésitations ($\chi^2 = 1.09$; $p < 0.58$; $V = 0.04$) sont tous deux apparus commne non significativement corrélés avec la durée. Ce résultat peut sembler étonnant de prime abord, mais semblable conclusion avait déjà été obtenue par [17, p. 75] au sujet du débit.

5. Le débit local est calculé en nombre de syllabes par seconde, dans la suite sonore située entre la pause précédente et la pause suivante.

TABLE 5: Table de contingence pour l'ensemble du corpus : nombre de pauses non-finales (in) et finales (end) d'*urc*et en fonction de la présence ou non d'une *pds*

		urc_pos		
	pds	in	end	Total
PAUSES COURTES	0	190	40	230
	1	0	0	0
PAUSES MOYENNES	0	59	37	96
	1	196	211	407
PAUSES LONGUES	0	1	10	11
	1	4	58	62
TOTAL		450	356	806

Enfin, suite aux réflexions de Grosjean [15] sur la différence existant entre des pauses situées en fin d'unité syntaxique ou incluses dans celle-ci, nous avons étudié la répartition des pauses en fonction des groupes rythmiques, des séquences et des unités de rection. Parmi ces trois variables, c'est la position par rapport aux frontières d'unité de rection, l'unité syntaxique maximale, qui se révèle la plus associée avec la durée ($\chi^2 = 143.28$; $p < 0.001$; $V = 0.42$). Les séquences ($\chi^2 = 90.54$; $p < 0.001$; $V = 0.34$) et les groupes rythmiques ($\chi^2 = 59.25$; $p < 0.001$; $V = 0,27$) se révèlent moins corrélés avec la durée. On peut toutefois s'interroger sur la nature de cette association. En effet, il semble vraisemblable que certaines pauses soient porteuses de fonctions syntaxiques et sémantiques et qu'elles se démarquent de pauses plus mécaniques. Toutefois, la Table 5 nous révèle que seules 17% des pauses "courtes" se situent à des frontières d'unité de rection, alors que 94% des pauses "longues" se rencontrent dans cette position. Il semblerait donc que ce soient les pauses destinées à la respiration qui sont également porteuses des fonctions d'organisation syntaxique et sémantique, tendance déjà rapportée par [11].

Bien que significativement corrélés avec la variable réponse (la durée), les variables *pds*, *urc_pos*, *rgr_pos* et *seq_pos* sont également corrélées entre elles – du moins, en ce qui concerne les trois dernières. C'est pourquoi, nous avons conçu un modèle qui prenne en compte l'ensemble des variables explicatives. La durée étant une variable ordinale à trois niveaux, nous avons utilisé une régression logistique ordinale (cumulative logit) [6]. Cette méthode de régression nous permet d'observer la quantité d'information apportée par chacun des prédicteurs lorsqu'il est pris en compte avec l'ensemble des autres variables explicatives. Les résultats obtenus brillent par leur évidence et par leur reproductibilité au niveau des styles. En effet, pour l'ensemble du corpus, on obtient, pour le modèle complet, un R^2 de 0,603, ce qui signifie que l'ensemble de nos variables prédictives permet d'expliquer 60% de la variabilité de la durée discrétisée. Le modèle est donc plutôt bon. De plus, lorsqu'on entraîne un modèle ne comprenant que *pds* et *urc_pos*, on obtient un R^2 de 0,595, révèlant que la quantité d'information propre des autres variables est quasi nulle.

Notons enfin qu'ajouter une variable *locuteur* au modèle réduit permet d'atteindre un R^2 de 0,64, ce qui démontre qu'une – faible – part des caractéristiques de la durée s'explique par les variations personnelles. Notre corpus ne permet toutefois pas de distinguer si cet effet est lié à l'origine

6. Pour appliquer cette technique statistique décrite par Agresti [1, p. 275], nous avons employé le package Design de R.

géographique ou à une variation propre au locuteur.

5. CONCLUSION

Dans cette étude, nous rapportons, à l'instar d'études précédentes [8, 16], une distribution multimodale pour le logarithme de la durée des pauses en français. Par conséquent, contrairement à ce qui a pu être avancé dans la littérature pour le français, la transformation logarithmique ne permet pas de retomber sur une simple distribution gaussienne de la durée des pauses. Aussi, l'utilisation de méthodes non-paramétriques reste préférable à celle de techniques paramétriques, telles que l'ANOVA ou le test-T de Student. De plus, nous avons montré que la répartition de la durée des pauses varie selon qu'on analyse le corpus globalement ou par style. Ainsi, tandis qu'on observe une trimodalité pour l'ensemble des pauses du corpus, l'analyse par style révèle une tendance à la bimodalité (à l'exception du style JPA).

Nous avons ensuite recherché certains facteurs explicatifs de la durée des pauses. Les éléments les plus fortement corrélés avec la durée des pauses sont la présence ou l'absence de prises de souffle et la position de la pause au milieu ou à la fin d'unités syntaxiques (unités de rection, séquences et groupes rythmiques). Certaines variables, telles que la présence ou l'absence d'hésitation et le débit local, ne s'avèrent pas explicatives. En utilisant, un modèle de régression logistique, nous pouvons conclure que l'ensemble de nos variables prédictives permet d'expliquer 60% de la variabilité de la durée des pauses.

En sus des résultats analytiques obtenus par l'analyse statistique, le recours à un modèle logistique permet également d'envisager une valorisation de ces résultats dans un système de synthèse de la parole multi-styles. Les durées moyennes des différents types de pauses (pauses courtes, moyennes ou longues) peuvent aider à une prédiction plus juste de la durée des pauses en fonction du style de parole à synthétiser. La proportion importante de prises de souffle à travers les différents styles et le taux élevé d'hésitations pour certains styles nous encouragent à intégrer ce type de pauses dans le synthétiseur. Le modèle logistique, quant à lui, n'est pas réutilisable comme tel. En effet, les variables les plus explicatives (*pds* et *urc_pos*) ne sont pas directement observables à partir du système de synthèse. Seule la variable groupes rythmiques peut être utilisée directement.

Outre cette perspective applicative, nous envisageons une observation plus détaillée de nos données. En effet, si quelques résultats obtenus sont éloquents (notamment la forte corrélation entre pds et durée des pauses), certaines pistes restent encore à explorer : quel facteur pourrait expliquer la présence de pauses longues sans prise de souffle ? Comment expliquer l'apparition d'un troisième mode pour certains styles ou pour certains locuteurs ? Quelles variables supplémentaires pourraient être utilisées pour obtenir un meilleur modèle de prédiction ?

6. REMERCIEMENTS

Cette étude est financée par la Région wallonne (projet No.0616422 : « Expressive. Système automatique de diffusion vocale d'information dédicacée : synthèse de la parole expressive à partir de textes balisés »).

RÉFÉRENCES

[1] A. Agresti. *Categorical Data Analysis. 2nd edition*. Wiley-Interscience, New York, 2002.

[2] M. Avanzi, A.-C. Simon, J.P. Goldman, and A. Auchlin. C-PROM : Un corpus de français parlé annoté pour l'étude des proéminences. In *Proceedings of JEP*, Mons, Belgium, mai 2010. http ://sites.google.com/site/corpusprom.

[3] R. Beaufort and A. Ruelle. eLite : système de synthèse de la parole à orientation linguistique. In *Proceedings of JEP*, pages 509–512, 2006.

[4] M. Bilger and E. Campione. Propositions pour un étiquetage en séquences fonctionnelles. *Recherches sur le Français Parlé*, (17) :117–136, 2002.

[5] C. Blanche-Benveniste, M. Bilger, C. Rouget, and K. van den Eynde. *Le français parlé : études grammaticales*. Éditions du CNRS, Paris, 1990.

[6] P. Boersma and D. Weenink. *Praat : doing phonetics by computer (Version 5.1.24)*, 2010. http ://www.praat.org.

[7] R. Boîte, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich. *Traitement de la parole*. Presses Polytechniques et Universitaires Romandes, Lausanne, 2000.

[8] E. Campione and J. Véronis. A Large-Scale Multilingual Study of Silent Pause Duration. In *Proceedings of the Speech Prosody 2002 Conference.*, pages 199–202, Aix-en-Provence, 2002.

[9] M. Candea. *Étiquetage semi-automatique de l'intonation dans les corpus oraux : algorithmes et méthodologie*. PhD thesis, Université de Provence, Aix-en-Provence, 2001.

[10] L. Degand and A.C. Simon. On identifying basic discourse units in speech : theoretical and empirical issues, Discours. *Discours*, 2009. http ://discours.revues.org/index5852.html.

[11] D. Duez. La fonction symbolique des pauses dans la parole de l'homme politique. *Faits de langues*, 13 :91–97, 1999.

[12] C. Fraley and A.E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458) :611–632, 2002.

[13] J.-P. Goldman. *EasyAlign : a semi-automatic phonetic alignment tool under Praat*, 2008. http ://latlcui.unige.ch/phonetique .

[14] F. Grosjean and M. Collins. Breathing, pausing and reading. *Phonetica*, 36(2) :98–114, 1979.

[15] F. Grosjean and A. Deschamps. Analyse des variables temporelles du francais spontané. *Phonetica*, 26(3) :129–156, 1972.

[16] K.M. Rosen. Analysis of speech segment duration with the lognormal distribution : A basis for unification and comparison. *Journal of Phonetics*, 33(4) :411–426, 2005.

[17] B. Zellner. *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* PhD thesis, Faculté des Lettres, Université de Lausanne, Lausanne, Suisse, 1998.

Expressions des états mentaux/cognitifs et affectifs : Prosodie des productions vocales minimales – des *grunt* et *burst* à l'interjection

Vanpé Anne[1,2] *, Aubergé Véronique*[1,2]

[1] GIPSA-lab, Département Parole et Cognition – Université de Grenoble - CNRS
2 Laboratoire d'Informatique de Grenoble, GETALP - CNRS
{anne.vanpe;veronique.auberge}@gipsa-lab.inpg.fr

ABSTRACT

Face to face interaction is a dynamic process where the human is continuously communicating (backchannel and feedback). A large panel of states values is expressed via the multimodal channels: mental/cognitive, emotions, attitudes, intentions. We name it Feeling of Thinking. The present study is specific to the minimal acoustic non lexical productions, from fillers, grunts, bursts to interjections. Following an ethology-based methodology, such productions of six French subjects, emotionally induced in HMI, are investigated for pragmatic cues and prosody. The ratio of pre-lexical *vs.* non phoneme and non phonetic items is studied and related to speaker properties and strategies. Prosodic complexity arise from "mouth noises" to interjections. It points out the role of the prosody control in a possible link to "double articulation" lexical items.
Keywords: affect burst, interjection, filler, expressive prosody, pragmatic prosody, speaker behavior, Feeling of Thinking, feedback, backchannel, face to face interaction.

1. INTRODUCTION

La prosodie d'items lexicaux isolés (interjections, *fillers, grunt,* ou autres « bruits vocaux » émotionnels) intéresse depuis longtemps. Toutefois, ces divers éléments sont étudiés soit pour leur fonction émotionnelle et sont alors considérés comme des éléments particuliers ne faisant pas partie de la parole ; soit pour leurs fonctions pragmatiques, et forment alors une catégorie lexicale [2].

Par ailleurs, à travers la nécessité, pour les technologies du TAL écrit et oral et le domaine de l'informatique affective, de modéliser des interactions écologiques, réalistes et surtout personnifiées, la prosodie a acquit un statut important en interaction [2]. Diverses expressions dans les différentes modalités, et en particulier les productions vocales minimales, ont été observées dans le phénomène de *backchannel* lors du *feedback* de l'auditeur (*cf.* les travaux du réseau Humaine D6d) dans des tâches d'interaction homme/homme ou machine. Ces éléments apparaissent lors des processus cognitifs du sujet liés à la tâche et alors qu'il exprime « en ligne » et en continu dans la modalité visuelle [8] ses états mentaux, émotionnels ou attitudinaux : ce que nous nommons *Feeling of Thinking* - *FoT*. Les objectifs à long terme de ce travail sont nombreux et assez ambitieux. Ce qui est présenté ici en est une étude préliminaire :

- modéliser pourquoi, comment et quand la modalité acoustique est utilisée lors d'une tâche cognitive, mais aussi comprendre dans quels cas elle est liée à l'organisation de l'interaction (*e.g.* le lien entre forme d'un rire et son contexte d'apparition [4]) ou liée à des actes communicatifs.

- la complexité prosodique est-elle liée à une classification phonétique telle que : pré-phone (pas dans l'API) ; pré-phonème (phone de l'API, mais pas dans la langue des locuteurs) ; item phonologique (phonèmes de la langue du locuteur) ? Cette considération est liée à l'apparition de la « double-articulation » pour le mécanisme lexical.

- Quels sont les « niveaux » de complexité prosodique des items relevés, *i.e.* les dimensions contrôlées : qualité de voix/sons, durée /timing, intensité, F0 ? Cette complexité prosodique est-elle liée aux fonctions communicatives et pragmatiques (expressions émotionnelles, d'affects sociaux, d'états mentaux, d'actes de parole) ?

- la « qualité de sons » des items de type « pré-phone »est-elle directement liée au contrôle prosodique de la qualité de voix ? Est-ce qu'elle peut être considérée comme icône expressive (« symbolisme des sons » [5] et *affect bursts* [6] pour les expressions émotionnelles) ?

- quels sont les éléments idiolectaux vs. pragmatiques/du langage (*i.e.* variations intra- vs. inter-sujets) ?

- sous quelles conditions les formes visuelles associés à ces items acoustiques peuvent être reliées aux théories des *appraisals* (Scherer), de la préparation à l'action (Frijda) ou plus généralement des expressions émotionnelles ? Comment sont intégrées les modalités audio et visuelle (redondantes, additives, combinées) ?

Pour amorcer ce travail, nous avons sélectionné 6 sujets (3 hommes et 3 femmes, âgés de 25 à 45 ans, avec un niveau d'éducation moyen à élevé) du corpus émotionnellement induit E-Wiz/Sound Teacher [1]. Ces sujets ont été retenus pour la variabilité de leurs réactions affectives et cognitives aux inductions et de leur personnalité perçue naïvement (d'introverti à extraverti). Ils ont auto-annoté leur propre enregistrement en langage vernaculaire, laissant apparaître des descriptions d'états affectifs et cognitifs complexes (partie 2). De plus, le signal audio-visuel a été étiqueté par deux experts en suivant une méthodologie issue de l'éthologie (partie 3). Pour tous les sujets, l'interaction est surtout composée de communication visuelle complétée par des productions acoustiques pré-lexicales, et non de parole. La partie 4 décrit enfin ce qui nous amène à étudier les productions vocales minimales et en fait une première analyse pragmatique et phonétique.

2. Du *Feeling of Knowing* au *Feeling of Thinking*

Cette étude porte sur l'observation du corpus spontané d'interaction personne-machine E-Wiz/Sound Teacher (*cf.* [1] pour la description complète), avec pour tâche prétexte l'apprentissage phonétique de sons des langues du monde. Un paradigme de magicien d'Oz a permis d'influencer les états des sujets, en leur donnant régulièrement de faux résultats : phase 1-bons résultats ; phase 2-très bons résultats ; phase 3-mauvais résultats ; phase 4-très mauvais résultats et avertissement d'une possible dégradation de ses capacités phonétiques. Lors de la tâche, les sujets sont en train de lire, de réfléchir ou de produire de la parole. Leur productions de parole peuvent être des réponses monosyllabiques (isolés) ou des commentaires libres. L'interaction n'est pas humanisée (tâche directe) et le sujet n'a ainsi aucune raison d'envoyer des expressions destinées à influencer les états de la machine : des expressions d'états mentaux, d'humeurs et d'émotions sont ainsi attendues, à l'inverse de celles d'attitudes ou d'intentions, pourtant présentes et décrites dans les auto-annotations des sujets.

Dans une étude pilote, Swerts *et al* [7] ont décrit une tâche mnésique lors de laquelle souvent, le sujet « sentait qu'il savait » la réponse et l'exprimait de manière multimodale. C'est ce qu'ils ont appelé « Feeling of Knowing », *FoK*. Dans notre expérience, un large panel d'états mentaux et affectifs, ainsi que de processus cognitifs (y compris des décisions de prendre la parole) est exprimé : par analogie au *FoK*, nous utilisons l'expression « Feeling of Thinking »-*FoT* [4][8] pour désigner ces phénomènes.

3. Une méthodologie générale issue de l'éthologie

Étiqueter les formes des expressions, autant qu'annoter la valeur de ces expressions sont des points fondamentaux de notre méthodologie. Ces tâches peuvent être guidées par une théorie, ou à l'inverse, guidées par les données dans une approche inductive.

L'auto-annotation naïve nous permet de nous échapper des pré-supposés théoriques difficilement évitables lors de la description de valeurs d'émotions, états mentaux, etc. Ces auto-annotations sont ainsi destinées à être des éléments d'échange « naturels » lors de tests perceptifs auprès de juges naïfs et ont ainsi été utilisées lors de la validation perceptive de l'étiquetage du signal audio-visuel [8].

En parallèle, l'étiquetage des mouvements de la face et du buste, ainsi que de la parole a été réalisé par deux chercheurs « naïfs », sans aucune hypothèse de théorie du geste ou des émotions, ou de pragmatique. Ils ignoraient tout des sujets, de la tâche, des auto-annotations et du contexte d'induction. Leur contrainte était de « découper » le plus finement possible les signaux audio et visuels (« Icônes minimales Gestuelles/ Vocales »), sans essayer de deviner l'information affective et cognitive pouvant être sous-jacente des signaux.

Même si certaines études étudient déjà en détail des mouvements significatifs subtiles, en particulier pour la face, la plupart de ces modèles semblent a priori insuffisants pour décrire certains événements pouvant être rencontrés dans notre corpus.

Ainsi, un protocole issu de l'éthologie, c'est à dire utilisant des éthogrammes (inventaire des comportements d'une espèce) a d'abord été appliqué pour étiqueter notre corpus. Chaque étiquette a été définie selon : des critères de direction, localisation, vitesse, durée, symétrie, répétition, intensité et/ou amplitude pour les événements faciaux / gestuels ; des critères phonétiques et acoustiques pour les événements vocaux (voir 4.2). Un consensus systématique a été établi entre les deux chercheurs, commençant par l'utilisation d'« icônes » primitives communes pour les mouvements du buste, de la face, et les événements vocaux. L'éditeur d'annotation de vidéo utilisé pour cette étape de notre étude a été ANVIL[1], développé par Kipp au DFKI.

4. Première analyse des événements vocaux

4.1. Des événements visuels aux événements vocaux

A la suite de l'annotation, la pertinence des paramètres visuels étiquetés a été validée par des tests perceptifs [8]. Nous avons entre autres étudié dans quelle mesure l'information de *FoT* est portée par le statique ou la dynamique. Les résultats ont montré que globalement, même s'il n'est pas écologique de percevoir une image statique, le bénéfice d'information du statique au dynamique dépendait de la nature de l'information, puisque la dynamique a, dans certains cas, perturbé l'information de *FoT* apportée par le stimulus. De la même manière que pour les images statiques, une vidéo sans signal audio n'est pas complète d'un point de vue écologique. D'autant plus que les événements vocaux semblent être d'une grande importance dans l'expression du *FoT*.

Par ailleurs, l'amplitude des mouvements ne semble pas pouvoir être directement considérée comme un paramètre mais doit être reliée au comportement du locuteur : *e.g.* les mouvements peu amples d'un sujet introverti peuvent être très informatifs. Par analogie, qu'en est-il de l'audibilité des événements vocaux?

En parallèle, certains gestes ou événements vocaux sans valeur affective intrinsèque, pourraient voir cette dernière révélée à travers leur motif temporel, leur rythmicité au cours de l'interaction. Il s'agirait alors de considérer non plus des événements ponctuels, mais le comportement global du locuteur, comme l'ont montré Carlier *et al* [3] concernant le comportement de joueurs de tennis.

Finalement, dans une étude préliminaire [4], il a été montré que les sujets sont capables de replacer un rire dans un énoncé, en terme d'organisation (*e.g.* choisir si le rire apparaît au début, milieu ou fin d'énoncé). Ainsi, un rire

[1] Documentation complète sur ANVIL : Kipp, M., 2004, "Gesture Generation by Imitation: From Human Behaviour to Computer Character Animation", Boca Raton, Florida, Dissertation.com.

pourrait soit indiquer le contexte, soit être intrinsèquement contextuel. Et les interjections et *bruits de bouche* ?

Les différentes perspectives apportées par ces questions (personnalisation du comportement, motif temporel, pragmatique et organisation du comportement) sont ainsi à étudier pour les événements vocaux, en complément d'analyses acoustiques / prosodiques. Dans notre corpus, les événements vocaux sont présents en grand nombre et peuvent être très différents de part leurs caractéristiques acoustique / articulatoire. Comme pour les événements visuels, nous voulons savoir quels paramètres sont communicativement pertinents pour l'expression du *FoT*. C'est pourquoi dans un premier temps, nous ne négligeons aucun paramètre.

4.2. Critères pour l'étiquetage

Pour étiqueter ces événements vocaux, nous avons d'abord distingué les interjections (éléments pré-lexicaux construits avec des phonèmes de la langue et dont il existe le plus souvent une transcription conventionnelle dans les dictionnaires de la langue), des *bruits de bouche* (construits de *phones vs. non-phones* : *bursts*, bruits de respiration, clicks, etc.). Ainsi, la signification socialement acceptée, en lien avec l'orthographisation, devient ici un critère pour la classification. Nous pouvons même observer que certains événements vocaux sont produits avec des variations phonologiques (*e.g.* l'interjection française orthographisée « pfff » peut être produite [pf:] ou avec un trill bilabial). Nous utilisons ici cette dichotomie interjection *vs. bruits de bouche*, même si notre but à long terme est d'ordonner tous ces items non-lexicaux (de la « deuxième articulation ») sur un continuum où leur lexicalisation et leur contrôle prosodique augmentent : bruit comme qualité de voix par lui-même + contrôle de la durée + contrôle de l'intensité / de la F0. Le contrôle des phones commencerait ainsi par le contrôle prosodique significatif.

Ensuite, pour la présente étude, nous avons classé les interjections selon des critères phonétiques/articulatoires : phonème vocalique (étiqueté « V », *e.g.* [ø:] « euh »), phonème consonantique (« C », *e.g.* [m:] « mmh ») ou combinaison de phonèmes vocaliques et consonantiques (« CV », *e.g.* [bø:] « beuh », et « VC », *e.g.* [ø:m:] « euh mmh »). Nous avons étiqueté « Comb » (« combinaison ») les interjections composées de plus d'un phonème vocalique ou consonantique (*e.g.* [bɛ̃m:] « ben mmh », [ula] « ouh là »). Nous avons donc autant étiqueté des événements vocaux complexes, que des plus simples.

De l'autre côté, seules deux distinctions fondées sur des paramètres articulatoires ont été ici utilisées pour classer les événements vocaux autres que les interjections, que nous nommons globalement « bruits de bouche » : s'ils sont produits par le sujet pendant une inspiration *vs.* pendant une expiration ; s'ils sont produits avec un flux d'air continu (*e.g.* grande inspiration, gémissement) *vs.* avec un flux d'air gêné ou bloqué au moins une fois (*e.g.* click, occlusion glottale, occlusion bilabiale). L'intérêt de ce dernier critère est qu'il implique une tension suivie d'un relâchement du sujet. Nous avons également relevé deux autres types d'items, « mouillés »: liés soit à une interaction entre langue et lèvres, soit à une déglutition.

Par ailleurs, nous avons classé chaque événement vocal selon la position de son occurrence par rapport à la production de parole. Nous l'avons noté : « 1 » lorsqu'il apparaît pendant la production de parole, « 0 » en dehors, « B » moins d'une seconde avant, « A » moins d'une seconde après, et « ~ » lorsqu'il apparaît pendant la production de parole, mais en dehors du flux.

4.3. Effets inter-sujets et analyse globale

Aucune corrélation n'apparaît entre le nombre global d'événements vocaux et la réaction des sujets aux inductions de l'expérience (Figure 1). Toutefois leur taux relatif par sujet pendant les trois dernières phases de l'expérience laisse apparaître deux groupes de sujets fonction de leur comportement (M_ et F_ respectivement pour sujet mâle et femelle) : F_T, M_N, M_J, qui se sont réellement inquiétés sur la dégradation de leurs capacités, et F_S, F_M, M_R, qui ont plutôt remis en question le logiciel. La phase 1, étant une phase d'apprentissage de la tâche sans induction émotionnelle recherchée, sépare quant à elle les stratégies des sujets par rapport à la tâche : F_T, F_S, M_R d'un côté *vs.* F_M, M_N, M_J de l'autre. Ces groupes ne peuvent reliés à l'âge, au sexe, au niveau d'éducation, ni la familiarité du sujet avec les IHM.

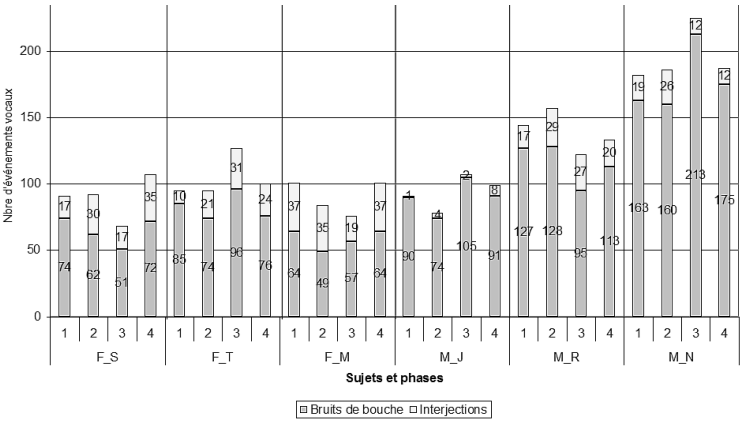


Figure 1. Comparaison inter-sujets du nombre d'événements vocaux (*bruits de bouche* et interjections) pendant les différentes phases du scénario.

Le nombre d'événements vocaux, et surtout de *bruits de bouche*, des hommes est globalement plus élevé que pour les femmes. Cela peut être relié à la quantité de parole (commentaires libres) des sujets. En effet, les trois femmes parlent plus longtemps et utilisent plus d'interjections pré-lexicales par rapport aux autres bruits de bouche, mais produisent globalement moins d'items non-lexicaux. Au contraire, les hommes parlent moins et semblent favoriser la production d'items non phonémiques et non phonétiques. Bien sûr, cette distinction homme/femme ne peut être généralisée avec six sujets, mais cette « tendance » au langage, avec peu de bruits de bouche et une proportion élevée d'interjections, pourrait peut-être être évaluée en tant qu'élément d'une stratégie communicative particulière.

4.4. Nature des interjections et « bruits de bouche »

Les répartitions temporelles des différents types d'interjections (cf. 4.2) montrent de grandes divergences entre elles. Les grandes tendances montrent toutefois : que la répartition par type des interjections consonantiques et celles plus complexes diffèrent selon les sujets et selon les phases du scénario ; que les interjections sont surtout des phonèmes vocaliques. Plus précisément, l'interjection la plus fréquente pour tous les sujets est l'item vocalique [ø:].

Table 1. Comparaison inter-sujets du nombre de *bruits de bouche* selon leur nature, et fonction de leur position temporelle par rapport à la production de parole.

Femelles	Inspiration						Expiration						Déglutit			Interaction langue-lèvres		
	continue			bloquée			continue			bloquée								
Sujet F_	S	T	M	S	T	M	S	T	M	S	T	M	S	T	M	S	T	M
0	3	9	5	23	43	22	6	21	3	11	19	10	17	12	11	9	8	7
1	1	0	1	0	0	1	15	44	1	1	14	2	0	0	0	0	0	0
~	36	40	21	24	19	30	7	3	1	13	4	3	0	1	9	0	0	2
Before	7	11	4	50	56	65	1	2	4	0	8	8	4	5	10	2	2	0
After	9	4	3	3	1	5	1	5	2	4	5	6	12	0	7	1	1	4
Total	**56**	**64**	**34**	**100**	**119**	**123**	**30**	**75**	**11**	**29**	**50**	**29**	**33**	**18**	**37**	**12**	**11**	**13**

Mâles	Inspiration						Expiration						Déglutit			Interaction langue-lèvres		
	continue			bloquée			continue			bloquée								
Sujet M_	J	R	N	J	R	N	J	R	N	J	R	N	J	R	N	J	R	N
0	17	22	69	48	85	98	17	40	27	22	19	52	4	28	44	5	20	7
1	1	1	1	2	1	24	99	5	63	6	0	18	0	0	1	0	1	0
~	6	37	23	13	20	13	2	1	10	1	6	6	0	3	6	0	0	1
Before	31	33	42	48	88	113	3	17	5	8	3	6	0	5	10	3	5	10
After	15	8	7	4	3	8	4	11	16	14	7	34	3	7	3	2	0	0
Total	**70**	**101**	**142**	**115**	**197**	**256**	**125**	**74**	**121**	**51**	**35**	**116**	**7**	**43**	**64**	**10**	**26**	**18**

Comme pour les interjections, la répartition des *bruits de bouche* selon leur nature (*cf.* 4.2) a été différente d'un sujet à l'autre. De plus, une autre répartition semble être particulièrement pertinente pour les *bruits de bouche* : leur position temporelle par rapport aux productions de parole. Les analyses montrent que les sujets utilisent différemment, en termes de répartition, les *bruits de bouche* inspirés et expirés. De la même manière, la répartition des bruits de bouche selon leur type de flux d'air (cf. 4.2, « continue » *vs.* bloqué ou gêné -« bloquée »-) diffère selon les sujets (Table 1) : par exemple alors que pour tous les sujets nous relevons plus de *bruits de bouche* « bloqués » en inspiration, ceux produits en expiration sont plutôt « continus » pour F_T, M_R et M_J et bloqués ou répartis pour les autres sujets ; de plus, les hommes s'opposent aux femmes par leur plus grande production d'inspiration, en particulier produits en dehors de la production de parole.

A la suite de cette étude préliminaire, nous nous focalisons maintenant sur les relations entre types d'induction, tours de parole et productions d'événements vocaux.

5. Conclusion et perspectives

Avec une approche empirique et inspirée de l'éthologie, certains éléments de comportement ont été établis à travers l'observation de 6 sujets aux personnalités différentes, issus du corpus expressif spontané IHM Sound Teacher.

La relation entre éloquence/durée de parole et choix d'items pré-lexicaux phonologisés semble apparaître comme une « tendance au langage » : les sujets utilisant le moins la parole produisent le plus d'items non phonologisés, mais porteurs de prosodie. La distinction homme/femme pour cette tendance ne sera certainement pas systématique, mais élargir à plus de sujets l'étude de cette tendance comme stratégie communicative particulière serait intéressant.

La nature des différents bruits de bouche montre que pour les items non-phonétiques, la prosodie peut apparaître « seule », sans « double-articulation », et que la « qualité de sons » est la qualité de voix. Cette observation peut être à la fois rapprochée de l'idée de « symbolisme des sons » et des motivations physiologiques pour l'expression des émotions primaires [5]. La relation au *FoT* est montrée par la stratégie utilisée par les locuteurs. Il reste à étudier plus finement les relations entre la nature des items produits, et *FoT* et personnalité. Comme il l'a été montré pour les rires [4], la cohérence entre timing pragmatique et nature des items est claire et doit maintenant être expliquée.

Ces éléments de la continuité entre prosodie « pure » et « double-articulation » sont très dynamiques en *feedback*. L'utilisation de ce type d'outil communicatif en interaction pourrait être une clé pour mieux comprendre le fondement de la prosodie. Étudier pourquoi, comment et où les différents niveaux de complexité de ces items apparaissent à l'intérieur des expressions visuelles nous en donnera certainement une trace, et il s'agit de nos travaux actuels.

Bibliographie

[1] V. Aubergé, N. Audibert et A. Rilliard. De E-Wiz à C-Clone. Recueil, modélisation et synthèse d'expressions authentiques. *Revue d'Intelligence Artificielle -"Interactions émotionnelles"*, 20(4-5), 499-528, 2006.

[2] N. Campbell. Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language. *Languages Resources and Evaluation*, 39, 109-118, 2004.

[3] G. Carlier and C. Graff. Unpredictability as a counter strategy: An analysis of elite matches. *Journal of Sport Sciences*, to be published, 2006.

[4] F. Loyau. *Expressions des états mentaux et émotionnels de l'humain en interaction : ébauches du "Feeling of Thinking"*. Thèse de Doctorat en Sciences Cognitives. INP Grenoble, 2007.

[5] J.J. Ohala. The frequency codes underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols & J.J. Ohala (Eds.), *Sound symbolism*, Cambridge University Press 325-347, 1994.

[6] K.R. Scherer. Affect bursts. In S.H.M. van Goozen, N.E. van de Poll & J.A. Sergeant (Eds.), *Emotions*, Hillsdale, NJ, Lawrence Erlbaum, 161-193, 1994.

[7] M. Swerts and E. Krahmer. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1), 81-94, 2005.

[8] A. Vanpé et V. Aubergé. Prosodie visuelle expressive du *Feeling of Thinking*. Perception statique ou dynamique ?. *WACA*, Paris, 2008.

Reconnaissance automatique du locuteur embarquée dans un téléphone portable

Anthony Larcher, Christophe Lévy, Driss Matrouf, Jean-François Bonastre

Université d'Avignon et des Pays de Vaucluse,
Laboratoire Informatique d'Avignon (UPRES 931),
F-84018 Avignon, France
{anthony.larcher, christophe.levy, driss.matrouf, jean-francois.bonastre}@univ-avignon.fr

ABSTRACT

Embedded speaker recognition in mobile devices involves a limited amount of computing resources. However, the performance of state-of-the-art systems are usually evaluated without any limitation of material resources. In this paper, we evaluate several downscaled configurations of the LIA UBM/GMM speaker verification system. The impact of scalability is evaluated in terms of memory resources and computational time. We propose two downscale configurations which allow a good compromise between resource consumption and performance degradation. Experiments performed on the Banca database show that the memory allocation and the computational time decrease of about 70% and 88% when the error rate raises from $3,48\%$ to $4,77\%$ or still comparable to the baseline.

Keywords: speaker recognition, scalability, embedded systems

1. Introduction

Au regard des dernières campagnes d'évaluation internationales organisées par le NIST [7], la Reconnaissance Automatique du Locuteur (RAL) a réalisé de réels progrès ces dernières années. Ces derniers sont dus à l'introduction de nouvelles méthodes telles que le Factor Analysis ([5], [6]) ou la Nuisance Attribut Projection ([8]) qui ont notamment permis de compenser les dégradations induites par l'usage de différents microphones et/ou de différentes liaisons téléphoniques, ou encore de limiter l'influence de la variabilité due à l'environnement acoustique dans lequel est utilisé le système. Ces avancées, ainsi que la maturité atteinte par les systèmes actuels, rendent envisageable le développement d'applications grand-public sécurisées par la biométrie vocale.
La multiplication des systèmes embarqués, observée ces dernières années (téléphones portable, GPS, PDA, ...), a ouvert de nouveaux marchés. Cependant, le développement du matériel et des réseaux s'est accompagné d'une croissance équivalente du nombre d'applications disponibles sur ces terminaux. Aussi, les contraintes matérielles induites par ces clients légers constituent une problématique majeure pour l'intégration des technologies de reconnaissance du locuteur.
Cet article propose une étude des performances d'un système état-de-l'art dans un contexte de ressources matérielles limitées. Nous considérons l'influence des principaux paramètres scalables qui peuvent être modifiés afin de satisfaire aux ressources restreintes d'un système embarqué. Différentes configurations du système UBM/GMM développé à partir de la plateforme Mistral (évalué lors des principales évaluations internationales de vérification du locuteur [2]) sont comparées en fonction de leurs performances, mais également en fonction des ressources qu'elles nécessitent tant en termes de temps CPU que d'occupation mémoire.
Dans un premier temps, nous donnons un exemple de la capacité des téléphones portables actuels. Ensuite, le système de reconnaissance développé à partir de la plateforme biométrique MISTRAL[1] [3] est présenté dans la section 3. La partie 4 contient les détails concernant le protocole expérimental utilisé. Les paramètres scalables du système ainsi que les performances obtenues sont commentés dans la partie 5. De cette analyse, deux configurations particulières du système de vérification du locuteur ont été sélectionnées : une solution nécessitant le moins de ressources possibles mais qui a des performances non optimales et une configuration représentant le meilleur compromis. Enfin, la dernière partie présente quelques conclusions issues de ce travail.

2. Exemple de ressources

Cette étude se situe dans le contexte du projet européen MOBIO[2] au cours duquel un système de vérification du locuteur doit être implémenté sur un téléphone portable. Le portable choisi pour cette intégration est le Nokia© N900.

vendor id	ARM
model name	Cortex-A8
CPU MHz	600
RAM	256 MB
CPU cores	2

Tab. 1: Caractéristiques du téléphone Nokia© N900 sur lequel doit être intégré le système de vérification du locuteur dans le projet européen MOBIO.

Les ressources présentées dans le tableau 1 correspondent à celles disponibles pour l'ensemble des applications actives sur le téléphone à un instant donné et non pas aux ressources disponibles pour la seule application de vérification du locuteur.

3. Description du système

Les différentes configurations du système UBM/GMM présentées par la suite sont basées

[1] http://mistral.univ-avignon.fr/
[2] http://www.mobioproject.org/

sur le système développé au sein du Laboratoire Informatique d'Avignon (LIA) à partir de la plateforme biométrique MISTRAL [3]. Les performances de ce système, évaluées lors des dernières campagnes internationales de vérification du locuteur NIST-SRE [2], le placent au niveau des systèmes état-de-l'art.

Ce système est basé sur le paradigme UBM/GMM associé au Latent Factor Analysis [6] ; ce qui permet de prendre en compte la variabilité session intra-locuteur. Les paramètres acoustiques utilisés sont issus d'une analyse cepstrale en banc de filtres (obtenus avec SPRO [4]). Les vecteurs de paramètres utilisés sont composés des 19 coefficients statiques (c), des 19 coefficients dérivés du premier ordre (Δc), de 11 coefficients dérivés du second ordre ainsi que du coefficient dérivé (du premier ordre) de l'énergie (ΔE). Une sélection des trames de plus haute énergie est appliquée de façon standard (pour séparer les trames *parole* des trames *non-parole*) avant d'appliquer une normalisation (soustraction de la moyenne et normalisation de la variance) au niveau de chaque fichier.

Le modËèle du monde (UBM) comprend 512 distributions Gaussiennes et la matrice de covariance est supposée diagonale. Chaque modèle de locuteur est dérivé de l'UBM par une adaptation basée sur le critère du Maximum A Posteriori (MAP). Il faut enfin noter qu'en raison du contexte (RAL embarquée), le choix a été fait de ne pas normaliser les scores. En effet, les normalisations classiques, type ZTNorm, sont consommatrices de temps CPU et d'espace mémoire.

4. Protocole expérimental

Les différentes configurations présentées dans ce papier ont été évaluées sur la base de données Banca [1]. Cette base comprend les enregistrements vidéo de 52 personnes réparties en deux sous-groupes G1 et G2 contenant chacun 13 hommes et 13 femmes.

Chaque locuteur présent dans la base BANCA a participé à 12 sessions d'enregistrement dans différentes conditions avec différentes caméras. Les quatre premières sessions sont enregistrées avec une luminosité et un environnement sonore contrôlés. Les sessions 5 à 8 et 9 à 12 correspondent respectivement à des conditions « dégradées » et « adverses ». Ces différentes conditions d'enregistrement permettent d'obtenir une plus grande variabilité au niveau de l'arrière plan sonore.

Cette étude, préalable à l'implémentation du système de reconnaissance du locuteur au sein du téléphone portable, a été réalisée sur un ordinateur et non sur le téléphone[3]. Son but n'est pas seulement d'évaluer les taux d'erreurs du système de RAL, mais bien de les rapprocher de l'occupation mémoire et du temps de calcul consommés. Les caractéristiques du PC utilisé sont présentées dans le tableau 2.

Le temps de calcul évalué avec la commande système `time` a permis de déduire le temps CPU nécessaire à l'exécution du processus (le chargement des différents modèles et des données est inclus dans le temps mesuré). La quantité de mémoire nécessaire a été estimée en utilisant le profileur Valgrind qui permet de mesurer le pic de mémoire allouée. Dans le cadre d'une implémentation sur système embarqué, la notion de maximum de mémoire allouée à un instant t est plus importante que la taille totale allouée. Enfin, l'estimation des ressources nécessaires a été réalisée à partir d'un sous-ensemble représentatif des tests de la base de données BANCA (les taux d'erreurs, eux, sont donnés pour l'ensemble du corpus).

[3]Le Nokia© N900 n'est pas encore sur le marché à l'heure actuelle, mais de premiers essais ont pu être réalisés par l'intégrateur partenaire du projet.

vendor id	GenuineIntel
model name	Intel(R) Core(TM)2 Quad CPU Q9300
CPU MHz	2000
cache size	3072 KB
CPU cores	4 **(1 seul est utilisé)**
bogomips	4987.44

Tab. 2: Caractéristiques du PC sur lequel ont été effectués les tests présentés dans cet article.

5. Paramètres scalables

Cette partie présente l'influence de trois paramètres scalables sur les performances du système :
- le nombre de distributions Gaussiennes des GMM ;
- la taille des vecteurs acoustiques ;
- la proportion d'échantillons traités en phase de test.

Ces trois paramètres, choisis car ils impactent fortement les besoins du système en termes de mémoire et de temps de calcul, influent sur deux facteurs directement liés aux ressources utilisées :
- le nombre de paramètres à stocker pour chaque modèle, exprimé par :

$$nb_{param} = nb_{gauss} \times (2 \times tailleVectAc + 1) \quad (1)$$

où nb_{param} correspond au nombre de paramètres à stocker pour chaque modèle GMM (UBM + locuteurs), nb_{gauss} est le nombre de composantes Gaussiennes du modèle et $tailleVectAc$ est la taille des vecteurs de paramètres utilisés ;
- le nombre de fonctions de log-vraisemblance (qui constituent l'essentiel du coût de calcul), donné par :

$$nb_{log-vrais} = 2 \times nb_{gauss} \times tailleVectAc \times nb_{trames} \quad (2)$$

où $nb_{log-vrais}$ est le nombre de fonctions de log-vraisemblance à calculer et nb_{trames} correspond au nombre de trames traitées de la séquence de test.

5.1. Nombre de composantes du modèle du monde

Au regard des équations 1 et 2, le nombre de composantes Gaussiennes du modèle du monde détermine de façon explicite la mémoire nécessaire au stockage des modèles de locuteurs. Le tableau 3 présente les résultats obtenus pour des modèles comprenant de 512 à 32 distributions Gaussiennes.

La réduction du nombre de distributions des modèles entraîne une augmentation du taux d'égales erreurs (EER) mais permet de réduire significativement le temps de calcul et la mémoire nécessaire. L'utilisation de modèles à 128 Gaussiennes permet par exemple, de conserver des performances comparables à celles

du système de référence, tout en divisant par 3 la mémoire allouée par le système et par 4 le temps de calcul. Le système avec 32 Gaussiennes permet, lui, de réduire le temps de calcul par un facteur 14 et les allocations mémoire par 5 tout en conservant un taux d'erreurs compris entre 4% et 5%.

# distributions	**512**	**256**	**128**	**64**	**32**
G1 (EER %)	3,48	2,19	3,86	4,23	5,15
G2 (EER %)	2,94	3,32	3,32	2,19	3,85
mem. (MB)	7,84	4,29	2,70	1,90	1,50
mem. rel. (%)	100	57	36	25	20
temps CPU (s)	2,06	1,07	0,53	0,27	0,15
temps rel. (%)	100	52	25	13	7

Tab. 3: Évolution des performances (EER) et des ressources (temps CPU et mémoire) utilisées en fonction du nombre de composantes du modèle. La consommation des ressources est donnée de manière relative par rapport au système de base.

5.2. Taille du vecteur acoustique

La dimension des vecteurs acoustiques utilisés est en lien direct avec l'espace mémoire et le temps de calcul nécessaires puisqu'elle apparaît dans les équations 1 et 2. Différentes dimensions de vecteurs acoustiques ont été testées. Ces configurations diffèrent également par la nature des coefficients utilisés (c, Δc, $\Delta\Delta c$ ou ΔE). Considérant le nombre très important des combinaisons possibles, nous avons choisi de ne retenir que quatre configurations (en plus de celle de référence), détaillées dans le tableau 4. Ce tableau présente les performances du système de vérification du locuteur utilisant ces différentes paramétrisations comparées au système de référence.

# paramètres	**50**	**41**	**30**	**25**	**20**
# c	**19**	**15**	**10**	**15**	**10**
# Δc	**19**	**15**	**10**	**10**	**10**
# $\Delta\Delta c$	**11**	**11**	**10**		
# ΔE	**1**				
G1 (EER %)	3,48	4,77	3,48	5,52	5,15
G2 (EER %)	2,94	3,85	4,23	3,85	4,23
mem. (MB)	7,84	6,60	5,48	4,99	4,46
mem. rel. (%)	100	88	73	67	60
temps CPU (s)	2,06	1,92	1,80	1,79	1,71
temps rel. (%)	100	95	87	86	83

Tab. 4: Évolution des performances (EER) et des ressources (temps CPU et mémoire) utilisées en fonction de la taille du vecteur acoustique (512 distributions par GMM). La consommation des ressources est donnée de manière relative par rapport au système de base.

La réduction de la dimension des vecteurs acoustiques réduit de manière significative l'espace mémoire utilisé ainsi que le temps de calcul nécessaire à la vérification du locuteur. À nombre de paramètres équivalent, il semble que la suppression des coefficients dynamiques $\Delta\Delta c$ entraîne une dégradation importante du taux d'égales erreurs ; en effet le taux d'erreurs moyen (sur G1 et G2) passe de 3,85% pour le système 30 coefficients ($10c + 10\Delta c + 10\Delta\Delta c$) à 4,65% pour le système 25 coefficients ($15c+10\Delta c$), ce qui représente une augmentation relative du taux d'erreurs de plus de 20%.

5.3. Sélection de trames

Les paramètres acoustiques fournis au système de reconnaissance du locuteur sont extraits de façon classique à la fréquence d'une trame toutes les 10ms. Dans cette sous-partie, la réduction est opérée sur la proportion de vecteurs acoustiques qui sont traités par le système pour le calcul de la vraisemblance. Il est important de noter que même si tous les vecteurs ne sont pas utilisés pour calculer le score du test, tous ces vecteurs doivent être extraits. En effet, un sous-échantillonnage réalisé durant la phase d'extraction des paramètres ne permettrait plus de calculer les paramètres dynamiques (Δc et $\Delta\Delta c$).

% de trames traitées	**100**	**50**	**25**
G1 (EER %)	3,48	3,86	4,23
G2 (EER %)	2,94	3,48	4,60
mem. (MB)	7,84	7,84	7,84
mem. rel. (%)	100	100	100
temps CPU (s)	2,06	1,20	0,68
temps rel. (%)	100	58	33

Tab. 5: Évolution des performances (EER) en fonction du nombre de trames utilisées pour la calcul de la vraisemblance du modèle de locuteur. La consommation des ressources (temps CPU et mémoire) est donnée de manière relative par rapport au système de base.

Les résultats présentés dans le tableau 5 sont obtenus en ne traitant qu'une trame sur deux ou une trame sur quatre. Le traitement d'une trame sur n permet de réduire le temps de calcul de façon considérable. Cependant, les performances du système sont fortement dégradées si le ratio entre les trames reçues et les trames traitées atteint 0,25. Dans ce cas, le taux d'erreurs moyen sur G1 et G2 passe de 3,21% (avec 100% des trames traitées) à 4,41% (avec un quart des trames traitées) soit une augmentation relative de près de 40%.

Toutefois, la sélection ici opérée consiste en un sous-échantillonnage régulier et pourrait sans doute être améliorée par l'utilisation de critères de sélection plus pertinents. Une sélection périodique présente, néanmoins, l'avantage de ne nécessiter aucune analyse des vecteurs de paramètres à traiter.

5.4. Conclusion

Dans les trois sous-parties précédentes, nous avons étudié l'influence qu'avait, indépendamment les uns des autres, trois paramètres : le nombre de composantes des modèles GMM, la taille du vecteur acoustique et le nombre de trames acoustiques traitées.
Deux configurations, correspondant à 2 situations distinctes, sont détaillées ci-après :

système minimal cette solution correspond à celle qui nécessiterait le moins de ressources ;

meilleur compromis cette configuration est celle qui représente le meilleur compromis entre les ressources matérielles et les performances.

Les performances relatives à chaque configuration sont présentées dans le tableau 6.

% Systèmes	**ref.**	**minimal**	**compromis**
G1 (EER %)	3,48	7,72	4,77
G2 (EER %)	2,94	7,34	2,94
mem. (MB)	7,84	1,37	2,19
mem. rel. (%)	100	17	28
temps CPU (s)	2,06	0,04	0,24
temps rel. (%)	100	1,7	11,7

Tab. 6: Comparaison des performances obtenues par le système LIA de référence, sa configuration minimale et le meilleur compromis. La consommation des ressources (temps CPU et mémoire) est donnée de manière relative par rapport au système de base.

Système minimal Le système minimal correspond à celui pour lequel les valeurs minimales de chaque paramètre ont été choisies, *i.e.* les modèles GMM comprennent 32 gaussiennes, les vecteurs acoustiques contiennent 20 coefficients ($20c$ et $20\Delta c$) et la vraisemblance est estimée en utilisant une trame sur quatre. Dans cette configuration, le système a des performances nettement en deçà du système de référence. Nous pouvons noter une augmentation relative du taux d'erreurs moyen de 130% ; cependant le taux d'erreurs reste proche de 7% ce qui en fonction de l'application finale choisie peut s'avérer suffisant. Dans le même temps, cette solution nécessite beaucoup moins de ressources. En effet, le temps CPU est divisé par 60 et le pic mémoire passe de 7,84Mo à 1,37Mo soit une baisse relative de 83%.

Meilleur compromis Pour le projet MoBio, l'intégration est prévue dans le téléphone Nokia© N900, pour lequel les ressources disponibles sont supérieures à celles requises par le système minimal. Un compromis entre les 3 paramètres a donc été choisi afin d'obtenir un système ayant de meilleures performances que le système minimal tout en nécessitant moins de ressources. Dans cette configuration, les modèles GMM contiennent 128 composantes Gaussiennes, les vecteurs acoustiques sont composés de 30 coefficients (10 statiques, 10 dynamiques de premier ordre et 10 dynamiques de second ordre) et la vraisemblance est estimée en utilisant une trame sur deux. Ce système obtient des performances satisfaisantes. En effet, le taux d'erreurs moyen passe de 3,21% à 3,84%, soit une hausse relative de 20% alors que dans le même temps les ressources requises sont nettement diminuées : le temps CPU est divisé par 10 et le pic mémoire est divisé par 4.

6. Conclusion et perspectives

Au cours de ces dernières années, les systèmes de reconnaissance automatique du locuteur ont fait des progrès significatifs, à tel point que des applications grand-public deviennent envisageables. Ces applications soulèvent, dans le contexte de l'embarqué, de nouvelles problématiques jusqu'alors peu considérées dans les approches classiques (où les ressources ne sont généralement pas limitées). Dans ce travail, nous avons proposé deux configurations du système UBM/GMM de vérification du locuteur développé au LIA. Ces configurations, faisant varier trois des paramètres majeurs des systèmes de reconnaissance du locuteur, sont adaptées au contexte embarqué et particulièrement aux ressources restreintes.
La première configuration proposée nécessite uniquement 28% de l'espace mémoire et 12% du temps CPU utilisés par le système état de l'art du LIA (reposant sur la plateforme Mistral) alors que le taux d'erreurs n'augmente que de 20% (restant sous le seuil des 4%). Une seconde solution, plus « extrême », a été proposée pour des applications moins sécurisées. Cette configuration obtient un taux d'erreurs inférieur à 8% mais ne nécessite que 17% de l'espace mémoire et 1,7% du temps CPU requis par le système de référence.
De futurs travaux viseront à développer un système scalable dynamique, s'adaptant aux ressources disponibles sur le téléphone portable à un instant donné tout en garantissant les meilleures performances possibles.

Remerciements

Cette étude a été en partie financée par le projet européen MoBio[4]. Ce projet a pour objectif le développement d'une solution pour la biométrie multi-modale embarquée sur un téléphone portable.

Références

[1] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, et al. The BANCA database and evaluation protocol. *Lecture Notes in Computer Science (LNCS)*, 2688 :625–638, 2003.

[2] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoît Fauve, and John S.D. Mason. ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2008. http ://mistral.univ-avignon.fr/.

[3] Eric Charton, Anthony Larcher, Christophe Lévy, and Jean-Francois Bonastre. Mistral : open source biometric platform. In *Symposium on Applied Computing (ACM)*, Sierre (Switzerland), march 2010.

[4] G. Gravier. SPro : speech signal processing toolkit. *Software available at http ://gforge. inria. fr/projects/spro.*

[5] Patrick Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Factor analysis simplified. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, 2005.

[6] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, and Jean-Francois Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *International Conference on Speech Communication and Technology*, 2007.

[7] M. Przybocki and A.F. Martin. NIST speaker recognition evaluation chronicles. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*. ISCA, 2004.

[8] A. Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for svm speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 629–632, 18-23, 2005.

[4] `http://www.mobioproject.org/`

Regroupement des occurrences des mots hors-vocabulaire répétés en vue de leur modélisation pour la transcription d'émissions radio

Frederik Stouten, Irina Illina, Dominique Fohr

INRIA-LORIA
Campus Scientifique, BP 239 54506 Vandoeuvre-lès-Nancy Cedex, France
{stoutenf, illina, fohr}@loria.fr
http://www.loria.fr

ABSTRACT

This paper describes a novel technique to cluster Out-Of-Vocabulary (OOV) word tokens in a LVCSR system used for transcribing broadcast news speech data. The system is composed of two blocks: (1) an OOV word detector and (2) a clustering module working on the detected OOV word segments. This combination allows a more reliable detection of *repeated* OOV words than would be possible with the OOV detector only. In the paper we focus our attention on the second part of the system i.e. the clustering algorithm. This algorithm is based on the estimation of the entropy. The proposed algorithm gives better performance than a classical incremental clustering algorithm based on a distance threshold.
Keywords: LVCSR, OOV, clustering

1. INTRODUCTION

Le problème des mots hors vocabulaire (OOV) est un problème important des systèmes de reconnaissance de la parole. En effet, le nombre de mots reconnus par le système est limité par la taille du vocabulaire utilisé.

Les solutions proposées pour résoudre ce problème peuvent être classifiées en 2 catégories:

(1) Modéliser les mots OOV soit avec des modèles de fragments (unités sous-lexicales) [1, 2] soit avec des modèles « poubelle ». Ces méthodes ont l'avantage d'être flexibles et peuvent être utilisées pour des langues flexionnelles.

(2) Détecter les mots OOV en utilisant, par exemple, des probabilités à posteriori ou des mesures de confiance [3, 4, 5]. Ces méthodes sont bien adaptées pour les mots OOV d'origine étrangère comme les noms propres.

La prise en compte des mots OOV est particulièrement importante dans le cas de la transcription d'émissions de radio car les bulletins d'informations contiennent beaucoup de nouveaux noms propres, importants pour la compréhension.

Dans un travail précédent [6], nous avons proposé un système pour la détection des mots OOV inspiré par [5]. Ce système utilise des indices acoustiques (mesures de confiances) calculés à partir de 3 systèmes de reconnaissance et des indices linguistiques. Ces informations sont réunies ensemble dans un vecteur. La décision finale est prise par un réseau de neurones.

Dans cet article, nous proposons d'exploiter le fait que 38% des mots hors vocabulaire sont prononcés plus d'une fois dans une fenêtre temporelle de moins d'une minute. Notre but est de regrouper des occurrences des mots hors-vocabulaire répétés en vue de leur modélisation pour la transcription d'émissions radio. Cela permettra d'améliorer la détection des mots OOV répétés. Nous proposons un algorithme de regroupement (*clustering*) incrémental fondé sur l'estimation de l'entropie.

La section 2 décrit l'algorithme de regroupement, la section 3 présente le corpus de parole utilisé. Les résultats expérimentaux sont présentés section 4.

2. MÉTHODOLOGIE

L'algorithme de regroupement proposé est fondé sur des concepts de la théorie de l'information [7, 8]. Afin d'avoir une bonne compréhension de l'algorithme de regroupement nous allons d'abord introduire les concepts de base.

2.1. Définitions

Soit C_i ($i \in \{1, \ldots, N\}$) un ensemble de *N* classes. Chaque classe contient les occurrences des mots appartenant à cette classe. Pour une classe *C*, les occurrences (segments) correspondantes seront notées $\{o_1^C, o_2^C, \ldots, o_n^C\}$, où *n* est le nombre de segments dans la classe *C*.

Algorithme de regroupement

A chaque étape, une occurrence de mot doit être classifiée. Seulement deux cas sont possibles: soit ajouter cette occurrence à une classe existante, soit créer une nouvelle classe avec cette occurrence. La décision choisie doit augmenter une *fonction objectif*.

Nous avons choisi comme *fonction objectif*:

$$R = 1 - H / H_{max} \quad (1)$$

H représente l'entropie totale et H_{max} est l'entropie maximale de toutes les classes. Pour augmenter *R*, il y a deux possibilités: soit diminuer *H* (en ajoutant la nouvelle occurrence à une classe existante), soit augmenter H_{max}(en créant une nouvelle classe). Dans la suite, nous allons décrire comment estimer *H* et H_{max}.

Estimation non paramétrique de l'entropie d'une classe

Soit $f(o^C)$ la densité de probabilité d'une observation appartenant à la classe *C*. L'entropie de Shannon est définie ainsi :

$$H = -\int f(o^C)\, ln\, f(o^C)\, d\, o^C$$

Une estimation non paramétrique de *H* est donnée par [9]:

$$\hat{H} = -1/n \sum_i ln\ \hat{f}(o_i^C)$$

où $\hat{f}$est l'estimateur non paramétrique du noyau de la densité [10] et est donné par :

$$\hat{f}(o^C) = \frac{1}{nh} \sum_i K(\frac{o^C - o_i^C}{h}) \qquad (2)$$

K(x) représente la fonction du noyau et *h* est un paramètre de lissage. Nous avons fixé *h* à 1 pour notre application. Habituellement, le noyau est choisi comme la densité d'une fonction gaussienne, mais nous avons fait un choix différent : fonction avec deux arguments $K(o_i^C, o_j^C)$ (cf. section suivante). L'estimation non paramétrique de l'entropie de la classe *C* est donnée par l'équation suivante :

$$\hat{H} = -\sum_i w_i\, ln \sum_j w_j K\,(o_i^C, o_j^C) \qquad (3)$$

où w_i, w_j sont les poids $(\sum_i w_i = 1)$.

Fonction noyau

La fonction noyau $K\,(o_i, o_j)$ est définie pour deux segments o_i et o_j comme le résultat d'une comparaison dynamique (DTW). La fonction locale de similarité de cette DTW est :

$$s_{i,j}(t, t') = \exp\left[\frac{-1}{2\sigma_{local}^2} \sum_m (o_{i,m}(t) - o_{j,m}(t'))^2\right] \qquad (4)$$

avec $o_{j,m}(t)$ la $m^{\text{ième}}$ composante du vecteur de paramètre de la trame *t* de o_j.

Estimation non paramétrique de l'entropie totale

En utilisant $\hat{f}(o) = \sum_C \hat{f}(o^C)$ et en faisant l'approximation que $K(o_i^C, o_j^{C'})$ est négligeable pour des mots appartenant à des classes différentes $C \neq C'$, nous obtenons une approximation de l'estimation non paramétrique de l'entropie totale $\hat{H}_{\text{tot}}$:

$$\hat{H}_{tot} \approx -\sum_C \sum_i w_i^c \ln \sum_j w_j^c K\,(o_i^c, o_j^c) \qquad (5)$$

où w_i^c, w_j^c sont les poids $(\sum_C \sum_i w_i^c = 1)$.

Entropie maximale d'une classe

L'entropie maximale d'une classe ne peut pas être calculée mathématiquement car elle peut être infinie. Pour cette raison nous allons estimer une borne inférieure *BI* (*BI<1*) pour le noyau *K(x,y)*. En remplaçant *K(x,y)* par la borne inférieure *BI*, nous obtenons une expression de l'entropie maximale d'une classe :

$$\hat{H}_{max} = -ln\, BI \qquad (6)$$

Entropie maximale pour toutes les classes

En utilisant l'approximation (5) pour l'entropie totale, l'entropie totale maximale pour un ensemble de *N* classes est donnée par :

$$\hat{H}_{max,tot} = ln\, \frac{N}{BI} \qquad (7)$$

Nous pouvons noter que l'entropie maximale pour toutes les classes augmente quand le nombre de classes augmente.

Dans la section suivante, nous décrivons l'algorithme incrémental de regroupement.

2.2. Regroupement incrémental

Supposons que nous devons classer un nouveau segment *o* à l'étape *e*. Nous allons considérer deux cas.

Premier cas: ajouter le segment à une classe existante

Calculons l'augmentation dR^{modif} de la fonction objectif *R* résultant de la diminution de l'entropie totale $\hat{H}_{tot}$. Soit C_{max} la classe pour laquelle l'entropie totale diminue le plus par l'ajout de la nouvelle séquence d'observation à cette classe :

$$C_{max} = argmax_C \left[\hat{H}_{tot,e-1} - \hat{H}_{tot,e}^C\right] \qquad (8)$$

$H_{tot,e}^C$ est l'entropie totale quand la nouvelle observation a été ajoutée à la classe *C*. Dans ce cas dR^{modif} est obtenu par :

$$dR^{modif} = \frac{\hat{H}_{tot,e-1} - \hat{H}_{tot,e}^{C_{max}}}{\hat{H}_{max,tot}} \qquad (9)$$

Deuxième cas: création d'une nouvelle classe

Ici, l'entropie totale reste la même, mais l'entropie maximale de toutes les classes change. De façon similaire, dR^{creer}peut être calculée. Si $dR^{modif} > dR^{creer}$, alors le nouveau segment est ajouté à la classe C_{max}, sinon une nouvelle classe est créée. L'entropie d'une classe à un seul élément serait nulle. Pour résoudre ce problème, nous proposons de donner à l'entropie de cette nouvelle classe la valeur arbitraire *E*.

L'algorithme a deux paramètres libres *BI* et *E* qui doivent être appris. Le paramètre σ_{local} utilisé par la DTW est fixé arbitrairement.

2.3. Algorithme de référence

Nous allons comparer les performances de notre nouvel algorithme avec celles d'un algorithme de référence. L'algorithme de référence est similaire à celui de [11] pour le regroupement de locuteurs. Cet algorithme de référence est fondé sur le calcul de la distance *D* entre le segment *o* et la classe *C* :

$$D^2(o,C)=K(o,o)+B-2\sum_i w_i K(o,o_i^c) \qquad (10)$$

K(o,o) vaut 1. *B* est le biais :

$$B=\sum_i\sum_j w_i\, w_j K(o_i^c,o_j^c) \qquad (11)$$

L'algorithme cherche la classe C_{min} pour laquelle la distance *D* est minimum. Si cette distance est inférieure à un seuil *T* (valeur estimée sur le corpus d'apprentissage, cf. section suivante), alors cette séquence d'observation est ajoutée à la classe C_{min}. Dans le cas contraire, une nouvelle classe est créée.

3. CONDITIONS EXPÉRIMENTALES

Le corpus de parole utilisé consiste en fichiers audio extraits du corpus radio ESTER [12]. Pour l'apprentissage nous utilisons la parole de 10 locuteurs (158540 occurrences de mots). Nous parlons en terme de « locuteurs » car souvent dans les émissions de radio un mot hors vocabulaire est répété plusieurs fois par le même locuteur. Pour aligner ce corpus d'apprentissage nous utilisons l'alignement automatique.

Sur les segments d'apprentissage, nous avons déterminé les valeurs des paramètres *BI* et *E (*pour notre algorithme de regroupement) et *T* (pour l'algorithme de référence) qui maximisent la *F-mesure* moyenne :

$$F=\sum_i \frac{|L_i|}{N} max_j\{F(L_i,C_j)\} \qquad (12)$$

L_i correspond à la classification idéale en connaissant l'identité des mots à classifier. $|L_i|$ est le nombre d'éléments de la *i*-ème classe. La *F-mesure* d'une classe est calculée de la façon suivante :

$$F(L_i,C_j)=\frac{2\times Rec(L_i,C_j)\times Prec(L_i,C_j)}{Rec(L_i,C_j)+Prec(L_i,C_j)} \qquad (13)$$

$Rec(L_i,C_j)$ est le rappel et $Prec(L_i,C_j)$ la précision de la classe idéale *i* évaluée en utilisant la classe *j*. Les valeurs optimales de paramètres libres sont *BI=0.0125, E = 4.0* et *T=1.178*, $w_i^c = \frac{1}{\sum_c n^c}$, $i \in \{1,\dots,n^C\}$

Pour la paramétrisation des fichiers audio, nous utilisons des coefficients MFCC calculées sur 9 trames consécutives (réduction à 40 coefficients par HLDA).

Deux expériences ont été réalisées: (1) une évaluation de l'algorithme de regroupement et (2) une évaluation de la combinaison d'un détecteur de mots hors-vocabulaire et de l'algorithme de regroupement proposé.

4. EXPÉRIENCES

Regroupement de segments OOV

Pour l'évaluation de l'algorithme de regroupement, les fichiers audio de 20 locuteurs ont été extraits du corpus ESTER. Ce corpus de test comprend 2034 occurrences de mots OOV (par rapport au lexique de 60K mots).

Pour chaque fichier du locuteur de test, l'alignement automatique est effectué pour trouver les segments correspondants aux mots. Pour évaluer l'algorithme de regroupement nous ne gardons que les segments correspondant aux occurrences de mots OOV. Puis nous effectuons le regroupement de ces segments. Après le regroupement, la *F-mesure* est calculée entre les classes obtenues par le regroupement C_j et les classes de référence L_i (toutes les occurrences d'un mot OOV forment une classe).

La première expérience consiste à classifier une suite de *M* occurrences (segments) de mots OOV, arrivant au fur et à mesure de l'alignement automatique. Nous avons fait varier *M* de 10 à 100.

Table 1: Performances de regroupement mesurées sur les segments OOV en fonction de la valeur de *M*.

***M*(#segms OOV)**	***F-mesure*(%) Alg. proposé**	***F-mesure* (%) Alg. de réf.**
10	**89.2 (+-7.0)**	81.5 (+-8.0)
20	**83.2 (+-6.2)**	75.7 (+-7.6)
30	**82.3 (+-5.7)**	74.1 (+-7.8)
40	**79.9 (+-5.6)**	70.4 (+-8.1)
50	**79.7 (+-5.1)**	67.8 (+-8.3)
60	**77.6 (+-5.2)**	65.9 (+-9.0)
70	**75.9 (+-4.6)**	63.6 (+-8.8)
80	**75.7 (+-4.4)**	62.2 (+-9.2)
90	**75.7 (+-4.1)**	61.5 (+-9.3)
100	**74.7 (+-4.4)**	60.5 (+-9.4)

Chaque expérience a été répétée 20 fois en changeant chaque fois la liste des occurrences des mots hors vocabulaire. En moyenne, chaque mot hors vocabulaire est prononcé 1.7 fois. Pour chaque expérience la *F-mesure* est calculée. La table 1 présente la moyenne et l'intervalle de confiance (à ±95%) de ces *F-mesures* pour les 20 expériences. Nous observons que notre algorithme obtient de meilleures performances que l'algorithme de référence.

Plus *M* augmente, plus la différence entre les résultats de la méthode proposée et de la méthode de référence est significative.

Combinaison du détecteur OOV et de l'algorithme de regroupement

Dans cette partie, nous combinons notre détecteur de mots hors vocabulaire et le nouvel algorithme de regroupement.

Pour l'évaluation de la combinaison (détecteur OOV + regroupement), le corpus de test se compose de 6 bulletins d'informations (2h15min, 23k mots) et est le même que dans [6].

Pour évaluer la précision et le rappel, nous nous intéressons ici uniquement aux mots hors vocabulaire qui sont répétés. (Nous comptons une erreur si un mot OOV répété est classé dans une classe à un seul élément ou si un mot OOV non répété est classé dans une classe à plusieurs éléments).

Nous obtenons, pour ces OOV répétés, une précision de 21.9% et un rappel de 28.1%.

5. CONCLUSIONS ET PERSPECTIVES

Dans cet article, nous avons proposé une nouvelle approche pour détecter les mots hors vocabulaire répétés. Nous avons ajouté un module de regroupement incrémental des mots OOV, fondé sur l'estimation de l'entropie.

Le nouvel algorithme a été évalué sur le corpus radiophonique ESTER. Les résultats montrent que les performances de ce nouvel algorithme sont meilleures que celles d'un algorithme classique incrémental.

Le module de regroupement des occurrences de mots hors vocabulaire répétées pourra nous permettre de mieux trouver les frontières de ces occurrences et de construire les modèles des mots OOV pour ensuite améliorer les taux de reconnaissance. Une possibilité envisageable pour ces modèles est une séquence de phonèmes.

RÉFÉRENCES

[1] M. Bisani and H. Ney. Open vocabulary speech recognition with flat hybrid models. In *Proceedings of Interspeech*, pages 725-728, 2005

[2] L. Galescu. Recognition of out-of-vocabulary words with sub-lexical language models. In *Proceedings of Eurospeech*, pages 249-252, 2003

[3] H. Lin, J. Bilmes, D. Vergyri and K. Kirchhoff. OOV detection by joint word/phone lattice alignment. In *Proceedings of the Automatic Speech Recognition and Understanding (ASRU) Workshop*, pages 478-483, 2003

[4] C. White, G. Zweig, L. Burget, P. Schwarz and H. Hermansky. Confidence estimation, OOV detection and language ID using phone-to-word transduction and phone-level alignments. In *Proceedings of ICASSP*, pages 4085-4088, 2008

[5] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky and J. Cernocky. Combination of strongly and weakly constrained recognizers for reliable detection of OOVs. In *Proceedings of ICASSP*, pages 4081-4084, 2008

[6] F. Stouten, D. Fohr and I. Illina Detection of OOV words by combining acoustic confidence measures with linguistic features. In *Proceedings of the Automatic Speech Recognition and Understanding (ASRU) Workshop*, pages 371-375, 2009

[7] R. Jenssen, K.E. Hild, D. Erdogmus, J. Principe and T. Eltoft. Clustering using Renyi's entropy. In *Proceedings of the International Joint Conference on Neural Networks*, pages 523-528, 2003.

[8] E. Gokcay and J. Principe Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24, 2:158-170, 2002

[9] I. Ahmad and P. Lin A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Transactions on Information Theory*, volume 22, 3:372-375, 1976.

[10] E. Parzen. On the estimation of a probability density function and the mode. In *Ann. Math. Stat.*, volume 32, pages 1065-1076, 1962.

[11] A. Vandecatseye, J.-P. Martens. A fast accurate and stream-based speaker segmentation and clustering algorithm. In *Proceedings of Interspeech*, pages 941-944, 2003

[12] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait and K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques, *Proceedings Journées d'Etude sur la Parole*, April 2004.

Détection et correction des disfluences dans le dialogue oral arabe spontané

Younès Bahou, Abir Masmoudi, Lamia Hadrich Belguith

ANLP Research Group - Laboratoire MIRACL, FSEGS – Université de Sfax, Tunisie.
bahou_younes@yahoo.fr, masmoudiabir@gmail.com, l.belguith@fsegs.rnu.tn

ABSTRACT

The disfluencies inherent in spontaneous speaking are a real challenge for speech understanding systems. Thus, we propose in this paper, an original method for processing disfluencies in the context of automatic Arabic speech understanding. Our method based on a robust and partial analysis of Arabic oral utterances (conceptual segments analysis) is effective for the treatment of such phenomena. This method has been tested through the understanding module of SARF system, an interactive vocal server for Tunisian railway information.

Keywords: Disfluencies, conceptual segments, pattern matching, spontaneous Arabic speech understanding.

1. INTRODUCTION

Comme indique leur étymologie, les disfluences correspondent à toute interruption ou perturbation de la *fluence*, c'est-à-dire du cours normal de la production orale spontanée [5]. Ainsi, le terme de disfluences regroupe un certain nombre de phénomènes comme les répétitions, les autocorrections, les amorces, les faux départs, les hésitations, etc.

Dans ce papier, nous proposons une méthode originale pour le traitement des disfluences (en particulier, les autocorrections, les répétitions, les hésitations et les amorces) dans le cadre de la compréhension automatique de la parole arabe spontanée. Cette méthode est basée sur une analyse robuste et partielle par segments conceptuels des énoncés arabes. L'originalité de notre méthode réside dans l'utilisation des segments conceptuels. En effet, un énoncé étiqueté sémantiquement passe par deux niveaux de traitement : ***i)*** le découpage en segments conceptuels et ***ii)*** la détection et la correction des disfluences.

Ce travail entre dans le cadre de la réalisation du serveur vocal interactif SARF ([1], [2]) offrant des renseignements sur le transport ferroviaire tunisien (*e.g.*, horaire du train, tarification, etc.) en langue arabe standard moderne.

2. TRAVAUX ANTÉRIEURS

Suite à notre étude de l'état de l'art sur le traitement des disfluences, nous proposons de répertorier les travaux existants selon trois approches à savoir, l'approche de *Standford Research Instute* (SRI), l'approche stochastique et l'approche linguistique.

2.1. Approche de SRI

L'approche développée au sein de *Standford Research Instute* [3] constitue l'un des premiers travaux sur les disfluences dans un cadre applicatif. La première étape de ce travail consiste à proposer un schème d'annotation des disfluences qui combine la simplicité à la finesse nécessaire pour la représentation des différentes formes des disfluences. Il s'agit, ensuite, de combiner l'analyse syntaxique et sémantique (afin de réduire la surgénération des patrons) avec la technique de la reconnaissance des patrons (*pattern matching*) afin de détecter et de corriger les répétitions simples et les anomalies syntaxiques simples [3].

L'inconvénient principal de cette combinaison est qu'elle est incompatible avec les approches d'analyse partielle qui sont les plus adaptées au traitement de l'oral. De plus, cette approche rend le module de traitement des disfluences complètement dépendant de l'analyseur syntaxique et par conséquent elle réduit considérablement sa portabilité.

2.2. Approche stochastique

Cette approche a été proposée par Heeman et Allen dans le cadre du projet TRAINS au sein de l'université de *Rochester* [7]. La première étape de ce travail consiste à proposer une version modifiée du schème d'annotation des chercheurs de SRI. Ainsi, le schème proposé ne permet pas le partage de la zone remplacée dans le cas de disfluences imbriquées. L'idée principale de ce travail est basée sur l'usage de certains types de disfluences (*i.e.*, hésitations, termes d'édition, etc.) comme indices des occurrences des autocorrections. Pour cela, les auteurs mettent en place un modèle statistique permettant d'associer à la probabilité d'occurrence d'une disfluence donnée, la probabilité d'apparition d'une autocorrection. Il s'agit, ensuite, de détecter et de corriger les disfluences en utilisant des patrons et des règles pour éviter leur surgénération.

Dans son travail, Kurdi adopte cette approche pour le traitement des disfluences (extra-grammaticalités) dans le cadre de la réalisation du

système CORRECTOR [8]. L'idée principale est de combiner les patrons et la technique de n-grammes lors du traitement des disfluences. Ainsi, Kurdi utilise à la fois une technique de la reconnaissance des patrons et une analyse syntaxique et sémantique superficielle. Aussi, il utilise des règles pour éviter la surgénération des patrons [8].

Le travail de Bove s'inscrit dans le cadre d'un travail d'équipe sur l'analyse morphologique et syntaxique du français parlé [5]. Il adopte une méthode "hybride" pour la reconnaissance des patrons (basée sur les catégories morpho-syntaxiques du corpus) couplée à un calcul de n-grammes pour détecter les différents phénomènes de disfluences. Une dernière étape consiste à regrouper les énoncés disfluents précédemment analysés en syntagmes minimaux non récursifs (ou *chunks*). Le corpus final est ainsi segmenté en *chunks* de l'écrit d'une part, à côté des *chunks* disfluents d'autre part [5].

2.3. *Approche linguistique*

Cette approche a été proposée par Core et Schubert dans le cadre général de l'analyse robuste des dialogues au sein de l'université de *Rochester* [6]. La particularité de ce travail est l'introduction d'informations linguistiques (notamment la syntaxe) dans le traitement des disfluences, d'une manière originale.

L'idée principale est basée à la fois sur l'utilisation d'un modèle statistique et l'analyse syntaxique de la totalité de l'énoncé. Pour cela, les disfluences détectées dans le module statistique sont analysées à l'aide de méta-règles syntaxiques dédiées spécialement à cette tâche. L'ajout des méta-règles syntaxiques s'est révélé très coûteux [6]. En effet, l'ajout des méta-règles syntaxiques multiplie le temps de calcul d'environ trois fois.

3. LA MÉTHODE PROPOSÉE

Avant de détailler les étapes de la méthode que nous proposons, nous jugeons nécessaire de présenter le corpus d'étude utilisé.

3.1. *Corpus d'étude*

Vu que les ressources linguistiques arabes sont très rares, nous étions amenés à créer notre propre corpus d'étude selon la technique de *Magicien d'Oz*. Ce corpus de 11 heures d'enregistrement est constitué de 300 dialogues (soit 7590 énoncés arabes) [2].

Le corpus d'étude nous a permis d'étudier la fréquence de chaque type de disfluences que nous envisageons de traiter à savoir, les autocorrections, les répétitions, les hésitations et les amorces. La figure 1 résume les résultats obtenus.

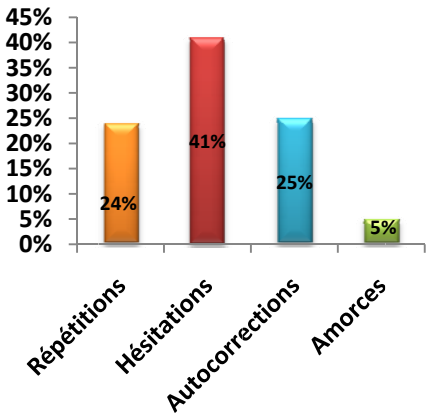


Figure 1 : Distribution des disfluences dans le corpus d'étude.

Aussi, ce corpus nous a aidé à recenser les segments conceptuels (85 segments conceptuels) ainsi que les patrons (201 patrons avec ou sans marqueurs de rectification).

3.2. *Étapes de la méthode proposée*

Nous proposons de traiter les disfluences au niveau de la compréhension de la parole vu que nous ne disposons pas d'un logiciel pour la reconnaissance de l'oral arabe avec un code source accessible (*open source*). Ainsi, la méthode proposée est basée sur l'approche de SRI et utilise la technique de la reconnaissance des patrons et une segmentation conceptuelle des énoncés. L'originalité de cette méthode réside dans l'utilisation des segments conceptuels pour détecter les disfluences. À notre connaissance, c'est le premier travail qui propose l'utilisation des segments conceptuels pour le traitement des disfluences. Généralement, les travaux qui se basent sur une analyse partielle des énoncés utilisent les *chunks* (purement syntaxiques) qui ont prouvé leur performance dans le traitement de l'écrit. Cependant, pour le traitement de l'oral et pour un domaine limité, comme le notre, nous jugeons que l'utilisation des segments conceptuels (purement sémantiques) est plus intéressante et peut donner de bons résultats. Aussi, la segmentation conceptuelle joue un rôle important dans la résolution du problème posé par la présence des disfluences imbriquées.

La méthode que nous proposons consiste ainsi en deux principales étapes à savoir, l'étape de découpage des énoncés arabes en segments conceptuels et l'étape de détection et de correction des disfluences dans ces énoncés.

Afin de bien expliquer notre méthode, nous prenons à titre d'exemple l'énoncé (1). Cet énoncé contient des disfluences qui seront détectées puis corrigées durant les étapes de notre méthode.

⟨Station, صفاقس⟩ ⟨Marq_Station_Départ, من⟩ ⟨Type_Train, عادي⟩ ⟨Marq_Train, قطار⟩ ⟨Mot_Ref_Horaire, وقت⟩ ⟨Demande, أراد⟩ (1)

⟨Station, تونس⟩⟨Marq_Station_Arrivée, إلى⟩⟨Station, صفاقس⟩⟨Marq_Station_Départ, من⟩⟨Type_Train, سريع⟩⟨Marq_Rectification, عفو⟩

- **Découpage en segments conceptuels**

La première étape consiste à segmenter l'énoncé en segments conceptuels (composés des classes de mots exprimant une même unité de sens [4]).

Ces segments, recensés à partir du corpus d'étude, sont de trois types [4] : les *illocutoires* qui font référence à la théorie des actes de langage (*i.e., Demande, Début_Dialogue*, etc.), les *référentiels* permettant de représenter le domaine de l'application (*i.e., Heure_Départ, Départ, Destination*, etc.) et les *rebuts* regroupant les mots et les groupes de mots considérés comme inutiles pour la compréhension des énoncés (*i.e., Bruit, Digression*, etc.).Pour les segments conceptuels illocutoires, nous proposons d'ajouter un nouveau segment contenant les disfluences que nous nommons *Disfluent*. Ce découpage est basé sur la liste des segments conceptuels, les pré-marqueurs et les post-marqueurs existants dans l'énoncé ainsi que sur les étiquettes sémantiques des mots de l'énoncé.

À ce stade d'analyse, un raffinement des étiquettes sémantiques est obligatoire pour une bonne description des cas sémantiques. Ce raffinement est basé essentiellement sur l'étiquette et la position du mot dans l'énoncé. En effet, notre méthode prend en considération le contexte du mot au sein de l'énoncé. Par exemple, l'étiquette *Nombre* peut avoir plusieurs raffinements possibles selon le contexte du mot (*i.e., Prix, Heure, Minute,* etc.).

Ainsi, l'énoncé (2) est le résultat du découpage en segments conceptuels de l'énoncé (1).

$$\begin{bmatrix} \text{عادي من صفاقس} \\ [EAdy\ mn\ SfAqs] \\ (\ normal\ de\ Sfax) \end{bmatrix}_{Reparandum} \quad \begin{bmatrix} \text{قطار} \\ [qTAr] \\ (\ train) \end{bmatrix} \qquad (3)$$

$$\begin{bmatrix} \text{سريع من صفاقس} \\ [sryE\ mn\ SfAqs] \\ (rapide\ de\ Sfax) \end{bmatrix}_{Altération} \quad \begin{bmatrix} \text{عفو} \\ [Efw] \\ (pardon) \end{bmatrix}_{Terme\ d'édition}$$

À ce niveau d'analyse, la technique de la reconnaissance des patrons est appliquée.

Ces patrons concernent le cas d'une autocorrection, d'une répétition, d'une hésitation, d'une amorce ou d'une combinaison des quatre types.

Ces patrons reposent sur l'identification des suites de mots du *reparandum* et de l'altération qui se répètent d'une manière identique (**M**), qui sont repris (mots différents jouant le même rôle sémantique : **R**) ou qui sont ajoutés (mots neutres : **X**). S'y ajoute éventuellement un terme d'édition (**ET**) et un point d'interruption noté par une barre verticale (|).

$$\begin{matrix} \text{قطار عادي من صفاقس} & & \text{عفو سريع من صفاقس} \\ M2\ \ M1\ \ R1 & | & ET\ \ R1\ \ M1\ \ M2 \end{matrix} \qquad (4)$$

Pour la correction proprement dite, l'altération est gardée, cependant le terme d'édition ainsi que le *reparandum* sont supprimés. Le segment résultat, passe par une analyse similaire à celle de l'étape de découpage en segments conceptuels afin de déterminer le type du segment résultat.

$$\left\{ \begin{matrix} \text{قطار عادي من صفاقس عفو سريع من صفاقس} \\ [qTAr\ EAdy\ mn\ SfAqs\ Efw\ sryE\ mn\ SfAqs\ ] \\ (train\ normal\ de\ Sfax\ pardon\ rapide\ de\ Sfax) \end{matrix} \middle| Disfluent \right\} \quad \left\{ \begin{matrix} \text{أراد وقت} \\ [OrAd\ wqt] \\ (vouloir\ horaire) \end{matrix} \middle| Demande\_Horaire \right\} \qquad (2)$$

$$\left\{ \begin{matrix} \text{إلى تونس} \\ [IlY\ twns] \\ (vers\ Tunis) \end{matrix} \middle| Destination \right\}$$

- **Détection et correction des disfluences**

Lors de cette étape les segments disfluents (détectés au niveau de l'étape précédente) subissent une annotation semblable à celle proposée par Bear et *al.* [3], ensuite, ils seront corrigés.

Le segment disfluent est décrit comme la succession d'un *reparandum* (partie du segment qui sera corrigée par la suite), d'un *terme d'édition* optionnel (marqueur de reprise) et d'une *altération* (partie qui corrige ou complète le *reparandum*) comme illustré dans l'exemple (3).

Le segment (5) représente le segment disfluent de l'énoncé (2) après sa correction.

$$\left\{ \begin{matrix} \text{من صفاقس} \\ [mn\ SfAqs] \\ (de\ Sfax) \end{matrix} \middle| Départ \right\} \left\{ \begin{matrix} \text{قطارسريع} \\ [qTAr\ sryE] \\ (train\ rapide) \end{matrix} \middle| Type\_Train \right\} \qquad (5)$$

4. MISE EN ŒUVRE DE LA MÉTHODE ET ÉVALUATION

La méthode proposée dans ce papier, a été implémentée dans le module de compréhension du système SARF (Serveur vocal Arabe des Renseignements sur le transport Ferroviaire). Nous l'avons programmée avec le langage JAVA sous l'environnement JBuilder 2007. Les patrons et les

segments conceptuels sont regroupés dans des fichiers XML.

Pour le corpus d'évaluation, nous l'avons construit selon la même technique du *Magicien d'Oz* utilisée pour la construction du corpus d'étude. Le corpus d'évaluation est constitué de 2535 énoncés (soit 32520 mots) prononcés d'une façon spontanée. Ces énoncés sont de différents types (859 énoncés contiennent des autocorrections, 738 énoncés contiennent des répétitions, 342 énoncés contiennent des amorces, 388 énoncés contiennent des hésitations et 208 énoncés contiennent des disfluences imbriquées).

Nous avons évalué le module de compréhension de SARF et les mesures de rappel, de précision et de *F-Measure* que nous avons obtenus sont respectivement 79.23%, 74.09% et 76.57% ; Le temps moyen d'exécution d'un énoncé, de 12 mots, est au environ de 0.394 seconde. Le taux d'erreurs que nous avons obtenu est de 12.63%. Notons que le taux d'erreur obtenu lors de l'évaluation de ce module sans tenir compte du traitement des disfluences [2] est de 18.54% ; ce qui fait une diminution de 5.91% d'erreurs, chose que nous jugeons satisfaisante.

Les cas d'échec s'expliquent principalement par une mauvaise assignation des étiquètes sémantiques aux mots d'un énoncé ce qui peut entrainer un mauvais découpage et par la suite provoquer des erreurs de détection des disfluences. La mauvaise attribution des étiquettes sémantiques est due principalement à la confusion entre les marqueurs de négation et les marqueurs de rectification. À titre d'exemple, le mot « لا » [*lA*] (non) peut jouer deux rôles sémantiques à savoir, un marqueur de négation ou un marqueur de rectification.

Un autre cas d'échec est dû à la présence des énumérations dans les énoncés. En effet, l'aspect structurel d'une énumération est très proche de celui d'une autocorrection. À titre d'exemple, le cas d'une énumération de deux types de billet à savoir, « ذهاب » [**hAb*] (aller simple) et « ذهاب-ايا » [**hAb-AyAb*] (aller-retour) que notre système considère comme étant un cas d'autocorrection d'un type de billet par un autre.

5. CONCLUSION ET PERSPECTIVES

La compréhension de la parole spontanée est un thème de recherche désormais classique mais il intéresse encore beaucoup de chercheurs et de nombreux progrès restent à faire dans ce domaine. Cependant la parole arabe a fait l'objet de très peu de travaux de recherche en comparaison avec d'autres langues telles que l'anglais et le français.

Dans ce papier, nous avons proposé une méthode assez originale pour le traitement des disfluences dans le cadre de la compréhension de la parole arabe spontanée. Cette méthode se base sur l'approche de RSI et combine la technique de la reconnaissance des patrons avec celle de la segmentation conceptuelle. Les résultats que nous avons obtenus sont encourageants (*F-Measure* égale à 76.57%).

Comme perspectives, nous envisageons d'étudier les phénomènes de négation et d'énumération en vue d'apporter des solutions quant à leur détection pour ne pas les confondre avec les disfluences.

BIBLIOGRAPHIE

[1] Y. Bahou, A. Bayoudhi et L. Hadrich Belguith. Gestion de dialogue oral Homme-machine en arabe. *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles, TALN'09*, Senlis, France, 2009.

[2] Y. Bahou, L. Hadrich Belguith, A. Ben Hamadou. Towards a Human-Machine Spoken Dialogue in Arabic. *LREC'08, Workshop HLT within the Arabic World: Arabic Language and local languages processing Status Updates and Prospects*, Marrakech, Morocco, 2008.

[3] J. Bear and J. Dowding and E. Shriberg. Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Deleware, USA ,1992.

[4] C. Bousquet. Compréhension Robuste de la Parole Spontanée dans le Dialogue Oral Homme-Machine – Décodage Conceptuel Stochastique. *Thèse de doctorat à l'Université de Toulouse III–Paul SABATIER*, France, 2002.

[5] R. Bove. A Tagged Corpus-Based Study for Repeats and Self-Repairs Detection in French Transcribed Speech. *Proceedings of the 11th International Conference on Text, Speech and Dialogue, TSD'08,* Brno, Czech Republic, 2008.

[6] M. Core and L. Schubert. A Model of Speech Repairs and Other Disruptions. *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems,* Cape Cod, MA, USA, 1999.

[7] P.A. Heeman and J.F. Allen. Combining the Detection and Correction of Speech Repairs. *In Proceedings of International Conference of Spoken Language Processing,* Philadelphia, PA, USA, 1996.

[8] Z. Kurdi. Contribution à l'analyse du langage oral spontané. *Thèse de doctorat à l'université Joseph Fourier*, France, 2003.

Annotation automatique en syllabes d'un dialogue oral spontané

Brigitte Bigi[1]*, Christine Meunier*[1]*, Roxane Bertrand*[1]*, Irina Nesterenko*[2]

Laboratoire Parole et Langage[1],
CNRS & Aix-Marseille Universités,
5, avenue Pasteur,
13100 Aix-en-Provence France,

Institut National des Langues[2]
et Civilisations Orientales
2 rue de Lille
75343 Paris cedex 07 France

e-mail : brigitte.bigi@lpl-aix.fr,christine.meunier@lpl-aix.fr,roxane.bertrand@lpl-aix.fr, irina.nesterenko@inalco.fr

ABSTRACT

This paper proposes a solution to identify automatically syllable boundaries in the particular context of spontaneous speech. The main goal consists in identifying syllables from a continuous stream of phonemes. At first, phoneme classes are defined to be as well-suited as possible to reduce the problem complexity. Secondly, a few number of general rules are defined. Finally, some exception rules allows to adapt the problem to the specific context of spontaneous speech. The proposed system is evaluated and compares favorably to the only two existing other systems, for French, with significant improvements.
Keywords : syllable, phoneme, segmentation, rules.

1. INTRODUCTION

La syllabe est une des unités fondamentales de la parole et une unité structurale de grande importance dans la production et la perception du langage. Toutefois, la caractérisation de la syllabe tant sur le plan articulatoire qu'acoustique reste très difficile. Elle est néanmoins souvent considérée comme l'unité de production dans laquelle les phénomènes de coarticulation sont plus saillants [7]. Par ailleurs, elle est une unité fondamentale dans l'organisation prosodique du français, langue « syllable timed » par opposition aux langues « stressed timed » comme l'anglais [9]. Même si cette classification est parfois remise en cause [2], la réalité cognitive de la syllabe (intégrée dans la compétence des locuteurs) nécessite de la prendre en compte dans les analyses portant sur la parole continue.

Cet article s'inscrit dans le cadre du développement d'annotations multimodales de dialogues oraux spontanés. A partir de l'annotation en phonèmes, notre but consiste à proposer un outil pour identifier automatiquement les segments syllabiques. L'intérêt de ce niveau d'annotation sur *un corpus de parole continue non contrôlée* est double. Dans un premier temps, il permet d'obtenir des statistiques sur la structure et la fréquence des syllabes du français en parole non contrôlée et après la réorganisation syllabique (re-syllabation) qu'entraîne la parole continue (non dépendante du lexique) et les phénomènes de réduction spécifiques à ce type de parole. Notamment, les phénomènes de réduction tels que déformation, assimilation, élision de phonèmes, extrêmement importants en parole naturelle non contrôlée, font apparaître des séquences de segments illégales d'un point de vue phonotactique (ex. : clusters de consonnes), mais également non prévisibles dans les combinaisons de joncture de mots. Si ces structures sont peu fréquentes, elles nécessitent néanmoins qu'on y porte intérêt. Disposer de cette annotation permet enfin d'envisager les analyses segmentales (structures acoustiques des phonèmes) et supra-segmentales (organisation rythmique et accentuelle du discours) en fonction du découpage syllabique. Par exemple, cette annotation permettra de mettre en perspective le rôle et la hiérarchie des frontières lexicales et/ou syllabiques dans la coarticulation. Ces phénomènes ont pu être évalués en parole contrôlée mais sont encore ignorés en parole non contrôlée.

Cet article concerne la proposition d'un système de détection automatique des frontière syllabiques en parole spontanée. Dans [4], quelques outils et la description du corpus ont été présentés. Le corpus - Corpus of Interactional Data (CID)[1], est un enregistrement audio-vidéo de dialogues spontanés entre deux locuteurs français natifs (8 heures, 8 paires de locuteurs). Une transcription orthographique enrichie (TOE) a été réalisée et corrigée manuellement. A partir de cette TOE, un convertisseur graphème-phonème suivi d'un aligneur permettent d'obtenir une phonétisation de bonne qualité. Cette segmentation en phonèmes, rend possible la réalisation d'une recherche automatique des frontières de syllabes.

2. TRAVAUX EXISTANTS

A notre connaissance, il existe deux outils *libres* de recherche des syllabes à partir de phonèmes du français. *syllabation.awk*, développé par C. Pallier [8] est un script, librement diffusé, écrit en langage *awk*. Les phonèmes y sont regroupés en quatre classes : voyelles, semi-voyelles, liquides, et consonnes. Une douzaine de règles de segmentation sont définies selon les suites de phonèmes et leur classe. Ce segmenteur a été appliqué avec succès sur des mots isolés, plus spécifiquement sur les lexiques Brulex et Lexique[2]. Le second système libre repose également sur un ensemble de règles de segmentation appliquées sur les suites de phonèmes. *syllabify2.praat* est une partie du logiciel *EasyAlign*, développé par J-P. Goldman [6], sous la forme de scripts Praat [5]. Cinq classes de phonèmes y sont définies : voyelles, semi-voyelles, liquides, une classe comportant les consonnes p t k b d g f v et une classe comportant les consonnes s ʃ z ʒ m n ŋ ɲ. Ce système inclue également le silence afin de traiter la syllabation de corpus oraux. Une soixantaine de règles y ont été développées. Enfin, [1] propose une étude de la syllabation du français parlé (radio). La syllabation est réalisée avec 13 règles par l'outil GRAPHON+. Les phonèmes sont groupés en 5 classes : voyelles, semi-voyelles, liquides, occlusives et autres consonnes.

[1] http ://crdo.up.univ-aix.fr
[2] http ://www.lexique.org

3. SYSTÈME PROPOSÉ

3.1. Principes généraux

Le système que nous avons développé s'inspire de ceux décrits précédemment en ce sens qu'il consiste à définir un ensemble de règles de segmentation entre phonèmes, selon leur classe. Cependant, avant la définition de règles, nous proposons les deux principes suivants :

Principe 1 : une syllabe contient une seule voyelle.

Principe 2 : une pause est une frontière de syllabe. Dans le corpus que nous utilisons, les pauses silencieuses supérieures à 200 ms ont été annotées automatiquement et les pauses inférieures à ce seuil ont été annotées manuellement lors de l'étape de transcription.

Ces deux principes résument le problème de syllabation que nous considérons en la recherche de frontières de syllabes entre deux voyelles.

3.2. Classes de phonèmes

Les classes que nous proposons sont les suivantes :

V - Voyelles : i e ɛ a ɑ ɔ o u y ø œ ə ɛ̃ ɑ̃ ɔ̃ œ̃

G - Semi-voyelles : j ɥ w

L - Liquides : l ʁ

O - Occlusives : p t k b d ɡ

F - Fricatives : s z ʃ ʒ f v

N - Nasales : m n ŋ ɲ

La lettre en gras est le symbole utilisé dans ce papier pour désigner la classe. De plus, le symbole **X** fait mention de l'un des G, L, O, N ou F (en d'autres termes, X renvoie à un phonème qui n'est pas une voyelle).

La répartition des consonnes en trois classes est un choix majeur qui réduit largement la complexité du développement des règles. La division en un nombre plus petit de classes de consonnes impliquerait le développement, comme dans le segmenteur syllabation2.praat, d'un très grand nombre de règles pour traiter les cas particuliers. Une subdivision plus importante nous est apparue inutile car elle montrait de fortes redondances dans l'élaboration des règles, sans gain de performance.

3.3. Règles de syllabation

Afin d'augmenter la généricité de l'approche, nous proposons d'organiser les règles en deux types : des règles générales décrites dans la table 1, applicables à tout type de situation, et des règles exceptions qui peuvent être modulées selon le problème abordé (dialogue spontané ou non, etc.) décrites dans la table 2.

La première règle générale applique le principe 1 selon lequel il n'y a qu'une voyelle par syllabe. La deuxième règle reflète la tendance universelle à privilégier les syllabes ouvertes, en conséquence de quoi la consonne est assignée à la seconde syllabe. Les règles générales 4, 5 et 6 satisfont la règle du « Maximum Onset Principle » pour laquelle, dans un groupe de consonnes intervocaliques, le maximum de consonnes doivent être attribuées à l'onset de la seconde syllabe plutôt qu'à la coda de la première. La troisième règle doit être considérée en fonction des règles exceptions 1, 2 et 3 (table 2), dans lesquelles les clusters intervocaliques sont constitués de deux consonnes. Ces règles montrent que pour un cluster de deux consonnes le « Maximum Onset Principle » doit être appliqué prudemment et en fonction du principe de sonorité, c'est-à-dire de la nature des consonnes du cluster. Ainsi, la règle générale ne s'applique que lorsque le principe de sonorité est violé au sein du cluster.

TAB. 1: Règles générales

	Séquence	Règle	Exemples
1	VV	V.V	poëte : po.ɛt
			il y a un : i.ɑ.œ̃
			en haut : ɑ̃.o
2	VXV	V.XV	limité : li.mi.te
			et donc on : e.dɔ̃.kɔ̃
3	VXXV	VX.XV	jardin : ʒɑʁ.dɛ̃
			comme ça : kɔm.sɑ
			parce qu'il : pɑs.ki
4	VXXXV	VX.XXV	avec moi : ɑ.vek.mwa
			cheval noir : sə.vɑl.nwɑʁ
5	VXXXXV	VX.XXXV	il se présentait : il.spʁe.zɑ̃.te
6	VXXXXXV	VXX.XXXV	alors je crois : ɑ.lorʒ.krwɑ

TAB. 2: Règles exceptions

	Séquence	Règle	Exemples
1	VXGV	V.XGV	baignoire : be.nwɑʁ
			spéciaux : spe.sjo
			tu vois : tu.vwɑ
2	VFLV	V.FLV	découvre : de.ku.vʁə,
3	VOLV	V.OLV	il trouve : i.tʁuv
			mais de la : me.dlɑ
4	VFLGV	V.FLGV	effroyable : e.fʁwɑ.jɑbl
5	VOLGV	V.OLGV	incroyable : ɛ̃.kʁwɑ.jɑbl
6	VOLOV	VOL.OV	connaître tu : ko.netʁ.ty
			capable parce : kɑ.pɑbl.pɑs

Les règles exceptions 4 et 5 sont liées au statut particulier des clusters Obstruante + Liquide + Glide en français, ces groupes étant le plus souvent homosyllabiques. La règle exception 6 est une exception au « Maximum Onset Principle ». Cette exception est largement motivée par la nature de la parole continue et le fait qu'il n'existe pas, à notre connaissance, de modèles permettant de combiner systématiquement frontière syllabique et frontière lexicale. Ainsi, un cluster tel que Plosive + Liquide + Plosive n'existe pas en interne de mot en français ; malgré tout il apparaît dans notre corpus en raison des phénomènes de réduction caractéristiques de ce type de parole.

Les règles que nous proposons suivent les principes usuels bien connus dans le domaine de la phonologie et peuvent être appliqués à notre corpus comme à d'autres. En ce sens, notre but n'est pas de proposer des règles spécifiques à notre corpus mais bien un principe général, possiblement adaptable à tout autre type de corpus.

Finalement, ce travail a été implémenté sous forme de classes java. Le programme utilise un fichier de configuration qui décrit la liste des phonèmes et leur classes, ainsi que la liste de toutes les règles. Ces paramètres peuvent ainsi être facilement modifiés. La version diffusée en GPL,

nommée *LPL-Syllabeur-v2.1.jar* prend en entrée un fichier praat de phonèmes encodés en SAMPA et rend en sortie un fichier praat de syllabes.

4. EVALUATION

4.1. Description du corpus

Le système décrit dans ce papier a été utilisé pour l'annotation du Corpus of Interactional Data (CID) [3, 4]. Le CID est un corpus de 8 heures d'enregistrements audio et video de dialogues français en parole spontanée (1 heure d'enregistrement par session). Chaque dialogue est constitué de deux participants du même sexe qui étaient amenés à converser autour de l'un des deux thèmes suivants : les conflits dans leur environnement professionnel ou bien les situations insolites qu'ils ont pu connaître. Ces consignes n'étaient pas strictes et les participants ont souvent fait évoluer la conversation vers d'autres thématiques.

L'une des caractéristiques majeure de la parole spontanée est l'écart manifeste que l'on peut observer entre des réalisations standards (« orthographic token ») et la production réelle des locuteurs. Les élisions et réductions telles que « je suis » produit [ʃɥi] ou « je ne sais pas » produit [ʃepɑ] sont extrêmement fréquentes et peuvent être extraites d'un lexique de variantes prototypiques. Mais un corpus conversationnel tel que le CID présente de surcroît de nombreuses variantes non prototypiques qu'il est illusoire de vouloir stocker dans un lexique ([3]). Ces phénomènes rendent la détection automatique de syllabe tout à fait spécifique. Par conséquent, et compte-tenu de l'aspect modulaire du syllabeur proposé, nous avons décidé de résoudre les cas particuliers les plus fréquents de la manière suivante : les enchaînements de phonèmes suivants fs, pt, sk (excepté quand pVsk) n'ont pas été éclatés, car ils correspondent à des unités lexicales très fréquentes (petit, parce que, puisque, faisait, etc).

4.2. Protocole d'évaluation

Dans cette section, nous présentons les résultats d'une comparaison établie entre deux syllabations manuelles et la syllabation automatique du CID. Le corpus de test représente environ 7 minutes d'un dialogue, soit environ 2000 mots (653 d'un locuteur, 1238 du second). Pour les évaluations, nous avons choisi de ne pas prendre en considération les deux principes énoncés dans la section 3.1. Selon le premier principe, les séquences VV ont une segmentation évidente qu'il n'est pas utile d'évaluer. De même, dans l'estimation des performances, nous n'avons pas inclus les cas concernés par le second principe : une voyelle suivie d'une pause, une pause suivie par une autre pause et une pause suivie d'une voyelle.

Les évaluations portent sur 1646 frontières pour lesquelles une décision doit être prise. La table 3 apporte des précisions sur les règles qui sont utilisées et pour lesquelles le système est évalué.

Les segmentations manuelles ont été réalisées par deux annotateurs qui disposaient des phonèmes, et de la transcription. Il est important de rappeler ici que contrairement aux annotateurs, le système automatique ne s'appuie que sur les phonèmes, sans la transcription orthographique. Le taux d'accord inter-annotateur est de 98,60 %, ce qui signifie que 23 frontières proposées sont différentes, sur les

TAB. 3: Statistiques sur l'utilisation des règles (1646 frontières)

Nombre	Règle		Nombre	Règle
1165	VXV			
435	VXXV	dont	54	VXGV
			17	VFLV
			73	VOLV
43	VXXXV	dont	0	VFLGV
			4	VOLGV
			4	VOLOV
3	VXXXXV			

1646 possibles. Ces 23 désaccords se répartissent différemment selon le nombre de consonnes qui sépare deux voyelles : plus leur nombre est important, plus le désaccord augmente, comme on le constate ci-après :
– 5 désaccords en VXV, soit 0,43 % des 1165 cas,
– 12 désaccords en VXXV, soit 2,76 % des 435 cas,
– 5 désaccords en VXXXV, soit 11,63 % des 43 cas,
– 1 désaccord en VXXXXV, soit 33,33 % des 3 cas.

4.3. Qualité de la syllabation

Afin de situer notre proposition par rapport aux systèmes existants, nous avons (1) implémenté les règles de C. Pallier dans un fichier de configuration de notre système, (2) idem pour les règles de la table 1 de [1] et (3) adapté l'outils de J-P. Goldman de sorte qu'il soit applicable à notre corpus (en particulier l'encodage des phonèmes). Enfin, nous avons évalué l'ensemble des outils sur le corpus de test.

La table 4 montre l'ensemble des performances des systèmes automatiques, par rapport à chacun des annotateurs (mentionnés Annot1 et Annot2). On y constate que la syllabation que nous proposons offre un gain relatif d'environ 30-35 % par rapport aux systèmes de l'état de l'art. Plus encore que ce gain de performance, l'outil que nous avons développé a vocation à être modulable à volonté en fonction du contexte (encodage des phonèmes, formats des fichiers, etc.), et diffusé librement.

TAB. 4: Nombres de désaccords et pourcentages

	Annot. 1	Annot. 2
syllabation.awk	74	84
(de [8])	4,50 %	5,10 %
graphon+	85	92
(de [1])	5,16 %	5,59 %
syllabify2.praat	67	75
(de [6])	4,07 %	4,56 %
LPL-syllabeur-v2.1.jar	43	53
(système proposé)	**2,61 %**	**3,22 %**

4.4. Analyse des différences

Comme présenté en table 3, on observe que 97,21 % des frontières syllabiques sont concernées par les règles suivantes : VXV, VXXV, VXGV, VOLV et VFLV. Ces règles ne présentent pas d'ambiguïté (voir les détails en table

5). En conséquence, les résultats de la syllabation automatique devraient être totalement conformes aussi bien à une segmentation selon des règles phonotactiques qu'aux intuitions des auditeurs. Le problème majeur pour la syllabation reste les cas où deux voyelles sont séparées par plus de deux consonnes. Ces occurrences sont rares (2,61 %) et la plus fréquente est VXXXV (dans laquelle sont exclus les cas où C2 est une plosive ou une fricative). Notre proposition (à l'instar du système de règles de Goldman) est de mettre la frontière syllabique entre C1 et C2. Les éventuelles erreurs de syllabation seront corrigées après le passage d'un expert.

TAB. 5: Désaccords entre le système proposé et les annotateurs

	Annot. 1	Annot. 2
VXV	5	4
	0,43 %	0,34 %
VXXV+exceptions	26	32
	5,98 %	7,36 %
VXXXV+exceptions	11	15
	25,59 %	34,88 %
VXXXXV	1	2
	33,33 %	66,67 %
Total	43	53

Nous proposons ci-après (table 6) quelques exemples des syllabations produites par notre système comparativement à celles des annotateurs. Il est à noter qu'une grande partie des différences observées entre les syllabations manuelle et automatique sont concentrées aux jonctures de mots. Il semble que les annotateurs humains soient influencés par les frontières lexicales lorsque la syllabation est complexe, comme « parcs c'est » dans l'exemple. Le syllabeur ne tient pas compte des frontières lexicales car il ne dispose pas de cette information. Il n'est pas question de décider *a priori* quelle est la meilleure syllabation, toutefois, l'information fournie par le syllabeur (sans influence lexicale) offre la possibilité d'évaluer les rôles des frontières lexicales et syllabiques indépendamment dans nos futures analyses phonétiques.

TAB. 6: Exemple de syllabation manuelle et automatique

Transcription	et donc on mange sur la baignoire donc c'est c'est ça
Phonèmes	e d ɔ̃ k ɔ̃ m ɑ̃ ʒ s y ʁ l a b e n w a ʁ d ɔ̃ k s e s e s a
Classes	V O V O V N V F F V L L V O V N G V L O V O F V F V F V
Syllabes (Auto & Annot)	e . dɔ̃ . kɔ̃ . mɑ̃ʒ . syʁ . la . be . nwaʁ . dɔ̃ k . se . se . sa
Transcription	non dans les parcs c'est un peu limité
Phonèmes	n ɔ̃ d ɑ̃ l e p ɑ ʁ k s e t œ̃ p ə l i m i t e
Classes	N V O V L V O V L O F V O V O V L V N V O V
Syllables Auto	nɔ̃ . dɑ̃ . le . pɑʁ . kse . tœ̃ . pə . li . mi . te
Syllables Annot1 & Annot2	nɔ̃ . dɑ̃ . le . pɑʁk . se . tœ̃ . pə . li . mi . te
Transcription	il expliquait pas vraiment ce qu' il y avait dedans
Phonèmes	i l e k s p l i k e p ɑ v r e m ɑ̃ s k i j a v e d ɑ̃
Classes	V G V O F O L V O V O V F L V N V F O V O G V F V O V
Syllabes Auto	i . lek . spli . ke . pɑ . vre . mɑ̃ . ski . ja . ve . dɑ̃
Syllabes Annot1	i . lek . spli . ke . pɑ . vre . mɑ̃ . ski . ja . ve . dɑ̃
Syllabes Annot2	i . leks . pli . ke . pɑ . vre . mɑ̃ . ski . ja . ve . dɑ̃

5. CONCLUSION

Dans cet article, nous avons proposé un outil permettant la syllabation, à partir de phonèmes, d'un corpus de dialogue oral spontané. Ses performances ont été évaluées et jugées satisfaisantes pour les tâches ultérieures auxquelles il est dédié. Nous avons regroupé les phonèmes dans des classes et défini un ensemble de règles générales suivies d'exceptions afin de trouver les frontières pertinentes entre les syllabes. Cet outil est destiné à être appliqué sur l'ensemble du CID afin d'être mis en relation avec les autres niveaux d'annotation existants. Mais ce niveau d'annotation va permettre également d'effectuer de nouvelles annotations telles que celles par exemple de la structure syllabique (onset, noyau, coda) qui n'a pas été beaucoup étudiée sur ce type de parole.

RÉFÉRENCES

[1] M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel. Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, 46 :119–139, 2005.

[2] C. Astesano. *Rythme et accentuation en Français*. L'Harmattan, coll. Langue et parole, Invariance et variabilité stylistique, 2001.

[3] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. Le cid - corpus of interactional data. *Traitement Automatique des Langues*, 49(3) :105–134, 2008.

[4] P. Blache, R. Bertrand, and G. Ferré. Creating and exploiting multimodal annotated corpora : the toma project. *Multimodal Corpora*, LNAI 5509 :38–53, 2009.

[5] P. Boersma and D. Weenink. Praat : doing phonetics by computer, http ://www.praat.org.

[6] J-P. Goldman. http ://latlcui.unige.ch/phonetique/, 2007.

[7] V-A. Kozhevnikov and L-A. Chistovitch. *Speech Articulation and Perception*, volume 30. Washington D.C. : Joint Publications Research Service, 1965.

[8] C. Pallier. Syllabation des représentations phonétiques de brulex et de lexique, http ://www.pallier.org, 1999.

[9] K-L. Pike. *The intonation of American English*. The University of Michigan Press in Linguistics 1, 1945.

Suffixes complexes : quand c'est fini, ça recommence…

Rémi Godement et Philippe Martin

CLILLAC-ARP EA3967, UFRL, Université Paris Diderot
30, rue du Château des Rentiers, 75013 Paris, France
remi.godement@linguist.jussieu.fr; philippe.martin@linguist.jussieu.fr

ABSTRACT

Recently, prosodic studies have started to deal with the interactions between intonation and macrosyntax, allowing a better interpretation of spontaneous speech data. This paper focuses on a particular macrosegment, the *suffix*, which follows the kernel macrosegment. After introducing the theoretical framework and a few regular types of suffixes, we present some interesting examples where macrosegments apparently planned as suffixes turn out to be realized as prefixes (macrosegments placed *before* the kernel). This eventually leads us to propose the use of an enriched type of unit: the *complex suffix*, itself composed of a prefix and a kernel.

1. INTRODUCTION

Les études sur l'intonation se sont récemment réorientées sur les rapports entre prosodie et macrosyntaxe, la macrosyntaxe permettant de mieux analyser les séquences habituellement observées dans la production de parole spontanée. En effet, paradoxalement, bien que les théories syntaxiques « classiques » censées correspondre aux intuitions linguistiques du sujet parlant natif aient connu un développement considérable depuis les 50 dernières années, elles ne parviennent que difficilement à rendre compte des réalisations de parole non préparée.

Après un rapide rappel des principes de l'analyse macrosyntaxique, inspirée des travaux de C. Blanche-Benveniste [1], nous nous intéressons ici à l'une des unités (*macrosegments*) employées dans cette analyse, le *suffixe*, qui se trouve à la suite du *noyau*. Nous en examinons des exemples canoniques, puis des exemples mettant en lumière les variations de planification syntaxique faites par le locuteur : un changement dynamique (se produisant en cours de réalisation) amène un macrosegment apparemment prévu comme un suffixe à être finalement réalisé comme un préfixe (macrosegment *précédant* le noyau). Ces variations seront analysées dans leur interaction avec l'intonation, et en particulier les variations mélodiques à l'endroit des syllabes accentuées, selon les lignes directrices données dans [5]. Le raisonnement nous conduira non pas à considérer que le suffixe s'est transformé en préfixe mais plutôt à envisager l'existence d'un macrosegment enrichi : le *suffixe complexe*, lui-même composé d'un préfixe et d'un noyau.

2. MACROSYNTAXE

Décrite très succinctement, la macrosyntaxe rend compte de la structure d'un énoncé par une analyse en macrosegments, correspondant, en première approximation, à une expansion maximale des relations syntaxiques « classiques ». Un macrosegment est donc « bien formé » au sens de la grammaire classique. Vu sous l'angle de grammaires de dépendance, les frontières des macrosegments correspondent à une rupture d'une dépendance envers une unité adjacente, vers la gauche pour la frontière gauche du macrosegment, et vers la droite pour la frontière droite. Ainsi, dans un énoncé tel que *moi mon père il est président*, présenté classiquement comme une dislocation à gauche, l'analyse macrosyntaxique distingue trois macrosegments : *moi*, *mon père* et *il est président*. De même, dans *il est président mon père*, apparaissent deux macrosegments : *il est président* et *mon père*.

Cependant, on peut aussi privilégier la notion de *macrosegment prosodique* et considérer qu'un énoncé bien formé syntaxiquement comme *j'ai pris mes clés et je suis parti* est constitué non pas d'un mais de deux macrosegments, et ce pour des raisons prosodiques (ex : contour montant puis descendant, pause au milieu).

Dans l'analyse macrosyntaxique, un macrosegment, appelé *noyau*, a un statut particulier en ce qu'il peut apparaître seul comme énoncé car il est bien formé à la fois syntaxiquement (au sens classique) et prosodiquement. On peut en changer la modalité (de déclarative en interrogative, de positive en négative, etc.) et il constitue une unité illocutoire autonome. En isolant le noyau dans un enregistrement par un éditeur de signal, on doit donc également percevoir un énoncé complet, se terminant par un contour conclusif déclaratif ou interrogatif (ou une variante impérative, implicative, de doute ou de surprise [5]).

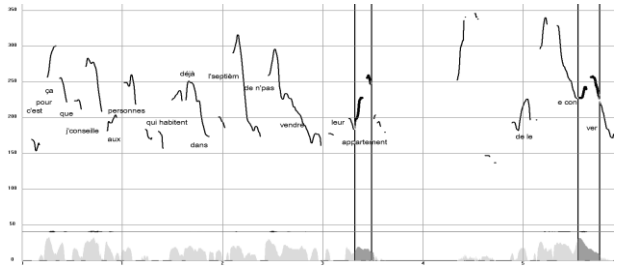
Les macrosegments placés devant le noyau sont appelés *préfixes* [1], à l'intérieur du noyau *incises* (ou parenthèses), et après le noyau *postfixes* ou *suffixes*. Postfixes et suffixes se différencient par le type de relation (manifestant donc une dépendance à gauche) qui les unit au noyau qui précède : les postfixes, qui rappellent une information, sont liés au noyau par la structure prosodique (en particulier par le contour final, plat dans le cas déclaratif et montant dans le cas interrogatif). Les suffixes, qui ajoutent une information, se rattachent au

noyau par une relation syntaxique (ou éventuellement sémantique « forte »).

C'est alors la structure prosodique, instanciée essentiellement par des contours mélodiques à l'endroit des syllabes accentuées, qui va déterminer *in fine* l'agencement hiérarchique des unités macrosyntaxiques, le noyau contenant l'information prosodique relative à la modalité de l'énoncé.

3. EXEMPLES CANONIQUES DE SUFFIXES

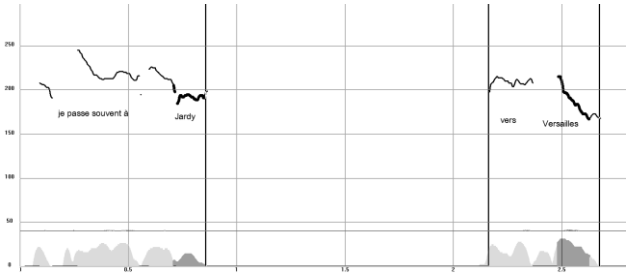
Les exemples canoniques qui suivent permettront de mieux comprendre les principes d'analyse macrosyntaxique ainsi que l'effet de la structure prosodique sur l'agencement des macrosegments. Tous les exemples sont extraits du CFPP2000 (Corpus de Français Parlé Parisien des années 2000, [3]), corpus *« composé d'un ensemble d'interviews sur les quartiers de Paris et de la proche banlieue »* à partir d'un *« questionnaire portant sur la ville »* [3].



(1) c'est pour ça qu'je conseille aux personnes qui habitent déjà dans l'septième de n'pas vendre leur appartement [rires] – de le conserver [rires] (CFPP2000, 07-02)

Le suffixe est *de le conserver*. Il est lié syntaxiquement au long noyau qui le précède par la préposition *de*. Le contour intonatif est le même sur les deux macrosegments, et il est intéressant de noter qu'ils sont également suivis du même rire (ou soupir). Le suffixe copie non seulement la prosodie du noyau mais aussi ses éléments extralinguistiques. On retrouve ce phénomène dans l'exemple (4) ci-contre.

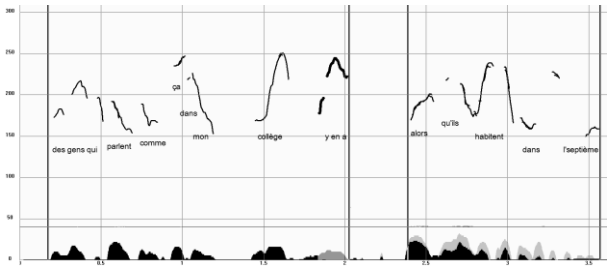
Ici, on peut débattre sur le rôle informationnel de *de le conserver* : s'agit-il d'un ajout, comme le sont habituellement les suffixes, ou d'une simple reprise ? Syntaxiquement, il s'agit d'une reprise : le syntagme repris est le complément d'objet direct du verbe *conseiller*, le groupe prépositionnel *de n'pas vendre leur appartement*. Mais on peut plutôt interpréter *de le conserver* comme un ajout dans le sens où la reformulation de *de n'pas vendre leur appartement* n'a pas tant pour but de le corriger que de rajouter, justement, un trait d'humour à celui que constitue le noyau.



(2) j'suis partie au Haras de Jardy – vers Versailles (CFPP2000, 07-02)

Dans cet énoncé, une pause relativement longue sépare le noyau *j'suis partie au Haras de Jardy* du suffixe *vers Versailles.* On peut donc interpréter ce dernier comme une réponse à la question qui pourrait être posée suite au noyau («Où c'est ?», «C'est vers où ?»). La locutrice anticipe la question par un suffixe. Elle n'est pas la seule : une seconde locutrice, que l'on entend dans le fond, dit «*vers Versailles*» exactement en même temps qu'elle.

Il arrive d'ailleurs qu'un noyau et un suffixe soient vraiment réalisés par deux locuteurs différents, comme en (3) :

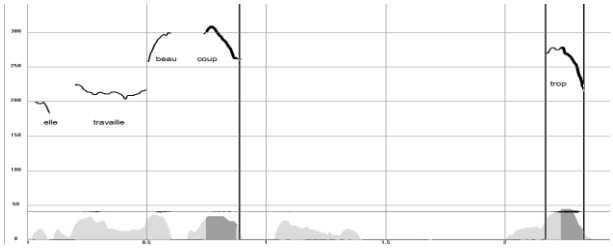


(3) A – des gens qui parlent comme ça dans mon collège y en a oui

B – alors qu'ils habitent dans le septième (CFPP2000, 07-02)

La locutrice B juge incomplet l'énoncé de la locutrice A et y apporte l'information manquante.

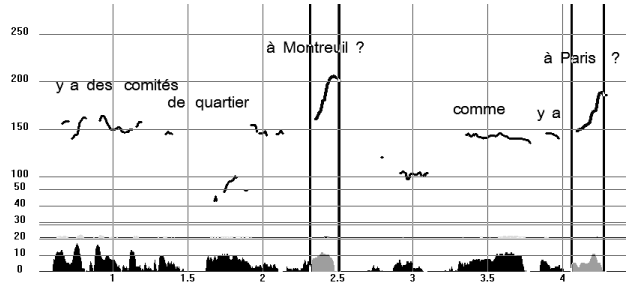
L'énoncé (4) présente un exemple de suffixe composé d'un seul mot :



(4) elle travaille beaucoup [rires] – trop [rires] (CFPP2000, 07-02)

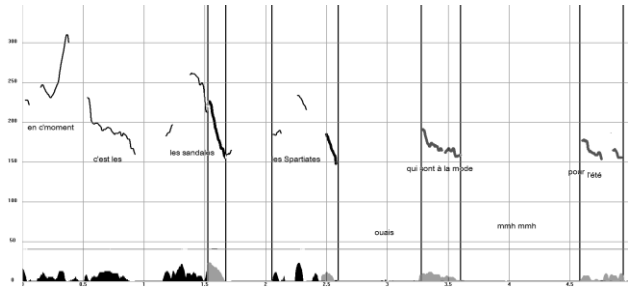
Étant composé d'un seul mot, il ne dispose d'aucune marque syntaxique pour le relier au noyau, si ce n'est lui-même, et c'est ce qui se passe : l'adverbe *trop* modifie le verbe *travaille*. Mais il ne le modifie pas tout seul : il est

d'abord relié à l'adverbe *beaucoup*. En effet, ici *trop* ne remplace pas *beaucoup* mais forme avec lui un groupe adverbial. Autrement dit, si l'on répétait la phrase deux fois, on n'aurait pas *Elle travaille beaucoup*. *Elle travaille trop*. mais *Elle travaille beaucoup. Elle travaille beaucoup trop*. C'est ce qui rend l'énoncé remarquable : avant et après la réalisation de *trop*, l'élément modifié par *beaucoup* n'est pas le même : dans *elle travaille beaucoup*, c'est *travaille*, mais dans *elle travaille beaucoup trop*, c'est *trop*. La locutrice réorganise la structure syntaxique de la phrase sans s'en rendre compte, peut-être parce que *beaucoup trop* a un statut de quasi-locution et que le premier adverbe appelle facilement le second.



(5) y a des comités de quartier à Montreuil ? comme y a à Paris ? (CFPP2000, Mo-01)

Cette fois, le noyau et le suffixe sont interrogatifs. Le contour est montant à chaque fois.



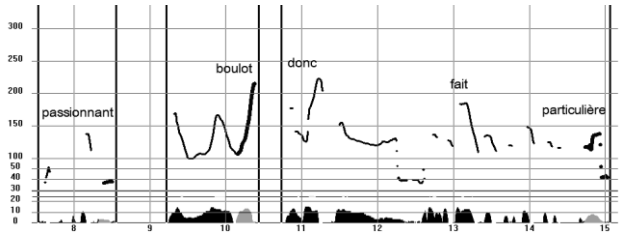
(6) en c'moment c'est les les sandales – les Spartiates – qui sont à la mode – pour l'été (CFPP2000, 07-02)

Cet exemple est complexe : il s'agit d'une suite Noyau + Suffixe + Postfixe + Postfixe. Le macrosegment *les Spartiates* ne contient pas de marque syntaxique le reliant au noyau, mais c'est un suffixe pour deux raisons : a) son contour intonatif, qui reprend celui de *les sandales* et b) il s'agit bien un ajout puisqu'il apporte une précision (sur la nature des sandales : *Spartiate* est hyponyme de *sandale*)

Quant *à qui sont à la mode* et *pour l'été*, le fait qu'ils commencent respectivement par un pronom relatif et une préposition pourrait les apparenter à des suffixes, mais ils portent tous deux le contour plat typique des postfixes (l'information relève du rappel et non de l'ajout, d'où une mélodie à faible variation).

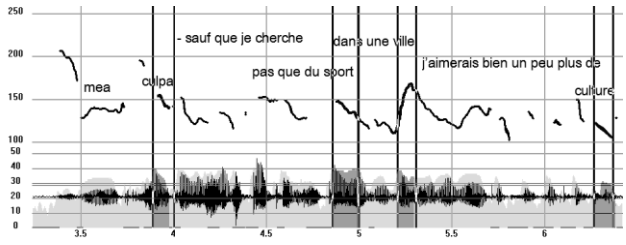
4. SUFFIXES COMPLEXES

A première vue, les exemples qui suivent [1] posent un problème d'analyse. Il s'agit de macrosegments qui, comme les suffixes, apparaissent à la suite d'un noyau, sont syntaxiquement dépendants de ce noyau et apportent une information nouvelle. Mais, contrairement aux suffixes, ils ne portent pas le même contour que celui du noyau, qui est descendant dans les exemples déclaratifs que nous avons sélectionnés. A la place, ils portent un contour montant *typique des préfixes* et sont suivis soit directement d'un noyau, soit d'un ou de plusieurs préfixes puis d'un noyau. On a donc une contradiction entre, du point de vue textuel, une séquence Noyau + Suffixe + Noyau et, du point de vue prosodique, une séquence Noyau + Préfixe + Noyau.



(7) c'était assez passionnant – <u>mais un énorme boulot</u> et donc euh ça s'est fait dans une atmosphère très particulière (CFPP2000, 07-03)

L'énoncé (7) commence par un noyau, *c'était assez passionnant*. Ce noyau est suivi d'une pause, puis d'un macrosegment que l'on prend d'abord pour un suffixe, et dont on attend par conséquent une intonation descendante, la même que celle du noyau. Mais sur la syllabe *–lot* de *boulot*, la locutrice place un contour montant et enchaîne avec un second noyau, *et donc euh ça s'est fait dans une atmosphère très particulière*.

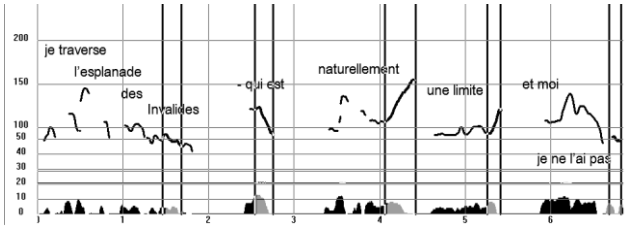


(8) au niveau du sport mea culpa <u>– sauf que je cherche pas que du sport dans une ville</u> j'aimerais bien un peu plus de culture (CFPP2000, KB-01)

Au départ, la configuration est la même qu'en (7) : un noyau puis un macrosegment qui commence comme un suffixe. *sauf que je cherche pas que du sport dans une ville* pourrait en effet très bien porter un contour final descendant, le même que celui de *mea culpa*, et l'énoncé s'arrêterait là. Mais le contour est finalement montant, ce

[1] Notons d'emblée que ces exemples sont minoritaires parmi toutes les réalisations de suffixes que nous avons rencontrées (environ 10%).

qui donne un préfixe, suivi d'un second noyau, *j'aimerais bien un peu plus de culture*.



(9) je traverse l'esplanade des Invalides – qui est naturellement une limite et moi je ne l'ai pas (CFPP2000, 07-03)

Encore une fois : noyau (*je traverse l'esplanade des Invalides*), macrosegment suffixe/préfixe (*qui est naturellement une limite)* puis à nouveau noyau (*et moi je ne l'ai pas*).

Comment analyser ces exemples ? Il ne s'agit pas de phénomènes de répétitions de lexique [2], ce qui serait le cas si on avait, pour reprendre l'énoncé (8) : *« Sauf que je cherche pas que du sport dans une ville. En plus du sport, j'aimerais bien un peu plus de culture »* Dans nos exemples, c'est la prosodie, et non la syntaxe, qui se charge de la progression thématique [4].

Une solution serait de considérer ces macrosegments comme des *suffixes se transformant en préfixes*. Ce qui ne serait pas si étonnant : on peut imaginer que le locuteur commence un suffixe pour ajouter une information au noyau, puis décide finalement d'enchaîner aussitôt avec autre chose, ou bien se rende compte que ce suffixe peut servir de préfixe à un nouveau noyau. Mais l'écoute et la lecture des énoncés ci-dessus, et d'autres encore, donne plutôt l'impression que ces macrosegments sont *dès le départ* reliés informationnellement au noyau qui suit, qu'il n'y a pas de changement d'avis du locuteur « en cours de route ». Et pourtant, syntaxiquement, c'est bien au noyau qui *précède* qu'ils sont reliés.

Pour résoudre cette contradiction, une analyse possible consiste à envisager l'existence d'un *suffixe complexe*, commençant par un *petit préfixe* et finissant par un *petit noyau* (avec éventuellement d'autres *petits préfixes*, voire des incises, au milieu). En appliquant cette analyse aux exemples ci-dessus, on obtiendrait :

(7) c'était assez passionnant
NOYAU

mais un énorme boulot | et donc euh […] particulière
préfixe | noyau
SUFFIXE COMPLEXE

(8) mea culpa
NOYAU

sauf que […] une ville | j'aimerais bien […] culture
préfixe | noyau
SUFFIXE COMPLEXE

(9) je traverse l'esplanade des Invalides
NOYAU

qui est naturellement une limite | et moi je ne l'ai pas
préfixe | noyau
SUFFIXE COMPLEXE

Le « suffixe complexe » justifie à la fois a) que le macrosegment par lequel il commence (celui qu'on a d'abord qualifié de « suffixe se transformant en préfixe ») soit syntaxiquement relié au noyau qui précède (puisqu'il fait partie du suffixe complexe) et b) qu'il soit cependant informationnellement relié au noyau qui suit (puisque c'est un préfixe). C'est désormais le suffixe complexe *dans son entièreté* qui est un ajout informationnel au noyau qui précède.

5. Conclusion

Nous avons examiné ici des cas particuliers de macrosegments apparemment hybrides entre suffixes et préfixes, ce qui nous a finalement amenés à proposer une catégorie enrichie de suffixes, le *suffixe complexe*, contenant un *petit préfixe* et un *petit noyau*. Il est possible que les autres macrosegments de base (noyau, préfixe, postfixe, incise) possèdent eux aussi des formes complexes contenant des sous-macrosegments. D'une manière générale, en affinant ses unités, le découpage macrosyntaxique pourrait – à l'aide de la structure prosodique, indispensable à ce découpage – rendre compte de manière encore plus efficace des phénomènes rencontrés en parole spontanée.

Références

[1] C. Blanche-Benveniste. *Approches de la langue parlée en français*. Ophrys, Paris, 2000.

[2] C. Blanche-Benveniste. « Répétitions de lexique et glissement vers la gauche », *Recherches sur le français parlé*, n°12, 1993.

[3] S. Branca-Rosoff, S. Fleury, F. Lefeuvre, M. Pires *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)* http://ed268.univ-paris3.fr/CFPP2000/

[4] B. Combettes. « Grammaire de phrase, grammaire de texte : le cas des progressions thématiques », *Pratiques*, n°77, pp.43-57, 1993.

[5] Ph. Martin. *Intonation du français*. Armand Colin, Paris, 2009.

Le F0 intrinsèque des voyelles est-il aussi suprasegmental ?

Olivier Piot

LLF (CNRS UMR 7110 & Université Paris 7)
olivier.piot@gmail.com

ABSTRACT

Intrinsic F0, i.e. the tendency for high vowels (such as /i/ and /u/) to have higher F0 than low vowels (such as /a/), has been found in every language in which it has been sought so far. In the present study, we tested it on the three French front vowels V = /i/, /e/ and /ɛ/, embedded in pseudo-words /d/V/d/V and /a/V/a/V. With the latter phonemic context, we wanted to assess the potential influence of articulatory movement amplitude to and from the target vowel, since in this particular context, and *contrary* to the former 'standard' one, this amplitude is higher for higher vowels. The main finding of this study is that the intrinsic F0 hierarchy, which is produced by 7 out of 8 speakers in the 'dVd' context, is *neutralized* or even *reversed* in the 'aVa' context. These results run counter the predictions of the current IF0 theories.

Keywords: Intrinsic F0, French.

1. INTRODUCTION

Il est bien connu, et largement accepté, que la fréquence fondamentale des voyelles hautes est plus élevée que celle des voyelles basses homologues. Cet effet, appelé F0 intrinsèque (désormais IF0), a été rapporté dans plus de 30 langues, et se retrouve dans les voyelles isolées, les mots isolés, ainsi que les mots insérés dans une phrase porteuse [1]. Plusieurs explications ont été proposées pour expliquer ce phénomène, qui sont d'ordre physiologique, acoustique et/ou cognitif (cf. [7] pour une étude détaillée). Tout d'abord, il a été démontré que l'activité musculaire crico-thyroïdienne est augmentée avec l'aperture des voyelles [2]. Et puisque cette activité est reconnue comme étant un paramètre majeur pour augmenter F0 [3], l'idée a été proposée que IF0 soit la simple conséquence d'une différence dans l'activité des muscles du larynx. Mais aucune explication satisfaisante de cette différence n'a encore été proposée. L'explication par une augmentation acoustique délibérée de la dimension de hauteur vocalique [4] est contredite par un certain nombre d'études. Tout d'abord, elle apparaît incompatible avec le fait, défendu par [1] comme étant universel, que IF0 disparaît dans le bas du registre tonal des locuteurs, ainsi qu'à la fin de leurs énoncés [6]. Il est plus que douteux qu'une telle amplification acoustique de la dimension de l'aperture vocalique soit moins utile dans le bas du registre tonal ou à la fin des énoncés. Ensuite, IF0 est déjà présent dans le babillage des enfants prélinguistiques [5], suggérant fortement qu'il s'agit d'une conséquence automatique de l'articulation. Enfin, l'argument basé sur les mesures EMG pour défendre un IF0 réalisé délibérément, est contredit par [7], qui montre que la relation EMG(CT)-F0 est significativement différente selon que la voyelle (isolée) est haute ou basse.

Outre cette hypothèse, encore insuffisamment intelligible, d'un IF0 résultant de différences dans les patterns d'activité des muscles laryngaux, l'explication la plus convaincante est certainement celle du 'tongue-pull' [8]. Elle suppose que pour les voyelles hautes, la langue exerce une traction sur le larynx par le biais de certains tissus laryngaux, ce qui tend les cordes vocales, augmentant ainsi F0. Cette hypothèse, en plus de sa valeur explicative, est en accord avec l'automaticité apparente de IF0. Mais elle est également contredite par certains faits. Tout d'abord, elle est en contradiction avec l'absence susmentionnée de IF0 dans le bas du registre tonal, ainsi qu'en fin d'énoncé. Ensuite, la hauteur de la langue n'est pas un bon prédicteur de IF0 en allemand, où les voyelles 'tense' et 'lax' ont des valeurs similaires de IF0, tandis que la hauteur de la langue est sensiblement plus élevée pour les premières que pour les secondes.

De toute évidence, de nouvelles données sont nécessaires pour développer cette discussion. Or, tandis que les études existantes sur IF0 ont été faites à partir de voyelles isolées, ou incorporées dans des syllabes de types C(C)(C)V(C)(C), aucun contexte phonémique n'a jusqu'à présent été utilisé dans lequel /i/ (par exemple) serait atteint par un geste lingual de plus grande amplitude que /e/. A ce sujet, notons que dans [11] est défendue l'hypothèse de l'existence de 'relations d'échange' entre les différents paramètres articulatoires de la parole, une activité augmentée sur un groupe musculaire entraînant, ceteris paribus, une diminution de l'activité disponible pour les autres groupes musculaires. Dans un tel contexte phonémique, ces relations d'échange pourraient favoriser la tendance inverse à celle de IF0. Par conséquent, on peut se poser la question des valeurs relatives du F0 de /i/ et /e/ dans des séquences de phonèmes comme /aia/ et /aea/, où cette condition serait a priori remplie. L'hypothèse du 'tongue-pull', ainsi que celle d'une planification délibérée, prédisent que la hiérarchie sur IF0 resterait inchangée dans ce contexte (i.e. F0(/i/)>F0(/e/)). Afin de tester ces prédictions, nous avons réalisé une expérience comparant les F0 des trois voyelles françaises antérieures non arrondies V = /i/, /e/ et /ɛ/, dans les deux contextes phonémiques /d/V/d/ et /a/V/a/.

2. METHODE

Parce que nous voulions étudier l'influence sur IF0 de l'amplitude articulatoire, et plus généralement de l'effort articulatoire, le débit a été inclus comme facteur dans l'expérience. Cependant, un débit plus élevé ne devait pas résulter d'un affect générant une plus grande excitation (cause *tonique*), comme avec la colère ou la joie, mais être suscité par le contexte de la communication (cause *phasique*) [9]. Nous avons donc utilisé trois contextes différents, dans lesquels le locuteur devait répondre à une question posée par l'expérimentateur. Dans la première condition, le locuteur devait répondre à la question d'une façon "neutre", "usuelle". Dans la deuxième condition, l'allocutaire était censé être un américain natif, ne parlant pas très bien français : le locuteur a été invité à utiliser un débit diminué afin de faciliter la perception de son allocutaire étranger. Dans la troisième condition, le locuteur devait utiliser un débit de parole augmenté, parce qu'il/elle pouvait se le permettre - l'allocutaire étant alors censé être un locuteur natif du français.

Les stimuli étaient les pseudo-mots /didi/, /dede/, /dɛdɛ/, insérés dans la phrase porteuse : “Il a dit /dVdV/ déjà.”, ainsi que /aiai/, /aeae/, /aɛaɛ/ insérés dans la phrase : “Il a dit /aVaV/ à haute voix.” Ces deux paradigmes sont par la suite nommés dVd et aVa (resp.). Les pseudo-mots cibles étaient écrits avec leur transcription phonétique. Les phrases porteuses ont été choisies afin que le phonème qui suit immédiatement le mot cible soit celui alternant avec V (i.e. /d/ dans le premier cas, /a/ dans le second). Ce choix a été fait afin d'étendre le cycle des mouvements linguaux à la syllabe qui suit le mot cible, diminuant ainsi l'influence de la coarticulation anticipatoire sur la réalisation du mot cible. F0 était mesurée au centre de la voyelle finale du mot cible, afin de limiter l'influence de la coarticulation avec le mot précédent. Pour chacun des trois débits de parole (normal, augmenté et abaissé), un bloc de 6 phrases, correspondant aux 6 mots cibles, était présenté 7 fois, chaque fois dans un ordre aléatoire différent (recalculé pour chaque locuteur), tout en assurant une alternance régulière des deux contextes dVd et aVa. L'expérience était divisée en 3 parties, chacune correspondant à l'un des 3 débits. La première partie était toujours celle avec un débit neutre, et l'ordre des deux autres parties était modifié après chaque session.

Les enregistrements ont été réalisés dans une chambre anéchoïque, avec un micro professionnel placé à environ 20 cm de la bouche du locuteur. Les phrases étaient écrites seules sur une feuille de papier plastifié, et se succédaient dans un classeur. L'expérimentateur demandait au locuteur "Il a dit quoi déjà ?" dans le contexte dVd, et "Il a dit quoi à haute voix ?" dans le contexte aVa, de façon à aiguillonner l'accentuation sur le seul élément d'information nouveau, i.e. le mot cible. Les locuteurs ont été informés du fait que, dans le contexte de cette communication simulée, le mot cible pouvait aussi bien être /dadi/ ou /dide/, c'est à dire que les deux syllabes étaient potentiellement informatives, et que la deuxième n'était pas inférable à partir de la première. L'intonation attendue était alors un 'hat pattern' sur le mot cible, ce que les locuteurs ont accepté comme étant une façon naturelle de répondre. Ainsi, la voyelle au centre de laquelle F0 était mesuré portait un contour intonatif 'haut-plat'. Aucune pause n'était admise entre le mot cible et le reste de l'énoncé, en raison de son influence potentielle sur la réalisation de la voyelle cible.

Huit locuteurs, quatre hommes (M1, M2, M3 et M4) et quatre femmes (W1, W2, W3 et W4), ont participé à cette expérience. Ils étaient tous étudiants en linguistique, monolingues de langue maternelle française (ayant vécu principalement en Ile de France), et de parents français. Un contexte de communication approprié était utilisé pour chacun des trois débits, afin de rendre la tâche expérimentale plus facile et réaliste. Avant l'enregistrement, un entraînement était fait sur quelques exemples, afin de s'assurer que le locuteur se sentait à l'aise avec la tâche, et qu'il la réalisait correctement. Au cours de la session d'enregistrement, chaque fois que l'expérimentateur ou le locuteur estimait que la tâche expérimentale n'avait pas été pleinement respectée, l'enregistrement (y compris la question précédente) était effectué de nouveau, jusqu'à ce qu'une réalisation acceptable soit produite. De tels "faux pas" pouvaient être causés par une erreur de prononciation, une intonation incohérente, une pause non désirée ou un bruit de fond. Les locuteurs ont globalement réussi à suivre les instructions sans difficulté. Cependant, il est arrivé qu'un énoncé soit accepté à tort par l'expérimentateur et par le locuteur. Dans ce cas l'énoncé a été exclu du corpus. Le nombre de ces exclusions, sur un nombre total de 126 énoncés par locuteur, était de 0 pour W2, M2, et W4, 2 pour W3, 3 pour M3, 6 pour W1 et M1, et 7 pour M4.

Les enregistrements ont été transférés numériquement du support DAT vers un disque dur au format originel (44,1 kHz, 16 bits). La voyelle finale de chaque mot cible a été étiquetée à la main en utilisant le programme PRAAT [10]. F0 a été mesurée au centre de la voyelle, en utilisant la fonction d'autocorrélation standard de PRAAT. Chaque fois qu'une valeur inattendue de F0 a été calculée, elle s'est toujours avérée être due à un calcul une octave en dessous de la valeur réelle, et a alors été corrigée manuellement.

3. RESULTATS

Les valeurs moyennes globales de F0 pour chaque locuteur sont représentées sur la figure 1. On peut constater que, dans le contexte dVd, la corrélation attendue entre F0 et aperture vocalique est observée pour chaque sujet, à l'exception de M1 qui ne produit pas de IF0 notable. Mais dans le contexte aVa, la hiérarchie classique sur IF0 apparaît être soit atténuée ou neutralisée (locuteurs W2, W3, M2 et M3), soit même inversée (locuteur W1, et peut-être W4, M1 et M4). Pour vérifier si ces tendances sont significatives ou non, des analyses ANOVA ont été effectuées avec F0 comme variable

dépendante, et le contexte phonémique ainsi que l'aperture des voyelles comme variables indépendantes. Des t-tests appariés ont également été utilisés pour évaluer l'influence de l'aperture dans le contexte aVa.

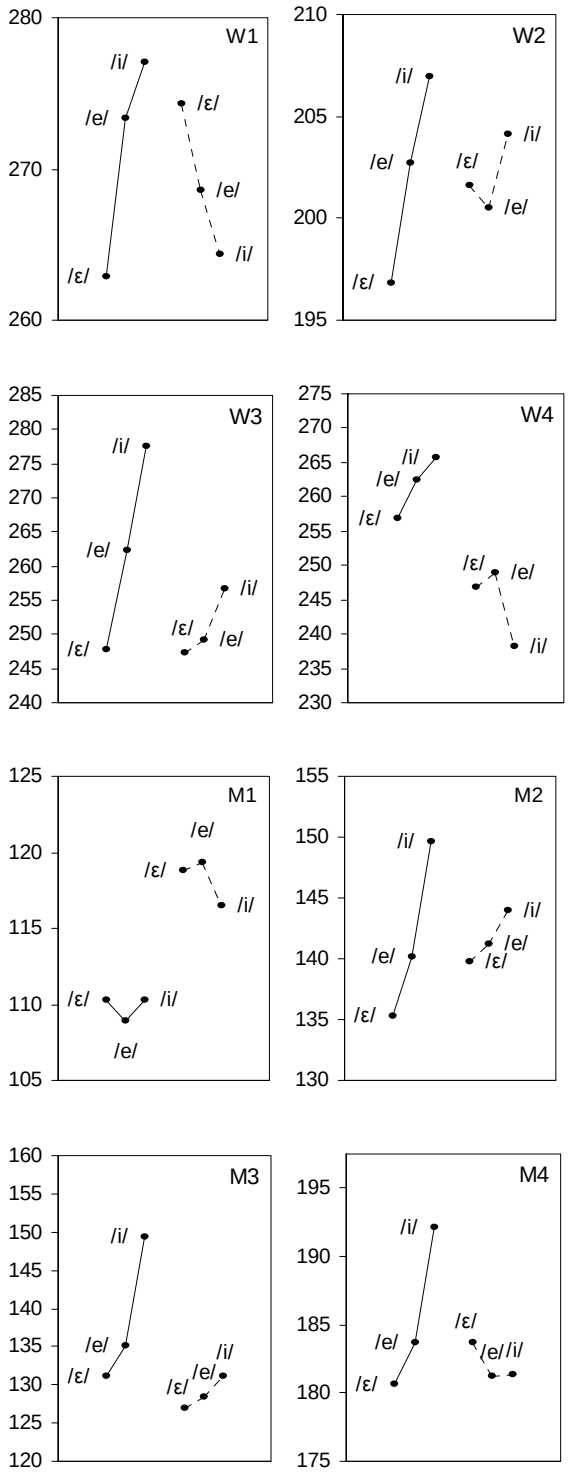


Figure 1 : F0 moyenne (en Hz) au centre des trois voyelles /ɛ/, /e/ et /i/, dans les contextes dVd (ligne continue) et aVa (ligne pointillée). Le nombre de mesures par point va de 18 à 21.

L'effet du contexte phonémique (aVa ou dVd) sur IF0 est confirmé pour tous les locuteurs sauf M1, à travers le terme d'interaction aperture-contexte de l'ANOVA à deux facteurs, comme on peut le voir sur le tableau 1. Dans le contexte dVd, la hiérarchie 'standard' sur IF0 est respectée pour tous les sujets [$F(2;55 \leq n \leq 60) \geq 5.5$, $p < 0.01$], sauf pour M1 [$F(2,55) = 0,48$, $p > 0,6$]. Cela signifie que M1 ne présente pas de IF0 dans ce contexte. Cela montre aussi que les 7 autres locuteurs présentent la hiérarchie attendue sur IF0 dans le contexte dVd, tandis que le contexte aVa tend significativement à inverser cette hiérarchie. En particulier, IF0 y est *inversé* pour W1 [$F(2,56) = 5.67$, $p < 0.01$], et cette inversion est avérée pour chacune des trois paires de voyelles ($p < 0,05$). La hiérarchie attendue est également *inversée* pour W4 dans le contexte aVa [$F(2,60) = 10.45$, $p < 0.0001$], bien que cette inversion ne soit observée que pour les paires /i/-/e/ et /i/-/ɛ/, et pas pour la paire /e/-/ɛ/ ($0.1 < p < 0.2$). Ajoutons que si M1 n'affiche pas de IF0 dans le contexte aVa [$F(2, 59) = 3.00$, $0.05 < p < 0.1$], ses F0(/e/) et F0(/ɛ/) y sont plus élevés que son F0(/i/) ($p < 0,01$).

Tableau 1 : Interaction aperture-contexte d'une ANOVA à deux facteurs, avec F0 comme facteur dépendant.

Locuteur	F	p
W1	$F(2,120) = 18.44$*	$p < 0.0001$
W2	$F(2,126) = 4.65$*	$p < 0.02$
W3	$F(2,124) = 21.37$*	$p < 0.0001$
W4	$F(2, 126) = 12.69$*	$p < 0.0001$
M1	$F(2,120) = 2.29$	$0.1 < p < 0.2$
M2	$F(2,126) = 5.85$*	$p < 0.01$
M3	$F(2,123) = 10.89$*	$p < 0.0001$
M4	$F(2,119) = 6.34$*	$p < 0.01$

Pour les locuteurs M2, M3, M4 et W2, la hiérarchie qui apparaît dans le contexte dVd n'est plus avérée dans le contexte aVa, où elle est donc *neutralisée* [$F(2,60) = 3.11$, $0.05 < p < 0.1$ pour M2; $F(2, 57) = 2.27$, $0.2 > p > 0.1$ pour M3; $F(2, 58) = 0.50$, $0.6 < p < 0.7$ pour M4; $F(2, 60) = 1.46$, $0.2 < p < 0.3$ pour W2]. La hiérarchie standard ne demeure ainsi que pour W3 [$F(2,58) = 8.38$, $p < 0.001$], et n'est donc qu'*atténuée* pour ce locuteur. Tous les locuteurs ont donc une tendance commune à l'inversion de la hiérarchie standard sur IF0 dans le contexte aVa, même si une inversion réelle n'est obtenue que pour W1, W4 (sauf sur la paire /e/-/ɛ/), et dans une moindre mesure pour M1.

Un point important à considérer est que, avec un débit suffisamment rapide, il y a un effet bien connu de 'target undershoot' qui fait ressembler l'articulation d'une séquence comme /ai/, par exemple, à une transition entre un /ɛ/ et un /e/, lorsque l'on écoute chaque voyelle isolément. Mais l'étude de l'influence du débit sur l'amplitude articulatoire, ainsi que sur l'effort articulatoire, sort du cadre de cette étude introductive.

4. DISCUSSION

L'effet d'inversion du contexte aVa sur IF0 est très robuste : il est significatif pour 7 des 8 locuteurs, tandis que le locuteur pour lequel cet effet n'est pas avéré ne réalise aucune différence de IF0 dans le contexte dVd, mais produit deux inversions de la hiérarchie dans le contexte aVa (F0(/i/)<F0(/e/) et F0(/i/)<F0(/ɛ/)). Cette robustesse permet d'affirmer l'existence d'un effet d'inversion du contexte aVa sur la hiérarchie du IF0. Ceci remet donc en cause la conception de IF0 comme étant un phénomène purement segmental.

Les deux principales hypothèses explicatives de IF0, celle du 'tongue-pull' et celle d'une 'planification délibérée', ont toutes les deux été contredites par des faits bien établis. Les résultats de la présente étude ajoutent donc à cette divergence entre faits et théorie. L'hypothèse du 'tongue-pull' ne peut pas expliquer l'effet d'inversion du contexte aVa sur la hiérarchie du IF0, car elle considère IF0 comme étant une fonction de la seule aperture. L'hypothèse d'une 'planification délibérée' est basée sur des prémisses largement indépendantes des phonèmes environnants : la différence acoustique entre voyelles est prétendument renforcée par IF0, mais la direction de ce renforcement dépend uniquement de la voyelle prononcée. Par conséquent, l'inversion ou la neutralisation de la hiérarchie du IF0 dans le contexte aVa va à l'encontre de ce principe.

Ces résultats demandent une explication, qui fait actuellement défaut. Les recherches expérimentales récentes sur le IF0 des voyelles ont montré que, pour subsister, les théories actuelles devaient subir certains raffinements. Or, même l'hypothèse du 'tongue-pull', qui est peut-être la plus compatible avec les données antérieures, est totalement incapable d'expliquer l'inversion de la hiérarchie du IF0 qui est produite par les locuteurs W1 et W4. Elle n'est pas non plus en mesure d'expliquer l'effet plus général d'inversion de la hiérarchie du IF0 généré par le contexte aVa. Sur ce thème de recherche, résolument déroutant, nous sommes invités à poursuivre la quête d'explications alternatives susceptibles de rendre compte des données. Dans la section 1, nous avons introduit l'hypothèse d'une relation entre F0 et l'effort articulatoire, qui pourrait peut-être contribuer à expliquer les résultats de la présente étude. Toutefois, nous devons encore étudier plus en détail l'influence de ce paramètre sur F0, en incluant aux données des mesures de fréquence formantique et de débit, ce qui sera la prochaine étape de cette étude.

5. CONCLUSION

Cette étude remet en question la nature *intrinsèque* du 'F0 intrinsèque' des voyelles à travers l'influence des phonèmes adjacents. Lorsque ces derniers sont la voyelle /a/, une tendance à l'inversion est observée sur la hiérarchie obtenue dans le contexte /dVd/, à tel point qu'elle disparaît. Elle est même inversée pour certains locuteurs, leur F0(/i/) étant significativement plus basse que leurs F0(/ɛ/) et F0(/e/). Ces résultats contribuent, avec d'autres études, à remettre en cause les hypothèses explicatives qui ont été proposées pour expliquer ce phénomène. Cette recherche doit maintenant être approfondie à partir de mesures additionnelles, ainsi que par la réalisation d'autres études expérimentales.

REFERENCES

[1] Whalen, D., H., Levitt, A., G., "The universality of intrinsic F0 of vowels" *Journal of Phonetics*, 23, pp. 349-366, 1995.

[2] Dyhr, N. The activity of the cricothyroid muscle, and the intrinsic fundamental frequency in Danish vowels. Phonetica 47, 141-154 (1990).

[3] Atkinson, J., E. Correlation analysis of the physiological factors controlling fundamental voice frequency. J.A.S.A. 63(1): 211-22 (1978).

[4] Fahey, R., P., Diehl, R., L. The missing fundamental in vowel height perception. Perception and Psychophysics 58 (5): 725-33 (1996)

[5] Whalen, D. H., Andrea G. Levitt, Pai-Ling Hsiao, and Iris Smorodinsky. Intrinsic F0 of vowels in the babbling of 6-, 9- and 12-month-old French- and English-learning infants. Journal of the Acoustical Society of America, 97, 2533-2539 (1995).

[6] C.H. Shadle. Intrinsic fundamental frequency of vowels in sentence context. J. Acoust. Soc. Am. 78, 1562-1567 (1985).

[7] Whalen, D. H., Gick, B., Kumada, M., & Honda, K. Cricothyroid activity in high and low vowels: Exploring the automaticity of intrinsic F0. Journal of Phonetics, 27, 125-142 (1998).

[8] Ohala, J., J., Eukel, B., W. Explaining the intrinsic pitch of vowels. In: R. Channon & L. Shockey (eds.), In honor of Ilse Lehiste. Ilse Lehiste Pühendusteos. Dordrecht: Foris. 207-215 (1987).

[9] Scherer, K., R.: "Vocal affect expression: a review and a model for future research", Psychol. Bulletin 99: 141-165 (1986).

[10] Boersma, P. & Weenink, D. (2005). Praat: doing phonetics by computer (Version 4.3.21) [Computer program]. Retrieved September 1, 2005, from http://www.praat.org/

[11] Piot, O. (2002). Vers une théorie unifiée de la prosodie du français et de l'anglais : des émotions à la phonologie; Ph. D. thesis (Université Paris III, Paris, France). http://www.cavi.univ-paris3.fr/ilpga/ED/student/stop.

Un changement de voix affecte t-il le processus de reconnaissance des mots parlés ?

Sophie Dufour, Noël Nguyen

Laboratoire Parole et Langage, CNRS et Université d'Aix-Marseille, Aix-en-Provence, France
5, Avenue Pasteur, 13604 Aix-en-Provence
sophie.dufour@lpl-aix.fr, noël.nguyen@lpl-aix.fr

ABSTRACT

According to McLennan and Luce [1], variability in talker identity affects spoken word recognition when processing is slow and effortful. In the present study, we tested this hypothesis by manipulating the neighbourhood density of target words in a repetition priming experiment. Both for words with few and many phonological neighbours, the amount of priming for repeated words was not affected by a voice change. Such observation supports the claim that abstract representations exist and underlie spoken word recognition.

Keywords: Variability, abstract representations, episodic models, neighbourhood density.

1. INTRODUCTION

C'est avec rapidité et sans aucune difficulté que nous parvenons à reconnaître les mots et ceci malgré la forte variabilité présente dans le signal de parole. Un mot n'est jamais produit deux fois exactement de la même façon et présente des différences substantielles sur le plan phonologique et/ou phonétique selon le locuteur, le contexte phonologique (phénomène de co-articulation) ou encore le débit de parole. Chaque mot se matérialise ainsi par une infinité de formes sonores différentes que l'auditeur doit ramener à une entité lexicale unique. Un problème majeur auquel est confronté notre système de perception est donc de reconnaître une même production ou un même mot sous différents modes de réalisation.

Selon la théorie « abstractionniste », les mots dans le lexique mental seraient stockés sous la forme de séquences linéaires consistant en des traits [2], des phonèmes [3] ou des syllabes [4]. Le signal de parole serait dans un premier temps converti en une séquence de segments discrets écartant ainsi tous les détails acoustiques fins non pertinents pour l'identification, et serait ensuite projeté sur les représentations symboliques abstraites stockées en mémoire. Au contraire, selon les modèles « épisodiques » [5], les mots seraient stockés sous la forme de traces acoustiques détaillées encodant ainsi des informations fines liées par exemple à la voix du locuteur. Chaque mot serait alors associé à de multiples « tokens » et reconnaître un mot consisterait à trouver l'appariement le plus proche dans une vaste collection d'exemplaires.

Cette recherche fait suite à une étude récente conduite par McLennan et Luce [1] qui ont examiné l'impact de la variabilité acoustique liée à un changement de voix et de débit de parole (lent / rapide) sur le processus de reconnaissance des mots. Pour ce faire, ils ont utilisé le paradigme d'amorçage de répétition et manipulé la difficulté de discrimination entre des mots et des non-mots dans une tâche de décision lexicale ainsi que le format de réponse (immédiat/différé) dans une tâche de répétition de mots. Précisons que le paradigme d'amorçage de répétition consiste à présenter dans un premier temps un bloc de mots aux participants sur lesquels ils doivent réaliser une tâche (e.g. décision lexicale ou répétition). Dans un second temps, un second bloc de mots (bloc cible) leur est présenté, la moitié des mots ayant déjà été rencontré dans le premier bloc, l'autre moitié n'ayant jamais été rencontré. Typiquement, les mots répétés sont reconnus plus rapidement que les mots non répétés. Un tel effet résulterait de l'activation répétée de la même représentation lexicale en mémoire. L'atténuation de cet effet lors de la modification d'une dimension particulière (par exemple, un changement de voix) entre le premier et le second bloc indiquerait en accord avec les modèles « épisodiques » que le même mot prononcé par des voix différentes active différentes représentations lexicales et que des spécificités liées par exemple à la voix du locuteur seraient stockées en mémoire. Au contraire, aucune modulation de l'effet lors d'un changement de voix ou d'un débit de parole indiquerait en accord avec les théories « abstractionnistes » que le même mot prononcé de façon différente active la même représentation lexicale.

En tâche de décision lexicale, McLennan et Luce [1] ont montré une atténuation de l'effet d'amorçage de répétition lors d'un changement de voix ou de débit de parole lorsque la discrimination mots / non-mots était rendue difficile par l'utilisation de non-mots similaires à des mots (ex, bacov issue de bacon). Aucune atténuation dans l'effet d'amorçage de répétition n'a été observée lorsque la discrimination mots / non-mots était rendue facile par l'utilisation de non-mots ayant

peu de ressemblance avec des mots (ex, thushthudge). En tâche de répétition de mots, une atténuation de l'effet d'amorçage de répétition a été obtenue lorsque les participants devaient attendre l'apparition d'un signal pour donner leur réponse, mais pas lorsqu'ils devaient répondre immédiatement après l'apparition du mot. Suite à ces résultats, McLennan et Luce [1] en ont conclu que la variabilité dans le signal de parole liée à un changement de voix ou de débit affecte le processus de reconnaissance des mots parlés uniquement lorsque le traitement est lent et demande un certain effort. Notons que de tels effets sont compatibles avec des modèles dits hybrides [6] selon lesquels à la fois des informations spécifiques et abstraites seraient encodées en mémoire et où l'utilisation de l'une ou de l'autre type d'information dépendrait alors de la lenteur du traitement.

Dans cette étude, nous avons testé plus profondément l'hypothèse selon laquelle des effets liés à l'utilisation d'indices acoustiques émergeraient lorsque le traitement est lent et coûteux. Plutôt que de rendre difficile le traitement par le biais d'une manipulation de l'environnement lié à la tâche, nous avons directement manipulé la difficulté de traitement des mots eux-mêmes et utilisé comme McLennan et Luce [1] un paradigme d'amorçage de répétition dans lequel les mots étaient répétés soit par la même voix, soit par une voix de sexe différent. De façon à favoriser l'exploitation d'indices acoustiques liés à la voix du locuteur, les participants devaient réaliser une tâche de décision lexicale dans laquelle les non-mots ressemblaient fortement à des mots [1]. Comme, il est désormais bien établi que des mots ayant beaucoup de mots qui leurs sont phonologiquement proches (e.g. mots à forte densité de voisinage) sont reconnus plus lentement que des mots n'en ayant peu (e.g. mots à faible densité de voisinage) [7], la difficulté de traitement a été manipulée par le biais de la densité de voisinage phonologique. Notre hypothèse était que si les effets liés à l'utilisation d'indices acoustiques émergent lorsque le traitement est lent et coûteux une atténuation de l'effet d'amorçage de répétition devrait être observée au moins pour les mots difficiles à traiter et donc pour les mots ayant une forte densité de voisinage.

2. EXPÉRIENCE

2.1. Méthode

2.1.1. Participants

40 volontaires de l'Université de Provence ont participé à l'expérience. Tous étaient de langue maternelle française et n'ont rapporté aucun trouble de l'audition ou de la parole.

2.1.2. Matériel

Quarante mots cibles de structure syllabique CVC ont été sélectionnés à partir de VOCOLEX [8]. La moitié d'entre eux résidaient dans une forte densité de voisinage et l'autre moitié dans une faible densité de voisinage. Le voisinage phonologique a été calculé en comptabilisant pour chaque mot cible le nombre de mots qui peuvent être générés par addition, délétion ou substitution d'un phonème quelle que soit sa position [7]. Les caractéristiques des mots cibles sont fournies dans le Tableau 1. 20 mots additionnels appariés aux mots cibles en fréquence, en nombre de phonèmes et en nombre de voisins phonologiques ont été également sélectionnés.

Afin que chaque mot cible soit vu dans la condition répétée et dans la condition non répétée, et qu'un même participant ne voit pas plusieurs fois le même mot cible, deux listes expérimentales ont été créées. Chaque liste était constituée de deux blocs de stimuli. Le premier (bloc 1) était constitué de la moitié des 40 mots cibles et des 20 mots additionnels. Parmi les 20 mots cibles, 10 étaient de forte densité et les 10 autres de faible densité de voisinage. Le second (bloc2) était constitué des 40 mots cibles, la moitié étant les mots présents dans le premier bloc et l'autre moitié étant alors des mots cibles contrôles non répétés. Les listes ont été contrebalancées de sorte à ce qu'un même mot cible serve à la fois en contrôle et en répétition. Pour les besoins de la tâche, 60 non-mots monosyllabiques de structure CVC ont été ajoutés dans chacune des listes et crées en changeant seulement le premier ou le dernier phonème de mots existants. 40 étaient présentés dans le bloc 1 et les 20 autres dans le bloc 2. Le bloc 2 comprenait également 20 non-mots du bloc 1.

De façon à manipuler le changement de voix, les 2 listes expérimentales ont été par la suite divisées en 4 sous listes chacune et se constituaient de la façon suivante : a) bloc 1 voix masculine, bloc 2 voix masculine, b) bloc 1 voix féminine, bloc 2 voix féminine, c) bloc 1 voix masculine, bloc 2 voix féminine, d) bloc 1 voix féminine, bloc 2 voix masculine.

Table 1 : Caractéristiques des mots cibles.

	Mots à faible densité	Mots à forte densité
Nombre de voisins phonologiques	15	31
Fréquence[1]	58	55
Nombre de Phonèmes	3	3
Durée[2] voix masculine	565	565
Durée[2] voix féminine	565	565

Notes: [1] en nombre d'occurrences par million ; [2] en

millisecondes.

2.1.3. Procédure

Les stimuli ont été enregistrés par une locutrice et par un locuteur de langue maternelle française et ont été digitalisés à un taux d'échantillonnage de 22 kHz avec une résolution de 16 bits. Les participants munis d'un casque audio ont été testés individuellement dans une chambre insonorisée et les stimuli leur étaient présentés à un niveau sonore confortable. La présentation des stimuli était contrôlée par un ordinateur et les temps de réponse (TRs) étaient enregistrés à partir du début des stimuli. Pour chaque stimulus, les participants devaient indiquer le plus rapidement et le plus précisément possible si il constituait un mot ou non de la langue française, et devait fournir la réponse mot avec leur main dominante. La réponse du participant et le début de présentation du stimulus suivant étaient séparés par un délai de deux secondes. Les participants ont été testés sur une seule des sous listes expérimentales et ont commencé l'expérience avec 16 essais d'entraînement.

2.2. Résultats et Discussion

Les temps de réaction obtenus dans le bloc 2 ont été analysés. Pour chaque participant, les temps de réaction supérieurs à 2,5 écart-types au-dessus et en-dessous de la moyenne des temps de réaction dans chaque condition ont été exclus des analyses. Adoptant ce critère seulement 1.06% des données ont été rejetées. Les réponses incorrectes ont été également supprimées des analyses. Les temps de réaction moyens obtenus en fonction du type de cible et du changement de voix sont représentés dans la Figure 1 pour les mots à forte densité et dans la Figure 2 pour les mots à faible densité. Les erreurs ayant été peu nombreuses (moins de 5%), les analyses ont été effectuées seulement sur les temps de réaction. Des analyses de variance (ANOVAs) par sujets (F_1) et par items (F_2) ont été conduites avec le type de cible (répétée, contrôle), la densité de voisnage (faible, forte) et le changement de voix (sans, avec) comme variables.

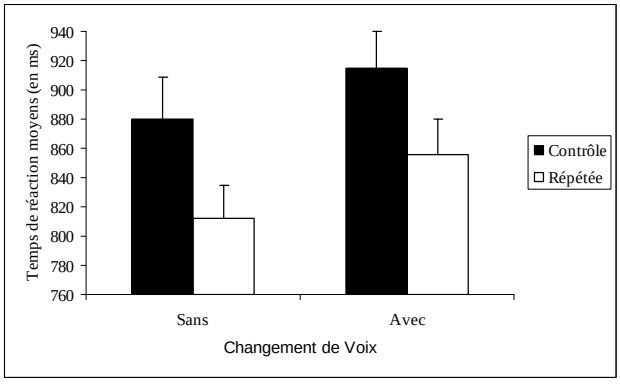


Figure 1 : Temps de réaction moyens (en ms) en fonction du type de cible et du changement de voix pour les mots à forte densité (les barres représentent les erreurs standards).

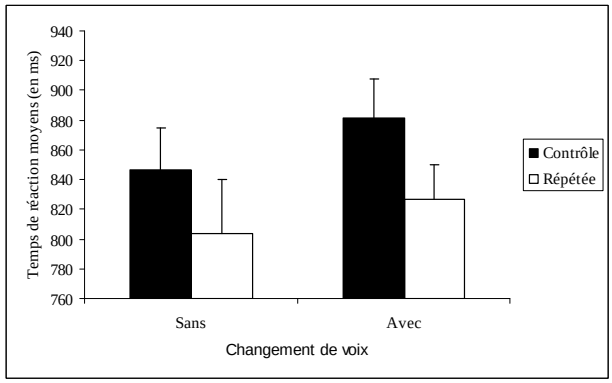


Figure 2: Temps de réaction moyens (en ms) en fonction du type de cible et du changement de voix pour les mots à faible densité (les barres représentent les erreurs standards).

Les temps de réponse étaient en moyenne plus rapides pour les mots cibles répétés (825 ms) que pour les mots cibles contrôles (881 ms). Cet effet était significatif à la fois par participants [$F_1(1, 38) = 66.30$, $p < .0001$] et par items [$F_2(1, 38) = 45.97$, $p < .0001$]. Les temps de réponse étaient en moyenne plus rapides pour les mots cibles résidant dans une faible densité de voisinage (840 ms) que pour ceux résidant dans une forte densité de voisinage (866 ms). Cet effet était significatif par participants [$F_1(1, 38) = 13.00$, $p < .001$] mais échouait à atteindre la significativité par items [$F_2(1, 38) = 1.82$, $p = .19$]. Les temps de réponse étaient en moyenne plus lents dans le cas d'un changement de voix (870 ms) que lorsque la voix restait identique (836 ms). Cet effet était signifiactif par items [$F_2(1, 38) = 18.28$, $p < .001$] mais pas par participants [$F_1(1, 38) = 0.87$, $p > .20$]. Que se soit pour les mots à forte ou à faible densité de voisinage, l'intéraction entre le type de cible et le changement de voix n'était pas significatif montrant ainsi aucune diminution de l'effet d'amorçage de répétition dans le cas d'un changement de voix [$F_1(1, 38) = 0.24$, $p > .20$; $F_2(1, 38) = 0.04$, $p > .20$ pour les mots à forte densité de voisinage; $F_1(1, 38) = 0.21$, $p > .20$; $F_2(1, 38) = 0.29$, $p > .20$ pour les mots à faible densité de voisinage.]

3. DISCUSSION GÉNÉRALE

L'hypothèse sous jacente à notre recherche était que si les effets liés à l'utilisation d'indices acoustiques émergent lorsque le traitement est lent et coûteux, une atténuation de l'effet d'amorçage de répétition devrait être observée pour des mots difficiles à traiter. De façon à manipuler la difficulté des mots, des mots à faible et à forte densité de voisinage ont été utilisés, les mots ayant beaucoup de voisins phonologiques étant généralement reconnus plus lentement que les mots ayant peu de voisins phonologiques [7]. Que se soit pour les mots à forte ou à faible densité de voisinage, aucune diminution dans la taille de l'effet d'amorçage de répétition n'a été observée lors d'un changement de

voix. Une telle observation argumente en faveur de l'existence de représentations abstraites et indiquerait qu'un même mot prononcé par différents locuteurs est susceptible d'activer la même représentation lexicale de base.

Comme nous l'avons vu précédemment, des modèles dit hybrides postulant la co-existence de représentations abtraites et détaillés ont été proposés [6]. En accord avec ce type de modèle, nous disposons dans la littérature de preuves expérimentales en faveur de l'un ou de l'autre type de représentations [9, 10]. L'existence conjointe de représentations abstraites et détaillées sous tendant la reconnaissance des mots parlés a clairement été mise en évidence dans l'étude de McClennan et Luce [1]. Comme ces auteurs nous avons favorisé l'exploitation d'indices liés à la voix des locuteurs en rendant difficile la discrimination entre les mots et les non-mots. Néanmoins dans notre étude aucun impact lié à un changement de voix sur le processus de reconnaissance des mots n'a été observé. Une différence entre notre étude et celle de McLennan et Luce [1] est relative à la durée de nos mots. En effet, ils ont été contrôlés de sorte à ce qu'il n y ait aucune différence de durée entre les mots prononcés par la voix masculine et ceux prononcés par la voix féminine. Cependant, bien que les auteurs précisent que leurs mots ont été jugés comme ayant été prononcés à un débit de parole normal, la durée moyenne des mots prononcés par la voix masuline et celle de ceux prononcés par la voix fémine différaient dans l'étude de McLennan et Luce [1]. Il se peut alors que certaines caractéristiques comme le débit de parole soient plus prépondérantes et aient plus d'impact sur le processus de reconnaissance des mots parlés. Notons en accord avec cette idée que Kittredge, Davis et Blumstein [11] ont récemment échoué à mettre en évidence une atténuation de l'effet d'amorçage sémantique lors d'un changement de voix avec des mots contrôlés en durée entre la voix masculine et féminine. D'avantage d'études sont donc nécessaires de façon à tester cette possibilité.

BIBLIOGRAPHIE

[1] C.T. McLennan and P.A. Luce. Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31: 306-321, 2005.

[2] W. D. Marslen-Wilson and P. Warren. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, 101: 653-675, 1994.

[3] J. L. McClelland and J. L. Elman. The TRACE model of speech perception. *Cognitive Psychology*, 18: 1 - 86, 1986.

[4] J. Mehler, J.Y. Dommergues, U. Frauenfelder, and J. Segui. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20: 298–305, 1981.

[5] S.D. Goldinger. Echoes of echoes? An episodic theory of lexical access. *Psychological Review,* 105: 251-279, 1998.

[6] J. Pierrehumbert. Exemplar dynamics: Word frequency, lenition, and contrast. *In J. Bybee & P. Hopper (Eds.), Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins, pages 137-157, 2001.

[7] P.A. Luce and D.B. Pisoni. Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing*, 19: 1–36, 1998.

[8] S. Dufour, R. Peereman, C. Pallier and M. Radeau. VoCoLex : une base de données lexicales sur les similarités phonologiques entre les mots français. *L'Année Psychologique*, 102: 725-746, 2002.

[9] S.D. Goldinger. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22: 1166-1183, 1996.

[10] J.M. McQueen, A. Cutler and D. Phonological abstraction in the mental lexicon. *Cognitive Science*, 30: 1113-1126, 2006.

[11] A. Kittredge, L. Davis and S.E. Blumstein. Effects of Nonlinguistic Auditory Variations on Lexical Processing in Broca, *Brain and Language*, 97: 25-40, 2006.

Temps de réaction et identification perceptuelle des langues étude préparatoire en vue de l'optimisation d'un protocole expérimental en IRMf

Melissa BARKAT-DEFRADAS, Jorge GUTIERREZ-CELAYA, Samia BELAID

[1] Laboratoire Praxiling UMR 5237 CNRS & Université de Montpellier
melissa.barkat@univ-montp3.fr ; jorge.gutierrez@univ-montp3.fr ; samia.belaid@etu.univ-montp3.fr

ABSTRACT

The aim of this study is to investigate to which extent Arabic listeners can assess some reliable perceptual characteristics from samples of speech of different durations so as to identify the Arabic dialect in which they are produced. Our goal is to determine the minimum duration threshold that is required to achieve correctly a perceptual identification task. Results – analyzed in terms of reaction time (RT), correct identification rate and error analysis – are discussed in the field of experimental research and will be used to draw up the experimental design of an ongoing fMRI study dealing with the perceptual treatment of linguistic distance applied to the Arabic dialectal continuum.

Keywords: Perceptual identification, Arabic dialects, Reaction time, Methodology, Experimental design, fMRI.

1. INTRODUCTION

La capacité de l'être humain à identifier et à discriminer perceptuellement des langues variées a donné lieu à de nombreux travaux (pour un état des lieux, voir [1]). Des expériences conduites en parole naturelle se sont ainsi intéressées à mesurer les performances d'identification de différentes langues indépendamment des conditions d'écoute et/ou des particularités linguistiques des langues proposées. Lorch et Maera [2] ont par exemple testé la capacité d'auditeurs britanniques à identifier 6 langues inconnues après 20 secondes d'écoute. Muthusamy et *al.* [3] se sont intéressés à la capacité de sujets monolingues anglophones à identifier 10 langues à partir de stimuli de durées variables (1-6 secondes). Le score moyen d'identification correcte varie en fonction de la durée des stimuli et selon que l'on inclut ou non les performances relatives à l'identification de la langue maternelle (LM). Cette étude démontre également l'effet de l'apprentissage sur l'amélioration des performances des sujets. Les résultats observés dans les deux études précédemment citées portent sur l'identification de langues plus ou moins éloignées. C'est pourquoi, au delà des scores d'identification correcte, l'analyse des erreurs s'avère particulièrement informative pour notre propos. En effet, il est intéressant de noter que les échecs à la tâche ont toujours concerné les langues les plus proches linguistiquement (i.e. mandarin/vietnamien, langues toutes deux tonales et/ou coréen/japonais, langues attestant une large part de vocabulaire commun). Suite à ces travaux, nous avons conçu une série d'expériences dont l'objectif était de tester l'effet des variables 'proximité' *vs*. 'distance' linguistique sur les performances d'identification. Le continuum dialectal arabe permet de tester de façon rigoureuse ce paramètre puisqu'il existe, sur ce domaine, une corrélation forte entre proximité linguistique et proximité géographique.

Depuis une dizaine d'années, l'imagerie par résonance magnétique fonctionnelle (IRM*f*), technique d'imagerie médicale permettant de visualiser précisément et en temps réel l'activité du cerveau, est devenue une technique couramment utilisée pour l'étude du traitement de la parole et du langage [4]. Cette technique – qui repose sur l'observation des modifications de concentration d'oxygène dans le sang – présente de nombreux avantages par rapport à d'autres techniques plus anciennes. En effet, outre sa bonne résolution spatiale et temporelle, cette technique s'avère totalement inoffensive et les résultats qu'elle permet d'observer ont souvent une portée considérable. Toutefois, l'IRM comporte certains inconvénients liés notamment au dispositif expérimental peu banal qu'elle requiert et aux contraintes qu'elle fait peser sur le sujet qui – bien que volontaire – se trouve le temps de l'expérimentation en bien mauvaise posture du fait du confinement dans le tunnel étroit de l'imageur et de la nécessité absolue d'éviter tout mouvement de la tête pouvant conduire à des activations parasites. Aussi, le choix du nombre de séquences d'enregistrements et du nombre d'essais constitue un facteur expérimental non négligeable. En effet, on présuppose qu'au delà de quelques minutes la fatigue et les mouvements du sujet risquent fort de détériorer la qualité des données. Toutefois, la lecture attentive des articles scientifiques relatifs à l'étude du langage par imagerie médicale révèle que cet aspect n'est pas toujours clairement présenté dans la partie méthodologique. C'est pourquoi une réflexion portant sur le temps d'expérimentation nous a semblé intéressante. Une expérience similaire conduite dans le scanner par Zhao et al. [5] a révélé que moins de 2,5 secondes sont suffisantes pour discriminer des langues connues par les sujets. L'expérience que nous présentons ici a pour objectif

d'étudier le temps nécessaire à l'identification perceptuelle de différents parlers arabes par des locuteurs naïfs. Les résultats observés nous permettront de concevoir au mieux le design expérimental d'une expérience en IRMf actuellement en cours de réalisation[1]. Nous avons ainsi conçu un paradigme expérimental dont la tâche était d'identifier *le plus rapidement possible* une série de langues/dialectes[2] plus ou moins proches du parler maternel.

2. TEMPS DE REACTION ET IDENTIFICATION DIALECTALE

L'expérience réalisée ici a pour objectif de déterminer le temps minimal nécessaire pour réaliser, avec le moins d'erreurs possibles une tâche d'identification dialectale. Les résultats observés devant permettre *in fine* l'optimisation de notre design expérimental.

2.1. Traitement perceptuel de la variabilité linguistique

Faisant suite à la caractérisation des parlers arabes aux niveaux segmental (Barkat-Defradas, 2000 [6]), supra-segmental (Hamdi & *al.*, 2005 [7]) et à leur identification perceptuelle à partir de parole naturelle (Barkat-Defradas 2001 [8]) ou synthétisée (Barkat et *al*., 1999 [9]), la présente étude a pour objectif de déterminer la durée minimale nécessaire à l'identification des parler arabes en fonction de la distance linguistique existant entre le dialecte maternel et les autres parlers[3]. Pour répondre à cette question, nous entendons concevoir un paradigme expérimental où des sujets arabophones natifs – stimulés auditivement à l'intérieur du scanner – auront à identifier la nature dialectale de différents échantillons de parole. En regard des résultats obtenus en perception, 3 conditions sont envisagées : (i) dialecte maternel, (ii) dialecte proche du dialecte maternel et (iii) dialecte éloigné du dialecte maternel. Pour chaque condition, 5 locuteurs différents seront enregistrés, chaque locuteur produira 6 échantillons de parole différents dans son dialecte maternel. Ceci conduit à un total de 90 échantillons de parole à traiter. On comprend dès lors que la détermination du temps minimal nécessaire pour réaliser la tâche d'identification est un facteur non négligeable lors de la conception du paradigme expérimental, notamment pour ce qui concerne la réduction du temps d'expérimentation. En effet, selon que les échantillons de parole présentent une durée moyenne de 1 à 6 secondes, la durée totale d'une session passe de 2,25 à 13,5 minutes en prenant en compte les intervalles inter stimulations (ISI) de durée aléatoirement variable (2-4 secondes) lesquels, couplés à une présentation aléatoire des stimuli, permettent d'améliorer l'amplitude de la réponse cérébrale du sujet en réduisant l'effet d'habituation à la tâche (Table 1).

Table 1 : Evolution du temps d'expérimentation en fonction de la durée des stimuli soumis à identification.

Durée du stimulus (sec)	Nombre d'essais	Durée des stimuli (sec)	Durée ISI (sec)	Durée totale d'une session (min)
1	90	90	45	2,25
2	90	180	90	4,5
3	90	270	135	6,75
4	90	360	180	9
5	90	450	225	11,25
6	90	540	270	13,5

2.2. Matériel, sujets, méthode

Dans le cadre de cette étude préparatoire, nous avons retenu les 3 mêmes conditions linguistiques que celles envisagées pour le protocole IRMf, soit : (i) dialecte maternel des sujets testés (i.e. arabe marocain), (ii) dialecte proche du dialecte maternel (i.e. arabe tunisien) et (iii) dialecte éloigné du dialecte maternel (i.e. arabe syrien). Afin d'éviter que les sujets testés ne s'appuient sur les caractéristiques vocales des sujets enregistrés pour constituer le matériel de stimulation auditive, 3 locuteurs par dialectes ont été enregistrés (i.e. 9 voix différentes) alors qu'ils se livraient à une tâche de narration dans leur dialecte maternel respectif. Chaque enregistrement présentait une durée totale de 10 minutes environ. Les enregistrements numériques (44KHz, 8bits, stéréo) ont été réalisés en chambre anéchoïque à l'aide d'un micro externe directionnel

[1] Recherche réalisée en collaboration avec le laboratoire Parole et Langage (UMR 6057 CNRS & Université de Provence) et le centre IRMf de l'Hôpital de la Timone de Marseille, avec le soutien du CNRS.

[2] En linguistique, il est difficile de distinguer clairement 'langues' et 'dialectes'. On dit généralement que lorsqu'il y a *intercompréhension* entre deux parlers on a affaire à deux '*dialectes*' de la même langue. Au contraire si cette intercompréhension n'existe pas, il s'agit alors de deux '*langues*' différentes. Cette définition n'est pas pleinement satisfaisante, en particulier lorsque l'on rencontre des situations de '*chaînes dialectales'*, où il peut y avoir intercompréhension entre une zone linguistique A et une autre zone B adjacente, de même entre une zone B et une zone C sans pour autant avoir une intercompréhension entre les locuteurs de la zone A et de la zone C, trop distinctes l'une de l'autre tant du point de vue géographique que linguistique.

[3] Le domaine arabophone s'avère particulièrement adapté à l'investigation de cette question. En effet, la plupart des travaux sur le traitement des langues chez le sujet bilingue comparent l'activité cérébrale générée par des langues parfois très proches (par exemple, le catalan et l'espagnol), parfois beaucoup plus éloignées (comme par exemple, le chinois et l'anglais). Il nous semble cependant très important de mieux contrôler ce paramètre pour examiner s'il existe une relation entre la proximité linguistique des langues traitées par le cerveau d'un individu et l'activation des zones cérébrales concernées par le traitement associé à L1 ou L2.

(microphone dynamique professionnel HQ) placé à une distance d'environ 15 cm de la bouche du locuteur. De ces enregistrements, nous avons extrait des échantillons de parole correspondant à des énoncés complets et dont la durée variait de 1 à 4 secondes (3 échantillons / durée). Aucun contrôle quant au contenu lexical des énoncés n'a été observé. Le matériel de stimulation est ainsi constitué de 36 échantillons différents. Nous avons utilisé ces échantillons pour concevoir une expérience perceptuelle d'identification dialectale avec mesure du temps de réaction (TR). L'interface expérimentale a été réalisée sous E-Prime. Les 36 échantillons ont été présentés par trois fois au cours de 3 sessions distinctes (i.e. 108 stimuli). Afin de familiariser les sujets au matériel et à la tâche, deux sessions d'entraînement, conduites à partir de voix (3) et d'échantillons sonores différents (9), ont préalablement été proposées aux sujets. Treize sujets volontaires arabophones natifs, locuteurs d'arabe marocain ont participé à l'étude. La consigne visuellement fournie aux sujets via l'écran d'accueil de l'ordinateur utilisé pour la passation de l'expérience était la suivante : *« vous allez entendre différents extraits de parole en arabe dialectal. Ces extraits sont produits soit en arabe marocain, soit en arabe tunisien, soit en arabe syrien. Vous devez devinez le plus rapidement possible de quel dialecte il s'agit. Si, selon vous, il s'agit d'arabe marocain, taper sur la touche étiquetée 'M', s'il s'agit d'arabe tunisien, taper sur la touche étiquetée 'T', s'il s'agit d'arabe syrien, taper sur la touche étiquetée 'S' ».* Trois des quatre touches de flèches directionnelles du clavier ont été garnies d'une étiquette autocollante portant les lettres précédemment citées. En référence à la géographie du monde arabe, nous avons attribué à la touche ← la valeur 'M' (le Maroc étant situé à l'Ouest du domaine), à la touche → la valeur 'S' (la Syrie étant située à l'Est du domaine) et enfin, à la flèche ↑ la valeur 'T' (la Tunisie étant située entre le Maroc et la Syrie). Cette motivation symbolique a été clairement explicitée aux participants. L'écoute des différents stimuli – présentés en ordre aléatoire pour chaque sujet et à l'intérieur de chaque session – a été faite via un casque Sennheiser HD580 à un niveau d'écoute confortable (60dB).

2.3. Résultats

Les résultats montrent d'une part que les scores d'identification sont élevés pour l'ensemble des parlers étudiés avec, pour l'arabe marocain (correspondant à la langue maternelle des sujets) 93,16% d'identification correcte, pour l'arabe tunisien, (correspondant au parler proche du dialecte maternel) 85,68% d'identification correcte et, pour l'arabe syrien (dialecte le plus éloigné – et donc linguistiquement le plus différent – du dialecte maternel) 88,03% d'identification correcte. D'autre part, l'analyse des erreurs en regard de la notion de distance *vs.* proximité linguistique montre que l'identification du dialecte le plus éloigné du dialecte maternel est une tâche moins complexe à réaliser que l'identification du parler le plus proche de la langue maternelle. En effet, les confusions (voire l'absence de réponse) concernent le plus souvent l'identification du tunisien (Table 2).

La difficulté à discriminer le dialecte maternel du parler qui lui est le plus proche transparait également sur le temps de traitement. Le temps de réaction moyen est significativement plus long pour l'identification du tunisien que pour l'identification du syrien ($p<.05$) (Figure 1).

Table 2. Matrice de confusion (en nombre d'essais / 1404 essais) et en % d'identification.

Stimuli	MA	TU	SY	Pas de réponse
MA	436 (93,1%)	21 (4,4%)	10 (2,1%)	1 (0,21%)
TU	34 (7,2%)	401 (85,6%)	30 (6,4%)	3 (0,64%)
SY	7 (1,4%)	47 (10%)	412 (88%)	2 (0,42%)

Les résultats montrent également un effet de la session sur les performances des sujets : les scores augmentent entre chaque session, révélant ainsi que l'apprentissage se poursuit au cours du temps ($p<.01$) (Figure 2).

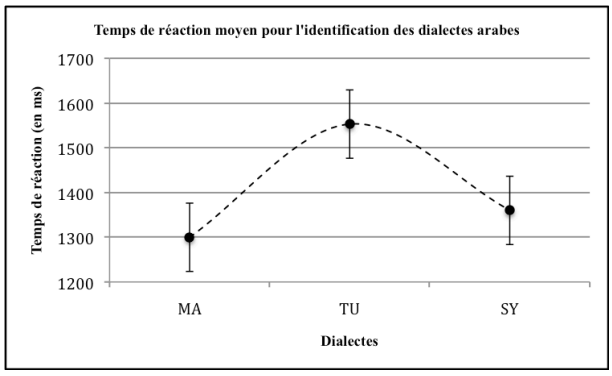


Figure 1. Temps de réaction moyen (en ms) pour l'identification en fonction du dialecte.

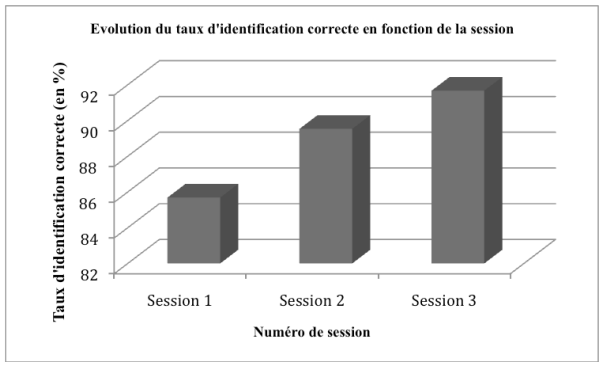


Figure 2. Evolution du taux d'identification (en %) en fonction de la session.

Il apparaît par ailleurs que les scores d'identification ne sont pas corrélés à la durée des stimuli (coeff. 0.05). Ce résultat est principalement lié au fait que nous avons explicitement demandé aux sujets de réaliser la tâche d'identification *le plus rapidement possible* (*cf.* consigne, section 2.2).

Enfin, afin de répondre à l'objectif méthodologique de cette étude, nous avons mis en regard les scores d'identification correcte et la durée moyenne normalisée du temps de réaction (Figure 3). Ces derniers résultats suggèrent l'existence d'un seuil de stimulation minimal à partir duquel il est possible d'observer des résultats d'identification probants : ≈ 1,75 seconde. Ce résultat est comparable à celui obtenu dans le cadre d'une étude antérieure sur la discrimination du dialecte américain [10].

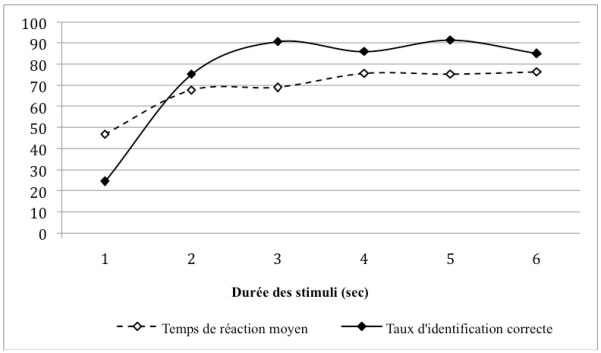


Figure 3. Temps de réaction moyen et taux d'identification correcte en fonction de la durée des stimuli.

3. CONCLUSIONS

Outre des scores d'identification significativement plus élevés pour l'identification de la langue maternelle, nos résultats ont également montré que les sujets obtiennent des scores supérieurs pour l'identification du parler le plus éloignés de leur dialecte d'origine. La notion de distance et/ou de proximité linguistique peut donc être considérée comme un critère d'étude pertinent dans la mesure où elle influe fortement sur les performances des sujets. La consigne proposée aux sujets ayant participé à cette étude, stipulant explicitement de répondre « *le plus rapidement possible* » peut expliquer que les sujets ont probablement réalisé la tâche sur la base des indices phonético-phonologiques seuls sans 'attendre' l'éventuelle occurrence d'un indicateur géolinguistique d'ordre lexical.

Les résultats de cette étude préparatoire nous ont conduits à envisager différents types d'amélioration pour notre protocole en IRMf. Concrètement, nous envisageons d'une part d'intégrer en amont une tâche comportementale d'identification perceptuelle. D'autre part, nous ferons le choix d'intégrer un parler proche du dialecte maternel qui soit suffisamment différent du dialecte d'origine pour ne pas porter à confusion, ce qui pourrait conduire à des chevauchements d'activations entre langue maternel et dialecte proche. Enfin, la détermination d'un point d'équilibre correspondant au temps minimal de stimulation nécessaire à la bonne réalisation de la tâche nous permet d'envisager l'augmentation du nombre de séquences et d'essais sans pour autant augmenter inutilement le temps d'expérimentation en IRMf.

4. REFERENCES

[1] M., Barkat-Defradas, I., Vasilescu, I. F., Pellegrino, F., Stratégies perceptuelles et identification automatique des langues: application au continuum dialectal arabe, *Revue Parole*, pp. 1-44, 2003.

[2] M., Lorch & P., Maera, "How People Listen to Languages They Don't Know », *Language Sciences*. Volume I, Number 4, pp. 343-353, 1989.

[3] Y., Muthusamy, N., Jain, R., Cole, "Perceptual benchmarks for Automatic Language Identification", in *Proceedings ICASSP*, 1994.

[4] A., Marchal, Ch., Cavé, *L'imagerie médicale pour l'étude de la parole*, Hermès Lavoisier, 2009.

[5] J., Zhao, H., Shu, L., Zhang, X., Wang, Q., Gong et P., Li, Cortical competition during language discrimination, *NeuroImage*, vol. 43, pp. 624-633, 2008.

[6] M., Barkat, Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes, Editions du Septentrion, Lille, 2000.

[7] R., Hamdi, M., Barkat-Defradas, E., Ferragne, F., Pellegrino, (2004), "Speech Timing and Rhythmic structure in Arabic dialects", in proceeding of 8th ICSLP, pp. 331-334, 2004.

[8] M., Barkat, Vers l'identification automatique des parlers arabes, *Revue Internationale 'Langues et Linguistique*', vol. 7, pp. 47-75, 2001.

[9] M., Barkat, J.J., Ohala, F., Pellegrino, "Prosody as a Distinctive Feature for the Recognition of Arabic Dialects." in *Proceedings of Interspeech*, Budapest, 1999.

[10] C., G., Clopper, A., R., Bradlow, Free classification of American English dialects by native and non-native listeners, *Journal of Phonetics*, vol. 37, pp. 436-451, 2009.

Le babillage et le développement des compétences temporo-articulatoires

Mélanie Canault[1], Pascal Perrier[2], Rudolph Sock[3] & Rafael Laboissière[4]

[1]Institut des Sciences et Techniques de la Réadaptation – Université Claude Bernard, Lyon 1– E.A. 4129 Santé, Individu et Société.
melanie.canault@recherche.univ-lyon1.fr
[2]GIPSA-lab, CNRS UMR 5216, Grenoble INP, Grenoble
Pascal.Perrier@gipsa-lab.grenoble-inp.fr
[3]Institut de Phonétique de Strasbourg – Université de Strasbourg – EA 1339, Linguistique, Langues et Parole, Composante Parole et Cognition
sock@unistra.fr
[4]« Espace et Action » U864 INSERM – Université Claude Bernard, Lyon 1, Bron
rafael.laboissiere@inserm.fr

ABSTRACT

Articulatory control emergence is a long process involving various parameters. At the early stage of babbling, strong temporal constraints are imposed on the young speaker's production system by the biological rhythm of the jaw. However, the baby gradually reorganizes the timing of his mandibular displacements. The observation of syllable duration in 11 subjects between 8 and 12 months, would suggest that, at 10 months, the baby starts decreasing the duration of his syllables in order to introduce a rhythm specific for speech.

Keywords: babbling, temporal development, jaw oscillation

1. INTRODUCTION

Le babillage, que l'on situe généralement entre l'âge de 7 et 12 mois, se voit marqué par l'apparition des premières formes syllabiques. Selon MacNeilage [22], ces dernières résulteraient de la superposition d'une oscillation mandibulaire rythmique au processus de vocalisation. La mandibule serait alors l'articulateur dominant [26] de cette période et elle imposerait de fortes contraintes temporelles sur le système de production. Un certain nombre de changements sont communément décrits à ce stade. Le plus connu d'entre eux reste le passage du babillage redupliqué, au babillage varié (10 mois). Le premier implique la répétition d'une même syllabe, et le second une variation des composantes syllabiques d'un cycle oscillatoire à un autre [22]. Ainsi, bien que l'on ne puisse pas affirmer que le contrôle articulatoire de la parole soit acquis, ce passage traduit l'émergence de compétences articulatoires nouvelles. Nous tenterons dans cet article de vérifier si le timing des productions est touché par cette évolution en nous attachant à l'observation de la durée syllabique.

2. L'ORGANISATION TEMPORELLE PRECOCE

Comme les poumons, la mandibule n'a pas l'activité de langage pour fonction première. Son système d'activation de base, prioritairement dédié à la nutrition, serait alors réaménagé pour cette activité.

2.1. Un rythme biologique

La contiguïté des activités motrices d'ingestion et de parole apparaît dans la similarité des patrons de mouvement impliqués, c'est-à-dire l'abaissement et l'élévation de la mandibule. Mais la parenté de l'activité mandibulaire de la mastication/déglutition et de la parole ne s'arrête pas là ; elle existe aussi au niveau cérébral ([25], [1], [10]). En effet, selon ces travaux, les deux fonctions activent des régions communes du cortex pré-moteur à savoir les aires de Brodmann 44 (recouvrant une partie de l'aire de Broca chez l'homme) et 6. Ces deux régions corticales interviendraient dans le contrôle moteur du cycle mandibulaire de l'hominidé et dans celui du cycle d'ingestion des mammifères.

La mandibule étant *a priori* le seul articulateur activement impliqué dans les productions du babillage ([14], [15]), il y a fort à parier que les premières syllabes soient générées sur son rythme biologique. Or, il faut rappeler que la nutrition et la parole sont deux activités fondées sur des schémas temporels distincts, puisque, chez l'adulte, les activités de nutrition seraient générées sur un rythme environ deux fois plus lent que celui de l'activité langagière. En effet, le rythme de mastication correspondrait à une fréquence variant de 1,5Hz [16] à 3Hz [24] et celui de la parole à des fréquences de 5Hz [16] à 6Hz ([28], [21]). Chez l'enfant, en revanche, le rythme de succion est plus lent. L'organisation temporelle de l'activité de succion présenterait deux modes, celui de la succion non nutritive, qui renverrait plutôt à une activité réflexe, et celui de la succion nutritive, caractéristique de la prise de nourriture ([29], [11], [17], [20], [13]). Le rythme oscillatoire de la succion non nutritive avoisinerait 2Hz et celui de la succion nutritive 1Hz ([6], [3], [12]).

2.2. L'accès à un rythme spécifique

Si l'on accepte, d'une part, que les premières syllabes s'organisent d'abord sur l'oscillation biologique de la

mandibule, et d'autre part, que la parole mature présente une fréquence oscillatoire différente de l'activité d'ingestion, alors le bébé doit se libérer du rythme naturel de la mandibule pour reconstruire un rythme spécifique à la fonction de parole.
Compte-tenu des valeurs présentées plus haut pour la fréquence oscillatoire de la mandibule impliquée dans les cycles d'ingestion (1–2Hz), et compte tenu du fait que l'activité langagière serait deux fois plus rapide que celle de la nutrition, le bébé devrait, avec l'âge, diminuer la durée de ses syllabes, c'est-à-dire mettre en place une fréquence oscillatoire de la mandibule plus rapide. Ainsi, l'oscillation mandibulaire entrant dans le comportement langagier s'engagerait sur un rythme qui se situe entre 2 et 4Hz (soit une valeur moyenne d'environ 3 Hz). C'est effectivement ce que montrent des auteurs [2], [18], [9], [8]) qui ont observé que les productions sonores de l'enfant semblaient se faire à un rythme de 2,5–3Hz. Notons qu'à ce rythme les oscillations mandibulaires de l'activité de parole restent environ deux fois plus lentes chez l'enfant que chez l'adulte ([19], [27]). Ainsi, pour accéder aux schémas temporels de la parole mature, le bébé devra encore augmenter le rythme de ses oscillations mandibulaires.

3. UNE ETUDE TRANSVERSALE

On peut alors se demander comment le mode fréquentiel de l'oscillation mandibulaire va se spécifier pour la fonction de parole. Va-t-il suivre nos prédictions en passant de 1–2 à 3–4 Hz ? Si oui, va-t-il suivre une évolution constante? Dans des travaux antérieurs ([4], [5]) sur l'émergence du contrôle segmental, nous avions pris comme paramètre d'observation celui de la variation temporelle, et avions montré qu'aux environs de 10 mois, le cycle mandibulaire, au préalable stable, affichait une forte variabilité pour ensuite retrouver une phase de stabilisation, accompagnée d'une diminution de sa durée, jusqu'à 13 mois. Notre objectif est de tester l'existence d'une période critique à la transition du babillage redupliqué et du babillage varié, à travers l'évolution temporelle de la syllabe.

3.1. Protocole et mesures

Une étude transversale des productions spontanées de jeunes locuteurs âgés de 8 à 12 mois, soit de 231 jours à 359 jours, nous a permis de tester l'évolution de la spécialisation temporelle de la mandibule, au stade du babillage. Les enregistrements acoustiques ont été réalisés chez 27 sujets, mais tous n'ont pas été inclus dans l'analyse. Seuls ceux ayant produit un nombre minimal de 10 syllabes CV ont été retenus (*cf.* Tabl. 1), c'est-à-dire un total de 11 sujets.
La syllabe correspondant, d'un point de vue articulatoire, à un cycle de fermeture et d'ouverture de la mandibule [22], les signaux recueillis ont donc été segmenté en cette unité ([4], [5]). La durée de chaque syllabe a ensuite été mesurée.

3.2. Résultats

Table 1 : données recueillies pour chaque sujet : âge en jours, nombre d'occurrences, durée moyenne et écart-type

Sujets	Age en jours	Nombre d'occurrences	Durée moyenne (ms)	Ecart-type (ms)
5	231	129	484	141
17	273	13	349	70
13	276	35	386	156
21	289	10	592	206
19	296	11	519	176
14	331	46	466	170
3	333	90	471	150
15	334	19	463	186
22	338	32	394	203
8	355	17	370	128
6	359	10	357	117

Notons que l'âge en jours des sujets a été retenu plutôt que l'âge en mois. Cela dans le but de nous affranchir du problème de quantification associé à l'établissement de tranches d'âges dans lesquelles insérer chaque sujet (ex 8 mois, 9 mois, 10 mois, 11 mois, 12 mois). En effet, une différence de 2 jours entre deux sujets pouvait impliquer un changement de classe, à notre sens, peu pertinent.
La visualisation de nos données peut se faire sous la forme de la figure 1.

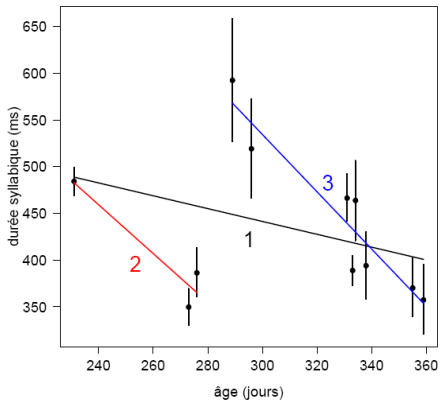


Figure 1 **:** Evolution de la durée syllabique (ms) au stade du babillage chez 11 sujets.

Chaque point correspond à la durée moyenne du cycle mandibulaire pour un sujet et la barre verticale qui lui est associée, à son erreur type. La droite « 1 » représente la régression linéaire pour l'ensemble des points. Les droites « 2 » et « 3 » représentent les régressions linéaires pour les deux ensembles de points de part et d'autre l'âge de rupture de 280 jours. Nous avons alors testé, par une régression linéaire, l'évolution de la durée syllabique sur l'ensemble de la période couverte (droite de régression 1 sur la Fig. 1). Celle-ci semble montrer une diminution générale de la durée syllabique avec l'âge, mais les analyses statistiques ($F [1,9] = 1,32$, $p > 0.28$) indiquent que cette diminution de la durée moyenne de la syllabe avec

l'âge n'est pas significative. En revanche, l'analyse des faits par un modèle non linéaire (discontinu, linéaire par morceaux, deux droites de régression) a permis la distinction de deux stades développementaux qui, au vu des connaissances théoriques, correspondent au stade du babillage redupliqué (âge < 280 jours) et du babillage varié (âge > 280 jours). L'âge optimal de transition entre ces deux stades est celui pour lequel les deux droites de régression (2 et 3 sur la Fig. 1) fournissent les plus petites erreurs résiduelles par rapport aux données. Nous avons ensuite étudié l'évolution temporelle de la syllabe au sein de chacune des deux zones statistiquement déterminées. Une première lecture qualitative des résultats obtenus s'oriente vers l'observation d'une diminution de la durée syllabique moyenne au sein de chaque stade développemental. Cependant, la tendance à la décroissance observée au stade du babillage redupliqué n'est pas statistiquement significative (droite de régression 2 sur la Fig.1, $F [1,1] = 8,51$, $p > 0.21$). Notons que cette observation mériterait d'être vérifiée en augmentant la taille de l'échantillon. Au contraire, la diminution observée au stade du babillage varié est quant à elle statistiquement significative (droite de régression 3 sur la Fig.1, $F [1,6] = 53,54$, $p < 0.001$). La durée syllabique moyenne, au début de cette deuxième phase, est située entre 500 et 600 ms (fréquence entre 1,7 et 2Hz), puis elle diminue jusqu'à venir se situer entre 350 et 400 ms (fréquence entre 2,5et 2,8Hz). Précisons que la droite de régression 3 a comme équation : $y = a(x\text{-}289)+b$, où y est la durée en ms est x l'âge en jours. La valeur 289 correspond à l'âge du premier enfant après la rupture. Les coefficients estimés sont (l'intervalle de confiance s'élevant à 95% est indiqué entre crochets) : a = -3.07 [-4,21 ; -1,92] ms/jour, b = 568 [515 ; 622] ms.

4. DISCUSSION ET CONCLUSION

Cette étude suggère que la diminution de la durée syllabique, c'est-à-dire l'accélération de la fréquence oscillatoire nécessaire à la mise en place du rythme spécifique de la parole à partir du rythme biologique plus lent de la succion/déglutition, ne se ferait pas de manière monotone. En effet, une tendance à l'accroissement de la durée syllabique semble apparaître à 10 mois. Une tentative d'explication de ce phénomène pourrait être l'hypothèse d'un retour au rythme physiologique quand les productions commencent à se diversifier et que le contrôle moteur se complexifie. Le retour à des patrons plus simples comme conséquence de la complexification de la tâche, a déjà été défendu dans la littérature sur l'acquisition du langage ([7], [23]). Les productions du babillage varié impliquent une reconfiguration de la quantité d'abaissement et d'élévation de la mandibule, de syllabe à syllabe. On pourrait alors considérer l'émergence de ces nouvelles manifestations comme une phase par laquelle le bébé explore les autres possibilités de sa mandibule. La modification des cibles mandibulaires d'une syllabe à l'autre introduirait une nouvelle complexité dans le champ articulatoire du bébé qui jouerait le rôle d'une perturbation du système redupliqué préalablement mis en place. Nous restons cependant prudents face à cette observation, car nous avons conscience que les sujets témoignant le mieux de ce phénomène sont ceux qui présentent le moins d'occurrences analysées. Mais la tendance n'en reste pas moins intéressante et reste une piste à creuser et à renforcer en augmentant l'échantillon, puisque l'accroissement observé entre les sujets les plus jeunes et les premiers sujets entrés dans le babillage varié, demeure lorsque l'on écarte ces derniers et que l'on confronte les sujets âgés de 9 mois (sujet 17 et 13) à ceux âgés de 11 mois (sujet 14, 3 et 15). A partir de 10 mois (289 jours), nous observons une diminution statistiquement significative de la durée syllabique jusqu'au terme de l'investigation (12 mois). La fréquence oscillatoire de la mandibule vient se situer entre 2,5Hz et 2,8Hz, et va dans le sens de nos hypothèses. Nous interprétons ces résultats comme une familiarisation, avec les nouvelles possibilités émergentes, qui peut alors laisser place à un mode fréquentiel spécifique plus rapide pour la fonction de parole.

Au terme de la période du babillage, le contrôle temporo-articulatoire semble émerger. Toutefois, il est loin d'avoir atteint sa maturité et doit encore s'affiner pour atteindre les caractéristiques temporelles de la parole de l'adulte. Nos jeunes apprentis locuteurs devront poursuivre leur apprentissage et continuer à augmenter leur rythme mandibulaire.

Remerciements à la Maison Interuniversitaire des Sciences de l'Homme d'Alsace (MISHA) et à l'ANR 07-CORP-018-01-DOCVACIM, 2008-2011.

BIBLIOGRAPHIE

[1] C. Abry, M. Stefanuto, A. Vilain and R. Laboissière. What can the utterance "Tan, Tan" of Broca's patient Leborgne tell us about the hypothesis of an emergent "babble-syllable" downloaded by SMA? In *Phonetics, phonology and cognition* J. Durand & B. Laks (eds), Oxford, University Press, pages 226-243, 2002.

[2] C. Bickley, B. Lindblom and L. Rough. Acoustic measures of rhythm in infants' babbling, or "All God's children got rhythm". In *Proceedings of the 12th International Congress on Acoustics*, Toronto, A6-4, 1986.

[3] P.M. Burke. Swallowing and the organization of sucking in the human newborn. *Child Development,* 48: 523-531, 1977.

[4] M. Canault, P. Perrier and R. Sock. L'émergence du contrôle segmental au stade du babillage : une étude acoustique. In *Actes des 26e Journées*

d'Etude sur la Parole, Dinard, pages 193-197, 2006.

[5] M. Canault. *L'émergence du contrôle articulatoire au stade du babillage : une étude acoustique et cinématique*. Thèse de doctorat NR, Université Marc Bloch, Strasbourg 2, 2007.

[6] C.K. Crook and L.P. Lipsitt. Neonatal Nutritive Sucking: Effects of taste stimulation upon sucking rhythm and heart rate. *Child Development,* 47 (2): 518-522, 1976.

[7] B.L. Davis, P.F. MacNeilage and C.L. Matyear. Acquisition of serial complexity in speech production: a comparison of phonetic and phonological approaches in first word production. *Phonetica,* 59: 75-107, 2002.

[8] J.K. Dolota, B.L. Davis and P.F. MacNeilage. Characteristics of the rythmic organization of vocal babbling: implications for an amodal linguistic rhythm. *Infant Behavior and development*, 31: 422-431, 2008.

[9] V. Ducey-Kaufmann. *Le cadre de la parole et le cadre du signe : un rendez-vous développemental*. Thèse de doctorat, Université Stendhal Grenoble III, 2007.

[10] L. Fogassi and P.F. Ferrari. Mirror neurons, gestures and language evolution. *Interaction Studies,* 5 (3): 345-363, 2005.

[11] C.R. Gallistel. *The organisation of action: a new synthesis*, Hillsdale, Erlbaum, 1980.

[12] E.C. Goldfield. *Emergent Forms Origins and Early Development of Human Action and Perception*. Oxford university press, 1995.

[13] E.C. Goldfield and P.H. Wolff. A dynamical systems perspective on infant action and its development. In *Theories of infant development* J.G. Brenner & A. Slater (eds), Oxford, Blackwell Publishing, pages 3-29, 2003.

[14] J.R. Green, C.A. Moore, M. Higashikawa and R.W. Steeve. The physiologic development of speech motor control: lip and jaw coordination. *Journal of Speech Language and Hearing Research,* 43: 239-255, 2000.

[15] J.R. Green, C.A. Moore and K.J. Reilly. The sequential development of jaw and lip control for speech. *Journal of Speech Language and Hearing Research,* 45: 66-79, 2002.

[16] U. Jürgens. Speech evolved from vocalization, not mastication. Commentary to MacNeilage P.F. (1998). The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences,* 21: 519-520, 1998.

[17] K. Kaye and A.J. Wells. Mother's jiggling and the burst-pause pattern in neonatal feeding. *Infant Behavior and Development*, 3:29-46, 1980.

[18] F.J. Koopmans Van Beinum. Cyclic effects on infant speech perception, early sound production, and maternal speech. In *Proceedings of the Institute of Phonetic Sciences (IFA)* 17, pages 65-78, 1993.

[19] D.P. Kuehn and K. Moll. A cinefluorographic investigation of CV and VC articulatory velocities. *Journal of Phonetics,* 4:303-320, 1976.

[20] J.P. Lecanuet. Des rafales et des pauses : les succions prénatales. *Spirale,* 22: 37-47, 2002.

[21] B. Lindblom. Economy of speech gestures. In *The production of speech* MacNeilage P.F. (Ed.). New York, Springer, pages 217-245, 1983.

[22] P.F. MacNeilage. The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences,* 21: 499-546, 1998.

[23] P.F. MacNeilage and B.L. Davis. Intersyllabic and word-level regularities in early acquisition. In *Proceedings of the 15th International Congress of Phonetics Sciences*, Barcelona, pages 383-386, 2003.

[24] T. Morimoto, T. Inoue, T. Nakamura and Y. Kawamura. Frequency dependent modulation of rhythmic human jaw movements. *Journal of Dental Research,* 68: 1310-1314, 1984.

[25] G. Rizzolatti, L. Fadiga, V. Gallese and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research,* 3: 131-141, 1996.

[26] J.L. Ruark and C.A. Moore. Coordination of lip muscle activity by two-year-old children during speech and nonspeech tasks. *Journal of Speech, Language, and Hearing Research,* 40: 1373-1385, 1997.

[27] B.L. Smith and T.E. Gartenberg. Initial observations concerning developmental characteristics of labio-mandibular kinematics. *Journal of the Acoustical Society of America,* 75 (5): 1599-1605, 1984.

[28] V.N. Sorokin, T. Gay and W.G. Ewan. Biomechanical correlates of jaw movements. *Journal of the Acoustical Society of America,* 68:S1, S32, 1980.

[29] P.H. Wolff. The serial organization of sucking in the young infant. *Pediatrics* 42 (6):943-956, 1968.

La syllabification de séquences VCV en irlandais : une étude de perception

Máire Ní Chiosáin et Pauline Welby†**

*An Coláiste Ollscoile Bhaile Átha Cliath
Scoil na Gaeilge, an Léinn Cheiltigh, Bhéaloideas Éireann agus na Teangeolaíochta
Áras Newman, Belfield, Baile Átha Cliath 4, Irlande
maire.nichiosain@ucd.ie, http://www.ucd.ie/icsifl/mairenc.html

†Laboratoire Parole et Langage
CNRS – Université de Provence
5 avenue Pasteur, BP 80975, 13604 Aix en Provence, France
pauline.welby@lpl-aix.fr, http://www.lpl.univ-aix.fr/~welby/

ABSTRACT

We examined whether Irish speakers syllabify intervocalic consonants as codas (e.g., *póca* 'pocket' /po:k.ə/ CVC.V), as claimed by many authors, but contrary to claims in phonological theory of a universal preference for syllables with onsets. We conducted a perception experiment using a part-repetition task and auditorily presented stimuli consisting of *VCV items* with a single medial consonant (C_m), varying in the length of V1 and the manner of articulation of C_m (e.g., *póca* /po:kə/ 'pocket', *lofa* /lofə/ 'rotten'), as well as *VCCV items* (e.g., *masla* /maslə/ 'insult', *canta* /ka:ntə/ 'English'). Response patterns were similar to those found for other languages: listeners preferred syllables with onsets, often treated C_m as ambisyllabic, syllabified C_m as a coda more often when V1 was short, and dispreferred stops as codas. Our findings are relevant to all research on Irish that makes reference to the syllable.
Keywords: syllabification, Irish, speech perception

1. INTRODUCTION

Comme toutes les langues celtiques, l'irlandais[1] comporte un grand nombre de propriétés typologiquement atypiques. Les plus connues concernent l'ordre de mots Verbe-Subjet-Objet (*Ólann Seán tae* 'Seán boit du thé') et les mutations initiales qui modifient la consonne initiale d'un mot selon le contexte (*Páras* /pa:rəs/ 'Paris' , *ó Pháras* /fa:rəs/ 'de Paris', *i bPáras* /ba:rəs/ 'à Paris'). Dans cette étude, nous avons étudié la structure syllabique de l'irlandais, dont le caractère supposé exceptionnel est souvent évoqué, mais qui a été jusqu'ici très peu exploré.

Selon les théories phonologiques, il existe une préférence universelle pour des syllabes commençant par une consonne d'attaque (voir, par exemple, [1]). Un mot comme *bateau* serait donc syllabifié /ba.to/ (CV.CV) plutot que /bat.o/ (CVC.V). À notre connaissance, on postule une syllabification qui privilégie des syllabes sans attaque que pour 4 langues : le kunjen [19] et l'arrernte [5] (des langues australiennes) ; le gaélique écossais [3] et l'irlandais. De nombreuses descriptions de l'irlandais proposent la structure CVC.V pour des mots comme *púca* 'fantôme' et *gúna* 'robe' [10,14,17], même si l'on propose aussi la structure plus attendue CV.CV [18]. Les motivations pour ce découpage ne sont pas toujours claires. De plus, pour certains auteurs [4,9,16], l'association phonologique de la consonne médiale n'est pas claire et ressemble à la notion d'ambisyllabicité, où la consonne pourrait appartenir à la fois à la première et à la deuxième syllabe.

Pour étudier cette question, nous nous sommes appuyées sur une tâche de syllabification explicite classique [7,20], la tâche de *répétition de syllabe.*

En plus d'élargir nos connaissances sur la syllabification en soi, le projet comporte aussi des implications pour toute recherche sur l'irlandais utilisant la syllabe, concept central dans la phonologie. Une étude récente d'alignement tonal [8], par exemple, a utilisé une syllabification CVC.V, un choix qui influence bien sûr les conclusions tirées à partir des résultats expérimentaux.

2. METHODE

2.1. Participants

Au total 18 locuteurs natifs de l'irlandais du Connemara ont participé à l'expérience, dont 16 remplissaient les critères d'inclusion à l'étude. Les participants retenus étaient tous étudiants à l'université et avaient un âge moyen de 26 ans. Chaque participant a reçu 15€ pour sa participation, qui a duré une heure environ.

2.2. Stimuli

Items VCV

Nous avons construit un corpus de 96 mots contenant des séquences VCV. Ces *items VCV* consistaient en des mots bisyllabiques de structure (C)C'V:CV ou (C)C'VCV. En irlandais, les mots monomorphémiques de cette structure ont quasiment tous le stress lexical sur la première syllabe. Il y avait 8 groupes de items VCV, qui variaient en

[1]Les chercheurs et les Irlandais appellent cette langue *Irish* (*irlandais*) et réservent le terme *Gaelic* (*gaélique*) pour sa langue sœur le gaélique écossais. *Irlandais* est aussi le terme officiel utilisé au sein de l'Union Européenne.

fonction de la longueur phonologique de la première voyelle (V1 : longue ou courte) et du mode d'articulation de la consonne médiale (C_m : plosive, fricative, liquide ou nasale) (par. Ex., *guta* /gutə/ 'voyelle', *póca* /po:kə/, 'poche', *lofa* /lofə/, 'pourri', *Nóra* /no:rə/ prénom de fille, *mála* /ma.lə/ 'sac', *dara* /darə/ 'deuxième', *gúna* /gu:nə/ 'dress', *cuma* /kumə/ 'apparence').[2] La voyelle de la deuxième syllabe (V2) était toujours courte.

Nous avons contrôlé les facteurs dont l'influence sur la syllabification a été démontrée, par exemple, en n'utilisant que des mots monomorphémiques et en évitant des formes fléchies et des mots dont l'orthographe comportait des doubles consonnes (par ex., *giolla* 'esclave').

Items VCCV

Nous avons également inclus 72 mots supplémentaires de structure (C)C'V:CCV et (C)C'VCCV pour donner plus de variété à la structure syllabique des mots et pour nous permettre de faire des analyses préliminaires sur la syllabification des séquences de consonnes intervocaliques. Ces *items VCCV* comprenaient 3 types de séquences de consonnes dont 36 items contenant des séquences de consonnes intervocaliques ne constituant pas un groupe d'attaque légal en irlandais (par ex., *Béarla* 'anglais', *fáilte* 'bienvenue'). L'inclusion d'items de cette structure « empêche les participants de segmenter toujours après la première voyelle » (notre traduction, [7], page 182), puisque pour ces items, les participants mettent les frontières toujours après la première consonne de la séquence. Faute de place, nous nous donnerons pas ici les résultats pour les items VCCV, mais nous nous concentrerons exclusivement sur les items VCV.

2.3. Procédure

La liste de tous les mots a été lue, dans un ordre randomisé, par un locuteur natif de l'irlandais du Connemara de 27 ans. L'enregistrement a été effectué par un enregistreur numérique Marantz PMD 660 avec un microcasque Shure SM10A à un taux d'échantillonage de 48 kHz. Chaque mot a été sauvegardé dans un fichier son individuel, à l'aide d'un script Praat [2].

Les stimuli (96 items VCV et 72 items VCCV) ont été pseudo-randomisés et divisés en 3 blocs, avec un nombre égal d'items de chaque type dans chaque bloc. Chacun des 3 blocs a été présenté dans deux conditions, *répétition de la 1ère partie* et *répétition de la 2ème partie*. Nous avons expliqué aux participants qu'ils allaient jouer à un jeu de langage dans lequel ils écouteraient des mots et répéteraient soit 'la 1ère partie' (« *an chéad chuid* ») soit 'la 2ème partie' (« *an dara cuid* ») de chaque mot (voir [7]). Comme dans les études antérieures sur les décisions de syllabification intuitives, nous avons expliqué les règles du « jeu » en donnant des exemples dans lesquels les « parties » correspondaient à des frontières de morphèmes. Par exemple, pour la 1ère partie, (« si je dis *cispheil* ['basketball'], vous direz *cis* ['panier (*basket*)'] ») et pour la 2ème partie, « si je dis *lánstad* ['point' (marque de ponctuation)], vous direz *stad* ['arrêt'] »). Nous n'avons jamais utilisé le terme *siolla(í)* 'syllabe(s)'.

Les stimuli ont été présentés à l'aide d'un script Praat par un ordinateur portable à travers des écouteurs Koss PRO3AA Titanium, avec un intervalle inter-stimulus de 3 secondes. Les réponses des participants ont été enregistrées par un enregistreur numérique Marantz PMD 660 avec un microphone Studio Projects B1 et ont été transcrites après l'expérience.

Après 2 sessions d'entraînement, les participants ont complété six blocs (3 blocs x 2 parties) expérimentaux avec une pause à la fin de chaque bloc. Pour chaque bloc successif, nous avons altérné la condition expérimentale (répétition de la 1ère ou de la 2ème partie).

3. RESULTATS

Les réponses ont été analysées de 2 façons différentes (voir [7]). Nous avons d'abord analysé les réponses pour la première partie (syllabe) et la deuxième partie (syllabe) séparément, en indiquant pour chaque réponse si la réponse contenait ou non la consonne médiale (C_m). Les réponses « 1ère partie » ont été codées soit *CV* soit *CVC* (par ex., *gúna* 'robe' – *gú-* ou *gún-*) quelle que soit la réponse du participant dans la condition « 2ème partie ». Les réponses dans la condition « 2ème partie » ont été codées soit CV, soit V (par ex., *gúna* – *-na* ou *-a*) soit *rime poétique*[3] pour les cas dans lesquels les participants ont donné des réponses telles que *-úna* pour *gúna*. Ensuite, nous avons analysé les réponses d'ambisyllabicité, c'est-à-dire, les cas pour lesquels la consonne médiale a été attribuée à la fois à la coda de la première syllabe et à l'attaque de la deuxième syllabe (une réponse « 1-2 » selon le terme de [20], par ex., pour *gúna* 'robe', réponse « 1ère partie » : *gún-*, réponse « 2ème partie » : *-na*). Les deux auteurs ont écouté les enregistrements et ont fait un codage indépendant. Les cas de désaccord étaient très rares (10 items, soit 0,37 % des données). Les résultats sont présentés dans la Table 1.[4]

3.1. Réponses « 1ère Partie »

Nous avons d'abord fait une ANOVA pour les réponses dans la condition « 1ère partie » en utilisant la transformation arcsinus de la moyenne des réponses CVC (les réponses dans lesquelles la consonne médiale a été syllabifiée comme coda de la première syllabe) comme variable dépendante et la longueur de V1 (longue, courte) et le mode d'articulation de C_m (plosive, fricative, nasale,

[2] Un accent aigu (*síneadh fada*) signifie une voyelle longue.

[3] Nous utilisons le terme *rime poétique* pour distinguer ce pattern de réponse du concept phonologique du rime syllabique (c'est-à-dire, le noyau d'une syllabe et les consonnes tautosyllabiques qui le suivent).

[4] La somme des pourcentages n'est pas toujours égale à 100 % à cause des arrondis.

Table 1 : Résultats du test perceptif (items VCV)

		Partie 1 (%)			Partie 2 (%)				Ambi (%)	
V1	C médiale	CV (0)	CVC (1)	autre	CV (1)	V (0)	Rime poétique	autre	oui	Non
longue	plosive	57,3	42,7	0	87,5	4,2	7,8	0,5	38,0	62,0
longue	fricative	37,0	62,5	0,5	88,5	4,2	6,8	0,5	55,2	44,8
longue	nasale	39,1	58,3	2,6	87,5	5,7	6,8	0	49,0	51,0
longue	liquide	53,6	45,3	1,0	90,1	1,6	7,3	1,0	40,1	59,9
courte	plosive	44,8	54,7	0,5	89,1	5,7	4,7	0,5	49,0	51,0
courte	fricative	21,4	78,6	0	83,9	7,3	6,8	2,1	66,7	33,3
courte	nasale	23,4	75,5	1,0	82,3	8,3	6,3	3,1	63,0	37,0
courte	liquide	18,8	79,7	1,6	80,7	9,4	6,3	3,6	62,5	37,5

liquide) comme variables indépendantes.

Les résultats montrent un effet principal de la longueur de V1 ($F1(1, 120) = 13,91$, $p < 0,001$, $F2(1,88) = 45,36$, $p < 0,001$). Les participants ont donné moins de réponses CVC quand V1 était longue (53 %) que quand elle était courte (73 %). Nous avons aussi observé un effet principal du mode de C_m ($F1(3,120) = 3,01$, $p < 0,05$, $F2(3,88) = 10.43$, $p < 0,001$). Les participants ont donné des réponses CVC pour 71 % des fricatives, 68 % des nasales, 63 % des liquides, mais uniquement 49 % des plosives. Il y avait également une interaction entre la longueur de V1 et le mode de C_m ($F2 = 3,30$, $p < 0,05$), mais cette interaction n'était pas significative dans l'analyse par sujets. ($F2 = 1,10$, $p = 0,35$). Nous constatons aucune préférence pour des syllabes ouvertes. Pour les analyses par items, les résultats de tests post hoc Scheffé montrent que les plosives se distinguaient des fricatives, des nasales et des liquides ($p < 0,01$ pour toutes les analyses). Pour les analyses par sujets, néanmoins, uniquement le test plosives vs. fricatives approchait la significativé ($p = 0,06$). Certains auditeurs ont préféré mettre la consonne médiale dans la première syllabe quand cette consonne n'était pas une plosive. Pour ces auditeurs, les plosives ne se comportent pas comme les autres modes de consonnes.

3.2. Réponses « 2ème partie »

Pour les réponses dans la condition « 2ème partie », nous avons effectué une deuxième ANOVA en utilisant la transformation arcsinus de la moyenne des réponses CV comme variable dépendante et la longueur de V1 et le mode d'articulation de C_m comme variables indépendantes. Il n'y avait aucun effet principal du mode de C_m. Par contre, il y avait un effet principal de la longueur de V1, significatif par items, mais pas par sujets ($F1(1,120) = 2,72$, $p = 0,10$; $F2(1,88) = 11,65$, $p < 0,01$). Les participants ont donné plus de réponses CV pour les mots avec des V1 longues (96 % des réponses qui n'étaient pas des rimes poétiques) que pour des mots avec des V1 courtes (91 %).

3.3. Réponses ambisyllabiques

Nous avons fait une troisième ANOVA en utilisant la transformation arcsinus de la moyenne des réponses ambisyllabiques comme variable dépendante et la longueur de V1 et le mode d'articulation de C_m comme variables indépendantes. À travers toutes les conditions, environ la moitié (53 %) des réponses étaient bisyllabiques. Comme la Table 1 le montre, les patterns observés pour l'analyse des réponses ambisyllabiques sont très similaires aux patterns pour les réponses « 1ère partie », ce qui n'est guère surprenant, puisque les participants ont donné 90 % de réponses CV pour la « 2ème partie ». Les effets principaux de la longueur de V1 ($F1(1,120) = 8,59$, $p < 0,01$, $F2(1,88) = 25,63$, $p < 0,001$) et du mode de C2 ($F1(1,120) = 2,01$, $p = 0,12$, $F2(3,88) = 6,45$, $p < 0,01$) étaient significatifs.

4. DISCUSSION

Nos résultats vont à l'encontre de l'idée que la syllabification de l'irlandais serait une exception typologique. Nous n'avons pas vérifié l'hypothèse selon laquelle les Irlandophones préféreraient produire des syllabes sans attaque ou syllabifier des consonnes intervocaliques uniquement en tant que codas. Pour les réponses Partie 2 (pour les items VCCV comme pour les items VCV), les participants ont préféré des réponses CV (+94 %) plutôt que des réponses V sans attaque. Pour les items VCV et VCCV, les participants ont préféré syllabifier les consonnes médiales comme coda, mais ils ont presque toujours aussi mis ces consonnes comme consonnes d'attaque dans leurs réponses Partie 2. Ce pattern d'ambisyllabicité est similaire à ce que d'autres chercheurs ont trouvé pour maintes autres langues. Comme le remarquent [12] et [15] l'ambisyllabicité semble été présagée dans les

descriptions des dialects de l'irlandais, même si cette notion théoretique n'avait pas encore été articulée.

En plus des résultats de notre test perceptif, certains patterns de la langue ne sont pas compatibles avec une préférence pour des syllables VC. Il existe notamment plusieurs processus dans lesquels une syllabe commençant par une voyelle prend une attaque de syllabe (*ag dul go hAix* 'va à Aix', cf. *Ag dul go Páras* 'va à Paris' ; *san Íoslainn* 'en Islande', cf. *sa Danmhairg* 'au Danemark').

Nos résultats s'accordent en général avec les résultats des études antérieures sur d'autres langues. Par exemple, des consonnes plus sonantes (par ex., les nasales et les liquides) sont plus susceptibles que des obstruentes d'être considérées comme des codas ou des consonnes ambisyllabiques, et des voyelles courtes ont tendance à attirer des consonnes.

Néanmoins, nous constatons une opposition entre les items de structure $VC_{plosive}V$ et les items ayant des consonnes d'autres modes d'articulation en position intervocalique (fricatives, nasales, liquides). Cette opposition est peu attestée dans d'autres langues (mais voir [11]), mais rappelle les règles de versification des poètes bardiques ([13]) et un éventuel rôle de la *résonance* dans la sonorité (proposé par [6]).

REMERCIEMENTS

Nous remercions Clíona Ní Chiosáin, Niall Ó Ciosáin, Anna Ní Ghallachair, Michal Boleslav Měchura, Brian Ó Raghallaigh, Kayla Reed, Michelle Tooher et John Walsh pour leur aide avec la mise en place de l'expérience, Sophie « den Oigheann » Dufour, Roibeard Espesser, Christine « Ní Mhuilleoir » Meunier et Serge « Pónaire » Pinto pour leurs conseils et leur assistance technique, les publics à nos présentations au congrès Formal Approaches to Celtic Linguistics (Tucson, É-U) et au Laboratoire Parole et Langage pour leur feedback précieux, et Foras na Gaeilge pour son soutien financier. Nous remercions également nos participants.

Nous sommes très reconnaissantes au regretté Ciarán Ó Con Cheanainn, qui a prêté sa voix aux enregistrements, et nous dédions cet article à sa mémoire.

BIBLIOGRAPHIE

[1] J. Blevins. The syllable in phonological theory. Dans J. Goldsmith (éd.), *The Handbook of Phonological Theory*, pages 206–244. Oxford: Blackwell, 1995.

[2] P. Boersma et D. Weenink. Praat: Doing phonetics by computer. logiciel. http://www.praat.org, 2008.

[3] C. Borgstrøm. The dialect of Barra in the Outer Hebrides. *Norsk Tidsskrift for Sprogvidenskap* 8:71–242, 1937.

[4] R. Breatnach. *The Irish of Ring, Co. Waterford: A Phonetic Study*. Dublin Institute for Advanced Studies (DIAS), Dublin, 1947.

[5] G. Breen. Arrernte: A language with no syllable onsets. *Linguistic Inquiry* 30:1–25, 1999.

[6] G. N. Clements. Does sonority have a phonetic basis? Comments on the chapter by Vaux. Dans E. Raimy et C. Cairns (éds.), *Contemporary Views on Architecture and Representations in Phonological Theory*, pages 165–175. Cambridge: MIT Press, 2009.

[7] A. Content, R. Kearns et U. Frauenfelder. Boundaries versus onsets in syllabific segmentation. *Journal of Memory and Language* 45:177–199, 2001.

[8] M. Dalton et A. Ní Chasaide. Tonal alignment in Irish dialects. *Language and Speech* 48:441–464, 2007.

[9] T. de Bhaldraithe. *The Irish of Cois Fhairrge, Co Galway: A phonetic study*. DIAS, Dublin, 1945.

[10] S. de Búrca. *The Irish of Tourmakeady, Co. Mayo: A phonemic study*. DIAS, Dublin, 1958.

[11] S. Gillis et G. de Schutter. Intuitive syllabification: universals and language specific constraints. *Journal of Child Language* 23:487–514, 1996.

[12] A. D. Green. The Prosodic structure of Irish, Scots Gaelic, and Manx. Thèse de doctorat, Cornell University, 1997.

[13] Knott, E. *An Introduction to Irish Syllabic Poetry of the Period 1200-1600*. Dublin: DIAS, 1974.

[14] R. Mhac an Fhailligh. *The Irish of Erris, Co. Mayo.* DIAS, Dublin, 1968.

[15] M. Ní Chiosáin. Topics in the phonology of Irish. Thèse de doctorat, University of Massachusetts, 1991.

[16] E. C. Quiggin. *A dialect of Donegal being the speech of Meenawannia in the parish of Glenties*. Cambridge University Press, Cambridge, 1906.

[17] S. Ó Searcaigh. *Foghraidheacht Ghaedhilge an Tuaiscirt* [La phonétique du gaélique du Nord]. Browne & Nolan, Belfast, 1925.

[18] M.-L. Sjœstedt. *Phonétique d'un parler irlandais de Kerry*. Librairie Ernest Leroux, Paris, 1931.

[19] B. Sommer. The shape of Kunjen syllables. Dans D. Goyvaerts (éd.), *Phonology in the 1980's*, pages 231–44. Story-Scientia, Gent, 1981.

[20] R. Treiman et C. Danis. Syllabification of intervocalic consonants. *Journal of Memory and Language* 27:87–104, 1988.

La densité des idées : une mesure pertinente de la dégradation linguistique chez les patients Alzheimer

Hye Ran Lee & Melissa Barkat-Defradas

Laboratoire Praxiling UMR5267-CNRS & Université de Montpellier
17, rue de l'abbé de l'Épée 34090 Montpellier
hlee1@etu.univ-montpellier3.fr ; melissa.barkat@univ-montpellier3.fr

ABSTRACT

Oral language in Alzheimer's disease (AD) participants and healthy older controls was analyzed using propositional analysis. The idea density was evaluated for its usefulness in discriminating between healthy elderly subjects and AD patients. Results suggest that scores of idea density accounts for a marker of pathological cognitive decline. This linguistic cue can be regarded as a reliable measure for future medical screening.

Keywords: Psycholinguistics, Discourse analysis, Oral language, Idea density, Alzheimer's disease.

1. Introduction

La démence sénile est caractérisée par l'affaiblissement progressif et global des fonctions cognitives incluant l'altération de la mémoire à court et/ou à long terme accompagnée d'au moins une aphasie, une apraxie, une agnosie ou un trouble des fonctions exécutives (American Psychiatric Association [1]). La dégradation de la capacité langagière est un élément caractéristique de la démence, 88-95% de patients éprouvent des déficits linguistiques (Bayles & Kazniak [2] ; Cardebat, Aithamon & Puel [3] ; Kemper & Altmann [4]). Le déficit linguistique est signalé dès la première description de caractéristiques clinico-pathologique de la pathologie (Alzheimer [5]). Le domaine lexico-sémantique du langage est le plus vulnérable dans la maladie d'Alzheimer (MA) (Appell, Kertesz & Fisman [6]). En effet, le phénomène anomique et la difficulté à trouver un mot sont fréquemment observés au début de la maladie et constituent un des symptômes les plus précoces de la maladie (Bayles & Tomoeda [7]). En revanche, les aspects phonologique et syntaxique sont relativement bien préservés jusqu'en fin d'évolution (Hart [8]; Bickel, Pantel, Eysenbach & Schröder [9] ; Kemper & Altmann [4]). Toutefois, des études récentes ont montré une simplification de la structure syntaxique dans le discours de patients MA (Kemper, LaBarge, Ferraro, Cheung & Storandt [10] ; Lyons, Kemper, LaBarge, Ferraro, Balota & Storandt [11] ; Lee, Barkat-Defradas, Gayraud [12]) ; Croot, Hodges, Xuereb, & Patterson [13] ont rapporté des cas de patients montrant un syndrome d'aphasie non fluente sans trouble de mémoire diagnostiqués – à l'examen *post mortem* – comme atteints de la maladie d'Alzheimer.

La recherche sur les troubles du langage associés à différents types de démence permet ainsi d'isoler les dysfonctionnements particuliers à chaque type de démence et à élargir la gamme des outils utiles pour le diagnostic différentiel de la MA. Par ailleurs, d'autres études ont révélé une relation étroite entre capacité linguistique et risque de développer la MA. En 1984, Brian Butterworth [14] a fait l'hypothèse que l'ancien président des E.U, Ronald Reagan, devait souffrir de cette pathologie, sur la base de l'analyse de ses discours électoraux et ce, dix ans avant que ce diagnostic ne soit effectivement confirmé par les médecins. De même, une étude consacrée aux écrits de l'écrivain britannique, Iris Murdoch, a révélé - à travers l'étude comparative de l'ensemble de sa production littéraire - une certaine détérioration linguistique laquelle était particulièrement remarquable dans son dernier ouvrage publié un an avant que l'auteure ne soit en effet diagnostiquée comme souffrant de la MA (Garrard, Maloney, Hodges & Patterson [15]). Enfin, une importante étude longitudinale et épidémiologique relative au vieillissement cognitif, (i.e. The Nun Study) a montré qu'une faible capacité linguistique durant le jeune âge – et déterminée par l'analyse d'écrits autobiographiques produits par des religieuses âgées de 18 à 32 ans – était associée à des performances cognitives moindres entre 75 et 93 ans (Snowdon, Kemper, Mortimer, Greiner, Wekstein & Markesbery [16]). Différentes études se sont ainsi attachées à examiner l'hypothèse selon laquelle des capacités linguistiques peu élevées signaleraient un développement cognitif et neurologique non optimal, et donc un facteur de sensibilité accrue au déclin cognitif lié à la maladie d'Alzheimer (Kemper, Grenier, Marquis, Prenevost & Mizner [17] ; K. Riley, D. Snowdon & W. Markesbery [18]).

Ces travaux sont intéressants dans une perspective préventive de la maladie d'Alzheimer. En effet, ils montrent que l'étude du langage pourrait permettre d'identifier les populations à risque avant même l'apparition des premiers symptômes de la maladie. Barkat-Defradas, Martin, Rico-Duarte & Brouillet [19] ont ainsi proposé un certain nombre d'indicateurs linguistiques pressentis comme étant *a priori* pertinents pour mesurer les capacités linguistiques : (i) la richesse lexicale (RL) (ii) la complexité syntaxique (CS) et, (iii) la densité des idées (DI). Contrairement aux autres compétences cognitives (e.g. mémoire, attention, vitesse de traitement), ces aspects du langage sont, semble-t-il, relativement résistants dans le vieillissement normal.

Ainsi, la présence d'un changement linguistique à ces différents niveaux dans le discours oral et/ou écrit des patients souffrant de la maladie d'Alzheimer peut être considérée comme un marqueur pertinent de déclin cognitif et pourrait contribuer d'une part à distinguer entre vieillissement cognitif normal et pathologique et d'autre part à un diagnostic précoce de la MA.

Parmi ces trois critères, nous nous intéressons, dans le cadre de cette étude, à la densité des idées (désormais DI) avec pour objectif d'évaluer la pertinence de la DI pour mesurer la dégradation linguistique associée à la MA.

En se basant sur les théories logiques et psychologiques, Kintsch ([20]) a développé un modèle d'analyse du traitement sémantique de l'information appelé analyse prédicative (désormais AP). L'auteur part du postulat que la forme dominante de la représentation cognitive du langage est de nature propositionnelle. Ainsi, « *Si l'on considère que la prédication qui s'exprime dans un message linguistique est une activité cognitive essentielle de l'homme et que, sous-jacent à la réalisation de surface, c'est-à-dire au mot, se trouve un concept, on peut estimer que l'analyse prédicative, outil de description sémantique des textes, est pour le psychologue la transcription d'une activité cognitive* » (Ghiglione, Kehenbosch & Landré [21]).

La proposition est la plus petite unité sémantique intégrée susceptible d'être traitée ou mémorisée. Un mot isolé seul ne suffit pas à créer une idée, c'est l'ensemble des propriétés et des relations s'y apportant qui permet d'appréhender et de produire la signification psychologique. Ainsi, une proposition est constituée d'un prédicat et d'un ou de plusieurs argument(s). Soit dans la phrase, « Le chien poursuivait un chat dans le jardin », les entités référentielle (être ou objet) « chien », « chat », « jardin » correspondent aux arguments; l'événement « poursuivre » et la relation dans l'espace « dans » correspondent aux prédicats qui définissent la relation entre les arguments. La représentation sémantique de cette phrase peut être codée comme:
P1. POURSUIVRE (chien, chat)
P2. DANS (P1, jardin)

La DI correspond à la quantification de cette activité cognitive. En ramenant le nombre de propositions sémantiques au nombre de mots produits dans le discours, on peut mesurer l'efficience de l'expression d'un sujet (Snowdon, Kemper, Mortimer, Greiner, Wekstein & Markesbery [16]; Vineeta & Bonnici [22] ; Baynes, Chand, Bonnici, Tomaszewski Farias [23]). Une valeur de DI élevée reflète l'aptitude d'un locuteur à exprimer efficacement ses idées ainsi que leurs interrelations complexes. En revanche, une valeur de DI faible peut révéler un discours peu efficient, du fait de l'utilisation d'un plus grand nombre de mots pour exprimer les idées essentielles. Ainsi, une faible densité des idées dans la production langagière peut indiquer l'altération de la capacité cognitive (Covington [24]).

L'efficacité de cette mesure a été validée à plusieurs reprises dans le domaine de la psycholinguistique appliquée en langue anglaise (Kintsch & Van Dijk [25] ; Thorson & Snyder [26]; Kemper, Grenier, Marquis, Prenevost & Mizner [17] ; Covington [24]). Nous proposons dans cette étude de l'appliquer au français.

2. Méthode

2.1. Sujets

11 patients cliniquement diagnostiqués comme étant atteints de la MA au stade léger et modéré – mesuré par MMSE (test neuropsychologique le plus utilisé dans le cadre du diagnostic démentiel, Folstein, Folstein & McHugh [27]) – et 11 sujets de contrôle (sujets âgés sains appariés aux patients en âge, sexe, et niveau socioculturel) ont participé à cette étude. Tous les patients (ou leurs tuteurs légaux) et tous les sujets âgés sains ont fourni un consentement signé et éclairé avant le démarrage de l'étude. Le profil des sujets est récapitulé dans la table 1.

Table 1 : Profil des sujets

Patients MA (n=11)	**Sujets âgés sains** (n=11)
Sexe 8 F + 3 H	Sexe 7 F + 4 H
Âge m= 78.27 (± 7.96)	Âge m = 78.27 (± 5.82)
Niveau socioculturel (grille de Poitrenaud) = 2.27 (±1.35)	Niveau socioculturel (grille de Poitrenaud) = 2.27 (±1.35)
MMSE = 22 (18< >26)	MMSE = 30

2.2. Procédure

Les données orales proviennent d'une série d'entretiens individuels semi-dirigés. Une narration libre de l'évocation d'un souvenir personnel a été proposée pour éliciter le discours. Après enregistrements, tous les entretiens ont été transcrits. Ce corpus a été tronqué de manière à garder environ 350 mots par transcription pour que les corpus soient comparables statistiquement. Dans un premier temps, un prétraitement manuel des données a été appliqué. Ce prétraitement consiste à marquer certaines caractéristiques spécifiques à l'oral telles que les phrases inachevées lesquelles sont inexploitables d'un point de vue automatique. Dans un second temps, un étiquetage grammatical automatique à l'aide de *TreeTagger* (Schmid [28]) a été effectué. Enfin, un calcul automatique de la DI a été réalisé à l'aide du programme *Densidées* dont le principe est de diviser le nombre de prédicats (typiquement les verbes, adjectifs, adverbes, prépositions, conjonctions) (Snowdon, Kemper, Mortimer, Greiner, Wekstein & Markesbery [16]) par le nombre total de mots (Lee, Gambette, Barkat-Defradas [29] ; Lee, Gambette, Maillé, Thuillier [30]).

3. Résultats

Les résultats indiquent que le groupe de patients atteints de MA a produit un score moyen de densité des idées inférieur à celui produit par le groupe de contrôle, 0.32 et 0.37 respectivement (voir Figure 1). La différence entre les deux groupes est significative (p= 0,003**). La densité des idées peut être dès lors considéré comme un critère fiable de différenciation des deux groupes de sujets et donc comme une mesure pertinente de dégradation linguistique liée à la MA

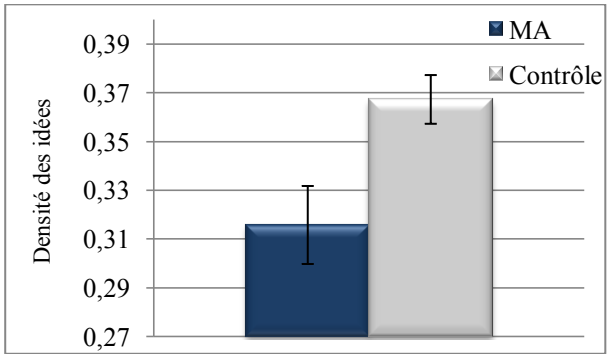


Figure 1 : Score moyen de la DI de groupe MA & groupe de contrôle.

S'il a été remarqué que ni l'âge ni le sexe n'influent sur le score de DI, le niveau socioculturel (mesuré en fonction du niveau d'éducation et de la profession exercée à l'aide du questionnaire de Kalafat, Hugonot-Diener, Poitrenaud [31]) est un facteur important à prendre en considération chez les sujets atteints de maladie d'Alzheimer. En effet, bien qu'aucune corrélation n'ait été observée chez la population de contrôle, on observe chez les patients MA des scores de DI différenciés selon le NSC. La dégradation linguistique – mesurée en termes de DI – est moindre pour les patients présentant un haut NSC que pour ceux de NSC inférieur (voir Figures 2 et 3).

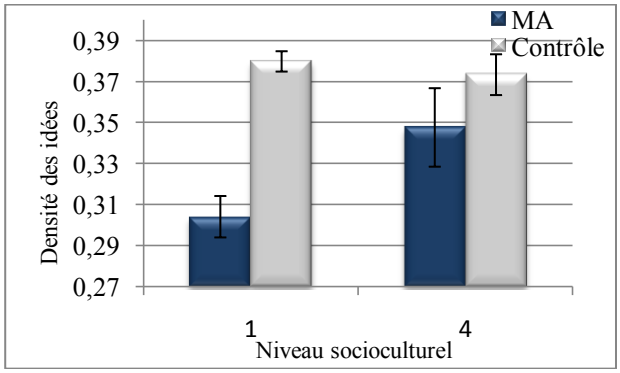


Figure 2 : Comparaison du score moyen de DI en fonction du NSC

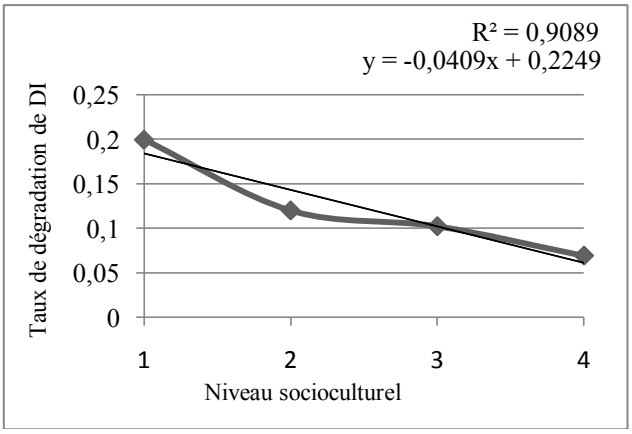


Figure 3 : Taux de dégradation de la DI en fonction du NSC chez les patients MA.

4. Discussion

L'analyse du discours oral de patients atteints de la MA *vs.* de personnes âgées saines a révélé que la densité des idées est une mesure sensible du déclin de la capacité langagière lié à cette démence. Cependant, il semble que le NSC influe sur cette mesure. La DI doit donc être pondérée avec tous les facteurs susceptibles de la faire varier indépendamment de la pathologie.

La dégradation de DI peut être le résultat de déficits sous-jacents comme l'altération de la mémoire sémantique, la fonction exécutive, la mémoire de travail. L'examen approfondi de données plus volumineuses et des variables mieux contrôlées permettraient d'éclairer la nature de cette détérioration.

5. Conclusion

L'étude du langage des personnes âgées saines et pathologiques permettrait de relever des variables pertinentes pour le diagnostic différentiel du vieillissement normal et pathologique.

Malgré la fréquence et l'importance des troubles du langage dans la maladie d'Alzheimer et des travaux consacrés à la dimension linguistique de cette pathologie, le niveau discursif – dont on sait qu'il présente des caractéristiques spécifiques chez le patient atteint de la MA – n'est, à l'heure actuelle, que peu étudié (Taler & Phillipes [32]). Compte tenu des limites diagnostiques des approches traditionnelles principalement fondées sur l'étude des performances mnésiques des sujets, la présente étude ouvre des perspectives intéressantes quant à l'intégration de ce niveau d'analyse dans le champ de l'évaluation des capacités cognitives des patients. Souvent négligée des praticiens impliqués dans le diagnostic et/ou dans la prise en charge des patients du fait d'une analyse longue, fastidieuse, coûteuse en temps et parfois subjective, l'analyse discursive des productions langagières envisagée d'un point de vue automatique peut ainsi contribuer à affiner le diagnostic précoce et/ou différentiel de la MA.

BIBLIOGRAPHIE

[1] American Psychiatric Association. *Diagnostic and statistical manual of Mental Disorders*. Washington DC: American Psychiatric Association, 1994.

[2] K. Bayles & A. Kazniak. *Communication and cognition in normal aging and dementia*. College Hill Press, London, UK, 1987.

[3] D. Cardebat, B. Aithamon & M. Puel. Les troubles du langage dans les démences de type Alzheimer In *Neuropsychologie clinique des démences : Évaluation et prises en charge,* pages 213-223, 1995.

[4] S. Kemper & L. Altmann. Dementia and language. In *Encyclopedia of neurosciences*, volume 3, pages 409-414, 2009.

[5] A. Alzheimer. Of a particular disease of the cerebral cortex. *Zentralblatt vur Nervenheilkunde und Psychiatrie*, 30: 177-179, 1907.

[6] J. Appell, A. Kertesz & M. Fisman. A study of language function in Alzheimer patients. *Brain and Language,* 17: 73–91, 1982.

[7] K. Bayles & C. Tomoeda. Caregiver report of prevalence and appearance order of linguistic symptoms in Alzheimer's patients. *The Gerontologist*, 3: 210-216, 1991.

[8] S. Hart. Language and dementia. *Psychological Medicine*, 18: 99–112, 1988.

[9] C. Bickel, J. Pantel, K. Eysenbach & J. Schröder. Syntactic comprehension deficits in Alzheimer's disease. *Brain and Language*, 71: 432–448, 2000.

[10] S. Kemper, E. LaBarge, R. Ferraro, H. Cheung & M. Storandt. On the preservation of syntax in Alzheimer's disease, *Archives of Neurology*, 50: 81–86, 1993.

[11] K. Lyons, S. Kemper, E. LaBarge, F. Ferraro, D. Balota & M. Storandt. Oral language and Alzheimer's disease: A reduction in syntactic complexity, *Aging and Cognition,* 1(4): 271–281, 1994.

[12] H. Lee, M. Barkat-Defradas & F. Gayraud. Le vieillissement normal et pathologique du langage: étude comparative des discours oraux. *6ème journées internationales de linguistique de corpus, Lorient,* 2009.

[13] K. Croot, J. Hodges, J. Xuereb & K. Patterson. Phonological and Articulatory Impairment in Alzheimer's Disease: A Case Series. *Brain and Language*, 75: 277-399, 2000.

[14] B. Butterworth. The Sunday Times, November 4, 1984.

[15] P. Garrard, L. Maloney, J. Hodges & K. Patterson. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128: 250-260, 2005.

[16] D. Snowdon, S. Kemper, J. Mortimer, L. Greiner, D. Wekstein & W. Markesbery. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life : findings from the Nyun Study, *JAMA*, 275: 528-532, 1996.

[17] S. Kemper, L. Greiner, J. Marquis, K. Prenevost & T. Mizner. Langage decline across life span : findings from the nun study, *Psychology and Aging*, 16(2): 227-239, 2001.

[18] K. Riley, D. Snowdon & W. Markesbery. Alzheimer's neurofibrillary pathology and the spectrum of cognitive function: findings from the Nun Study, *Anne Neurol*, 51: 567-577, 2001.

[19] M. Barkat-Defradas, S. Martin, L. Rico Duarte & D. Brouillet. Les troubles du langage dans la maladie d'Alzheimer. *28ème JEP, Avignon*, 2008.

[20] W. Kintsch. *The representation of meaning in memory*. Hillsdale, NJ, America, 1974.

[21] R. Ghiglione, C. Kehenbosch & A. Landré. *L'analyse cognitive-discursive*. Presse Universitaire de Grenoble, 1995.

[22] C. Vineeta & L. Bonnici. Quantifying Language Degradation in Alzheimer's Disease. *New Ways of Analyzing Variation 36 (NWAV 36)*, Philadelphia, 2007.

[23] K. Baynes, V. Chand, L. Bonnici & S. Tomaszewski Farias. Idea Density as a Measure of Communicative Skill in Alzheimer's Disease. *Midyear Metting*, 103, 2007.

[24] M. Covington. Idea Density: A potentially informative characteristic of retrieved documents. *Proceedings, IEEE SoutheastCon*, 2009.

[25] W. Kintsch & T. Van Dijk. Toward a model of text comprehension and production. *Psychological Review,* 85: 363-394.

[26] E. Thorson & R. Snyder. Viewer recall of television commercials : Predication from the propositional structure of commercial scripts. *Journal of Marketing Research*, 21: 127-136, 1984.

[27] M. Folstein, S. Folstein & P. McHugh. Mini-mental state. A pratical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12: 189-198, 1975.

[28] H. Schmid. Probabilistic Part-of-Speech tagging using decision trees. In *New Methods in language processing*, 154-164, 1994.

[29] H. Lee, P. Gambette & M. Barkat-Defradas. Utilisation de l'analyse textuelle automatique dans la recherché sur la maladie d'Alzheimer. Colloque international des jeunes chercheurs en didactique des langues et en linguistique, 2010.

[30] H. Lee, P. Gambette, E. Maillé & C. Thuillier. Densidées : calcul automatique de la densité des idées dans un corpus oral. *Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues, Québec*, soumis.

[31] M. Kalafat, L. Hugonot-Diener, J. Poitrenaud. Étalonnage français du MMS version GRECO. *Revue de neuropsychologie*, 13(2): 209-236, 2003.

[32] V. Taler & N. Phillipes. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. Journal of clinical and experimental neuropsychology, 30: 510-556, 2008.

Structure syllabique en portugais brésilien : une analyse typologique

Luciana Marques, Nathalie Vallée° et Didier Demolin**

Laboratoire de Phonologie, Université Libre de Bruxelles*, GIPSA-Lab. Grenoble, Université Stendhal°
lufmarques@terra.com.br, ddemoli@ulb.ac.be, nathalie.vallee@gipsa-lab.grenoble-inp.fr

ABSTRACT

This paper presents the results of a typological study of syllables in Brazilian Portuguese. The paper compares the results to the ULSID database made by Rousset [11]. Results are in agreement with the prototypical classifications of syllables. Brazilian Portuguese is similar to other languages of the ULSID database for the consonant-vowel.

1. INTRODUCTION

L'*UCLA Lexical and Syllabic Inventory Database* (ULSID) est constitué d'un ensemble de langues choisies pour la recherche d'universaux au niveau syllabique. Ce projet est développé au GIPSA-lab de l'université de Grenoble III et comprend à ce jour 22 inventaires lexicaux, chacun syllabé manuellement par au moins deux locuteurs natifs et transcrit selon les principes de l'API (2005). Les lexiques contiennent en moyenne 5 800 mots. L'ensemble des langues totalise environ 130 000 unités lexicales et plus de 300 000 syllabes. Ces langues sont diversifiées quant aux origines géographiques et par rapport à la famille linguistique auxquelles elles appartiennent.

2. MATÉRIEL ET MÉTHODE

Un ensemble de mots du portugais brésilien (PB) a été organisé de manière à se conformer au modèle ULSID. 5 000 mots, divisés en syllabes, sélectionnés à partir d'un sous-ensemble du corpus du projet DIRECT (http///www2.lael.pucsp.br/direct) ont été retenus. Ces mots ont été choisis selon les critères utilisés au GIPSA-lab, comme celui de ne retenir dans les inventaires lexicaux que les lemmes, afin de créer une base de données homogène [11]. Ces données ont été organisées de manière à les rendre analysables par un programme spécialement développé pour ce projet. Les mots sélectionnés ont ensuite été soumis à quatre types de transcription ou de traitement : phonétique selon les principes de l'API, en lieu et mode d'articulation et en cohortes (syllabes en C et V). La transcription phonétique a été faite par les auteurs de cette étude puisqu'il n'existait pas de données transcrites des mots. Les transcriptions ont été faites selon les principes mis au point par Cristórfaro-Silva [4] et Collischonn [3] pour le PB. Pour les phonèmes contrastifs, les symboles de l'API ont étés utilisés.

En PB, les phonèmes /s/ et /r/ ont différents allophones en position de coda, à cause de facteurs de variation sociolinguistique. Pour le phonème /s/ qui ne possède qu'un allophone en position de coda, le symbole [s] a été utilisé. Pour le phonème /r/, le symbole [R] a été utilisé. Pour les nasales, trois possibilités d'analyse se présentent : i) il existe une coda nasale; ii) il existe un appendice nasal qui termine la syllabe ; iii) il n'y a pas de coda nasale, mais seulement une voyelle nasale. Pour l'analyse adoptée dans cette étude, considérant l'absence de convergence dans les différentes études portant sur le statut de la nasalité en coda en PB, la troisième solution a été choisie [10]. À propos des diphtongues, Câmara [2] affirme que celles du PB sont descendantes, constituées d'une séquence voyelle +semi-voyelle. Dans ce travail, elles sont transcrites par une séquence VV afin de conserver la structure du timing. Selon Collischonn [3], les vélaires [kw] et [gw] sont des réminiscences du latin que le PB a tendance à éliminer. Cependant, ces séquences sont encore considérées comme des phonèmes uniques dans les descriptions. Elles ont donc été répertoriées comme telles dans la base de données. La classification en lieux et modes d'articulation des consonnes a été faite en suivant les propositions de l'API de 2005. Les voyelles sont décrites en termes de position antérieure, centrale ou postérieure et de degré d'aperture.

Le programme utilisé pour l'analyse des données est *Exploitation de Données Lexicales et Syllabiques*©. Il a été développé au GIPSA-lab, sur une plateforme MatLab spécialement construite pour le traitement statistique des données syllabiques.

3. RESULTATS

Cette recherche suit les mêmes étapes que la démarche utilisée au GIPSA-lab et décrite en détail par [11] pour étudier la typologie et les universaux des syllabes. Ces dernières sont analysées à partir des unités lexicales, des structures syllabiques, et de leurs occurrences et des dépendances qu'elles ont entre elles.

3.1 Unités lexicales

Cette section examine le nombre de syllabes par unité lexicale, la position de la syllabe dans le mot, le nombre de phonèmes par mot et par syllabe ainsi que les contraintes sur la concaténation des syllabes. Pour analyser le nombre de syllabes par unité lexicale, la forme canonique (FC), c'est-à-dire le rapport entre le nombre de syllabes par unité lexicale et le nombre de mots dans un inventaire donné, a été utilisée. Elle permet le classement des langues d'ULSID en 4

catégories, selon le nombre de syllabes contenu en moyenne dans les entrées lexicales (entre 1 et 4 syllabes).

La plupart des langues d'ULSID sont de type dissyllabique, la valeur de FC étant proche de 2. Le PB a une forme canonique de 3.09, et est donc de type trisyllabique. Le tableau suivant montre la distribution des syllabes pour cette langue.

Table 1: Distribution des syllabes par mot en PB.

Nombre de syllabes	Occurrences	%
1	103	2,06
2	1407	28,15
3	1940	38,82
4 ou +	1548	30,97
Total	4998	100,00

Le tableau 1 montre que les items lexicaux de trois syllabes sont majoritaires en PB et que les monosyllabes sont très peu présents, ce qui est attendu dans ce type de langue [11].

Lorsqu'est examinée la relation entre le nombre de phonèmes par syllabe et le nombre de syllabes par unité lexicale, 3 catégories de langues peuvent être établies selon le nombre de phonèmes dans chaque syllabe. La plupart des langues d'ULSID se trouvent autour de 2 et les syllabes de plus de 3 phonèmes sont rares. Le tableau 2 montre que la plupart des syllabes du PB sont constituées de deux phonèmes.

Table 2: Distribution des phonèmes par syllable.

Nombre de phonèmes	Occurrences	%
2	10445	67,47
3	3681	23,78
1	1220	7,88
4	130	0,84
5	5	0,03
Total	15481	100,00

En ce qui concerne la distribution des phonèmes par mot, la figure 1 montre une concentration entre quatre et huit phonèmes par mot pour le PB.

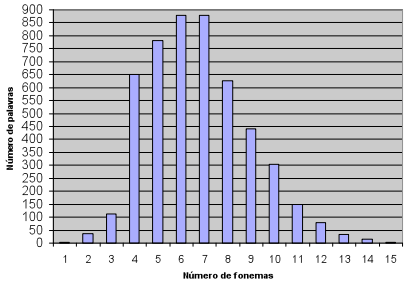


Figure 1: Distribution des phonèmes par mots en PB.

Selon la loi de Menzerath[9], il existe une sorte d'équilibre entre le nombre de phonèmes par syllabe et le nombre de syllabes par unité lexicale. Les données présentées plus haut montrent que le PB se conforment à cette tendance, comme les autres langues d'ULSID : plus le mot est long (apprécié en nombre de syllabes), moins complexes sont les syllabes. Le PB est centré autour de trois syllabes par mot, deux phonèmes par syllabe et de quatre à huit phonèmes par mot (max autour de 6-7 par lemme).

Un autre point important de ce travail est l'analyse des mots en cadres, c'est-à-dire de mots transcrits en cohortes syllabiques C et V. La plupart des cadres de mots dans les langues d'ULSID sont faits de deux syllabes, malgré la part de monosyllabes dans la base de données liés aux langues monosyllabiques (ex. wa) ou majoritairement monosyllabiques (ex. navaho).

Table 3: Distribution des cadres de mots en PB (cadres représentant plus de 2% des lemmes)

Cadre	Occurrences
CV.CV	10,86
CV.CV.CV	9,66
CV.CV.CV.CV	3,16
V.CV.CV	2,84
CVC.CV	2,64
CV.CV.CVV	2,60
CVC.CV.CV	2,38
CVV.CV	2,38
CCV.CV	2,18

Plusieurs cadres de mots du PB sont constitués de trois syllabes. Cependant, il y a deux cadres de mots qui émergent en PB et dans les langues d'ULSID : CV.CV et CV.CV.CV. Les syllabes complexes ne sont pas présentes dans les cadres de mots les plus fréquents, ce qui signifie que le PB favorise à la fois les cadres et syllabes simples, particulièrement les types CV.

L'analyse des langues d'ULSID et du PB montre que des structures sont plus présentes que d'autres et que quelques possibilités ne sont pas du tout utilisées. Ceci signifie qu'il existe probablement des contraintes sur leur choix.

3.2 Structures syllabiques

La première unité analysée dans la structure syllabique est la cohorte. Au travers d'une telle analyse, il est possible de trouver le cadre syllabique le plus fréquent en PB.

ULSID présente 19 cohortes différentes, CV et CVC étant les plus fréquentes, avec 85 % des cas selon Rousset [11], CV seul rendant compte de plus de 50 % des occurrences.

Table 4: Cohortes du PB par fréquence décroissante

Attaque vide	Attaque simple	Attaque complexe
V	CV	CCV
VC	CVV	CCVV
VV	CVC	CCVC
	CVVC	CCVVC
	CVCC	

Dans la table 4, le petit nombre de cohortes est étonnant, étant donné le nombre de syllabes attestées dans le lexique du PB (plus de 15 000, cf. Table 2). Les CV représentent environ 64 % des syllabes, sans aucun doute la cohorte la plus fréquente. Le PB compte plus de syllabes à attaques pleines que le contraire en conformité avec le principe d'attaque maximale ou

MOP de la phonologie (*Maximal Onset Principle*) selon lequel les attaques pleines sont préférées aux attaques vides [12].

En PB, et dans les langues ULSID, les clusters consonantiques sont peu fréquents (présents dans 5,14 % des syllabes en PB). Ce fait, et ceux mentionnés plus haut, conduisent à conclure que la fréquence des structures est inversement associée à leur complexité. Par conséquent, la présence d'une syllabe est plus probable si elle est peu complexe, proche de CV.

Un dernier point à considérer sur la structure des syllabes est l'efficacité syllabique (ES) estimée à partir du rapport entre le nombre de syllabes attestées dans le lexique et le nombre de syllabes différentes.

Rousset [11] a retenu 4 types de langues en rapport avec la notion d'ES, la troisième et la quatrième étant celles qui réutilisent le plus leurs syllabes dans les différents lemmes. Les langues de ces catégories usent de combinaisons de syllabes pour créer des mots et non des syllabes isolées. Au plus il y a de syllabes dans un mot, au plus elles se combinent et par conséquent sont réutilisées.

L'ES du PB est de 17, ce qui permet de mettre cette langue dans la catégorie 4. Le PB réutilise donc beaucoup ses syllabes. Le facteur ES montre que le PB favorise certains types de syllabes plus que d'autres en les réutilisant. Ce fait indique la présence d'unités syllabiques favorisées.

En PB le noyau vocalique le plus fréquent est /a/, suivi de /i/ et /u/. Ces voyelles correspondent aux extrémités du triangle vocalique et sont aussi les plus fréquentes dans les langues du monde [1]. Les voyelles nasales sont moins fréquentes que leurs contreparties orales. Ceci peut être du à leur plus grande complexité articulatoire et acoustique [5].

Les occurrences des consonnes ont été estimées selon leur position dans la syllabe. En attaque, la consonne la plus présente est la coronale sourde /t/ (11.95 % des syllabes), suivie de /k/ et /d/. Ces résultats sont conformes à la typologie des consonnes dans les inventaires phonémiques des langues [1] [8]. Le PB a un inventaire de 23 consonnes, toutes peuvent apparaître en position d'attaque alors que seuls les allophones de /s/ et /R/ peuvent apparaitre en coda. Le PB connait donc une très forte restriction de son inventaire consonantique en position de coda.

3.3 Occurrences et dépendances

Cette section présente une analyse des occurrences et dépendances entre les consonnes et les voyelles qui remplissent les positions d'attaque et de noyau dans les syllabes CV. Dans la lignée des travaux de [7] et dans le but d'établir l'existence ou non de liens forts entre attaque et noyau des syllabes CV, le projet ULSID utilise le rapport entre les différents types de syllabes attestés et les différents types de syllabes possibles. Si ce rapport est égal à 1, alors la langue utilise toutes ses possibilités. Au plus bas est le rapport, moins la langue utilise ses possibilités syllabiques. Dans les langues d'ULSID le rapport moyen des différentes langues est de 0.5 ce qui signifie qu'en général, elles utilisent autour de 50 % de l'ensemble des possibilités de syllabes CV. En PB le rapport est de 0.67, ce qui signifie que cette langue utilise autour de 67% des syllabes CV possibles.

Si l'efficacité syllabique atteste que le PB réutilise beaucoup ses syllabes dans la construction des lemmes (section 3.2) et qu'il est à présent attesté qu'il n'utilise pas toutes ses possibilités syllabiques CV, quelles sont les restrictions qui s'opèrent et quelles sont les syllabes favorisées ?

La méthode utilisée par [6] et [11] pour identifier le type de syllabe le plus fréquent et la relation entre ses constituants consiste à créer une matrice de leurs combinaisons, en groupant voyelles et consonnes selon leur lieu d'articulation, et ensuite à calculer le rapport des combinaisons attestées et celles attendues. Les rapports en dessous de 1 signifient que les combinaisons ne sont pas favorisées; les rapports égaux à 1 signifient qu'il n'est pas possible de faire une prédiction et les rapports au dessus de 1 indiquent que la combinaison est favorisée.

Table 5: Rapport entre syllabes CV attestées et attendues selon les différentes combinaisons de constituants. Les attaques sont en colonnes. Les cellules en grisé correspondent aux cooccurrences favorisées trouvées par [6] et [11]. En gras les coocurrences favorisées en PB.

Occurrences	Bi	Co	Ve	Autres
An	1,00	**1,10**	0,33	1,29
Ce	1,08	0,94	**1,27**	0,87
Po	0,91	0,94	**1,51**	0,80

Les ratios calculés montrent que les combinaisons entre consonnes coronales (Co) et voyelles antérieures (An) d'une part, et entre consonnes vélaires (Ve) et voyelles postérieures (Po) d'autre part, sont favorisées. Les voyelles centrales (Ce) après consonne labiale (Bi) sont moins favorisées, une attaque vélaire étant plus fréquemment trouvée devant noyau central. Seul ce dernier résultat ne confirme pas les résultats obtenus par [11] et [6].

Pour les deux premières combinaisons, il est possible de conclure que, dans les syllabes CV en PB, la combinaison entre attaque et noyau est favorisée quand les deux éléments sont réalisés sans déplacement antéro-postérieur de la langue entre le début et la fin de la syllabe, c'est-à-dire lorsque la syllabe est produite par un seul geste mandibulaire, confirmant ainsi un des éléments de la théorie Frame/Content [6] [7]. Cependant, la plus forte cooccurrence des consonnes vélaires et des voyelles centrales ne fait pas partie des

patrons prédits par cette théorie et aucune explication ne semble évidente pour le moment.

4. CONCLUSIONS ET PERSPECTIVES

Cet article présente une première tentative pour analyser les structures du PB, en vue de créer une typologie des syllabes dans cette langue. Le principal objectif était, en insérant le PB dans ULSID, de rendre les structures syllabiques de la langue analysable dans le contexte plus large de la recherche d'universaux phonologiques.

La méthodologie a consisté à créer une base de données lexicale en PB, de diviser chaque unité lexicale en syllabes, et de les traiter selon les procédures et critères du GIPSA-lab retenus pour le développement d'ULSID [11].

Les principaux résultats concernent la taille des unités lexicales, la taille et le contenu de la structure syllabique, les occurrences et les dépendances parmi les constituants à l'intérieur des syllabes CV. Le PB se conforme aux classifications les plus prototypiques, propose et ajoute aussi quelques éléments sur la forme des universaux d'organisation de la syllabe, tels que la sélection des segments sur une base articulatoire. Les résultats sur l'analyse des unités lexicales montrent qu'il y a un équilibre entre le nombre de syllabes dans un mot et le nombre de constituants de ses syllabes. Les unités syllabiques les plus communes ont trois syllabes, mais les syllabes constituant de telles unités varient. Le cadre de mot le plus courant a deux syllabes de type CV. Par rapport à la structure syllabique, la plupart des cadres du PB se conforment aux cadres relevés dans ULSID, en donnant des preuves d'une organisation plus générale à travers les langues. CV est le type syllabique préféré, ce qui confirme son caractère universel. Les voyelles les plus fréquentes des syllabes sont /a/, /i/, et /u/, ce qui correspond aux voyelles extrêmes du triangle vocalique, les consonnes sont /t/, /k/ et /d/. Ces deux résultats sont conformes aux analyses typologiques de [1] [8]. Les derniers résultats concernent les co-occurrences entre constituants dans les syllabes de type CV. Il se dégage de nos analyses certaines préférences très nettes pour les combinaisons consonne-voyelle suivantes : coronale-antérieure et vélaire-postérieure. La tendance à combiner consonne labiale avec voyelle centrale est moins nette. Les mêmes résultats, également trouvés dans les autres langues d'ULSID, sont des patrons syllabiques prédits par la théorie Frame/content [6] [7]. Par contre ce que ne prédit pas cette théorie, c'est la forte proportion de cooccurrence entre consonne vélaire et voyelle centrale relevée en PB. Néanmoins, ces résultats ajoute des éléments au fait qu'il existe, à propos des structures syllabiques, des principes généraux qui gouvernent les systèmes sonores des langues du monde. Ces principes ne sont probablement pas arbitraires, mais dus à des contraintes de production, puisque les syllabes préférées partagent, comme le souligne [11] les éléments de régions articulatoires identiques.

5. REMERCIEMENTS

Cette recherche a été soutenue par la Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

6. REFERENCES

[1] L. Boë, N. Vallée, P. Badin, J.-L. Schwartz, C. Abry. Tendencies in Phonological Structures: the Influence of Substance on Form. In: *Special Issue on Current Trends in Phonetics and Phonology*. Les Cahiers de l'ICP, Bulletin de la Communication Parlée Volume 5 : 35-55, 2000.

[2] J. M. Câmara Jr. *Estrutura da Língua Portuguesa.* Petrópolis, Ed. Vozes, 1979.

[3] G. Collischonn. *A Sílaba em Português.* In: Bisol, L (org.) Introdução a Estudos de Fonologia do Português Brasileiro. Porto Alegre: Edipucrs, 2005.

[4] T. Cristófaro-Silva. *Fonética e Fonologia do Português.* 6ªed. Contexto, São Paulo, 2002.

[5] B. Lindblom, I. Maddieson. Phonetic universals in consonant systems. In L.H. Hyman & Li C.N. (eds.), *Language, Speech and Mind*: 62-78. London and New-York: Routledge, 1988.

[6] P. MacNeilage, B. Davis. On the Origin of Internal Structure of Word Forms. *Sciences* 288: 527-531, 2000.

[7] P. MacNeilage, B. Davis. Motor mechanisms in speech ontogeny: phylogenetic, neurobiological and linguistic implications. *Current Opinion in Neurobiology* 11: 696-700, 2001

[8] I. Maddieson. *Patterns of Sounds.* Cambridge University Press, 1984.

[9] P. Menzerath. *Die Architektonik des deutsch Wortschatzes*. Bonn: Dummler, 1954.

[10] B. Raposo de Medeiros, M. D'Imperio, R. Espesser. La voyelle nasale en Portugais Brésilien et son appendice nasal : étude acoustique et aérodynamique. Journées d'Études sur la Parole, Avignon, 2008

[11] I. Rousset. *Structures Syllabiques e Lexicales des Langues du Monde.* Thèse - Université Grenoble III – Stendhal, Grenoble, 2004.

[12] D. Steriade. *Greek Prosodies and the Nature of Syllabification*. MIT: PhD Thesis. 1982.

Finalisation des phrases en lecture et en parole spontanée: le cas du Portugais brésilien

Waldemar Ferreira Netto
Fernanda Consoni
Daniel Oliveira Peres
Universidade de São Paulo – São Paulo – Brasil
walfnetto@usp.br, consoni.fernanda@usp.br, danielperes@usp.br
http://sites.google.com/site/exprosodia/

ABSTRACT

The aim of this work is to verify if the prosodic finalization in reading and spontaneous speech sentences on Brazilian Portuguese, can be interpreteted taking in consideration the principles of musical analyses based on western music could be observed in the prosody finalization. The ExProsodia, the routine for automatic analysis elaborated by Ferreira Netto [5] was the instrument to verify if a prosodic sequence could be interpreted as a set of harmonic sounds. These harmonic sounds have their trajectory leaving a medial dominant tone and arriving at a fundamental finishing tone. We observed that there is a tendency in the intonation to finish the melodic curve keeping a perfect cadence major, and it is clearer in the reading sentences.

Keywords: prosody; reading; spontaneous speech; intonation; music.

1. INTRODUCTION

Le but de cette étude est analyse la prosodie selon les principes de base de l'analyse musicale, en considérant l'hypothèse que les idiosyncrasies culturelles de la tradition musicale d'un groupe social se retrouvent dans la prosodie de ce groupe [6], [12]. Une des caractéristiques de l'exécution musicale est la possibilité de définir une suite de tons dont l'harmonie est établie à partir d'un ton fondamental [13]. De ce point de vue, une séquence mélodique est constituée à partir des différentes trajectoires qu'impliquent des départs et des rapprochements du ton fondamental, ce qui crée toujours l'expectative que le point final de cette trajectoire est le propre ton fondamental [7].

Si on tient compte de cette caractéristique des sons pour l'élaboration de la mélodie, on comprend que le même principe pourrait s'appliquer à la parole par rapport à la variation de la fréquence qui compose l'intonation.

L'idée principale selon laquelle une collection de sons qui sont en harmonie est définie à partir d'un ton fondamental, ne s'applique pas à la parole de la même façon, bien que ce soit le même phénomène. L'élaboration et la perception des fréquences dans la parole sont soumises à des modèles différents selon les poursuivis. T'Hart [14], par exemple, a vérifié que la perception des variations de fréquence produites dans la parole ont été de 1,5 tons ou 3 demi-tons (dt). Il s'agit d'un modèle différent de celui des interprétations mélodiques, car dans ce modèle on peut percevoir des variations minimes jusqu'à 0,1% du son qui est produit [10], [8]. En augmentant la marge d'erreur pour l'identification des fréquences dans la parole, nous avons considéré que celles-ci seraient comprises dans les bandes de fréquence et non dans celui d'un ton précis comme, par conséquent, dans la musique. Dans le procédé mélodique standard de la musicalité populaire occidentale, une cadence parfaite vient d'un ton moyen dominant, et se termine par un ton fondamental, environ 1,5 fois inferieur que le ton dominant moyen. La définition d'un ensemble harmonique qui forme la séquence prosodique devra, à son tour, être définie sur la base de ce qui pourrait être interprété comme la fin de la mélodie qui est composée par ces sons harmoniques. Pour cela, on peut prendre le principe de base proposé par Schoenberg [13] (op. cit, p., 162) qu'un son est dépendant d'un son qui est une quinte en dessous de celui-ci. Par conséquent, nous assisterons à la création d'un son fondamental, unité de finition, et un dominant, formant une cadence majeure parfaite. Ainsi, nous pensons qu'une séquence prosodique pourrait être interprétée comme un ensemble harmonique qui établit une unité définie par son achèvement sur le fondamental, une quinte en dessous des quelques sons prédominants dans la séquence formatrice.

2. MÉTHODOLOGIE

Pour tester ces principes, nous avons analysé 150 phrases produites par 40 sujets (masculins) 20 étions des présentateurs professionnels du journalisme-radio dont la F0 maximum= 187 Hz et minimum = 115 Hz (moyenne = 143Hz, l'écart type =24 Hz). Les autres 20 locuteurs produisaient de la parole spontanée dans des entretiens, leur F0 variait entre 142 Hz et 109 Hz, avec une moyenne de123 Hz et un écart type de 14Hz). En travaillant avec de la parole en lecture et dans les déclarations spontanées, la fréquence moyenne quell'on a obtenu en lecture est ce à quoi on pourrait s'attendre pour des voix d'hommes. Les valeurs sont comprises entre 111,5 et 132,6 Hz, comme présenté par Russo et Behlau [10] et celui de Andrade [1], entre 110 et 146,7 Hz. Toutes les phrases ont été recueillies à partir d'internet et

sélectionnées après vérification de l'intégrité du son, c'est à dire, pour vérifier qu'il n'y a pas de fréquences filtrées en dessous de 8000 Hz, et surtout au-dessous 100Hz. Les phrases analysées sont assertives et elles ont été segmentées de manière subjective en considérant les aspectes syntaxiques et sémantiques.

L'analyse a été faite par la routine de l'analyse automatique Exprosodia[i]. La routine ExProsodia segmente un fichier sonore qui a été converti en texte par l'utile Speech Filing System. La segmentation c'est fait á partir des seuils acoustiques que nous avons abordés précédemment, à savoir :

Seuil inférieur de fréquence : 50 Hz ;

Seuil supérieur de fréquence : 350 Hz ;

Seuil d'intensité : 2000 RMS

L'évaluation automatique des seuils de fréquence et d'intensité établit la première étape pour trouver des points qui peuvent être segmentés dans la ligne sonore de la parole. Un segment qui a ces caractéristiques sera catégorisé selon la variation de sa fréquence par rapport à une échelle de 5 bandes de fréquence, chacune comprenant 3 demi tons (dt), basés sur la fréquence moyenne de la parole, qui est le centre du ton moyen 3. La catégorisation de la fréquence sur une échelle de 5 niveaux est basé sur les études de t'Hart [14] qui ont montré que, pour la perception de la parole, seulement les variations supérieures à 3 demi tons sont validées [4]. Ainsi, la fréquences peut être organisée em 5 bandes et peut être exprimé en 5 groupes de notes musicales, par example : do do# re, re# mi fa, fa# sol sol#, la la # si, do do# re.

Un ensemble de mesures ont été faites pour verifier si le variation de 3 demi tons etait importante pour les sujets brésiliens. Consoni et collegues [2] ont verifié que la variation de 3 demi tons ascendants et de 4 demi tons descendants est bien consideré pour les sujets testés. Ferreira Netto et Consoni [3] ont observé que la variation de la fréquence dans les textes lus est plus aigüe que cellle de la parole spontanée.

En général, fréquence et intensité sont indépendants l'un de l'autre. La durée n'est pas obtenue directement par rapport à la fréquence ou à l' intensité. La durée est établie par rapport à u mantien de quelqu'un des 5 tons patrons pendant un certain temps. C'est-à-dire, une fréquence entre 50 et 350 Hz d'intensité > 2000 RMS et catégorisés dans une période prédéfinie. Ainsi, la durée - 20 ms, seuil inférieur, et 150 ms, seuil supérieur - est établie directement par rapport au mantien de 5 tons de base pour un locuteur.

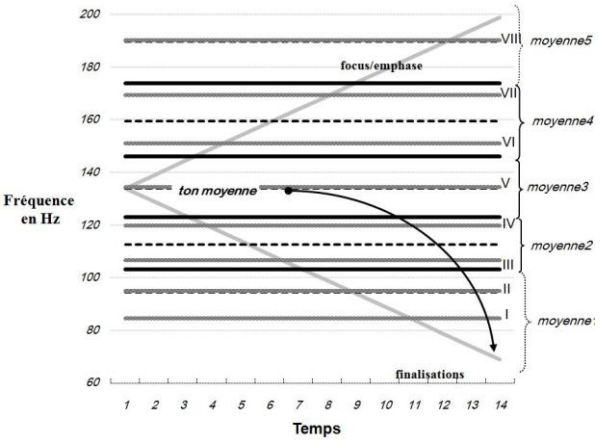


Figure 1: graphique qui represente une échelle de 5 tons. Les lignes pointillées montrent les valeurs moyennes pour chaque bande de fréquence. Les lignes continues plus sombres montrent les limites de chaque bande, comprises dans l'intervalle donné à droite.. Les lignes continues plus légères montrent les points précis de chaque degré de l'échelle diatonique occidentale, qui est définie dans Pierce [9]. Les degrés de l'échelle diatonique occidentale sont marqués en chiffres romains à droite et les degrés de l'échelle des 5 tons sont étiquetés à droite. Un ligne courbe dans le centre montre la tendance de finalisation des phrases et les lignes diagonales montrent le focus et l emphase pour la finalisation.

Pour évaluer la relation entre le ton moyen et le ton final, nous avons analysé séparément les cinq phrases de chaque locuteur, en convertissant des tons moyens finaux de chaque phrase en notes musicales correspondantes. Nous avons utilisé les valeurs données par Pierce [9], en établissant les valeurs moyennes de chaque intervalle de demi-tons, comme la limite pour la conversion des fréquences en notes musicales.

3. RÉSULTATS

La première analyse a examiné si les valeurs moyennes et finales des phrases ont été stables pour chaque sujet. Pour cela, nous avons compté la quantité de notes comprises entre la plus grave et la plus aigüe dans chaque phrase. Les différences entre lecture et parole spontanée ont été considérées. Dans l'ensemble, la moyenne de variation pour le ton moyen des sujets a été de 2,8 dt pour la lecture et de 4,2 dt pour la parole spontanée, ce qui représente une différence significative indiquée par l'analyse de variance ANOVA ($Fo=17,8>Fe=4,1$ e $P<0,001$). La moyenne de variation pour le ton final des mêmes sujets a été de 5 dt pour la lecture et de 7,4 dt pour la parole spontanée, ce qui montre une différence significative indiquée par l'analyse de variance ANOVA ($Fo=11>Fe=4,1$ e $P<0,001$). Dans les deux évaluations on a remarqué une différence significative ce qui indique une plus grande stabilité de la lecture par rapport à la parole spontanée.

Tableau 1- données relatives aux variations maximums de demi-tons dans les 5 phrases de chaque sujet.

Ton Moyen		Ton final	
Lecture	Pspontanée	Lecture	Pspontanée
3	3	4	3
3	5	5	8
2	4	4	10
1	6	3	5
2	6	5	10
3	3	4	7
2	4	7	9
3	3	2	11
2	5	3	9
3	3	6	10
2	4	4	2
6	3	5	9
2	5	4	12
2	3	7	5
3	5	5	6
3	3	4	8
5	4	7	7
2	5	8	6
3	5	9	6
3	5	4	6

À partir de cette analyse, nous avons vérifié si les valeurs correspondraient à une cadence majeure parfaite, à savoir, la relation entre un ton dominant et un ton fondamental. Par conséquent, pour chaque sujet, nous avons compté les occurrences de phrases et la relation entre dominant (ton moyen) et fondamental (ton final). Bien que la comparaison entre les données brutes ne différait pas de façon significative, elle a été trouvé dans la corrélation établie entre les moyennes mobiles d'occurrences de phrases en relation D>T (dominat>tonique) et le fait si cela se partage en lecture (catégorisé comme 0) ou en parole spontanée (catégorisé comme 1). L'analyse par le test de Spearman a démontré une corrélation négative : $R= -0,63$ et $p< 0,001$.

Tableau 2: La colonne de gauche montre les occurrences de la séquence D> T (dominant > tonique), avec une variation de ± 1 dt dans la série de 5 phrases de chaque sujet. La deuxième colonne indique la nature des données, avec 1 pour la lecture et 0 pour la parole spontanée. Les deux colonnes de droite montrent la moyenne mobile de quatre points pour l'occurrence de D> T et la nature des données.

Donnees D>T	Bruts Catégorie	Moyenne D>T	Mobile Catégorie
5	0		
5	1		
4	0		
4	0	4,5	0,25
4	0	4,25	0,25
4	0	4	0
4	0	4	0
4	0	4	0
3	0	3,75	0
3	0	3,5	0
3	0	3,25	0
3	0	3	0
3	1	3	0,25
3	1	3	0,5
3	1	3	0,75
3	1	3	1
2	0	2,75	0,75
2	0	2,5	0,5
2	0	2,25	0,25
2	0	2	0
2	0	2	0
2	0	2	0
2	1	2	0,25
2	1	2	0,5
2	1	2	0,75
2	1	2	1
2	1	2	1
1	0	1,75	0,75
1	0	1,5	0,5
1	1	1,25	0,5

4. CONCLUSIONS

Ces résultats indiquent qu'il existe une tendance vers la finalisation d'une cadence majeur parfaite de dominante-tonique, et le fait que cette tendance est plus marquée dans les finalisations des phrases produites dans les lectures que celles produites par la parole spontanée. À partir d'une analyse fondée sur les principes de l'analyse musicale, nous suggérons que la tendance pour la finalisation basée dans une cadence majeure parfaite de dominante-tonique est une indication qu'il existe une relation entre les faits communs de la musique ocidentalle et les variations de l'intonation de la phrase en Portugais du Brésil.

BIBLIOGRAPHIE

[1] Andrade, Luciana Mara de Oliveira (2003). *Determinação dos limiares de normalidade dos parâmetros acústicos da voz*. Dissertation de Master, Bioengenharia/USP, São Paulo

[2] Consoni, Fernanda; Ferreira Netto, Waldemar; Peres, Daniel Oliveira; Lassak, Amanda de Lima; Rosa, Renata Cezar de Moraes. Sensitivity to f0 variation in Brazilian Portuguese. *Poznan Linguistic Meeting, Poznan, Polônia, 2009*.

[3] Ferreira Netto, Waldemar; Consoni, Fernanda Estratégias prosódicas da leitura em voz alta e da fala espontânea. *ALFA: Revista de Lingüística*, volume 52, número 2, 2008.

[4] Ferreira Netto, W. *Decomposição da entoação frasal em componentes estruturadoras e semântico-funcionais.* X Congresso Nacional de Fonética e Fonologia/IV Congresso Internacional de Fonética e Fonologia, Niterói, RJ, 2008.

[5] Ferreira Netto, Waldemar, EXPROSODIA. *Revista da Propriedade Industrial – RPI*, n. 2038, pág. 167, 2008.

[6] Glaser, S. The missing link: Conections between musical and linguistic prosody. *Contemporary Music Review*, n. 19, v. 3, p. 131-154, 2000.

[7] Longacre, Robert E.; Chenoweth, Vida. Discurse as music. *Word*, n. 37, v. 1-2. p. 125-134, 1986.

[8] Menezes, Flo. *A acústica musical em palavras e sons.* São Paulo: Ateliê Editorial/Fapesp, 2003.

[9] Pierce, John R. *The science of musical sounds*. New York: Scientific American Library, 1987.

[10] Roederer, Juan G. *Introdução à física e à psicofísica da música*. Trad. Alberto Luis da Cunha. São Paulo: Edusp, 2002. Edição original de 1975.

[11] Russo, Ieda; Behlau, Mara. *Percepção da fala: análise acústica do português brasileiro.* São Paulo: Lovise, 1993.

[12] Schellenberg, Glenn E.; Trehub, Sandra E. Culture-general and culture-specific factors in the discrimination of melodies. *Journal of Experimental Child Psychology*, n. 74, p. 107–127, 1999.

[13] Schoenberg, Arnold. *Harmonia*. São Paulo: Editora Unesp, 2001. Edição original de 1921.

[14] T'hart, J. Differential sensitivity to pitch distance, particularly in speech. *Journal of Acoustical Society of the America*, n. 69, v. 3, p. 811-821, 1981.

[15] T'hart, J.; Collier, R.; Cohen, A. *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press, 1990.

[i] Les exemples des données obtenues par la routine ExProsodia. Pour faire l'analyse en cinq demin tons, la routine élimine les variations de la déclinassion qui se maintient droite. La première image est la représentation de la curve F0 en MIDI, et la deuxième, en cinq tons. Les images sont faites à partir de la l'analyse de la phrase: "os alunos fizeram prova de literatura." Les phrases du corpus sont disponible sur le website: http://sites.google.com/site/exprosodia/

Figure 2

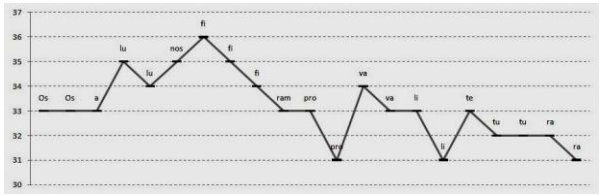
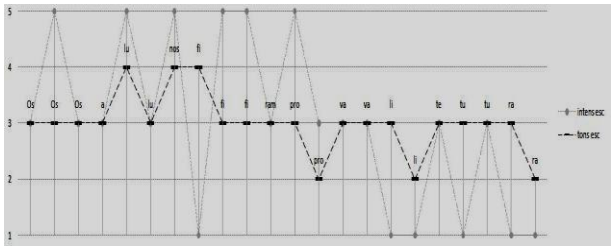


Figure 3



Stratégie d'apprentissage actif pour l'adaptation de modèles de compréhension dans un Système de Dialogue Oral déployé

Pierre Gotab [1], *Frederic Bechet* [2], *Geraldine Damnati* [3]

[1] LIA, Université d'Avignon, 339 chemin des Meinajaries, 84911 Avignon Cedex 9
pierre.gotab@univ-avignon.fr

[2] Aix Marseille Université - LIF-CNRS, Case 901, 163 avenue de Luminy, F-13288 Marseille Cedex 9
frederic.bechet@univ-avignon.fr

[3] France Telecom R&D - Orange Labs, 2 av. Pierre Marzin 22307 Lannion, France
geraldine.damnati@orange-ftgroup.com

ABSTRACT

Active learning can be used for the maintenance of a deployed Spoken Dialog System (SDS). This maintenance process is crucial as a deployed SDS evolves quickly. Knowledge-based approaches are often preferred as system designers can easily modify the SLU models to cover examples not yet seen. However as new examples are added to the annotated corpus, corpus-based methods can then be applied, replacing or in addition to the initial knowledge-based models. This paper describes an Active learning scheme used for automatically updating the SLU models of a deployed SDS.

Keywords: Active learning, deployed spoken dialog system, machine-learning, rules-based models, corpus-based models.

1. Introduction

La multiplication des Systèmes de Dialogue Oral (SDO) à grande échelle apporte de nouvelles opportunités pour les chercheurs du domaine de la compréhension de la parole. En effet, de tels systèmes peuvent être vus comme des sources illimitées de corpus de dialogue qui peuvent être utilisées pour améliorer chaque étape du système, depuis les modèles acoustiques jusqu'au gestionnaire de dialogue en améliorant la couverture des modèles ainsi qu'en les adaptant. En effet l'adaptation est essentielle car les SDO déployés ne sont pas statiques : des services sont ajoutés, certains disparaissent, d'autres peuvent être saisonniers, et des stratégies commerciales peuvent rapidement influencer le comportement des utilisateurs. Il est donc important d'être capable d'adapter les modèles à ces changements. L'apprentissage actif peut être utilisé à cette fin en sélectionnant automatiquement les exemples non couverts par les modèles.

L'apprentissage actif a été utilisé pour entraîner différents types de modèles utilisés dans un SDO, depuis les modèles acoustiques [5] et linguistiques [6] aux modèles de compréhension [9]. Dans [8], une étude intéressante montre que le fameux principe *"there is no data like more data"* peut être violé en sélectionnant les données selon leur *complétude*. Dans tous ces travaux antérieurs, deux objectifs sont visés : réduire l'effort d'annotation manuelle sans dégrader la performance des modèles, et trouver l'ensemble *optimal* de données à annoter parmi un large corpus de données non annotées. L'algorithme général de l'apprentissage actif consiste à apprendre les modèles initiaux sur un *bootstrap* annoté manuellement, puis à confronter ces modèles à un large ensemble de données non annotées afin d'en sélectionner un sous ensemble à l'aide d'un critère d'apprentissage actif arbitraire. Ce sous ensemble sera annoté manuellement et ajouté au *bootstrap*. Ce processus est répété jusqu'à ce qu'un critère d'optimalité soit atteint, que ce soit une mesure de performance sur un corpus de développement ou une borne maximum du nombre d'exemples à transcrire. Le but du critère d'apprentissage actif est de sélectionner les exemples qui vont maximiser la qualité des modèles tout en maintenant un effort d'annotation aussi faible que possible.

Les modèles à base de connaissances (grammaires manuelles ou règles d'inférence) sont souvent préférés car leurs concepteurs peuvent directement modifier les modèles afin de prendre en compte une modification du service avant même qu'il n'y ait de données reflétant cette modification. Cependant, à mesure que de nouvelles traces de dialogue sont collectées et ajoutées au corpus annoté, des méthodes à base de corpus peuvent être appliquées, en remplacement ou en addition aux modèles manuels initiaux. Un tel cycle de développement est décrit dans [7].

Les expériences rapportées dans cet article suivent les mêmes principes en montrant comment un critère d'apprentissage actif peut améliorer l'apprentissage de modèles de compréhension en remplacement de critères purement temporels.

Cet article est structuré ainsi : la section 2 décrit le système à base de règles déployé par France Telecom ; puis la section 3 présente la stratégie d'apprentissage actif proposée dans cette étude ; finalement, la section 4 rapporte les résultats obtenus sur un important corpus composé de dialogues réels obtenus au travers du SDO déployé par France Telecom, et montre le gain obtenu grâce à notre stratégie.

2. Modèles de compréhension de la parole

2.1. Le système à base de règles

Cette étude porte sur le service vocal "Le 3000" de France Telecom (FT3000). C'est le premier service vocal déployé à France Telecom à exploiter les technologies du traitement de la langue naturelle. Il a été déployé et ouvert au public en octobre 2005. Il permet

de souscrire à une trentaine de services et d'accéder à la gestion de sa ligne téléphonique.

Le système de reconnaissance automatique de parole (RAP) est basé sur un modèle de langage n-gram, puis l'analyse sémantique est composée de trois niveaux :

1. Le premier niveau transforme le treillis de mots de la RAP en treillis de concepts par le moyen d'un transducteur FSM (Finite State Machine) contenant toute la grammaire représentant les concepts de FT3000.
2. Le second niveau applique un ensemble d'environ 2600 règles logiques sur les concepts pour les transformer en interprétations structurées de la forme prédicat/argument. Les règles et leur ordre de priorité ont été écrits manuellement par les concepteurs du système.
3. Le troisième niveau sélectionne l'interprétation la plus probable au regard de l'état courant du dialogue. La sélection tient compte des scores donnés par les deux phases précédentes (score de RAP + priorité des règles).

Le modèle sémantique comporte plus de 400 concepts différents (entités nommées, commandes de dialogue, mots clés), 56 prédicats et 138 arguments.

Depuis son déploiement en 2005, des traces de dialogue ont été collectées, transcrites et annotées. Une telle source de données peut être utilisée par des méthodes à base de corpus. Par exemple, tous les chemins des transducteurs mot/concept peuvent être pondérés grâce à un taggueur HMM entrainé sur ce corpus. Nous avons proposé dans [2, 3] différentes stratégies mélangeant une approche statistique modélisant la *plausibilité* d'une interprétation avec l'approche à base de règles du système déployé modélisant l'*acceptabilité* d'une interprétation. Dans cet article nous proposons de construire à l'aide de l'apprentissage actif un modèle entièrement statistique imitant le modèle à base de règles actuel.

2.2. Utiliser des classifieurs pour prédire les structures prédicat/argument

Le nombre de combinaisons prédicat/argument est trop grand pour entraîner directement des classifieurs statistiques sur ces interprétations structurées. Nous avons alors séparé ce problème : prédire séparément les prédicats et les arguments, puis fusionner les deux prédictions pour obtenir l'interprétation structurée.

Pour un exemple donné, chaque tâche produit toutes les classes (prédicats ou arguments) probables associées à leur score de confiance. Connaissant la liste des combinaisons prédicat/argument valides, nous sélectionnons la combinaison qui maximise le produit des scores de confiance des deux prédictions.

Nous utilisons trois ensembles de paramètres pour entraîner les classifieurs :
– la sortie du système de RAP,
– les concepts obtenus par un taggueur HMM sur la sortie de RAP,
– le contexte de dialogue représenté par le numéro du tour et une étiquette caractérisant chaque état de dialogue.

3. Stratégie d'apprentissage actif

La stratégie d'apprentissage actif proposée dans cet article est tirée de [1]. Le critère utilisé pour sélectionner les exemples est basé sur le désaccord entre différents classifieurs. A cette fin nous utilisons deux classifieurs : *Icsiboost*[1] utilisant l'algorithme *AdaBoost* et *Liblinear* basé sur de la régression logistique.

Le critère d'apprentissage actif choisit d'abord les exemples sur lesquels les prédictions des deux classifieurs diffèrent, puis ceux obtenant le plus mauvais score de confiance.

– Soit un ensemble d'exemples annotés L, un ensemble non annoté U, une partition de test T et deux classifieurs C_1 et C_2.
– Un exemple $e \in L$ est un couple $e = (x, l)$ où x est le vecteur dont chaque dimension représente un paramètre, et l est la classe de cet exemple.
– Un exemple $e = (x, l)$ traité par un classifieur C produit $C(e) = (l', s)$ où l' est la classe prédite avec un score de confiance $s \in [0..1]$.
– Nous considérons que nous avons un Oracle, noté $O(e) = l$, qui correspond à l'annotateur humain, et qui peut fournir la classe correcte l d'un exemple e.
– La quantité de données à annoter manuellement à chaque itération peut être choisie arbitrairement, en fonction des ressources humaines disponibles par exemple. En pratique, la classe donnée par l'Oracle est obtenue en appliquant les règles manuelles sur la transcription manuelle. Dans les expériences présentées dans la section 4, étant donné que nous simulons un tel processus puisque en réalité toutes nos données sont annotées, nous avons arbitrairement choisi la quantité δ_i de corpus à ajouter à l'ensemble L à chaque itération i.
– Considérant que nous avons N itérations, nous avons N tailles de partition : $d_1, d_2, ..., d_N$ avec $d_i = \sum_{k=1}^{i} \delta_k$ et $d_N = |U|$.

Algorithme du processus de sélection des exemples :

- Entraîner C_1 et C_2 sur L
- Évaluer C_1 et C_2 sur T
- Pour k allant de 1 à N
  - Définir deux séquences de tuples $U^a = U^d = \emptyset$.
  - C_1 et C_2 classifient chaque $e = (x, l) \in U$:
    $C_1(e) = (l_1, s_1)$ et $C_2(e) = (l_2, s_2)$
    Remplit les séquences U^a et U^d :
    - Si $l_1 = l_2$ alors $U^a := U^a \cup \{(x, s_1 \times s_2)\}$
    - Sinon $U^d := U^d \cup \{(x, s_1 \times s_2)\}$
  - Trier U^a et U^d par produit $(s_1 \times s_2)$ ascendant
  - Ajuste la taille de U^d :
    - Si $|U^d| < d_k$ alors $U^d := U^d \cup (U^a_n)_{1 \leq n \leq d_k - |U^d|}$
    - Sinon si $|U^d| > d_k$ alors $U^d := (U^d_n)_{n \leq d_k}$
  - Pour chaque couple $(e, s) \in U^d$, $e = (x, l)$
    - $L := L \cup \{(x, O(x))\}$
    - $U := U \setminus \{e\}$
  - Entraîner C_1 et C_2 sur L
  - Évaluer C_1 et C_2 sur T

La stratégie d'apprentissage actif présentée ici peut être directement appliquée à un SDO déployé si le

[1] icsiboost - http ://code.google.com/p/icsiboost - une implémentation open-source de BoosTexter

module de compréhension est basé sur une approche statistique, comme celui présenté dans la section 2. Dans ce cas, le corpus d'apprentissage L est régulièrement augmenté avec les nouveaux exemples sélectionnés par l'apprentissage actif parmi l'ensemble des traces de dialogue collectées.

4. Expériences

Les expériences ont été menées sur le corpus collecté par le service FT3000 présenté dans la section 2 et composé de deux parties : un corpus d'apprentissage de 16600 dialogues collectés sur une période de 10 mois consécutifs par une sélection aléatoire de conversations, et un corpus de test de 2460 dialogues collectés pendant 10 jours consécutifs (postérieurs au corpus d'apprentissage). Les deux corpus sont transcrits manuellement mais seul le corpus de test est annoté manuellement avec les structures sémantiques prédicat/argument. Les annotations sémantiques du corpus d'apprentissage ont été obtenues par application du module de compréhension présenté dans la section 2.1. Les modèles de RAP ont été entraînés sur le corpus d'apprentissage et utilisés pour transcrire le corpus de test, avec un taux d'erreur mot de 40,2%. Toutes les expériences suivantes s'appuient sur la transcription automatique du corpus de test.

Dans les expériences suivantes, les paramètres d'entrée du classifieur *Icsiboost* sont les 2-grams de la sortie de la RAP et de la sortie du taggueur HMM, et l'état du dialogue. *Liblinear* utilise le sac de mots de la sortie de la RAP et du taggueur ainsi que l'état du dialogue.

Tab. 1: Description du corpus

Corpus	Appr.	Test
# dialogues	16600	2460
# tours de dialogue	50000	5370
# mots	81k	13k
# prédicats uniques	43	46
# structures prédicat/argument uniques	369	188
Word Error Rate (WER)	-	40.2%

De nombreux messages peuvent être considérés comme *vides* du point de vue du SDO :ce sont les messages ne contenant pas de parole ou bien de la parole hors domaine. C'est ce qu'on appelle le rejet. Ainsi le système peut faire trois types d'erreur : délétion (quand il rejette le message à tort), insertion (quand il affecte une structure sémantique à un message *vide*), et substitution. L'*Interpretation Error Rate* (IER) est défini comme la somme de ces trois types d'erreur divisée par la somme des interprétations valides (non *vides*). L'IER est suppérieur au classique SER (Sentence Error Rate) qui est relatif au nombre total de messages sans prendre compte leur validité sémantique. L'IER est considéré comme plus représentatif des performances d'un système comme ressenties par un utilisateur.

La table 2 compare les résultats FT3000 à base de règle sur le corpus de test avec ceux d'*Icsiboost* (I.) et de *Liblinear* (L.), entraîné sur l'intégralité du corpus d'apprentissage . Le premier système surpasse largement les classifieurs, ce qui s'explique en premier lieu par la quantité de connaissances utilisée (2600 règles) qui couvre bien plus que les exemples apparaissant dans le corpus d'apprentissage ; puis par la difficulté qu'ont les classifieurs à modéliser la classe *rejet*, résultant en un taux d'insertions élevé - en effet, un message peut être rejeté pour de multiples raisons, toutes représentées par une seule classe *rejet* ; et enfin car le corpus d'apprentissage des classifieurs est annoté par le système à base de règles lui-même, qui est donc une borne supérieure.

Tab. 2: Interpretation Error Rate (IER) du système à base de règles, d'*Icsiboost* (I.) et de *Liblinear* (L.)

Système	substitution	insertion	délétion	IER
Règles	9.7%	11.9%	1.9%	23.5%
I. Têtes	10.8%	16.8%	0.8%	28.5%
I. Arguments	9.0%	12.3%	3.7%	24.9%
I. Fusion	13.4%	16.8%	0.8%	31.1%
L. Têtes	8.3%	11.4%	3.6%	23.3%
L. Arguments	9.2%	16.8%	0.8%	26.9%
L. Fusion	14.3%	16.8%	0.8%	32.0%

La première expérience menée dans cette étude, appelée *Chronological*, souvent utilisée comme baseline, consiste à prendre les exemples dans l'ordre chronologique. A partir de l'intégralité du corpus ordonné chronologiquement L^o, nous avons extrait 10 ensembles d'apprentissage de L^{o_1} à $L^{o_{10}}$ de tailles respectives $d = (0.5k, 1k, 2k, 3k, 5k, 8k, 12k, 20k, 30k, 40k)$ avec $L^{o_n} = \{L^o_i | 1 \leq i \leq d_n\}$

La seconde expérience, appelée *Reverse chronological*, consiste à prendre l'ordre antichronologique : $L^r_i = L^o_{|L^o|-i+1}$ et $L^{r_n} = \{L^r_i | 1 \leq i \leq d_n\}$

La troisième expérience, utilisée comme baseline, est le résultat de 9 expériences *Random* sur 10 ensembles d'apprentissage $L^{b_n} = \{L^b_i | 1 \leq i \leq d_n\}$ où L^b est une séquence aléatoire du corpus d'apprentissage.

La dernière expérience, *Active-learning*, est l'application de la stratégie d'Active-learning précédemment décrite. 4 expériences ont été réalisées avec des *bootstraps* de 500 exemples pris aléatoirement parmi le corpus d'apprentissage, et considérant les autres exemples comme non annotés (U dans l'algorithme).

La figure 1 présente les résultats de ces expériences, exprimant l'IER en fonction de la quantité de données manuellement annotée.

Les résultats les moins bons sont obtenus par l'expérience *Chronological*. Cela s'explique par la différence temporelle entre les données d'apprentissage et celles de test. En prenant l'ordre chronologique, cette différence est toujours maximale, révélant une discordance entre l'état du service au moment de l'apprentissage et celui lors de la phase de collecte des données de test. Ce résultat est confirmé par l'expérience *Reverse chronological* : en minimisant la différence temporelle on obtient une amélioration très significative de la

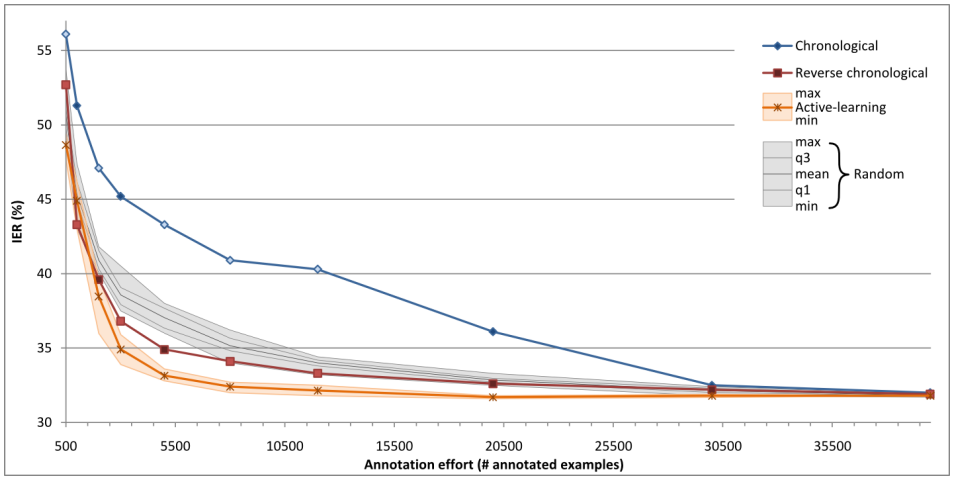


Fig. 1: Interpretation Error Rate (IER) en fonction de la taille du corpus annoté pour les différentes expériences

courbe d'apprentissage qui dépasse même la baseline *Random*. Il est intéressant de constater que n'importe quelle expérience aléatoire est meilleure que l'expérience *Chronological*, qui est pourtant la procédure la plus commune lors de la mise à jour des modèles d'un SDO déployé.

Bien que la différence temporelle soit clairement un facteur important, l'expérience d'*Active-learning* montre qu'un critère de sélection n'utilisant aucune information temporelle permet d'atteindre de meilleurs résultats que les autres expériences.

La progression de la courbe d'apprentissage de l'*Active-learning* est également bien plus rapide que toutes les autres : à 12000 exemples annotés (30% du corpus d'apprentissage), l'IER atteint une valeur comparable à celle obtenue avec l'intégralité du corpus d'apprentissage.

Dans [4], ce critère d'apprentissage actif a également été appliquée avec succès au système à base de règles, montrant des résultats comparables. Nous y proposons également une méthode réduisant l'effort et le temps d'analyse nécessaire à la rédaction des règles.

5. Conclusion

Nous avons présenté une méthode d'apprentissage actif, pouvant être appliquée à un SDO, consistant en l'apprentissage de modèles statistiques sur le corpus automatiquement annoté par le module de compréhension déployé. Les modèles obtenus sont alors utilisés au sein d'une stratégie d'apprentissage actif basée sur l'accord inter-classifieurs afin de sélectionner les nouveaux exemples à annoter à partir des traces de dialogue enregistrées par le système. Les expériences appliquées au corpus de France Telecom FT3000 ont montré que cette méthode surpasse clairement d'autres méthodes de sélection basées sur des informations temporelles ou aléatoires.

Remerciements

Ce travail est partiellement financé par le Conseil Régional PACA et par le 6e Framework Research Program de l'Union Européenne (UE), Projet LUNA, IST contrat n°33549. Les auteurs souhaitent remercier l'UE pour son soutien financier. Pour plus d'informations à propos du projet LUNA : `www.ist-luna.eu`.

Références

[1] I. Dagan and S.P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Int. Conf. on Machine Learning*, 1995.

[2] G. Damnati, F. Bechet, and R. De Mori. Spoken language understanding strategies on the france telecom 3000 voice agency corpus. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'07*, April 2007.

[3] G. Damnati, F. Bechet, and R. de Mori. Speaker turn characterization for spoken dialog system monitoring and adaptation. In *Spoken Language Technology Workshop, SLT'08*, Dec. 2008.

[4] P. Gotab, F. Bechet, and G. Damnati. Active Learning for rule-based and corpus-based Spoken Language Understanding models. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009.

[5] T.M. Kamm and G.G.L. Meyer. Selective sampling of training data for speech recognition. In *Int. Conf. on Human Language Technology Research, HLT'02*. San Diego CA, USA, 2002.

[6] G. Riccardi and D. Hakkani-Tur. Active learning : theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(4) :504–511, July 2005.

[7] D. Suendermann, K. Evanini, J. Liscombe, P. Hunter, K. Dayanidhi, and R. Pieraccini. From rule-based to statistical grammars : Continuous improvement of large-scale spoken dialog systems. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'09*, April 2009.

[8] D. Suendermann, J. Liscombe, K. Evanini, K. Dayanidhi, and R. Pieraccini. C5. In *Spoken Language Technology Workshop, SLT'08*, Dec. 2008.

[9] G. Tur, R.E. Schapire, and D. Hakkani-Tur. Active learning for spoken language understanding. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'03*, April 2003.

L'adaptation thématique d'un modèle de langue fait-elle apparaître des mots thématiques ?

Gwénolé Lecorvé, Guillaume Gravier, Pascale Sébillot

IRISA, 263 av. Gén Leclerc Campus universitaire de Beaulieu, 35042 RENNES, France
{gwenole.lecorve,guillaume.gravier,pascale.sebillot}@irisa.fr

ABSTRACT

Whereas topic-based adaptation of language models (LM) claims to increase the accuracy of topic-specific words within automatic speech recognition, this paper investigates why this wish is not always verified. After outlining the mechanisms of LM adaptation and automatic speech recognition, diagnosing elements are proposed along with solutions. In addition to a better accuracy on topic-specific words, results show better graph error rates and word error rates on a set of spoken documents with various topics.
Keywords: language models, automatic speech recognition, topic-based adaptation

1. Introduction

L'adaptation thématique des modèles de langue (ML) vise à pallier le manque d'adéquation des ML généralistes appris une fois pour toute sur des textes aux sujets variés face à un document contenant de la parole thématiquement spécifique. Dans le cadre largement répandu des ML statistiques, basés sur des n-grammes, de multiples méthodes ont déjà été étudiées pour cette tâche [1]. Leur idée principale est d'acquérir automatiquement à partir d'un corpus des informations sur un thème considéré avant de les utiliser pour ré-estimer les probabilités d'un ML généraliste. Ceci permet que plus d'importance soit accordée aux n-grammes contenant des mots thématiques, mots porteurs d'une notion du thème en question.

Dans la littérature, les informations sur un thème sont souvent modélisées soit par des probabilités n-grammes estimées sur le corpus thématique [2], soit par une distribution unigramme obtenue par une technique d'analyse sémantique latente [3]. Par ailleurs, beaucoup de travaux récents recourrent à des techniques basées sur le minimum d'information discriminante (MDI) pour l'étape de ré-estimation du ML généraliste. C'est notamment le cas dans nos travaux [5]. Toutefois, alors que la technique d'adaptation en elle-même est donc largement étudiée, la majorité des travaux comparent des taux d'erreur sans s'interroger sur l'impact précis des ML adaptés dans le complexe processus de reconnaissance automatique de parole (RAP). Il n'est pourtant pas rare d'observer dans des transcriptions l'absence de mots thématiques prononcés dans le document sonore, alors que ceux-ci sont pourtant dans le vocabulaire du système et que leurs probabilités ont été adaptées dans le ML. Laissant donc de côté le problème des mots hors vocabulaire, cet article cherche à diagnostiquer pourquoi l'effet de l'adaptation thématique peut ainsi ne pas se faire ressentir et propose des éléments de solutions portant sur le système de RAP et sur l'adaptation du ML afin de réduire le nombre d'erreurs faites sur des mots thématiques.

Après avoir rappelé le fonctionnement de nos méthode d'adaptation et système de RAP, la section 2 présente un diagnostic général du manque d'effet de l'adaptation thématique, qui met au jour des réglages inadaptés du système de RAP et une adaptation parfois trop faible du ML. Les sections 3 et 4 présentent alors des solutions respectives à ces deux problèmes. Enfin, dans la section 5, nous présentons une manière de concilier représentation thématique et qualité de la modélisation du langage.

2. Vue d'ensemble

L'adaptation thématique dans un processus de RAP vise à ré-estimer le ML et à relancer un nouveau décodage sur la base de ce nouveau ML adapté. Dans cette section, nous présentons un aperçu du fonctionnement de ces deux étapes avant de donner quelques éléments de diagnostic quant à leurs insuffisances pour l'adaptation thématique.

2.1. Ré-estimation du modèle de langue

Notre technique d'adaptation thématique est fondée sur la méthode MDI. L'originalité et l'intérêt de cette méthode est de définir le ML adapté que nous cherchons à calculer comme la solution d'un système de contraintes dont l'entropie relative avec un ML de départ est minimale. Notre approche vise à n'augmenter que les probabilités des mots thématiques. Aussi, le système de contraintes utilisé ne porte que sur les probabilités des n-grammes se terminant par des mots appartenant à la terminologie du thème considéré, appelés *termes* par la suite. Pour un n-gramme hw, il en découle l'écriture suivante des probabilités conditionnelles P_A d'un ML adapté :

$$P_A(w|h) = \frac{P_B(w|h) \times \alpha(w)}{\sum_{\hat{w} \in V} P_B(\hat{w}|h) \times \alpha(\hat{w})} \quad (1)$$

$$\text{avec } \alpha(w) = \begin{cases} \frac{P_a(w)}{P_B(w)} & \text{si } w \text{ est un terme,} \\ 1 & \text{sinon,} \end{cases} \quad (2)$$

où V est le vocabulaire du système, P_B est la distribution du ML à adapter et P_a est une distribution estimée sur un petit corpus thématique.

En pratique, pour alléger le calcul du coefficient de normalisation dans (1), la masse de probabilité des évènements observés $\mathcal{E}(h)$ pour chaque historique h est contrainte à être conservée durant l'adaptation. Il en découle une nouvelle expression de $P_A(w|h)$:

$$P_A(w|h) = \frac{P_B(w|h) \times \alpha(hw)}{Z(h)} \quad (3)$$

$$\text{avec } Z(h) = \frac{\sum_{h\hat{w}\in\mathcal{E}(h)} P_B(\hat{w}|h) \times \alpha(h\hat{w})}{\sum_{h\hat{w}\in\mathcal{E}(h)} P_B(\hat{w}|h)}. \quad (4)$$

Outre la qualité variable de nos terminologies, due à un apprentissage automatique, notre méthode peut souffrir du fait que la présence d'un terme t dans la terminologie du thème n'implique pas que ce mot ait une probabilité $P_a(t)$ élevée dans notre corpus thématique. Ce phénomène conduit alors à une adaptation trop faible de certains n-grammes.

2.2. Système de RAP

Notre système de RAP est composé de plusieurs passes dont nous présentons seulement les principales du point de vue de l'utilisation du ML. Dans un premier temps, partant de paramètres acoustiques extraits d'un signal contenant de la parole, des graphes de mots sont générés par un algorithme de recherche en faisceau en utilisant des modèles acoustiques triphones et un ML 3-gramme sur un vocabulaire de 65 000 mots. Les graphes ainsi obtenus représentent, sous une forme plus ou moins compacte, l'intégralité des hypothèses de transcription que l'on peut s'attendre à obtenir en fin de décodage. Dans un second temps, ces graphes de mots sont réévalués en utilisant des modèles acoustiques triphones plus fins et un ML 4-gramme. Sur les nouveaux graphes ainsi obtenus, plus petits, un algorithme de recherche du meilleur chemin (Viterbi dans notre système) puis un décodage par consensus sont appliqués pour finalement obtenir une transcription du signal de départ.

Comme la potentialité d'apparition d'un mot dans le résultat final d'un décodage est avant toute chose conditionnée par sa présence dans les graphes de mots, l'étape de création des premiers graphes est particulièrement importante. Dans notre système, celle-ci est implémentée selon le principe de programmation dynamique détaillé dans [7] : chaque graphe de mots est construit au fur et à mesure que les hypothèses de phrases sont chronologiquement explorées. Pour que cette exploration s'effectue en un temps raisonnable, les hypothèses partielles les moins prometteuses sont itérativement écartées grâce à une stratégie d'élagage qui tient en trois points. Tout d'abord, à chaque trame t du signal décodé, seules sont conservées les hypothèses dont le score est suffisamment proche du score de la meilleure hypothèse :

$$Q_h(s,t) > \delta_{AC} \times \max_{(h',s')} \{Q_{h'}(s',t)\} \quad (5)$$

où s est un état de la copie d'arbre lexical de l'historique h. La constante δ_{AC} est appelé *seuil acoustique*. Ensuite, le même type de seuillage est appliqué pour chaque hypothèse de fin de mot. Seules seront explorées les nouvelles copies d'arbre des historiques dont le score est suffisamment proche du score $Q_{LM}(t)$ de la meilleure hypothèse de fin de mot à la trame t :

$$Q_{h'}(s_0,t) > \delta_{LM} \times Q_{LM}(t) \quad (6)$$

avec

$$Q_{h'}(s_0,t) \simeq \max_{h,s_w} \left\{Q_h(s_w,t) \times P(w|h)^\lambda \times I^{-1}\right\}, \quad (7)$$

où w est un mot supposé se terminer à un état s_w, h' est le nouvel historique dont l'état initial est s_0, $P(w|h)$ est la probabilité du ML, λ est le poids du ML et I est la pénalité d'insertion d'un mot. Le facteur δ_{LM} est appelé *seuil linguistique*. Enfin, parmi ces hypothèses de fin de mot ayant survécu, seules sont conservées les M hypothèses ayant le meilleur score. En pratique, les constantes λ, I, δ_{AC}, δ_{LM} et M sont des valeurs empiriques généralement obtenues par la recherche d'un compromis optimal entre différents critères comme le taux oracle des graphes générés, la durée de cette génération et la taille des graphes.

2.3. Premiers éléments de diagnostic

Dans cet article, nos expériences sont basées sur 91 documents sonores thématiquement homogènes issus de 3h d'émissions d'actualités du corpus ESTER 1 [4]. Ces segments, provenant de 3 radios différentes et datés de la même période, sont variés en terme de thème (guerre en Irak, politique nationale et internationale, sports...) et de longueur (de 30 à 2 000 mots). Pour chaque segment, les ML généralistes de notre système sont adaptés selon la méthode décrite en 2.1 puis un décodage basé sur ces nouveaux ML adaptés est exécuté. Nous présentons un bref bilan de ce décodage quant à son impact sur les mots thématiques.

L'intérêt d'utiliser un ML adapté durant un décodage est de favoriser en sortie l'émergence d'hypothèses comportant des séquences probables dans le thème considéré, séquences jusqu'alors sous-estimées par le ML généraliste. Si cet effet est bien observé lorsque les mots thématiques à corriger sont déjà dans les graphes de mots générés en première passe avec le ML généraliste, il apparaît au contraire que l'impact de l'adaptation thématique est souvent insuffisant pour insérer dans les graphes de mots des hypothèses comportant de nouveaux mots thématiques. Nous identifions principalement deux raisons à cela. Tout d'abord, nous avons observé que les termes d'un thème font particulièrement les frais de l'élagage de l'espace de recherche utilisé lors de la génération des graphes de mots en première passe – ceci en dépit de l'importance linguistique accrue que leur apporte un ML adapté. Nous expliquons ce phénomène par la tendance des termes à être des mots rares, des mots techniques, des entités nommées..., caractéristique qui accentue les risques pour ces mots d'être mal prononcés ou mal phonétisés. Ensuite, il semblerait que, malgré l'adaptation du ML, certains n-grammes se terminant par un terme aient toujours des probabilités trop faibles pour que les hypothèses de phrase qui les contiennent survivent au seuillage linguistique. La suite de cet article revient sur ces deux problèmes et présente des solutions envisageables.

3. Favoriser le modèle de langue

Malgré l'importance plus grande donnée aux termes dans les ML adaptés, le calcul des scores Q_h accorde toujours la même importance au ML, conduisant les hypothèses de phrases contenant ces termes à être élaguées. À ce phénomène peuvent s'ajouter

Tab. 1: GER et WER pour les réglages d'origine et pour les nouveaux réglages, avec le ML généraliste (ML_B) ou le ML adapté (ML_A).

	Réglages d'origine		**Réglages modifiés**	
	ML_B	ML_A	ML_B	ML_A
GER	8.9	8.6 (-0.3)	8.5	**8.1 (-0.4)**
WER	21.8	21.0 (-0.8)	21.0	**20.5 (-0.5)**

Tab. 2: Exemple d'alignement d'un groupe de souffle de référence (Réf) pour les réglages d'origine et modifiés, avec un ML généraliste (ML_B) ou adapté (ML_A).

Réf : `cas probable de la maladie`
Réglages d'origine
ML_B : `cas probable de la MÊLÉES`
ML_A : `cas probable de la MALAISIE`
Réglages modifiés
ML_B : `cas probable de la MÊLÉES`
ML_A : `cas probable de la maladie`

des conditions acoustiques difficiles, par exemple une mauvaise phonétisation, une prononciation erronée ou encore un locuteur parlant avec un accent régional ou étranger. Pour pallier ce problème, nous proposons de rendre le seuillage acoustique plus tolérant et de donner plus d'importance au ML dans le calcul des scores Q_h. En pratique, ceci consiste à diminuer δ_{AC} et à augmenter λ. Par ailleurs, pour contrebalancer l'augmentation du nombre d'hypothèses actives engendrée par la diminution de δ_{AC}, le nombre d'hypothèses de fin de mot pour chaque trame M est abaissé de manière à ce que le temps global de calcul soit conservé par rapport aux réglages d'origine.

Le tableau 1 présente les taux d'erreur en mots sur les graphes en première passe (GER) et sur les transcriptions finales (WER) obtenues, avec ou sans adaptation thématique, pour les réglages d'origine et nos réglages modifiés. Tout d'abord, il ressort clairement que la nouvelle configuration produit des taux d'erreur nettement meilleurs. Si les gains sur le GER s'expliquent par une augmentation constatée du nombre d'hypothèses de phrase dans les graphes générés en première passe, les résultats en WER montrent que nos anciens réglages n'étaient pas optimaux. Ensuite, il apparaît que les gains initiaux impliqués par l'adaptation thématique (colonne 2) se cumulent en partie mais pas entièrement avec ceux obtenus par nos nouveaux réglages sans adaptation (colonne 3). Ceci s'explique par le fait que les nouveaux réglages permettent notamment de corriger des erreurs sur des termes. Toutefois, les gains reportés pour l'utilisation conjointe des nouveaux réglages et de l'adaptation thématique (colonne 4) sont le signe d'une certaine complémentarité entre ces deux mécanismes. Le tableau 2 illustre ce propos en présentant des alignements d'un groupe de souffle tiré d'un document parlant de la pneumonie atypique où le locuteur prononce « *malédie* » au lieu de « *maladie* ». Alors que « *maladie* » n'apparaît dans aucun des trois premiers cas, soit à cause du problème acoustique, soit à cause d'une probabilité linguistique trop faible, la combinaison de nos nouveaux réglages et de l'adaptation thématique permet d'obtenir la bonne sortie. En contrepartie, l'augmentation de λ semble provoquer la disparition de nombreux mots courts (prépositions, articles...). De plus, certains termes n'apparaissent toujours pas dans les graphes de mots. Nous pensons que ceci est dû à la probabilité linguistique trop faible que peut leur attribuer notre technique d'adaptation.

Tab. 3: WER pour différents ML utilisés lors de la création des graphes, puis avec un ML adapté avec P_a pour le reste du décodage. Entre parenthèses, le type de poids utilisé lors de l'adaptation du premier ML.

	Réglages d'origine	**Réglages modifiés**
ML_B	21.8	20.9
ML_A ($P_a(t)$)	21.0 (-0.8)	**20.5 (-0.4)**
ML_A ($K = 10^{-8}$)	21.5 (-0.3)	20.8 (-0.1)
ML_A ($K = 10^{-5}$)	21.6 (-0.2)	21.0 (+0.1)

Tab. 4: Exemple d'alignement des graphes d'un groupe de souffle de référence (Réf) pour différents ML utilisés lors de la création des graphes.

Réf : `accès à la frontière du libéria`
Réglages modifiés
ML_B : `À SERT la ANTIENNE du DÉLIRE`
ML_A (P_a) : `MERCI à la frontière NOUVELLE`
ML_A (10^{-8}) : `MERCI à la frontière LIBÉRIENNE`
ML_A (10^{-5}) : `MERCI à la frontière libéria`

4. Surpondérer les termes

Comme évoqué en 2.1, la probabilité $P_a(t)$ d'un terme t peut conduire à une adaptation trop faible et inhiber l'apparition d'hypothèses de phrases contenant ce terme dans les graphes de mots. Pour pallier ce problème, nous proposons alors d'utiliser des poids K arbitrairement élevés et identiques pour tous les termes. Le facteur de mise à l'échelle $\alpha(w)$ d'un n-gramme hw dans (2) se réécrit alors :

$$\alpha(w) = \begin{cases} \frac{K}{P_B(w)} & \text{si } w \text{ est un terme,} \\ 1 & \text{sinon.} \end{cases} \quad (8)$$

Les résultats sur le GER obtenus pour différentes valeurs de K nous montrent, d'une part, que des poids trop élevés dégradent vite les performances et qu'un poids $K \approx 10^{-8}$ semble optimal quel que soit le réglage. En poursuivant le reste du décodage avec un ML adapté avec P_a, nous obtenons les taux d'erreur de la table 3. Ces résultats sont reportés pour deux valeurs de K : la valeur optimale en terme de taux oracle (10^{-8}) et une valeur beaucoup plus grande (10^{-5}). Dans l'ensemble, il ressort que l'utilisation de poids fixes n'apporte rien voire dégrade les taux d'erreur obtenus avec un ML généraliste. Cependant, une analyse qualitative des résultats montre qu'une adaptation à poids fixes et forts produit bien, lors de la création des graphes, une apparition de termes qui n'apparaissaient jusqu'alors pas, même en utilisant une adaptation classique avec P_a (*cf.* exemple de la table 4).

Nous expliquons ce comportement par deux raisons. D'une part, nous pensons que certaines hypothèses introduites lors de la création des graphes ne survivent pas aux différents réglages d'élagage des étapes postérieures. Il faudrait alors modifier ces réglages, notam-

Tab. 5: GER mesurés après fusion des graphes de mots issus d'un premier décodage avec le ML généraliste et d'un second décodage avec différents ML.

1er décodage ► / **▼ 2nd décodage**		Réglages d'origine + ML_B	Réglages modifiés + ML_B
Réglages d'origine	+ ML_B	8.9	7.8 (-1.1)
	+ ML_A	8.5 (-0.4)	7.7 (-1.2)
Réglages modifiés	+ ML_B	7.8 (-1.1)	8.5 (-0.4)
	+ ML_A	**7.6 (-1.3)**	8.1 (-0.8)

ment encore une fois au niveau acoustique. D'autre part, la normalisation utilisée en pratique dans MDI (formule 4) ne permet pas une modification globale de la distribution d'un ML mais aboutit seulement à des modifications locales de celle-ci, historique par historique. Lorsque les poids considérés sont importants, ceci tend à rendre moins probables, en moyenne, les historiques adaptés que ceux qui ne le sont pas.

5. Fusionner les graphes de mots

Comme le montrent les expériences et observations précédentes, il est difficile de concilier au niveau d'une adaptation de ML les informations que l'on possède sur un thème avec celles plus généralistes d'un ML initial. Il semble alors intéressant de s'orienter vers une intégration *a posteriori* de celles-ci, notamment via la fusion de graphes de mots [6]. Pour cela, nous définissons la fusion de deux graphes de mots comme le graphe déterminisé représentant l'union de l'ensemble des phrases codées par chaque graphe. Nous appliquons cette méthode de fusion entre les graphes générés avec le ML généraliste et ceux obtenus grâce à un ML adapté.

Le tableau 5 présente les résultats GER obtenus par cette méthode de fusion pour différents couples de réglages. Il apparaît que les gains les plus importants sont ceux où sont utilisés deux réglages différents de l'algorithme de création des graphes (cellules grisées). Ces gains dépassent nettement les gains sans fusion de la table 1. L'effet de la fusion est tel que celui de l'adaptation thématique semble quasi gommé. Nos résultats sur le WER montrent cependant que ces différences sont lissées à la sortie du système. Seuls ressortent des écarts notables pour chaque couple de réglages entre l'utilisation d'un ML généraliste et d'un ML adapté.

Ces résultats quelque peu décevants nous apparaissent cependant comme logiques. En effet, notre méthode de fusion n'aboutit qu'à considérer l'union des hypothèses des graphes fusionnés et n'introduit donc aucune nouvelle hypothèse. Ainsi, quel que soit le ML utilisé, celui-ci privilégiera quasi toujours les mêmes hypothèses qu'il privilégiait déjà sans fusion. Il serait intéressant d'étudier l'impact de stratégies de fusion plus élaborées permettant d'introduire de nouvelles hypothèses, par exemple une stratégie basée sur la combinaison de réseaux de confusion.

6. Conclusions

Dans cet article, nous avons cherché à diagnostiquer l'impuissance à transcrire des mots thématiques dont souffre parfois l'intégration d'un ML adapté thématique dans le processus de RAP. Pour cela, deux mécanismes ont été proposés au niveau de la génération des graphes de mots en première passe : une prise en compte moins forte de l'acoustique au profit des scores du ML et une adaptation plus marquée du ML utilisé en première passe. De plus, pour concilier qualité de représentation thématique et modélisation du language, nous avons proposé une technique de fusion de graphes. Outre les gains GER et WER qui peuvent être relevés, ces mécanismes tendent à améliorer la transcription des mots thématiques au détriment de portions de texte plus généralistes.

Ces résultats ouvrent probablement la voie à de meilleurs résultats dans des tâches où les mots discriminants ont une importance particulière comme, par exemple, l'indexation. Au-delà de cela, il est intéressant de réfléchir plus en avant au principe d'adaptation thématique. En effet, ce dernier est trop souvent centré sur la tâche de ré-estimation du ML dont le résultat serait censé se suffire à lui-même. À notre sens, ce postulat est illusoire. Nous pensons que différentes solutions peuvent être envisagées. Par exemple, il serait bon de systématiser les mécanismes de fusion *a posteriori* ou de systèmes de RAP collaborant en parallèle. De même, il serait intéressant de réfléchir à une modification des algorithmes du processus de RAP pour pouvoir intégrer directement des modèles indépendants de différentes sources d'information. Enfin, bien que la question des mots hors vocabulaire n'ait pas été traitée dans cet article, celle-ci joue un rôle dans l'impact d'une adaptation thématique car la rareté et la spécificité générale des mots thématiques implique souvent leur absence dans le vocabulaire du système. L'adaptation de ce dernier est donc un enjeu majeur pour l'adaptation thématique d'un système.

Références

[1] J. R. Bellegarda. Statistical language model adaptation : review and perspectives. *Speech Communications*, 2004.

[2] M. Federico. Efficient language model adaptation through MDI estimation. In *Proc. Eurospeech*, 1999.

[3] M. Federico. Language model adaptation through topic decomposition and MDI estimation. In *Proc. ICASSP*, 2002.

[4] S. Galliano, É. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proc. Eurospeech*, pages 1149–1152, 2005.

[5] G. Lecorvé, G. Gravier, and P. Sébillot. Constraint selection for topic-based MDI adaptation of language models. In *Proc. Interspeech*, pages 368–371, 2009.

[6] X. Li, R. Singh, and R. M. Stern. Combining search spaces of heterogeneous recognizers for improved speech recognition. In *Proc. ICSLP*, pages 405–408, 2002.

[7] H. Ney and S. Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, pages 64–83, 1999.

L'échelle OME (Octave-MEdiane) : une échelle naturelle pour la mélodie de la parole.

Céline De Looze, Daniel Hirst

Laboratoire Parole et Langage, CNRS UMR 6057
Université de Provence, Aix-en-Provence
{celine.delooze, daniel.hirst}@lpl-aix.fr
http://www.lpl.univ-aix.fr

ABSTRACT

Fundamental frequency, the primary acoustic correlate of speech melody, is generally analysed and displayed using a linear scale (in Hertz) or a logarithmic one (usually in semitones), generally offset to an arbitrary reference level. In this paper we argue that a more natural scale for analysing speech is the OME (Octave-MEdian) scale, using the octave (8ve) as the basic unit, centred on the median of the speaker's range. We present results showing that a reasonable estimate of a speaker's pitch range can be obtained directly from the median.

Keywords: prosodie, mélodie, échelle naturelle, octave

1. INTRODUCTION

Bien qu'on puisse observer une certaine non-linéarité dans la perception de la hauteur des sons de la parole, la fréquence fondamentale reste néanmoins le corrélat acoustique principal de cette hauteur.

Des échelles psycho-acoustiques propres à l'étude de la parole ont été proposées, en particulier les échelles Mel, Bark et ERB. La pertinence relative de ces échelles, cependant, reste à établir. Une étude récente [16] montre, par exemple, que dans des tâches de réplication de contours, entre voix d'hommes et voix de femmes, une échelle logarithmique rend mieux compte de la performance des locuteurs qu'une échelle linéaire ou qu'une échelle psycho-acoustique.

L'échelle physique en Hertz (cycles par seconde) est souvent transformée dans des études sur la prosodie en une échelle logarithmique, généralement exprimée en demi tons avec une valeur de référence (appelé C0), fixée arbitrairement à 16.3516 Hz [17]. Fant et al [9], ont proposé l'unité *St* définie comme suit :

(1) $St = 12[\ln(Hz/100)/\ln 2]$.

avec le niveau de référence, donc, à 100 Hz.

Le demi ton n'a, cependant, rien de naturel comme unité de mesure. Il est, en effet, le produit d'une évolution complexe de la culture musicale occidentale classique, correspondant à la division de l'octave en 12 intervalles égaux, une idée qui avait été déjà décrite dans un traité publié en Chine en 1584 [13]. En Europe, la gamme à 12 demi tons égaux, dite *gamme à tempérament égal*, a été employée progressivement depuis le 18e siècle pour accorder les claviers de musique, en remplacement de la gamme *naturelle* ('just intonation') utilisée auparavant, ou encore de la gamme *bien tempérée* de Bach. Il s'agissait chaque fois de la recherche d'un compromis permettant de moduler d'une gamme à une autre, sans introduire de discordance majeure et sans avoir à changer de clavier.

Dans différentes civilisations à différentes époques, on observe l'utilisation de gammes de notes différentes. La plupart de ces gammes, cependant, ont en commun le fait que les noms des notes sont généralement les mêmes, quelle que soit l'octave. Ainsi, dans l'échelle classique occidentale, par exemple, la séquence *do ré mi fa sol la si do ré mi... etc* peut se répéter indéfiniment dans les limites physiques de la production sonore. Des travaux récents semblent établir que cette circularité (connue aussi sous le nom de répétition chromatique) a des bases physiologiques dans la perception des sons par des humains [4][5], y compris par des nouveaux-nés [15], et également par des singes [18].

En tout cas, c'est l'octave, et non le demi-ton, qui apparaît clairement comme l'intervalle naturel pour la *perception* des hauteurs des sons de parole et de la musique.

On a suggéré par ailleurs [10][11] qu'il y a peut-être également une explication physiologique pour l'octave et la demi-octave comme unité pour la *production* d'intervalles mélodiques. Hirst [10] rapporte une expérience où ces deux intervalles sont observés comme valeurs modales dans une tâche de production de contours variés sur des syllabes isolées, « oui » et « non ».

Bien que le mécanisme du contrôle de la fréquence fondamentale ne repose pas uniquement sur l'élongation des plis vocaux, on peut penser que dans la mesure où ceux-ce se comportent comme des cordes vibrantes, alors, suivant la loi de Mersenne, un doublement de la tension des plis vocaux correspondra à une montée mélodique d'une demi-octave. Ceci pourrait expliquer la raison pour laquelle les deux intervalles – octave et demi-octave – semblent être fréquents dans la production mélodique, malgré le fait qu'une montée ou une chute d'une octave sur une seule syllabe, par exemple, n'est certainement pas perçue dans sa totalité.

Dans la suite de ce travail, nous décrivons une étude menée sur 4 corpus, en anglais et en français, qui montre que, dans la production de la parole naturelle, les variations de la fréquence fondamentale seraient délimitées par l'octave supérieure et la demi-octave inférieure par rapport à la hauteur médiane de la voix d'un locuteur, ce qui nous mène à proposer une nouvelle échelle de mesure normalisée : l'OME. Pour plus de détails concernant cette étude, voir [8].

1. CORPUS

Quatre corpus représentant un total d'environ 2 heures de parole, ont été utilisés pour cette étude : les corpus PFC et CID (pour l'étude du français) et les corpus PAC et AIX-MARSEC (pour l'étude de l'anglais). Nous décrivons ces corpus ci-après :

PFC (*Phonologie du Français Contemporain*) [7] : Nous avons sélectionné 10 locuteurs français, originaires de Marseille, 6 femmes et 4 hommes. Nous avons choisi de leurs productions les lectures oralisées (lecture à voix haute d'un passage, de type article de journal régional). Cela correspond à environ 30 minutes d'enregistrement.

CID (*Corpus of Interactional Data* [2]) : Nous avons sélectionné 6 locuteurs français, originaires de Marseille, 3 hommes et 3 femmes. Les enregistrements correspondent à des conversations où les locuteurs évoquent des conflits professionnels ou des situations insolites dans lesquels ils se sont trouvés, un total de 30 minutes d'enregistrement.

PAC (*Phonologie de l'Anglais Contemporain*) [6] : Nous avons sélectionné 8 locuteurs anglais, 4 hommes et 4 femmes, originaires du Nord de l'Angleterre. Nous avons choisi de leurs productions les lectures oralisées (lecture à voix haute d'un passage, de type article de journal régional). Cela correspond à environ 25 minutes d'enregistrement.

AIX-MARSEC [1] : Les enregistrements correspondent à des extraits de la BBC des années 80. Nous avons sélectionné 51 locuteurs, 13 femmes et 38 hommes. 11 types de production sont représentés : commentaires, bulletins d'information, paroles publiques, émissions religieuses, reportages, fictions, poésies, dialogues, propagandes, etc., que nous qualifions de parole authentique, i.e. de la parole produite dans un but de communiquer avec un ou plusieurs auditeurs. Cela représente un total de 50 minutes d'enregistrement environ.

2. ESTIMATION DES PARAMETRES DE REGISTRE

Le registre d'un locuteur (i.e. la gamme tonale, ou espace tonal, effectivement utilisée dans un énoncé, à différencier de sa tessiture qui correspond à la plage totale des sons qu'un individu est capable d'émettre), est généralement défini par deux paramètres : sa hauteur et son étendue [14]. La *hauteur* est le plus souvent mesurée en termes de la moyenne ou de la médiane de la distribution de f0, ou alors en termes de la moyenne de points étiquetés au préalable comme représentant des cibles ou tons bas. *L'étendue* peut être mesurée en termes de la différence entre la valeur minimale et la valeur maximale produites dans un énoncé ou en termes de différence entre la moyenne des cibles hautes et la moyenne des cibles basses. Une mesure du registre fondée sur l'analyse de cibles tonales, peut être à la fois couteuse et source d'erreur, en particulier si ces cibles sont annotées manuellement.

Dans cette étude, nous mesurons la valeur du registre à partir de la *médiane* de la distribution de la f0, qui est plus stable que la moyenne (influencée, elle, par des valeurs extrêmes) et qui, du fait de sa nature non-paramétrique, est indépendante de l'échelle de mesure.

Afin d'éviter les problèmes inhérents aux mesures manuelles, nos mesures sont effectuées au moyen de l'algorithme MOMEL-INTSINT [12]. Le système INTSINT permet, à partir d'une modélisation automatique en cibles tonales (MOMEL), de coder ces cibles au moyen d'un alphabet de 8 symboles discrets. Les codages T(op) et B(ottom) délimitent les valeurs hautes et basses du registre du locuteur, le codage M(id) sa valeur moyenne. Les cibles H(igher), L(ower), S(ame), U(pstep) et D(ownstep) sont, en revanche, encodées en tenant compte de la cible qui les précède et sont définies comme étant, respectivement, plus haute, plus basse, égale, un peu plus haute et un peu plus basse que la cible qui les précède. Ce codage INTSINT est obtenu automatiquement à partir d'un ensemble de points cibles MOMEL avec un codage optimisé pour la totalité des points-cibles analysés.

Dans notre étude, chaque estimation de ces paramètres est obtenue à partir de 660 cibles tonales en moyenne (plus ou moins en fonction de la longueur de l'enregistrement).

Nous avons vu que la façon la plus commune de mesurer l'étendue du registre est d'utiliser la différence entre la moyenne des tons hauts et celle des tons bas. Il est intéressant de regarder la corrélation, d'une part, entre la médiane et la moyenne des tons bas (B), et, d'autre part, entre la médiane et la moyenne des tons hauts (T).

On observe, en effet, deux fortes corrélations (Figure 1). Pour la première, le coefficient de détermination (R^2) est de 0.92, pour la deuxième de 0.91. Il est donc possible à partir de la médiane de la distribution de la f0 de prédire les limites du registre d'un locuteur et ainsi son étendue. Les relations affines obtenues sont les suivantes :

$$(2) \qquad B = 0.741*\text{médiane} - 5.52$$
$$T = 1.537*\text{ médiane} + 3.75$$

Les tests de significativité des coefficients de régression donnent une probabilité critique < 2^e-16. En revanche, les probabilités critiques des constantes sont non significatives (0.161 et 0.659). Un ajustement des modèles sans la constante est donc calculé avec les valeurs :

$$(3) \qquad B = 0.706*\text{ médiane}$$
$$T = 1.561*\text{médiane}$$

Il est cependant important de vérifier s'il existe une interaction avec le sexe du locuteur, la langue ou le type de production. Cela revient à tester si les pentes de régression linéaires en fonction des niveaux de chaque facteur (homme/ femme ; anglais/français ; lecture/ parole authentique) sont significativement différentes. Les ANOVAs effectuées pour la prédiction de la moyenne des tons bas (B) à partir de la médiane révèlent qu'il n'y a pas d'effet de sexe (p=0.0917), ni de langue (p=0.170), ni de

type de production (p-val=0.134). Les ANOVAs effectuées pour la prédiction de la moyenne des tons hauts (T) rejettent également les effets de sexe du locuteur (p=0.381), de langue (p=0.274) ou encore de type de production (p=0.368).
Il est donc possible, à partir de la médiane, quelque soit le sexe du locuteur, aussi bien en anglais et en français et quelque soit le style de parole qu'il adopte, de prédire les limites (l'étendue) de son registre.
En fait, 4 [8] montre que la relation entre la hauteur du registre et son étendue est complexe. Dans le modèle donné en (3), l'étendue en Hz est strictement proportionnelle à la hauteur puisque le rapport entre T et B est fixe. Une corrélation encore plus forte est obtenue avec les valeurs sur une échelle logarithmique, ce qui donnerait comme modèle un registre dont l'étendue en octaves est proportionnelle à la hauteur, allant d'une octave pour une voix grave à un peu moins d'une octave et demi pour une voix aiguë.
Cette co-variation avait été soulignée par Ladd [14] dans sa définition du registre. L'auteur explique en effet que la difficulté d'admettre deux dimensions au registre vient du fait que ces deux dimensions co-varient.

3. L'ECHELLE OME (OCTAVE-MEDIANE), UNE ECHELLE NATURELLE NORMALISEE POUR LA MELODIE DE LA PAROLE

Il est intéressant de noter dans les relations définies en (2) que le coefficient 0.706 correspond presque exactement à une demi-octave (log2(0.706) = -0.502) et que le coefficient 1.561 est légèrement supérieur à une demi-octave (log2(1.561) = 0.642). Nous pouvons donc conclure que la moyenne des tons hauts et la moyenne des tons bas, i.e. les limites du registre d'un locuteur, se trouvent généralement dans une étendue située à peu près à plus et moins une demi-octave par rapport à la médiane.
La figure 1 donne une représentation graphique de la moyenne des tons bas (B) et la moyenne des tons hauts (T) en fonction de la médiane. Les régressions linéaires correspondantes sont tracées en lignes continues et les lignes en pointillés représentent les intervalles +octave, +demi-octave, unisson, -demi-octave et –octave par rapport à la médiane. La régression linéaire pour la moyenne des tons bas (B) se confond avec la demi-octave en-dessous de la médiane, celle de la moyenne des tons hauts (T) se situe entre l'intervalle demi-octave et octave au dessus de la médiane.
Ces intervalles musicaux, définis par rapport à la médiane, pourraient donc être utilisés pour estimer le registre d'un locuteur. Par ailleurs, ils nous permettent de proposer une échelle naturelle normalisée pour l'analyse et la visualisation de la mélodie de la parole définie en octaves, centrée sur la médiane, l'échelle OME (Octave-MEdiane).
La Figure 2 donne en exemple la phrase « J'ai des problèmes avec mon adoucisseur d'eau. » lue par un locuteur féminin et un locuteur masculin.

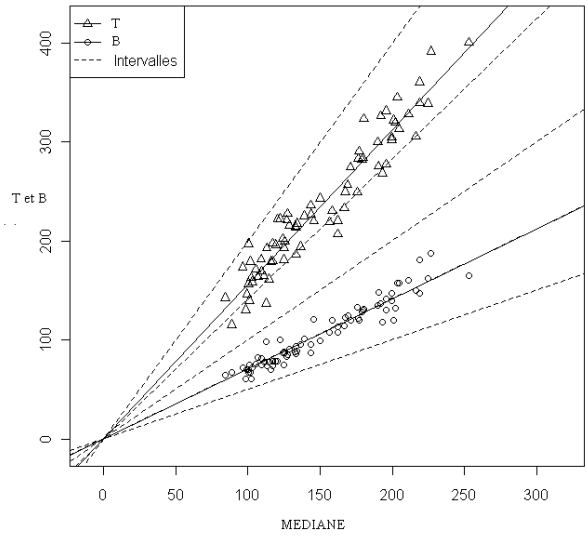


Figure 1 : *Représentation graphique de la moyenne des tons bas (B) et la moyenne des tons hauts (T) en fonction de la médiane ; en lignes continues, les régressions linéaires correspondantes ; en lignes pointillées, les intervalles +octave, +demi-octave, unisson, -demi-octave (cachée par la ligne continue correspondant à la régression linéaire des tons bas) et -octave par rapport à la médiane.*

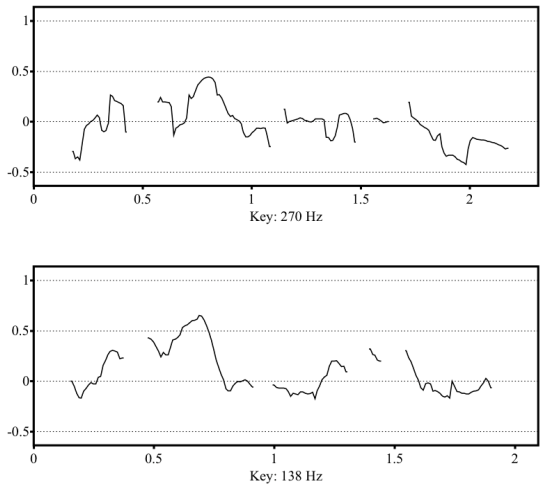


Figure 2 : *la phrase « J'ai des problèmes avec mon adoucisseur d'eau. » lue par un locuteur féminin (en haut) et un locuteur masculin (en bas) visualisée au moyen de l'échelle OME. Le niveau 0 correspond à la valeur médiane de fréquence fondamentale, les autres lignes horizontales indiquent la demi-octave en dessous la médiane et la demi-octave et l'octave au-dessus.*

La visualisation de ces deux énoncés est obtenue automatiquement à partir de la courbe de F0 au moyen du petit script Praat [3] donné en annexe. Il est à noter qu'avec cette technique, les paramètres optimaux pour

l'analyse de la fréquence fondamentale du locuteur sont automatiquement déterminés en fonction de la médiane.

4. CONCLUSIONS

On propose dans ce travail d'utiliser l'octave, plutôt que le demi-ton, comme unité de base définissant une échelle naturelle pour la prosodie de la parole. La hauteur de référence d'une telle échelle est donnée par la valeur médiane de la distribution analysée. On montre par ailleurs que dans les corpus étudiés la valeur minimale de cette échelle se situe à une demi-octave en-dessous de la médiane alors que le maximum se situe entre une demi-octave et une octave au dessus. L'échelle ainsi définie permet d'obtenir une visualisation automatique normalisée d'une courbe de fréquence fondamentale dont les limites sont déterminées à partir de la valeur médiane.

ANNEXE: SCRIPT PRAAT

```
##Praat script – Display pitch with OME Scale
## version 2010:01:18
## author <daniel.hirst@lpl-aix.fr>
min_f0 = 50
max_f0 = 750
time_step = 0.01
Erase all
mySound = selected("Sound")
To Pitch... time_step min_f0 max_f0
median = Get quantile... 0 0 0.5 Hertz
log_median = log10(median)
top_f0 = median*2
log_top = log10(top_f0)
bottom_f0 = median/sqrt(2)
log_bottom = log10(bottom_f0)
log_mid = log10(bottom_f0*2)
Remove
select mySound
To Pitch... time_step bottom_f0 top_f0
Draw logarithmic... 0 0 bottom_f0/1.1 top_f0*1.1 no
Draw inner box
One mark left... log_bottom no yes yes -0.5
One mark left... log_median no yes yes 0
One mark left... log_mid no yes yes 0.5
One mark left... log_top no yes yes 1
Marks bottom every... 1 0.5 yes yes no
Text bottom... yes Key: 'median:0' Hz
Remove
Select mySound
```

5. RÉFÉRENCES

[1] Auran, C., Bouzon, C. & Hirst, D.J. 2004. The Aix-MARSEC Project: An evolutive database of spoken British English, in *Proceedings of the 2nd International Conference on Speech Prosody* 2004, ISCA.

[2] Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. and Rauzy, S. 2007. Le CID-Corpus of Interactional Data-: protocoles, conventions, annotations. In *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence*, 25, 25-55.

[3] Boersma, P. & Weenink, D. 2010. *Praat: doing phonetics by computer* (Version 5.1.23) [Computer program]. Downloaded from http://www.praat.org/.

[4] Braun, M., Chaloupka, V. 2005. Carbamazepine induced pitch shift and octave space representation. *Hear. Res*. 210, 85-92.

[5] Braun, M. 2006. A retrospective study of the spectral probability of spontaneous otoacoustic emissions: Rise of octave shifted second mode after infancy. *Hear. Res*. 215, 39-46.

[6] Carr, P. Durand, J. 2003. *La Phonologie de l'Anglais Contemporain : usages, variétés et structure / The Phonology of Contemporary English: usages, varieties and structure*.

[7] Delais-Roussarie, E. & Durand, J. 2003. *Corpus et variation en phonologie du français: méthodes et analyses*, Presses Universitaires du Mirail, Toulouse.

[8] De Looze, Céline. 2009. *Analyse et interprétation de l'empan temporel des variations prosodiques en anglais et en français.* Thèse de doctorat 2009. LPL et Université de Provence, Aix-en-Provence.

[9] Fant, Gunnar; Anita Kruckenberg; Kjell Gustafson; Johan Liljencrants. 2002. A new approach to intonation analysis and synthesis of Swedish. in *Proceedings of the 1st International Conference on Speech Prosody*, Aix en Provence. 283-286.

[10] Hirst, D.J. 1981. Phonological implications of a production model of English intonation. *Phonologica* 1980, 195-201.

[11] Hirst, D.J. 1983. Structures and categories in prosodic representations. in Cutler & Ladd 1983. *Prosody : Models & Measurements* (Springer, Berlin), 93-109.

[12] Hirst, D.J. 2007. A Praat Plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. *ICPhS*. 1233-1236.

[13] Kuttner, Fritz A. 1975. Prince Chu Tsai-Yü's Life and Work: A Re-Evaluation of His Contribution to Equal Temperament Theory. *Ethnomusicology*, Vol. 19, No. 2 (May, 1975), pp. 163–206.

[14] Ladd, D.R. 1996. *Intonational Phonology*, Cambridge University Press, Cambridge, G.B.

[15] Liu J, Wang N, Li J, Shi B, Wang H 2009. Frequency distribution of synchronized spontaneous otoacoustic emissions showing sex-dependent differences and asymmetry between ears in 2- to 4-day-old neonates. *Int J Pediatr Otorhinolaryngol.* 2009 May;73(5):731-6

[16] Nolan, Francis. 2003. Intonational equivalence: an experimental evaluation of pitch scales. in *Proceedings of ICPhS 15*, Barcelona, 771-774.

[17] Young, R.W. 1939. Terminology for Logarithmic Frequency Units. *The Journal of the Acoustical Society of America*. 1939. 11: 134.

[18] Wright, Anthony A. ; Jacquelyne J. Rivera; Stewart H. Hulse; Melissa Shyan; Julie J. Neiworth.2000. Music perception and octave generalization in rhesus monkeys. J *Exp Psychol Gen*, Sep, Vol 129 No 3, 291-307

Recherche automatique d'hétéro-répétitions dans un dialogue oral spontané

Brigitte Bigi, Roxane Bertrand, Mathilde Guardiola

Laboratoire Parole & Langage, CNRS & Aix-Marseille Universités
5, avenue Pasteur, BP 80975, 13604 Aix en Provence, France
Mél : {brigitte.bigi,roxane.bertrand,mathilde.guardiola}@lpl-aix.fr

ABSTRACT

Other-repetitions are a device involving the reproduction by a speaker of what another speaker has just said. This paper proposes a solution to automatically detect other-repetitions in French conversational dialogue. A first step of the proposed system consists in finding all possible other-repetitions in the dialogue. A second step is used to select other-repetitions which need to be kept by combining rules with speaker statistics. This automatic detection, evaluated on a one hour dialogue, shows good results according to the expected objectives : recall is 1, and precision is about 80%.
Keywords : other-repetitions, automatic detection, conversational dialogue, French.

1. INTRODUCTION

Ce papier présente une méthode d'identification automatique des hétéro-répétitions ou répétitions diaphoniques [11], c'est-à-dire la reproduction par un locuteur 2 d'un énoncé ou d'une partie d'énoncé préalablement produit par un locuteur 1. La majorité des travaux relatifs aux répétitions concerne essentiellement les auto-répétitions [5], lesquelles sont fréquemment associées aux phénomènes de disfluences du discours (voir entre autres [6, 7]). Les hétéro-répétitions restent quant à elles encore peu explorées. A notre connaissance, il n'existe pas de travaux relatifs à leur identification automatique. Le but d'un tel outil est de proposer un nombre réduit de segments du dialogue susceptibles d'être des répétitions, de sorte que seuls ces derniers soient examinés manuellement. Disposer d'un tel outil se révèle donc une aide précieuse pour l'annotation de corpus de dialogues de plus en plus importants en taille. Il permet en outre de systématiser et de comparer des données nombreuses extraites de corpus variés de dialogues (débat, entretiens, etc.).

L'intérêt porté aux hétéro-répétitions s'inscrit dans le cadre plus large d'un projet sur l'imitation dans la parole (projet ANR SPIM[1]). Les répétitions, phénomène universel et particulièrement crucial dans la communication humaine, constituent en ce sens l'un des procédés les plus explicites pour explorer cette question, en témoignent les nombreux travaux menés en acquisition du langage notamment (pour une revue, voir [5]) et le rôle indiscutable de l'imitation pour améliorer les compétences linguistiques du petit enfant. D'un point de vue pragmatique, des résultats montrant que les répétitions joueraient davantage encore un rôle dans l'acquisition d'une compétence communicative (versus une compétence linguistique) chez l'enfant, ont toutefois conduit les auteurs à distinguer répétition et imitation [4]. Au delà de cette dernière, les répétitions serviraient donc d'autres objectifs communicatifs. Nous nous intéressons précisément aux hétéro-répétitions pour les différentes fonctions pragmatiques qu'elles revêtent dans l'interaction. Parmi d'autres, les répétitions constituent des procédés discursifs grâce auxquels les locuteurs convergent, s'alignent (ou pas) pour répondre au principe de coopération qui régit toute conversation. En effet, la conversation se définit notamment par un très haut degré de coopération qui n'exclut cependant pas des moments de tension, voire de compétition. Mais les participants à une conversation n'ont de cesse de collaborer pour construire ensemble à la fois le sens mais aussi la relation qui définit l'interaction en cours. Cette coopération est visible au travers notamment de procédés discursifs spécifiques, et renvoie à la question centrale du ménagement des faces en présence qui conditionne l'activité de régulation des discours.

Les backchannels - signaux multimodaux tels que *mh, ouais, hochement de tête, sourire*, etc. émis par l'interlocuteur en dialogue - participent à cette régulation en informant sur le processus d'écoute et de compréhension des discours [8]. Les hétéro-répétitions jouent dans ce cadre un rôle crucial. Certains auteurs, à l'image de Laforest [9], les considèrent comme une catégorie complexe[2] de backchannels en raison notamment des fonctions pragmatiques qui leur sont associées. Il existe diverses classifications fonctionnelles des signaux backchannels, dans lesquelles les deux classes de « continuer » (rôle d'accusé-réception) et d'« assessment » (rôle de prise de position) demeurent consensuelles [12]. Plus exhaustive et détaillée, la classification de Maynard [10] ajoute une fonction de compréhension, une fonction de support/adhésion au discours, et une fonction de demande de confirmation. La typologie établie par Perrin et al., dans [11] confirme un fonctionnement très similaire des hétéro-répétitions en français québécois à travers les 4 fonctions suivantes :

– « a taking into account function », signalant que l'interlocuteur a correctement entendu et interprété le discours précédent,
– « a confirmation request function », signalant un problème dans le discours,

[1] Imitation in speech : From sensori-motor integration to the dynamics of conversational interaction. http ://lpl-aix.fr/projet/169

[2] L'auteur oppose les régulateurs simples (*mh, ouais*, etc) qui ne sont pas considérés comme de réels tours de parole (au sens d'apport informatif) aux régulateurs complexes qui renvoient non seulement aux répétitions mais aussi aux reformulations, complétions, et métaquestions dont le statut de non tour s'avère plus délicat à établir.

– « a positive reply function », qui signale l'accord de l'interlocuteur avec le discours précédent,
– « a negative reply function », signalant le désaccord.

Dans une étude pragmatique et prosodique, [3] ont pu mettre en évidence une fonction spécifique des hétéro-répétitions qui joueraient un rôle fondamental dans la construction collective de séquences discursives humoristiques. Nous avons pour but d'étudier les hétéro-répétitions pour mettre ainsi à jour de nouvelles fonctions pragmatiques susceptibles de mieux caractériser certains moments de convergence dans l'interaction.

Dans la section suivante, nous présentons les critères formels de définition des répétitions sur la base desquels l'outil a été conçu. La prochaine section décrit la méthode implémentée pour la recherche automatique. Nous présentons enfin brièvement le corpus pour lequel l'outil a été développé, puis une évaluation de ce dernier, sur l'un des dialogues de ce corpus, et des exemples illustrant les catégories fonctionnelles proposées dans [11].

2. CRITÈRES FORMELS DE DÉFINITION DES RÉPÉTITIONS : ÉTAPE PRÉALABLE

Une observation préalable d'un dialogue a permis d'identifier les critères formels définitoires des hétéro-répétitions sur lesquels reposera l'implémentation de l'outil de recherche automatique. L'identification a été faite sur une base lexicale (répétition de mots), le long d'un continuum allant des répétitions lexicales à l'identique (répétition verbatim ou écho strict) aux répétitions partielles ou écho partiel (modification d'un item, changement de pronom, ajout d'une particule discursive, etc.). L'outil automatique prend ainsi en considération le fait qu'une répétition est en écho strict ou non (section 3.2). Les phénomènes de reformulation, parfois associés aux phénomènes de répétition, ont en revanche été éliminés. Un autre critère d'identification concerne la taille de la répétition. Nous nous sommes demandé dans quelle mesure on peut considérer que la reprise d'un seul mot constitue une réelle répétition. L'une des options a été de considérer comme relevant d'une répétition un mot qui s'avère relativement rare dans le dialogue. Nous verrons que l'outil de recherche automatique a permis d'affiner cette notion puisqu'elle a impliqué la prise en compte de ce critère de fréquence de mot chez l'un et l'autre des locuteurs (section 3.2). Enfin cette exploration manuelle préalable nous a permis de mettre à jour un autre critère définitoire des répétitions souvent omis lorsque l'on définit les répétitions comme étant la reproduction presque immédiate de l'énoncé précédent. Il s'agit de l'empan de discours sur lequel la répétition doit être recherchée. Nos premières observations ont permis de mettre à jour des répétitions (verbatim notamment), relativement longues, sur un empan assez important du discours, à savoir sur plusieurs tours de parole. L'outil automatique intègre ce critère fondamental (section 3.1).

3. RECHERCHE AUTOMATIQUE DES HÉTÉRO-RÉPÉTITIONS

Les travaux présentés dans [1] concernent la détection et la correction automatique des auto-répétitions dans un contexte de dialogue homme-machine. Une corresondance lexicale systématique permet de détecter un ensemble initial de répétitions candidates. Celles-ci sont ensuite analysées avec des informations de niveaux syntaxique, sémantique et acoustique afin de distinguer les véritables répétitions des fausses détections. Si la démarche est intéressante, la différence majeure avec le travail que nous présentons réside dans le fait que les répétitions recherchées dans [1] sont intra-locuteurs, par exemple « show me *flighths* daily *flights* to boston ». Nous proposons une recherche des énoncés « source » (énoncés sur lesquels s'ancrent les énoncés répétés), filtrés selon des règles incluant des statistiques, qui reprennent et systématisent les critères identifiés manuellement.

3.1. Recherche systématique

Pour effectuer la recherche systématique des répétitions, nous nous appuyons sur la segmentation du dialogue en IPUs - Inter-Pausal Units, blocs de parole bornés par des pauses silencieuses de 200 ms, ainsi que sur leur alignement sur le signal qui donne leur localisation temporelle. La recherche des hétéro-répétitions consiste à mettre en correspondance :
– les mots d'un locuteur d'une IPU donnée, qui correspond à une localisation temporelle dans le signal ;
– avec les mots de l'autre locuteur, sur une IPU de localisation temporelle proche et jusqu'à un nombre N d'IPU suivantes.

En théorie, toutes les IPUs qui ne sont pas sélectionnées par cette recherche systématique ne peuvent pas contenir de répétitions puisque les deux locuteurs n'emploient aucun vocabulaire commun. Or, nécessairement, en dialogue, un énoncé produit à la première personne par exemple sera répété par l'interlocuteur à la seconde, ce qui en fait des répétitions authentiques mais qui répondent aux contraintes du dialogue. Ainsi, dans le but d'obtenir un taux de faux rejets minimum, une lemmatisation permet de capter des répétitions qui ne l'auraient pas été par l'utilisation des mots orthographiés. Elle permet en effet de mettre en correspondance des verbes conjugués différemment par les deux locuteurs, des mots au singulier ou pluriel, tous les pronoms sujets sont ramenés à la forme *il*, etc. La lemmatisation permet d'une part de trouver des répétitions supplémentaires, d'autre part de renvoyer des segments plus longs, comme dans l'exemple ci-après. L'utilisation des mots permettrait de renvoyer une répétition de 2 mots :

source : non j' ai *pas voulu* non
écho : tu avais *pas voulu*

La lemmatisation permet de renvoyer une suite de 4 lemmes :

source : non *il avoir pas vouloir* non
écho : *il avoir pas vouloir*

Dans la suite de ce document, nous considérons non pas le lexème en tant que tel mais le lemme ; le « mot » fera référence à la lemmatisation de celui-ci.

Avec un empan temporel suffisamment large, cette première étape produit l'ensemble des répétitions possibles. L'intérêt d'un outil automatique réside aussi en sa capacité de ré-itérer les exécutions afin de déterminer les valeurs les plus appropriées. Pour notre corpus, après essai de différentes valeurs, nous avons choisi de chercher dans les 3 IPUs qui suivent l'émission. Cependant, les IPUs sélectionnées contiennent de nombreuses correspondances qui ne sont pas des répétitions (utilisation par les deux locuteurs du même déterminant, d'un même article par exemple). De fait, la seconde étape filtre les répétitions repérées.

3.2. *Filtrage des répétitions*

La sélection repose sur l'utilisation de deux règles :

Règle 1 Une répétition candidate est acceptée dans la mesure où elle contient au moins un mot pertinent du point de vue du locuteur qui répète.

Règle 2 Une répétition longue (au moins K mots) composée seulement de mots non pertinents sera acceptée, dès lors qu'elle est répétée telle quelle (en écho strict). Par expérience, la valeur optimale de K, pour notre corpus, est 3.

La règle 1 nécessite de préciser la notion de mot pertinent d'un locuteur. Introduire ce critère pour déterminer si une répétition doit être sélectionnée est essentiel. Sans ce critère, beaucoup de répétitions de mots considérés comme « usuels » seraient conservées. Il s'agit donc d'établir, soit une liste de mots pertinents, soit au contraire, une liste de mots non-pertinents. Il serait possible d'utiliser un lexique *a priori* des mots outils de la langue, par exemple, comme cela est déjà fait pour l'écrit. Cependant, nous avons constaté que certains mots peuvent être pertinents pour un locuteur, sans l'être pour un autre. De même, certains mots généralement considérés comme mots outils s'avèrent pertinents pour un locuteur car très peu employés. Ceci sera illustré dans la section relative à l'évaluation. Nous avons alors opté pour une sélection dynamique des mots pertinents : une liste est établie pour chaque locuteur, dans chaque dialogue.

On note $N_l(w)$, le nombre d'occurrences du mot w prononcé par le locuteur l, et $|V_l|$, la taille du vocabulaire (nombre de mots différents) du locuteur l. On définit enfin $P_l(w)$, la probabilité du mot w pour le locuteur l par :

$$P_l(w) = \frac{N_l(w)}{\sum_i^{|V_l|} N_l(w_i)}$$

On dira qu'un mot w est pertinent pour un locuteur l, si :

$$P_l(w) \leq \frac{1}{\alpha \times |V_l|}$$

Le coefficient α pourra être déterminé de façon empirique selon le type de corpus utilisé. Dans nos expériences, nous avons utilisé $\alpha = \frac{1}{2}$. D'un certain point de vue, on peut considérer qu'un mot est pertinent s'il n'est pas très fréquent pour ce locuteur, au regard de la richesse de son vocabulaire.

L'utilisation conjointe de règles simples et de statistiques directement déduites des IPUs de chaque locuteur constitue ainsi une solution simple pour filtrer automatiquement les hétéro-répétitions.

4. CONTEXTE APPLICATIF

4.1. *Description du corpus*

Nos données sont extraites du CID, corpus conversationnel composé de 8 dialogues. Chacun de ces dialogues dure une heure et implique deux participants. Un travail important autour de ce corpus a été l'élaboration d'un schéma d'annotation pour l'ensemble des niveaux de l'analyse linguistique (morpho-syntaxique, syntaxique, phonétique, prosodique, discursive) et des différentes modalités impliquées dans le dialogue (audio et mimo-gestuel) [2]. Nous nous limiterons à présenter ici le premier niveau d'annotation qu'est la transcription étant donné que l'outil d'identification des hétéro-répétitions se fonde exclusivement sur ce niveau. Sans entrer dans le détail, soulignons que cette transcription a été effectuée pour traiter des données multi-niveaux. Nous avons donc opté pour une transcription dite orthographique enrichie (TOE) afin de garantir à la fois un meilleur rendement de l'alignement phonétique nécessaire à l'analyse des module phonétique et prosodique, et le meilleur rendement pour les modules morpho-syntaxique et syntaxique qui ne peuvent être faits que sur des transcriptions orthographiques standard (utilisation de lexiques, etc.). Le CID a été automatiquement segmenté en IPUs ; chacune des 12950 IPUs est alors alignée phonétiquement indépendamment, ce qui évite de propager d'éventuelles erreurs au delà d'une IPU. Les transcriptions, effectuées sur les 8 heures, se sont déroulées en deux phases : la première a consisté en la transcription et la correction des dialogues par 2 annotateurs, la seconde a été effectuée par un expert qui a corrigé une nouvelle fois l'ensemble des transcriptions.

L'évaluation des performances de l'outil automatique est réalisée manuellement, elle porte sur l'un des dialogues du CID. Néanmoins, ce dialogue contient un nombre suffisant d'hétéro-répétitions pour que cette évaluation soit pertinente. Les locuteurs sont mentionnés par AB et CM. La table 1 présente quelques valeurs textométriques pour ces locuteurs. On constate que le vocabulaire de AB est plus riche que celui de CM. En effet, même si AB intervient moins dans le dialogue (nombre d'occurrences inférieur) que CM, son vocabulaire est plus large et le nombre d'hapax[3] est également plus élevé.

TAB. 1: Description du corpus en terme de vocabulaire

Loc.	Vocab.	Occ.	Hapax	Occ. max	Mot max
AB	1183	6619	671	184	on
CM	1066	7681	561	501	ouais

Selon la définition de la pertinence, proposée en section 3.2, un mot est pertinent s'il est apparu moins de 16 fois pour AB, 20 fois pour CM. Ainsi, le mot *petit*, présent 21 fois pour AB, mais 8 fois pour CM, est pertinent seulement pour CM. Inversement, le mot *voilà* est prononcé 8 fois par AB et 27 fois par CM ; il est donc pertinent seulement pour AB. De même, le mot *vachement* est apparu 2 fois pour AB, 20 fois pour CM. Des mots comme *le, de, pouvoir, truc* ne sont pas pertinents pour les deux locuteurs.

4.2. *Evaluation du système automatique*

La valeur de rappel a été obtenue par un expert qui a examiné manuellement toutes les répétitions trouvées dans l'empan temporel choisi. Comme nous le souhaitions, notre système a obtenu une valeur de rappel égale à 1 (aucune répétition omise). La valeur de précision a également été évaluée sur le dialogue AB-CM, grâce à l'analyse manuelle de deux experts sur les répétitions filtrées.

La table 2 indique le nombre de répétitions trouvées puis filtrées par le système automatique ainsi que le nombre d'entre-elles qui ont été validées manuellement par les experts. Dans le cas AB émet - CM répète, la précision du

[3] mots présents une seule fois

système automatique est de 80,17 %. Dans le cas AB répète - CM émet, elle est de 74,79 %. Il est important de noter que les IPU présentent de nombreux chevauchements de localisations, et dans ce cas, la même répétition est repérée dans les 2 sens .

TAB. 2: Répétitions automatiques et évaluation

Locuteurs	Nombre de répétitions		
	trouvées	filtrées	validées
AB émet - CM répète	860	116	93
AB répète - CM émet	1251	119	89

L'évaluation de cet outil est donc satisfaisante au regard des objectifs pour lesquels il a été conçu. Ses performances ne sont pas remises en question par les experts. Les différences sont en effet liées à des critères non automatisables. Par exemple, une répétition est pertinente si elle a un caractère ostensif, c'est-à-dire qu'elle est le fruit d'une intention du locuteur qui manifeste par cette répétition son intention de citer les mots en question [11]. En outre, un tel critère suppose que la répétition comporte une réaction de l'interlocuteur sur ce qu'il cite.

4.3. Illustrations

Cette section présente quelques exemples de répétitions que l'outil a trouvé dans le dialogue AB-CM. Le premier exemple montre une répétition proposée par l'outil de détection automatique, car `oui c' était` est une répétition de 3 mots non pertinents en écho strict (application de la règle 2). Celle-ci n'a cependant pas été validée par l'expert :

CM *oui c' était* `un insolite donc un peu malaise quand même quoi un peu désagréable quoi quand même ouais`

AB *oui c' était* `pas on n' était pas tristes mais on était on était` *coincés* `obligés alors qu' on était en vacances et`

CM `ouais ouais ouais ouais ouais`

CM *coincée* `ouais`

Par ailleurs, cet exemple montre à nouveau l'intérêt d'utiliser des lemmes, car ils permettent de repérer la répétition du verbe `coincer`. L'exemple ci-après illustre également l'intérêt d'utiliser des lemmes :

CM `et` *il contrôlait pas*

AB *il* `a` *pas contrôlé*

L'exemple suivant montre un cas où la répétition est formulée sur deux IPUs.

CM `c'était La Rochelle`

AB `non` *c'était à Poitiers*

CM *c'était à*

CM *à Poitiers* `tu as été` *à Poitiers* `ouais`

CM `ouais`

Enfin, si la grande majorité des répétitions détectées puis validées par les experts concernent moins de 6 mots, nous en avons également identifiées de très longues comme celle reproduite ci-dessous qui présente un écho strict de 11 items :

CM *et y a de la neige qui était rentrée dans la* `chambre et`

AB *et y a de la neige qui était rentrée dans la* `fenêtre oui un peu`

5. CONCLUSION

Cet article concerne la recherche des hétéro-répétitions en conversation. La démarche qui nous a amené à les identifier ainsi que l'outil automatique qui permet de les repérer ont été présentés. De nombreux exemples, ainsi qu'une évaluation de la méthode, permettent de valider l'approche automatique ainsi que la pertinence de son utilisation au sein d'un dialogue oral spontané. A court terme, cet outil permettra de constituer une base de données conséquente d'hétéro-répétitions puisque nous souhaitons l'utiliser sur plusieurs types de corpus (comme les débats, les entretiens, les dialogues orientés-tâche, etc.) en vue de comparer les répétitions en termes formels et fonctionnels.

RÉFÉRENCES

[1] J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *30th annual meeting on Association for Computational Linguistics*, pages 56–63, Newark, Delaware, 1992.

[2] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. Le cid - corpus of interactional data. *Traitement Automatique des Langues*, 49(3) :105–134, 2008.

[3] R. Bertrand and B. Priego-Valverde. Does prosody play a specific role in conversational humor ? *Pragmatics and Cognition*, 18(2), 2010.

[4] M-W. Casby. A pragmatic perspective of repetition in child language. *Journal of Psycholinguistic Research*, 15(2) :127–140, 1986.

[5] H. Chiung-chih. Other-repetition in mandarin child language : A discourse pragmatic perspective. *Journal of Pragmatics*, doi :10.1016/j.pragma.2009.08.005, 2009.

[6] J. Cole, M. Hasegawa-Johnson, C. Shih, H. Kim, E-K. Lee, H-Y. Lu, Y. Mo, and T-J. Yoon. Prosodic parallelism as a cue to repetition disfluency. In *Disfluency in Spontaneous Speech Workshop*, Aix-en-Provence, France, 2005.

[7] T-S. Curl, J. Local, and G. Walker. Repetition and the prosody-pragmatics interface. *Journal of Pragmatics*, 38 :1721–1751, 2006.

[8] J-E. FoxTree. Listening in on monologues and dialogues. *Discourse Processes*, 27(1) :35–53, 1999.

[9] M. Laforest. Le back-channel en situation d'entrevue. *Québec : CIRAL/Recherches sociolinguistiques*, 2, 1992.

[10] S. Maynard. Japanese conversation : Self-contextualization through structure and interactional management. Ablex, Norwood, NJ, 1989.

[11] L. Perrin, D. Deshaies, and C. Paradis. Pragmatic functions of local diaphonic repetitions in conversation. *Journal of Pragmatics*, 35 :1843–1860, 2003.

[12] E-A. Schegloff. Discourse as an interactional achievement : Some uses of uh huh and other things that come between sentences. *Analyzing discourse : Text and talk*, pages 71–93, 1982.

Démarcation lexicale en français : profils prosodiques sur grand corpus

Rena Nemoto[1], *Martine Adda-Decker*[1], *Jacques Durand*[2]

[1]LIMSI/CNRS (UPR351)
BP 133 F-91403 Orsay Cedex, France
{madda,nemoto}@limsi.fr, http ://www.limsi.fr
[2]CLLE-ERSS/CNRS UMR 5263
Université de Toulouse-Le Mirail, 5, allées Antonio Machado F-31058 Toulouse Cedex 9, France
jacques.durand@univ-tlse2.fr, http ://w3.erss.univ-tlse2.fr

ABSTRACT

The goal of this paper was to investigate whether regularities of French fundamental frequency (f_0) profiles can contribute to localize word boundaries. 13 hours of broadcast news speech were investigated using automatic processing (lexical and phonemic alignment, f_0 extraction, parts of speech (POS) tagging). Average f_0 profiles were calculated on word and phrase (determiner-noun) levels and examined according to their syllabic lengths. f_0 profiles of lexical words and phrases presented word-final accentuation. For a given syllabic length, nouns and noun phrases featured very different f_0 profiles indicating word boundary information. In particular, the `determiner noun` f_0 profiles showed word-initial accentuation on lexical words after a determiner. Further phrase types, speaking styles and languages will be investigated in future studies.
Keywords: fundamental frequency, phrasing, French, broadcast news

1. Introduction

Est-ce que le mot classique de l'écrit correspond à une entité repérable à l'oral ? Contrairement à l'écrit, le mot à l'oral s'inscrit dans un signal acoustique qu'on qualifie de continu. Bien que le signal de parole comporte de nombreuses discontinuités, elles ne correspondent en général pas aux frontières de mot, mais sont plutôt induites par des propriétés intrinsèques des sons élémentaires, typiquement des plosives, ou de leurs enchaînements. Comment pouvons-nous alors extraire les frontières de mot d'un tel signal ? Quels indices peuvent être perçus ? Les travaux empiriques récents démontrent que la notion classique de mot n'est pas suffisante d'un point de vue cognitif. Les locuteurs peuvent autonomiser des unités qui peuvent être plus petites que le mot classique et inversement ils semblent mémoriser des unités plus larges que le mot classique On peut donc considérer la structure habituellement attribuée par les linguistes au lexique mental comme une approximation utile et, en fait, indispensable à l'appréhension de diverses régularités observables. En même temps, il semble évident que les relations multidimensionnelles entre les entités lexicales sont plus complexes que les divisions traditionnelles considérées en linguistique. En reconnaissance automatique de la parole, les hypothèses de frontière de mot sont induites davantage par les niveaux supérieurs (le lexique et la cooccurrence des mots), que par des indices acoustiques. En nous appuyant sur les instruments issus du traitement automatique de la parole, nous nous proposons d'examiner à grande échelle les indices portés par la langue parlée, permettant de faire des hypothèses fortes sur les frontières de mot. Ces indices peuvent se situer à un niveau abstrait, comme par exemple les niveaux morpho-phonologique et phonotactique, ou à un niveau du signal physique, dans les réalisations acoustico-phonétique et prosodique des frontières de mots [3]. Nous allons concentrer nos efforts sur ces derniers indices acoustiques. Les questions qui sous-tendent ce travail sont alors les suivantes : (i) existe-t-il des indices acoustiques permettant de faire l'hypothèse de frontière de mot ? (ii) si oui, quels sont ces indices ?

2. Démarcation lexicale en français

Il y a une longue tradition de négation de l'importance du mot en français au niveau phonique. Cette tradition remonte au moins aux observations de Palsgrave [9] qui soulignait que le français est une langue où les consonnes finales ou de liaison s'enchaînent pour des raisons euphoniques aux mots qui suivent s'ils commencent par une voyelle et où l'on forme des groupes rythmiques de cinq ou six mots qui sont traités comme s'ils n'en formait qu'un. À une époque plus récente, Pulgram ([10],[11]) a établi une opposition entre les langues Nexus (langues germaniques comme l'anglais) et les langues Cursus (comme le français). On peut noter par ailleurs que dans la tradition prosodique héritée de Nespor et Vogel [8], le mot phonologique en français ne correspond pas au mot orthographique puisque les préfixes sont séparés des bases par une frontière de mot. On ne s'étonnera donc pas que certains spécialistes aillent même jusqu'à défendre la thèse que le mot en français n'est qu'une projection de la norme orthographique (Laks [7]). Si le mot lexical n'a pas la prégnance en français qu'il peut avoir dans d'autres langues, on peut néanmoins s'interroger sur sa négation complète au niveau phonologique. Si on examine la phonologie et la phonétique du français de plus près dans la diversité des usages et des variétés, il y a néanmoins un ensemble de faits qui militent en faveur de la thèse que les mots lexicaux (noms, adjectifs, verbes, adverbes) sont signalés de diverses façons. Plusieurs arguments peuvent être tirés de la phonotactique. En français standard, on sait par exemple que si on rencontre les séquences /tl/ ou /dl/ on n'est pas en début de mot lexical en dehors de quelques noms propres empruntés (e.g. **Tl**emcen) à d'autres langues. Il nous a donc semblé important de vérifier à grande échelle si les mots lexicaux n'avait

pas plus de prégnance que ne le prétend la tradition. Ainsi, différents chercheurs ont pu montrer qu'une montée [6] ou un coude [14] de f_0 représente souvent un indice robuste de début de mot lexical. Dans ce travail nous nous intéressons plus particulièrement au lien entre mots et contours prosodiques. Dans la section suivante nous décrivons le corpus ainsi que la méthodologie mise en œuvre. Dans la section 4 nous présentons des premiers résultats en termes de profils prosodiques moyens.

3. Corpus et méthodologie

3.1. Corpus

Ce travail part de la parole journalistique du corpus TECHNOLANGUE-ESTER [4] transcrit manuellement. Nous utilisons un total de 13 heures effectives de parole avec 165k mots (tokens) et 14k entrées lexicales (types) distinctes, provenant de locuteurs masculins. Le style de parole est dit « préparé » pour une grande partie du corpus.

Tab. 1: Description quantitative du corpus en termes de mots (tokens) de longueur syllabique n, pour $n = 0 - 4$. Les comptes sont séparés en fonction de schwa final réalisé (s=1,bas) ou non (s=0,haut). Concernant *class syll.* : n=longueur syllabique du mot ; s=présence(1)/absence(0) d'un schwa final.

n	*class syll.* n_s	*#mots*	*exemples*
0	0_0	12.6k	`l'` ; `d'` ; `de`
1	1_0	72.2k	`vingt` ; `reste`
2	2_0	36.0k	`beaucoup` ; `journal`
3	3_0	16.0k	`notamment`
4	4_0	6.1k	`présidentielle`

n	*class syll.*	*#mots+ /ə/*	*exemples*
0	0_1	12.3k	`de` ; `le` ; `que`
1	1_1	39.2k	`reste` ; `test`
2	2_1	2.1k	`ministre`
3	3_1	0.7k	`véritable`
4	4_1	0.2k	`nationalistes`

3.2. Méthodologie

Nous décrivons les procédures de traitement et d'annotation effectuées permettant les analyses proposées sur la démarcation lexicale (Figure 1).

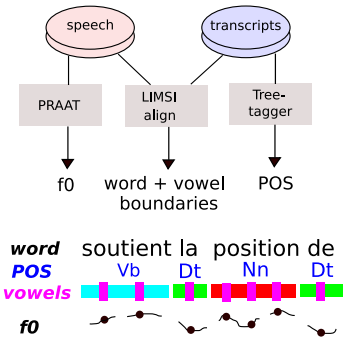


Fig. 1: Traitement : chaque voyelle est étiquetée par une valeur moyenne de f_0, par sa durée, par le rang à l'intérieur du mot et par sa POS.

Alignement lexical et phonémique : Les données ont été alignées automatiquement avec le système du LIMSI [5] produisant des segmentations en mots et en phones. Le dictionnaire de prononciation permet des schwas optionnels en fin de mot dès lors que sa prononciation se termine par une consonne.

Annotation en POS : Le corpus transcrit est étiqueté en parties du discours (**POS** : **P**art **O**f **S**peech) à l'aide de TREETAGGER [12], afin de pouvoir localiser des syntagmes pertinents.

Longueur syllabique des mots ; rang syllabique des voyelles : À chaque mot est associé une longueur syllabique, qui correspond au nombre de voyelles de sa prononciation (hors schwa final). Ainsi le mot `président` est de longueur syllabique 3, car il comporte 3 voyelles /e/, /i/, et /ã/ respectivement de rangs syllabiques 1, 2 et 3 (cf. table 1). Un schwa final réalisé ne changera pas cette longueur syllabique, mais pour les mesures ultérieures, nous considérons cependant les mots se terminant par un schwa dans une classe prosodique à part. Chaque voyelle (en dehors de schwa final) est annotée par son rang syllabique. Longueur et rang syllabiques permettront d'organiser des classes de mots supposées pertinentes d'un point de vue prosodique en général, et de l'intonation en particulier. La table 1 montre la composition du corpus en termes de mots de longueur syllabique n. Les mots à une syllabe sont les plus fréquents, et la fréquence des mots diminue en fonction de la longueur syllabique. La partie inférieure correspond à des mots se terminant par un schwa réalisé.

Contour de f_0 ; valeur de f_0 par segment vocalique : Les contours de f_0 ont été extraits avec PRAAT [2]. La fréquence fondamentale du signal a été mesurée toutes les 5 ms comme dans [1]. Pour chaque segment phonémique, dont les bornes proviennent de l'alignement automatique, un *taux de voisement* est défini comme le rapport entre le nombre de mesures où f_0 est défini par le nombre total de mesures (qui correspond donc à la durée du segment (en ms) divisé par 5). Si le *taux de voisement du segment* est supérieur à 70%, nous faisons l'hypothèse qu'à la fois la prononciation, la segmentation automatique et l'extraction de f_0 sont correctes, et la valeur de fréquence fondamentale qui lui est affectée correspond à la moyenne des mesures. Cette valeur sera ensuite utilisée dans le calcul des profils de f_0. Le taux de voisement sert de filtre (toutes les voyelles du mot doivent satisfaire la contrainte des 70%) et nous rejetons ainsi environ 10% des mots. Les mesures en Hz sont converties en demitons en normalisant par rapport à une fréquence de référence arbitraire de 120 Hz, souvent considérée comme la fréquence moyenne des voix masculines. Si nous nous intéressons aux hauteurs ou écarts perçus, différents travaux ont montré par exemple que des différences de 3 demitons jouent un rôle dans les situations communicatives, des différences plus faibles pouvant cependant déjà contribuer à la perception de la démarcation lexicale [13].

Profil de f_0 : À chaque segment de voyelle est associée une valeur de f_0, qui correspond à la moyenne de trois mesures relevées au centre du segment vocalique. Nous pouvons alors calculer pour chaque mot un profil de f_0, qui correspond simplement à un contour de f_0 stylisé, obtenu en reliant par des segments de droite

les valeurs de f_0. Ce profil peut ainsi être calculé pour chaque mot ou pour des classes de mots.

Dans le corpus ainsi préparé, à chaque mot prononcé est associée sa transcription orthographique et phonémique avec durées correspondantes, ainsi que sa partie du discours. Chaque voyelle est annotée avec les informations : f_0 moyen, durée et rang syllabique dans le mot. Ces annotations sont ensuite exploitées pour les calculs de profils prosodiques de mots et syntagmes.

4. Travail expérimental et résultats

Nous considérons des classes de mots de même longueur syllabique (distinguant les cas avec ou sans schwa final). A l'intérieur d'une classe, les mesures de f_0 de voyelles de même rang sont moyennées. Afin d'examiner les indices prosodiques portés par le signal de parole sur les frontières de mot, nous calculons ensuite des profils moyens de f_0 de mots et de classes de mots (e.g. la classe de tous les mots de la même longueur syllabique sans schwa final réalisé) et de quelques syntagmes particuliers, en particulier le syntagme nominal. Dans cette étude, nous avons limité l'étude à des mots sans schwa final.

4.1. Profils de f_0 moyens

Mots lexicaux Nous nous proposons de calculer les profils de f_0 moyens pour tous les mots lexicaux d'au plus 4 syllabes (et de 2 syllabes pour les mots grammaticaux). La figure 2 montre les profils moyens stylisés pour les mots lexicaux (sans schwa final) en fonction de leur longueur syllabique (centrés à droite : les fins de mots se superposent dans la figure). A partir des profils moyens de la figure 2, nous pourrons faire des observations et élaborer quelques hypothèses sur l'intonation du français de notre corpus journalistique. Pour des raisons de place nous ne mettons pas les profils pour les mots à schwa final. Pour ces derniers, les profils sont similaires à ceux présentés, si on néglige le schwa final. Pour la syllabe additionnelle à schwa final, la f_0 moyenne descend plus bas que la f_0 de la syllabe pénultième de la figure 2. Pour les mots grammaticaux, les f_0 moyennes mesurées sont proches de celles des syllabes initiales de mots lexicaux. La figure 2 montre ainsi les profils moyens des classes de mots de longueurs syllabiques n, sans schwa final. On peut faire les observations suivantes :

(i) La $f0$ moyenne est nettement plus élevé pour la syllabe finale n, que pour les syllabes antérieures (pénultièmes, antépénultièmes...).
(ii) Pour les mots trisyllabiques et plus, la différence de $f0$ entre 2 voyelles consécutives est maximale pour les syllabes pénultième et finale ; ce *delta* tend à croître avec la longueur syllabique du mot.
(iii) La f_0 moyenne des monosyllabes est au moins aussi haute que celle de la finale des mots plus longs.
(iv) L'accentuation initiale ne se manifeste que faiblement sur les profils moyens.

Syntagme nominal Dans cette partie, nous étendons la mesure du profil de f_0 moyen au syntagme nominal, où nous nous limitons aux bigrammes de type `déterminant nom`. Est-ce que le profil moyen d'un syntagme `déterminant nom` de longueur n se distingue de celui d'un nom de longueur n ? Si oui, cette différence pourrait alors contribuer à la démarcation lexicale. Nous ne présentons que les bigrammes sans schwa final. La figure 3 montre à gauche les profils moyens des mots étiquetés `nom` (31k occ.), très similaires à ceux de la figure 2. À droite se trouvent les profils moyens des syntagmes nominaux sélectionnés (13k occ.). Pour chaque profil moyen on peut observer que, contrairement aux observations concernant les mots lexicaux seuls, la différence de hauteur est maximale entre le début du syntagme (ici déterminant monosyllabique) et la fin (syllabe finale du nom). L'accent est marqué sur la première syllabe d'un mot nominal suivi par un déterminant. Ces résultats suggèrent que les minima des profils de f_0 (à l'intérieur d'une fenêtre temporelle de quelques syllabes) permettent de repérer des frontières de syntagmes, au moins pour le cas des syntagmes nominaux de type `déterminant nom`. Des études plus détaillées sont en cours pour étayer ces premières observations.

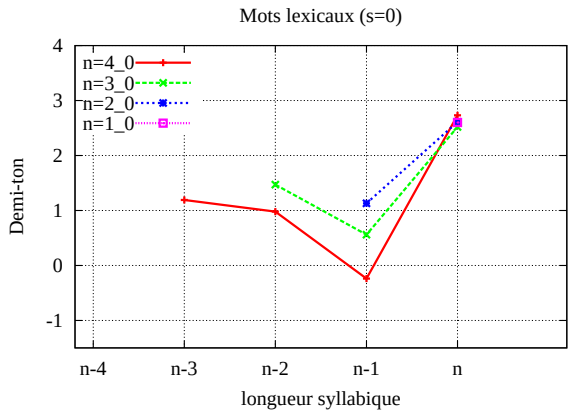


Fig. 2: Profils de f_0 moyens des mots lexicaux pour des longueurs syllabiques n (1-4 syll.).

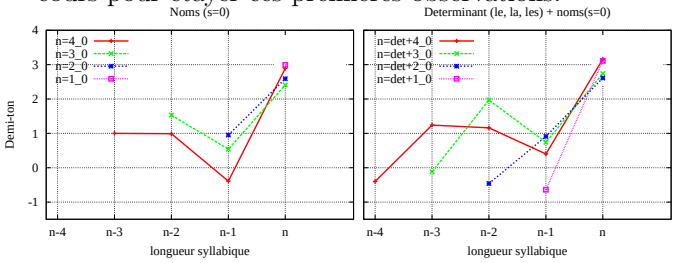


Fig. 3: Profils de f_0 moyens pour des longueurs syllabiques n. **Gauche :** `nom` **Droite :** syntagme `déterminant nom`

4.2. Distributions de Δf_0 inter-vocalique

Les profils moyens des substantifs et des syntagmes nominaux montrent des tendances moyennes. Afin de vérifier si ces tendances traduisent l'évolution de f_0 effectivement réalisée, nous calculons les distributions des différences de f_0 (Δf_0) entre deux voyelles consécutives. Si la f_0 du déterminant correspond à un minimum, la Δf_0 doit être négative pour une proportion importante d'échantillons ; s'il y a un accent initial sur le début des mots polysyllabiques, une proportion élevée d'échantillons doit présenter une Δf_0 positive pour la première voyelle du nom et négative pour la voyelle interne... La variation de f_0 entre deux voyelles consécutives V_{k-1} et V_k est calculée

comme $\Delta f_0(k) = f_0(V_k) - f_0(V_{k-1})$. Ces mesures de Δf_0 en demi-tons (DT) sont divisées en 3 groupes :**Fall**, **Stable** et **Rise.**

Fall	**Stable**	**Rise**
$\Delta f_0 \leq -1$ (DT)	$\Delta f_0 \in]-1 \quad +1[$ (DT)	1 (DT) $\leq \Delta f_0$

La figure 4 (en 4 parties) montre les distributions de Δf_0 pour le syntagme nominal (déterminant monosyllabique + nom n-syllabique, avec $n = 1 - 4$). Environ 80% des voyelles de déterminant présentent une chute de f_0 supérieure à 1 DT par rapport à la voyelle du mot précédent. Ce pourcentage, stable pour les 4 graphes, ne semble pas dépendre de la longueur du nom suivant. Pour le graphe **déterminant - monosyllabe** (haut-gauche), on a également 80% des occurrences avec une montée de f_0 de plus d'un demi-ton sur la syllabe unique du nom. Concernant le graphe **déterminant - bisyllabe** (haut-droite), on peut remarquer des proportions de montée de f_0 aussi importantes sur les deux syllabes (1_2_0 et 2_2_0). Ceci est cohérent avec les profils moyens et des analyses plus fines sur les contextes droits sont prévus. Pour les graphes **déterminant - trisyllabe** (bas-gauche) et **déterminant - 4-syllabe** (bas-droite), on peut observer au moins 60% des échantillons analysés avec une montée de f_0 sur la syllabe initiale du nom confirmant la tendance à l'accent initial, et 60% (respectivement 70%) de chute de f_0 sur les syllabes pénultièmes (2_3_0) (respectivement 3_4_0), préparant ainsi la réalisation d'un accent final.

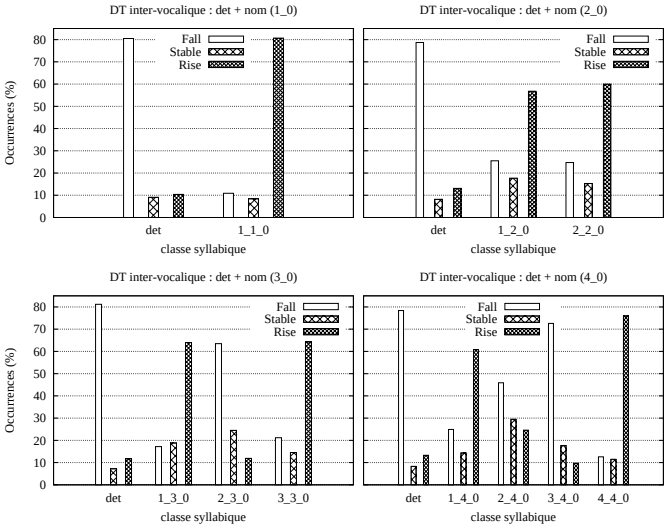


Fig. 4: Histogrammes de f_0 inter-vocalique sur syntagmes `dét nom` de longueur syllabique n. **Haut gauche :** det+monosyllabe. **droite :** det+bisyllabe. **Bas gauche :** det+3-syllabe. **droite :** det+4-syllabe.

5. Conclusion

Dans cette contribution nous avons exploité 13 heures de parole journalistique (165k mots)de voix d'hommes afin d'étudier quelques stratégies de démarcation du mot en français en exploitant conjointement des transcriptions phonémiques et lexicales, ainsi que des annotations prosodiques et morphosyntaxiques. Nous avons présenté une méthodologie reposant sur des profils de f_0 moyen calculés sur des classes de mots et de syntagmes, afin mettre en évidence des régularités sur le contour prosodique qui contribuent à repérer les frontières de mot à l'oral.
Les résultats décrits semblent d'ores et déjà confirmer que le mot est une entité dont la prégnance ne provient pas seulement des conventions orthographiques. L'oral offre des stratégies de démarcation permettant de reconstruire des formes de mot sous-tendant des mots grammaticaux et des lexèmes.
Dans le futur, nous envisageons d'étudier en détail à l'intérieur de contours de f_0 syllabique et d'éteindre d'autres langues et d'autres styles de parole.

6. Remerciements

Les travaux présentés ont été partiellement financés dans le cadre des projets *AMADEO* du RTRA DIGITEO et par le programme de OSEO *Quaero.*

Références

[1] M. Adda-Decker, C. Gendrot, and N. Nguyen. Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *TAL*, 49(3), 2008.

[2] P. Boersma and D. Weenink. Praat : doing phonetics by computer [computer program], from http ://www.praat.org/. Technical report, 2005.

[3] J. Durand. Mot et phonologie en français. In *Colloque Les structures des français en contact*, Université Tulane. La Nouvelle Orléans, Etats-Unis, 26-28 juin 2008.

[4] S. Galliano et al. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proc. Interspeech*, Lisbonne, Septembre 2005.

[5] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk. Where Are We In Transcribing French Broadcast News ? In *Proc. Interspeech*, Lisbonne, 2005.

[6] S.-A. Jun and C. Fougeron. *Intonation : Analysis, Modelling and Technology*, chapter A phonological model of French intonation. Kluwer, 2000.

[7] B. Laks. La liaison et l'illusion. *Langages*, 158 :101–125, 2005.

[8] M. Nespor and I. Vogel. *Prosodic Phonology.* Foris, Dordrecht, 1986.

[9] J. Palsgrave. *L'esclarcissement de la langue françoyse, composé par maistre jehan palsgrave, angloys, natyf de Londres et gradué de Paris.* Honoré Champion 2003, Paris, 1530.

[10] E. Pulgram. Prosodic systems : French. *Lingua, Elsevier*, 13 :125–144, 1965.

[11] E. Pulgram. *Syllable,word,nexus,cursus.* Mouton, La Haye, 1970.

[12] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, 1994.

[13] J. 't Hart. Differential sensitivity to pitch distance, particularly in speech. *Journal of Acoustical Society of America*, 69(3) :811–821, 1981.

[14] P. Welby. The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication*, 49 :28–48, 2007.

Détection semi-automatique des syllabes proéminentes avec une segmentation automatique en pseudo-syllabes

Philippe Martin

CLILLAC-ARP EA3967, UFRL, Université Paris Diderot
philippe.martin@linguist.jussieu.fr

ABSTRACT

To assert a prominence character to syllables is pivotal to evaluate prosodic theories, especially those which predict the phonetic features of melodic contours (rise, fall, height, etc.) located on those syllables. The advent of large corpora of spontaneous speech induced the elaboration of automatic prominence detectors, usually requiring syllabic segmentation. In order to avoid the sometimes problematic segmentation into phonetic units, we introduce an algorithm for prominence detection operating without syllabic segmentation on readily available phonetic properties of speech, at the exemption of spectral properties.

1. INTRODUCTION

La détection des syllabes proéminentes est une opération importante en analyse prosodique. La plupart des théories phonologiques actuelles décrivent les événements prosodiques principalement à l'endroit des syllabes accentuées, où ces événements se manifestent par des traits acoustiques de niveau, de montées ou de descentes mélodiques.

D'autre part, en parole spontanée, les corpus analysés atteignent des durées considérables qui rendent difficiles sinon impossible un traitement manuel exhaustif. Aussi la détection automatique des proéminences syllabiques tend de plus en plus à se substituer à l'identification perceptive par des opérateurs humains. Un processus automatique présente de plus l'avantage de délivrer l'opérateur d'une tâche qui, pour le français du moins, est loin d'être évidente [1], et de mettre en œuvre un processus plus facilement contrôlable qui ne fait pas intervenir des éléments extérieurs aux propriétés acoustiques de proéminences [2]. Des réalisations de détection automatique de syllabes proéminentes ont été proposées récemment et implémentées dans des logiciels tels que Praat [5], [3] ou WinPitch [2].

2. MÉTHODES EXISTANTES

La méthode d'identification de proéminence syllabique la plus répandue à ce jour est le prosogramme [5], qui procède par une stylisation graphique des variations mélodiques à l'intérieur d'un segment selon une valeur de glissando attribuée à ce segment. Si le contour mélodique, décrit par sa durée et sa variation de Fo, présente un glissando inférieur à un certain seuil, on lui substitue une ligne horizontale (donc un niveau) placée aux 2/3 de la variation ; si le glissando est supérieur ou égal au seuil, il est remplacé par une variation linéaire reliant les valeurs aux extrémités du contour. Le seuil de glissando a été établi par Rossi [6] d'abord pour des sons purs puis pour des voyelles synthétiques. Cette représentation intègre donc implicitement un modèle simplifié de la perception combinant la durée et la variation de fréquence fondamentale. À partir du prosogramme, on peut ensuite appliquer des critères de proéminences, soit par inspection visuelle ou automatiquement.

Malgré son usage assez répandu, la détection de proéminence par prosogramme présente plusieurs problèmes :

1. Le seuil de glissando est en fait un paramètre ajustable, afin de tenir compte des variations possibles étant survenues à l'origine lors des tests de perception établissant ce seuil ;
2. La stylisation étant linéaire ne tient pas compte des formes mélodiques convexes, concaves, en cloche, fréquemment observées dans des réalisations régionales ou idiosyncratiques ;
3. Le seuil de glissando est constant pour un enregistrement déterminé, alors que des variantes de réalisation par un même locuteur sont possibles dans des espaces de temps relativement courts ;
4. La fiabilité du procédé dépend de celle de la segmentation en syllabes et en voyelles. Une segmentation automatique doit en général être soigneusement vérifiée et corrigée visuellement (surtout pour de la parole spontanée), ce qui peut s'avérer couteux en temps et diminuer d'autant l'intérêt du traitement automatique.

Diverses réalisations tentent de pallier ces difficultés. Ainsi, des implémentations telles que [7] opèrent un ajustement du paramètre de seuil en prenant en compte les différences de glissando entre syllabes successives plutôt que la valeur liée à un seul segment. D'autres algorithmes [9] procèdent sans segmentation en syllabe de manière à éviter les erreurs fréquentes de la segmentation automatique.

3. PRINCIPES SOUS-JACENTS

3.1. Fenêtre de proéminence

On sait que la proéminence est relative à une certaine fenêtre temporelle de la phrase ou du discours. La

mémoire auditive n'est pas capable de retenir et de classer plus qu'un nombre limité d'occurrences de durées syllabiques, d'intensité ou de fréquence fondamentale, et opère dans une fenêtre limitée dans le temps. Pour un débit normal, le nombre de 7 syllabes a été suggéré [11], ce qui justifie le nombre de pseudo-syllabes choisi par défaut pour une fenêtre de proéminence, mais d'autres valeurs ou une durée temporelle peuvent être également adoptées.

On propose alors une fenêtre de calcul de proéminence, fenêtre glissante ordonnant à chaque pas les pseudo-syllabes en degrés de proéminence décroissante. Le glissement de cette fenêtre permet de s'assurer que le calcul de la moyenne de la proéminence relative s'effectue pendant une durée suffisante. Ceci correspond sensiblement à la procédure manuelle décrite en [4], dans laquelle les opérateurs ont eu la possibilité de réécouter à volonté les séquences précédant et suivant les syllabes testées pour établir leur jugement.

3.2. Durées et variations de Fo

La proéminence perçue résulte d'un mécanisme complexe d'intégration auditive des paramètres acoustiques tels que la durée et la variation de fréquence fondamentale, mais aussi d'événements phonétiques comme la présence d'une consonne voisée en fin de syllabe, ainsi que des événements linguistiques comme l'appartenance de la syllabe à une classe d'unités accentuables (verbes, noms, adjectifs et adverbes). Puisque la méthode présentée n'utilise pas ces informations, les index de proéminence de durée et de Fo sont utilisés séparément à ce stade.

3.3. Proéminence relative

Les valeurs absolues de durée et de changement de Fo ont souvent été interprétées de manière erronée à partir de la stylisation des contours de Fo [5], mais des valeurs relatives sont maintenant couramment employées dans des systèmes récents. Ce principe est utilisé ici par le choix d'un seuil au dessus duquel les pseudo-syllabes sont sélectionnées. Si une transcription phonétique est disponible, la règle des 7 syllabes peut être appliquée, donnant le nombre minimal de syllabes accentuées et donc de pseudo-syllabes proéminentes.

4. IMPLEMENTATION

Pour tenter de résoudre les problèmes évoqués plus haut, le procédé présenté ici opère sans segmentation syllabique préalable, tient compte des variations mélodiques non linéaires, et intègre une propriété des séquences syllabiques de 7 syllabes de comporter au moins une syllabe proéminentes [11].

4.1. Durée

La segmentation en pseudo-syllabes, c'est-à-dire en approximation de syllabes sans segmentation explicite en voyelles et consonnes, est généralement opérée en élaborant un histogramme des valeurs échantillonnées d'intensité. Cet histogramme présente normalement une distribution sensiblement bimodale. On en retient alors les valeurs du mode supérieur.

Une autre méthode consiste à sélectionner les pics d'intensité présentant une chute suffisante (par exemple – 6 dB) sur leurs deux versants [10]. C'est cette dernière méthode qui a été retenue : la durée du segment dont l'intensité est supérieure à ce seuil est considérée comme durée de la pseudo-syllabe. En cas de plateau dans la courbe d'intensité, on utilise alors un point de remontée de la courbe à l'intérieur du plateau et situé à moins de 6 dB du pic.

4.2. Fréquence fondamentale

Les variations de fréquence fondamentale correspondant aux pseudo-syllabes sont ensuite définies comme suit (Fig. 1) : à partir de la valeur de Fo à l'endroit du pic d'intensité, on suit la courbe de Fo vers la gauche et vers la droite jusqu'à ce que la frontière de la pseudo-syllabe soit atteinte, ou que la valeur de Fo soit nulle (fin de voisement) ou invalide. On évite ainsi les valeurs de Fo erronées ou manquantes fréquemment rencontrées en début ou en fin de voisement. Si la valeur de Fo au pic est nulle, la pseudo-syllabe est écartée.

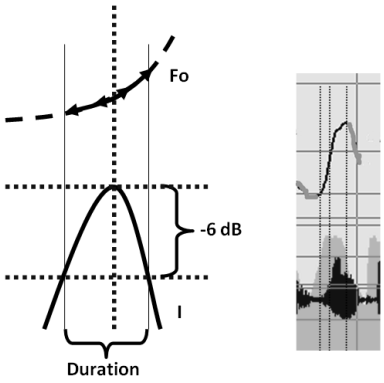


Figure 1 : Durée de la pseudo-syllabe et segment de Fo

Pour prendre en compte les variations non linéaires de Fo, la courbe mélodique est « linéarisée » en prenant des valeurs toutes les 50 ms et en additionnant les valeurs absolues des différences de Fo prises en leurs extrémités. Ainsi, les variations en forme de cloche sont mieux prises en compte (elles seraient quasi annulées si on prenait seulement les valeurs de Fo à leurs extrêmes).

Chaque pseudo-syllabe est donc représentée par deux valeurs : a) la durée à -6 dB d'intensité par rapport au pic et b) la variation corrigée de Fo.

4.3. Fenêtre de proéminence

On définit une « fenêtre de proéminence » glissante, contenant 7 pseudo-syllabes consécutives. Pour chaque fenêtre, on calcule ensuite un index de proéminence séparé pour la durée, la variation de Fo et la différence d'intensité entre pseudo-syllabes. Pour chaque fenêtre de proéminence les pseudo-syllabes sont ordonnées de n à 1. La Fig. 2 donne un exemple avec n = 7 et un classement

de durée. La fenêtre de proéminence est alors déplacée vers la droite et les pseudo-syllabes sont à nouveau classées. Le processus est répété jusqu'à la fin de la séquence de pseudo-syllabes pour chacun des paramètres de durée, Fo et intensité..

4.4. Classement des index de proéminence

L'index de classement de chaque pseudo-syllabe est ensuite additionné et ensuite normalisé pour les 7 positions de la fenêtre glissante pour tenir compte des classements partiels effectués en début et en fin de séquence (cf. Fig. 2 ci-dessous).

La Fig. 3 ci-dessous donne un exemple similaire mais réalisé sur des valeurs de changement de Fo à l'intérieur de la pseudo-syllabe.

Les Fig. 4 et 5 montrent les pseudo-syllabes retenues par ce processus, avec un seuil de 5,6 pour la durée (Fig. 4) et de 5,8 pour la variation de Fo (Fig. 5). Ces seuils ont été choisis arbitrairement de manière à n'obtenir qu'un nombre prédéterminé de syllabes proéminentes.

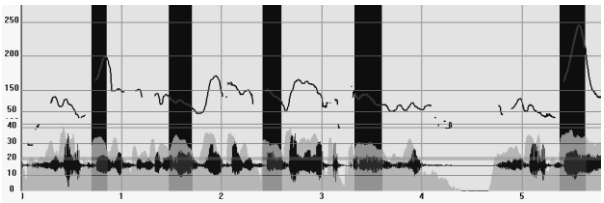


Figure 4 : pseudo-syllabes retenues comme proéminentes en durée avec un seuil de 5,6

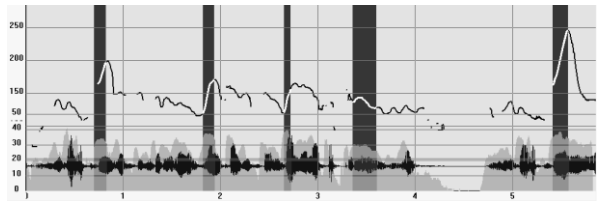


Figure 5 : pseudo-syllabes retenues comme proéminentes en variation de Fo avec un seuil de 5,8

4.5. Un exemple

Considérons l'exemple suivant extrait du corpus CFPP 200, [12]) :

j'y vais à pied je suis chez moi je m'conditionne dans mon appartement en me disant j'y vais à pied

Cet énoncé contient un total de 25 syllabes. Le critère basé sur les pics d'intensité a permis de détecter 22 pseudo-syllabes (cf. 3.1).

Les 5 syllabes perçues auditivement comme les plus proéminentes sont :

*j'y vais à **PIED** je suis chez **MOI** **je** m'condi**TION**ne **dans** mon apparte**MENT** en m'di**SANT** j'y vais à **PIED***

Les 5 accents lexicaux sont ***PIED***, ***MOI***, *m'condi**TION**ne*, *di**SANT***, ***PIED***, et les 2 syllabes avec accent secondaire ***je*** et ***dans***.

Les 5 pseudo-syllabes avec les plus grandes durées relatives :

PIED, ***MOI***, *m'condi**TION**ne*, *apparte**MENT***, ***PIED***.

Les 5 pseudo-syllabes avec les plus grandes variations de Fo :

PIED, ***je***, ***dans***, *apparte**MENT***, ***PIED***.

Une seule syllabe perçue auditivement comme proéminente *m'di**SANT*** manque dans la liste obtenue par l'union des pseudo-syllabes classées comme proéminentes par la durée et le changement de Fo

En comparant les durées et les changements de Fo des pseudo-syllabes les plus proéminentes de l'exemple avec a) la proéminence phonologique prédite à partir de la transcription orthographique et b) la proéminence perçue auditivement, on obtient les résultats consignés dans la table 1, dans laquelle les proéminences sont indiquées par X, phonologiques pour la ligne *lex*,

Table 1 : Syllabes perçues comme proéminentes selon des critères phonologiques (accent lexical et secondaire) et acoustique par fenêtre glissante pour la durée et la variation de Fo.

	PIED	***MOI***	***je***	***TION***	***dans***	***MEN***	***SANT***	***PIED***
lex	X	X		X		X	X	X
sec			X		X			
dur	**X**	**X**		**X**		**X**		**X**
Fo	**X**		**X**		**X**	**X**		**X**

5. CONCLUSIONS

Les décisions d'experts en détection de proéminence syllabique résultent d'interactions complexes d'ordre phonétique, phonologique, lexical, syntaxique et sémantique [1]. L'obtention de résultats similaires par un processus automatique présuppose l'accès aux mêmes informations, ce qui n'est certainement pas le cas d'algorithmes n'utilisant que l'information acoustique. On en conclut que la détection automatique des proéminences syllabiques (ou pseudo-syllabiques) ne peut être qu'un outil pour le linguiste permettant un accès plus rapide et efficace aux données possiblement pertinentes.

Le processus présenté n'exige pas de segmentation préalable en syllabes (mais il peut bien sur intégrer cette information si elle est disponible). Les résultats préliminaires sont très encourageants et ouvrent la voie à la détection assistée de la proéminence, pour laquelle l'opérateur reste le seul juge in fine. La méthode peut être aisément adaptée pour incorporer d'autres paramètres comme la pause ou la différence d'intensité entre pseudo-syllabes consécutives, etc.

6. RÉFÉRENCES

[1] Martin, Philippe (2005) La transcription des proéminences accentuelles : mission impossible ? *Revue PFC*, septembre 2005.

[2] Avanzi, Mathieu et Philippe Martin (2007) Un outil pour la détection automatique des proéminences accentuelles dans les corpus oraux, *XXV CILPR 2007*, Innsbruck, 3-8 septembre 2007.

[3] Goldman, Jean-Philippe (2007) *EasyAligner: a semi-automatic phonetic alignment tool under Praat*. http://latcui.unige.ch/phonetique.

[4] Avanzi, Mathieu, Jean-Philippe Goldman, Anne Lacheret-Dujour, Anne-Catherine Simon & Antoine Auchlin (2007) Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé, *Cahiers of French Language Studies,* 13/2, 2-30.

[5] Mertens, Piet (2004) Un outil pour la transcription de la prosodie dans les corpus oraux *Traitement Automatique des langues* 45 (2) 109-130.

[6] Rossi, Mario (1978) Interaction of intensity glides and frequency glissandos, Language and Speech (21) 4, 384-396.

[7] Obin, Nicolas, Goldman, J.-P., Avanzi, M. & Lacheret-Dujour, A. (2008) Comparaison de trois outils de détection automatique des proéminences en français parlé *Actes des 27èmes journées d'étude sur la parole (JEP 08)*, Avignon, 8-13 juin 2008. non paginé

[8] Goldman, Jean-Philippe & Avanzi, M. (2007) Vers un algorithme de détection (semi-)automatique des proéminences en français parlé *Actes des 7èmes Rencontres des Jeunes Chercheurs sur la Parole (RJCP07)*, Paris, 05-06 juillet 2007, 84-87.

[9] Mertens, Piet & Christophe d'Alessandro (1995) *Pitch contour stylization using a tonal perception model,* Proc. 13th International Congress of Phonetic Sciences, Vol. 4, 228-231.

[10] Martin, Philippe (1979) Automatic Location of Stressed Syllables in French *Proceedings of the International Congress of Phonetics*, Miami, 1091-1094.

[11] Wioland, François (1985) *Les structures rythmiques du français*, Slatkine-Champion, Paris.

[12] Branca-Rosoff, Sonia (2009) Corpus de Français Parlé Parisien (CFPP 2000) http://ed268.univ-paris3.fr/syled/ressources/Corpus-Parole-Paris-PIII/

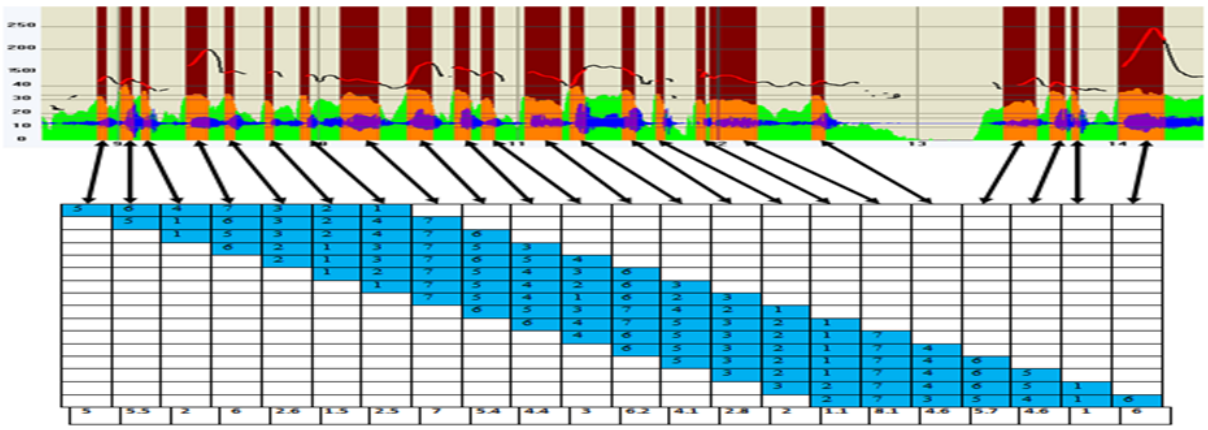


Figure 2 : Proéminence des durées relatives calculée sur une fenêtre glissante de 7 pseudo-syllabes. La somme normalisée des valeurs relatives donne la proéminence relative sur tout le signal

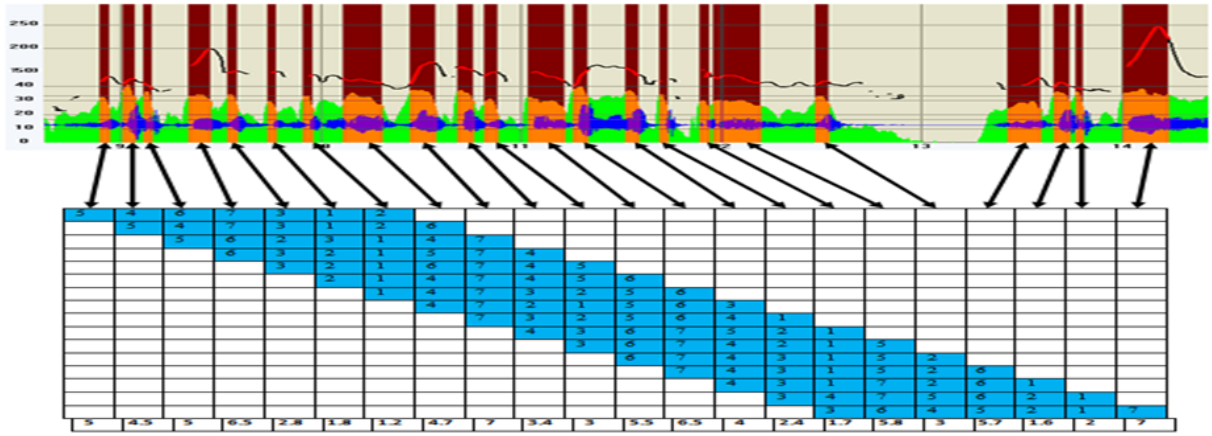


Figure 3 : Proéminence des changements de Fo calculée sur une fenêtre glissante de 7 pseudo-syllabes. L'addition des valeurs relatives donne la proéminence relative sur tout le signal

Comparaison des propriétés acoustiques de la parole lue, préparée et conversationnelle en français

Jean-Luc Rouas[1], *Mayumi Beppu*[2], *Martine Adda-Decker*[1]

[1] LIMSI-CNRS UPR 3251, France
[2] Dept. Computer Science, Tokyo Institute of Technology, Japan
rouas@limsi.fr, beppu@ks.cs.titech.ac.jp, madda@limsi.fr

ABSTRACT

In this paper, we investigate the acoustic properties of vocalic phonemes in three speaking styles : Read speech, broadcast news and casual spontaneous speech. Our aim is to understand better why speech recognition systems still fail to achieve good performances on spontaneous speech. Using Nakamura's method [9], we use classical speech recognition features, MFCC, and try to represent the effects of the speaking styles on the spectral space. We happen to find some cues, and it also seems that phone duration also plays an important role regarding spectral reduction, especially for spontaneous speech.

Keywords: speaking styles, speech recognition

1. Introduction

Ce travail fait suite aux recherches effectuées par Nakamura et al. [9] sur les styles de parole en Japonais. Dans cet article, nous nous focaliserons sur les styles de parole en Français, tout d'abord en reproduisant certaines expériences de Nakamura et en proposant ensuite de nouvelles analyses. Les motivations pour ce travail sont décrites dans la section 2. La méthode d'analyse est détaillée dans la section 3. Les résultats obtenus par Nakamura sur le Japonais sont rappelés dans la section 4. Les données utilisées pour l'analyse du Français sont décrites dans la section 5. Les expériences effectuées sur les données en Français sont détaillées dans la section 6. Enfin, nous comparons les résultats obtenus sur le Français et sur le Japonais lorsque cela peut être fait.

2. Motivations

Après s'être principalement penchés sur la transcription de la parole lue, les systèmes de reconnaissance automatique de la parole obtiennent aujourd'hui de très bons résultats sur des données radiophoniques, habituellement légèrement au dessus de 90% [3]. Comme le montrent les évaluations NIST des deux dernières décennies, ces systèmes ont bénéficié de nombreuses années de recherches dédiées à la transcription automatique de données de parole lue ou radiophonique.

Toutefois, lorsque ces systèmes sont confrontés à de la parole conversationnelle spontanée, les performances se dégradent nettement. Cette chute drastique des performances peut s'expliquer par le fait que des différences majeures existent entre la parole lue et la parole spontanée, à la fois en termes linguistiques et acoustiques [4].

Un des challenges subsistant pour la reconnaissance automatique de la parole est d'apporter aux systèmes la possibilité de traiter toutes les qualités de parole, y compris la parole conversationnelle spontanée. Dans cet objectif, il est important d'analyser quelles sont les différences les plus notables entre les styles de parole en considérant les paramètres habituellement utilisés dans les systèmes de transcription automatique, les paramètres acoustiques.

De ce point de vue, les distributions spectrales des voyelles et des syllabes en parole continue sont bien plus réduites que lorsque celles-ci sont prononcées isolément. Ce phénomène est appelé réduction spectrale. Il a également été observé dans l'espace paramétrique lors d'une comparaison entre parole spontanée et parole lue, en utilisant des données formantiques provenant d'un locuteur [11]. L'étude de Nakamura [9] sur le Japonais a confirmé ce résultat en utilisant un corpus plus étendu.

En ce qui concerne le Français, nous savons que la parole lue et la parole spontanée sont structurellement différentes. Les syllabes complexes ont tendance à se simplifier, et la disparition des consonnes de fin de mot et des voyelles des syllabes non accentuées sont fréquentes pour la parole spontanée en Français [1].

C'est ce phénomène que nous estimons être responsable d'une partie des erreurs de transcription en parole conversationnelle que nous nous proposons d'étudier dans cet article.

Un autre phénomène pouvant être caractéristique de la parole spontanée est l'extension de variance spectrale. Ce phénomène a été caractérisé pour le Japonais dans [9].

Nous présentons dans la section suivante les méthodes de calcul permettant de quantifier l'intensité de ces phénomènes.

3. Méthode de calcul et formules

3.1. Paramètres

Sur des fichiers audio numérisés avec un taux d'échantillonnage de 16 kHz, un jeu de 12 coefficients ceps-

traux (MFCC) sont extraits en utilisant une fenêtre de 25 ms avec un recouvrement de 10 ms. Aux vecteurs MFCC sont ajoutées leurs dérivées premières et secondes, ainsi que les premières et secondes dérivées de la log-énergie. Les vecteurs résultants sont de dimension 38. Ce sont les paramètres les plus classiquement utilisés en reconnaissance de parole.

3.2. Ratio de réduction de l'espace spectral

Afin de caractériser la réduction de l'espace spectral, nous devons nous munir d'un corpus de référence. Ce corpus, dénoté dans les formules suivantes R, est dans notre cas un corpus de parole lue.

L'étendue de l'espace spectral est estimée en faisant la différence entre la moyenne des vecteurs MFCC pour un phonème p donné et la valeur moyenne sur l'ensemble des phonèmes du corpus.

Le ratio est mesuré en divisant l'estimateur de l'étendue de l'espace spectral d'un corpus X par celui du corpus de référence R. Nous utilisons ici des distances euclidiennes.

En d'autres termes, le calcul peut être résumé par la formule suivante :

$$red_p(X) = \frac{\|\mu_p(X) - Av(\mu_p(X))\|}{\|\mu_p(R) - Av(\mu_p(R))\|} \quad (1)$$

avec μ_p la valeur moyenne des vecteurs MFCC du phonème p du corpus X, $\mu_p(R)$ la moyenne des vecteurs MFCC pour le phonème p pour la parole lue (corpus R), Av indique la valeur moyenne.

3.3. Ratio d'extension de la variance spectrale

De la même manière, nous définissons le ratio d'extension de la variance spectrale par rapport à un corpus de référence R.

La variance spectrale est estimée comme étant la somme des variances de chacun des coefficients MFCC pour l'ensemble des réalisations d'un phonème p.

Le ratio d'extension de la variance est alors le rapport entre la variance spectrale du phonème p corpus X et celle du même phonème pour le corpus R.

Il est calculé en utilisant la formule suivante :

$$ext_p(X) = \frac{\sum_{k=1}^{K} \sigma_{pk}^2(X)}{\sum_{k=1}^{K} \sigma_{pk}^2(R)} \quad (2)$$

avec K la dimension des vecteurs MFCC, $\sigma_{pk}^2(X)$ le $k$ième élément du vecteur de variance obtenu pour le phonème p avec le style de parole X.

4. Caractérisation acoustique des styles de parole en Japonais [9]

Dans [9], les différents styles de parole sont étudiés grâce à de grands corpus. Les données utilisées lors de ces expériences proviennent du "Corpus of Spontaneous Japanese" (CSJ) [8] et du "Japanese Newspaper Article Sentence" (JNAS) [6].

Ces bases de données couvrent différentes conditions discursives, incluant des monologues, des dialogues et de la parole lue. Pour les expériences suivantes, seules quelques conditions, considérées comme les plus représentatives d'un style de parole, sont utilisées : Parole lue, Présentations académiques, Présentations informelles, Dialogue.

Les expériences menées en utilisant sur ces bases de données montrent que la réduction de l'espace acoustique des MFCC est observable pour quasiment tous les phonèmes dans les trois styles de parole par rapport à la parole lue. Ce phénomène est plus marqué lorsque l'on considère les situations de dialogue.

Les résultats obtenus avec la mesure du ratio d'extension de variance spectrale montrent bien une augmentation de la variance pour quasiment tous les phonèmes dans les trois styles de parole.

5. Données en Français

Trois bases de données comprenant de la parole lue, préparée et conversationnelle ont été utilisés. Pour la parole lue, nous utilisons le corpus BREF [7], qui est composé de textes journalistiques lus. Le corpus BREF comporte plus de 100 heures de parole lue par 120 locuteurs. Les textes ont été sélectionnés à partir du journal Le Monde afin de couvrir un vocabulaire large (plus 20000 mots) et un nombre important de contextes phonétiques.

Le corpus de parole préparée est constitué des parties de développement et de test des campagnes d'évaluation ESTER (2003-2004) et ESTER2 (2007-2008) [5]. Ce corpus a une durée d'environ 50 heures. Les données sont composées d'émission télévisées ou radiophoniques provenant de France mais également de plusieurs pays francophones (Maghred et Afrique) : France Inter, Radio France Internationale, Radio Télévision du Maroc, France Info.

Les données de parole spontanée proviennent du corpus NCCFr [10]. Ce corpus est composé de conversations informelles entre amis. D'une durée de 36 heures, ce corpus comprend 23 paires de locuteurs (24 hommes et 22 femmes), chaque conversation durant approximativement 90 minutes. Ce corpus a été enregistré en 2007 à Paris.

Chacun de ces corpus sera dénommé dans la suite de ce document par les acronymes BREF, ESTER, NCCFr respectivement. L'ensemble des corpus inclut des transcription orthographiques manuelles qui ont été phonétisées et alignées automatiquement.

6. Expériences

La première expérience est une mesure des durées des phonèmes automatiquement alignés pour les différents corpus. L'objectif est d'une part de vérifier le répartition des durées des phonèmes selon les styles de parole et, d'autre part, de sélectionner les phonèmes ayant les durées les plus représentées.

Le résultat de cette expérience est donné sur la figure 1.

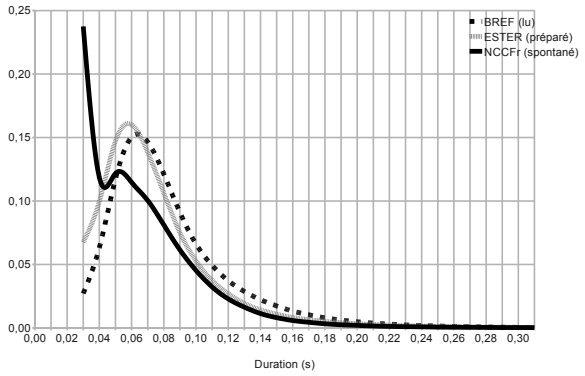


Fig. 1: Distribution des durées des phonèmes pour chaque corpus. Abscisses : durée en secondes, Ordonnées : estimation de la probabilité

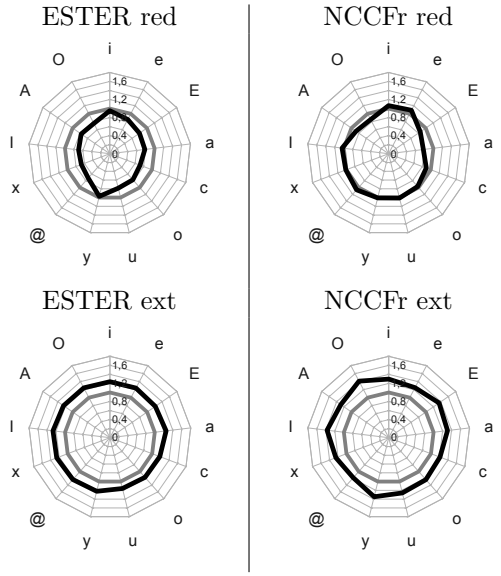


Fig. 2: ratio de réduction spectrale (red) and ratio d'extension de variance (ext) pour les corpus ESTER and NCCFr comparés avec BREF

Comme le montre cette figure, les distributions pour BREF et ESTER ont des formes semblables, avec un léger décalage vers la gauche pour ESTER. Ce décalage montre l'effet du débit de parole, normalement plus élevé dans le style journalistique et la parole spontanée que pour la lecture.

La forme de la distribution pour le corpus NCCFr est toutefois assez différente, avec un pic important vers les durées de l'ordre de 30 ms (durée minimale d'un phonème dans l'alignement automatique). Cela peut s'expliquer par le fait que les données considérées sont en parole conversationnelle spontanée, donc sujettes aux phénomènes de délétions de certains phonèmes. Une étude plus détaillée de ce phénomène est effectuée dans [2].

Les valeurs médianes des distributions sont cependant similaires pour chacun des corpus (autour de 60 à 70 ms). Pour la suite des expériences, nous avons dans un premier temps décidé de ne considérer que les segments phonétiques ayant des durées comprises entre 40 et 120 ms.

6.1. Ratio de réduction & extension de variance

Le ratio de réduction de l'espace spectral des phonèmes et le ratio d'extension de variance sont calculés sur nos données en utilisant les formules décrites dans les sections 3.2, équation 1 et 3.3, équation 2. Le corpus de référence R est pour l'ensemble des expériences menées ci-après le corpus BREF. Pour des raisons de clarté, nous avons décidé de ne représenter que les figures décrivant les résultats obtenus pour les voyelles.

Comme nous pouvons le voir sur la première ligne de la figure 2, la réduction de l'espace spectral est observée principalement pour le corpus ESTER. Pour ce corpus, quasiment tous les phonèmes voient leur espace spectral réduit, à part /i/ et /y/. Sur notre corpus de parole conversationnelle, NCCFr, l'espace spectral semble quasiment identique à celui du corpus de référence, à l'exclusion de certains phonèmes qui voient leur étendue spectrale légèrement diminuée.

Les résultats sur la mesure de l'extension de la variance spectrale sont en revanche cohérents avec nos prédictions. Nous pouvons observer, sur la deuxième ligne de la figure 2, une augmentation nette de la variance spectrale pour ESTER, et une augmentation encore plus importante pour NCCFr.

6.2. Influence de la durée des segments phonémiques

Au vu des résultats quelque peu étonnants de la section précédente, nous avons effectué des expériences supplémentaires pour évaluer la réduction de l'espace spectral pour les segments courts (en dessous de 40 ms) et les segments longs (au dessus de 120 ms).

La figure 3 montre les résultats des expériences effectuées selon ces deux conditions. Nous pouvons observer ici que les segments courts, notamment pour NCCFr, voient leur étendue spectrale très réduite. Malgré la durée des segments considérés, ce phénomène n'est pas à négliger car les segments courts sont nombreux dans notre corpus de parole conversationnelle. Pour les segments de durée importante, la réduction de l'espace spectral n'est pas très importante.

Le ratio d'extension de variance a également été calculé pour les deux conditions (figure 4). Ces figures montrent que la variance spectrale des phonèmes augmente pour les deux conditions de durée. L'extension paraît plus importante pour les segments courts de parole conversationnelle par rapport aux segments longs.

7. Discussion

Nakamura [9] obtient des résultats montrant des effets assez nets de la réduction de l'espace spectral et de l'extension de la variance spectrale en Japonais.

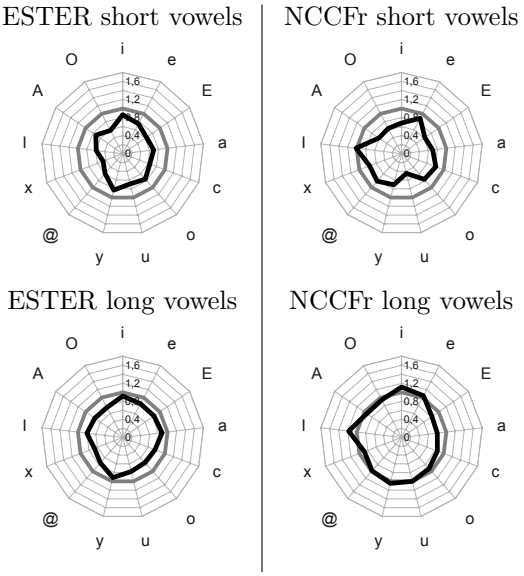


Fig. 3: Ratio de réduction spectrale pour les corpus ESTER et NCCFr en utilisant BREF comme référence. première ligne : segments courts (<40ms), deuxième ligne : segments longs (>120ms)

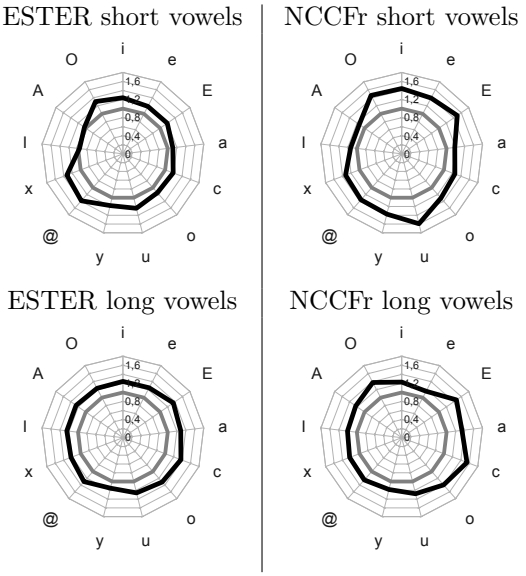


Fig. 4: Ratio d'extension de variance spectrale pour les corpus ESTER et NCCFr en utilisant BREF comme référence. première ligne : segments courts (<40ms), deuxième ligne : segments longs (>120ms)

Sur ses données, l'espace spectral est de plus en plus réduit lorsqu'il analyse des données de plus en plus spontanées. De la même manière, la variance spectrale est toujours plus importante pour les corpus de parole spontanée.

Dans nos expériences sur le français, nous avons pu reproduire les mêmes résultats pour les mesures de variance spectrale. Cependant, les tests effectués sur les segments de durée "normale" vont quelque peu à l'encontre des résultats de Nakamura : l'espace spectral est moins étendu pour la parole journalistique que pour la parole spontanée.

Nous avons pu toutefois mesurer une réduction spectrale importante pour les segments courts, majoritaires dans le cas de la parole spontanée.

C'est certainement sur ces segments, déjà difficiles à reconnaître de manière automatique à cause de leur faible durée, que nous devrons focaliser nos efforts afin d'améliorer les performances des systèmes de transcription automatique.

Références

[1] M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel. Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, 46(2) :119–139, 2005.

[2] M. Adda-Decker, C. Gendrot, and N. Nguyen. Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues*, 49, 2008.

[3] P. Fousek, L. Lamel, and J.-L. Gauvain. Transcribing Broadcast Data Using MLP Features. In *InterSpeech'08*, pages 1433–1436, Brisbane, Australia, September 22-26 2008.

[4] S. Furui. Recent advances in spontaneous speech recognition and understanding. In *ISCA & IEEE workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003.

[5] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, M. Mostefa, and K. Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Language Evaluation and Ressources Conference*, 2006.

[6] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi. Jnas : Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of the acoustical society of Japan*, 20(3) :199–206, 1999.

[7] L. Lamel, J. L. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french. In *Eurospeech*, 1991.

[8] K. Maekawa. Corpus of spontaneous japanese : its design and evaluation. In *ISCA & IEEE workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003.

[9] M. Nakamura, K. Iwano, and S. Furui. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*, 22 :171–184, 2008.

[10] F. Torreira, M. Adda-Decker, and M. Ernestus. The nijmegen corpus of casual french. *Speech Communication*, in press.

[11] R. J. J. H. van Son and L. C. W. Pols. An acoustic description of consonant reduction. *Speech Communication*, 28(2) :125 – 140, 1999.

Méthodes basées sur les HMMs et les GMMs pour l'inversion acoustico-articulatoire en parole

Atef Ben Youssef, Viet-Anh Tran, Pierre Badin, Gérard Bailly

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Université de Grenoble
961 rue de la Houille Blanche, BP 46, F-38402 Saint Martin d'Hères cedex, France
{Atef.Ben-Youssef, Viet-Anh.Tran, Pierre.Badin, Gerard.Bailly}@gipsa-lab.grenoble-inp.fr

ABSTRACT

Two speech inversion methods are implemented and compared. In the first, multistream Hidden Markov Models (HMMs) of phonemes are jointly trained from synchronous streams of articulatory data acquired by EMA and speech spectral parameters; an acoustic recognition system uses the acoustic part of the HMMs to deliver a phoneme chain and the states durations; this information is then used by a trajectory formation procedure based on the articulatory part of the HMMs to resynthesise the articulatory data. In the second, Gaussian Mixture Models (GMMs) are trained on these streams to associate directly articulatory frames with acoustic frames in context. Over a corpus of 17 minutes uttered by a French speaker, the RMS error was 1,66 mm with the HMMs and 2,25 mm with the GMMs.

Keywords: speech inversion, EMA, GMM, HMM

1. INTRODUCTION

L'inversion en parole a été longtemps basée sur l'analyse par la synthèse. Mais depuis une décade, des techniques d'apprentissage plus sophistiquées sont apparues, grâce à la disponibilité de corpus importants de données articulatoires et acoustiques produites par des dispositifs tels que l'Articulographe Electro-Magnétique (EMA) ou des dispositifs de suivi de marqueurs basés sur la vidéo classique ou infrarouge.

Au moins deux classes de modèles statistiques de production de parole se trouvent dans la littérature récente: les modèles de Markov cachés (HMMs) [1], [2], et les modèles de mélanges de Gaussiennes (GMMs) [3]. En plus de la différence structurelle entre HMMs et GMMs, on peut noter que les HMMs utilisent explicitement les informations phonétiques et l'organisation temporelle, tandis que les GMMs agrègent simplement le comportement multimodal de segments de parole similaires.

Hiroya & Honda [1] ont développé une méthode qui estime les mouvements articulatoires à partir du son à l'aide d'un modèle de production de parole basé sur les HMMs. Pour chaque phone, le modèle comprend un HMM des paramètres articulatoires dépendant du contexte et un associateur linéaire qui transforme les paramètres articulatoires en spectre de parole pour chacun des états du HMM. Les modèles sont construits à partir d'observations acoustiques et articulatoires simultanées acquises par EMA. En utilisant le modèle de production, la séquence des états HMM est déterminée en cherchant le maximum de vraisemblance de la séquence de spectres de parole. Les paramètres articulatoires sont ensuite déterminés en cherchant le maximum de l'estimation a posteriori des paramètres articulatoires pour un spectre de parole donné et la séquence des états HMM. L'erreur RMS moyenne obtenue est de 1,73 mm.

Toda et coll. [3] ont décrit une approche statistique à la fois pour le mapping articulatoire vers acoustique et le mapping inverse acoustique vers articulatoire sans information phonétique. Ils modélisent la densité de probabilité conjointe des trames acoustiques et articulatoires en contexte par un modèle GMM entraîné sur une base de données parallèles acoustiques et articulatoires. Ils utilisent deux techniques différentes pour établir le mapping GMM. Avec un critère d'erreur quadratique moyenne minimum (MMSE) sur une fenêtre acoustique de 11 trames et 32 composantes pour le GMM, ils obtiennent des erreurs RMS d'inversion de 1,61 mm pour une locutrice, et de 1,53 mm pour un locuteur. L'utilisation d'une méthode de maximum de vraisemblance (MLE) avec 64 composantes gaussiennes, réduit les erreurs à 1,45 mm pour la locutrice, et à 1,36 mm pour le locuteur.

Les études décrites ci-dessus ne permettent pas de déterminer la méthode d'inversion optimale, puisque les données, les locuteurs et les langues ne sont pas comparables. En outre, les corpus ainsi que les conditions d'apprentissage et de test ne sont pas non plus comparables. Ainsi, l'objectif du présent travail est de comparer, *ceteris paribus*, la méthode HMM utilisée dans [2] avec une méthode GMM similaire à celle de [3] en utilisant les critères MMSE et MLE pour la méthode GMM.

2. DONNEES

Pour cette étude préliminaire, nous avons utilisé un corpus déjà enregistré par un unique locuteur français [4], composé de deux répétitions de 224 séquences VCV, deux répétitions de 109 paires de mots de structure CVC différant par un seul trait, 68 phrases courtes et 20 phrases longues. Les phones sont d'abord étiquetés à partir du signal audio et de la transcription phonétique associée, à l'aide d'une procédure d'alignement forcé basée sur des HMMs. Les étiquettes et les frontières de phones sont

ensuite corrigées manuellement. Les 36 phonèmes sont : [a ɛ e i y u o ø ɔ œ ɑ̃ ɛ̃ œ̃ ɔ̃ p t k f s ʃ b d g v z ʒ m n ʁ l w ɥ j ə _ __], où _ et __ sont respectivement les pauses internes courtes et les pauses longues en début et fin de phrase. Au total, le corpus, dont les longues pauses ont été exclues, contient approximativement 100.000 trames (~17 mn) correspondant à 5132 allophones.

Les données articulatoires ont été acquises à l'aide d'un Articulographe Electro-Magnétique (EMA) qui permet de suivre dans le plan médiosagittal des points cutanés des articulateurs à l'aide de petites bobines électro-magnétiques. Pour cette étude, six bobines ont été utilisées: l'une attachée aux incisives inférieures, trois autres attachées à la pointe, au milieu, et à l'arrière de la langue, et les deux dernières attachées à la limite entre la peau et le vermillon des lèvres supérieure et inférieure. Le signal de parole a été enregistré à 22050 Hz, de manière synchrone avec les coordonnées des bobines EMA enregistrées à 500 Hz, et filtrées passe-bas à 20 Hz afin de réduire le bruit.

3. MODELES HMM

Nous rappelons les expériences menées en [2]. Pour l'apprentissage des HMMs, les vecteurs de traits acoustiques sont composés de 12 coefficients cepstraux en échelle Mel (MFCC) et du logarithme de l'énergie, estimés à partir du signal sous échantillonné à 16 kHz sur des fenêtres de 25 ms à une fréquence de trame de 100 Hz ; ces vecteurs sont complétés par les dérivées premières temporelles. Les vecteurs de traits articulatoires sont composés des 12 coordonnées *x* et *y* des six bobines actives, ainsi que leurs dérivées premières. Les trajectoires EMA sont sous-échantillonnées à 100 Hz pour être synchrones avec les vecteurs acoustiques.

Différentes variantes ont été testées: phonèmes sans contexte *(no-ctx)*, avec contexte gauche *(L-ctx)* ou droit *(ctx-R)*, et avec contextes gauche et droite *(L-ctx-R)*. Une méthode de regroupement hiérarchique basée sur la matrice des distances de Manhattan entre les coordonnées des bobines pour chaque paire de phonèmes, a permis de définir six classes cohérentes pour les contextes vocaliques ([a ɛ ɛ̃ | ø œ œ̃ | e i | y | u | o ɔ ɑ̃ ɔ̃]) et dix classes pour les contextes consonantiques ([p b m | f v | ʁ | ʃ ʒ | l | t d s z n | j | ɥ | k g | w]). Le schwa, et les pauses courtes et longues ([ə _ __]) ne sont pas pris en compte comme contextes.

Nous avons utilisé des modèles HMM gauche-droite à trois états, avec une gaussienne par état et une matrice de covariance diagonale. Les procédures d'apprentissage et de test ont été réalisées avec la boîte à outils HTK [5]. Pour l'apprentissage, le critère de maximum de vraisemblance (ML) était implémenté sous forme de maximisation de l'espérance (EM). Les vecteurs de traits acoustiques et articulatoires sont considérés comme deux flux dans la procédure multi-flux de HTK. Les modèles HMMs obtenus sont ensuite séparés en *HMMs articulatoires* et *HMMs acoustiques*.

Un modèle de langage bigramme considérant les séquences de phones en contexte est appris sur l'ensemble du corpus. L'inversion est réalisée en deux étapes : la première effectue une reconnaissance phonémique basée sur les HMMs acoustiques, et fournit la séquence des allophones reconnus, ainsi que la durée de chaque état. Une procédure d'héritage permet de remplacer un HMM en contexte manquant dans le corpus d'apprentissage par le HMM le plus proche [2]. La seconde étape resynthétise les trajectoires articulatoires à partir de ces informations à l'aide de la procédure de formation de trajectoire proposée par Zen *et al.* [6].

La méthode est ensuite évaluée par la procédure du *jack-knife* : les données sont séparées en 5 parties approximativement homogènes du point de vue de la répartition des phones ; chaque partie est tour à tour utilisée pour évaluer les performances des modèles HMM appris sur le restant des données. Les performances sont évaluées sur l'ensemble des 5 résultats par (1) la racine carrée des erreurs quadratiques moyennes (RMSE), (2) les coefficients de corrélation (*r*) entre données mesurées et données estimées, et (3) les taux de reconnaissance et de précision agrégés sur l'ensemble du corpus.

Les taux de reconnaissance / précision obtenus varient entre 88,90 / 68,99 % en l'absence de contexte et la meilleure performance de 93,66 / 80,9 % obtenue pour des phones en contexte droit. La procédure d'héritage de HMMs manquant permet de gagner entre 1 et 5 % sur les performances de reconnaissance. Le modèle de langage pour la reconnaissance permet de passer de taux de reconnaissance / précision de 72,29 / 34,22 % à 93,66 / 80,90 %. Cette amélioration spectaculaire a cependant une faible influence sur les performances puisque l'on passe seulement, en contexte droit, de 1,83 à 1,66 mm pour la RMSE et de 0,90 à 0,92 pour la corrélation. On voit sur la Table 1 que l'utilisation de contextes augmente très sensiblement les performances (sauf pour le contexte droit et gauche pour lequel la reconnaissance est nettement moins bonne, vraisemblablement dû à la taille trop petite du corpus).

Afin d'estimer la contribution du processus de formation de trajectoire à l'erreur RMSE de l'inversion complète, nous avons aussi synthétisé les trajectoires en utilisant un alignement forcé des états basés sur étiquettes originales, émulant ainsi un étage de reconnaissance parfaite (voir

Table 1 : RMSE (mm) et coefficient de corrélation *r* pour la méthode HMM. (1) : avec étape de reconnaissance parfaite ; (2) inversion complète.

	no-ctx		**L-ctx**		**ctx-R**		**L-ctx-R**	
	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*
(1)	1,91	0,90	1,55	0,93	1,55	0,93	**1,40**	**0,94**
(2)	2,07	0,87	1,72	0,91	**1,66**	**0,92**	1,91	0,89

Table 1). Le niveau relativement élevé de ces erreurs montre que la majeure partie de l'erreur globale (entre 70 et 90 %) est due à l'étape de formation de trajectoire qui lisse en excès les mouvements prédits et ne capture pas de manière appropriée les patrons de coarticulation.

4. MODELES GMM MULTIMODAUX

Nous avons mis en œuvre une mise en correspondance basée sur les GMMs en utilisant le critère de minimum de l'erreur quadratique moyenne (MMSE), souvent utilisé pour la conversion de voix. En outre, afin d'améliorer la précision de l'inversion, nous avons ajouté une étape d'optimisation basée sur l'estimation du maximum de vraisemblance (MLE) [3]. Les trajectoires de paramètres cibles ayant les propriétés statiques et dynamiques adéquates sont déterminées en combinant les estimations locales de la moyenne et de la variance pour chaque trame $p(t)$ et ses dérivées $\Delta p(t)$ par la relation explicite entre les paramètres statiques et dynamiques (*p. ex.* $\Delta p(t) = p(t) - p(t-1)$). Pour chaque trame, le vecteur de traits est la concaténation d'un vecteur articulatoire de 24 composantes (12 coordonnées EMA et leurs dérivées), et d'un vecteur acoustique de 24 composantes. Afin de prendre en compte le contexte acoustique [3], [7], de 9 à 17 vecteurs acoustiques (12 MFCC, énergie log) sont prélevés de manière équirépartie dans une zone temporelle contextuelle de taille variable, et réduits à 24 composantes par Analyse en Composantes Principales. Nous avons fait varier le nombre de composantes gaussiennes de 8 à 64 et la zone contextuelle d'une taille phonémique (~90 ms) à une taille syllabique (~170 ms). Chaque gaussienne est représentée par une matrice de covariance pleine (48×48), un vecteur de moyennes (48) et son coefficient de pondération.

La Table 2 montre les performances pour les différentes expériences déterminées par la même méthode *jack-knife* sur les mêmes parties. L'erreur quadratique moyenne (RMSE) diminue lorsque le nombre de composantes augmente, et atteint un optimum pour une fenêtre contextuelle de 110 ms. L'explication la plus plausible est qu'une fenêtre de taille de diphone contient de manière optimale les traits phonétiques locaux nécessaires à l'inversion. La meilleure précision d'inversion est finalement obtenue pour une fenêtre de 110 ms et 64 composantes qui semblent constituer la meilleure représentation des 36 phonèmes. Nous avons noté par ailleurs que l'étape supplémentaire d'optimisation par MLE augmente les performances de l'ordre de 5 %.

Table 2 : RMSE (mm) et coefficient de corrélation *r* pour la méthode GMM en fonction du nombre de Gaussiennes (# mix) et de la taille du contexte *ctw* (ms).

#mix	**8**		**16**		**32**		**64**	
ctw	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*
90	2,68	0,78	2,61	0,80	2,38	0,83	2,32	0,84
110	2,68	0,78	2,54	0,80	2,37	0,83	**2,25**	**0,85**
130	2,66	0,78	2,51	0,81	2,36	0,83	2,27	0,85
150	2,66	0,78	2,50	0,81	2,44	0,82	2,32	0,84
170	2,65	0,78	2,44	0,82	2,41	0,82	2,29	0,84

5. COMPARAISONS ET COMMENTAIRES

La Figure 1 compare les trajectoires originales et estimées des ordonnées des bobines EMA pour les systèmes étudiés.

Au vu de la littérature, il est surprenant que nos résultats d'inversion basés sur les HMMs soit significativement plus précis ($p < 10^{-6}$) que ceux basés sur les GMMs (1,66 mm *vs.* 2,25 mm) : dans les deux expériences les plus abouties, Hiroya & Honda [1] trouvent 1,73 mm avec des HMMs (ce qui est proche de nos résultats) comparé aux 1,36 – 1,45 mm trouvés par Toda et coll. [3] avec des GMMs. Même en prenant en compte le fait que ces deux expériences sont basées sur des sujets et des langues différentes, la différence est telle que nous ne nous attendions pas à de tels résultats. Nous n'avons pour l'instant pas d'explication à cette divergence.

Nos deux systèmes peuvent cependant être améliorés. L'inversion à base de HMMs pourrait inclure un traitement plus sophistiqué de l'asynchronie articulatoire / acoustique en introduisant des modèles de retard qui se sont révélés efficaces pour la synthèse multimodale par HMMs [8]. Le système basé sur les GMMs pourrait être amélioré en considérant d'autres techniques de réduction de la dimensionnalité telles que l'Analyse Discriminante Linéaire (LDA) qui sont assez efficaces pour l'inversion

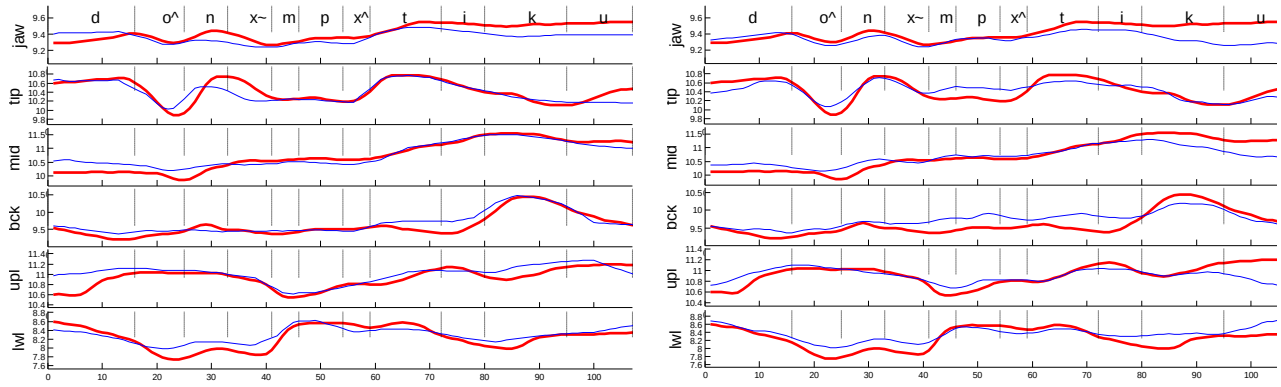


Figure 1 : Comparaison des trajectoires originales (traits épais) et prédites (traits fins) des ordonnées des 6 points de chair EMA. A gauche: inversion et synthèse basée-HMM avec des modèles en contexte droit. A droite: synthèse par GMM avec une fenêtre contextuelle de 110 ms centrée sur la trame courante et un mélange de 64 Gaussiennes.

basée sur les HMMs [7]. Les deux systèmes pourraient aussi gagner à incorporer de l'information visuelle en entrée et à inclure de manière plus intime cette information additionnelle dans le processus d'optimisation qui va considérer la cohérence multimodale entre les paramètres d'entrée et de sortie: les lèvres sont clairement visibles, et la position de la mâchoire est accessible de manière indirecte à partir des mouvements faciaux.

6. CONCLUSIONS ET PERSPECTIVES

Nous avons mis en œuvre et comparé deux techniques d'inversion acoustico-articulatoire en parole qui diffèrent par la façon dont elles capturent et exploitent la cohérence multimodale a priori entre son et articulation. Plusieurs réserves peuvent cependant être faites à propos de ces premières expériences.

Le système basé sur les HMMs bénéficie de la phonotactique du langage cible. Notons que le Français possède un inventaire syllabique riche : nous pouvons ainsi imaginer que les résultats obtenus avec des langues présentant des structures phonologiques très différentes telles que le Japonais, le Polonais ou l'Espagnol présentant des complexités syllabiques diverses pourraient conduire à des résultats différents.

Les mesures objectives globales pourraient ne pas refléter entièrement le comportement spécifique des phones qui peut avoir un impact majeur sur une évaluation subjective de l'articulation générée. La précision de la reconstruction est bien naturellement de la plus haute importance pour l'évaluation, mais d'autres éléments tels que la précision de la récupération d'éléments cruciaux comme les constrictions du conduit vocal sont également très importants.

Nous avons montré [4] que les sujets ont des performances très diverses en *lecture linguale*, et que cette performance augmente avec l'entraînement. Notons ainsi que le réalisme du mouvement pourrait compenser le manque de précision des détails de forme: la cinématique des trajectoires calculées pourrait être plus importante pour la perception que la précision des trajectoires elles-mêmes.

Finalement, les résultats de cette étude vont nous permettre de développer un système de tuteur pour la correction phonétique [9], dans lequel les mouvements articulatoires reconstruits seront utilisés pour piloter une tête parlante virtuelle 3D avec tous les degrés de liberté possibles [10].

7. REMERCIEMENTS

Nous remercions Christophe Savariaux et Coriandre Vilain pour les enregistrements EMA, et Tomoki Toda (NAIST, Japon) pour la mise à disposition du logiciel de GMM. Ce travail a été partiellement financé par le projet ANR-08-EMER-001-02 *ARTIS* et le projet Franco-japonais PHC SAKURA *CASSIS*.

BIBLIOGRAPHIE

[1] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Trans. Speech and Audio Processing,* vol. 12, pp. 175-185, March 2004.

[2] A. Ben Youssef, P. Badin, G. Bailly, and P. Heracleous. Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models, in *Interspeech 2009*, Brighton, UK, pp. 2255-2258, 2009.

[3] T. Toda, A. W. Black, and K. Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication,* vol. 50, pp. 215-227, 2008/3 2008.

[4] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly. Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication,* vol. 52, pp. 493-503, 2010.

[5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK Book. Revised for HTK Version 3.4 December 2006, 2006.

[6] H. Zen, K. Tokuda, and T. Kitamura. An introduction of trajectory model into HMM-based speech synthesis, in *Fifth ISCA ITRW on Speech Synthesis (SSW5)*, Pittsburgh, PA, USA, pp. 191-196, 2004.

[7] V.-A. Tran, G. Bailly, H. Loevenbruck, and C. Jutten. Improvement to a NAM captured whisper-to-speech system, in *Interspeech*, Brisbane, Australia, pp. 1465-1468, 2008.

[8] O. Govokhina, G. Bailly, and G. Breton. Learning optimal audiovisual phasing for a HMM-based control model for facial animation, in *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.

[9] P. Badin, G. Bailly, and L.-J. Boë. Towards the use of a Virtual Talking Head and of Speech Mapping tools for pronunciation training, in *Proceedings of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning*, Stockholm, Sweden, pp. 167-170, 1998.

[10] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka. An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data, in *Vth Conference on Articulated Motion and Deformable Objects (AMDO 2008, LNCS 5098)*, F. J. Perales and R. B. Fisher, Eds. Berlin, Heidelberg, Germany: Springer Verlag, pp. 132–143, 2008.

Décodage interactif de la parole

Grégory Senay, Georges Linarès, Benjamin Lecouteux, Stanislas Oger, Thierry Michel

Laboratoire Informatique d'Avignon,
CERI, Université d'Avignon et des Pays de Vaucluse,
Xtensive Technologies
{gregory.senay, georges.linares, benjamin.lecouteux, stanislas.oger}@univ-avignon.fr
thierry.michel@xtensive.com

ABSTRACT

Speech recognition technology suffers from a lack of robustness which limits its usability for fully automated speech-to-text transcription, and manual correction is generally required to obtain perfect transcripts. In this paper, we propose a general scheme for semi-automatic transcription, in which the system and the transcriptionist contribute jointly to the speech transcription. In order to reduce the correction time, we evaluate various strategies aiming to guide the transcriptionist towards the critical areas of transcripts. These strategies are based on graph density-based criterion and two semantic consistency criterion ; using a corpus-based method and a web-search engine. Results show semantic information must be integrated into the interactive decoding process.

Keywords: semi-automatic transcription, speech recognition, transcription aids, speech understanding

1. Introduction

Malgré les efforts réalisés par la communauté scientifique ces vingts dernières années, les performances des systèmes de reconnaissance automatique de la parole (RAP) sont étroitement liées à différents paramètres, notamment les conditions acoustiques, la couverture lexicale et, plus généralement, l'adéquation entre le corpus d'apprentissage et les conditions d'utilisation. Dans un contexte réel d'utilisation, le manque de robustesse des systèmes conduit à réaliser une correction manuelle des transcriptions. Le coût de la transcription doit donc intégrer cette étape de post-traitement manuel.

Dans cet article nous proposons un processus de RAP dans lequel système et opérateur collaborent à l'élaboration d'une transcription conforme aux besoins de l'application.

Plusieurs études ont évalué l'intérêt d'utiliser un système de RAP pour améliorer la productivité des transcripteurs [1]. La plupart d'entre elles propose un processus séquentiel se décomposant en deux étapes, la première consiste à utiliser un système de RAP pour produire une hypothèse de transcription (éventuellement un réseau de confusion), et la seconde consiste à corriger manuellement ces hypothèses. Les résultats montrent un gain assez variable : le temps moyen de correction est réduit de 20 à 80%, en fonction des performances du système de RAP initial.

Ici, nous proposons d'utiliser un processus interactif dans lequel le système et l'opérateur collaborent à l'écriture des transcriptions. Cette interactivité repose sur l'exploitation par le système de RAP des actions correctives de l'opérateur, ce qui lui permet de produire automatiquement une meilleure hypothèse, qui sera à son tour corrigée. Dans cette boucle du processus de correction, la position dans la phrase où les corrections sont appliquées peut être crutiale pour l'efficacité du redécodage ; nous proposons différentes stratégies de guidage du transcripteur basées sur des mesures de confiance et des critères de consistance sémantique.

La section suivante décrit l'algorithme de décodage intéractif que nous proposons. Les differéntes stratégies de guidage du correcteur sont ensuite détaillées, puis évaluées dans le section 4. La dernière section propose un bilan général de l'approche proposée et présente quelques perspectives.

2. Décodage intéractif

2.1. Principe

La meilleure hypothèse de reconnaissance ne constitue qu'une partie de l'information que le système de RAP possède sur les données à transcrire. En effet, le système évalue généralement un grand nombre d'hypothèses concurrentes. Proposer des alternatives au correcteur pourrait améliorer son efficacité [9], par exemple en lui évitant la saisie de certains mots. Cette approche présente néanmoins un certain nombre de difficultés, en particulier la grande diversité de variantes possibles qui de plus ne diffèrent souvent que de quelques mots [4]. Comme proposé dans [2], nous utilisons une représentation basée sur des réseaux de confusion (RC), qui sont plus compacts que les treillis de mots, et de fait plus *lisibles*. Avec une telle représentation, les actions correctives sont réduites à de simples actions d'édition du réseau : sélection d'un mot, suppression ou ajout d'une alternative manquante.

Chaque action corrective effectuée sur le réseau de confusion est suivie d'un redécodage de celui-ci, guidé par l'historique des corrections de l'utilisateur. L'objectif de ce redécodage contraint par les corrections est d'améliorer la transcription, en particulier au voisinage des corrections dont le contexte linguistique se trouve changé. Cette modification locale peut de plus changer globalement le segment de transcription, car

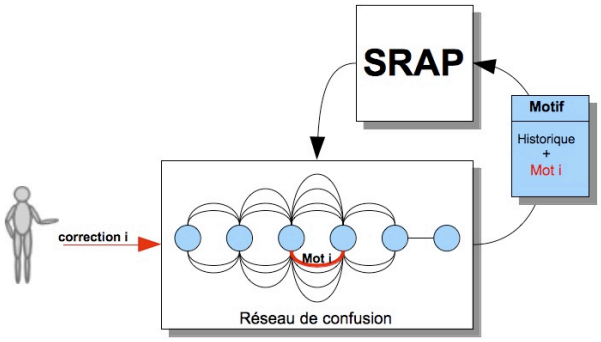


Fig. 1: Decodage interactif

la contrainte locale peut induire des modifications qui se propagent.

2.2. Décodage contraint par un motif

Le principe général du décodage contraint est de fournir au décodeur un motif de phrase, dans lequel les corrections effectuées par l'utilisateur (cf Figure 1) apparaissent comme des mots figés et les zones à redécoder comme des jokers que le système doit trouver. Techniquement, le décodage contraint est implémenté par l'algorithme de décodage guidé que nous avons présenté dans des travaux précédents, en particulier pour l'alignement de transcriptions imparfaites [6] ou la combinaison de systèmes [7]. Nous proposons donc un décodage guidé par le motif issu des corrections appliquées par le correcteur au RC.

Considérant le scénario de correction incrémentale des transcriptions semi-automatiques, on peut penser que l'ordre et les zones dans lesquels les corrections sont appliquées sont suceptibles de modifier sensiblement l'efficacité du décodage. Nous allons évaluer différentes méthodes visant à guider le correcteur vers les zones qui sont supposées être les plus profitables au système en terme de réduction du taux d'erreur mot (TEM) pour chaque acte correctif.

3. Stratégies d'orientation du correcteur

Nous proposons d'exploiter les capacités d'autodiagnostic d'un système pour orienter le correcteur vers les zones qui sont susceptibles d'être utiles au redécodage. La stratégie la plus naturelle est une progression de gauche à droite ; suivant le sens de lecture. On peut aussi penser que les corrections des zones massivement erronées seront particulièrement utiles au système. Une dernière stratégie consiste à estimer les zones sémantiquement incohérentes, dont la correction permet d'améliorer l'intelligibilité de la transcription. De plus, cette approche dans le contexte d'une application de compréhension ou d'indexation, privilégie la qualité sémantique à la classique mesure du TEM.

3.1. Utilisation de la densité du graphe de mots

Les mesures de confiance ont pour but d'estimer la probabilité qu'un mot de la transcription soit correct. Notre objectif est d'utiliser ces mesures pour orienter la correction, afin de maximiser le gain en TEM pour chaque action corrective. La largeur du graphe à un instant t a souvent été utilisée comme un indicateur pour l'estimation des scores de confiances [5], une "explosion" de la largeur étant caractéristique d'une situation d'incertitude de l'algorithme.

En s'inspirant de cette mesure, nous évaluons une stratégie dans laquelle le correcteur est orienté prioritairement vers les zones les plus développées du graphe de mots. Le système sélectionne automatiquement la zone la plus dense du graphe, et propose au transcripteur de la corriger. Chaque action corrective produit un motif de la phrase qui est réinjecté dans le système de reconnaissance via un décodage guidé. L'efficacité de cette méthode est estimée en calculant le gain en TEM issu de cette dernière phase de décodage.

3.2. Utilisation de la consistance sémantique sur un corpus fermé

Cette méthode a pour but d'estimer la consistance sémantique de la transcription par l'utilisation d'un très grand corpus de nouvelles journalistiques. Le principe général est de considérer qu'un segment est inconsistant s'il est significativement distant de la dépêche du corpus la plus similaire. Une fois la plus proche dépêche trouvée, la transcription est découpée en fenêtres plus courtes (10 mots pertinents en moyenne) ; pour chacunes d'entres elles, leurs consistances sémantiques sont évaluées et le correcteur est dirigé en premier vers la fenêtre de moindre consistance.

La similarité entre le segment de transcription et les dépêches repose sur une mesure *Cosinus* [10].Afin de prendre en compte uniquement des mots significatifs, la transcription et le corpus sont lemmatisés et filtrés via une stop-liste contenant les mots les plus fréquents du corpus. Les scores obtenus sont utilisés pour guider le correcteur vers la fenêtre du segment ayant le score le plus faible.

Les données utilisées pour nos expériences sont issues du corpus Français Gigaword qui est une archive de données journalistiques acquise pendant plusieurs années par le Linguistic Data Consortium (LDC). Le corpus a été collecté à partir de deux sources internationales distinctes. La première est l' *Agence France-Presse*, qui fournit des données journalistiques entre 1994 et 2006. Cette première partie contient plus de 480000 mots différents. La seconde vient de l'*Associated Press French Service*, couvrant la même période et contenant environ 180000 mots différents. Tout ce contenu est écrit dans un style journalistique : les dépêches sont relativement courtes ; avec en moyenne une quinzaine de mots par phrase. Les documents sont structurés en paragraphes. Chacun d'entre eux correspond à une dépêche, se focalisant sur un sujet précis. Le corpus contient environ 2 millions de phrases et 250 millions de mots.

3.3. Utilisation de la consistance sémantique à l'aide du Web

L'idée de cette approche est de détecter des ruptures sémantiques dans une phrase en estimant à quel point les mots porteurs de sens sont cohérents entre eux dans leur contexte. Nous considérons que les mots porteurs de sens sont ceux différents des mots-outils de la langue. Nous proposons d'utiliser le Web car il offre une couverture très large de la langue. Le Web est alors utilisé comme une très grande base de documents dans laquelle chacun est vu comme un sac de mots proches d'un point de vue sémantique.

Pour mesurer à quel point un mot porteur de sens est cohérent avec son contexte, et donc à quel point il apparaît dans ce contexte sur le Web, nous proposons d'utiliser la probabilité d'apparition d'un mot dans un document Web, sachant que les mots de son contexte y apparaissent. Cette probabilité est formalisée dans l'équation 1, pour un mot w_i et une probabilité d'ordre n, et donc avec un contexte gauche de taille $n-1$, noté $\psi_i^n = w_{i-n-1}, \ldots, w_{i-1}$:

$$P_s(w_i|\psi_i^n) = \frac{WC(\psi_i^n, w_i)}{WC(\psi_i^n)} \quad (1)$$

Avec $WC(\psi_i^n, w_i)$ le nombre de documents dans lesquels les mots ψ_i^n et w_i apparaissent quelque soit leur position dans le document. Comme avec un modèle n-gramme classique, lorsqu'un mot n'apparaît dans aucun document web où apparaissent les mots de son contexte, on se replie sur la probabilité obtenue avec un contexte plus court, pondérée par un coefficient de repli déterminé empiriquement :

$$\hat{P}_s(w_i|\psi_i^n) = \begin{cases} P_s(w_i|\psi_i^n), \text{ si } WC(\psi_i^n, w_i) > 0 \\ \alpha \cdot P_s(w_i|\psi_i^{n-1}), \text{ sinon} \end{cases} \quad (2)$$

La mesure de cohésion sémantique d'ordre n de la phrase est donc :

$$SC(w_1 \ldots w_i) = P_s(w_2|w_1) \times P_s(w_3|w_1, w_2) \times P_s(w_i|\psi_i^n) \quad (3)$$

Les résultats expérimentaux obtenus avec cette approche sont présentés dans la section 4.2.

4. Expériences

Le système de RAP utilisé est SPEERAL [8], développé au *Laboratoire Informatique d'Avignon*, et la manipulation des RC et des treillis de mots est réalisée avec la boite à outils SRILM [11] développée par l'entreprise *SRI International*. Les expériences sont conduites sur les données de développement de la campagne d'évaluation ESTER 2005 [3], qui est composée de 8 heures de journaux d'informations français provenant de 4 radios différentes. Le décodage initial des données est obtenu avec le système rapide en deux fois le temps réel, sans seconde passe comprenant l'adaptation des modèles acoustiques au locuteur. Dans ces conditions de décodage, le TEM est de 32,6% sur l'ensemble du corpus.

4.1. Protocole

La correction Gauche-Droite, sans décodage réactif, est simulée en utilisant un alignement fourni par l'outil de mesure de la campagne ESTER : *Sclite*[1]. Chaque erreur rencontrée est marquée par *Sclite* comme insertion, suppression ou substitution. L'action corrective est appliquée sur l'hypothèse en fonction de l'indication fournie par cet outil. Les corrections sont effectuées suivant les différentes stratégies de guidage de la correction : Gauche-Droite (GD-ID), densité de graphe (DG-ID), correction sémantique par l'approche basée sur le corpus (Corp-ID) et par l'approche basée sur le Web (Web-ID).

Nous évaluons, dans la partie 4.2, les performances en mesurant la réduction du TEM après chaque action corrective faite dans un segment. Le nombre de corrections effectuées dans un segment est limité à 20. Afin de mettre en valeur les performances de notre approche dans différentes situations, les segments sont séparés en deux classes : la première contient ceux dont le TEM du décodage initial est inférieur ou égal à 40% et la seconde ceux dont le TEM du décodage initial est supérieur à 40%.

4.2. Résultats

Tab. 1: *TEM selon le nombre d'actions correctives, pour les segments dont la transcription initiale est ≤ à 40% de TEM*

#c	1	3	10	20
Manuelle	25.22	22.98	17.23	9.44
GD-ID	24.28	**20.82**	**11.88**	**5.26**
DG-ID	26.58	25.38	16.62	11.76
Corp-ID	**23.90**	21.15	13.93	8.51
Web-ID	24.33	21.10	12.21	7.40

Tab. 2: *TEM selon le nombre d'actions correctives, pour les segments dont la transcription initiale est > à 40% de TEM*

#c	1	3	10	20
Manuelle	55.91	54.05	47.81	40.14
GD-ID	54.95	49.77	37.71	**25.36**
DG-ID	57.51	53.52	44.05	36.99
Corp-ID	54.19	49.37	39.06	29.54
Web-ID	**51.88**	**48.32**	**37.49**	29.49

Dans les tableaux 1 et 2, nous présentons les résultats obtenus en terme de TEM, pour deux classes de segments dont chacun a subi un, trois, dix et vingt actes correctifs (#c). Ces deux classes représentent respectivement 46% et 54% du corpus. Nous constatons que le décodage réactif améliore le TEM dans toutes les configurations par rapport à la correction manuelle (qui pour rappel est sans redécodage).

Pour la première classe (TEM ≤ 40%), la comparai-

[1] Vous pouvez trouver la dernière version de l'outil à cette adresse : http ://www.itl.nist.gov/iad/mig/tools/

son entre les différentes stratégies guidées montre que la correction par densité de graphe (DG-ID) est plutôt inefficace, son rendement étant nettement moins bon que la méthode classique Gauche-Droite (Manuelle - sans décodage interactif). Les stratégies basées sur la sémantique obtiennent de meilleurs résultats, plus particulierement l'approche Web (Web-ID). Néanmoins, elles restent légèrement moins performantes que le décodage interactif utilisant une méthode Gauche-Droite (GD-ID), qui serait probablement plus confortable pour le correcteur.

Le tableau 2 présente les résultats obtenus sur les segments massivement erronés (TEM $>$ 40%). Les résultats obtenus par le guidage sémantique du correcteur sont, dans ce cas, sensiblement meilleurs que toutes les autres stratégies guidées lors des 10 premières corrections. En particulier, la méthode basée sur la consistence sémantique à l'aide du Web (Web-ID) qui procure un gain très significatif dès la première correction (-3,07% absolus). On peut noter que cette méthode est la plus efficace pour 10 corrections, le gain est alors de 10,32% absolus comparativement à la méthode manuelle.

Globalement, l'approche basée sur le Web obtient de meilleurs résultats que la méthode basée sur le corpus, en dépit du fait que les données d'ESTER (journaux radiophoniques) sont étroitement liées au corpus que nous utilisons, composé de dépêches d'informations. Les bénéfices de l'approche Web pourraient être encore plus importants sur les tâches qui se sont pas couvertes par le corpus.

5. Conclusion et perspectives

Nous avons présenté et évalué une approche intéractive du décodage de la parole qui vise à minimiser le coût global de la transcription. L'idée principale est d'alterner des phases de correction et de transcription automatique qui prennent en compte les corrections de l'utilisateur. Considérant que la correction d'une zone bien précise de la transcription pouvait être particulièrement rentable durant un décodage réactif, nous avons proposé différentes stratégies afin d'orienter le correcteur vers les zones "critiques" de la transcription.

Les résultats montrent que le décodage interactif apporte une amélioration notable dans l'efficacité de la correction, comparé à une correction uniquement manuelle. De plus, la comparaison entre les différentes stratégies d'orientation du correcteur permet de tirer plusieurs conclusions. Focaliser les actions correctives sur les zones denses du graphe de recherche est peu efficace. Une des raisons de ce résultat décevant est qu'un système en situation d'échec majeur ne peut pas se raccrocher à un faible nombre de corrections pour se remettre sur le chemin d'un décodage réussi. L'intégration d'un mot correct dans une zone dans laquelle le système hésite n'a pas l'efficacité d'une correction par bloc telle qu'elle est réalisée lors d'une correction gauche-droite. Les stratégies de guidage sémantique n'améliorent pas clairement les parties du corpus où le taux d'erreur mot est bas, mais elles deviennent très efficaces lorsque la transcription est de mauvaise qualité.

Ces résultats ouvrent des perspectives intéressantes dans des contextes d'indexation par le contenu ou plus généralement, d'interprétation ou d'analyse de la parole. Nous envisageons une stratégie permettant d'identifier et de corriger uniquement les parties de la transcription mal décodées afin de les rendre correctement indexables ou interprétables. C'est dans cette perspective que nous développerons ce travail.

Références

[1] Thierry Bazillon, Yannick Estève, and Daniel Luzzati. Manual vs assisted transcription of prepared and spontaneous speech. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

[2] G. Riccardi D. Falavigna, R. Gretter. Acoustic and word lattice based algorithms for confidence scores. pages 1621–1624, 2002.

[3] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *European Conference on Speech Communication and Technology, Interspeech*, Lisbon, Portugal, 2005.

[4] J.Ogata and M.Goto. Speech repair : Quick error correction just by using selection operation for speech input interfaces. In *International Conference on Speech Communication and Technology, Interspeech*, pages 133–136, Lisboa, Portugal, 2005.

[5] Thomas Kemp and Thomas Schaaf. Estimating confidence using word lattices. In *Proc. Eurospeech '97*, pages 827–830, Rhodes, Greece, 1997.

[6] Benjamin Lecouteux, Georges Linarès, J.F. Bonastre, and Pascal Nocera. Imperfect transcript driven speech recognition. In *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA, 2006.

[7] Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Guillaume Gravier. Generalized driven decoding for speech recognition system combination. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Las Vegas, USA, 2008.

[8] Georges Linarès, Dominique Massonié, Pascal Nocera, and Christophe Lévy. The lia speech recognition system : from 10xrt to 1xrt, 2007.

[9] Hiroaki Nanjo, Yuya Akita, and Tatsuya Kawahara. Computer assisted speech transcription system for efficient speech archive. In *Western Pacific Acoustic conference*, Seoul, Korea, 2006.

[10] C. Van Rijsbergen. Information retrieval. Newton, MA, USA, 1979. Butterworth-Heinemann.

[11] A. Stolcke. SRILM-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, USA, 2002.

Reconnaissance vocale basée sur les phonèmes voisés

Matthieu Duvinage, Jean-Yves Parfait

Laboratoire TCTS, Université de Mons (UMONS) - MULTITEL
20, Place du Parc, 7000 Mons, Belgique - Parc Scientifique Initialis, 7000 Mons, Belgique
matthieu.duvinage@umons.ac.be - parfait@multitel.be

ABSTRACT

This paper describes a new approach for robust speech recognition. Inspired from mask theory, the Frame Dropping technique consists in using only highly voiced frames in the decoding process. The idea is to be more robust to noise by using more stable parts. This method was assessed on a database of isolated words in additive noise provided by the Signal Processing Information Base (NOISEX-92). The test showed that there is a deterioration in clean environment. However, the results are quite close to the baseline in adverse conditions. Moreover, it paves the way for global reduction of complexity.
Keywords : speech recognition, Frame Dropping, additive noise, multimodal, embedded systems

1. Introduction

La reconnaissance vocale a depuis quelques dizaines d'années été sources de recherches intensives. Les performances ont ainsi évoluées de manières éloquentes. Cependant, les excellentes performances mises en avant le sont dans le cas de signaux peu bruités. En effet, en présence de bruit, les résultats s'écroulent rapidement non seulement à cause de l'ajout d'un autre signal perturbateur mais également dû à l'effet Lombard. Celui-ci est connu pour notamment allonger la durée des phonèmes voisés (et particulièrement les voyelles) ainsi qu'augmenter leur Rapport Signal à Bruit local (RSB) en vue d'une meilleure intelligibilité [3, 4].

De plus, comme indiqué dans [9], étant donné la multiplication des systèmes embarqués dont les caractéristiques de faibles ressources en font des systèmes compliqués à appréhender, on est en droit de se demander s'il n'est pas possible de réduire le nombre de fenêtres à traiter afin d'obtenir un système globalement moins complexe, et ce, en n'utilisant que de l'information hautement fiable. Ceci permettrait également de ne pas utiliser l'information peu fiable pouvant impliquer des mauvais choix locaux du chemin de Viterbi, c'est-à-dire avec une probabilité très faible pour le phonème correct. Ceci peut encore être pire dans le sens où cette décision locale peut faire en sorte que l'algorithme ne sera plus capable de retrouver le chemin correct des suites des méthodes d'élagage.

Ces dernières remarques mettent le doigt sur la difficulté du problème : trouver de l'information hautement fiable. Dans ce cadre, les phonèmes voisés s'avèrent être une bonne option de par leurs caractéristiques de renforcement en milieu bruité. De plus, ils sont caractérisés par une plus grande stabilité et une allure périodique plus facilement détectable dans un environnement fortement bruité, au contraire des phonèmes non-voisés beaucoup plus proche du bruit.

En outre, une approche multimodale est envisagée. La reconnaissance vocale sur des systèmes embarqués confrontés à des milieux fortement bruités peut tenir compte de la possibilité de l'utilisateur d'interagir. L'évaluation du système se fera donc sur base des N-meilleures solutions parmi lesquelles l'utilisateur peut choisir la réponse correcte.

C'est donc dans cette optique que le comportement d'un système de reconnaissance vocale basé sur les fenêtres fortement voisées (technique de Frame Dropping) est étudié. Ce système est inspiré des méthodes de masques décrites dans [6, 7, 8] qui utilisent soit le rapport signal à bruit (RSB) soit le voisement pour réduire l'espace de recherche. Ce papier propose une nouvelle approche différente des méthodes existantes d'utilisation du voisement basées respectivement sur l'incorporation dans le vecteur de caractéristiques et dans le modèle de Markov [10, 5]. Le présent papier s'organise donc en plusieurs sections. La section 2 explique le choix de l'algorithme de voisement ainsi que son principe. La section 3 décrit brièvement comment ce système est implémenté dans un système existant. La section 4 illustre les résultats obtenus. Les conclusions et travaux futurs sont finalement exposés.

2. Algorithme d'estimation de voisement : YIN

Comme il est voulu de n'utiliser que des fenêtres correspondants à des fenêtres hautement fiables, donc hautement voisées selon notre contexte, il est nécessaire d'avoir un estimateur avec un très faible taux de faux positifs en définissant la classe voisée comme positive. Le choix de l'algorithme YIN est guidé par sa robustesse au bruit étudiée préalablement. Comme l'estimateur YIN fournit une valeur continue et non un choix, le choix est réalisé en calibrant le seuil de décision pour des taux de faux positifs de 1%, 5% et 10% par une analyse en courbe de ROC. Ceux-ci ont été obtenus sur une base de données indépendante segmentée en phonèmes à la main par une experte linguiste. Cette base était destinée à la synthèse vocale et a une couverture phonétique bien répartie.

Par définition, un signal périodique de période T_0 est invariant pour chaque décalage temporel T_0 : $x_t - x_{t+T_0} = 0, \forall t$. Etant donné que la période est la valeur T_0 minimale qui vérifie cette équation, si la période est inconnue, la fonction de différence suivante est nulle pour cette période :

$$d_t(\tau = T_0) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2 = 0$$

où W est la moitié de la taille de la fenêtre utilisée dans l'extraction des caractéristiques. Si le signal n'est pas parfaitement périodique, il suffit de déterminer pour quelle période d_t est minimum. Afin de normaliser les valeurs comme proposé dans [2], l'estimateur final pour la fenêtre n est :

$$d^n_{YIN}(\tau) = \begin{cases} 1 & \text{si } \tau = 0 \\ \dfrac{d_t(\tau)}{(\frac{1}{\tau}) \sum_{j=1}^{\tau} d_t(j)} & \text{autrement} \end{cases}$$

3. Frame Dropping

Dans cette section, l'astuce mathématique qui a permis d'utiliser un système de reconnaissance vocale actuel basé sur un algorithme de Viterbi existant pour implémenter le Frame Dropping est décrite.

Afin de ne pas tout réimplémenter, le Frame Dropping est implémenté d'une manière similaire à [1]. D'un point de vue conceptuel, il faut supprimer la fenêtre non fiable et donc, il faut qu'elle n'ait plus de poids dans la décision du chemin de Viterbi, c'est-à-dire de ne plus tenir compte ni de la vraisemblance de cette fenêtre, ni de l'incrément de temps lié à la présence de celle-ci.
D'un point de vue pratique, cette méthode a été implémentée en mettant toutes les valeurs de vraisemblance à une constante pour la fenêtre correspondante. Pour chaque état j du modèle de Markov et à un temps t, la vraisemblance de chaque chemin du Viterbi est calculée en multipliant les probabilités de transition a_{ij} entre les états et la probabilité d'émission b_j selon ce chemin. La vraisemblance partielle $\sigma_{j,t+1}$ peut s'exprimer :

$$\sigma_{j,t+1} = \max_i [\sigma_{i,t} a_{ij}][b_j(o_t)] \quad (1)$$

Grâce à cette définition, comme proposé dans [1], l'astuce est donnée par :

$$\sigma_{j,t+1} = \max_i [\sigma_{i,t} a_{ij}][b_j(o_t)]^{\gamma_t} \quad (2)$$

où γ_t permet de supprimer mathématiquement la fenêtre. Dans notre cas, γ_t est une variable binaire ($\gamma_t = 1$ si la fenêtre est voisée, $\gamma_t = 0$ dans le cas contraire). De plus, vu que les probabilités de transition a_{ij} sont unitaires, l'influence d'une fenêtre non-voisée est annulée lors de l'algorithme de Viterbi.

4. Evaluation des performances

Dans cette section, le système originel est d'abord décrit. Ensuite, la base de données utilisée est définie ainsi que la méthode d'évaluation. Finalement, les performances obtenues sont discutées.

Le système originel est basé sur un système hybride ANN/HMM utilisant pour chaque fenêtre de 30 ms (décalage de 10 ms) des coefficients jRASTA-PLP et fournissant les N-meilleures solutions. Les bruits utilisés viennent de la base NOISEX-92 (obtenu de Signal Processing Information Base[1]) : un bruit stationnaire de voiture à bande étroite, un bruit non-stationnaire de parole à large bande et un autre bruit non-stationnaire courant enregistré dans une usine. Ces bruits ont été ajoutés à la base de données dans des rapports signaux à bruit moyen de 5 à 15 dB.

La base de données est composée de 1611 enregistrements qui couvrent un vocabulaire de 154 mots français dits de manière isolée. Le bruit fut ajouté à des degrés de rapport signal à bruit différents, et donc, la base de données ne présente pas d'effet Lombard. Ceci représente une limitation de l'étude présentée mais voulu afin d'isoler les effets. L'étude de l'effet Lombard pour cette approche se fera prochainement. De plus, aucune détection d'activité de voix ni de techniques d'élagage n'ont été utilisées.
En ce qui concerne la méthode d'évaluation, celle-ci se base sur la possible interaction de l'utilisateur avec le système. Le système peut donc fournir une liste des N-meilleures réponses à la requête et l'utilisateur choisira la bonne solution. Il faut donc que la solution apparaisse dans la liste, d'où l'importance de l'ordre et la méthode adaptée d'évaluation : la précision à N. Celle-ci se calcule par le rapport du nombre de bonnes réponses fournies sur le nombre d'essais, considérant qu'une bonne réponse est définie par la présence du mot demandé en requête dans les N-meilleurs éléments retournés par le système (dans notre cas, la liste des N-meilleures solutions est de 10).

Les résultats obtenus par la méthode développée sont globalement moins bons que le système originel mais des tendances peuvent être dégagées. Tout d'abord, comme montré sur la figure 1, pour un taux de faux positifs plus faible, c'est-à-dire un nombre plus bas de fenêtres utilisées comme montré à la table 1, la performance diminue. Dans un environnement sans bruit, comme les phonèmes non-voisés représentent une information pertinente vu leur taux de reconnaissance élevés, il est logique de voir les performances diminuer. Néanmoins, comme escompté, dans un environnement fortement bruité, l'écart entre les différents cas devient plus faible par l'utilisation d'information plus fiable. Ceci tend à dire qu'en milieu bruité, il vaut mieux utiliser l'information pertinente. Cependant, l'approche proposée ne conclut pas que le masque basé sur le voisement suffit ou est le bon choix.
Par exemple, pour un bruit de parole à un RSB de 5 dB, les performances s'écroulent. Cependant, les résultats des différentes approches sont quasi-identiques et restent au dessus de la chance (précision à 10 de $\frac{10}{154} = 6.49$ %). En ce qui concerne le bruit de voiture, celui n'affecte pas vraiment les performances et n'a pas été présenté. Ceci s'explique peut-être par la grande robustesse des coefficients jRasta-PLP à ce genre de bruits stationnaires et en basses fréquences. De plus, quand N augmente, l'écart se réduit entre la nouvelle approche et le système originel sauf pour

[1] `http://spib.rice.edu/spib/select_noise.html`

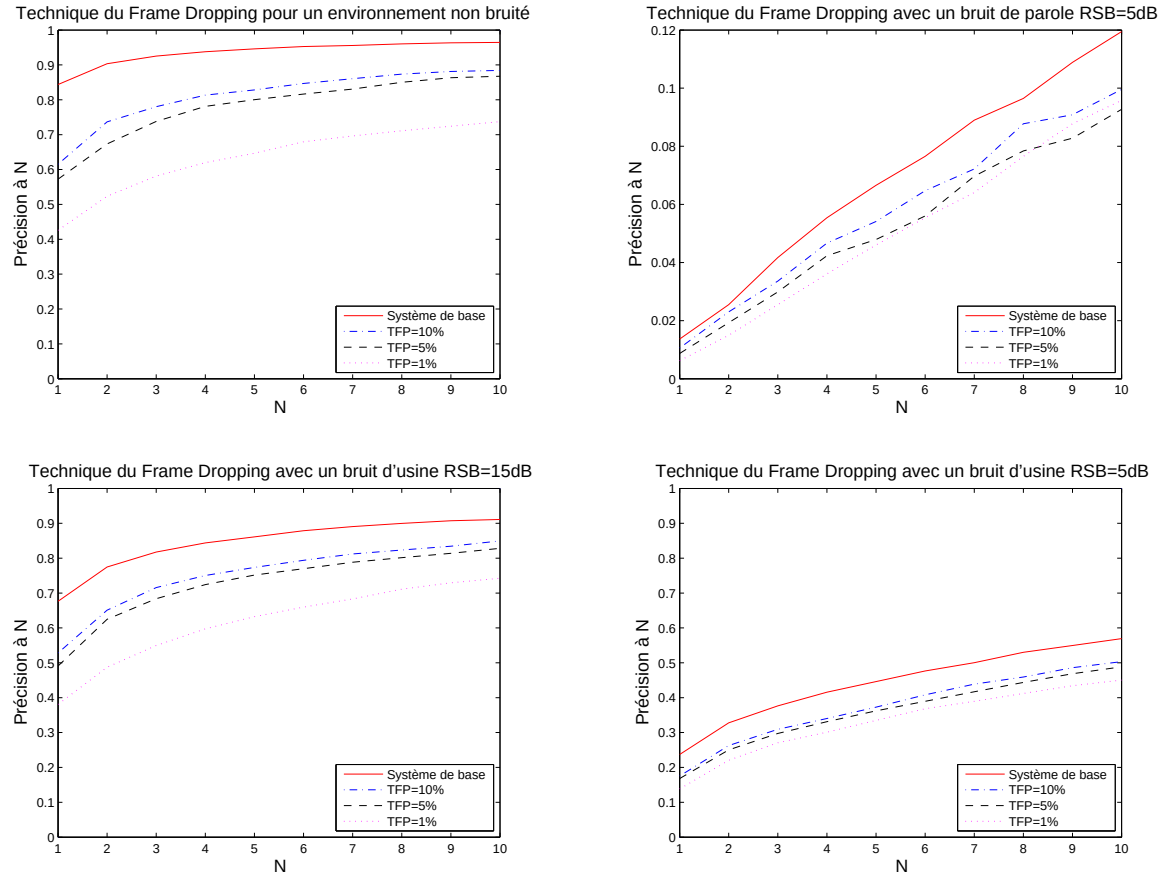


Fig. 1: Au dessus : à gauche, la technique du Frame Dropping amène une diminution substantielle de la précision dans un environnement non bruité. Néanmoins, à droite, la performance s'écroule pour un bruit de parole mais la technique proposée donne des résultats similaires. En dessous : on remarque que les courbes du Frame Dropping se rapprochent du système originel lorsqu'on se situe à des RSB plus bas.

deux cas montrés à la figure 1. Ceci tend à indiquer que la simplification du modèle HMM utilisé augmente la confusion entre mots (deux mots peuvent avoir un modèle plus proche de par le principe de l'approche) pour les N faibles mais permet d'extraire les éléments qui discernent le plus fortement ceux-ci d'où la diminution de l'écart pour les N élevés.
Enfin, la perte de performance peut-être compensée par l'approche multimodale. L'intervention de l'utilisateur peut permettre d'obtenir des résultats similaires à un système basé sur une précision à 1. En effet, on peut voir sur la figure 1 qu'il existe un nombre N_0 du Frame Dropping qui permet d'avoir des performances identiques au système classique ($N = 1$) sauf pour TFP=1% en milieu non bruité.

En ce qui concerne la précision résultante, la robustesse et la complexité du modèle, cette méthode a un bon comportement. En effet, même si la précision globale est affectée par la suppression de fenêtres, le modèle résultant est plus robuste à certaines variabilités dues à la coarticulation, potentiellement à un environnement bruité et aux mauvaises prononciations dans les transitions voisées/non-voisées.
L'effet de coarticulation est moins important car seule la partie la plus stable des phonèmes voisés est prise en compte, c'est-à-dire au milieu des phonèmes. Ceci pourrait laisser entrevoir des performances similaires entre système à contexte et sans contexte, impliquant la possibilité de ne pas devoir calculer les caractéristiques de toutes les fenêtres.
En ce qui concerne le bruit, bien qu'il n'est pas possible de conclure à l'heure actuelle, celui-ci serait moins dérangeant pour deux raisons. L'effet Lombard, bien que non présent ici, augmente l'intelligibilité des phonèmes voisés (en allongeant la longueur et en augmentant le RSB local pour la plupart des phonèmes voisés) en droite ligne avec la procédure proposée. Ce propos doit être nuancé par le désaccord entre modèles et données résultant du décalage en fréquence. D'autres part, le système n'utilise plus les phonèmes déjà semblables au bruit à la base qui provoquaient de nombreuses erreurs dans le réseau de neurones, et par conséquent dans la recherche du chemin de Viterbi. Ce concept permettrait d'être plus robuste aux techniques d'élagages agressifs.
De plus, cette approche supprime aussi les transitions entre éléments voisés et non-voisés non stable ce qui rend le système plus robuste à cette partie.

Finalement, une telle approche diminue la complexité du décodage en diminuant le nombre de fenêtres comme montré à la table 1. De surcroit, comme déjà dit, une étude plus profonde pourrait peut-être montrer que l'utilisation de contexte n'est plus nécessaire dans cette approche et d'où la possibilité de ne cal-

Tab. 1: Le nombre de fenêtres utilisées augmente avec le taux de faux positifs (TFP). Les enregistrements ont une durée fixe de deux secondes sans détection d'activité de voix.

Frame Dropping	TFP=1 %	TFP=5 %	TFP=10 %
Rapport fenêtres utilisées/totales de parole	28 %	48.5 %	61.4 %
Rapport fenêtres utilisées/totales	11 %	19 %	24.25 %

culer les caractéristiques des fenêtres que pour les fenêtres suffisamment voisées. Comme proposé dans [2], l'algorithme YIN peut être optimisé au niveau de la complexité. En outre, d'après l'analyse des résultats, il pourrait être envisagé de ne plus utiliser de détection d'activité de voix. Tout ceci consiste en des pistes afin d'obtenir un système globalement moins complexe, c'est-à-dire ne compensant pas qu'uniquement le surcout de calcul de l'algorithme YIN.

5. Conclusions

Ce papier a étudié une nouvelle approche de Frame Dropping basée sur l'utilisation d'information hautement fiable. Dans ce papier, l'information fiable a été définie comme une fenêtre hautement voisée. Ce concept se base sur sa structure plus facilement reconnaissable dans des milieux fortement bruités et sur l'effet Lombard vis-à-vis des phonèmes voisés.
Les résultats ont montré que cette technique amène une dégradation dans les performances. Cependant, l'écart diminue par rapport au système originel avec un milieu de plus en plus bruité. Ceci tend à confirmer que l'utilisation d'un masque efficace avec de l'information hautement fiable permet d'amener le plus d'information utile dans cet environnement. Le Frame Dropping est aussi plus robuste face aux variabilités telles que la coarticulation et les transitions de phonèmes voisés/non-voisés.
De plus, par l'interaction de l'utilisateur, il est possible d'obtenir des résultats similaires à la méthode originelle à des RSB élevés.
En outre, cette approche possède un potentiel élevé de diminution de complexité, notamment par la possibilité d'utiliser des réseaux de neurones sans contexte, de ne pas calculer les coefficients pour toutes les fenêtres et de ne plus devoir utiliser une détection d'activité de voix poussée.
Finalement, cette approche serait logiquement plus robuste à un élagage agressif.

6. Travaux futurs

D'abord, en ce qui concerne les travaux futurs, il est évidemment indispensable d'étudier cette approche sur une base de données où l'effet Lombard est présent ainsi qu'un élagage agressif afin d'évaluer le réel potentiel au niveau performance. De plus, d'autres informations comme le RSB pourrait être incorporées dans la décision de fiabilité. Cette décision pourrait de surcroit être continue et non binaire comme le permet le cadre théorique développé.
Ensuite, une étude plus précise au niveau de la complexité du système obtenu en étudiant toutes les voies de simplifications évoquées semblent aussi être une voie de recherche pour le futur.
Finalement, vu le comportement au niveau du nombre de fenêtres utilisées, il est en droit de se demander si l'utilisation d'une détection de voisement ne serait pas utile dans une détection d'activité de voix.

Références

[1] Alexis Bernard and Abeer Alwan. Joint channel decoding - viterbi recognition for wireless applications. In *in Proceedings of Eurospeech*, pages 2703–6, 2001.

[2] Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *JASA : Journal of the Acoustical Society of America*, 111 :1917–1930, 2002.

[3] R. Hajislam, Y. Anglade, J.-C. Junqua, and J.-M. Pierrel. Etude acoustique du réflexe Lombard en vue de la reconnaissance de la parole produite en milieu bruité. *Journal de Physique IV*, 04(C5) :C5–485–C5–488, 1994.

[4] John H. L. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Commun.*, 20(1-2) :151–173, 1996.

[5] Peter Jančovič and Münevver Köküer. Incorporating the voicing information into hmm-based automatic speech recognition in noisy environments. *Speech Commun.*, 51(5) :438–451, 2009.

[6] Christopher Kermorvant, Christopher Kermorvant, Andrew Morris, and Andrew Morris. A comparison of two strategies for asr in additive noise : Missing data and spectral subtraction, 1999.

[7] D. O'Shaughnessy and H. Tolba. Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision. In *ICASSP '99 : Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*, pages 413–416, Washington, DC, USA, 1999. IEEE Computer Society.

[8] Michael L. Seltzer, Bhiksha Raj, and Richard M. Stern. Classifier-based mask estimation for missing feature methods of robust speech recognition. In *Proc. ICSLP*, pages 538–541, 2000.

[9] Tan Zheng-Hua and Lindberg Børge. *Automatic Speech Recognition on Mobile Devices and over Communication Networks (Advances in Pattern Recognition)*, chapter 1, pages 1–21. Springer, 2008.

[10] András Zolnay, Ralf Schlüter, Ralf Schl Uter, and Hermann Ney. Extraction methods of voicing feature for robust speech recognition. In *Proceedings of Eurospeech*, pages 497–500, 2003.

Modèles de langage probabilistes et possibilistes basés sur le Web

*Stanislas Oger, Vladimir Popescu, Georges Linarès**

Laboratoire Informatique d'Avignon, Université d'Avignon, France
{stanislas.oger, vladimir.popescu, georges.linares}@univ-avignon.fr

ABSTRACT

Language models are usually built either from a closed corpus, or by using World Wide Web retrieved documents, which are considered as a closed corpus themselves. In this paper we propose several other ways of using this resource for language modeling. We first start by improving an approach consisting in estimating n-gram probabilities from Web search engine statistics. Then, we propose a new way of considering the information extracted from the Web in a probabilistic framework. Then, we also propose to rely on Possibility Theory for effectively using this kind of information. We compare these two approaches on two automatic speech recognition tasks : (i) transcribing broadcast news data, and (ii) transcribing domain-specific data, concerning surgical operation film comments. We show that the two approaches are effective in different situations.
Keywords : language modeling, world wide web, possibility measure, automatic speech recognition

1. Introduction

La qualité des modèles de langage (ML) n-gramme repose essentiellement sur la taille et la qualité des corpus sur lesquels ils sont appris. La couverture linguistique ne peut pas être exhaustive avec un corpus fermé, quel qu'il soit, et à fortiori sur un domaine de spécialité.

Le Web a été largement utilisé pour l'estimation de ML ces dernières années. La plupart du temps, les approches proposées consistent à collecter sur le Web des documents textuels relatifs au contexte linguistique de l'application et à estimer des ML n-grammes sur ces documents [6, 1, 3]. Plus rarement, il est proposé d'estimer les probabilités n-grammes en utilisant directement les statistiques des moteurs de recherche Web [8, 10].

Différents travaux montrent que, d'une façon générale, les ML Web sont bon marché mais moins performants que les modèles appris sur corpus, particulièrement parce que les distributions statistiques des séquences de mots sur le Web ne sont pas fiables [8].

Par contre, le Web est assez exhaustif, et le fait qu'un n-gramme existe ou pas sur le Web pourrait être une information pertinente à intégrer dans un ML. L'intégration des n-grammes impossibles a été étudiée dans [2], où les auteurs produisent automatiquement des n-grammes impossibles et proposent des méthodes pour les intégrer au ML. Un des principaux obstacles évoqué par ces auteurs concerne la génration a-priori des n-grammes impossibles.

Dans cet article, nous présentons un cadre pour estimer des probabilités à partir des statistiques d'un moteur de recherche Web. Nous comparons les résultats obtenus avec ceux de l'approche classique qui utilisent des corpus fermés. Nous reconsidérons la façon dont l'information issue du Web est intégrée au ML : au lieu d'estimer des probabilités n-grammes Web, une mesure *possibiliste* Web [5] est introduite. Nous proposons plusieurs stratégies pour intégrer l'information issue de cette nouvelle mesure dans les ML classiques. Finalement, ces dernières sont évaluées dans le cadre de la RAP sur deux corpus : un corpus broadcast news Français (ESTER), et un corpus spécialisé contenant des vidéos d'interventions chirurgicales (AVISON).

2. Les Probabilités n-grammes Web

Les probabilités n-gramme s'estiment généralement à partir des comptes de séquences de mots issus d'un corpus. L'utilisation du Web pour le calcul de ces probabilités nécessite donc de connaître, pour une séquence de mots donnée, sa fréquence dans au moins une partie des documents Web. Pour cela les statistiques d'un moteur de recherche Web sont utiles : la plupart de ceux-ci fournissent le nombre de documents répondant à une requête donnée, qui peut être une séquence de mots de type n-gramme. A partir du nombre de documents contenant une séquence de mots, se déduit le nombre de n-grammes. Cette approche est présentée dans [10], cependant on peut noter que d'une part les auteurs ne présentent pas de résultat en utilisant directement le nombre de documents comme une estimation du nombre de n-grammes et que d'autre part ils utilisent les statistiques Web uniquement dans le cadre d'un repli du ML initial et n'essayent pas d'utiliser uniquement les probabilités Web comme score linguistique. C'est ce que nous allons présenter ici.

2.1. Estimation des probabilités n-grammes à partir du Web

Nous proposons d'estimer les probabilités n-grammes Web en utilisant directement les statistiques du nombre de document contenant un n-gramme donné. On notera ψ_i^n l'historique de taille $n-1$ du mot w_i : $\psi_i^n = w_{i-n+1} \dots w_{i-1}$. Ainsi, pour obtenir la probabilité Web d'un n-gramme, nous utilisons la formule 1 :

$$P_{\text{web}}(w_i|\psi_i^n) = \frac{H(\psi_i^n, w_i)}{H(\psi_i^n)} \quad (1)$$

Avec $H(S)$, le nombre de documents contenant la séquence de mots S fournie par le moteur de recherche et n est l'ordre du modèle n-gramme. Cependant, la formule 1 n'est pas directement utilisable car elle attribue une probabilité nulle aux séquences de mots qui ne sont pas présentes sur le Web. Ce problème se contourne généralement en enlèvant une part de la masse de probabilité attribuée aux évènements observés pour la redistribuer aux évènements non observés. Etant donné que nous ne disposons pas des statistiques Web nécessaires pour calculer un repli état de l'art tel que Kneser-Ney

*Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (ANR), par l'intermédiaire du projet AVISON (ANR-007-014).

[7], nous interpolons notre distribution avec les distributions d'ordre inférieures. Cette technique est simple, mais a fait ses preuves [7]. Le calcul d'une probabilité se fait donc grâce à la formule 2 :

$$P^*_{\text{web}}(w_i|\psi_i^n) = \alpha_1 \cdot P_{\text{web}}(w_i|\psi_i^n) + \alpha_2 \cdot P_{\text{web}}(w_i|\psi_i^{n-1}) + \ldots + \alpha_n \cdot P_{\text{web}}(w_i) \quad (2)$$

Avec α_i un réel positif tel que $\sum_{i=1}^{n} \alpha_i = 1$. Cette formulation présente tout de même une difficulté relative à l'estimation de la probabilité unigramme. Dans le cadre du Web la fréquence d'un mot est le nombre de documents Web qui le contiennent, et la taille du corpus correspond au nombre total de documents indexés par le moteur de recherche. Pour avoir une estimation de ce dernier, nous avons utilisé le nombre de documents qui contiennent le mot le plus fréquent de la langue étudiée (pour l'anglais le mot *the* et pour le français le mot *le*), en espérant ramener tous les documents de la langue étudiée indexés par le moteur de recherche. La probabilité unigramme ainsi calculée ne sera jamais nulle si on s'assure que tous les mots de notre vocabulaire sont au moins présents dans un document du Web.

Nous avons proposé une estimation de la probabilité n-gramme Web qui n'admet aucune probabilité nulle, même pour des séquences non vues. Nous allons maintenant voir comment cette source d'information peut être intégrée dans le calcul des probabilités linguistiques.

2.2. Les probabilités Web comme coefficient de repli

Une première approche, qui a été proposée dans [10], consiste à utiliser la probabilité Web comme repli du ML appris sur corpus. En faisant cela, une confiance élevée est accordée aux probabilités n-grammes estimées sur corpus, tandis que les probabilités Web sont considérées comme plus fiables que les probabilités de repli. Soit $U_{\psi_i^n}$ l'ensemble des mots "w_i" d'historique "ψ_i^n", de taille $n-1$, pour lesquels le ML initial doit se replier. La formule 3 représente cette approche :

$$\hat{P}(w_i|\psi_i^n) = \begin{cases} \alpha \cdot P_{\text{LM}}(w_i|\psi_i^n) + \\ \quad (1-\alpha) \cdot P^*_{\text{web}}(w_i|\psi_i^n), \\ \quad \text{si } w_i \in U_{\psi_i^n} \\ \beta \cdot P_{\text{LM}}(w_i|\psi_i^n), \text{sinon} \end{cases} \quad (3)$$

avec α un coefficient positif, déterminé empiriquement, et β un facteur de normalisation, défini ainsi :

$$\beta = \frac{1 - \sum_{u \in U_{\psi_i^n}} \hat{P}(u|\psi_i^n)}{1 - \sum_{u \in U_{\psi_i^n}} P_{\text{LM}}(u|\psi_i^n)} \quad (4)$$

2.3. Les probabilités Web comme ML à part entière

Un autre moyen d'utiliser le Web pour construire des ML est de considérer comme fiables les probabilités estimées à partir du Web, et donc de ne pas interpoler ces probabilités avec le ML appris sur corpus. Cette approche est formalisée dans l'équation 5 :

$$\hat{P}(w_i|\psi_i^n) = P^*_{\text{web}}(w_i|\psi_i^n) \quad (5)$$

Cette seconde approche est justifiée lorsque le corpus utilisé pour apprendre le ML initial est trop petit ou mal adapté à la tâche considérée, comme c'est le cas pour le corpus AVISON. Dans la section 4.2, nous présentons les résultats expérimentaux de ces deux approches sur une tâche de RAP.

3. Estimation des possibilités avec le Web

De nombreux travaux montrent comment tirer parti des statistiques issues des séquences de mots sur le Web. Par contre, le fait qu'un n-gramme n'existe pas sur le Web pourrait être une information pertinente à intégrer dans un ML. A notre connaissance, cette approche n'a jamais été étudiée dans la littérature. La théorie des possibilités [5] offre un cadre théorique pour modéliser cette information.

3.1. La théorie des possibilités

La théorie des possibilités est un cadre mathématique permettant de manipuler l'incertain résultant d'une connaissance incomplète [5]. En ce sens, elle est un complément à la théorie des probabilités. Bien que conçue à l'origine pour formaliser la notion d'incertitude linguistique [5], un cadre plus formel lui a récemment été donné, en s'appuyant sur des concepts de la théorie de la mesure [4], et permettant d'utiliser un ensemble de mesures quantitatives pour manipuler les connaissances incomplètes.

Une première notion fondamentale est la notion de *distribution possibiliste*, qui est une fonction π qui associe à chaque élément d'un ensemble d'évènements E, une valeur de l'intervalle unité $[0;1]$. Cette fonction représente l'information distinguant ce qui est plausible de ce qui l'est moins. Par convention, pour un événement e, on a : (i) si $\pi(e) = 0$, l'événement e est impossible, et (II) si $\pi(e) = 1$ l'événement e est totalement possible (plausible).

Comme pour la théorie des probabilités, une mesure possibiliste peut être construite à partir d'une distribution possibiliste, si l'ensemble des événements est fini [4]. Une mesure de possibilité Π peut être définie sur un ensemble d'événements E tel que $\Pi(E) = \max_{e \in E} \pi(e)$, avec π la distribution de possibilités de E. $\Pi(E)$ évalue dans quelle mesure l'ensemble E est cohérent avec la connaissance π.

3.2. Mesure possibiliste basée sur le Web

Dans cette section nous allons montrer comment obtenir une mesure possibiliste pour des séquences de mots, en utilisant des statistiques issues du Web.

La mesure possibiliste doit représenter la possibilité qu'une séquence de mots existe. Pour cela, nous nous appuyons sur l'existence ou non de cette séquence et des sous-séquences la composant sur le Web. Nous entendons ici par *existence* sur le Web le fait qu'il existe au moins un document web contenant la séquence de mots en question. L'idée est que plus de longues sous-séquences de la suite de mots existent sur le Web, plus la suite de mots est possible. Cependant, il est nécessaire de borner la recherche de sous-séquences pour obtenir une mesure fiable. En effet, plus le corpus considéré pour calculer la mesure possibiliste est petit (ici le Web), moins la non-existence de séquences longues sera significative. C'est l'ordre du modèle qui servira de borne.

Tout d'abord, pour l'ordre désiré n du ML, nous construisons récursivement un ensemble distinct de distributions de possibilités π_n à π_1, selon l'équation 6 :

$$\pi_n(W) = \frac{|W_n \cap \text{Web}_n| + \alpha \cdot |W_n \backslash \text{Web}_n| \cdot \pi_{n-1}(W)}{|W_n|} \quad (6)$$

avec W une séquence de n mots ou plus, W_n l'ensemble des séquences de mots de taille n composant W, Web_n est l'ensemble des séquences de mots de taille n sur le Web, $\backslash$ est l'opérateur de différence ensembliste, et $0 \leq \alpha \leq 1$ est un coefficient de repli. La condition terminale de la récursion est $\pi_0(W) = 0$.

Pour une séquence de mots W , la valeur résultante de

cette formule est le nombre de sous-séquences de taille n de W présentes sur le web, normalisée par le nombre total de sous-séquences de taille n de W. La masse de possibilité perdue par l'absence de sous-séquences de taille n sur le Web est redistribuée aux évènements de la distribution possibiliste d'ordre inférieur.

Les distributions possibilistes que l'on vient de définir nous permettent de construire un ensemble de mesures possibilistes correspondantes Π_n, à l'aide de la formule 7 :

$$\Pi_n(\Theta) = \max_{W \in \Theta}(\pi_n(W)) \quad (7)$$

avec Θ un ensemble de séquences de n mots ou plus. Si Θ a un seul élément W, alors $\Pi_n(\{W\}) = \pi_n(W)$.

3.3. Possibilités comme facteur de repli

La mesure de possibilité précédemment définie nous renseigne sur la confiance que l'on peut avoir sur l'existence d'une séquence de mots. Si l'on accorde au corpus d'entraînement d'un ML une confiance plus élevée qu'au Web, alors tous les n-grammes vus dans ce corpus sont totalement possibles ($\pi_n(\psi_i^n, w_i) = 1$). Par contre, les n-grammes composés grâce aux stratégies de repli sont sujet à controverse. Nous proposons donc de pondérer la probabilité qu'accorde le ML aux n-grammes non vus dans le corpus d'entrainement par la possibilité estimée à partir du Web. Cette approche est formalisée dans l'équation 8 :

$$\hat{P}(w_i|\psi_i^n) = \begin{cases} \Pi_n(\{\psi_i^n, w_i\}) \cdot \alpha(\psi_i^n) \cdot P(w_i|\psi_i^{n-1}), \\ \quad \text{si } w_i \in U_{\psi_i^n} \\ \beta \cdot P_{\text{LM}}(w_i|\psi_i^n), \text{ sinon} \end{cases} \quad (8)$$

avec $\alpha(\psi_i^n)$ le coefficient de repli du ML original. De cette manière, la masse de probabilité attribuée à tort à des événements impossibles du point de vue du Web sera redistribuée aux événements vus dans le corpus d'apprentissage, par l'intermédiaire du coefficient β définit dans l'équation 4. Le résultat de cette approche est présenté dans la section 4.3.

3.4. La mesure possibiliste seule

Dans l'approche précédente, les possibilités sont vues comme moyen de repli lorsque le ML n'est plus compétent. Cependant, la mesure possibiliste peut être vue comme une mesure linguistique à part entière et être utilisée seule, par exemple en combinaison avec des scores acoustiques en RAP. A partir de la formule 6 qui définit la mesure possibiliste d'une séquence de mots à partir du Web, il est possible d'obtenir la possibilité d'une séquence de mots de taille variable telle que les hypothèses produites par un système de RAP. La possibilité d'une hypothèse S^m, de taille m avec $m \geq n$, est donc :

$$\Pi_n(S^m) = \pi_n(S^m) \quad (9)$$

Les résultats de ces approches sont présentés dans la section 4.4.

4. Résultats expérimentaux

Les approches proposées dans cet article pour utiliser l'information présente sur le Web pour construire des ML sont évaluées sur deux tâches de RAP : (i) la transcription de broadcast news français, avec un ML initial appris sur un corpus approprié, et (ii) la transcription de commentaires audio relatifs à un domaine spécialisé, en anglais, avec un ML initial appris avec peu de données.

4.1. Configuration expérimentale

Pour évaluer notre approche sur les deux tâches de transcription, nous avons utilisé le système de RAP du LIA, SPEERAL [9]. Ce système est basé sur un décodeur A*, utilise des modèles de Markov cachés pour la modélisation acoustique et un ML 3-gramme.

Pour la tâche de transcription d'émissions radio, nous avons utilisé 6 heures du corpus de test de la campagne d'évaluation ESTER 2005. Le ML initial est un 3-gramme estimé sur un corpus composé de 200M de mots extraits du journal "Le Monde" et d'environ 1M de mots issus du corpus d'entrainement d'ESTER 2005. La technique de lissage de Kneser-Ney modifiée est utilisée et le lexique est composé des 65k mots les plus fréquents de ces corpus. Le taux d'erreur mot (TEM) de ce système sur le corpus de test, sans adaptation au locuteur, est de 24.4%.

Pour la tâche de transcription dans un domaine de spécialité, nous avons utilisé 2 heures du corpus anglais AVISON, qui contient des commentaires audio concernant la chirurgie assistée par la robotique. le ML 3-gramme initial est obtenu en interpolant un ML broadcast news appris sur le corpus HUB4, avec un ML estimé sur toute les transcriptions de référence disponibles dans le corpus d'entrainement d'AVISON. La taille du lexique et la technique de lissage sont les mêmes que précédemment. Ce système obtient un TEM de 33.8% sur le corpus de test d'AVISON.

Intégrer directement les estimateurs Web présentés dans l'algorithme de recherche du moteur de RAP nécessiterait un nombre considérable de requêtes Web. Pour contourner ce problème, nous effectuons un décodage des 100 meilleures hypothèses (100-best) avec un ML initial appris sur corpus, et nous utilisons les techniques Web proposées ici pour réordonner ces hypothèses. Ce problème est contourné de la même manière dans [10]. Le moteur de recherche Google est utilisé pour effectuer les requêtes Web.

4.2. Résultats des probabilités n-grammes Web

Nous présentons ici les résultats de l'approche présentée dans la section 2, qui utilise les probabilités Web.

La partie gauche du tableau 1, intitulée "Web seul", montre les résultats de l'approche présentée dans la section 2.3, qui consiste à utiliser le ML probabiliste Web seul. Le tableau contient les résultats de modèles Web d'ordres (n) différents, pour les deux tâches considérées : ESTER et AVISON. Des ML initiaux d'ordres élevés ($n \geq 3$) ont été construits pour les comparer aux modèles Web. Les résultats de ces nouveaux ML initiaux sur la tâche de réordonnancement des 100-best sont présentés dans la partie droite du tableau 1, intitulée "ML initiaux".

On observe que le ML Web 3-gramme donne des résultats semblables au ML initial sur le corpus AVISON, et un peu moins bons sur le corpus ESTER. Par contre, lorsque l'ordre du ML Web augmente, on observe une diminution significative du TEM, alors qu'augmenter l'ordre du ML initial n'améliore que très peu les résultats. Le ML Web permet de réduire le TEM de 2% absolus sur le corpus AVISON et de 0.7% sur le corpus ESTER, pour $n = 6$, comparé au ML initial d'ordre 6. Ces résultats indiquent que les statistiques issues du Web ne sont pas fiables pour les n-grammes trop petits ($n \leq 3$).

La partie centrale du tableau 1, intitulée "Repli Web", contient les résultats de la seconde approche, décrite dans la section 2.2, qui consiste à utiliser les n-grammes Web comme coefficients de repli, avec un coefficient $\alpha = 0.9$ qui a accordé un poids de 90% aux n-grammes Web.

Comme pour la précédente approche, les résultats sont meilleurs avec les ML Web d'ordres élevés sur les deux corpus. Avec $n = 6$ on obtient une diminution de 2.2% absolue du TEM sur le corpus AVISON, par rapport au ML initial. Cela indique que les probabilités Web d'ordre faible sont moins fiables que les probabilités de repli du ML initial.

Tab. 1: TEM [%] du réordonnancement des 100-best avec différents ML, pour les corpus AVISON et ESTER, en fonction de l'ordre n des ML.

	Web seul		Repli Web		ML initiaux	
n	ESTER	AVISON	ESTER	AVISON	ESTER	AVISON
3	24.8	33.7	24.7	33.7	24.4	33.8
4	23.7	32.8	24.2	32.0	24.2	33.5
5	23.6	32.5	24.2	31.3	24.2	33.3
6	23.5	31.3	24.1	31.1	24.2	33.3

4.3. Les possibilités comme coefficient de repli

Cette section présente les résultats de l'approche décrite dans la section 3.3, qui consiste à utiliser la mesure possibiliste Web comme un coefficient de repli. On voit dans la partie gauche du tableau 2, intitulée "Repli poss.", les performances de cette approche sur les corpus AVISON et ESTER, en fonction de l'ordre n du modèle.

On ne constate pas d'amélioration sur le corpus ESTER. En revanche, le corpus AVISON profite d'une amélioration significative, même pour les modèles d'ordres faibles, ce qui confirme la mauvaise qualité des probabilités de repli du ML initial d'AVISON. On constate une baisse absolue de 1.7% du TEM avec le modèle possibiliste d'ordre 6.

4.4. Les possibilités Web seules

Cette section présente les résultats de l'approche qui consiste à utiliser la possibilité Web comme seul score linguistique. Elle est décrite dans la section 3.4. Les possibilités Web sont combinées aux scores acoustiques en utilisant un facteur d'échelle déterminé de manière empirique.

La partie droite du tableau 2, intitulée "Poss. seules", contient les résultats expérimentaux de l'approche qui consiste à calculer un score possibiliste sur les hypothèses entières en utilisant directement la formule 6. Le TEM obtenu sur les corpus ESTER et AVISON sont présentés, en fonction de l'ordre n du modèle.

Sur le corpus ESTER, on constate une amélioration du TEM de 0.5% absolu avec le modèle possibiliste d'ordre 5, alors qu'une amélioration de 0.7% est obtenue avec les probabilités Web décrites dans la section 4.2. Par contre sur le corpus AVISON, l'amélioration du TEM est de 2.9% absolu avec $n = 6$, ce qui est mieux que l'amélioration obtenue avec les n-grammes Web sur ce corpus. Ces résultats indiquent que les probabilités Web apportent une amélioration par rapport aux ML bien appris, et les possibilités Web améliorent les ML appris avec peu de données.

5. Conclusion

Nous avons proposé deux manières d'utiliser les données Web comme source d'information linguistique : (i) un cadre probabiliste utilisant les comptes de documents des moteurs de recherche Web, et (ii) un cadre possibiliste pour tirer parti de l'existence ou non d'une séquence de mots sur le Web.

Les résultats montrent que la méthode proposée pour

Tab. 2: TEM [%] des ML possibilistes obtenus sur ESTER et AVISON, en fonction de l'ordre n des ML.

	Repli poss.		Poss. seules	
n	ESTER	AVISON	ESTER	AVISON
3	24.5	32.1	24.1	31.8
4	24.3	31.9	23.8	31.5
5	24.3	31.9	23.7	30.6
6	24.3	31.7	23.9	30.4

estimer des probabilités n-grammes Web est meilleure que la méthode classique d'apprentissage des ML sur corpus ou que l'utilisation du Web comme facteur de repli. De plus, nous avons prouvé que la mesure possibiliste Web comme seul score linguistique améliore significativement les résultats dans le cadre du domaine de spécialité où nous disposons de peu de données d'entraînement. Les probabilités Web permettent d'obtenir une réduction absolue du TEM de 0.7% sur ESTER et les possibilités Web permettent de réduire de 2.9% absolu le TEM sur le corpus de spécialité AVISON. Nous avons montré que la théorie des possibilités nous fournit des outils efficaces pour manipuler l'information provenant du Web dans le contexte de la RAP.

Nous projetons maintenant d'essayer d'intégrer les mesures Web plus tôt dans le processus de RAP en modifiant directement les treillis de mot afin de couper les hypothèses les moins possibles ou les moins probables.

Références

[1] A. Berger and R. Miller. Just-in-time language modelling. In *Proc. ICASSP*, volume 2, pages 705–708, 1998.

[2] A. Brun, D. Langlois, K. Smaïli, and J.-P. Haton. Improving statistical language models by removing impossible events. In *Proc. SPECOM*, 2001.

[3] I. Bulyko, M. Ostendorf, and A. Stolcke. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proc. HLT-NAACL*, volume 2, pages 7–9, 2003.

[4] G. de Cooman. Possibility theory I : the measure- and integral-theoretic groundwork. *International Journal of General Systems*, 25 :291–323, 1997.

[5] D. Dubois. Possibility theory and statistical reasoning. *Computational Statistics and Data Analysis*, 21 :47–69, 2006.

[6] M. Federico and N. Bertoldi. Broadcast news LM adaptation over time. *Computer Speech & Language*, 18(4) :417–435, 2004.

[7] J.T. Goodman. A bit of progress in language modeling extended version. Technical report, Microsoft Research, 2006.

[8] M. Lapata and F. Keller. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2 :3, 2005.

[9] P. Nocéra, C. Fredouille, G. Linarès, D. Matrouf, S. Meignier, JF Bonastre, D. Massonié, and F. Béchet. The LIA's french broadcast news transcription system. In *SWIM*, 2004.

[10] X. Zhu and R. Rosenfeld. Improving trigram language modeling with the world wide web. In *Proc. ICASSP*, volume 1, pages 533–536, 2001.

Découverte non supervisée de mot(if)s dans le signal de parole

Armando Muscariello, Guillaume Gravier, Frédéric Bimbot

IRISA UMR 6074 & INRIA Rennes
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
{armando.muscariello,guillaume.gravier,frederic.bimbot}@irisa.fr

ABSTRACT

We propose a method to automatically discover repeating acoustic patterns in speech signals in an unsupervised manner, allowing variability between occurences of a pattern. The resulting patterns, known as audio motifs, are mostly words or sequences of words characterics of the audio content. In this paper, we formalize the problem of motif discovery in speech signals and describe a practical solution using DTW and exploiting the local repetitiveness of motifs. Experimental results on the motif discovery task are provided on a large radio broadcast news corpus. We also propose a refinement of the DTW-based method to account for more variability.
Keywords : motif discovery, audio keyword, unsupervised learning, *data mining*, DTW

1. INTRODUCTION

Dans de nombreuses applications, il est utile de résumer un contenu afin d'en permettre une appréhension rapide. Ainsi, pour les textes, on a généralement recours à quelques mots ou phrases clés tandis qu'en vidéo, on utilise des images clés présentées sous forme d'icônes. En revanche, appréhender un contenu audio directement à partir du signal reste problématique. Dans le cas de contenus oraux, il est évidemment possible d'utiliser une transcription automatique pour se ramener au cas du texte. Mais le processus de transcription automatique est coûteux et parfois peu fiable. La détection de mots clés, ou *word spotting*, présente une alternative intéressante mais limitée à une liste de mots prédéfinis.

Nous étudions ici une approche radicalement différente basée sur la découverte de motifs dans le signal pour faire émerger des icônes sonores correspondant à des mots ou des locutions caractéristiques d'un contenu. La découverte de motifs sonores consiste à détecter à partir du signal des éléments acoustiques réccurents présentant éventuellement un certain degré de variabilité, sans aucune forme de connaissance *a priori*, tant sur le plan acoustique que linguistique. Par exemple, dans le cas de la parole, les mots ou locutions qui se répètent sont des motifs typiques que nous souhaitons voir émerger.

Il convient de bien distinguer la *découverte* de motifs de la *recherche* de motifs. Dans le premier cas, les motifs ne sont pas définis *a priori* tandis que dans le deuxième cas, il s'agira de retrouver un motif connu et défini à l'avance, par exemple par une occurence de référence. Par ailleurs, il est également important de noter que nous souhaitons développer des approches non supervisées dans lesquelles aucune forme d'apprentissage n'intervient. En particulier, nous ne souhaitons utiliser ni modèle de langage, ni modèle acoustique prédéfinis.

Dans le domaine audio, quelques travaux récents s'intéressent au problème de la découverte de motifs. En particulier, Herley propose un algorithme de découverte de motifs sonores quasi invariants pour la découverte d'éléments récurrents (génériques, publicités, *etc.*) dans un flux télévisé [1]. De récents travaux sur la découverte de mots dans le signal de parole relève le défi de la variabilité des motifs [5, 4, 3]. Les approches proposées dans [5] et [4] s'appuient sur un algorithme en deux passes : une première passe vise à détecter des fragments similaires qui sont regroupés dans une passe suivante. Dans [3], nous proposons une approche combinant la stratégie en une passe de [1] avec les méthodes de comparaison de séquences basées sur l'alignement temporel dynamique (DTW). Dans cet article, nous étendons l'approche présentée dans [3] afin d'accroître la robustesse de l'algorithme à la grande variabilité du signal de parole.

Nous formalisons tout d'abord le problème de la découverte de motif avant de détailler l'architecture générale de l'approche proposée. Nous détaillons à la section 4 différentes méthodes pour la comparaison de deux séquences sonores. Les résultats expérimentaux sont rassemblés dans la section 5.

2. FORMALISATION DU PROBLÈME

De manière tout à fait générique, la découverte de motifs consiste à trouver dans un ensemble de données ϕ toutes les paires de segments disjointes, de longueur minimal L_{min}, suffisamment proches. Formellement, on cherche les paires ϕ_a^b, ϕ_c^d telles que

$$H(\phi_a^b, \phi_c^d) < \epsilon \ , \tag{1}$$

où H est une mesure de la distance entre les deux segments, sous les contraintes $b-a > L_{\text{min}}$ et $a < b < c < d$.

Ainsi formulée, la découverte de motifs a pour but de trouver des paires de segments similaires, regroupant ainsi deux occurences d'un même motif. Une étape supplémentaire de *clustering* est ensuite nécessaire pour grouper l'ensemble des occurences d'un motif. Une telle considération nous amène à envisager le problème de découverte de motifs comme un problème de *clustering* se limitant aux portions de signal qui se répètent au moins une fois. Une telle approche s'applique aussi bien lors d'un traitement *a posteriori*, par exemple avec une stratégie multipasse lorsque l'ensemble des données est accessible [5, 4], que pour un

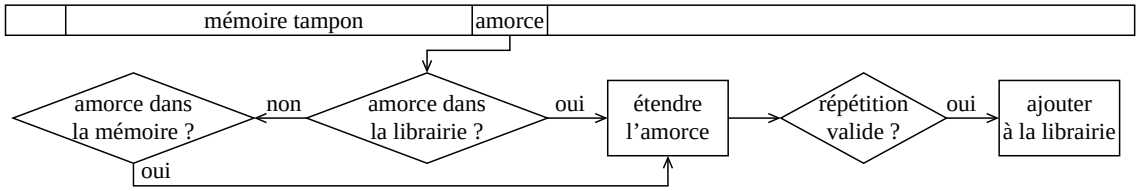


FIG. 1: Schéma de principe de la segmentation du flux et de la recherche pour une amorce donnée.

traitement en flux [1, 3]

Du point de vue conceptuel, nous pouvons décomposer la découverte de motifs en quatre tâches élémentaires : représentation, segmentation, détection et validation. La *représentation* consiste à choisir les descripteurs utilisés pour représenter le signal. La *segmentation* recouvre l'organisation du processus en terme de segmentation des données et d'organisation de la recherche. En effet, une recherche exhaustive de toutes les paires vérifiant (1) n'est bien évidemment pas possible et le recours à une forme de segmentation s'avère indispensable. En particulier, le premier choix à effectuer est celui de la stratégie en une ou plusieurs passes. Enfin, les deux dernières tâches sont directement liées à la comparaison de segments et à la découverte des motifs. La *détection* consiste à identifier les répétitions ϕ_a^b, ϕ_c^d susceptibles de correspondre à deux occurences d'un motif. La *validation* permet par la suite de décider si deux répétitions correspondent en effet à un motif. Cette dernière tâche revient à vérifier (1). Bien que conceptuellement différentes, les tâches de détection et de validation peuvent se résumer en une seule si la même métrique H est utilisée pour les deux.

3. ARCHITECTURE GÉNÉRALE

Nous proposons une approche permettant un traitement en flux des données, dérivée de l'approche ARGOS [1] pour la segmentation. L'idée générale consiste à construire séquentiellement, de manière incrémentale, un catalogue de motifs à partir des données vues comme un flux. Dès lors qu'une nouvelle répétition est trouvée et validée, une nouvelle entrée est créée dans le catalogue, permettant ainsi de retrouver ultérieurement d'autres occurences de ce motif.

La détection des répétitions exploite la notion d'amorce, une amorce correspondant à un segment court, de taille fixé, dans le flux. Une amorce est vue comme un fragment de motif potentiel dont on cherche, dans la phase de détection, à trouver une répétition. Si une répétition de l'amorce est trouvée, on étend alors les segments répétés pour déterminer la répétition la plus longue possible. Cette répétition est ensuite validée comme occurence d'un motif dès lors que les deux segments sont suffisamment proches et insérée dans le catalogue. Afin de limiter le coût calculatoire et de permettre un traitement en flux, la recherche d'une répétition d'une amorce $\phi_t^{t+\delta}$ est limité au passé immédiat $\phi_{t-\Delta}^t$ conservé dans une mémoire tampon. La taille de l'amorce est étroitement liée à la taille minimum des motifs. En effet, l'amorce correspond à un hypothétique fragment de motif et, dans la mesure où l'on cherche une répétition de l'amorce complète, il est important qu'elle ne contienne pas de signal n'appartenant pas au motif lorsque l'amorce est effectivement un fragment de motif. Pour garantir cette propriété, on fixe $\delta = L_{\min}/2$.

Les étapes de l'algorithme sont illustrées par la figure 1. Pour une amorce donnée $\phi_t^{t+\delta}$, on cherche dans un premier temps si cette amorce fait parti d'un motif connu, référencé dans le catalogue, ce dernier étant initialement vide. Si oui, on étend alors l'amorce pour vérifier qu'elle correspond au motif référencé dans le catalogue, remettant à jour le modèle du motif dans le catalogue le cas échéant. Dans nos travaux, le modèle de chaque motif est obtenu par moyennage des occurences trouvées. Si aucun motif du catalogue ne correspond, on cherche dans la mémoire tampon si il existe une répétition de l'amorce de manière à trouver deux occurences candidates pour un nouveau motif par extension de l'amorce. Si un nouveau motif est ainsi découvert, il est ajouté au catalogue après validation. L'algorithme se poursuit ensuite à partir d'une nouvelle amorce localisée soit juste après l'amorce courante si aucun motif n'a été trouvé, soit juste après l'occurence de motif trouvé.

4. DÉTECTION ET VALIDATION

Dans le cadre de segmentation que nous venons de présenter, les tâches de détection et de validation interviennent à deux niveaux, lors de la comparaison avec les entrées du catalogue et lors de la recherche d'une répétition dans la mémoire tampon. Nous décrivons tout d'abord une technique de détection de motifs candidats utilisant une variante segmentale de la technique d'alignement temporel dynamique (DTW) avant de discuter de la validation des répétitions comme occurences d'un motif.

4.1. Détection par DTW segmentale

Rappelons tout d'abord que la phase de détection d'une répétition à partir d'une amorce est un processus en deux étapes. On cherche une répétition de l'amorce – dans le catalogue ou dans la mémoire tampon – avant d'étendre la correspondance de manière à trouver le fragment répété le plus long possible. Nous rappelons ici le principe général de ces deux étapes décrites en détail dans [3].

Considérons une amorce $\phi_t^{t+\delta}$ à rechercher dans un segment χ de longueur $l \gg \delta$. Cette recherche se fait par un algorithme de DTW dans lequel les contraintes de début et fin d'appariement sont relachées, de manière à trouver le fragment de χ apparié au mieux avec l'amorce. Le résultat est un segment χ_s^e tel que sa distance à l'amorce, normalisée par la longueur du chemin d'appariement, notée $D_{\text{DTW}}(\phi_t^{t+\delta}, \chi_s^e)$, est minimum. Les deux segments sont considérés comme une répétition si $D_{\text{DTW}}(\phi_t^{t+\delta}, \chi_s^e) < \epsilon_1$.

La deuxième étape vise à étendre au maximum à gauche et à droite l'appariement existant en s'appuyant sur les points extrêmes. Si l'on prend pour exemple le cas de l'extension à droite (*i.e.*, vers le futur) à partir des deux points $(\chi_e, \phi_{t+\delta})$, on cherche par DTW la meilleure exten-

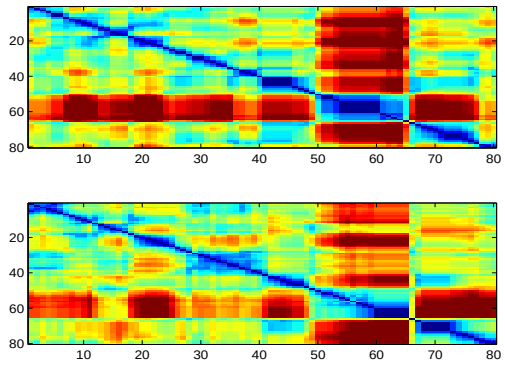


FIG. 2: Exemple de matrices d'autosimilarité d'un motif pour deux locuteurs (masculin en haut, féminin en bas).

sion vers $(\chi_{e+1}, \phi_{t+\delta+1})$, $(\chi_{e+1}, \phi_{t+\delta})$ et $(\chi_e, \phi_{t+\delta+1})$. Le processus d'extension se poursuit tant que D_{DTW} le long du nouvel appariement est inférieure à ϵ_1. Le résultat est une paire de segments, $\phi_{t-\beta_a}^{t+\delta+\alpha_a}$ et $\chi_{s-\beta_b}^{e+\alpha_b}$ telle que $D_{\text{DTW}}(\phi_{t-\beta_a}^{t+\delta+\alpha_a}, \chi_{s-\beta_b}^{e+\alpha_b}) < \epsilon_1$, correspondant à une hypothèse de motif qu'il convient de valider.

L'étape de validation consiste à évaluer (1). La distance D_{DTW} peut être directement utilisée comme métrique H. Cependant, afin d'éviter de valider deux segments différents, cette stratégie requiert un seuil ϵ_1 très petit, limitant ainsi la variabilité tolérée entre occurences d'un motif. En particulier, nous avons observé que cette approche ne permet pas de retrouver des occurences d'un motif par différents locuteurs. Utiliser un seuil ϵ_1 plus élevé autorise une plus grande variabilité au prix d'un nombre plus élevé de fausses détections, c'est-à-dire de détection de répétitions ne correspondant pas à deux occurences d'un motif.

4.2. Validation par matrices d'autosimilarité

Pour pallier au problème précédent, nous proposons une étape de validation exploitant la comparaison de matrices d'autosimilarité. La matrice d'autosimilarité d'une séquence χ_a^b est la matrice carrée $\Phi(\chi_a^b)$ des distances entre points χ_i et χ_j. Clairement, les matrices d'autosimilarité de différentes occurences d'un motif présentent une forte ressemblance visuelle comme illustré par la figure 2. C'est cette ressemblance – interprétable comme une distance entre les autocorrélations plutôt qu'entre les séquences elles-mêmes – que nous souhaitons mesurer et utiliser pour la validation.

La comparaison des matrices d'autosimilarité requiert une normalisation de la longueur des séquences χ_a^b et χ_c^d à comparer, normalisation s'appuyant sur la fonction optimale d'appariement des deux séquences. Étant données les deux séquences normalisées de longueur l, $\tilde{\chi}_a^b$ et $\tilde{\chi}_c^d$, plusieurs métriques sont possibles. La plus simple consiste à prendre la norme l_1 normalisée, soit $D_{\text{SSM}}(\chi_a^b, \chi_c^d) = |\Phi(\tilde{\chi}_a^b) - \Phi(\tilde{\chi}_c^d)|/l^2$. Cette distance reste cependant très dépendante des valeurs absolues des éléments des matrices et ne reflète que peu la similarité visuelle. Afin de prendre en compte la structure spatiale des matrices d'autosimilarité, nous avons recours à une technique basée sur les histogrammes de gradients orientés [2][1]. L'idée générale d'une telle approche est que l'apparence locale d'une matrice d'autosimilarité se caractérise bien par la distribution des gradients d'intensité locaux. Chaque matrice est ainsi transformée en un vecteur de caractéristiques locales, composé des histogrammes des gradients d'intensités pris localement en divers points. La distance entre deux matrices est alors définie comme la norme l_1 entre leurs vecteurs de caractéristiques et notée D'_{SSM}.

Les deux métriques D_{SSM} et D'_{SSM} apportent des informations complémentaires sur la structure des matrices d'autosimilarité. La première mesure directement la différence d'intensité entre les entrées de la matrice. En revanche, la seconde est invariante à l'ajout d'une constante à chaque entrée de la matrice. De plus, en ne se limitant pas à des informations ponctuelles, elle permet de prendre en compte une information plus complexe. En partique, on utilisera donc en parallèle les deux métriques pour valider une répétiton comme occurence d'un motif si $D_{\text{SSM}}(\chi_a^b, \chi_c^d) < \epsilon_2$ et $D'_{\text{SSM}}(\chi_a^b, \chi_c^d) < \epsilon_3$.

5. RÉSULTATS

Nous évaluons tout d'abord l'approche par DTW segmentale pour la découverte de mots dans un flux de parole avant de présenter des résultats préliminaires sur les distances D_{SSM} et D'_{SSM}.

5.1. Découverte de mots dans un flux

Nous avons artificiellement créé un flux de 10 h de signal par concaténation de dix enregistrements d'une heure chacun, dans l'ordre chronologique. Les six premières heures (2 h x 3 chaînes) ont été enregistrées sur une période de 15 jours, les quatre premières correspondant au même jour. Les quatre dernières heures, provenant de 4 chaînes différentes, correspondent à une période de 2 jours, éloignée de 18 mois de la première période. Le choix des données répond à deux considérations majeures. D'une part, on trouve de nombreux mots ou séquences de mots présentant à la fois des répétitions à court terme (au sein d'un reportage par exemple) et à long terme (reportage sur le même sujet mais sur une autre station le même jour ou le lendemain). D'autre part, nous disposons sur ces données d'alignements phonétiques permettant de faire correspondre les motifs découverts au niveau acoustique avec une transcription phonétique.

Dans toutes les expériences, le signal est représenté par des vecteurs de 12 MFCC, plus l'énergie, extraits à une fréquence de 100 trames par seconde.

La qualité des motifs découverts est évaluée au niveau phonétique. Rappelons que le résultat du processus de découverte de motifs est un catalogue de motifs, C_i, chacun caractérisé par ses occurences. La transcription phonétique permet d'associer à chaque occurence j de C_i sa transcription phonétique $C_p(i,j)$. Le motif C_i peut alors être représenté au niveau phonétique par son centroïde, défini comme l'élément $C_p(i,j)$ le plus proche de toutes les occurences du motif. La précision d'un motif correspond alors à la proportion d'occurences suffisamment proche du centroïde. Le rappel est défini par rapport à l'ensemble des chaînes phonétiques suffisamment proches du centroïde de C_i dans la transcription phonétique du flux.

[1]Nous tenons à remercier Émilie Dexter et Patrick Pérez qui ont aimablement mis leurs programmes à notre disposition.

TAB. 1: Précision/Rappel (en %) pour la détection de locutions clés dans un flux de 20 minutes

locution	D_{DTW}	$+D_{\text{SSM}}$	$+D'_{\text{SSM}}$
Jean Marie Le Pen	33 / 59	40 / 59	56 / 59
vingt-et-un avril	18 / 71	22 / 71	43 / 71
extrême droite	17 / 57	25 / 57	67 / 57
France	11 / 43	18 / 39	22 / 35

Pour découvrir des motifs correspondant à des mots ou séquences de mots, nous avons fixé la taille de l'amorce à 0,3 s et celle de la mémoire tampon à 120 s. Le seuil ϵ_1 a été réglé empiriquement de manière à obtenir un bon compromis entre rappel, précision et temps de calcul. Sur les 10h de signal, nous avons trouvé environ 300 motifs, avec une précision de 85 % et un rappel de 25 %. Les motifs trouvés sont donc peu entâchés d'erreurs mais la DTW permet difficilement de grouper des occurences d'un motif qui présente une trop grande variabilité, expliquant ainsi le faible rappel. En particulier, la DTW est très dépendante du locuteur et les occurences d'un même motif par différents locuteurs ne sont pas détectées comme un unique motif mais plutôt comme autant de motifs séparés. Augmenter le seuil ϵ_1 permettrait d'augmenter le rappel au prix d'une forte baisse de la précision. En effet, les motifs dans le catalogue sont représentés par la forme moyenne des occurences trouvées pour ce motif. Augmenter ϵ_1 engendre alors un nombre accru de fausses détections qui viennent déteriorer la représentation des motifs dans le catalogue.

De manière qualitative, les motifs trouvés correspondent principalement à des mots ou des courtes séquences de mots. Par ailleurs, plusieurs motifs sans contenu linguistique sont également trouvés. C'est notamment le cas des inspirations et des *jingles*.

Finalement, il convient de souligner que le temps de calcul pour le traitement des 10h de signal a été d'environ 13h. Même si des optimisations permettrait de décroître de manière significative le temps de calcul, ces chiffres mettent en évidence la difficulté du passage à l'echelle de notre algorithme dans le cas de la découverte de mots. En effet, la taille du catalogue de motifs croît rapidement pour ce type de données, ralentissant ainsi l'algorithme. Ainsi, nous avons mesurer que le temps de traitement en fonction du temps dans le flux est une fonction exponentielle (de la taille du catalogue).

5.2. Utilisation des matrices d'autosimilarité

Avant d'utiliser les métriques D_{SSM} et D'_{SSM} pour la découverte de motifs, nous les avons tout d'abord validé dans un cadre de recherche de motifs connus. Nous avons artificiellement construit un signal de 20 minutes par concaténation de six reportages sur le thème du 21 avril 2002, provenant de radios (et donc de locuteurs) différentes. Quatre locutions clés – Jean-Marie Le Pen, vingt-et-un avril, extrême droite, France –, caractérisées par une occurence de référence chacune, sont recherchées dans les 20 minutes de signal.

Les résultats, en terme de rappel et précision des occurences retrouvées, sont présentés dans le tableau 1. L'algorithme de DTW segmental présenté à la section 4.1 peut être utilisé pour cette recherche (colonne 2), l'occurence de référence du motif à rechercher jouant le rôle d'amorce. Les occurences trouvées pour chaque motif sont ensuite validées en utilisant la distance D_{SSM} (colonne 3), éventuellement complétée par D'_{SSM} (colonne 4). Ces résultats mettent clairement en évidence l'intérêt d'une mesure entre matrices d'autosimilarité pour la validation des motifs, permettant ainsi une amélioration substantielle de la précision pour un rappel constant (à l'exception du motif « *France* », très court). Les occurences trouvées correspondent bien à différents locuteurs, tant masculin que féminin.

Des premières expériences sur l'utilisation des distances entre matrices d'autosimilarité pour la tâche de découverte de motif sur ce court extrait de 20 minutes confirment l'intérêt de ces distances. En utilisant conjointement les deux distances, la précision augmente de 52 % à 66 % et le rappel de 42 % à 51 % par rapport à la seule DTW segmentale. Par ailleurs, l'analyse qualitative des résultats montre que des occurences du motif par différents locuteurs sont retrouvées pour certains motifs, comme « *élevage* »ou « *poisson* ».

6. CONCLUSION

Nous avons proposé une approche pour la découverte non supervisée de motifs sonores dans le signal de parole. La plupart des motifs retrouvés correspondent à des mots ou des séquences courtes de mots qui peuvent être utilisés comme mots clés sonores pour caractériser ou indexer un signal. La méthode utilisant l'alignement temporel dynamique permet de detecter des mots clés avec une bonne précision mais présentent un rappel faible. La combinaison de l'alignement temporel dynamique avec la comparaison des matrices d'autosimilarité permet d'améliorer la découverte de motif au prix d'un effort calculatoire supplémentaire. Ce travail ouvre de nombreuses perspectives, tant pour améliorer la méthode que pour intégrer la découverte de motifs dans des applications d'indexation de documents oraux. En particulier, deux problèmes nous semblent cruciaux. D'une part, le passage à l'échelle reste problématique. Par ailleurs, afin d'utiliser efficacement les motifs découverts, il convient de les caractériser afin de ne conserver que ceux qui décrivent effectivement un contenu linguistique.

RÉFÉRENCES

[1] C. Herley. ARGOS : Automatically extracting repeating objects from multimedia streams. *IEEE Transactions on Multimedia*, 8(1) :115–129, Feb. 2006.

[2] I. Junejo, E. Dexter, I. Laptev, and P. Pérez. Cross-view action recognition from temporal self-similarities. In *Proc. European Conf. on Computer Vision*, pages 293–306, 2008.

[3] A. Muscariello, G. Gravier, and F. Bimbot. Audio keyword extraction by unsupervised word discovery. In *Proc. Interspeech*, pages 2843–2846, 2009.

[4] A. Park and J. R. Glass. Unsupervised pattern discovery in speech. *IEEE Trans. on Acoustic, Speech and Language Processing*, 16(1) :186–197, Jan. 2008.

[5] L. ten Bosch and B. Cranen. A computational model for unsupervised word discovery. In *Proc. Interspeech*, pages 1481–1484, 2007.

Estimation d'enveloppes spectrales contraintes temporellement pour la conversion de voix

Elizabeth Godoy [1], *Olivier Rosec*[1], *Thierry Chonavel* [2]

[1] Orange Labs, Lannion, France
[2] Telecom Bretagne, Signal & Communication Department, Brest, France
{elizabeth.godoy, olivier.rosec}@orange-ftgroup.com,thierry.chonavel@telecom-bretagne.eu

ABSTRACT

This paper presents a new approach to estimating the speech spectral envelope that is adapted for Voice Conversion (VC). In particular, we represent the spectral envelope as a sum of peaks that evolve smoothly in time, within a phoneme. We highlight important properties of our proposed spectral envelope estimation and illustrate its potential for use in a VC context. We analyse natural speech using the proposed methods and we compare results with those from a more traditional frame-by-frame cepstrum-based analysis. Subjective comparisons of synthesized speech quality, as well as implications of this work in future research are also discussed.
Keywords: spectral envelope, voice conversion

1. INTRODUCTION

L'estimation de l'enveloppe spectrale est un sujet récurrent en traitement de la parole de part l'importance que revêt cette information dans des applications telles que le codage de la parole ou encore la transformation et la conversion de voix. La pratique courante consiste à effectuer cette estimation trame par trame en considérant chacune de ces trames isolément. Pour ce faire, des méthodes à base de prédiction linéaire ou de cepstre sont couramment employées en vue de générer des paramètres d'intérêt tels que les LSF (Line Spectral Frequencies) ou les coefficients cepstraux [1]. Cette stratégie est adaptée dans de nombreux contextes applicatifs, mais elle trouve néanmoins ses limites dans des applications nécessitant des transformations de l'enveloppe spectrale variant dans le temps [2]. C'est par exemple le cas en conversion de voix où l'on cherche à modifier l'enveloppe spectrale d'un locuteur source de telle sorte que le signal résultant semble avoir été prononcé par le locuteur cible désiré. L'un des problèmes cruciaux dans un tel cadre est de préserver la cohérence temporelle du signal converti. Le processus de conversion de voix est classiquement découpé en trois étapes : premièrement l'analyse visant à extraire des paramètres d'intérêt (*e.g.* l'enveloppe spectrale), deuxièmement l'apprentissage d'une fonction de conversion établissant le lien entre les espaces acoustiques des locuteurs source et cible et troisièmement la transformation proprement dite. Des travaux récents ont porté sur la restitution de trajectoires continues via des méthodes d'interpolation [3]. Ces travaux se concentrent uniquement sur les étapes d'apprentissage et de transformation sans remettre en cause la phase d'analyse.

Dans cet article, nous suggérons d'introduire des contraintes lors de l'analyse de façon à extraire des paramètres d'enveloppe spectrale variant continûment et dont l'évolution temporelle peut être facilement contrôlée lors de l'étape de conversion. Plus précisément, sur la base d'une représentation de l'enveloppe spectrale par somme de pics gaussiens [4], nous prenons en compte les dépendances temporelles entre trames au sein de chaque phone afin de modéliser des trajectoires de pics spectraux. Une telle représentation revêt plusieurs propriétés intéressantes. Tout d'abord, les pics spectraux obtenus sont fortement liés aux formants, ce qui les rend pertinent du point de vue de la perception. En second lieu, cette représentation offre une grande flexibilité dans un contexte de modification de signaux de parole, car elle permet un contrôle fin de la position de la largeur et de l'amplitude de chacun des pics spectraux. Enfin, la modélisation de trajectoires spectrales est cruciale pour pouvoir rendre compte de l'évolution des résonnances du conduit vocal, celle-ci se faisant généralement de manière lisse pour la plupart des sons voisés (ceux-là mêmes qui portent une grande partie de l'identité vocale). En résumé, nous proposons une nouvelle méthode d'estimation de l'enveloppe spectrale adaptée à la conversion de voix en ce sens qu'elle fournit des paramètres liés à la perception, localisés en fréquences et dont l'évolution temporelle peut être contrôlée.

Dans cet article, nous examinons également le potentiel de notre méthode d'analyse à discriminer des événements acoustiques. Ceci est particulièrement important en conversion de voix afin de faire ressortir des différences de timbre pertinentes entre deux locuteurs. Pour cela nous suggérons une métrique basée sur la distance entre les amplitudes des pics spectraux.

L'article est structuré comme suit. La section 2 présente le modèle de pics spectraux utilisé et la méthode d'analyse proposée. Dans la section 3, nous analysons le potentiel de l'approche à effectuer une analyse-synthèse de haute qualité et examinons également ses capacités en terme de discrimation d'événements acoustiques. La section 4 conclut nos propos et fournit des perspectives à cette étude.

2. ESTIMATION DE L'ENVELOPPE SPECTRALE CONTRAINTE TEMPORELLEMENT

2.1. Modélisation des pics spectraux

Notre modélisation est basée sur l'hypothèse que l'amplitude de l'enveloppe spectrale à la fréquence *f*, $S_i(f)$, de la trame d'indice *i* s'écrit comme la somme de M_i gaussiennes :

$$S_i(f) = \sum_{m=1}^{M_i} a_m^i N(f; \mu_m^i, \sigma_m^i)$$

Les paramètres $\{a_m^i, \mu_m^i, \sigma_m^i\}$ désignent respectivement l'amplitude, la position et la variance du pic d'indice *m*. Cette modélisation est à rapprocher de celle proposée dans [4]. Mais à la difference des travaux reportés dans [4] et [5], notre analyse est basée sur la transformée de Fourier Discrète (TFD) plutôt que sur des versions de l'enveloppe spectrale obtenue par lissage spectral, prédiction linéaire ou modélisation cepstrale qui peuvent d'emblée altérer les caractéristiques spectrales du signal. De plus, notre méthode d'estimation des paramètres diffère radicalement de la méthode basée sur l'algorithme EM décrite dans [4]. Nous suggérons ici d'estimer les positions et amplitudes des pics par le biais d'un algorithme de détection de pics utilisant un masque fréquentiel dont la taille est ajustée en fonction du support fréquentiel considéré et de la fréquence fondamentale. Cette taille variable permet une bonne résolution fréquentielle sur les basses fréquences (particulièrement importante sur le plan de la perception) tout en évitant de modéliser les harmoniques elles-mêmes. Le nombre de pics détectés dépend de ce masque est varie donc selon la trame analysée. Nous autorisons cependant un nombre maximum de 20 pics par trame. Une fois que les positions et amplitudes des pics ont été déterminées, nous calculons la variance de chaque pic en considérant la partie du spectre au voisinage immédiat du pic en question. De cette manière nous évitons les interférences entre pics qui pourraient conduire à des sur-estimation des amplitudes. Plus précisément, entre deux pics nous déterminons le point du spectre généré par les deux pics de manière équiprobable. Ce point permet de calculer les variances de l'échantillon à droite et à gauche respectivement des pics de gauche et droite. La variance d'un pic est alors la moyenne de ces variances à gauche et à droite. Un soin tout particulier doit être accordé au premier pic spectral pour ne pas que sa variance soit influencée par les valeurs de la TFD en deçà de f0. L'enveloppe est ainsi calculée en assurant une interpolation lisse entre les amplitudes des premier et deuxième pics.

La Figure 1 montre les spectres d'amplitude obtenus par TFD, par notre méthode et par une modélisation par cepstre discret d'ordre 40 et calculé sur une échelle Bark conformément à [6]. Les deux approches permettent de capturer la forme globale de l'enveloppe spectrale, la modélisation cepstrale se révélant davantage apte à modéliser les détails. Pour pouvoir apprécier les différences entre les deux méthodes, il convient tout d'abord de noter que la méthode du cepstre discret n'introduit pas de fortes contraintes sur la forme de l'enveloppe spectrale. Notre modélisation cherche quant à elle à capturer les pics spectraux principaux puis à estimer des trajectoires entre ces pics. Ainsi, nous sacrifions donc une certaine précision à l'échelle de la trame, dans le but de faire émerger une structure spectrale à l'échelle phonémique.

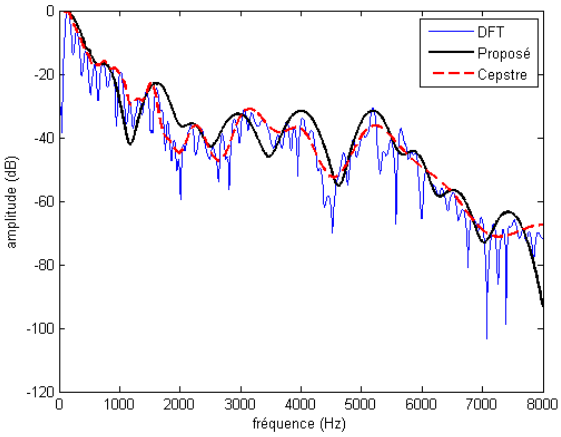


Figure 1: Spectre d'amplitude d'une trame de parole obtenu par TFD (en bleu) et envelopes spectrales estimées par la méthode du cepstre (en pointillé rouge) discret et la méthode proposée (en noir).

2.2. Suivi de l'évolution des pics spectraux à l'échelle phonémique

Un point clé de notre méthode est de contraindre le processus d'estimation de l'enveloppe spectrale par la prise en compte explicite de l'évolution temporelle des pics spectraux à l'échelle d'un phone. Pour cela, nous supposons qu'une segmentation associée au signal de parole est disponible. L'hypothèse sous-jacente est que les pics spectraux ne doivent pas varier trop drastiquement d'une trame à l'autre pour un phone donné. Nous introduisons donc une étape de suivi des pics spectraux préalablement déterminés.

Nous nous intéressons tout d'abord à la partie stable d'un phone. Cette zone est définie pour la majorité des phonèmes (à l'exception toutefois des consonnes plosives), et peut être localisée au voisinage du milieu du phone. Nous définissons alors arbitrairement cette partie stable comme le triplet de trames constitué de la trame centrale d'un phonème et de ses voisines immédiates. Notons que si le phone est déclaré voisé nous effectuons une analyse pitch-synchrone ; dans le cas contraire l'analyse est faite avec un pas de *10* ms. L'intérêt de cette partie stable est que le signal peut y être considéré comme stationnaire et que les effets de coarticulation y sont minimums. L'analyse y est donc a priori plus fiable. Nous suggérons alors d'exploiter ces propriétés afin d'obtenir des points d'ancrage pertinents pour guider notre

procédure d'analyse. Nous commençons alors par une analyse de ces trois trames prises individuellement. Nous sélectionnons les deux trames les plus proches sur la base d'une distance euclidienne entre spectres d'amplitude déduits des représentations spectrales et nous alignons les pics de ces deux trames. Cet alignement est réalisé en minimisant localement la différence entre les positions fréquentielles des pics. La moyenne des paramètres de ces pics alignés définit alors les paramètres de la trame centrale du phone. Ensuite, nous faisons de part et d'autre de cette trame stable une analyse de proche en proche jusqu'aux frontières de phone. Pour une trame donnée, cette analyse est contrainte de manière à ne sélectionner que les pics suffisamment proches de ceux détectés pour la trame adjacente précédemment analysée. Cette méthode d'analyse assure ainsi une évolution graduelle des paramètres à l'échelle d'un phone, tout en privilégiant les parties sur lesquelles l'estimation de l'enveloppe spectrale est la plus fiable, *i.e.* les parties stables.

La figure 2 montre une vue 3D de la séquence de spectres d'amplitude déduits des enveloppes spectrales estimées sur un exemple de réalisation acoustique du phonème 'A'. Notons que nous n'avons représenté que la partie du spectre comprise entre 0 et 4 kHz et les amplitudes sont normalisées. Sur cette figure nous observons que la modélisation proposée fournit une représentation temps-fréquence lisse permettant clairement d'identifier des trajectoires spectrales sur l'ensemble du phone. A l'inverse, la modélisation par cepstre discret met en évidence des variations sporadiques et discontinues de l'enveloppe spectrale.

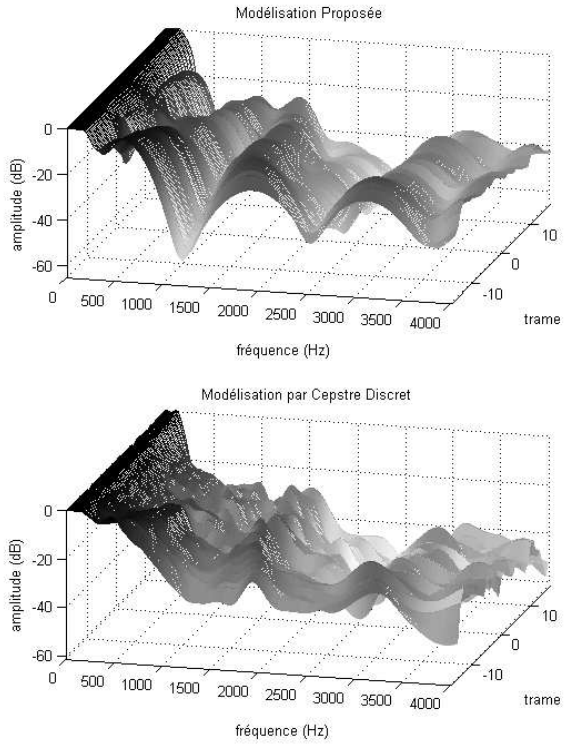


Figure 2: Une séquence de spectres d'amplitude dans une réalisation du phonème 'A'. Le centre du phone est à 0.

3. ILLUSTRATION DU POTENTIEL DE LA METHODE PROPOSEE

3.1. Qualité accessible en analyse-synthèse

Dans cette section nous appliquons notre nouvelle méthode d'analyse de l'enveloppe spectrale dans un contexte d'analyse-synthèse. Notre expérience porte sur 4 corpus de parole (i.e. 4 voix, 2 féminines et 2 masculines) utilisés dans le système de synthèse vocale Baratinoo de France Télécom. Ces corpus ont été enregistrés en studio par des locuteurs professionnels. Les signaux sont échantillonnés à 16 kHz, étiquetés, segmentés en phones et pitch-marqués. Pour l'analyse et la resynthèse des signaux nous utilisons un modèle HNM (Harmonic plus Noise Model) [6]. Dans cette expérience, nous avons fixé la fréquence maximale de voisement des trames voisées à 8 kHz. Les amplitudes des harmoniques sont obtenues par échantillonnage de l'enveloppe spectrale. Les phases utilisées à la synthèse sont celle estimées lors de l'analyse HNM. Cette méthodologie nous permet ainsi de comparer uniquement l'effet de l'estimation de l'enveloppe spectrale sur le résultat de la synthèse. Sur cette base nous avons donc procédé à un test d'écoute informel dont l'objectif était de comparer la méthode d'estimation proposée à celle du cepstre discret. Ce test d'écoute a montré une légère dégradation de la qualité en analyse-synthèse. Une telle dégradation était prévisible dans la mesure où, comme mentionné précédemment, l'enveloppe estimée par notre méthode est moins précise. Des tests supplémentaires faisant intervenir des modifications de timbre variant dans le temps sont nécessaires pour pouvoir mieux apprécier le potentiel de notre méthode en tant que brique d'analyse-modification-synthèse. Un objectif est bien entendu de tester cette approche dans le cadre de la conversion de voix.

3.2. Caractérisation des espaces acoustiques

Comme mentionné précédemment, un point clé en conversion de voix est de déterminer le lien entre les espaces acoustiques des locuteurs source et cible. Bien entendu, pour qu'un tel lien puisse être établi, il faut pouvoir disposer d'un espace de représentation adéquat, c'est-à-dire permettant de regrouper des événements acoustiques perçus comme similaires et donc également capable d'offrir une discrimination claire entre sons perçus différemment.

L'une des motivations de ce travail était qu'en imposant une continuité temporelle dès l'étape d'analyse, nous favoriserons de fait le regroupement de trames adjacentes au sein de classes acoustiques similaires, tout en garantissant une évolution temporelle suffisamment lisse du degré d'appartenance à ces classes acoustiques. Ainsi, indépendamment de la méthode utilisée pour la conversion de voix (quantification vectorielle, mélange de gaussiennes, réseaux de neurones, …) il apparaît crucial de mesurer les capacités de classification et de discrimination des modèles sous-jacents.

Pour cela des métriques adaptées doivent être mises en œuvre, car c'est au final ces métriques qui seront utilisées dans les étapes d'apprentissage voire de conversion elle-même. Dans le cas d'une modélisation cepstrale, la distance classiquement employée est la distance euclidienne entre vecteurs de coefficients cepstraux [6]. Notre représentation est de taille variable, ce qui proscrit l'utilisation de mesures euclidiennes pour comparer deux trames acoustiques. C'est pourquoi nous proposons une mesure adaptée à notre espace de représentation de l'enveloppe spectrale.

Le calcul de la distance entre deux trames se fait en deux étapes. La première étape vise à aligner les pics spectraux des deux trames de façon à disposer d'un espace de représentation commun. Cet alignement est réalisé via l'algorithme mentionné en section 2.2. Dans un second temps, la distance euclidienne entre les vecteurs composés des log-amplitudes des pics alignés est calculée.

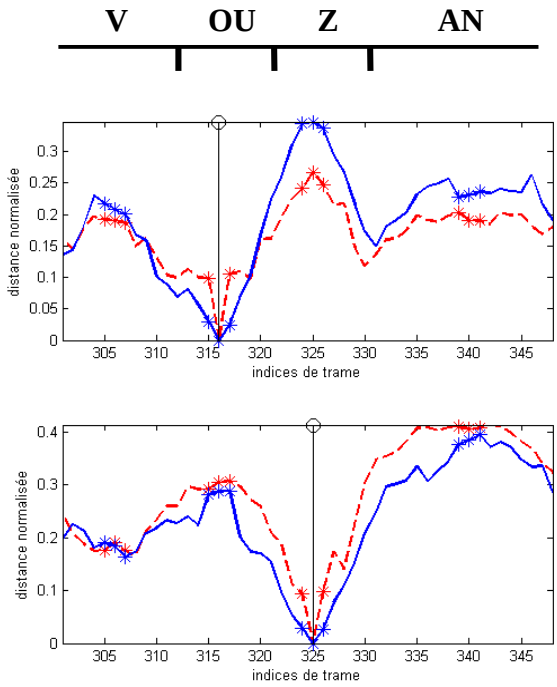


Figure 3: Distances normalisées pour la séquence phonétique 'V-OU-Z-AN'. Les trames de référence sont indiquées par une barre verticale et les trames stables sont marquées par une astérisque. La métrique proposée (en bleu) est comparée à la distance euclidienne entre coefficients cepstraux (courbe rouge en tirets).

La Figure 3 montre un exemple d'utilisation de cette métrique pour comparer une trame de référence à l'ensemble des trames correspondant à la séquence phonétique 'V-OU-Z-AN'. La distance proposée est comparée à une distance cepstrale classique. Notons que ces deux distances sont normalisées par l'énergie de la trame de référence. Les courbes obtenues font apparaître un comportement similaire dans les deux cas. Des travaux futurs consisteront à comparer de manière plus approfondie ces deux types de métriques, et ceci notamment dans un contexte d'apprentissage pour la conversion de voix.

4. Conclusions et Perspectives

Dans cet article, nous avons proposé une nouvelle méthode d'estimation de l'enveloppe spectrale de signaux de parole basée sur une modélisation et un suivi de l'évolution temporelle de pics spectraux. Nous avons illustré quelques propriétés intéressantes de cette méthode, notamment le fait qu'elle fournisse une enveloppe spectrale évoluant de manière lisse et régulière à l'échelle d'un phone. Nous avons également introduit une métrique permettant la comparaison de deux trames sur la base de la représentation proposée et comparé notre analyse à une approche cepstrale plus traditionnelle.

Les avantages de notre méthode d'estimation de l'enveloppe spectrale devraient apparaître de manière plus évidente dans un contexte de conversion de voix. En particulier, nos prochains travaux viseront à explorer les capacités de la méthode proposée à modéliser les espaces acoustiques des locuteurs source et cible.

Références

[1] Turk, O., and Arslan, L., "Robust processing techniques for voice conversion," Computer Speech and Language 20, 2006, 441-467.

[2] *Springer Handbook of Speech Processing*, Editors Benesty, J., Sondhi M. & Huang Y., Springer, 2008.

[3] Nguyen, B. and Akagi, M., "Spectral Modification for Voice gender Conversion Using Temporal Decomposition," Journal of Signal Processing, Vol. 11, No. 4, pp. 333-336, July 2007.

[4] Zolfaghari, P., Watanabe, S., Nakamura, A. and Katagiri, S. "Bayesian Modelling of the Speech Spectrum Using Mixture of Gaussians," in Proc of ICASSP '04, pp. 553-556.

[5] Nguyen, B., "Studies on Spectral Modification in Voice Transformation," Ph.D. diss, Japan Advanced Institute of Science and Technology, March 2009.

[6] Stylianou, Y. "Harmonic Plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification," Ph.D. diss., ENST, Paris, France, Jan. 1996.

Analyse et Modification de la Qualité Vocale basée sur l'Excitation

Thomas Drugman, Baris Bozkurt, Thierry Dutoit

TCTS Lab - Faculté Polytechnique - Université de Mons
31, Boulevard Dolez - 7000 Mons - Belgique

ABSTRACT

This paper investigates the differences occuring in the excitation for different voice qualities. Its goal is two-fold. First a large corpus containing three voice qualities (modal, soft and loud) uttered by the same speaker is analyzed and significant differences in characteristics extracted from the excitation are observed. Secondly rules of modification derived from the analysis are used to build a voice quality transformation system applied as a post-process to HMM-based speech synthesis. The system is shown to effectively achieve the transformations while maintaining the delivered quality.
Keywords : Speech Analysis, Speech Synthesis, Voice Quality, Glottal Source, Voice Modification

1. INTRODUCTION

Depuis les débuts de la recherche en synthèse de parole, l'analyse et la modification de qualité vocale (ou timbre vocal perçu) ont attiré un intérêt scientifique particulier [9]. L'analyse de qualité vocale trouve des applications dans divers domaines du Traitement de la Parole, tels que la synthèse paramétrique de parole de haute qualité, la synthèse de parole expressive/émotionnelle, l'identification du locuteur, la reconnaissance d'émotions, l'analyse de prosodie, etc...En raison de la disponibilité de revues tels que [4] et de la limitation d'espace, une présentation des méthodes d'analyse de la qualité vocale ne sera pas détaillée dans cet article.

Pour analyser la qualité vocale sur un corpus de parole, il est d'usage d'estimer des paramètres spectraux directement à partir du signal de parole, comme par exemple les amplitudes relatives d'harmoniques, ou encore le rapport signal-sur-bruit (HNR). Bien que les variations de qualité vocale soient principalement considérées comme étant contrôlées par la source glottique, l'estimation de cette dernière est bien souvent considérée comme étant problématique et donc évitée dans les procédures d'estimation de paramètres sur d'importants corpus. Dans ce travail, nous adoptons une approche essentiellement orthogonale en étudiant les différences présentes dans la source glottique estimée via un algorithme automatique quand un locuteur donné produit différentes qualités vocales. En se basant sur une analyse paramétrique de la contribution glottique (Section 2), nous examinons l'utilisation de l'information extraite d'un important corpus pour modifier, dans un synthétiseur de parole basé sur des HMMs (Section 3), la qualité vocale d'autres bases de données.

2. ANALYSE DE LA QUALITÉ VOCALE BASÉE SUR L'EXCITATION

Le but de cette partie est de mettre en exergue les différences présentes dans l'excitation quand un locuteur donné produit différentes qualités vocales. La base de données De7 utilisée dans cette étude a initialement été développée par Marc Schroeder dans un but de synthèse de parole expressive par diphones [13]. Cette base de données contient 3 qualités vocales (modale, douce et tendue) prononcées par une locutrice allemande, avec environ 50 minutes de parole par qualité vocale. Dans la Section 2.1, les méthodes d'estimation du flux glottique et de paramétrisation de celui-ci sont présentées. L'harmonicité de la parole est étudiée via la fréquence maximale de voisement en Section 2.2. En tant que caractéristique perceptuelle importante, le tilt spectral est analysé en Section 2.3. La Section 2.4 compare les *résidus propres* [8] estimés pour différentes qualités vocales. Finalement la Section 2.5 quantifie la séparabilité entre les 3 qualités vocales pour les attributs d'excitation ainsi extraits.

2.1. Source Glottique

Nous avons récemment montré que le cepstre complexe permet d'estimer efficacement le flux glottique [6]. Cette méthode a pour but de séparer les composantes à minimum et à maximum de phase du signal de parole. En effet il a été montré précédemment [5] que la parole est un signal à phase mixte où la contribution à maximum de phase (c-à-d anticausale) correspond à la phase ouverte glottique, alors que la composante à minimum de phase est liée à la transmittance du conduit vocal. Isoler la composante à maximum de phase permet donc une estimation fiable de la source glottique, ce qui peut être réalisé par le cepstre complexe. La phase ouverte de la glotte est ensuite paramétrisée par 3 attributs : la fréquence du formant glottique (F_g), le Quotient d'Amplitude Normalisé (NAQ, [2]) et le Quotient de Quasi-Ouverture (QOQ, [1]).

La fréquence du formant glottique est extraite par la méthode décrite dans [3]. La Figure 1(a) montre les histogrammes de F_g/F_0 pour les 3 qualités vocales. Des différences significatives entre les distributions sont observées. Entre autres, il s'avère qu'une voix plus tendue (douce) est traduite par une fréquence de formant glottique plus élevée (basse). La présence de deux modes pour les voix modale et tendue peut être également remarquée sur cette figure. Ceci peut être expliqué par le fait que la source glottique estimée peut comprendre une *ondulation résiduelle* dans les domaines temporel et fréquentiel. Cette

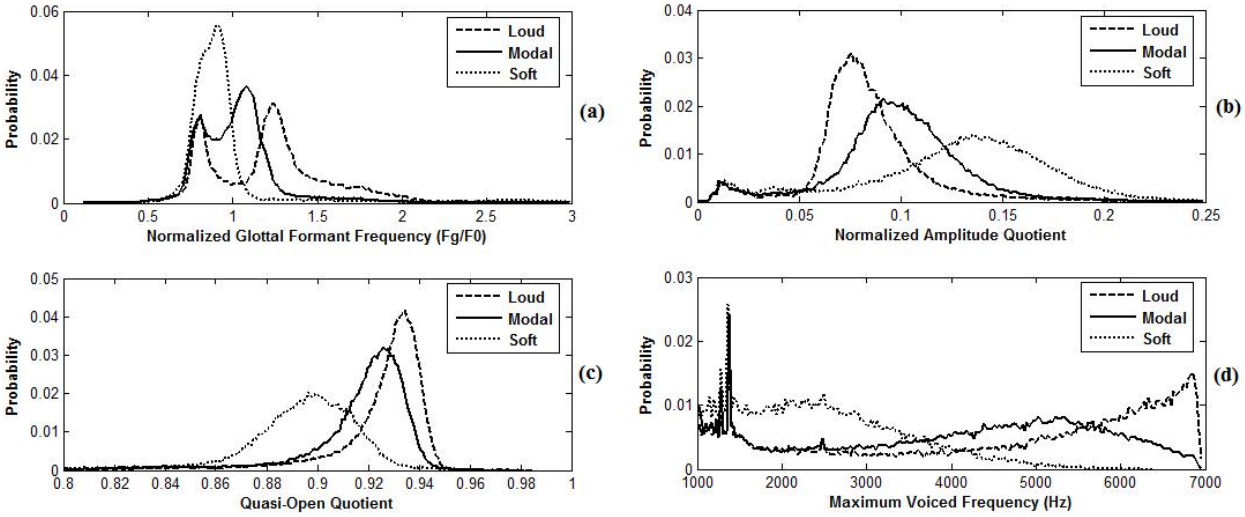


FIGURE 1: Histogrammes, pour les 3 qualités vocales, de *(a) :* la fréquence du formant glottique normalisée F_g/F_0, *(b) :* le Quotient d'Amplitude Normalisé NAQ, *(c) :* the Quotient de Quasi-Ouverture QOQ, et *(d) :* la fréquence maximale de voisement F_m.

dernière peut avoir 2 causes possibles : une séparation incomplète entre F_g et le premier formant F_1 [3], et/ou une interaction non-linéaire entre le conduit vocal et la glotte [12]. Cette ondulation résiduelle peut alors perturber la détection du formant glottique et de cette façon expliquer le pic parasite dans l'histogramme de F_g/F_0 (pic pour les valeurs de F_g/F_0 inférieures à 1).

Dans des travaux précédents [2], [1], Alku et al. ont proposé le Quotient d'Amplitude Normalisé et le Quotient de Quasi-Ouverture comme 2 paramètres temporels utiles pour caractériser respectivement les phases de fermeture et ouverte du flux glottique. Ces paramètres sont ici extraits via la librairie Aparat [11], librement accessible, à partir de la source glottique estimée par le cepstre complexe. Les Figures 1(b) et 1(c) illustrent les histogrammes de ces 2 attributs pour les 3 qualités vocales. Des différences notables entre ces distributions peuvent être observées.

2.2. *Fréquence Maximale de Voisement*

Certaines approches, telles que le Modèle Harmonique plus Bruit (HNM, [14]), considèrent que la parole peut être modélisée par une composante non-périodique au-delà d'une fréquence donnée. Dans le cas du HNM, cette *fréquence maximale de voisement* (F_m) démarque la frontière entre 2 bandes spectrales distinctes, où respectivement des modélisations harmonique et stochastique sont supposées être valides. Plus F_m est élevée, plus l'harmonicité est forte, plus la présence de bruit dans la voix sera faible. Dans ce travail, F_m est estimée par l'algorithme décrit dans [14]. La Figure 1(d) montre les histogrammes de F_m pour les 3 qualités vocales. Il peut être noté qu'en général la voix douce a une fréquence maximale de voisement basse (résultant de sa nature *soufflante*), et que plus l'effort vocal est marqué, plus la parole est harmonique, et par conséquence plus F_m est grande.

2.3. *Tilt Spectral*

Le tilt spectral de la parole est connu pour jouer un rôle important dans la perception de la qualité vocale [16]. Pour capturer cet attribut essentiel, un spectre moyenné est calculé sur le corpus entier (pour une qualité de voix donnée) par un processus indépendant de la prosodie et des variations du conduit vocal. Pour cela, les trames de parole voisée sont extraites par un fenêtrage de Hanning long de 2 périodes de pitch et centré sur un Instant de Fermeture Glottique (GCI). Les positions des GCIs sont déterminées par la technique décrite dans [7]. Ces trames sont ensuite ré-échantillonnées sur un nombre fixé de points et normalisées en énergie. Le spectre moyen est finalement obtenu en moyennant les spectres des trames ainsi normalisées. Ce spectre moyen d'amplitude contient donc un mélange des contributions moyennes de la glotte et du conduit vocal. Le spectre moyenné est illustré en Figure 2 pour les 3 qualités vocales. Puisque ces spectres moyens ont été calculés pour le même locuteur et que l'ensemble d'enregistrements utilisé est phonétiquement équilibré (les formants du conduit vocal auront donc tendance à se contrebalancer), il est raisonnable de penser que les principales différences entre eux sont liées au tilt spectral de la qualité vocale considérée. Entre autres il peut être remarqué que plus l'effort vocal est important, plus le contenu spectral dans la bande [1kHz-5kHz] est riche.

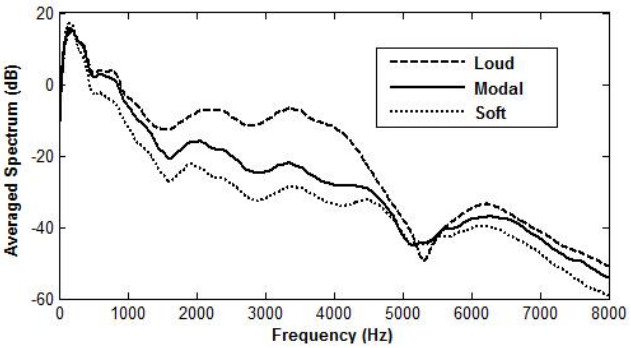


FIGURE 2: Spectre moyenné pour les 3 qualités vocales.

2.4. *Résidus propres*

Nous avons proposé dans [8] de modéliser le signal résidu (obtenu par filtrage LPC inverse) en décomposant des

trames de résidu pitch-synchrones sur une base orthonormée. Il a été également montré que le premier vecteur propre ainsi obtenu (appelé *résidu propre*) permet d'améliorer le naturel en synthèse paramétrique de parole. Comme les résidus propres seront utilisés dans l'application de modification de qualité vocale (Section 3), la Figure 3 illustre la forme d'onde de ce signal selon la qualité vocale produite. On peut noter que les conclusions tirées en Section 2.1 à propos de la phase ouverte glottique sont corroborées. On peut en effet observer que plus l'effort vocal est important, plus la réponse de la phase ouverte du résidu propre est rapide.

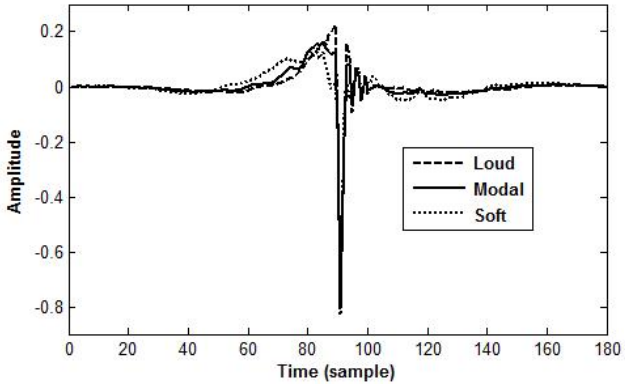


FIGURE 3: Premier résidu propre pour les 3 qualités vocales.

2.5. Séparabilité entre Distributions

Des différences importantes entre les distributions d'attributs ont été présentées dans les sections précédentes. Celles-ci sont d'ailleurs en accord avec les conclusions exposées dans d'autres études [9], [2], [16]. Dans cette section, nous quantifions combien ces différences entre les qualités vocales sont importantes. Pour cela, la divergence de Kullback-Leibler (KL) est connue pour mesure la séparabilité entre 2 fonctions de densité discrètes A et B [10]. Mais puisque cette mesure est non-symétrique (et n'est donc pas une vraie distance), sa version symétrisée, appelée divergence de Jensen-Shannon [10], est généralement préférée. Elle consiste en la somme de 2 mesures KL :

$$D_{JS}(A,B) = \frac{1}{2}(\sum_i A(i)\log_2\frac{A(i)}{M(i)} + \sum_i B(i)\log_2\frac{B(i)}{M(i)}) \quad (1)$$

où M est la moyenne des 2 distributions ($M = 0.5 * (A + B)$). Le Tableau 1 montre les résultats pour les 4 attributs présentés précédemment. Entre autres, on peut remarquer que les voix tendues et douces sont fortement séparables, alors que la qualité tendue est plus proche de la voix modale que la qualité douce ne l'est. On voit également que F_g et NAQ sont hautement informatives pour l'annotation automatique de qualités vocales.

3. MODIFICATION DE QUALITÉ VOCALE

Nous avons proposé dans un travail précédent [8] le Modèle Déterministe plus Stochastique (DSM) du signal résidu. Selon cette approche, l'excitation est divisée en 2 bandes spectrales distinctes délimitées par la fréquence maximale de voisement F_m. La partie déterministe concerne le contenu basses-fréquences et est modélisée par le résidu propre tel que décrit en Section 2.4. Quant à la composante stochastique, il s'agit d'un bruit hautes-fréquences filtré similairement à ce qui est fait dans le modèle HNM [14]. Le signal résidu est ensuite passé dans un filtre de type LPC pour obtenir la parole synthétique.

	F_g	NAQ	QOQ	F_m
$D_{JS}(T,M)$	0.196	0.118	0.035	0.076
$D_{JS}(T,D)$	0.353	0.371	0.279	0.297
$D_{JS}(M,D)$	0.175	0.194	0.175	0.215

TABLE 1: Divergence de Jensen-Shannon entre les 3 qualités vocales (M = Modal, T = Tendu, D= Doux) pour les 4 caractéristiques de l'excitation extraites.

Cette section a pour but d'appliquer les modifications de qualité vocale comme un post-traitement à la synthèse de parole basé sur des HMMs [15], tout en utilisant la modélisation DSM du signal résidu. Plus précisément, un synthétiseur basé sur des HMMs est entraîné sur un corpus de voix modale pour un locuteur donné. Le but est ensuite de transformer la voix de synthèse afin qu'elle soit perçue comme étant douce ou tendue, tout en évitant une dégradation de qualité globale dans la parole produite.

Puisqu'aucun enregistrement de voix expressive n'est disponible pour le locuteur considéré, les modifications sont extrapolées à partir des prototypes décrits pour le locuteur De7 en Section 2, en faisant l'hypothèse que d'autres locuteurs modifient leur qualité vocale de la même façon. Trois principales transformations sont ici considérées :

- Les résidus propres présentés en Section 2.4 sont utilisés pour la partie déterministe du modèle DSM. Ces formes d'onde véhiculent implicitement les modifications de la phase ouverte glottique qui ont été soulignées en Section 2.1.
- La fréquence maximale de voisement F_m est, pour une qualité de voix donnée, fixée selon la Section 2.2 en prenant sa valeur moyenne : 4600 Hz pour la voix tendue, 3990 Hz pour la modale (ce qui confirme les 4 kHz que nous avons utilisé dans [8]), et 2460 Hz pour la qualité douce.
- Le tilt spectral est modifié en utilisant l'inverse du processus décrit en Section 2.3. Pour cela, le spectre des segments voisés est transformé, dans le domaine pitch-normalisé, par un filtre exprimé comme le ratio entre les modélisations auto-régressives des spectres moyennés des qualités vocales source et cible (voir Figure 2). A la synthèse, les trames de résidu sont ensuite ré-échantillonnées à la fréquence fondamentale cible. Cette dernière transformation est donc pitch-dépendante.

Pour évaluer la technique, 10 personnes ont participé à un test subjectif. Le test consistait en 27 phrases générées par notre système pour 3 locuteurs (2 masculins et 1 féminin). Un tiers de ces phrases a été converti vers une voix plus douce, et un autre tiers vers une voix plus tendue. Pour chaque phrase, il a été demandé aux participants d'évaluer l'effort vocal perçu (0 = très doux, 100 = très tendu), et de donner un score MOS selon leur appréciation de la qualité générale. Les résultats sont détaillés dans le Tableau

2 avec leur intervalles de confiance à 95%. De manière intéressante, il peut être noté que les modifications de qualité vocale sont perçues comme souhaité, alors que la qualité globale n'est sensiblement pas altérée (bien que les sujets aient une tendance naturelle à préférer les voix douces).

	Effort vocal	Scores MOS
Modal vers Doux	36.11 ± 2.60	3.189 ± 0.145
Modal	52.89 ± 2.82	3.017 ± 0.147
Modal vers Tendu	72.11 ± 2.60	2.606 ± 0.146

TABLE 2: Evaluation de l'effort vocal perçu (0 = voix très douce, 100 = voix très tendue) et scores MOS pour les 3 versions, ainsi que leurs intervalles de confiance à 95%.

4. CONCLUSION

Nous avons montré dans cette étude qu'un algorithme d'estimation du flux glottique [6] peut être utilisé de manière efficace pour l'analyse de la qualité vocale sur un important corpus de parole, où la plupart de la littérature sur l'estimation du flux glottique se base sur des tests avec des voyelle soutenues. Nous avons étudié les variations de paramètres d'excitation pour différentes qualités vocales et conclu que les 2 attributs F_g et NAQ caractérisant le flux glottique sont hautement informatifs pour l'annotation de qualité vocale. De plus, nous avons montré que l'information extraite d'une base de données peut être appliquée à d'autres corpus de parole dans un but de modification de qualité vocale, tout en maintenant dans une application de synthèse paramétrique de parole une qualité globale relativement élevée.

5. REMERCIEMENTS

Thomas Drugman est supporté par le Fonds National de la Recherche Scientifique (FNRS). Les auteurs aimeraient également remercier M. Schroeder pour la base de données De7, ainsi que Y. Stylianou pour nous avoir fourni l'algorithme d'extraction de la fréquence maximale de voisement.

RÉFÉRENCES

[1] M. Airas and P. Alku. Comparison of multiple voice source parameters in different phonation types. In *Proc. Interspeech*, 2007.

[2] P. Alku, T. Backstrom, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. In *JASA*, volume 112, pages 701–710, 2002.

[3] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. A method for glottal formant frequency estimation. In *Proc. Interspeech*, 2004.

[4] C. D'Alessandro. Voice source parameters and prosodic analysis. In *Method in empirical prosody research*, pages 63–87, 2006.

[5] B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *Proc. ISCA ITRW VOQUAL03*, pages 15–19, 2003.

[6] T. Drugman, B. Bozkurt, and T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proc. Interspeech*, 2009.

[7] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[8] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. Interspeech*, 2009.

[9] D. Klatt and L. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. In *JASA*, volume 87, pages 820–857, 1990.

[10] J. Lin. Divergence measures based on the Shannon entropy. In *IEEE Trans. on Information Theory*, volume 37, pages 145–151, 1991.

[11] [Online]. TKK Aparat main page. In *http ://aparat.sourceforge.net/*.

[12] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. In *IEEE Trans. on Speech and Audio Processing*, volume 7, pages 569–586, 1999.

[13] M. Schroeder and M. Grice. Expressing vocal effort in concatenative synthesis. In *Proc. 15th International Conference of Phonetic Sciences*, pages 2589–2592, 2003.

[14] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. In *IEEE Trans. Speech and Audio Processing*, volume 9, pages 21–29, 2001.

[15] K. Tokuda, H. Zen, and A. Black. An HMM-based speech synthesis system applied to english. In *Proc. IEEE Workshop on Speech Synthesis*, pages 227–230, 2002.

[16] O. Turk, M. Schroeder, B. Bozkurt, and L. Arslan. Voice quality interpolation for emotional text-to-speech synthesis. In *Proc. Interspeech*, 2005.

Indices acoustiques de phonémicité et d'allophonie dans la parole adressée aux enfants

Alejandrina Cristià[1], *Amanda Seidl*[2], *and Kristine H. Onishi*[3]

[1] Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS - DEC - ENS - CNRS
29, rue d'Ulm - 75005 Paris, France
[2] Speech, Language, and Hearing Sciences, Purdue University
500 Oval Drive - 47907 West Lafayette, IN, USA
[3] Psychology, McGill University
1205 Dr Penfield Avenue - H3A 1B1 Montreal, Qc, Canada
alecristia@gmail.com ; aseidl@purdue.edu ; kris.onishi@mcgill.ca

ABSTRACT

Since infants' loss of sensitivity to allophonic contrasts occurs by 11 months of age, it cannot be lexical knowledge that allows them to determine a contrast's allophonic or phonemic status. Instead, it may be that the acoustic instantiation of phonemic contrasts is better than that of allophonic ones. Moreover, this difference may be exaggerated in speech to infants. Acoustic analyses of the correlates of nasality and tenseness in a corpus of American English show that the phonemic contrast (tenseness) is more marked than the allophonic one (nasality). However, this difference is not further enhanced in speech to young children.
Keywords : speech perception acquisition, infant-directed speech, allophones.

1. Introduction

Plusieurs études ont montré que les enfants perdent la sensibilité aux sons qui ne sont pas représentés dans leur langue maternelle, et qu'ils deviennent au contraire plus sensibles à ceux qui sont représentés dans cette langue (e.g., [12]). Le cas des allophones est particulièrement intéressant, car il s'agit de sons présents dans la langue maternelle, mais qui n'y jouent pas un rôle contrastif. Pour prendre un exemple, en anglais les voyelles sont nasalisées avant les consonnes nasales appartenant à la même syllabe ; ainsi, la voyelle /aɪ/ (comme dans le mot *I* « je ») est nasale dans *I'm* « je suis », mais orale dans *I'd* « j'avais ». Contrairement à la nasalité, le trait tendu/relâché (Advanced Tongue Root, ATR) est contrastif en anglais, car il existe des paires minimales comme /bit-bɪt/ (*beet* « betterave » ; *bit* « un peu »). Donc, même si il y des voyelles nasales et orales, tendues et relâchées en anglais, elles ont un statut différent et leur traitement est nettement différent chez les adultes. Dans notre étude, nous examinons une hypothèse se proposant d'expliquer comment les enfants peuvent apprendre à être plus sensibles aux distinctions phonémiques et moins sensibles aux distinctions allophoniques.

1.1. Apprendre des phonèmes et des allophones

Du point de vue linguistique, on détermine le statut allophonique ou phonémique d'une paire de sons avec l'existence de paires minimales, comme *beet* et *bit*. Pourtant, des travaux récents montrent que les enfants perdent la sensibilité aux sons qui ont un rapport allophonique, par rapport aux contrastes phonémiques, dès l'âge de 11 mois [11]. À cet âge, les enfants ne disposent pas dans leur lexique des paires minimales suffisantes pour déterminer la phonémicité de ces contrastes [2]. Par conséquent, il semble probable que d'autres informations présentes dans le signal acoustique soient utilisées pour faire cette distinction.

1.2. Indices acoustiques de phonémicité et d'allophonie

Une possibilité serait d'utiliser les informations distributionnelles. Dans l'exemple ci-dessus, les voyelles nasales se trouvent seulement avant des consonnes nasales, et les voyelles orales apparaissent dans toutes les autres positions. Des études computationelles ont démontré que cette stratégie est robuste, mais seulement si elle est limitée par la ressemblance des sons impliqués [10]. C'est pour cette raison que nous explorons ici la manière dont la ressemblance acoustique pourrait faciliter l'apprentissage de la distinction phonémique/allophonique. Plus précisément, les rapports phonémiques et allophoniques pourraient être représentés différemment dans la parole adressée aux enfants, de sorte que les sons impliqués dans un rapport allophonique deviendraient plus semblables entre eux que les sons impliqués dans un contraste phonémique. Par exemple, de la même manière que les anglophones ne prononcent pas le contraste entre voyelles nasales et orales de manière nette [3], les cibles nasales et orales pourraient être acoustiquement plus proches que les cibles tendues et relâchées. Et cela, d'autant plus qu'il s'agirait de parole adressée aux enfants, comme nous le détaillons dans la prochaine section.

1.3. Parole adressée aux enfants (PAE)

C'est un fait bien établi que les gens parlent de façon très différente quand ils s'adressent aux enfants (la PAE) que lorsqu'ils s'adressent à d'autres adultes (la parole adressée aux adultes, PAA) [4]. La plupart des études ont établi l'existence de différences dans des dimensions paralinguistiques ; par exemple, les mères utilisent une fréquence fondamentale plus haute dans la PAE que dans la PAA [5]. Une deuxième série d'études essaye de montrer que les mères améliorent également les catégories phonétiques dans la PAE [6], et quelques chercheurs soutiennent que cette stratégie a des effets importants sur l'acquisition du langage

(voir, par exemple, Native Language Magnet de [7]). Néanmoins, ces recherches ne sont pas concluantes, car d'autres études suggèrent qu'il n'y a pas d'amélioration dans tous les cas, ou même, que quelques catégories peuvent être détériorées (par exemple, [1]). Toujours est-il que d'après des théories comme le Native Language Magnet, on pourrait prédire que les contrastes phonémiques s'améliorent dans la PAE en augmentant la distance acoustique entre les sons dans la PAE par rapport à la PAA. En revanche, cette amélioration n'aurait pas lieu pour les sons avec un rapport allophonique.

2. Méthode

2.1. Plan général et prédictions

Pour explorer ces hypothèses, nous avons enregistré des mères anglophones américaines discutant avec leurs enfants et un adulte à propos d'un ensemble de jouets. Les jouets en question avaient été choisis car ils présentaient des voyelles nasales, orales, tendues, et relâchées dans des positions phonologiques contrôlées. Ce corpus était codé et soumis à des analyses acoustiques. Nos hypothèses étaient testées avec des Analyses de Variance (ANOVAs), une pour chaque contraste, avec Registre (PAE, PAA) et Trait [ou ATR (tendu, relâché) ou Nasalité (nasal, oral)] comme facteurs. Deux groupes de variables dépendantes étaient considérés ; pour commencer, les corrélats acoustiques de chaque trait. De plus, afin d'avoir le même nombre de variables dépendantes pour chaque trait, mais également afin de choisir des dimensions vraiment orthogonales, le deuxième groupe de variables dépendantes était constitué des deux premières composantes résultant des analyses en composantes principales.[1]

L'ANOVA donne des valeurs F et p pour chaque facteur ainsi que pour leur interaction. Etant donné nos connaissances préalables sur la façon dont les mères américaines parlent à leurs enfants, nous avons supposé qu'il y aurait un effet principal du Registre sur quelques dimensions acoustiques ; par exemple, les voyelles seraient plus longues dans la PAE (car le débit des mères est plus faible ; Bernstein Ratner [15]). En dehors de ces effets, notre intérêt principal portait sur l'effet du facteur Trait, indicatif de la distance acoustique entre les représentants du contraste concerné ; mais aussi l'interaction, pouvant suggérer que l'implémentation acoustique dépendrait du Registre. Concernant ces effets, nous avons formulé **deux hypothèses sur l'implémentation acoustique des contrastes phonémiques et allophoniques. Premièrement, les contrastes phonémiques seront mieux implémentés que les sons avec un rapport allophonique.** Par conséquent, ils auront des valeurs p plus faibles pour le facteur Trait dans les ANOVAs pour l'ATR que dans celles pour la nasalité. **Deuxièmement, les mères vont exagérer la distance acoustique entre catégories dans la PAE, mais seulement pour le contraste phonémique.** Il en résultera que l'interaction Registre x Trait sera significative dans les ANOVAs pour l'ATR, mais pas pour la nasalité.

[1] Par ailleurs, l'analyse en composantes principales établit automatiquement les composantes (dimensions décorrelées) qui optimisent la répresentation de la variance pour les facteurs considérés dans chaque analyse. Ces composantes orthogonales sont des *combinaisons* des facteurs originaux, donc non réductibles au facteur le plus influant. Par consequent, les deux corrélats le plus significatifs dans les premières ANOVAs (avec les corrélats acoustiques comme variables dépendantes) peuvent ne pas être identiques aux facteurs principaux des composantes principales.

2.2. Corpus : équipement et procédure

Dix mères américaines, dont les enfants avaient environ 4.5 mois (3.95-4.9 ; M = 4.42) ont été enregistrées avec un microphone AKG WMS40 Pro Presenter Set Flexx UHF Diversity CK55 Lavalier, sur un Marantz Professional Solid State Recorder PMD660ENG. Pour nous assurer que les mères ne soient pas excessivement concentrées sur leur prononciation, nous leur avons demandé une tâche particulière qui consistait à expliquer comment ranger les objets en catégories. Ceux-ci avaient été spécifiquement choisis car ils présentaient les deux groupes de voyelles dans des positions phonologiques similaires, et dans des mots de fréquence proche (par ex., « pesto/basil » ; « dancer/tassle »). Après avoir expliqué la tâche à effectuer, l'expérimentatrice a laissé mères et enfants seuls pendant 35-45 minutes puis est retournée discuter avec la mère pour obtenir les mêmes mots que dans la séquence PAE.

2.3. Analyses acoustiques

Les voyelles ont été codées en utilisant Praat (www.praat.org). Les fichiers-sons codés ont ensuite été soumis à une analyse sur le logiciel Praat qui mesure plusieurs dimensions dans le milieu de chaque voyelle (disponible sur http ://sites.google.com/site/acrsta/). Ces dimensions correspondent aux corrélats acoustiques associés au trait tendu (F1, F2, durée ; [8]) ainsi qu'à la nasalité (fréquence et bande passante F1, A1P0 et A1P1, qui représentent respectivement la différence d'amplitude entre F1 et le premier et deuxième formants nasaux ; ces corrélats ont été choisis parce que Chen [4] a déjà montré leur valeur pour le trait nasalité dans une comparaison entre l'anglais et le français québécois, cf. la discussion).

3. Résultats

Le nombre de voyelles marquées pour chaque catégorie était assez équilibré, avec un total de 918 données complètes pour l'analyse de la nasalité (430 nasales ; 488 orales) et 1030 pour le trait ATR (594 tendues ; 436 relâchées). Les résultats des ANOVAs pour les dimensions acoustiques sont résumés dans les Tables 1-2 et ceux des composantes principales dans la Table 3. On remarque ainsi que ces analyses ont confirmé notre prédiction des effets du Registre sur certaines caractéristiques acoustiques, comme, par exemple, le fait que les voyelles seraient plus longues dans la PAE.

3.1. Hypothèse 1 : Phonémique mieux qu'allophonique

Les résultats de ces analyses sur les dimensions acoustiques et aussi ceux sur les deux premières composantes (cf. section 2.1) confirment notre première hy-

pothèse : le trait ATR semble mieux marqué que la nasalité du point de vue des différences statistiques, car les valeurs p sont plus basses.

Tab. 1: Valeurs F et p (dans l'exposant) pour les corrélats acoustiques du trait ATR. L'exposant indique l'ordre de grandeur négatif (base 10) pour p arrondi à l'unité la plus proche, tel que $p = 0$ .001 est répresenté par ^{4} ; $p = 0.01$ par ^{3} ; etc. F1 = frequence du F1 ; F2 = frequence du F2.

	F1	F2	Durée
Registre (R)	2.49	3.53^2	14.98^5
Trait (T)	411.84^{16}	72.95^{16}	51.11^{12}
R x T	3.43^2	8.63	0.13

Tab. 2: Valeurs F et p (dans l'exposant) pour les corrélats acoustiques du trait Nasalité. F1 b. = bande passante.

	F1	F1 b.	A1P0	A1P1
Registre (R)	1.49	5.38^3	0.00	0.01
Trait (T)	336.79^{16}	4.88^3	19.35^5	48.35^{12}
R x T	1.90	0.12	0.89	0.41

Tab. 3: Valeurs F et p (dans l'exposant) pour les deux premières composantes principales. Facteur = facteur avec le plus de poids.

	ATR		Nasalité	
	CP1	CP2	CP1	CP2
Facteur	F2	F1	F1 b.	F1
Registre (R)	2.24	0.07	5.39^2	2.62
Trait (T)	77.67^{16}	347.23^{16}	5.08^2	337.24^{16}
R x T	8.48^3	5.84^2	0.13	1.73

3.2. Hypothèse 2 : Phonémique, mais non allophonique, amélioré dans PAE

Comme nous en avions fait l'hypothèse, on trouve également quelques interactions Registre x Trait pour l'ATR, mais pas pour la nasalité. Ce résultat semblerait soutenir notre deuxième hypothèse : il y aurait une implémentation acoustique différente dépendant du registre, mais seulement dans le cas phonémique. Néanmoins, une inspection plus profonde des données montre que ces interactions n'indiquent pas forcément une amélioration. Dans les analyses sur les dimensions acoustiques, seul F2 montre une interaction, ce qui est contraire à notre prédiction (i.e., la différence entre les tendances centrales pour les voyelles tendues et relâchées sur F2 est de 182 Hz pour la PAE et 383 Hz pour la PAA ; et la valeur p pour le facteur Trait a un ordre de grandeur 10 fois plus faible dans la PAA que dans la PAE). Naturellement, la CP1 montre une détérioration puisque F2 a une influence importante sur cette composante. Au contraire, on trouve une amélioration spécifique du registre PAE sur CP2. Une représentation graphique de ces interactions est donnée dans la Figure 1, montrant la tendance centrale sur CP1 et CP2 pour les voyelles tendues et relâchées. De plus, ces tendances centrales sont projetées sur chaque axe pour faciliter l'inspection de l'interaction. Dans ces projections, on peut voir que les voyelles tendues et relâchées sont plus différentes (donc, le contraste est amélioré) dans la PAE sur CP2, mais elles sont plus proches (le contraste est détérioré) sur CP1.

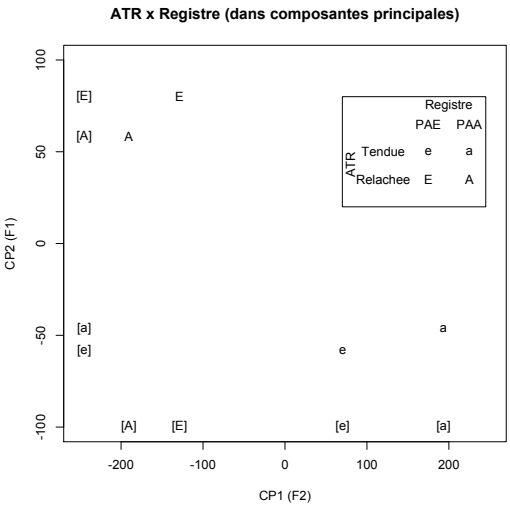


Fig. 1: Tendances centrales pour les voyelles tendues et relâchées, dans la parole adressée aux enfants (PAE) et dans la parole adressée aux adultes (PAA). Les projections sur chaque axe sont montrées entre crochets.

4. Discussion

Etant donné que les enfants traitent les contrastes phonémiques et allophoniques de manière différente dès l'âge de onze mois, il est probable qu'il y ait des informations directement codées dans la parole qui permettent aux enfants de décider quels sont, entre tous les contrastes présents dans la parole, ceux qui sont phonémiques et ceux qui sont allophoniques. Cette expérience a exploré deux hypothèses. La première était que la dissimilitude acoustique est plus grande dans les contrastes phonémiques que dans les sons qui ont un rapport allophonique. Les résultats apportent des preuves positives sur ce point, puisque les voyelles tendues et relâchées étaient plus différentes entre elles que les voyelles nasales et orales.

Cependant, il y a une explication alternative : le trait ATR est toujours mieux marqué acoustiquement que le trait Nasalité. Pour être sûr que la différence dans le codage acoustique correspond au statut phonémique/allophonique d'un contraste, il faudrait étudier une deuxième langue où la nasalité est phonémique et l'ATR allophonique. Le français québécois correspond à cette situation : les voyelles sont relâchées dans les syllabes fermées, sinon elles sont tendues ; et les voyelles nasales et orales sont contrastives.

Nous sommes actuellement en train d'enregistrer des mères francophones à Montréal pour tester cette explication alternative. Les résultats préliminaires de 3 mères vont dans la même direction que les résultats des mères américaines : le facteur Trait a des valeurs p plus basses dans les ANOVAs pour le contraste phonémique dans le français québécois (la nasalité) que celles pour les sons avec un rapport allophonique dans cette langue (l'ATR) (par ex., ; la valeur pour CP1 et pour CP2 est respectivement plus faible de 2 et 9 ordres de grandeurs).

Ces résultats suggèrent que les différences dans l'implémentation acoustique pourraient attirer l'attention des enfants sur les contrastes phonémiques et les éloigner des sons avec un rapport allophonique. Cette information pourrait être utile de deux manières. Tout d'abord, encoder le statut phonémique acoustiquement serait très utile étant donné l'énorme sensibilité des enfants pour les distributions des corrélats acoustiques [9]. De plus, la méthode des distributions complémentaires a quelques limitations, car elle ne peut pas expliquer les cas de variation libre, où deux ou plusieurs allophones peuvent occuper la même position phonologique (comme les deux prononciations du mot anglais « atom », l'une avec une occlusive, l'autre avec un flap). D'autres études doivent explorer la méthode présentée ici dans des cas de variation libre, car les différences dans l'implémentation acoustique peuvent être les seuls indices de phonémicité ou d'allophonie que les enfants trouvent dans la parole.

Il y a, pourtant, une importante limitation dans cette étude, car la comparaison des valeurs p restait une mesure relative et imprécise. Des études futures devraient incorporer des méthodes plus complexes mais plus précises (comme par exemple celles utilisées pour comparer la distance génétique entre espèces).

Notre deuxième hypothèse concernait la façon dont les contrastes phonémiques et allophoniques étaient implémentés dans la parole adressée aux enfants (la PAE) par rapport à celle adressée aux adultes (la PAA). Vues les dernières théories sur l'implémentation acoustique des contrastes linguistiques, nous avions prédit que les contrastes seraient améliorés dans la PAE. De plus, nous avions supposé que ceci ne concernerait que le trait phonémique, l'ATR. Les données relatives à la première partie de cette hypothèse n'étaient pas concluantes, vu que les mères ont amélioré le contraste sur certaines dimensions mais pas sur les autres. Au contraire, la deuxième partie de cette hypothèse semblait être soutenue par les données, car il n'y avait pas de telles interactions Registre x Trait pour le trait allophonique. Clairement, ce sujet devrait être exploré plus profondément, car les résultats présents ne soutiennent pas les théories actuelles sur la manière dont la parole adressée aux enfants facilite l'acquisition phonologique.

En résumé, ces travaux avaient pour but de tester l'hypothèse que la parole contient des informations pouvant aider les enfants à apprendre si un contraste est phonémique ou allophonique. Les résultats ont montré que les contrastes phonémiques peuvent être plus clairement marqués acoustiquement que ceux qui sont allophoniques. Néanmoins, ils ne sont pas forcément exagérés dans la parole adressée aux enfants. D'une manière générale, il y a sans doute quelques informations qui pourraient contribuer à modeler la perception des enfants, et aider à expliquer comment les enfants montrent des performances détériorées avec des sons qui ont un rapport allophonique avant la fin de la première année de leur vie.

Références

[1] J. A. Baran, M. Zlatin Laufer, and R. Daniloff. Phonological contrastivity in conversation : A comparative study of Voice Onset Time. *Journal of Phonetics*, 5 :339–350, 1977.

[2] M. C. Caselli, E. Bates, P. Casadio, J. Fenson, L. Fenson, L. Sanderl, and J. Weri. A cross-linguistic study of early lexical development. *Cognitive Development*, 10 :159–199, 1995.

[3] M. Chen. Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, 102, 1997.

[4] C. A. Ferguson. Baby talk in six languages. *American Anthropologist*, 66 :103–114, 1964.

[5] A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. Boysson-Bardies, and I. Fukui. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16 :977–1001, 1989.

[6] P. K. Kuhl, J. E. Andruski, I. A. Chistovich, E. V. Kozhevnikova, V. L. Ryskina, E. I. Stolyarova, U. Sundberg, and F. Lacerda. Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277 :684–686, 1997.

[7] P. K. Kuhl, B. T. Conboy, S. Coffey-Corina, D. Padden, M. Rivera-Gaxiola, and T. Nelson. Phonetic learning as a pathway to language : new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B*, 363 :979–1000, 2008.

[8] P. Ladefoged and I. Maddieson. *Sounds of the World's Languages*. Blackwell Publishers, Cambridge, MA, 1996.

[9] J. Maye, J. F. Werker, and L. A. Gerken. Infant sensitivity to distributional information can effect phonetic discrimination. *Cognition*, 82 :B101–B111, 2002.

[10] S. Peperkamp, R. LeCalvez, J. P. Nadal, and Emanuel Dupoux. The acquisition of allophonic rules : Statistical learning with linguistic constraints. *Cognition*, 101 :B31–B41, 2006.

[11] A. Seidl, A. Cristià, A. Bernard, and K. H. Onishi. Allophones and phonemes in infants' learning of sound patterns. *Language Learning and Development*, 5 :191–202, 2009.

[12] J. F. Werker and R. Tees. Cross-language speech perception : Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7 :49–63, 1984.

Prosodie et discrimination d'expressions émotionnelles actées *vs.* spontanées

Nicolas Audibert[1,2]*, Véronique Aubergé*[2,3]*, Albert Rilliard*[4]

[1] Université d'Avignon, CERI/LIA, Avignon, France
[2] GIPSA Lab, UMR 5216 CNRS/Univ. Grenoble [3] LIG UMR 5217 CNRS/Univ. Grenoble, France
[4] LIMSI-CNRS, Orsay, France
nicolas.audibert@univ-avignon.fr, veronique.auberge@imag.fr, albert.rilliard@limsi.fr

ABSTRACT

This paper presents the first results of an acoustic analysis of 12 pairs of monosyllabic acted *vs.* spontaneous expressions of satisfaction, irritation and anxiety produced by 4 speakers, discriminated and rated for emotional intensity differences in previous perceptual experiments. Acoustic features were extracted, compared within each pair and correlated with perceptual ratings. Results show significant correlations between general F0 level difference in the pair and perceived emotional intensity difference, but cannot account for the observed variability in discrimination scores. The influence of F0 contours shape of selected stimuli is discussed.

Keywords: expressive speech, acted emotion, spontaneous emotion, acoustic features, F0 contours

1. INTRODUCTION

La plupart des études menées sur la parole expressive se sont appuyées sur des expressions simulées par des acteurs amateurs ou professionnels (voir Scherer [1] pour une discussion des différents paradigmes), en raison de la possibilité offerte par cette méthode de contrôler le contenu lexical des énoncés produits et les conditions d'enregistrement. La représentativité des corpus de parole expressive actés pour la modélisation d'expressions réalistes d'affect a toutefois été remise en question (voir notamment Campbell [2]), ce qui a conduit à un effort accru de recueil et d'analyse de corpus expressifs spontanés. En dépit de cette prise de conscience des limites des corpus actés, très peu d'études les ont comparés directement à la parole expressive spontanée. Parmi les rares travaux qui se sont attachées à une telle comparaison, Aubergé *et al.* ont montré que des expressions audiovisuelles spontanées d'amusement induites par des plaisanteries visuelles inattendues pouvaient être perceptivement discriminées, avec une variabilité très importante entre les performances des différents juges [3].

Pour les besoins d'une étude de la typicalité d'expressions vocales d'émotions menée par Laukka *et al.* [4], 47 énoncés acté et 146 énoncés spontanés produits par 6 acteurs ont été extraits du corpus de parole expressive EWiz/Sound Teacher [5], validés et notés en termes d'intensité émotionnelle perçue. Cette évaluation a montré une intensité perçue supérieure pour les énoncés actés *vs.* spontanés. Dans une précédente étude [5], nous avons extrait de ce corpus 24 paires d'expressions audiovisuelles actées *vs.* spontanées de satisfaction, d'irritation et d'anxiété produites par 4 acteurs semi-professionnels, appariées par locuteurs et par émotion. Une expérience de discrimination perceptive a montré que des sujets naïfs sont globalement capables de discriminer les expressions actées *vs.* spontanées, avec un important effet inter-juge qui confirme les résultats obtenus sur l'amusement dans [3]. Un protocole similaire a permis d'obtenir des jugements de différence d'intensité émotionnelle, confirmant la perception d'une intensité supérieure pour les expressions actées. La confrontation entre ces mesures a montré qu'en dépit qu'une corrélation significative entre ces mesures, les différences d'intensité émotionnelle perçue ne peuvent expliquer entièrement les performances de discrimination.

Nous étudions dans cet article les propriétés acoustiques des énoncés qui composent les 12 paires monosyllabiques actées *vs.* spontanées, et examinons les différences intra-paires des paramètres extraits à la lumière des résultats perceptifs obtenus en condition de présentation audio. Les formes et valeurs d'ancrage des contours de fréquence fondamentale sont ensuite discutées.

2. PAROLE EXPRESSIVE ACTÉE *VS.* SPONTANÉE

Le corpus expressif E-Wiz/Sound Teacher [5] a été enregistré avec une technique de Magicien d'Oz, dans laquelle le sujet croit interagir avec une interface personne-machine complexe alors que le comportement de l'application est en réalité contrôlé à distance par l'expérimentateur. Il s'agit ici d'une application présentée comme un logiciel novateur d'aide à l'apprentissage des langues étrangères basé sur un système de reconnaissance vocale, pour lequel de derniers tests seraient nécessaires avant sa commercialisation. L'interaction avec le système est contrainte par un langage de commande composé des mots monosyllabiques français [bʁik], [ʒon], [ʁuʒ], [sabl] et [vɛʁ] ainsi que de la commande [paʒsʮivãt], réparties au cours du déroulement du scénario afin de pouvoir collecter des énoncés identiques exprimant diverses valeurs affectives. Les performances attribuées aux 17 sujets ayant participé à l'expérience ont été manipulées afin d'induire chez eux des émotions positives puis

négatives, les productions recueillies étant étiquetées dans un premier temps par les sujets eux-mêmes à partir de l'enregistrement vidéo, afin de tirer parti de leur mémoire autobiographique et permettre ainsi un étiquetage plus fin, avant une étape de validation perceptive par un panel de juges naïfs. Un protocole spécifique a été mis en place pour les 7 sujets qui étaient également acteurs, à qui il a été demandé immédiatement après l'enregistrement de reproduire les affects ressentis lors de l'expérience sur les mêmes énoncés, les expérimentateurs mettant l'accent sur le fait que ces affects devaient être exprimés de la façon la plus similaire possible au ressenti dans l'expérience. Les acteurs recrutés pour cette tâche pratiquent le théâtre de rue et/ou d'improvisation, et ont déclaré que le dispositif expérimental leur fournissait des conditions optimales pour développer leur jeu d'acteur.

3. DISCRIMINATION ET DIFFERENCE D'INTENSITE EMOTIONNELLE

La table 1 récapitule les résultats obtenus en condition de présentation audio dans les expériences de discrimination perceptive et de jugement de différences d'intensité émotionnelle menées précédemment [5] pour les 12 paires d'énoncés monosyllabiques, les résultats obtenus pour les autres conditions et pour les paires d'énoncés de longueur supérieure n'étant pas discutés ici. Les productions actées et spontanées de deux hommes (M1 et M2) et deux femmes (F1 et F2) composent ces paires, construites en sélectionnant autant que possible des stimuli exprimant un niveau d'intensité de l'émotion similaire d'après les jugements perceptifs obtenus dans [4]. Les scores supérieurs à 50 correspondent respectivement à une paire discriminée au-dessus du niveau du hasard par 33 juges naïfs francophones, et jugés comme expriment le même niveau d'intensité émotionnelle par 32 juges naïfs francophones.

Table 1 : Scores perceptifs obtenus dans [5] pour les 12 paires monosyllabiques en condition audio-seule. 1ère valeur : taux de discrimination, 2ème valeur (entre parenthèses) : différence d'intensité émotionnelle. Les scores significativement différents de 50 (t-tests) sont indiqués par * (p<.05) ou ** (p<.01).

Locuteur	Anxiété	Irritation	Satisf.	Global
F1	63** (73**)	63** (64**)	43 (51)	56* (63**)
F2	58* (55*)	67** (60**)	57 (52)	61** (55**)
M1	49 (57*)	44 (46)	60* (63**)	51 (55**)
M2	56 (52)	72** (85**)	71** (72**)	66** (70**)
Global	56** (59**)	62** (63**)	58** (59**)	59** (61**)

4. MESURES ACOUSTIQUES

4.1. Méthodes

Les mesures acoustiques ont été extraites au moyen de Praat [6], à partir de frontières phonémiques placées manuellement. Les paramètres acoustiques utilisés par Banse et Scherer [7], dont les travaux font référence pour l'étude des expressions vocales d'émotions, ont été retenus. Certaines mesures ont été adaptées aux spécificités de nos données. Ce jeu de paramètres a été complété par des mesures spectrales additionnelles. Etant donné que la distance de la bouche des locuteurs au microphone n'a pu être gardée constante au cours de l'enregistrement du corpus, les mesures d'intensité acoustique, qui pourraient être biaisées, n'ont pas été conservées. Afin de nous assurer que les différences d'intensité acoustique intra-paires ne biaisent pas les jugements perceptifs de discrimination et de différence d'intensité émotionnelle, nous avons néanmoins vérifié que ces différences soient non-significatives au sens d'un t-test apparié, ce qui est bien le cas, tant pour une mesure effectuée sur l'ensemble de l'énoncé (p=.348) que limitée à la voyelle (p=.110).

Fréquence fondamentale et durée

La détection des pulsations glottiques réalisée par la méthode d'autocorrélation de Praat a été corrigée manuellement afin de garantir une mesure précise de F0, selon une méthode adaptée de Xu [8]. Les valeurs extraites en Hertz ont été converties en demi-tons. Les moyennes et écarts-types de F0 ainsi que les 25èmes et 75èmes centiles ont été extraites pour les segments vocaliques et pour l'ensemble de la portion voisée de énoncés. L'attaque, l'étendue (différence entre max et min) et la déclinaison (différence entre les valeurs finales et initiales) ont été extraites comme descripteurs complémentaires des variations de F0. Les mesures cycle-à-cycle de perturbation (jitter et shimmer) ont également été extraites des segments vocaliques.

Cette étude étant centrée sur les énoncés monosyllabiques, les mesures de débit ont été remplacées par une mesure de la durée de l'énoncé, complétée par la proportion de la durée de la voyelle.

Mesures spectrales

Les mesures spectrales utilisées dans [7] ont été extraites du spectre voisé à long terme (LTAS) des portions voisées et non-voisées. Les paramètres suivants ont été extraites du LTAS voisé : l'index HammI défini comme la différence entre le maximum d'énergie dans les bandes 0-2kHz et 2-5kHz ; la pente spectrale au dessus de 1 kHz DO1000 ; la répartition de l'énergie entre hautes et basses fréquences, avec une fréquence de coupure de 500Hz (PE500) et 1kHz (PE1000) ; la proportion d'énergie dans 9 bandes de fréquences distinctes entre 125Hz et 8kHz. La proportion d'énergie dans 9 bandes de fréquences distinctes entre 125Hz et 8kHz a également été extraite du LTAS non-voisé.

Des descripteurs plus généraux de la distribution spectrale ont également été extraits des segments vocaliques : centre de gravité, dissymétrie et kurtosis. Le ratio HNR, qui évalue la proportion de l'énergie acoustique correspondant aux composantes harmoniques du signal et permet ainsi d'estimer la périodicité d'un signal de parole liée à la qualité de voix [9], a également été extrait sur les segment vocaliques et l'ensemble de l'énoncé. La valeur des 3 premiers formants a été extraite semi-automatiquement des points centraux des voyelles.

Comparaisons intra-paires

Les différences intra-paires entre les stimuli actés et spontanés ont été calculées pour chaque paire et chaque paramètre extrait. Ces différences sont exprimées sous forme de proportion relative pour les paramètres calculés selon une échelle linéaire, tandis que les valeurs sont soustraites dans le cas des échelles logarithmiques, en prenant dans les deux cas le stimulus spontané comme référence.

La construction des paires n'a pu être effectuée en conservant systématiquement le même énoncé au sein d'une paire, 7 des 12 paires présentant des énoncés différents. En conséquence, la plupart des paramètres extraits ne peuvent être directement comparés que pour deux paires de [ʒon], une de [bʁik], une de [ʁuʒ] et une de [sabl], dans lesquelles la comparaison intra-paire est effectuée sur la même voyelle et reste donc valide. Toutefois, les variations microprosodiques intrinsèques de F0 en français étant selon Di Cristo d'amplitude très inférieures à celles dues à la prosodie linguistique et expressive [10], les mesures fondées sur la distribution de F0 (et elles seules) ont été comparées pour les 12 paires.

4.2. Résultats

Une série de t-tests appariés indique une différence intra-paire significative pour plusieurs mesures de F0 : l'attaque et la moyenne sur la voyelle ($p<.01$), l'écart-type sur l'énoncé ($p<.01$), l'étendue sur l'énoncé ($p<.01$), les 25èmes et 75èmes centiles sur l'énoncé ($p<.05$) et la voyelle ($p<.01$). En revanche la différence de déclinaison est non-significative indiquant globalement une F0 plus élevée pour les stimuli actés sur l'ensemble de l'énoncé. Les valeurs de F2 sont également significativement plus faibles pour les stimuli actés ($p<.05$).

Les corrélations entre chaque différence intra-paire et les jugements de différence d'intensité émotionnelle ont été calculées. Ces corrélations sont significatives pour la F0 moyenne ($r=.710$; $p=.01$), l'attaque de F0 ($r=.617$, $p<.05$), et pour les 25ème ($r=.723$; $p<.01$) et 75ème centiles ($r=.661$, $p<.05$) de F0 sur la voyelle, ainsi que pour le shimmer ($r=-.927$; $p<.05$). Elles le sont sur l'ensemble de l'énoncé pour les différences de HNR ($r=-.924$; $p<.05$), de proportion d'énergie spectrale entre 600 et 800Hz sur les portions voisées ($r=-.887$; $p<.05$), de centre de gravité spectral ($r=-.963$; $p<.01$).

En revanche, seules la dissymétrie ($r=.984$; $p<.01$) et le kurtosis ($r=.925$; $p<.05$) de la distribution spectrale sont significativement corrélés aux scores de discrimination.

5. COMPARAISON DES CONTOURS DE F0

La figure 1 présente les contours de fréquence fondamentale de 4 paires sélectionnées. Ces paires correspondent aux productions de locuteurs F1 et M1, qui selon les résultats de la discrimination perceptive en condition audio parviennent le mieux à simuler des expressions vocales d'émotions similaires à des expressions spontanées. Pour chacun, une paire discriminée au dessus du niveau du hasard (1 et 3) et une paire pour laquelle les scores de discrimination sont au niveau du hasard (2 et 4) ont été retenues. Les contours présentés dans ces figures sont normalisés pour permettre une comparaison de leurs formes indépendamment des variations de durée.

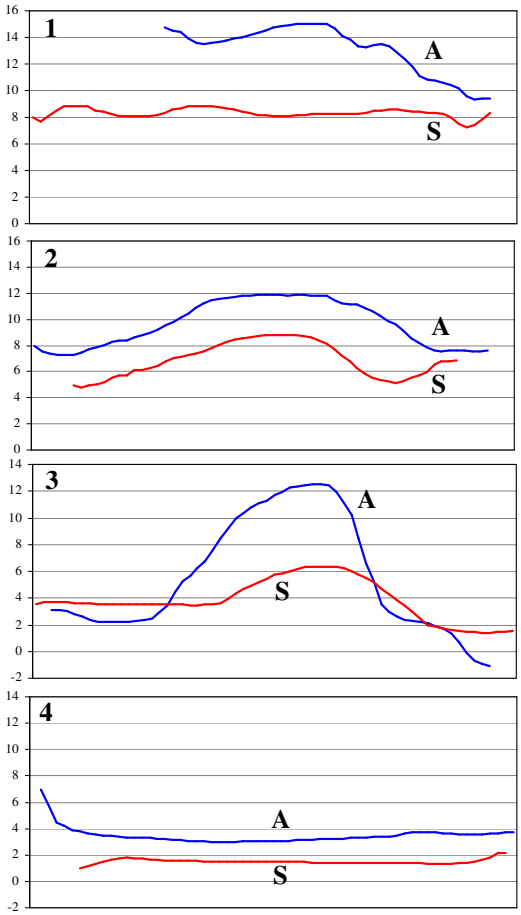


Figure 1 : Contours de F0 pour 4 paires sélectionnées. 1 : anxiété, locutrice F1 ; 2 : satisfaction, F1 ; 3 : satisfaction, M1 ; 4 : irritation, M1. Les expressions actées sont notées A, les expressions spontanées S. Les échelles sont en demi-tons, 0 dt = 100Hz.

Une première observation est que ces contours confirment un niveau général de F0 plus élevé pour les expressions actées. Cette tendance est en cohérence avec

les résultats obtenus par Bänziger et Scherer [11], qui ont lié l'étendue et le niveau général de F0 à la dimension émotionnelle d'activation, qui dans un tel contexte peut être assimilée à l'intensité de l'émotion. Si l'écart est plus important pour les paires bien discriminées, il demeure néanmoins plus élevé d'environ 1 demi-ton pour les stimuli actés des paires qui ne sont pas discriminées à un niveau supérieur au hasard. Les formes de contours observées pour les expressions de satisfaction sont très similaires à celles précédemment observées pour des expressions de joie et satisfaction actées et spontanées produites par le locuteur M2 [12].

6. DISCUSSION

Les comparaisons intra-paires indiquent qu'une même émotion exprimée spontanément *vs.* simulée par le même locuteur diffère principalement par le niveau général de F0. Le second formant est également plus bas pour les stimuli actés, ce qui constitue un résultat inattendu, tout particulièrement pour la satisfaction. En effet, ce résultat semble en contradiction avec les valeurs de F2 plus élevées pour des sourires mécaniques *vs.* spontanés observées par Tartter et Braun [13]. En revanche, tandis que l'exagération fréquemment liée à l'acté aurait pu conduire à étirement plus important pour ces expressions de satisfaction, ce résultat attendu ne se vérifie pas au niveau des différences entre valeurs de F3 mesurées.

Les corrélations suggèrent un lien entre les performances de discrimination et la distribution spectrale dont la caractérisation, à défaut d'être directement interprétable en termes de production ou perception de parole, pourrait contribuer à l'identification des productions actées plus proches de productions spontanées. Elles indiquent également que les stimuli jugés plus intenses sont réalisés avec des perturbations et apériodicités moindres. Ces stimuli sont de plus réalisés avec un niveau général de F0 plus élevé, correspondant d'après Bänziger et Scherer [11] à un niveau d'activation plus élevé.

La comparaison des contours de F0 sur un sous-ensemble sélectionné de paires confirme cette tendance, et indique que F0 demeure plus élevée, quoique dans une proportion moindre, dans le cas d'expressions actées qui ne sont pas discriminées des expressions spontanées à un niveau supérieur au hasard.

Afin d'explorer plus avant le rôle des contours de F0 dans la différence d'intensité émotionnelle perçue, le niveau général de F0 pourra être manipulé en synthèse tout en conservant la forme générale du contour afin d'évaluer de manière systématique son impact sur les jugements d'intensité émotionnelle. Une telle expérience de resynthèse devra être menée précautionneusement, en prenant en compte les effets de la manipulation de F0 sur l'intensité acoustique perçue et le biais induit sur les jugements d'intensité émotionnelle. Dans l'hypothèse où un lien fort entre niveau de F0 et intensité émotionnelle perçue serait confirmé, un protocole similaire pourrait ensuite permettre une évaluation des performances de discrimination d'expressions actées *vs.* spontanées par des juges naïfs en gelant les variations de l'intensité émotionnelle. Etant donné que les contours de F0 des expressions de joie et de satisfaction présentent de façon récurrente une forme « en cloche » dans notre corpus [12], et véhiculent l'essentiel de l'information affective pour ces affects [14], cette future étude pourrait ainsi se focaliser sur des expressions de satisfaction.

BIBLIOGRAPHIE

[1] K. R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication,* 40(1-2):227-256, 2003.

[2] N. Campbell. Databases of Emotional Speech., *1st ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, 34-38, 2000.

[3] V. Aubergé and M. Cathiard. Can we hear the prosody of smile? *Speech Communication*, 40(1-2):87-97, 2003.

[4] P. Laukka, N. Audibert and V. Aubergé. Exploring the graded structure of vocal emotion expressions. In Hancil, S. (ed.), *The Role of Prosody in Affective Speech*, Peter Lang, 241-258, 2009.

[5] N. Audibert, V. Aubergé and A. Rilliard. Discrimination perceptive d'expressions émotionnelles actées *vs.* spontanées. *TSI, numéro spécial ACA*, sous presse.

[6] P. Boersma and D. Weenink. Praat: doing phonetics by computer (Version 5.1.20). Téléchargé depuis http://www.praat.org/, 2009.

[7] R. Banse and K. R. Scherer. Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology*, 70(3):614-636, 1996.

[8] Y. Xu. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, 27:55-105, 1999.

[9] L. Eskenazi, D. G. Childers, and D. M Hicks. Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research*, *33*, 298–306, 1990.

[10] A. Di Cristo. *De la microprosodie à l'intonosyntaxe*. Université de Provence, 1985.

[11] T. Bänziger and K. R. Scherer. *The role of intonation in emotional expressions.* Speech Communication, 46:252-267, 2005.

[12] V. Aubergé, N. Audibert and A. Rilliard. Acoustic morphology of expressive speech: What about contours? *2nd Speech Prosody*, 91-95, 2004.

[13] V. C. Tartter and D. Braun Hearing smiles and frowns in normal and whisper registers. *JASA*, 96 (4):2101-2107, 1994.

[14] N. Audibert, D. Vincent, V. Aubergé and O. Rosec. Expressive Speech Synthesis: Evaluation of a Voice Quality Centered Coder on the Different Acoustic Dimensions. In *Proc. 3rd Speech Prosody*, 2006.

Une petite histoire de l'analyse harmonique de la parole

Bernard Teston

Laboratoire Parole et Langage, UMR 6057, Aix-Marseille Université

ABSTRACT

Since the beginning of the harmonic analysis theory by Joseph Fourier in 1822, until his universal practices with the technics of digital signal processing in the seventies, many machines and methods have been proposed to make acoustical analysis of the speech signals. Mechanical, optical, electronical and acoustical devices have been realised successfully with a remarquable creative imagination and considerable efforts from speech scientist to make progress the knowledge of the secrets of the physical nature of the speech

I. INTRODUCTION

Dans le courant du 18ème siècle, les grammairiens étudient l'évolution historique des langues isolément sur des textes et les lettres de leurs alphabets. Plus rares, sont ceux qui s'intéressent aux sons qu'elles représentent c'est-à-dire, à la parole et à ce qui est dit plutôt qu'écrit. Ce sont pour l'essentiel des éducateurs de sourds-muets, partisans de l'oralité et des dialectologues phonéticiens. Cela oblige à appliquer des méthodes nouvelles empruntées à la physiologie (étude des articulations) et à la physique (acoustique des sons). Les savants de l'époque des lumières ne peuvent cependant que constater que les sons, sont des phénomènes physiques si fugaces que la connaissance de leurs structures et de leurs mécanismes créateurs ne pourra se développer qu'à la condition d'inventer des techniques nouvelles pour les capturer les analyser et les restituer. La fin du 18ème siècle connaît le développement d'outils mathématiques, dont certains vont permettre de modéliser des phénomènes acoustiques, ainsi qu'un engouement particulier pour les automates parlants. Il va se maintenir durant tout le 19ème siècle et ces synthétiseurs de voix et de parole, vont tester les modèles de production des segments phonétiques. De nombreux savants vont travailler alors sur la représentation graphique de la parole, pour la rendre visible, l'enregistrer, afin de pouvoir l'étudier dans le détail, l'analyser et même la reproduire.

Analyser un son ou plus généralement un signal, c'est étudier sa structure, c'est-à-dire l'amplitude et l'organisation de ses différents constituants. On peut voir l'acte fondateur de l'analyse harmonique dans les travaux de Daniel Bernouilli qui en 1753, a l'idée d'exprimer la solution du problème des cordes vibrantes au moyen de séries trigonométriques. Compte-tenu de la complexité structurelle et temporelle de la parole qui en fait un objet si particulier, son analyse harmonique représente un des aspects les plus fascinants de l'instrumentation de mesure avec notamment la notion de spectre évolutif. C'est cette histoire que nous nous proposons d'exposer.

II. L'ANALYSE HARMONIQUE

Au tout début du 19ème siècle, Joseph Fourier étudie la propagation de la chaleur dans les corps solides. Il la modélise au moyen d'une équation aux dérivés partielles parabolique, fonction unique mais difficile à manier mathématiquement. Il a alors l'idée originale pour l'époque, d'exprimer certaines fonctions périodiques sous la forme de séries trigonométriques infinies. Dans un but de simplification, il remplace son équation aux dérivées partielles par une série de fonctions trigonométriques, sinus et cosinus beaucoup plus maniable et dont la somme reconstitue la fonction initiale.

La proposition mathématique des séries de Fourier tient en deux formules :

- D'une part, celle de l'analyse (1), qui consiste à déterminer la suite des coefficients de Fourier *Cn* de la fonction périodique *f(x)*.

$$c_n = \int f(x)\, c_n e^{-i\,n\,x} \frac{dx}{2\pi} \quad (1)$$

- D'autre part, celle de la synthèse (2) qui permet de retrouver la fonction périodique f(x) à partir de la suite de ses coefficients de Fourier Cn.

$$f(x) = \sum c_n e^{i\,n\,x} \quad (2)$$

Analyse et synthèse sont deux aspects complémentaires de la théorie des séries de Fourier qui établit une correspondance entre la fonction périodique et les coefficients de Fourier. Publiée en 1822, cette nouvelle façon de décrire les fonctions périodiques permet de leur appliquer des opérations bien plus simples, particulièrement la dérivation. C'est à partir de la théorie des séries de Fourier que s'est développé le domaine mathématique de l'analyse harmonique qui est devenue l'outil fondamental pour l'étude des fonctions périodiques.

Certes, avant Fourier, les séries trigonométriques étaient connues, et Daniel Bernoulli avait eu l'idée de les utiliser dans la solution du problème des cordes vibrantes mais uniquement pour l'analyse de son cas particulier. Fourier a été le premier à généraliser les séries trigonométriques comme outil s'appliquant à toutes les fonctions périodiques mais surtout à considérer l'analyse et la synthèse (1) et (2) comme un tout indissociable. Les mathématiciens contemporains ont souvent reproché à Fourier son manque de rigueur, mais son but a été de fournir des outils mathématiques utiles à l'explication des phénomènes naturels et pour cela, on peut le considérer comme le créateur des mathématiques appliquées.

En 1843, Georg Simon Ohm émet l'hypothèse que tout son est une fonction périodique du temps décrite « à la Fourier » par une somme de sons purs constituée par une fréquence fondamentale (qui correspond à la hauteur du son) et de ses multiples (les harmoniques).

Au milieu du 19ème siècle, l'analyse harmonique permet donc aux acousticiens de calculer des spectres de son. Comme généralement on n'en connaît pas la fonction mathématique, il faut en avoir des images fidèles et cela ne sera possible qu'à la suite de nombreuses évolutions des phonautographes [1]. L'analyse harmonique va cependant se révéler avec les travaux de Herman von Helmholtz, dans la controverse sur l'origine du timbre des voyelles qui va perdurer pendant plus d'un siècle. En 1830, Robert Willis énonce que le timbre est défini par une hauteur et un signal laryngien particuliers à chaque voyelles. C'est la théorie du « fixed pitch ». Charles Wheatstone l'approuve en 1837 en la complétant par une influence partielle de la cavité buccale sur le timbre des voyelles. C'est la théorie du « relative pitch ».

Franciscus Donders propose en 1858 une théorie totalement opposée aux précédentes. Pour lui, le timbre des voyelles est uniquement défini par la configuration anatomique du conduit vocal supra glottique, hypothèse qu'il formule à la suite d'expériences minutieuses sur la voix chuchotée.

A la même époque, Helmholtz, qui poursuit des recherches sur les mécanismes de l'audition, développe une théorie de la résonance acoustique des cavités. Il propose un modèle de résonateur sphérique dont on peut calculer les caractéristiques acoustiques, fréquence de résonnance et sélectivité (coefficient de surtension), en fonction des ses dimensions (diamètre et orifice) ; les résonateurs de Helmholtz. Ce dernier, adepte de la théorie d'Ohm et de l'analyse harmonique de Fourier, va confirmer en 1852 les hypothèses de Donders au moyen de sa théorie des résonateurs qu'il applique au conduit vocal. Il réalise l'analyse de Fourier en hurlant des voyelles près du sommier d'un piano et en appréciant tactilement les vibrations des cordes qui résonnent à la fondamentale et à certains de ses harmoniques avec des amplitudes variables. Il réalise également la synthèse des voyelles au moyen d'un banc de ses résonateurs, en les excitant avec un signal périodique piloté par un diapason. Il constate que les voyelles peuvent être différenciées à hauteur constante par une distribution particulière de trois harmoniques dans le spectre de fréquence. C'est la première mention du concept de formant. Il précise ainsi l'hypothèse de Donders sous la forme de la « théorie du timbre ». Ses résonateurs ainsi que son synthétiseur de voyelles sont construits et diffusés par Rudolf Koenig, un de ses anciens élèves facteur d'instruments d'acoustique à Paris. Dès lors, de nombreux travaux vont se succéder pour tenter de réaliser l'analyse harmonique des sons, sans connaître la forme de leur signal.

III. LES ANALYSEURS DE KÖNIG

En 1862, Koenig, inspiré par les flammes chantantes, a l'idée de mesurer l'amplitude de la pression acoustique dans des résonateurs au moyen de la flamme d'un bec alimenté en gaz de ville par l'intermédiaire d'une capsule manométrique dont la membrane au contact de l'air vibrant, module la pression du gaz.

Il réalise une batterie de huit résonateurs de Helmholtz équipés de capsules manométriques dont les mouvements des flammes sont ralentis grâce à des miroirs tournants, pour pouvoir être observés, car leurs images sont fugaces. Sans développer les détails des caractéristiques acoustiques d'un tel dispositif, il nous faut reconnaître en lui l'origine de tous les analyseurs « temps – fréquence » dont la structure va perdurer jusqu'au traitement numérique du signal. Les résonateurs sont des filtres passe bande, les capsules manométriques sont des détecteurs d'énergie et les miroirs tournants, une base de temps. Très vite, Koenig étend son analyseur à quatorze résonateurs cylindriques, dont les fréquences peuvent être ajustées La bande d'analyse de 100 à 1200 Hz est assez réduite. La sensibilité d'un tel appareil est médiocre, et il est nécessaire, pour la parole, de crier très fort et très près des orifices des résonateurs, pour observer quelque chose. Tel quel, il a été diffusé dans de nombreux laboratoires universitaires (où on peut encore en rencontrer quelques uns) et a été utilisé comme outil pédagogique jusqu'au milieu du 20ème siècle.

Koenig a imaginé ses capsules manométriques couplées à un résonateur (tuyau d'orgue ou de Helmholtz), dans lesquels la pression acoustique est relativement forte. En champs libre, elles fonctionnent très mal. En 1882, il parle dans une embouchure connectée à une capsule manométrique et constate que la forme des flammes est plus complexe que celles obtenues avec des résonateurs et qu'elles semblent donner une représentation en relation avec la structure acoustique des voyelles et qui permet de les différencier. Ainsi, dans la description de chaque période on peut distinguer de une à quatre flammes de différentes largeurs et amplitudes, dont le positionnement dans l'image du cycle périodique varie en fonction du timbre du son. René Marage, un élève d'Etienne-Jules Marey, réussit en 1895 à fixer l'image des flammes sur un film grâce à plus de lumière et à la caméra de son maitre Il peut étudier ainsi les phénomènes de distorsions acoustiques dans les tubes et les cornets utilisés comme aides auditives avec des résultats cohérents avec les connaissances actuelles. Il a également utilisé les flammes pour une étude du timbre des voyelles ainsi que Nichols et Merritt en 1898. Mais ces flammes ne correspondent qu'approximativement à la structure harmonique du spectre. Elles sont de surcroit difficiles à interpréter (un de ces derniers auteurs a même déclaré qu'il avait plus de difficultés à discerner des indices dans les flammes que les haruspices de l'antiquité dans les entrailles de poulets). De fait, les flammes manométriques sont peu précises et se prêtent mal au calcul. Elles ont été peu utilisées et ravalées bien vite au niveau des curiosités par l'analyse de Fourier à partir de la représentation graphique du signal de parole.

IV.L'ANALYSE HARMONIQUE GRAPHIQUE

Depuis Léon Scott de Martinville, qui obtient en 1858 les premières traces d'un signal de parole que l'on pouvait voir et conserver avec son phonautographe, de nombreuses tentatives se sont succédées pour en améliorer les performances [1]. Ce n'est que dans le dernier quart du 19ème siècle que la finesse et la précision des traces sont jugées suffisantes pour que l'on puisse calculer les coefficients de Fourier correspondants aux composantes harmoniques sur une période de signal. C'est Schneebeli qui en 1878 réalise les premiers calculs de spectres de Fourier sur des signaux de voyelles tracés par son phonautographe. Ces calculs sont laborieux et très longs. Compte-tenu de l'intérêt de la méthode, de nombreux efforts sont déployés pour simplifier les calculs tout en conservant une bonne précision. Ludimar Herman propose en 1898 une méthode graphique basée sur l'utilisation de feuilles millimétrées pré-imprimées et d'une table de calcul, qui permettent de diviser par dix (approximativement) le temps de calcul des 20 premiers harmoniques. Cette méthode est la favorite des phonéticiens des débuts du 20ème siècle, Rousselot et

Scripture en particulier, pour analyser les signaux de parole fournis par des phonautographes optiques ou des kymographes [1].

Devant la grande quantité de calculs longs et fastidieux, nécessaires à l'obtention d'un spectre de Fourier « à la main », de nombreux savants proposent des solutions pour les réaliser de manière semi-automatique au moyen de dispositifs mécaniques. William Thomson propose en 1878, le premier analyseur harmonique mécanique pour calculer les mouvements des marées, et les variations météorologiques de température et de pression. Il fonctionne très bien mais est très complexe, volumineux et très cher à construire. En 1894, Olaus Henrici imagine une méthode mécanique beaucoup plus simple qui est immédiatement exploitée par la firme suisse Coradi. Léger, portable et d'un prix accessible, cet instrument permet dans sa version la plus précise le calcul du 100ème harmonique. Il obtient très vite une diffusion universelle dans les communautés scientifiques et techniques qui ne cessera qu'en 1960. Il est suivi plus tard par le planimètre spécialisé de Mader-Ott encore plus simple et moins onéreux mais moins performant en rapidité de calcul.

Malgré l'aide mécanique, l'analyse de Fourier reste fastidieuse à appliquer. D'abord il est nécessaire d'agrandir la trace des signaux de parole par la photographie, si l'on veut obtenir une bonne précision. Ensuite, les harmoniques sont calculés période par période et les masses de calculs, bien que très diminuées demeurent considérables, au point qu'elles doivent être sous traitées (souvent auprès de communautés monastiques). L'analyse harmonique mécanique va se perpétuer jusque vers 1950 pour les signaux lents ou transitoires. Pour les signaux de parole, elle est supplantée rapidement dans les années 1930 par l'analyse électronique qui, associée au microphone, magnétophone et oscilloscope, s'éloigne de Fourier pour quelques décennies.

V. LES ANALYSEURS ELECTRONIQUES

Depuis l'invention du téléphone par Graham Bell en 1878 on sait comment transformer une variation de pression acoustique en un signal électrique. On sait également que les résonateurs de Helmholtz peuvent être modélisés par des filtres passe-bande au moyen de circuits électriques RLC (capacités, inductances, résistance). Mais pour réaliser un modèle électrique de l'analyseur acoustique de Koenig, il manque une fonction essentielle; l'amplification, qui n'apparait que dans les années 1920 avec l'essor technologique de la radioélectricité. Plusieurs familles d'analyseurs électroniques vont alors se succéder jusqu'à l'apparition des techniques de traitement numérique du signal sur ordinateur dans les années 1970.

5.1. Les analyseurs à bancs de filtres

Ils sont dans la filiation directe de l'analyseur acoustique de Koenig dont ils simulent la même structure : des filtres passe-bande en parallèle et contigus suivis de détecteur d'énergie suivis d'un oscilloscope ou d'un enregistreur graphique pour visualiser les spectres. Les filtres, de largeur de bande relative constante (ΔF/F constant) se sont stabilisés rapidement à une largeur de bande 1/3 d'octave. Avec 22 filtres la bande d'analyse est comprise entre 36 Hz et 18 kHz. Les analyseurs les plus simples sont les analyseurs séquentiels dont les filtres sont commutés cycliquement en synchronisation avec l'enregistreur graphique. Le plus représentatif en est le Bruel et Kjaer 2101 associé à l'enregistreur graphique 2305. Avec ce type d'appareil, le signal analysé (segment d'un signal variable ou transitoire) doit être répété en boucle sur un magnétophone. En associant un détecteur d'énergie à chaque filtre, et en commutant rapidement les canaux sur un scope synchronisé, on dispose d'un analyseur en temps réel bien plus pratique. Le premier de ces instruments est proposé par Freistedt en 1935 mais ne sera commercialisé que plus tard par Siemens. Il est suivi dans les années 1950 par toute une gamme d'appareils proposés par plusieurs constructeurs dont ; General-Radio, Hewlett-Packard et Bruel et Kjaer. L'ultime évolution de cette famille, le modèle 3347 de ce dernier constructeur, a été exploitée jusqu'à la fin des années 1970. Ce type d'analyseur a surtout été utilisé par les phonéticiens pour des études sur la perception.

5.2. Les analyseurs à filtre continu

Contrairement aux précédents, ces analyseurs utilisent un filtre unique à bande étroite en ΔF/F constant, dont on fait varier de manière continue la fréquence centrale sur la largeur de la bande d'analyse. Ils ont été développés pour répondre à la nécessité de disposer d'une plus grande résolution spectrale que celle du 1/3 d'octave pour discriminer dans le spectre des harmoniques très proches. La variation de fréquence est effectuée soit au moyen d'un filtre RC (amplificateur sélectif) dont on fait varier les résistances, soit selon le principe de l'hétérodyne en faisant varier la fréquence d'un oscillateur. Les analyseurs du premier type, dont on fait varier la fréquence du filtre en synchronisation avec un enregistreur graphique, ont été proposés par plusieurs constructeurs dont Bruel et Kjaer. Le signal doit être répété en boucle au moyen d'un magnétophone et l'analyse est beaucoup plus lente qu'avec un analyseur 1/3 d'octave séquentiel, en revanche, elle est beaucoup plus précise. Le premier analyseur hétérodyne pour l'analyse de la parole est proposé par Grutzmacher. Dans ce type d'analyseur, la fréquence du filtre est fixe et sa largeur de bande peut être choisie en fonction de la définition spectrale désirée. L'analyse est en ΔF constant et l'exploration spectrale est obtenue par battement avec une fréquence variable dans la gamme d'analyse, générée par un oscillateur local. Le balayage de l'oscillateur est synchronisé avec la base de temps d'un oscilloscope ou d'un enregistreur graphique pour visualiser le spectre.

5.3. Les analyseurs à compression de temps

Ils ont été imaginés pour accélérer les temps d'analyse des analyseurs à filtre continu et se rapprocher de l'analyse en temps réel. Leur principe consiste à enregistrer un signal de courte durée, et de le reproduire en boucle beaucoup plus rapidement. On comprime ainsi le temps d'analyse et on multiplie la gamme de fréquence par le rapport de compression de temps. On applique ensuite une analyse hétérodyne. Le premier dispositif fonctionnant sur ce principe est le célèbre Sona-graph d'un rapport de compression temporelle de 10. Les ultimes évolutions des analyseurs à compression de temps sont les Spectral Dynamics et Saicor à la fin des années 1960. D'un rapport de compression de 1000, ils furent les seuls à analyser un signal en temps réel mais disparurent très vite, remplacés par les premiers transformateurs de Fourier numériques.

5.4. Le Sona-graph ou sonagraphe

Cet analyseur, véritable légende de l'analyse acoustique de la parole a été développé dans les laboratoires de la Bell dans le but d'étudier les distorsions des signaux de parole pour améliorer les communications radiotéléphoniques. Il a été imaginé dés le début des années 1930 par Ralph Potter avec les trois principes suivants : la compression de temps, l'analyse hétérodyne et des résultats d'analyse en présentation temps fréquence sous la forme de spectres évolutifs. Son développement est très long et en grande partie, occulté par la seconde guerre mondiale. Le premier prototype de faisabilité fonctionne en 1940 mais n'est présenté par Potter qu'en 1945. Toute une équipe autour de Walter Koenig améliore et stabilise l'appareil original après de multiples mises au point de la boucle magnétique et du système de gravure des spectres qui sont les deux parties les plus originales de l'appareil.

La boucle est constituée par une piste magnétique située sur la tranche d'un disque qui tourne à la vitesse d'un tour toutes les 2,2 secondes (courant alternatif de 50 Hz) durant l'enregistrement du signal. Pendant l'analyse, le disque tourne 10 fois plus vite. La durée du signal est donc comprimée d'un rapport 10 et la bande d'analyse dilatée d'autant, soit de 200 Hz à 200 kHz, qui sont les limites haute et basse de variation de l'oscillateur hétérodyne. Le spectre évolutif s'inscrit sur un cylindre vertical solidaire du disque d'enregistrement assurant une parfaite correspondance temporelle entre le signal et le spectre. La hauteur du cylindre correspond à la bande d'analyse, dont le balayage s'effectue de bas (20 Hz) en haut (20kHz), avec un pas de 50 Hz par tour, en 160 tours. Si à chaque instant de la boucle qui le reproduit, le signal contient de l'énergie au 1/10 de la fréquence du filtre d'analyse, une pointe électrique brûle plus ou moins un papier spécial disposé sur le cylindre, en fonction de l'intensité du signal dans le filtre. A chaque tour du cylindre, la pointe s'élève, et son mouvement fait varier la fréquence de l'oscillateur local. Le spectre évolutif du signal enregistré sur la boucle est présenté avec le temps en abscisse et les fréquences en ordonnée. L'intensité des composantes spectrales est donnée par le degré de noirceur de leurs traces.

Avec une largeur de bande d'analyse de 20 Hz à 20 kHz, sur une durée de 2,2 secondes, un filtre d'analyse étroit (40 Hz pour une meilleure résolution fréquentielle et visualiser les harmoniques) ou large (300 Hz pour une meilleure résolution temporelle et mieux visualiser les formants), et surtout un remarquable système de gravure du spectre évolutif, le Sona-graph est parfaitement adapté à l'étude de la dynamique des signaux de parole. En 1951, Bell cède la licence d'exploitation à la société Kay Electric qui va le distribuer sous le nom de Sona-graph pendant près de 40 ans identique, à quelques détails prés, à l'original fonctionnant avec des tubes à vide (Zoom des fréquences, courbe d'intensité, image en courbes de niveaux). Sa version transistorisée, le modèle 6061A, apparaît en 1964. Il devient très vite le standard des phonéticiens et sera le plus diffusé. A sa suite, Kay devenu Elemetric propose en 1978 le type 7800, une version dans laquelle la boucle est une mémoire circulante numérique qui permet d'afficher le signal de parole synchrone avec le spectre présenté sur le même type d'inscripteur à papier brulé. Il permet également d'avoir accès à certaines mesures quantitatives.

Le sonagraphe et ses documents les sonagrammes ont eu une importance considérable dans les études sur la parole pendant une quarantaine d'années et si de nos jours il est en grande partie oublié il nous a laissé le terme générique d'un *instrument permettant l'analyse et la représentation graphique d'un son*. Il nous a laissé également la représentation temps fréquence et jusqu'aux valeurs de la sélectivité des filtres d'analyse étroit et large. Le développement des techniques numériques de traitement des signaux l'a rapidement évincé dans les laboratoires.

VI. L'ANALYSE HARMONIQUE NUMERIQUE

Dés les années 1960, la diffusion des ordinateurs de calculs scientifiques a permis le développement du domaine du traitement numérique du signal et une véritable explosion des applications de l'analyse harmonique de Fourier avec la transformée de Fourier discrète (DFT). Une des principales avancées dans ce domaine est la proposition par Jones Cooley et John Tuckey en 1965 de l'algorithme de la transformée de Fourier rapide (FFT) qui permet une diminution considérable du temps de calcul, au prix il est vrais de quelques restrictions. Parallèlement, les progrès des circuits de calculs numériques ainsi que leur miniaturisation permettent au milieu des années 1970 de développer des instruments de mesure spécialisés, capables d'effectuer des analyses de Fourier en temps réel dont le DSP Sona-graph de Kay Elemetric est l'ultime représentant. Leur apparition suit de quelques années à peine celle des analyseurs à compression de temps rapides, mentionnés précédemment. Ils sont peu utilisés par les phonéticiens car onéreux, et leur existence est éphémère à cause du traitement numérique du signal qui se diffuse rapidement dans les laboratoires sur les mini-ordinateurs, les stations de travail puis les PC [2], sous la forme des éditeurs de signaux, qui regroupent toutes les fonctions essentielles pour l'analyse de la parole [3]. Ce sont des outils que maintenant nous utilisons tous à des degrés divers, sans bien en connaître les origines, et nos aînés qui les ont imaginé et auxquels nous sommes tous redevables.

Les lecteurs qui désireraient approfondir certains aspects de cet exposé pourront consulter avec profits les ouvrages [4], [5] et [6].

BIBLIOGRAPHIE

[1] B. Teston. A la poursuite du signal de parole. *Actes, Journées d'Etude sur la Parole 26ème JEP*. 7-10. 2006.

[2] B. Teston. A la poursuite du signal de parole : Suite et fin. Actes*, Journées d'Etude sur la Parole 27ème JEP*. 397-400. 2006.

[3] P. Martin. *Phonétique acoustique : Introduction à l'analyse acoustique de la parole*, Armand Colin, Paris, p. 163, 2008.

[4] C. Escudier, H. Gazanhes, Tachoire et V. Torra. *Des cordes aux ondelettes.* Publications de l'Université de Provence. Aix. p. 482. 2001.

[5] T. M. Hankins, R.J. Silverman. *Instruments and the imagination*, Princeton University Press, Princeton, p. 337, 1995.

[6] R. T. Beyer, *Sounds of our time*, Spinger, New-York ,p. 444, 1.999.

Les degrés de laryngalisation dans l'espagnol parlé au Yucatan, Mexique: manifestations acoustiques et physiologiques et processus phonétiques

Antonia Colazo-Simon

Laboratoire Phonétique et Phonologie (UMR 7018) CNRS/ Paris 3-Sorbonne Nouvelle
19 rue des Bernardins 75005 Paris
simonantonia@hotmail.com

ABSTRACT

Laryngealised phonation requires the simultaneous activation of the larynx, glottis and vocal cords. Recent studies demonstrate that the production of the glottal consonant can be mapped onto a continuum spanning from a slightly constricted to a strongly constricted glottis. In spite of the modern instrumental techniques, it remains fairly difficult to define with precision the possible degrees of laryngeal activity. In this paper, we attempt to explain the phenomenon of laryngealisation in the Spanish dialect spoken in Yucatan, Mexico, and offer an account of the various degrees of laryngealisation (glottalisation) in this language by means of two instrumental techniques, spectrography and glottography.

Keywords: laryngealisation, larynx, glottis, vocal cords, glottography.

1. INTRODUCTION

Depuis les premiers instruments utilisés par Rousselot dans ses « Principes de phonétique expérimentale » [11], les techniques instrumentales adaptées aux recherches phonétiques n'ont cessé de se perfectionner et d'ouvrir de nouveaux axes de recherche dans l'étude de la parole.

Des travaux récents ont montré que le mode vibratoire des cordes vocales (Colazo-Simon [2]), l'évaluation de la qualité de voix (Michaud [11]) et/ou les relations entre les paramètres de source glottique et la qualité vocale dans la voix parlée et chantée (Henrich [7]), pouvaient être décrits avec précision au moyen de la laryngographie (ou glottographie). Des recherches laryngoscopiques (Esling, Fraser et Harris [3]) ont montré, par ailleurs, que la consonne glottale se produit sur un continuum allant d'une occlusion totale à une fermeture partielle de la glotte.

Comment représenter alors les différents degrés de laryngalisation quand l'Alphabet Phonétique International (API) ne propose que le signe phonétique [ʔ] pour représenter les sons produits par une occlusion dans le larynx et range parmi la voix craquée tous les autres sons que compte la phonation laryngalisée?

La laryngalisation dans la variété espagnole parlée au Yucatan, Mexique

La réflexion menée dans l'article est centrée sur les manifestations acoustiques et physiologiques de la laryngalisation et se donne pour objectif de classifier les différents degrés de l'activité laryngale observés dans les productions laryngalisées de la variété linguistique espagnole parlée au Yucatán, Mexique. Dans cette langue, le contact, depuis plus de cinq siècles, avec la langue maya est notable sur le plan phonétique dans les emprunts (toponymes, cuisine, etc.), les mots hybrides qui se créent lorsque l'espagnol ne recourt pas à l'emprunt mais aussi dans les mots d'origine espagnole avec l'apparition de la laryngalisation ou *saltillo* (petit saut glottique) [9], [6].

Selon l'état de la glotte, nous interprétons les phénomènes laryngaux tantôt comme une occlusive glottale [ʔ], tantôt comme une laryngalisation selon que le segment glottal est placé en fin de mot (ou en coda syllabique) ou en contexte intervocalique (entre deux voyelles identiques et/ou entre deux voyelles distinctes).

Au-delà de l'intérêt scientifique que peut représenter ce travail pour les chercheurs qui travaillent sur le contact de langues, cette étude revêt un intérêt supplémentaire car elle souligne l'absence de symbole ou signe diacritique dans l'Alphabet Phonétique International pour représenter les différents sons pouvant être produits au cours de la phonation laryngalisée.

Nous indiquons tout d'abord la méthodologie adoptée (corpus, locuteurs, instruments et mesures effectuées). Nous présentons et comparons les différentes manifestations de la laryngalisation dans les mots et phrases espagnols. Nous rendons compte, pour finir, des variabilités observées dans la production de la laryngalisation telle qu'elle est réalisée dans l'espagnol parlé au Yucatan.

2. METHODOLOGIE

La production de la laryngalisation -qui se manifeste par un petit saut glottique « saltillo », généralement donné comme un « coup de glotte », est examinée ici dans 30 mots et séquences espagnols aussi bien en position finale de mot, comme à l'intervocalique entre voyelles identiques [V_1V_1] et distinctes [V_1V_2]. Les données analysées proviennent de six hispanophones yucatèques (3 hommes et 3 femmes), âgés entre 25 et 50 ans, tous originaires de la même région du Yucatán, Mexique.

Les données audio et laryngographiques ont été enregistrées simultanément et ont fait l'objet de plusieurs mesures sur les logiciels d'analyse acoustique *Praat* et *Sound Forge.* Sur le signal acoustique, nous avons relevé l'amplitude et les valeurs formantiques des voyelles adjacentes au segment glottal afin de voir si ce dernier exerce une influence sur les formants de la voyelle qui le précède ou le suit. Parallèlement, nous avons observé la qualité de la voix, l'état de la glotte et la durée des cycles glottiques, sur le signal physiologique émis par les vibrations des cordes vocales et captés, in situ par le laryngographe (Fourcin et Abberton [5]). Cet appareil permet d'étudier les mouvements d'ouverture et de fermeture de la glotte lors de la phonation. La méthode consiste essentiellement à placer de part et d'autres du cartilage thyroïdien deux électrodes dans lesquelles circule un courant de haute fréquence (200.000 Hz) à très faible intensité.

2.1. Les divers degrés de laryngalisation

La distinction fondamentale entre les productions laryngalisées observées dans la variété espagnole parlée au Yucatán, repose sur le degré d'accolement des cordes vocales. Une classification en catégories phonétiques est établie ici à partir des réalisations acoustiques et physiologiques observées sur les signaux audio et laryngographiques. De fait, en espagnol du Yucatán, la laryngalisation se produit sur un continuum allant d'une constriction faible - rapprochement des aryténoïdes permettant aux cordes vocales de se fermer sur leur partie postérieure- à une fermeture complète de la glotte.

La phonation laryngalisée répond à une constriction du mécanisme aryépiglottique qui couvre la glotte par-dessus et d'arrière en avant (Esling [4]). Ce phénomène relie hiérarchiquement l'occlusive épiglottale, à la pharyngalisation, à l'occlusive glottale, à la laryngalisation (voix craquée). Les deux premières n'apparaissent pas dans les présentes données (bien que chez certains locuteurs une laryngalisation forte s'assimile parfois à la qualité pharyngalisée (Assadi [1]). Les autres productions dominent ces données et leur distribution systématique parmi nos six locuteurs yucatèques nous permet de voir comment une forte constriction laryngienne se relâche pour se convertir en une constriction moyenne ou faible. Toutes ces configurations glottales entraînent une arythmie dans les cycles glottiques. Dans l'étude, les formes de laryngalisation, présentant une configuration faiblement fermée de la glotte, sont regroupées sous la catégorie «phonation laryngalisée». Cette catégorie réunit les réalisations présentant une constriction faible, moyenne ou forte de la glotte. Selon le degré d'accolement, nous aurons donc une phonation laryngalisée (voix craquée, *creaky voice*), entre autres (Maddieson [10] et Ladefoged et Maddieson [8]), moyennement laryngalisée ou faiblement laryngalisée (voix faiblement craquée, *slighty creaky voice*). Les réalisations classées sous la catégorie «occlusive glottale» sont celles qui ont les aryténoïdes rapprochés au maximum l'un de l'autre afin d'empêcher les cordes vocales de vibrer (glotte ouverte). Les cycles glottiques sont irréguliers non seulement en durée, mais aussi en amplitude d'accolement.

Laryngalisation entre voyelles (V_1V_2)

L'objectif de cette expérience était d'observer et comparer les productions de la laryngalisation à l'intervocalique entre deux voyelles différentes chez les six locuteurs yucatèques.

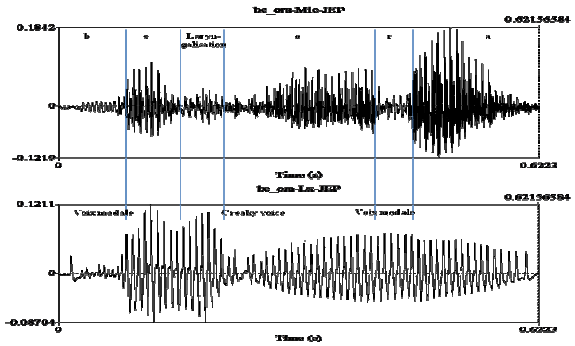


Figure 1: représentation (en haut : signal audio ; en bas : signal laryngographique) de la laryngalisation à l'intervocalique dans le mot «ahora» > /be'ora/ par un locuteur.

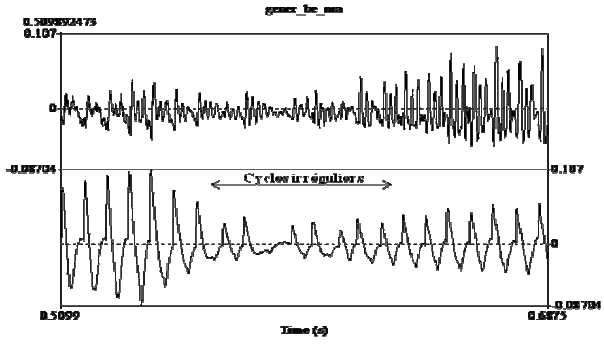


Figure 2 : agrandissement (en haut : signal audio ; en bas : signal laryngographique) des cycles irréguliers dans *be'ora*.

Résultat

Après rapprochement des courbes des signaux laryngographiques des six locuteurs pour les mêmes séquences présentant une laryngalisation entre deux voyelles distinctes, on constate que la laryngalisation se manifeste par une arythmie dans la durée des cycles glottiques et une variation en amplitude d'accolement. Chez cinq locuteurs (1, 3, 5, 6, 7), le segment identifié glottal entre deux voyelles distinctes est le résultat de la résonance produite par la vibration des cordes vocales sous l'influence du constricteur laryngien et, de ce fait, ne peut en être isolé. C'est ce que montrent les expériences que nous avons menées et qui ont consisté à diminuer le volume (-20 dB) pour comparer à l'écoute les deux voyelles puis à supprimer ce segment. Nos résultats laissaient toujours entendre la fusion des deux voyelles et non pas la laryngalisation. Ceci peut s'expliquer par le fait qu'à l'intervocalique la voyelle placée avant le segment glottique se termine généralement par une fermeture brusque et totale de la glotte et celle qui le suit commence toujours par une attaque dure.

Laryngalisation entre voyelles (V_1V_1)

La laryngalisation entre deux voyelles identiques se manifeste aussi par une arythmie dans la durée des cycles glottiques et une variation en amplitude d'accolement.

Résultat

Les locuteurs ont montré des tendances différentes pour la production de la laryngalisation entre deux voyelles identiques. Les locutrices ont majoritairement réalisé la laryngalisation par une forte constriction comme on peut le voir dans la production de la locutrice 1 que l'on peut voir sur la figure présentée ci-dessous (exemple de laryngalisation à l'intervocalique entre 2 voyelles identiques). Cette production est caractéristique de la voix craquée (*creaky voice*). Les cordes vocales ne sont pas suffisamment accolées pour parvenir à produire une fermeture complète qui donnerait une occlusive glottale.

Dans tous les exemples analysés en contexte intervocalique, les mesures montrent que la glottalisation n'affecte ni la voyelle précédente ni celle qui suit. La laryngalisation entre deux voyelles distinctes comme entre deux voyelles identiques se traduit sur le signal audio par une baisse de l'intensité et une interruption de la fréquence fondamentale. Cette interruption dans la régularité des cycles glottiques indique que les vibrations ne sont pas périodiques pendant cette phase. En effet, les périodes ne sont plus régulières, mais sur le signal laryngographique, on constate qu'il y a encore des cycles et que cette laryngalisation se manifeste par une arythmie dans les derniers cycles glottiques. L'interruption dans la régularité des vibrations des cordes vocales observées sur le signal audio ne doit pourtant pas faire illusion : la consonne glottale n'est pas produite par un « arrêt » glottique, mais par de « petits bonds » des cycles glottiques. D'un cycle à un autre, l'arythmie temporelle varie parfois du simple au double. La cavité buccale est fermée, mais l'occlusion de la glotte ne l'est pas totalement. Il est possible d'envisager que les cordes vocales se séparent très légèrement l'une de l'autre tout en maintenant le contact. Les analyses acoustiques et laryngographiques montrent qu'en contexte intervocalique la séquence arythmique placée entre des voyelles de caractère identique ou distinct correspond toujours à une laryngalisation.

Il arrive parfois que la laryngalisation ne soit pas réalisée en particulier lorsque la voyelle est suivie d'une consonne comme dans les exemples suivants : «mata» > /ma:ta/ ; «queso» > /que:so/ ; «malo» > /ma:lo/ ou encore «campo» > /ca:mpo/. Dans ce contexte, la note un allongement de la voyelle chez les six locuteurs.

Laryngalisation en fin de mot

En ce qui concerne la séquence arythmique de fin de mot, celle-ci peut être marquée soit par un ralentissement soit par une interruption brusque des cycles glottiques laissant la glotte ouverte ou fermée.

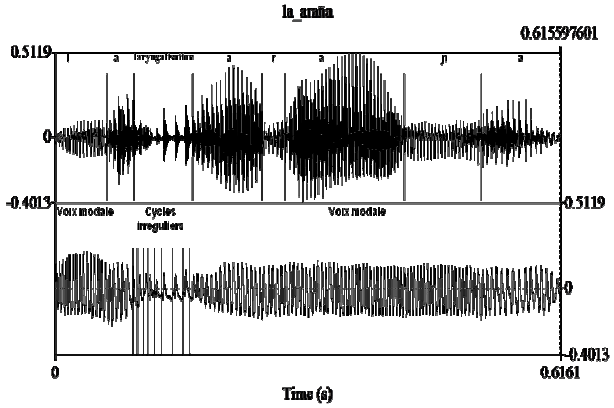


Figure 3: représentation des signaux audio (haut) et laryngographique (bas) de la laryngalisation dans le segment «la araña» prononcé par une locutrice

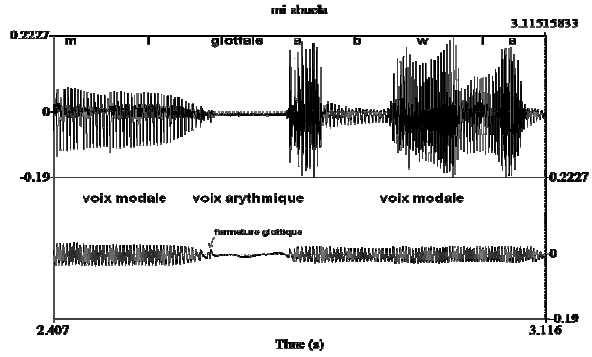
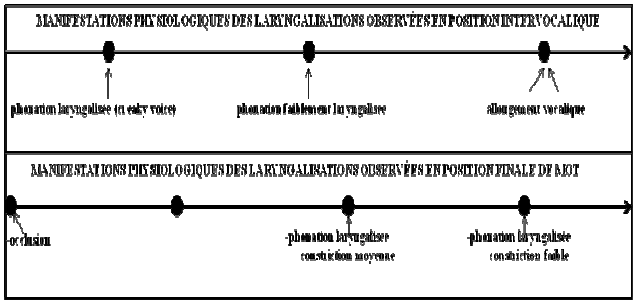


Figure 4: Exemple de laryngalisation (glotte fermée) en finale de mot dans la séquence «mi abuela» prononcée par une locutrice.

Lorsque la laryngalisation se caractérise par une fermeture glottique partielle (baisse de l'énergie et du fondamental) et une irrégularité dans les vibrations mais qu'à partir du dernier cycle, bien qu'il y ait encore du signal, il y a moins d'harmoniques hautes, nous sommes en présence d'une phonation laryngalisée. En revanche, lorsque l'arythmie se termine par une fermeture glottique nette à la fin du dernier cycle glottique comme par exemple dans les mots espagnols *si* et *no*, réalisés respectivement [siʔ] et [noʔ], celle-ci correspond à une occlusive glottale.

Les séquences <voyelle atone suivie d'une voyelle tonique de timbre différent>, comme par exemple dans les phrases «*no podía uno salir*», «*cuatro años*» ou encore, «*este otro*» *pour ne donner que quelques exemples*, se montrent les plus propices à l'apparition de la laryngalisation. Le tableau suivant décrit le continuum des réalisations de la laryngalisation observées dans l'espagnol parlé au Yucatan. En finale de mot en espagnol, on trouve aussi bien l'occlusive glottale /ʔ/ comme la phonation laryngalisée produite par une constriction forte.

Table 1: Classification des réalisations de la laryngalisation en position intervocalique et en fin de mot.



3. CONCLUSION

L'apparition de la laryngalisation dans l'espagnol parlé au Yucatán, montre l'influence phonétique qu'une langue minorée (maya) peut exercer sur la langue nationale (espagnol). Bien que la laryngalisation dans l'espagnol du Yucatán constitue une réalité phonétique, le phénomène non phonologique. L'article montre que cette laryngalisation peut apparaître selon plusieurs modalités en fonction de sa position dans le mot. Il divise le résultat phonétique de la laryngalisation dans l'espagnol parlé au Yucatan en deux catégories : laryngalisation à l'intervocalique et laryngalisation en finale de mot. Ainsi, à l'intervocalique, celle-ci peut se présenter sous la forme d'une phonation laryngalisée ou faiblement laryngalisée. Lorsque la voyelle est suivie d'une consonne où dans les emprunts mayas, la laryngalisation n'a pas lieu, mais il se produit un allongement de la voyelle. En position finale de mot, la laryngalisation se manifeste sur un continuum allant de la fermeture complète à une série de constrictions (forte, moyenne, faible) de la glotte. Cette classification permet entre autres d'apporter une précision sur les variations de l'état de la glotte. Elle soulève aussi la question de la représentation phonétique des divers degrés de laryngalisation qui s'y rattachent et par là même la nécessité pour l'Alphabet Phonétique International de matérialiser les découvertes des phonéticiens à travers des signes phonétiques supplémentaires.

BIBLIOGRAPHIE

[1] ASSADI Sh. S. (2003) *Les phénomènes de glottalisation en persan (langue littéraire/langue parlée)*, thèse de doctorat, Université Sorbonne nouvelle-Paris III.

[2] COLAZO-SIMON A. (2007) *Les phénomènes glottaux en situation de contact linguistique :maya et espagnol du yucatán, Mexique*, thèse de doctorat, Université Paris III-Sorbonne Nouvelle.

[3] ESLING J. H., FRASER K. E. et HARRIS J. G. (2005) « Glottal stop, glottalized resonants, and pharyngeals : A reinterpretation with evidence from a laryngoscopic study of Nuuchahnulth (Nootka) », *Journal of Phonetics*, 33, 383-410.

[4] ESLING J. H. (2006) « States of the glottis », *Encyclopedia of Language and Linguistics* (2e edn.), Oxford, K. Brown ed. Elsevier, 9, p. 425-442.

[5] FOURCIN A. J. et ABBERTON E. (1971), « First Applications of a New Laryngograph », Medical and Biological Illustration, 21, p. 172-182.

[6] GARCÍA FAJARDO J. (1984) *Fonética del español hablado en Valladolid*, México, UNAM.

[7] HENRICH N. (2001) *Étude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception*, thèse de doctorat de l'Université Paris VI.

[8] LADEFOGED P. et MADDIESON I. (1996, 2004) *The Sounds of the World's Languages*, Oxford, Cambridge, Blackwell Publishing LTD.

[9] LOPE BLANCH J. M. (1987) *Estudios sobre el español de Yucatán*,México, UNAM.

[10] MADDIESON I. (1984) *Patterns of sounds*, Cambridge, Cambridge University Press.

[11] MICHAUD A. (2004) « Final consonants and glottalization : new perspectives from Hanoi Vietnamese », *Phonetica*, 61(2-3), p. 1-28 et p. 119-146.

[12] ROUSSELOT P. J. (abbé) (1924), *Principes de phonétique expérimentale*, Didier, H., Paris, Tomes I et II.

Adaptation Autonome de Modèles Acoustiques Pour la Transcription Automatique de Réunions Multilingues

Sethserey Sam[1,2], *Laurent Besacier*[1], *Eric Castelli*[2]

[1]LIG Laboratory, UMR CNRS 5524 BP 53, 38041 Grenoble Cedex 9, France
[2]MICA research center, UMI CNRS 2954, HUT, Hanoi, Vietnam
{sethserey.sam, Laurent.besacier}@imag.fr, eric.castelli@mica.edu.vn

ABSTRACT

We found several challenges in automatic speech transcription system for multilingual meetings. Firstly, the dialog occurs between native and non-native speakers. Secondly, the non-native speakers come from different parts of the world. Thirdly, no data is available to bootstrap the acoustic models. We propose some autonomous online and offline acoustic model adaptation approaches to deal with the above challenges and to improve the performance of the phone recognizers used for automatic transcription purpose. Experiments show that our adaptation approach (hybrid-interpolation with MLLR based on PR-VSM language observer) can provide about 4% absolute gain in phone accuracy compared to the multilingual baseline system and it is even better than the performance of the supervised monolingual systems.

Keywords: ASR, multilingual acoustic modeling, autonomous acoustic models

1. INTRODUCTION

Dans le contexte de la transcription de la parole de type « réunions multilingues », nous trouvons de nombreux défis intéressants: 1) Le dialogue est entre les locuteurs natifs et non natifs ; 2) Les accents des locuteurs non natifs sont variés. Par exemple, l'anglais parlé par des français est remarquablement différent de l'anglais parlé par des vietnamiens. Selon [1], les locuteurs empruntent des caractéristiques acoustiques de leur langue maternelle dans la parole non-native. 3) Il est difficile de trouver suffisamment de données de type « réunion multilingue » pour l'adaptation des modèles acoustiques.

Comme première réponse à ces défis, nous proposons une technique d'adaptation des modèles acoustiques multilingues (MA-Mult) appelée « adaptation autonome » pour améliorer la performance d'un système de transcription de parole de type « réunion multilingue » dans lequel 3 langues sont étudiées : anglais (EN), français (FR) et vietnamien (VN).

L'adaptation autonome signifie que le modèle acoustique (multilingue) est automatiquement réadapté lui-même pour mieux décoder le flux de parole. Ce processus d'adaptation est fait en utilisant un module d'observation de la langue parlée (OL). L'objectif de cet observateur est d'assigner une probabilité à chaque langue candidate et d'utiliser cette information pendant la phase d'adaptation du modèle acoustique (MA-Mult). A notre connaissance, une telle approche fondée sur un « observateur », n'a pas encore été proposée pour l'adaptation acoustique de modèles multilingues.

L'article est organisé de la façon suivante. La Section 2 présente le recueil des données de test. Le système « baseline », le processus autonome, l'observateur de langue et l'adaptation du modèle acoustique sont détaillés dans les sections 3, 4, 5 et 6 respectivement. La Section 7 présente les résultats expérimentaux. La Section 8 conclut cet article.

2. DONNÉES DE TEST

Nous avons extrait les paroles natives et non-natives de trois langues (EN, FR et VN) à partir du corpus « réunion multilingue de MICA » [2]. Tout d'abord, les flux des signaux de parole sont segmentés en locuteurs. Ensuite, seuls les segments de plus de 3 secondes sont sélectionnés pour nos expériences. Une segment sélectionné contient une seule langue parlée (la parole native ou non-native).

Table 1 : Quantité de données de test (valeur en seconde) utilisée dans nos expériences.

Langue	Native/Non-native	Qté de données
EN	EN_fr	715
	EN_vn	56
FR	FR_fr	251
	FR_vn	279
VN	VN_vn	219
TOTAL		**1520**

Dans la 2e colonne du Tableau 1, les termes en lettres majuscules désignent les langues parlées et ceux en minuscules dénotent l'origine des locuteurs (par exemple, EN_fr signifie la langue anglaise parlée par les locuteurs français). Au final, nous avons des données de test d'environ 26 minutes dans laquelle 69 % représente de la parole non-native.

3. SYSTÈME « BASELINE »

Toutes les expériences décrites dans cet article sont établies en utilisant les outils de décodage *Sphinx3* [3].

Notre « baseline » est un système de reconnaissance acoustico-phonétique multilingue (RP-Mult). Le modèle acoustique multilingue (MA-Mult) du système « baseline » est créé en combinant les 3 modèles acoustiques monolingues (MA-Mono) existants. Les 3 MA-Monos sont EN (anglais), FR (français) et VN (vietnamien) et ils sont entrainés respectivement sur les corpus WSJ [4], BREF120 [5], et VNSpeechCorpus [6]. La combinaison de ces 3 MA-Monos s'est faite en utilisant la méthode ML-sep [7]. Cela signifie qu'il n'existe pas de données à partager, à travers les 3 MA-monos combinés. En outre, notre modèle multilingue (MA-Mult) est un modèle indépendant du contexte qui contient au total 124 phonèmes: EN (40 phonèmes), FR (43 phonèmes) et VN (41 phonèmes). Un phonème est représenté par un HMM de 3 états de 16 Gaussiennes. Dans cet article, nous n'étudions que l'adaptation du modèle acoustique pour améliorer la performance du système « baseline ». Les expériences décrites plus loin concernent donc uniquement le décodage acoustico-phonétique.

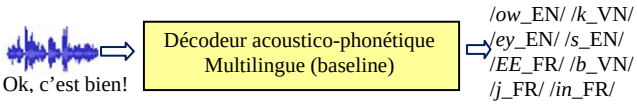


Figure 1 : Un example de décodage en utilisant notre système « baseline ».

En Figure 1, chaque phonème en sortie est étiqueté par la langue à laquelle il appartient. Notre concept d'adaptation autonome utilisant un observateur de langue démarre de ce constat initial.

4. PROCESSUS.D'ADAPTATION AUTONOME

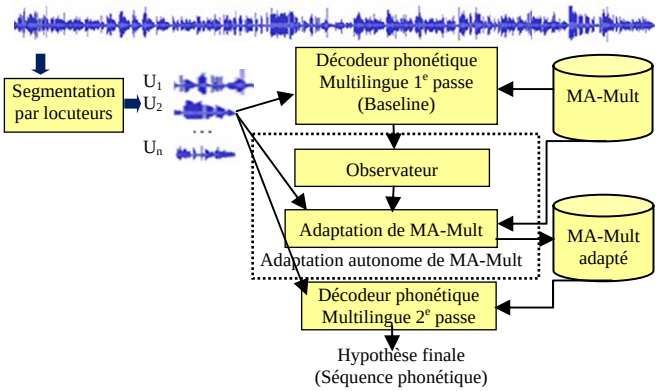


Figure 2 : Le processus d'adaptation autonome du modèle acoustique multilingue.

En Figure 2, U_1,.., U_n sont les données de test mentionnés dans la Section 2. Le module « Observateur » fournit les informations de langue de chaque segment U_i en générant les probabilités de langues selon deux méthodes décrites dans la section 5. Puis l'adaptation du modèle MA-Mult se réalise en utilisant ces probabilités. Enfin, le décodage phonétique deuxième passe utilise le MA-Mult adapté pour décoder U_i.

Par conséquent, notre « Observateur » est au cœur du processus d'adaptation autonome de MA-Mult.

5. OBSERVATEUR DE LANGUE (OL)

Il est important de préciser le fait que l'observateur de langue est plus qu'un simple module d'identification de la langue. L'observateur de langue attribue un ensemble de probabilités pour les 3 langues possibles, pour chaque segment de parole. Ces probabilités seront utilisées pendant la phase d'adaptation de MA-Mult. En outre, nous pensons que l'observateur donne des informations riches sur le segment de parole en train d'être décodé (la langue parlée ainsi que l'origine du locuteur). Par exemple, si l'observateur de langue donne: P (EN) = 0.5, P (FR) = 0,4 et P (VN) = 0.1, le segment pourra être considéré comme de l'anglais prononcé par un français (ou vice-versa).

5.1. Méthode 1 : Phone Equiprobable

Nous proposons une méthode appelée « Phone Equiprobable » (PhEquiPro) afin de générer les probabilités de langues de chaque segment. La méthode PhEquiPro fournit les probabilités de langues en utilisant l'hypothèse du système « baseline » (Figure 2) par la formule suivante:

$$P(L_i) = \frac{n(L_i)}{N} \qquad (1)$$

où *P, Li, n, N* représentent respectivement, la probabilité, la langue (FR ou EN ou VN), le nombre de phonèmes d'une langue dans la séquence phonétique, et le nombre total de phonèmes dans la séquence phonétique.

Par exemple, si « *h*_EN *e*_EN *l*_FR *o*_EN» est l'hypothèse du « système baseline » lors du décodage d'un segment "Bonjour". Avec l'Equation 1, PhEquiPro produit les distributions (probabilités) de langues de la façon suivante: P(EN) = 3/4, P(FR) = 1/4 et P(VN) = 0.

5.2. Méthode 2 : PR-VSM

Comme une alternative à notre approche simple PhEquiPro, une approche phonotactique de reconnaissance de langue appelée « *Phone Recognizer-Vector Space Modeling* » (PR-VSM) [8] a été également utilisée. Dans ce cas, le système « baseline » est utilisé comme le décodeur phonétique (la partie *frontend* du PR-VSM). Dans la partie « *back-end* », chaque modèle vectoriel (VSM) qui représente une langue, est entrainé en utilisant 6 heures (2 heures par langue) des corpus suivants: WSJ (EN), BREF120 (FR) et VNSpeechCorpus (VN). Comme probabilité de langue de l'observateur, nous utilisons ici la probabilité a posteriori du reconnaisseur PR-VSM selon chaque modèle de langue. Pour plus de détails sur cette méthode, voir [8].

6. ADAPTATION DU MODÈLE MULTILINGUE

Pour une adaptation de MA-Mult autonome sans besoin de données externes, nous proposons deux approches d'adaptation: l'adaptation « *online* » et « *offline* ». L'adaptation « *online* » signifie que seul le segment de

parole actuel est disponible pour le processus d'adaptation. D'autre part, l'adaptation « *offline* » peut utiliser à la fois le segment actuel et toutes les informations des segments déjà décodés dans notre réunion.

6.1. Adaptation « online »

Maximum likelihood linear regression (*MLLR*)

MLLR est simple et connue pour être robuste pour l'adaptation non supervisée puisqu'elle peut être appliquée avec des petites quantités de données d'adaptation. L'hypothèse du système « baseline » du segment en cours de décodage est utilisée pour l'adaptation du modèle MA-Mult original. Dans cet article, toutes les adaptations *MLLR* ne se basent que sur l'adaptation des moyennes des modèles acoustiques (Mean Only Adaptation) [9].

Interpolation de modèles acoustiques (*INTER*)

Parce que la parole non-native existe souvent dans les données de test (69%), nous étudions également la technique d'adaptation entre modèles acoustiques issus de plusieurs langues. Cette technique est proposée dans [1] et est une version modifiée de l'interpolation de modèles. Elle a été validée pour la reconnaissance de la parole non-native. Dans ce cas, le modèle acoustique de la langue la plus probable selon l'observateur est considérée comme le modèle principal (MA_{cible}) et les autres sont considérés comme les modèles sources. Parce que nous avons plus de deux modèles à interpoler au total, nous proposons de faire l'interpolation en deux fois successivement où 2 modèles acoustiques (le MA_{cible} et un des modèles sources) sont interpolées à chaque fois (les formules plus détaillées qui représentent l'interpolation de deux modèles acoustiques sont présentées dans [1]). Enfin, le MA-Mult adapté est créé en combinant les 2 modèles acoustiques interpolés, en utilisant la méthode de combinaison LM-sep, déjà mentionnée à la Section 3. Nous formulons finalement l'interpolation de plusieurs modèles acoustiques comme suit :

$$MA - Mult_{adapté} = \sum_{i=1}^{m-1} [1 - P(L_i)].MA_{cible} + P(L_i).MA_{L_i} \quad (2)$$

Où.P, L_i représentent respectivement la probabilité (Équation 1) d'une des deux langues sources. Σ représente la combinaison des modèles acoustiques interpolés, et m est le nombre total de langues impliquées.

Nous étudions également l'interpolation des modèles acoustiques suivie par *MLLR* (*INTER-MLLR*) en appliquant simplement une passe d'adaptation *MLLR* au modèle acoustique interpolé.

6.2. Adaptation « offline »

MLLR par langue (*Same Language MLLR, SL-MLLR*)

Ce processus d'adaptation se fait en 3 étapes successives: 1) pour tous les segments déjà décodés, nous les regroupons en fonction des meilleurs probabilités de langue fournies par le module « observateur » (donc on a au total 3 groupes: EN, FR et VN), 2) ensuite pour chaque groupe, on utilise tous les segments de parole et les hypothèses du décodeur 1^{e} passe correspondant pour faire l'adaptation *MLLR*; 3) enfin, le segment en cours est décodé avec le modèle acoustique multilingue adapté.

Phone Mapping-MLLR (*PM-MLLR*)

La seule différence entre *PM-MLLR* et *SL-MLLR* est que, *PM-MLLR* remplace, dans l'hypothèse utilisée pendant l'étape MLLR, tous les phonèmes des autres langues par les phonèmes similaires de la langue la plus probable selon l'observateur. La substitution de phonèmes (*phone mapping*) entre les trois langues est faite en se basant sur le tableau API (Alphabet Phonétique International).

Par exemple, si l'hypothèse du système « baseline » d'un segment est «*h*_EN *e*_EN *l*_FR *o*_EN", la langue détectée sur ce segment fourni par l'observateur est EN. Cela signifie que la technique *MP-MLLR* remplace le phonème /*l*_FR/ par le phonème similaire /*l*_EN/. Après le processus de substitution, *PM-MLLR* effectue le même processus d'adaptation en 3-étapes que *SL-MLLR*.

7. RÉSULTATS EXPÉRIMENTAUX

7.1. Système multilingue « Baseline » vs. les décodeurs monolingues

Table 2 : La performance (% phonèmes corrects reconus) du système« baseline » vs. celles de 3 RP-Monos (supposant une idéntification des langues parfaite).

Native/Non-native	RP-Mult(Baseline)	RP-Mono (oracle)
EN_fr	39.8	44.1
EN_vn	38.5	40.7
FR_fr	41.8	48.7
FR_vn	40.8	44
VN_vn	43.3	50.3
Moyen	**40.84**	**44.56**

Le Tableau 3 compare la performance entre le système « baseline » et les trois autres reconnaissances.acoustico-phonétiques monolingues (RP-Mono) EN, FR et VN dont les modèles acoustiques sont entrainées respectivement sur les corpus WSJ, BREF120 et VNSpeech-Corpus mentionnés dans la Section 3. Dans le cas RP-Mono, la comparaison est effectuée en utilisant les données de test du Tableau 2 en supposant une parfaite identification de la langue parlée pour chaque segment à décoder. Puis, le bon modèle acoustique (EN, FR ou VN) est utilisé pour décoder tous les segments de même langue (native ou non-native). Par exemple, les segments FR_fr et FR_vn sont décodés par RP-Mono en français.

Il est important de rappeler que les données de test contiennent 69% de parole non-native ce qui explique pourquoi la performance de reconnaissance phonétique des systèmes RP-Mult (baseline) et RP-Monos sont relativement faibles.

7.2. Observateur : PhEquiPro vs. PR-VSM

Dans le tableau 3, nous évaluons notre observateur de langue (OL) en fonction de deux métriques: 1) Taux correct d'identification de la langue parlée (LID, chiffres à gauche de "/"): ici le max des probabilités à posteriori de OL est utilisé pour choisir la langue la plus probable. 2) Taux correct d'identification de la langue parlée et de l'origine du locuteur (LID+ORG, chiffres à droite de "/") : un test est considéré comme un succès si le max des probabilités à posteriori donné par OL est soit la langue parlée soit la langue maternelle du locuteur (dans ce dernier cas, la deuxième langue la plus probable doit être la langue parlée sinon on considère le test comme un échec).

Table 3 : Evaluation de différent types d'observateur basée sur les taux corrects LID/LID+ORG [%].

Native/Non-native	PhEquiPro	PR-VSM
EN_fr	**94.34/100**	89.68/100
EN_vn	**66.67/66.67**	61.9/61.9
FR_fr	89.28/89.28	**96.43/96.43**
FR_vn	0/0	**17.39/30.44**
VN_vn	**55.62/55.62**	50.28/50.28
Moyen	68.83/70.8	**69.18/74.03**

Comme présenté dans le Tableau 3, la performance des deux observateurs (PhEquiPro et PR-VSM) est très comparable en termes de LID mais celle de PR-VSM donne de meilleures performances en terme de LID+ORG. En outre, la performance du français parlé par les Vietnamiens est plus faible comparativement à d'autres langues, car les vietnamiens impliqués ont un niveau de français très limité. Donc, on pourrait probablement conclure que l'identification de la langue est basée non seulement sur la langue parlée, mais aussi sur l'origine du locuteur ainsi que sur les connaissances de la langue parlée du locuteur.

7.3. Adaptation de MA-Mult : online vs. offline

Table 4 : La performance (% phonèmes corrects reconus) de différentes techniques d'adaptation (utilisant l'observateur PR-VSM).

Native/Non-native	Baseline	ONLINE			OFFLINE	
		MLLR	INTER	INTER-MLLR	SLI-MLLR	PM-MLLR
EN_fr	39.8	39.6	45.7	**45.7**	41.8	42.7
EN_vn	38.5	38.2	**43.9**	43.7	35.2	41.65
FR_fr	41.8	43.1	40.7	41.3	43.4	**44.3**
FR_vn	40.8	**41.3**	38.3	39.6	39.5	41.2
VN_vn	43.3	**43.5**	42.85	43.15	41.1	37.3
Moyenne	40.84	41.2	44.22	**44.68**	41.96	42.38

Grâce aux performances des systèmes présentées dans le Tableau 4, nous pouvons faire les commentaires suivants:

- L'adaptation autonome de type « interpolation », fondée sur l'observateur PR-VSM, améliore sensiblement la performance du système quand les segments de parole sont non-natifs mais dégrade celle de parole native. Ce résultat montre, cependant, que l'adaptation autonome, a un fort potentiel pour le décodage non supervisé de la parole non-native ;

- L'adaptation *INTER-MLLR* fondée sur l'observateur PR-VSM offre de meilleures performances (44.68%) que les systèmes RP-Monos (44.56%), dans lesquels une parfaite identification de la langue parlée est considérée pour tous les segments avant de les décoder. Cela confirme que pour les réunions multilingues, qui contiennent souvent de la parole non-native, l'adaptation autonome d'un modèle multilingue que nous avons proposée est un choix prometteur.

8. CONCLUSION

Nous avons exploré une approche autonome pour l'adaptation des modèles acoustiques multilingues afin d'améliorer la performance du système de transcription multilingue. L'avantage de cette approche est que le modèle acoustique s'adapte automatiquement pendant le décodage et sans l'aide de données externes. En outre, notre adaptation proposée (*INTER-MLLR* basée sur l'observateur *PR-VSM*) peut fournir plus de 4% d'amélioration absolue par rapport à la performance du système « baseline » et elle est même meilleure que les performances obtenues par les systèmes de décodages phonétiques monolingues supervisés.

BIBLIOGRAPHIE

[1] T.P. Tan, and L. Besacier, Modeling context and language variation for non-native speech recognition, INTERSPEECH, 1429-1432, 2007.

[2] S. Sam, Vers des modèles autonomes pour la reconnaissance automatique de la parole multilingue, RJCP, Avignon, 2009.

[3] http://cmusphinx.sourceforge.net

[4] Doug B. Paul and Janet M. Baker, The Design for the Wall Street Journal-based CSR Corpus, In Proceedings of the DARPA SLS Workshop, 1992.

[5] L.F. Lamel, J.L. Gauvain and M. Eskénazi. BREF, A Large Vocabulary Spoken Corpus for French. In Proc. Eurospeech, pp. 505-508, 1991.

[6] V.B. Le, D.D. Tran, E. Castelli, L. Besacier, J-F.Serignat , "Spoken and written language resources for Vietnamese", LREC, Lisbon, 2004

[7] T. Schultz, K. Kirchhoff, Multilingual Speech Processing, Academic Press, 2006.

[8] H. Li, B. Ma, and C.H Lee, A Vector Space Modeling Approach to Spoken Language Identification, in IEEE Transactions on Audio, Speech and Language Processing, 2007.

[9] Leggetter C. J. and Woodland P. C., Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Computer Speech and Language, 9:171-186, 1995.

Perception de vocoïdes postérieurs fermés synthétisés : l'effet de la configuration labiale et de la position de la langue sur les auditeurs francophones et japonophones

Takeki Kamiyama

Laboratoire de Phonétique et Phonologie (UMR7018), CNRS / Sorbonne Nouvelle
19 rue des Bernardins, 75 005 Paris, France
takeki.kamiyama@univ-paris3.fr
http://lpp.univ-paris3.fr/equipe/takeki_kamiyama.htm

ABSTRACT

Japanese /u/ is produced with a fronter tongue constriction with less closed lips than French /u/. Such an articulatory configuration results in higher F1 and F2 (F2 > 1000 Hz) than French /u/. Variations in tongue position and lip configuration were simulated using VTCalc (Maeda [10]). A perceptual identification test was conducted using these synthesised vowels with 16 native French-speaking listeners and 16 native Japanese-speaking listeners. The results show that the back tongue position and a high degree of labiality were both essential for French speakers to perceive French /u/, whereas Japanese speakers perceived Japanese /u/ in zones where French speakers identified /u/ and /ø/. These findings explain the fact that Japanese-speaking learners of French have difficulty in distinguishing /u/ and /ø/ in perception and production.

Keywords: Japanese /u/, French /u/, tongue position, lips, articulatory synthesis, identification

1. INTRODUCTION

Il est largement connu que des phonèmes de différentes langues qui sont transcrits phonémiquement avec le même symbole ne correspondent pas nécessairement aux mêmes réalisations phonétiques. C'est le cas du /u/ du français, du japonais et de l'anglais. La réalisation phonétique du /u/ japonais est moins arrondie, plus antérieure, et avec une constriction moins importante que celle du /u/ français (Uemura [14] pour le japonais, Bothorel *et al*. [1] pour le français : Figure 1), et souvent notée [ɯ] en transcription phonétique large. La différence articulatoire entre le /u/ français et le /u/ japonais donne lieu à la différence acoustique suivante : le /u/ français en contexte isolé est caractérisé par deux premiers formants proches (il s'agit d'une voyelle focale) et inférieurs à 1000 Hz (Figure 2, gauche ; CALLIOPE [3] pour le contexte [pu] ; Gendrot et Adda [4] pour des données de la parole continue journalistique). Le F2 du /u/ japonais est supérieur à 1000 Hz (Figure 2, milieu ; Sugitô [13], Mokhtari et Tanaka [11] pour 22 mots) et non regroupé avec F1 (il s'agit d'une voyelle « acoustiquement centrale » : formants situés approximativement à équidistance ; Vaissière [15]). Le /u/ (comme dans « *goose* ») de l'anglais américain est caractérisé par une légère diphtongaison (Wells [16], entre autres), par un F2 plus élevé que celui du /u/ français (Hillenbrand *et al*. [5], entre autres), et les deux premiers formants ne sont pas regroupés, comme en japonais (Figure 2, droite).

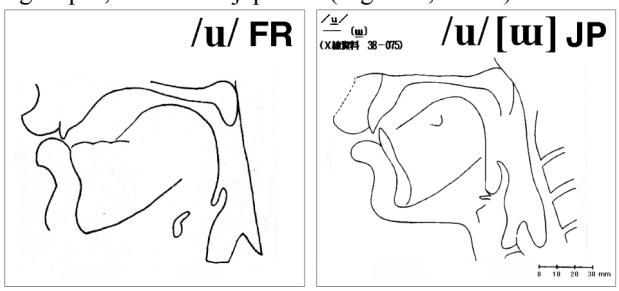


Figure 1 : Profil sagittal du /u/ français (gauche : Wioland [17]) et du /u/ japonais (droite : Uemura [14]) en contexte isolé.

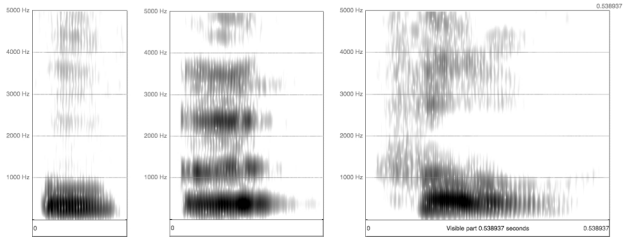


Figure 2 : Spectrogramme à bande large (Praat [1] ; largeur de fenêtre : 5 ms) du /u/. À gauche, un /u/ français prononcé par un locuteur francophone natif dans la phrase cadre « *je dis* /u/ *comme dans loup* » (Kamiyama [7]) ; au milieu, un /u/ japonais prononcé par un locuteur japonophone natif dans la phrase cadre « *sore-wa* /u/ *to iimasu* [on appelle cela /u/] » (Kamiyama [7]) ; à droite, un /u/ dans le mot « *who* » prononcé par un locuteur américain (Jones *et al*. [6]).

Comment les deux paramètres articulatoires, la configuration des lèvres et la position de la langue, influencent-ils sur le résultat acoustique et la perception des auditeurs francophones et japonophones ? Nous avons simulé les variations de ces deux paramètres articulatoires en utilisant le synthétiseur articulatoire VTCalc (Maeda [10]).

2. POSITION DE LANGUE ET LABIALITÉ : SYNTHÈSE ARTICULATOIRE

Le synthétiseur articulatoire VTCalc calcule les caractéristiques acoustiques correspondant à un profil sagittal, depuis la glotte jusqu'aux lèvres, spécifié par les 7 paramètres suivants :

1) Position de la mâchoire : *-jaw-*
2) Position du dos de la langue : *-tongue-*
3) Forme de la langue : *-shape-*
4) Position de l'apex de la langue : *-apex-*
5) Aperture des lèvres : *-lip_ht-*
6) Protrusion des lèvres : *-lip_pr-*
7) Hauteur du larynx : *-larynx-*

Trois séries de continuum (avec des pas espacés de façon régulière sur les dimensions articulatoires) ont été créées en partant des paramètres articulatoires d'un [u] français (Figure 3) et en en modifiant quelques-uns :

1. modification de l'ouverture des lèvres et de l'avancement de la langue : entre un [u] et un [ø] français et son extension (vers [e]), en modifiant les paramètres *jaw*, *tongue*, *shape*, *apex*, *lip_ht* et *lip_pr* (Figure 4, haut) ;

2. avancement de la langue uniquement : [u] français avec la langue mise progressivement vers l'avant, en modifiant les paramètres *jaw*, *tongue*, *shape* et *apex* (Figure 4, milieu) ;

3. ouverture des lèvres uniquement : [u] français avec les lèvres progressivement moins arrondies et protruses, en modifiant les paramètres *lip_ht* et *lip_pr* (Figure 4, bas).

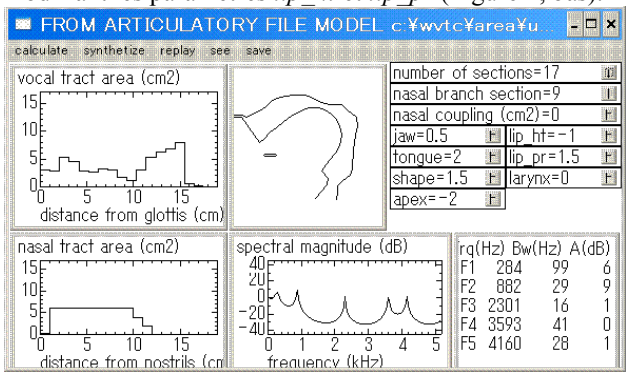


Figure 3 : Simulation d'un [u] à partir des données articulatoires d'un [u] français sous VTCalc (Maeda [10]). Notons les deux fortes constrictions (labiale et vélaire) et le F2 inférieur à 1000 Hz.

Figure 4 : Simulation de : 1) avancement de la langue et ouverture des lèvres (en haut) ; 2) avancement de la langue seulement (au milieu) ; 3) ouverture des lèvres seulement (en bas). Profil sagittal et vue frontale des lèvres.

Nous observons les effets acoustiques suivants sur ces 3 séries de simulation (Figure 5) :

Série 1 : augmentation de F1, F2 (notamment de F2) ;
Série 2 : augmentation de F2 ;
Série 3 : augmentation de F1, F2 (notamment de F1).

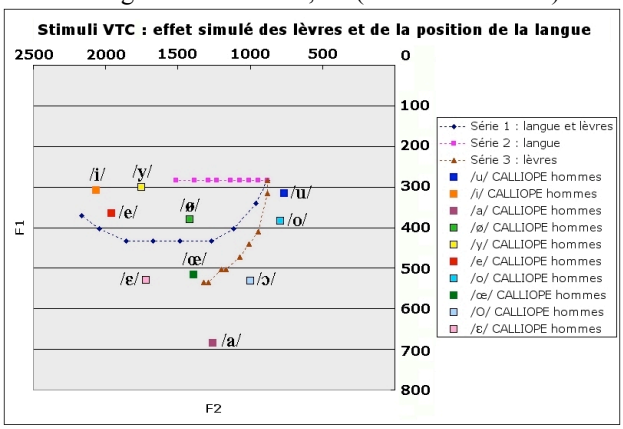


Figure 5 : Simulation d'avancement de la langue et d'ouverture des lèvres (série 1 : losanges), d'avancement de la langue seulement (série 2 : carrés), et d'ouverture des lèvres seulement (série 3 : triangles) : effet acoustique sur le plan F1-F2, en comparaison avec les valeurs formantiques de CALLIOPE [3] (contexte [pV] ; [pVʀ] pour [i ɛ a ɔ œ] ; 10 hommes).

Quels sont les effets de ces deux paramètres articulatoires sur la perception des auditeurs natifs du français et du japonais ?

3. PERCEPTION D'AUDITEURS FRANCOPHONES ET JAPONOPHONES

Une expérience d'identification perceptive a été réalisée en utilisant les vocoïdes synthétisés présentés dans la section précédente.

3.1. Méthode

Vingt-cinq stimuli ont été ainsi obtenus (= 9 x 3 – 2, celui d'un [u] français étant commun aux trois séries : Figure 5). Ces stimuli ont été disposés dans un ordre semi-aléatoire.

Deux groupes d'auditeurs ont participé à cette expérience de perception. Le premier était composé de 16 auditeurs natifs du japonais n'apprenant pas le français. Ils étaient âgés de 24 à 64 ans. Tous étaient originaires de Tokyo et de ses environs. Aucun d'entre eux n'avait vécu plus d'un an à l'étranger. Le second groupe réunissait 16 auditeurs natifs du français, âgés de 18 à 38 ans, résidant en région parisienne lors de l'expérience.

L'expérience a été effectuée sur un ordinateur. L'auditeur disposait d'une souris et d'un casque. Les tâches de l'auditeur étaient d'écouter les stimuli, de les identifier comme une des voyelles de sa langue maternelle (français

ou japonais), et de les évaluer sur une échelle de 1 (mauvais) à 5 (bon). Il était informé qu'il allait entendre des voyelles prononcées par des étrangers.

Les auditeurs japonais disposaient comme choix de réponse non seulement des cinq voyelles du japonais (/i e a o u/) mais aussi de trois syllabes avec le /j/ (/ja jo ju/). Ces dernières ont été ajoutées afin de permettre la réponse /ju/, séquence utilisée dans des mots d'emprunt pour représenter la voyelle /y/ du français (ex. Hugo /ygo/ > /'juRgoR/ [jɯːgoː], ou /'jugoR/ [jɯgoː] ; voyelle suivie de /R/ représente une voyelle longue). Ces voyelles et syllabes ont été transcrites en syllabaires *katakana* dans l'ordre des syllabaires japonais (アイウエオ, ヤユヨ : /a i u e o/, /ja ju jo/).

L'expérience était précédée par un entraînement pour familiariser l'auditeur avec les tâches. Pendant l'expérience, chaque auditeur a entendu les stimuli dans 4 ordres différents (25 stimuli x 4 répétitions = 100 réponses par auditeur).

3.2. Résultats

Les résultats sont présentés ici sous forme de réponse modale (la réponse la plus fréquemment observée) pour chacun des stimuli et par groupe d'auditeurs (Figure 6) et de nombre de réponse de chaque stimulus (0-64) de la série 1 (modification de la langue et des lèvres : Figure 7).

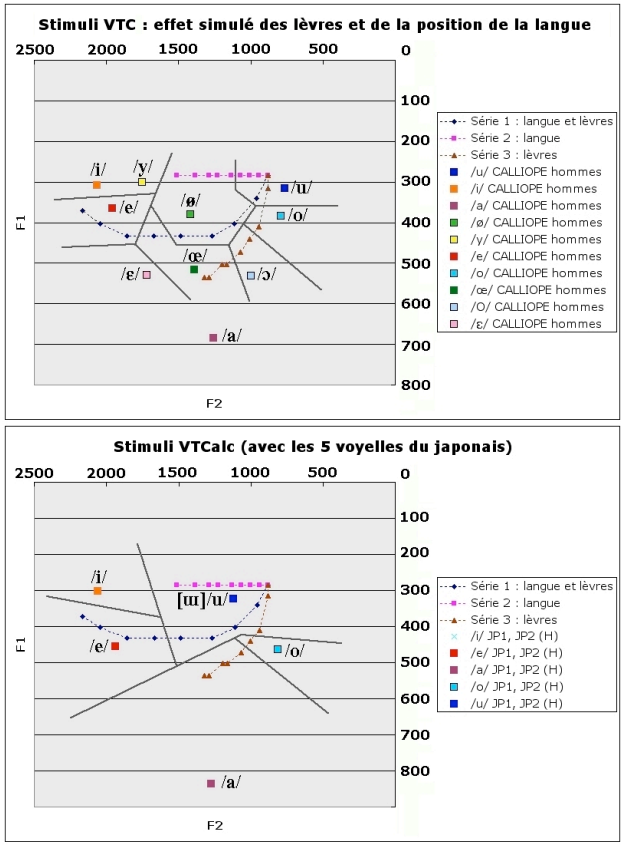


Figure 6 : Répartition de l'espace acoustique F1-F2 selon la réponse modale des stimuli :1) (en haut) auditeurs francophones (64 réponses par stimulus : 16 auditeurs x 4 réponses), en comparaison avec les valeurs formantiques de CALLIOPE [3] (contexte [pV] ; [pVʀ] pour [i ɛ a ɔ œ] ; 10 hommes) ; 2) (en bas) auditeurs japonophones (64 réponses par stimulus : 16 auditeurs x 4 réponses), en comparaison avec les valeurs formantiques de Kamiyama [7] (contexte : « *sore-wa* /V/ *to iimasu* [on appelle cela /V/] » ; 2 hommes x 6 répétitions).

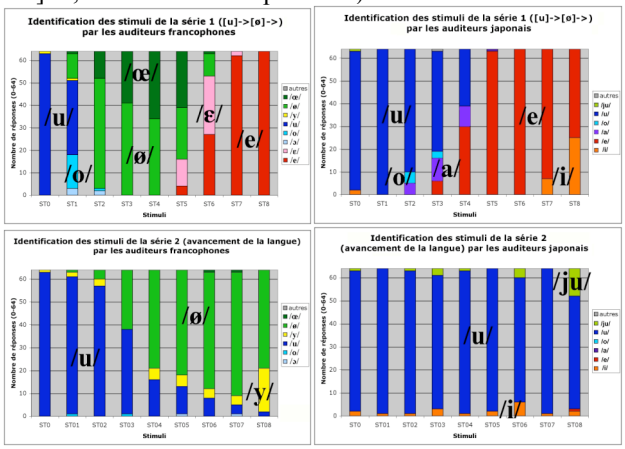


Figure 7 : Nombre de réponses (0-64) d'identification des stimuli des série 1 (langue et lèvres) et 2 (langue) : à gauche, 16 auditeurs francophones ; à droit, 16 auditeurs japonophones. 64 réponses par stimulus (16 auditeurs x 4 réponses).

Les résultats obtenus auprès des auditeurs francophones sont les suivants : i) un fort degré de labialité et la position postérieure de la langue sont nécessaires pour que les auditeurs francophones perçoivent le /u/ français ; ii) des vocoïdes synthétisés en tant que voyelles postérieures fermées non-arrondies ont été perçus majoritairement comme /œ/ ; iii) les stimuli qui ont un F2 supérieur à 1000 Hz ne sont pas perçus comme un bon exemplaire du /u/ français, mais plutôt comme /ø/.

Par contraste, les auditeurs japonais (n'apprenant pas le français) qui ont identifié les mêmes stimuli comme voyelles du japonais ont perçu généralement le /u/ japonais dans les zones où les auditeurs francophones ont entendu /u/ et /ø/, ce qui suggère que les auditeurs japonophones perçoivent le /u/ japonais dans une zone articulatoire et acoustique (sur l'axe de F2) plus large.

En ce qui concerne les voyelles perçues en seconde position, nous observons deux cas de figure (Figure 7) : 1) d'une part, on trouve des voyelles qui ont les mêmes caractères d'antériorité et de labialité que la réponse majoritaire, avec une différence d'aperture d'un seul degré (ex. /ø/ et /œ/) ; 2) d'autre part, quand la perception majoritaire passe d'une catégorie à une autre (ex. /u/->/ø/ dans la série 2 : position de la langue), c'est l'un des deux qui constitue la réponse majoritaire, l'autre étant en seconde position.

4. DISCUSSION ET CONCLUSION

Les différences de comportements entre les auditeurs francophones et japonophones observées dans cette expérience de perception se résument comme suit. Dans

les zones articulatoires et acoustiques où les auditeurs francophones montrent un passage du /u/ au /ø/, les auditeurs japonophones entendent en général un /u/ japonais, avec une évaluation relativement bonne. Les francophones sont très sensibles à des écarts par rapport au /u/ français aussi bien sur le plan articulatoire (avancement de la langue et délabialisation) que sur le plan acoustique (F2 élevé), tandis que la zone du /u/ japonais est plus large, notamment au niveau de F2.

Ces résultats illustrent des caractères non-linéaires et non-univoques des correspondances entre les domaines articulatoire, acoustique et perceptif. Les voyelles « acoustiquement centrales » (F2 situé vers 1500 Hz, à mi-chemin entre F1 et F3 : Vaissière [15]) ont été observées dans plusieurs séries de continuum synthétisés : 1) autour de [ø] français (série 1) ; 2) avec les lèvres de [u] et la langue en avant (série 2) ; 3) avec la langue de [u] (postérieure) et les lèvres ouvertes (séries 3). Ces vocoïdes ont été tous identifiés majoritairement comme voyelles antérieures arrondies /ø/ (les deux premiers cas) et /œ/ par les auditeurs francophones, quel que soit le lieu d'articulation. Il a été effectivement montré que différentes configurations articulatoires peuvent correspondre à des sons acoustiquement similaires (compensation pour la labialité : Savariaux *et al*. [12]) et que la labialité des voyelles centrales n'est pas toujours identifiée auditivement (Lisker et Rossi [9]).

Les résultats de la présente étude expliquent également le fait que les apprenants japonophones ont des difficultés à distinguer /u/ et /ø/ sur les plans de la perception et de la production et que le /u/ français isolé prononcé par les apprenants japonophones (probablement à un lieu d'articulation proche de celui du /u/ japonais) est identifié majoritairement comme /ø/ par les auditeurs francophones natifs (Kamiyama et Vaissière [8], Kamiyama [7]).

REMERCIEMENTS

L'auteur remercie Jacqueline Vaissière et Antonia Colazo-Simon ainsi que les deux relecteurs anonymes pour leurs commentaires et suggestions sur les versions antérieures de cet article.

BIBLIOGRAPHIE

[1] P. Boersma, D. Weenink. *Praat: doing phonetics by computer* (logiciel). http://www.praat.org/

[2] A. Bothorel, P. Simon, F. Wioland, and J.-P. Zerling. *Cinéradiographie des voyelles et consonnes du français*. Publications de l'Institut de Phonétique de Strasbourg, Strasbourg, 1986.

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

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

[5] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97(5): 3099-3111, 1995.

[6] D. Jones, P. Roach, J. Hartman, and J. Setter. *Cambridge English Pronouncing Dictionary* (17th edition). Cambridge University Press, Cambridge, UK, 2006.

[7] T. Kamiyama. *Apprentissage phonétique des voyelles du français langue étrangère chez des apprenants japonophones*. Thèse de doctorat de phonétique, Université Paris 3 – Sorbonne Nouvelle, 2009.

[8] T. Kamiyama and J. Vaissière. Perception and production of French close and close-mid rounded vowels by Japanese-speaking learners. *Revue AILE – LIA*, 2: 9-41, 2009.

[9] L. Lisker and M. Rossi. Auditory and visual cueing of the [Rounded] feature of vowels. *Language and Speech*, 35(4): 391-417, 1992.

[10] S. Maeda. A digital simulation method of the vocal tract system. *Speech Communication*, 1(3-4): 199-229, 1982.

[11] P. Mokhtari and K. Tanaka. A Corpus of Japanese Vowel Formant Patterns. *Bulletin of Electrotechnical Laboratory* 64 (special issue): 57-66, 2000.

[12] C. Savariaux, P. Perrier, J.-P. Orliaguet, and J.-L. Schwartz. Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis. *The Journal of the Acoustical Society of America* 106(1): 381-393, 1999.

[13] M. Sugitô. *Ôsaka - Toukyou akusento onsei jiten CD-ROM: kaisetsuhen* [Dictionnaire CD-ROM sonore d'accent des japonais d'Osaka et de Tokyo : commentaires]. Maruzen, Tokyo, 1995.

[14] Y. Uemura (Kokuritsu Kokugo Kenkyûjo [Institut national de la langue japonaise]). *Nihongo no boin, shiin, onsetsu: chouon undou no jikken-onseigakuteki kenkyuu* [Voyelles, consonnes et syllabes en japonais : étude en phonétique expérimentale sur les mouvements articulatoires]. Shûei shuppan, Tokyo, 1990.

[15] J. Vaissière. *La phonétique*. Presses Universitaires de France, Paris, 2006.

[16] J. C. Wells. *Accents of English*. Cambridge University Press, Cambridge, UK, 1982.

[17] F. Wioland. *Prononcer les mots du français : des sons et des rythmes*. Hachette, Paris, 1991.

Influence de la décision voisé/non-voisé dans l'évaluation comparative d'algorithmes d'estimation de F_0

François Signol, Jean-Sylvain Liénard, Claude Barras

LIMSI-CNRS – Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
Bâtiments 502bis et 508, Université Paris-Sud, 91400 Orsay Cedex, France
{francois.signol, jean-sylvain.lienard, claude.barras}@limsi.fr
http ://www.limsi.fr/Individu/{signol, lienard, barras}

ABSTRACT

This paper aims at pointing out the impact of the voiced/unvoiced decision on the results obtained by any pitch estimation algorithm. To be equitable, a comparative evaluation must provide the voided/unvoiced state in terms of overvoiced rate and undervoiced rate. The interaction between the detection and the estimation is usually underestimated in the pitch evaluation framework. Its consequence is a large bias in the comparative evaluation. This paper illustrates this phenomenon through the use of two comparative evaluation methodologies on 25 minutes of voiced speech. PEA could be improved not only from the F0 estimation part but also from the voiced/unvoiced decision part. This raises the problem of quantifing properly the voicing strength of a speech frame.
Keywords: F_0 estimation, voiced/unvoiced decision, comparative evaluation.

1. Introduction

Le problème d'estimation de la F_0 d'un signal monopitch est un problème ancien dont les principes sont donnés dans [6] et qui reste encore largement ouvert. Dans le domaine de la parole, il existe un grand nombre d'algorithmes d'estimation de pitch (AEP). L'évaluation des AEP est une étape indispensable et l'usage qui est fait des résultats obtenus conditionne la complexité du processus d'évaluation. L'évaluation la plus simple est celle qui consiste à donner des indicateurs de performance objectifs pour un AEP donné à un instant donné. Le processus se complique lorsque l'objectif est de quantifier les progrès apportés au même AEP dans le temps. Le processus d'évaluation le plus compliqué est certainement celui dont le but est de fournir une comparaison des performances d'AEP différents. Il impose en effet d'être *équitable* alors que les AEP testés peuvent être très hétérogènes dans leur but et leur fonctionnement. Une évaluation comparative doit au minimum uniformiser les corpus qui sont de même nature de parole (lue, spontanée, etc.), l'intervalle de recherche de la F_0 et les instants temporels d'estimation. Un léger décalage peut aboutir à plusieurs pour-cents d'erreur en plus (ou en moins). Ces points peuvent paraître évidents mais leur influence sur les résultats est souvent sous-estimée. Dans la mesure du possible, il est aussi préférable d'homogénéiser la durée des trames et de prendre en compte les éventuels post-traitements correctifs (prise en compte du contexte temporel).

De nombreuses évaluations de la qualité des AEP existent. C'est dans le domaine musical que l'évaluation est la plus dynamique sous l'impulsion de MIREX [1]. Cette campagne d'évaluation annuelle propose de nombreux travaux concernant les meilleures méthodologies à suivre pour évaluer des AEP multipitch. L'analyse de cette littérature met en évidence une analogie entre le problème de détection de l'étendue temporelle d'une note dans une tâche de transcription automatique de piano et celui de la décision voisé/non-voisé (VnV) en parole [5]. De telles campagnes d'évaluation n'existent pas encore pour la parole. Il existe néanmoins des travaux concernant l'évaluation comparative d'AEP [3][10][9]. Dans ce domaine, l'estimation de F_0 sur des signaux monolocuteur relativement neutres d'expressivité et enregistrés en condition de laboratoire, les meilleures performances sont de l'ordre de 1% de taux d'erreurs grossières (GER). C'est ce GER qui est généralement utilisé comme critère principal de comparaison de performance. Le GER quantifie le nombre de F_0 de référence que l'AEP n'a pas réussi à correctement estimer selon une certaine tolérance. Cette tolérance est différente entre le domaine musical et la parole. En parole, la valeur typique est de 20% d'écart relatif alors qu'en musique elle est de 3%. Les 3% de tolérance du domaine musical s'expliquent facilement par le fait que l'espace des fréquences est discrétisé en notes espacées d'un demi-ton (6%). Ces notes n'existant pas en parole (hormis la voix chantée qui n'est pas traitée dans ce papier), une tolérance de 3% n'est pas nécessairement la plus adaptée. Le choix de 20% peut sembler élevé mais s'explique par le fait que la grande majorité des erreurs d'estimation de F_0 sont des erreurs d'octave ou de sous-octave.

Les évaluations ne donnent pas souvent la performance des AEP dans la détection de F_0 (décision VnV). Or, l'estimation de F_0 est dépendante de cette décision. Cette dépendance est différente selon les AEP. Dans certains cas, la décision VnV est préalable à l'estimation de F_0 (ex. YIN, SWIPE, PAL présentés dans la section 3), dans d'autres cas, la décision VnV est entremélée à l'estimation de F_0 au sein d'un processus de suivi de F_0 (programmation dynamique ou autre) comme dans PRAAT [2] [1] par exemple. L'interaction entre la détection de F_0 et l'estimation de F_0 est un fait connu mais son influence sur le GER est

1. Musical Information Retrieval Evaluation eXchange : `http://www.music-ir.org/mirex/`
2. Fonction classique *To pitch (ac)...*

très souvent sous-estimée. Si la décision VnV n'est pas identique pour les AEP comparés, l'équité de l'évaluation comparative n'est plus garantie et le taux de GER ne suffit plus à la comparaison des AEP.

L'article est organisé de la manière suivante : la section 2 présente les métriques classiques utilisées pour l'évaluation d'AEP. La section 3 détaille les corpus, les AEP et la manière dont les références F_0 ont été calculées. Dans la section 4, l'influence de la décision VnV sur l'estimation de F_0 est présentée. Dans la section 5, ce point crucial est justifié par l'intermédiaire de deux méthodologies d'évaluation, la première ne donnant pas d'informations sur la qualité de la décision VnV et la seconde uniformisant d'emblée cette décision. L'objectif de ce papier est double : d'une part il s'attache à démontrer qu'il est impératif de prendre en compte l'interaction entre décision VnV et estimation de F_0 et d'autre part, il se propose de clarifier les manières de procéder pour réaliser une évaluation comparative équitable.

2. Métriques d'évaluation

Les métriques présentées dans cette section sont proposées dans l'article [9]. Deux processus sont à évaluer : la décision VnV qui renvoie à la détection de F_0 et l'estimation de F_0.

La métrique concernant la décision VnV compare la décision VnV faite par la référence et celle faite par l'AEP. Une trame peut prendre deux états : voisé ou non-voisé. Deux types d'erreurs sont possibles : l'erreur de sous-voisé (faux-rejet) dans laquelle la trame de référence est voisée alors que l'AEP est non-voisé, ou bien l'erreur de sur-voisé (fausse alarme) pour laquelle la trame de référence est non-voisée alors que l'AEP la considère voisée. Le taux d'erreurs de sur-voisé OVR est formalisé dans l'équation 1. Il correspond au nombre de trames en erreur de sur-voisé rapporté au nombre de trames de référence non-voisées.

$$OVR = \Omega[R_U \cap H_V]/\Omega[R_U] \quad (1)$$

$\Omega[...]$ désigne l'opérateur cardinal ensembliste. $\cap$ désigne l'opérateur d'intersection ensembliste. R_U désigne l'ensemble des trames de références non-voisées et H_V désigne l'ensemble des trames données voisées par l'AEP (hypothèses). Le taux d'erreurs de sous-voisé est donné dans l'équation 2. Il correspond au nombre de trames en erreur de sous-voisé rapporté au nombre de trames de référence voisées.

$$UVR = \Omega[R_V \cap H_U]/\Omega[R_V] \quad (2)$$

R_V désigne l'ensemble des trames de référence voisées et H_U désigne l'ensemble des trames données non-voisées par l'AEP. Le couple des taux d'erreurs (OVR, UVR) permet de connaître le point de fonctionnement d'un AEP en termes de décision voisé/non-voisé.

La métrique concernant l'estimation de F_0 classiquement utilisée est le taux d'erreurs grossières noté GER. Il s'agit du nombre de trames dont la valeur F_0 d'hypothèse est distante de plus de 20% en écart relatif absolu de la valeur F_0 de référence. Le GER est calculé sur les trames qui sont déclarées voisées par la référence et par l'AEP. Le GER est donné dans l'équation 3. R_E désigne l'ensemble des trames de référence mal estimées.

$$GER = \Omega[R_E]/\Omega[R_V \cap H_V] \quad (3)$$

3. Matériau

Cette section détaille les matériaux indispensables à toute évaluation comparative d'AEP : les corpus utilisés, la manière dont sont générées les valeurs F_0 de référence et les AEP comparés.

3.1. Corpus de parole

Trois corpus de parole sont utilisés et totalisent environ 50 minutes de parole. Il s'agit des corpus Bagshaw, Keele et Mocha, tous trois choisis car ils contiennent les Electro-Glotto-Grammes (EGG) associés à chaque signal permettant de calculer la vérité terrain F_0 ("*groundtruth*"). Ces trois corpus sont comparables en nature de parole : lue et peu expressive. Bagshaw [3] comporte deux locuteurs, un homme et une femme, prononcant chacun 50 phrases courtes. Le corpus contient environ cinq minutes de parole. Keele [4] comporte dix locuteurs, cinq hommes et cinq femmes, prononcant chacun le même énoncé lu ("*The north wind and the sun...*"). Le corpus contient environ cinq minutes de parole. Mocha [5] comporte deux locuteurs, un homme et une femme, prononçant chacun 460 phrases courtes. Le tout forme un corpus contenant 25 minutes de segments voisés.

3.2. Références F_0

Les valeurs F_0 de référence sont extraites de manière automatique à partir de l'EGG. Pour cela, un algorithme simple utilisant l'autocorrélation est employé. Un post-traitement correctif permet de supprimer les erreurs d'octave ou de sous-octave. Les résultats automatiques obtenus ont été vérifiés manuellement afin de contrôler leur validité. Cette annotation automatique a pour avantage de s'affranchir du travail d'annotation manuel qui peut être fastidieux.

3.3. AEP évalués

Trois AEP sont évalués. Ce choix est justifié par le besoin de paramétrer à volonté la décision VnV ce qui n'est pas toujours le cas selon les AEP. Ils sont également tous purement trame-à-trame (pas de post-traitement). YIN [4] est un algorithme temporel s'appuyant sur la *squared difference function* ou SDF décrite dans l'équation 4. YIN produit des fonctions de pitch dans lesquelles la fréquence du minimum local de plus faible amplitude est considérée comme hypothèse F_0. $x(n)$ est un signal discret de N échantillons. τ varie entre 0 et $N-1$.

$$SDF(\tau) = \sum_{n=0}^{N-1} [x(n) - x(n+\tau)]^2 \quad (4)$$

SWIPE [2] et PAL [7] sont des algorithmes fréquentiels qui calculent le produit scalaire entre un spectre

3. www.cstr.ed.ac.uk/research/projects/fda
4. www.liv.ac.uk/Psychology/hmp/projects/pitch.html
5. www.cstr.ed.ac.uk/research/projects/artic/mocha.html

d'amplitude (noté $|X|$) et une fonction noyau (notée K). Pour SWIPE, cette fonction est une ondelette cosinusoïdale décroissante et pour PAL, il s'agit du peigne alterné [7]. Les fonctions de pitch produites (notée Φ) quantifient une force de périodicité de chaque fréquence dans l'intervalle de recherche. Le maximum local le plus fort est considéré comme hypothèse F_0. L'équation 5 donne la valeur de la fonction de pitch pour une fréquence F_c donnée.

$$\Phi(F_c) = \sum_{n=0}^{N-1} |X(nF_c)|.K(F_c, f) \quad (5)$$

4. Décision voisé/non-voisé et GER

Cette section montre l'impact de la décision VnV sur le taux de *GER*. En effet, si l'AEP est paramétré de manière à ne considérer que les trames dont le voisement est fort, alors l'estimation de la F_0 sur ces trames est probablement correcte. Par contre, si l'algorithme évalué estime la F_0 sur n'importe quelle trame, alors il considère des trames dont le voisement est plus litigieux et sera davantage sujet aux erreurs. La figure 1 montre ce comportement sur l'algorithme SWIPE. Le comportement est similaire pour les deux autres algorithmes. Ces valeurs sont obtenues sur les trois corpus comptant 25 minutes de parole voisée (cf. section 3). Comme attendu, la probabilité d'erreur augmente fortement lorsque la force de périodicité diminue. Cette probabilité représente le nombre d'erreurs à 20% rapportées au nombre total de valeurs à estimer (correctes+erreurs) pour une force de périodicité donnée. Or, la manière de calculer la force de périodicité diffère d'un algorithme à un autre. Ainsi, le comportement de chacun des algorithmes quant à la décision VnV est différent. Dans ce contexte de théorie de la décision, les courbes ROC ou DET sont classiquement utilisées. Les courbes DET [8] données figure 2 illustrent les différences des trois AEP en termes de détection de F_0 sur le corpus de Mocha. L'objectif n'étant pas ici de donner un classement des AEP, les trois algorithmes sont rendus anonymes et remplacés par A, B et C.

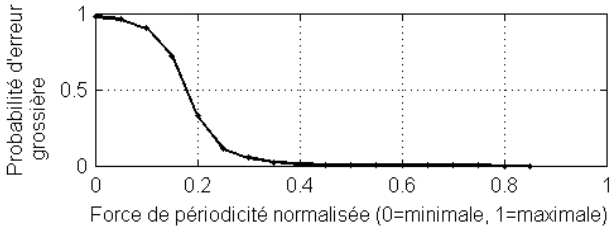


Figure 1: Probabilité d'erreurs grossières en fonction de la force de périodicité.

Il reste enfin à montrer l'influence de la décision VnV sur l'estimation de F_0. Ce point est l'objet de la figure 3. Les courbes présentent le taux *GER* en fonction du point de fonctionnement de la décision VnV pour les trois AEP. Comme attendu, lorsqu'un AEP favorise le sous-voisement il fait moins d'erreurs grossières (OVR/UVR inférieur à 1). Ainsi, il est possible d'être artificiellement meilleur que d'autres AEP si le point de fonctionnement de l'AEP testé lui est favorable. Mais cette supériorité n'est que la conséquence d'une sélection implicite de trames "bien voisées". Or, c'est justement dans le traitement des trames litigieuses que se manifeste la qualité véritable d'un AEP. Pour éviter ce biais, il est nécessaire soit de donner systématiquement les taux d'OVR et d'UVR, soit de mettre en œuvre une méthodologie d'évaluation qui uniformise les OVR et UVR pour tous les AEP comparés.

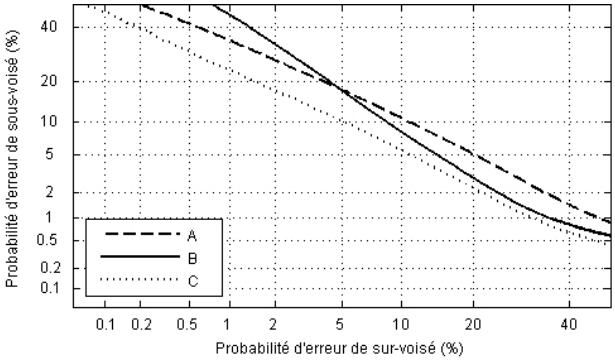


Figure 2: Courbe DET obtenue sur Mocha.

5. Evaluation comparative

Cette section permet de confirmer par l'expérience ce qui a été vu dans la section 4. Pour cela deux méthodologies d'évaluation sont utilisées. La première est nommée méthodologie *standard* et la seconde est nommée méthodologie *sans sous-voisé*. Cette section montre que les résultats bruts de la méthodologie standard sont à prendre avec précaution et qu'il faut toujours mettre en parallèle les résultats d'OVR et UVR avec le *GER*.

5.1. Méthodologies

Standard (STD) Utilisée dans [2], elle consiste à évaluer les AEP dans les versions préconisées par les auteurs. Le principal avantage est sa simplicité et le fait de pouvoir la mettre en place quel que soit l'AEP (boîtes noires). La décision VnV n'est pas uniforme selon les AEP. Ceci implique que les trames évaluées ne sont pas forcément les mêmes. A la limite, les résultats obtenus par deux AEP différents pourraient être les mêmes alors que les trames évaluées sont toutes différentes. Pour être complète et équitable, cette méthodologie **doit** renseigner le taux d'erreurs de sous-voisé et le taux d'erreurs de sur-voisé.

Sans sous-voisé (SSV) Utilisée dans [3], elle est la plus équitable et consiste à régler la décision VnV des AEP comparés de manière à ce qu'aucune erreur de sous-voisé ne soit faite. Ainsi tous les AEP sont évalués sur les mêmes trames et ces trames sont les trames voisées de la référence. La méthodologie impose un taux d'UVR à 0% et un taux d'OVR à 100%. Cette méthodologie n'est pas toujours applicable car elle nécessite de pouvoir régler la décision VnV, ce qui n'est pas toujours le cas.

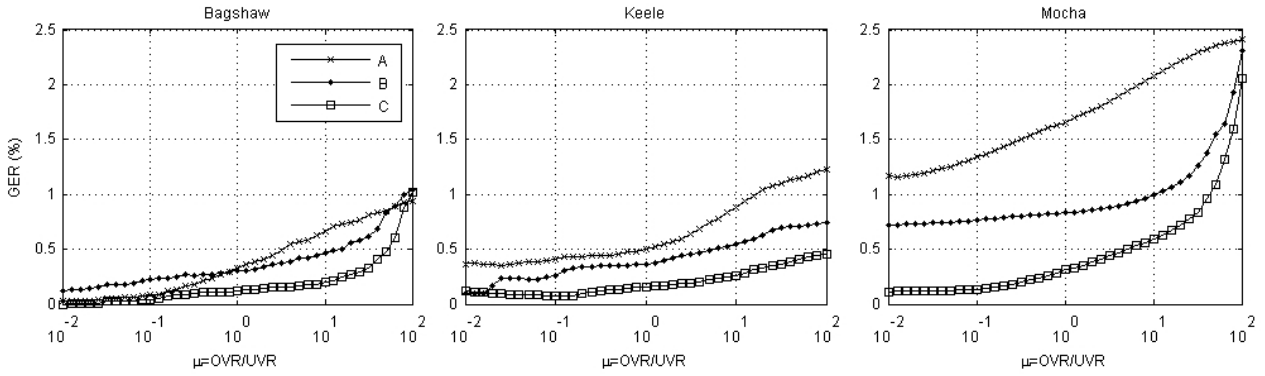


Figure 3: Évolution du GER en fontion du rapport OVR/UVR. L'échelle des abcisses est semi-logarithmique. L'abcisse 10^0 correspond à l'*Equal Error Rate*.

5.2. Résultats

Les résultats obtenus sur les 25 minutes de parole voisée des trois corpus sont récapitulés dans la table 1. Les deux méthodologies sont appliquées. Pour la méthodologie standard, les seuils de décision VnV sont fixés aux valeurs préconisées par les auteurs : 0.2 pour YIN et SWIPE et 1 pour PAL. L'intervalle de recherche de la F_0 est $[60, 500Hz]$.

Table 1: Récapitulatifs des résulats obtenus.

	STD			SSV
	OVR(%)	UVR(%)	GER(%)	GER(%)
A	3.99	21.56	0.38	1.36
B	18.76	3.03	0.45	0.76
C	11.97	2.72	0.19	0.53

Dans la méthodologie standard, les taux d'erreurs grossières sont au-dessous de 0.5%. Le point de fonctionnement ($\mu = OVR/UVR$) de A vaut 0.19, celui de B vaut 6.19 et celui de C vaut 4.40. B est l'algorithme qui donne les moins bonnes performances. Néanmoins, ce résultat doit être considéré avec les taux d'erreur OVR et UVR pour être complet car les μ sont différents. A favorise le sous-voisement contrairement à B et C. Il est donc avantagé dans son taux de GER. C'est effectivement ce qui transparait dans la colonne SSV puisqu'il apparait que le B n'est plus l'algorithme donnant la plus faible performance. B est légèrement moins performant que C alors que A fait environ deux fois plus d'erreurs grossières que les deux autres algorithmes.

6. Conclusions

Cet article montre clairement l'influence de la décision voisé/non-voisé (détection de F_0) sur le taux d'erreurs grossières (estimation de F_0). Il est indispensable de donner les taux d'OVR et d'UVR reflétant la décision VnV des AEP. Si seul le GER est fourni, l'équité de l'évaluation s'en trouve affectée. Ce travail montre que la qualité d'un AEP ne se limite pas à une pure estimation de F_0 mais aussi à la décision VnV. Il y a donc deux axes d'améliorations des AEP : travailler sur l'estimation de F0 et travailler sur la décision VnV. Ce dernier point soulève le problème de vérité terrain en termes de décision voisé/non-voisé et de quantification de la force de voisement.

Références

[1] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA 1993*, volume 17, pages 97–110, 1993.

[2] Arturo Camacho and John G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3) :1638–1652, 2008.

[3] Alain de Cheveigne and Hideki Kawahara. Comparative evaluation of f0 estimation algorithms. In *Eurospeech 2001*, volume 4, pages 2451–2454, 2001.

[4] Alain de Cheveigne and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4) :1917–1930, 2002.

[5] Nuno Fonseca and Anibal Ferreira. Measuring music transcription results based on a hybrid decay/sustain evaluation. In *ESCOM 2009*, pages 119–124, Finland, 2009.

[6] Wolfgang Hess. *Pitch Determination of Speech Signals : Algorithms and Devices.* Springer-Verlag, Germany, heidelberg edition, 1983.

[7] Jean-Sylvain Liénard, François Signol, and Claude Barras. Speech fundamental frequency estimation using the alternate comb. In *Interspeech 2007*, Antwerpen, Belgium, 2007.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Eurospeech 1997*, pages 1895–1898, 1997.

[9] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5) :399–418, 1976.

[10] Peter Veprek and Michael S. Scordilis. Analysis, enhancement and evaluation of five pitch determination techniques. *Speech Communication*, 37(3-4) :249–270, 2002.

Etude des caractéristiques des collections de documents pour les évaluations de système de questions-réponses

Guillaume Bernard, Martine Adda-Decker, Sophie Rosset

LIMSI-CNRS
Université Paris-Sud XI
{gbernard,madda,rosset}@limsi.fr

ABSTRACT

In this paper, we propose some measures to evaluate the proximity between different tasks in question answering on speech transcripts evaluation campaign. These measures are based on word distances between elements of question and expected answer. The results of these measures seem to indicate that the document type is not the most important aspect. For instance, the language used for the task seems to be more relevant for task proximity.
Keywords spoken corpus for questions-answering system, measures of task proximity

1. Introduction

Les systèmes de questions-réponses (QR) [10] ont pour objectif de fournir des réponses à des questions formulées par un utilisateur. Les réponses sont aussi bien extraites de documents textuels que de sources audio. Contrairement à un système de recherche d'information, les questions sont posées en langue, et non pas sous la forme d'une liste de mots-clés. De plus, le système retourne une réponse courte plutôt qu'une liste de documents. Par exemple, la réponse attendue à la question *Combien d'années Nelson Mandela passa-t-il en prison ?* est 27 ans. Les campagnes d'évaluations des systèmes QR [8] permettent d'observer les progrès effectués. On fournit un ensemble de questions aux participants, et les systèmes doivent chercher la bonne réponse (si elle existe) dans une collection de documents, qui est également fournie. L'origine des documents a une importance sur les problématiques rencontrées par les participants au cours de ces évaluations. Par exemple, les articles journalistiques sont structurés en phrases, le plus souvent clairement délimitées. Les journaux radio-diffusés transcrits automatiquement utilisent eux aussi une notion de phrases mais ces dernières sont délimitées automatiquement, et donc très éloignées des phrases des articles journalistiques.

La campagne QAst [9, 8] (Question-Answering on Speech Transcriptions) a pour objectif d'évaluer les performances des systèmes QR sur des transcriptions orales et sur différentes langues. QAst permet donc à la fois d'évaluer le comportement des différentes méthodes par rapport à des documents oraux ainsi que la polyvalence des systèmes sur différents langues. Cependant, la parole n'est qu'une modalité, et divers genres peuvent être concernés. Ainsi, les transcriptions orales pour le français proviennent de documents journalistiques (Broadcast News), tandis que celles de l'espagnol sont extraites du parlement européen (EPPS). Les échanges lors des sessions du parlement européen utilisent de la parole préparée. Les données en anglais sont elles de deux genres : parole préparée provenant des sessions du parlement européen également et de la parole spontanée provenant de séminaires et de réunions.

L'objectif de cet article est de présenter des mesures permettant d'évaluer les différences entre les types de documents utilisés pour chaque langue. Nous présentons tout d'abord la campagne d'évaluation QAst, ainsi que les caractéristiques à priori des différents corpus. Nous expliquons ensuite les mesures d'évaluation appliquées, ainsi que les résultats obtenus sur les différentes collections de documents. Ces résultats sont ensuite analysés, et nous présentons nos conclusions et perspectives.

2. QAst, une campagne d'évaluation multi-lingue

QAst est une campagne d'évaluation multi-lingue dont les collections de documents sont des transcriptions orales. Nous avons choisi dans cet article de nous baser sur les éditions 2008 et 2009 de la campagne. Trois langues sont traitées : l'anglais, l'espagnol et le français [9, 8]. Nous avons pour ces deux éditions 4 types de documents différents. Le premier genre est le séminaire. Une personne seule parle pendant que les autres présents dans la salle écoutent et interviennent parfois mais rarement. Il s'agit du corpus *CHIL* [3] (anglais). Le second est la réunion de travail. Plusieurs personnes parlent ensemble, et souvent en même temps, se coupent la parole, etc. Il s'agit du corpus *AMI* [1] (anglais). Le troisième correspond à des émissions d'informations de la radio. Il s'agit du corpus *ESTER* [5] (français). Enfin le dernier genre se rapporte à des enregistrements de sessions du Parlement Européen en anglais et en espagnol. Chaque parlementaire fait à son tour un discours préparé pendant quelques minutes, le président de séance s'assurant du bon déroulement de la session. Il s'agit du corpus *TC-STAR* [7]. Les corpus *CHIL* et *AMI* ne sont présents que dans l'édition 2008, tandis que les émissions journalistiques de radios en français (Broadcast News) et les sessions du parlement Européen en anglais et en espagnol (EPPS) sont présentes en 2008 et 2009.

Les questions sont différentes d'une année sur l'autre.

Deux approches pour créer chaque corpus de questions ont été utilisées. En 2008, les questions étaient créées par un évaluateur à partir des documents. En 2009, l'idée était d'avoir des questions plus spontanées. On fournissait à des utilisateurs des extraits de documents. Il leur était alors demandé de créer des questions oralement sur des informations en relation avec le passage mais dont la réponse n'était pas présente.

La question à laquelle nous nous intéressons ici est de savoir si ces différences peuvent être importantes dans le cadre des systèmes de QRs. En particulier, nous voulons vérifier l'importance du type de document (parole préparée, spontanée etc.), l'importance de la langue et l'impact de la méthodologie d'acquisition des questions. Nous avons donc effectué des mesures qui sont présentées dans la section suivante 3.

3. Mesures des caractéristiques des collections de documents

Trois mesures sont présentées ci-dessous. Ces mesures se basent principalement sur les deux caractéristiques suivantes : la répétition des éléments clef d'une question, et leur distance par rapport à la réponse attendue. Chacune de ces mesures est effectuée avec une comparaison stricte et des transformations (synonymie, forme lemmatisée, etc ...).

3.1. Distance entre les éléments de la question et la réponse

L'objectif de cette mesure est de fournir une distance "globale" entre les éléments de la question et la réponse, pour ainsi évaluer la répartition des informations dans les documents. Cette distance est une distance *physique* qui est calculée en comptant le nombre de mots entre chaque élément de la question et la réponse. Pour chaque question d'une collection de documents nous calculons la distance globale entre les éléments de la question et la portion du document contenant la bonne réponse. Seul les éléments définis comme critiques sont utilisés dans le calcul. Les éléments considérés comme étant important sont des entités nommées (classiques, étendues et non-spécifiques) et des expressions à mots multiples. La mesure globale de la question est donc la moyenne des distances de chaque élément. De ce fait, la mesure globale d'un corpus est la moyenne des distances globales de chaque question. Les deux exemples ci-dessous expliquent comment la mesure globale est calculée pour deux questions. Dans le premier exemple, la réponse correcte à la question *Quelle organisation belge a été déclarée criminelle ?* est *Vlaams Blok*. Les distances ont été calculées entre cette réponse et chaque élément critique de la question : *belge*, *organisation* et *criminelle*. Les distances correspondantes en mots sont 7, 2 et 3. La mesure globale est donc de 4.

Quelle **organisation belge** *a été déclarée* **criminelle** *?*

La Court suprême **belge** *a confirmé un précédent arrêt déclarant que Vlaams Blok est une* **organisation criminelle**.

Le deuxième exemple est basé sur un passage de texte plus long. La bonne réponse à la question *Quel chef politique de Palestine est mort récemment ?* est *Arafat*. Les éléments critiques sont *mort*, *Palestine*, *politique* et *chef*. Les distances correspondantes sont 0, 11, 36 et 35, et la mesure globale est donc de 20.

Quel **chef politique** *de* **Palestine** *est* **mort** *récemment ?*

La **mort** *d'Arafat implique que nous allons avoir à présent une nouvelle élection en* **Palestine**. *L'Union Européenne a déclaré à Israel que le dialogue entre les deux pays est important. Il est nécessaire d'avoir un nouveau* **chef politique** *aussi vite que possible.*

3.2. Répétitions des éléments de la question

L'idée de cette mesure est d'évaluer la présence des éléments des questions dans les documents d'une collection. Dans un premier temps, on compte le nombre d'occurrences de chaque élément de la question ainsi que les transformations associées. La mesure donne la moyenne pour une question du nombre d'occurrences des éléments. Dans un deuxième temps, la distance en mots entre chaque occurrence de chaque élément est calculée. Ainsi, une question est représentée par deux mesures : la moyenne des occurrences des éléments, et la moyenne de la distance entre chaque occurrence. Par exemple :

Où y a -t-il eu un **attentat** *en* **Irak** *?*

Mais on commence par l' actualité en **Irak** *; quelques 400 tonnes d' un explosif très puissant se sont volatisées . L'* **Irak** *c' est également la violence quotidienne , avec notamment cet* **attentat** *à Bagdad , capitale de l'***Irak** *, qui visait un convoi militaire australien 12 personnes ont été tuées .*

Dans la question évaluée, le système identifie deux éléments critiques : *attentat* et *Irak*. Par rapport au passage évalué, ces deux éléments auront respectivement 1 et 3 répétions. Cette question a donc une valeur de répétition moyenne de 2. *attentat* n'ayant qu'une seule occurrence, la distance de répétition de la question est calculée seulement à partir de celle de *Irak*. Entre la première occurrence et la seconde, on a une distance de 12, et entre la seconde et la troisième 15. La fréquence moyenne de répétition est donc de 13.

4. Évaluation et analyse

4.1. Présentation des résultats

Nous avons calculé ces différentes mesures sur l'ensemble des corpus et des questions. En ce qui concerne les répétitions (distance ou nombre de répétitions), les résultats ne montrent pas de différences imputables au type de documents ou à la langue. Par contre, on observe pour ce qui est des distances entre éléments de la question et réponse attendue des différences notables. Ces différents résultats sont montrés dans les figures 1 et 2. Le tableau 1, quant à lui, montre l'évolution de la distance moyenne entre deux campagnes

d'évaluations (méthode de création des questions différentes).

	Broadcast News		
	Moy	Ecart type	Δ
2008	45	100	+98
2009	143	431	
	EPPS anglais		
	Moy	Ecart type	Δ
2008	97	284	+39
2009	136	310	
	EPPS espagnol		
	Moy	Ecart type	Δ
2008	381	851	-359
2009	22	73	

Tab. 1: Tableau de l'évolution de la distance moyenne entre les éléments de la question et la réponse attendue sur les corpus Broadcast News et EPPS.

4.2. Analyse des résultats

Ces mesures ont été appliquées sur nos différents corpus de manière à identifier les caractéristiques et en déterminer lesquels partageaient des spécificités communes. Le premier aspect qui nous intéresse est de savoir si la langue d'un corpus a plus ou moins d'importance que son type. La figure 1 présente les résultats en terme de distance entre éléments de la question et réponse attendue obtenus en comparant EPPS anglais et espagnol (2008 et 2009). Les mesures sont clairement différentes entre les corpus selon la langue. Par exemple, les distances moyennes sont plus faibles en anglais qu'en espagnol en 2008. C'est moins vrai pour les données 2009. Il semble donc que la langue a un impact plus important que le type de documents. Et la différence entre 2008 et 2009 a peut-être un rapport avec la manière dont les questions ont été créées. Si la langue est plus importante que le type de documents, alors nous devrions avoir des mesures similaires pour une même langue.

Nous avons donc comparé le corpus EPPS anglais avec les corpus CHIL et AMI. Nous avons là aussi appliqué la mesure de distance entre les éléments d'une question et la réponse. La figure 2 montre les résultats de cette comparaison. Les corpus EPPS anglais et CHIL ont des résultats assez proches, particulièrement dans les premières classes de valeurs. Par contre, le corpus AMI a des résultats différents. On peut donc supposer que la langue est plus importante que le type de corpus, mais que le type de corpus a aussi un impact. CHIL et EPPS anglais sont des corpus assez similaires, et contiennent de la parole semi-préparée. CHIL est un corpus de séminaires, où le discours est donc semi préparé, et EPPS anglais un corpus du parlement européen, où le discours est préparé, alors que le corpus AMI contient des enregistrements de réunions où la parole est spontanée et surtout se trouvent de nombreuses interruptions au sein d'un message.

Enfin, nous voulions vérifier si la manière dont les questions sont créées avait un impact. Nous avons donc appliqué cette mesure sur les corpus des éditions 2008 et 2009 de Broadcast News et EPPS anglais et espagnol. Les résultats de cette étude sont présentés dans le tableau 1. On peut observer de grandes différences entre les résultats obtenus sur l'édition 2008 et les résultats obtenus sur 2009. Cependant, cet impact est peu prévisible. On observe que la distance moyenne a quasiment doublé pour Broadcast News et EPPS anglais. Cependant, la distance de EPPS espagnol, qui était bien plus élevée dans l'édition 2008, a énormément chuté. De plus, les écarts types obtenus sur chaque corpus sont bien plus élevés que les moyennes observées. On en déduit que quelque soit la valeur moyenne d'un corpus, il existe à chaque fois un certain nombre de questions avec une distance globale très élevée. Ces observations sont corrélées par la distribution des valeurs des figures 1 et 2.

4.3. Impact possible sur les systèmes de questions-réponses

Les distances entre éléments de la question et réponse attendue jouent un rôle dans la segmentation en passage des documents traités par les systèmes QR. La segmentation des documents en passage est un aspect important du fonctionnement général des systèmes QR. L'objectif est de simplifier l'extraction de la réponse. Selon le système, un passage peut être une simple phrase, ou bien plusieurs lignes. Dans le cas de l'oral, on construit généralement des blocs proches de phrases classiques. INAOE [6] découpe les documents en passages de 24 mots. 12 mots des passages adjacents sont englobés. UPC [4] définie les passages comme étant des segments où deux mots-clés consécutifs sont séparés par au plus w mots. Dans le système du LIMSI [2], les documents sont sélectionnés grâce à un Descripteur De Recherche (DDR) contenant les éléments de la question supposés critiques pour trouver la bonne réponse. Les passages sont ensuite extraits du document par rapport à une fenêtre fixée préalablement par tuning sur les données de développement. Dans ces trois méthodes, la segmentation des documents nécessitent donc que les éléments de la question soient relativement proches entre eux, ou que les phrases (ou blocs dans le cas de l'oral) aient une taille définie ou encore qu'il y ait une ressemblance entre les données de développement (celles de l'année précédente) et les données de test. Or les questions des éditions 2008 et 2009 ont été créées différemment. Par ailleurs, les écarts types élevés obtenus sur chaque corpus montrent que la répartition des éléments des questions par rapport aux réponse est problématique pour ces méthodes de segmentation. Pour ces trois méthodes, la taille des passages extraits est déterminée selon des paramètres fixés préalablement. L'écart type de chaque corpus étant bien plus élevé que la mesure globale, les passages contenant la bonne réponse ne seront pas toujours extraits.

5. Conclusions et perspectives

Nous avons présenté des mesures pour tenter de caractériser des corpus de questions-réponses. Notre objectif était de vérifier, par ces mesures, si une typologie *a priori* des documents suffisait pour caractériser ces données dans un cadre de questions-réponses. Cela ne semble pas être le cas. On constate en particulier que

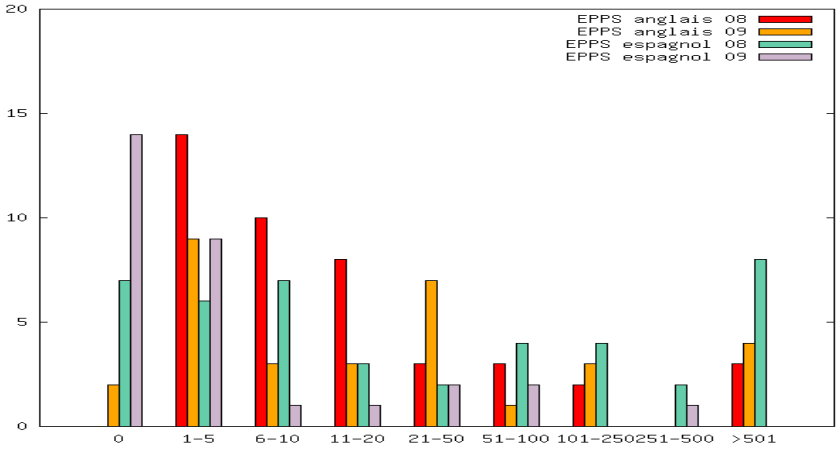


Fig. 1: Distribution des distances moyennes - Comparaison 0809 EPPS

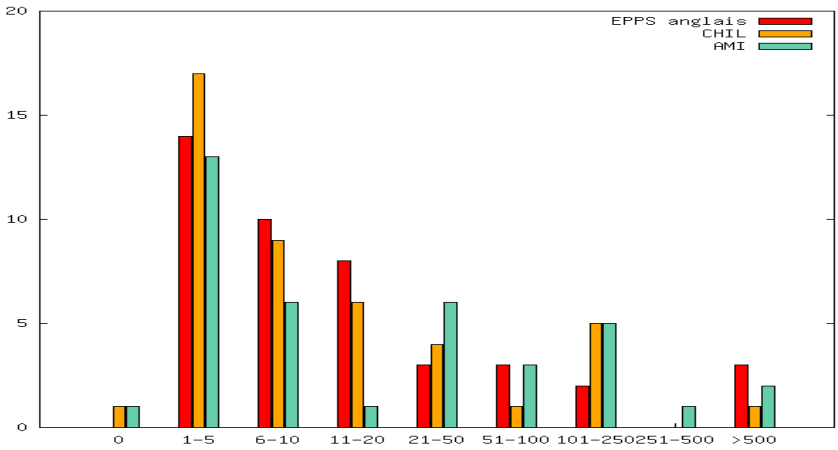


Fig. 2: Distribution des distances moyennes - Comparaison 08 EPPS anglais/CHIL/AMI

l'impact de la langue semble plus important que la proximité entre deux collections (EPPS anglais et espagnol plus éloignés que EPPS anglais et CHIL par exemple). Il est évident que ces mesures demandent à être développées. En particulier, il serait intéressant d'évaluer la présence d'expressions référentielles (anaphores par exemple).

6. Remerciements

Ce travail a été partiellement financé par OSEO par le biais du programme Quaero.

Références

[1] AMI. The ami meeting corpus. http ://www.amiproject.org, 2005.

[2] Guillaume Bernard, Sophie Rosset, Olivier Galibert, Eric Bilinski, and Gilles Adda. The limsi participation to the qast 2009 track. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, October 2009.

[3] CHIL. The european project chil. http ://chil.server.de, 2007.

[4] P. Comas and J. Turmo. Robust question answering for speech transcripts : Upc experience in qast 2009. In *Working Notes of CLEF 2009 Workshop*, Corfu, Greece, October 2009.

[5] S. Galliano, E. Geoffrois, G. Gravier, J.F. Bonastre, D. Mostefa, and K. Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of LREC'06*, Genoa, 2006.

[6] A. Reyes-Barragan, L. Villasenor-Pineda, and M. Montez y Gomez. Inaoe at qast 2009 : Evaluating the usefulness of a phonetic codification of transcriptions. In *Working Notes of CLEF 2009 Workshop*, Corfu, Greece, October 2009.

[7] TC-Star. http ://www.tc-star.org, 2004-2008.

[8] Jordi Turmo, Pere Comas, Sophie Rosset, Olivier Galibert, Nicolas Moreau, Djamel Mostefa, Paolo Rosso, and Davide Buscaldi. Overview of qast 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, October 2009.

[9] Jordi Turmo, Pere Comas, Sophie Rosset, Lori Lamel, Nicolas Moreau, and Djamel Mostefa. Overview of qast 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September 2008.

[10] E. M. Voorhees and L. P. Buckland. The fifteenth text retrieval conference proceedings (trec 2006). In Voorhees and Buckland, editors, *NIST Special Publication 500-272*, 2006.

Exploitation des segmentations en locuteurs pour la détection de rôle : application à des émissions radiodiffusées

Benjamin Bigot, Isabelle Ferrané et Julien Pinquier

IRIT - Université de Toulouse
118, route de Narbonne - 31062 Toulouse Cedex 9 - France
{bigot, ferrane, pinquier}@irit.fr

ABSTRACT

In this paper, we consider that detecting speaker role to enrich interaction sequences between speakers is a first step to find high-level events like interviews or debates. We assume that speaker roles can emerge from parameters extracted from speaker segmentations without taking any prior information into account. Each speaker is then represented by a feature vector carrying temporal, signal and prosodic information. We study how methods for dimensionality reduction and classification can help to recognize speaker roles. This method is applied to the ESTER2 corpus and our best result reaches about 72% of well recognized roles (that corresponds to nearly 79% of speech duration).
Keywords: speaker segmentation, speaker role detection, temporal and prosodic features.

1. Introduction

Trouver dans un contenu audiovisuel des événements de haut-niveau comme les séquences d'interviews, de débats, de reportages, ou de documentaires nécessite de trouver un moyen de franchir le fossé entre l'extraction de données bas-niveau et la détection automatique de descripteurs ou de concepts haut-niveau. Depuis quelques années cette problématique fait l'objet de travaux de recherche, notamment en recherche de documents sur le web [15], en création de résumé et en recherche d'information dans les vidéos de sport [11] mais aussi en analyse de contenus audio [3]. Dans ce contexte, nous nous intéressons tout particulièrement à la détection de rôle des locuteurs en vue d'enrichir et de caractériser les séquences d'interaction entre intervenants. Dans cet article, nous présentons nos motivations et quelques travaux de l'état de l'art relatifs à la reconnaissance de rôles. Nous décrivons ensuite la méthode générique que nous proposons, pour l'appliquer au corpus de la campagne d'évaluation ESTER2.

2. Motivations

2.1. Détection de séquences d'interaction

L'une des premières applications de la détection des zones d'interaction est la structuration de contenu. Par interaction, nous entendons ici échanges de propos entre deux locuteurs ou plus. De telles séquences correspondent souvent à des interviews ou des débats et constituent également des éléments structurant du flux audiovisuel. Détecter et caractériser de telles séquences permettent également de se focaliser sur les échanges verbaux, plus informels ou moins bien structurés car susceptibles de contenir de la parole conversationnelle et/ou spontanée. Afin d'enrichir la description de ces zones particulières, nous proposons d'ajouter des informations sur les rôles des locuteurs ou intervenants. Ces travaux trouvent un cadre applicatif par le biais du projet ANR EPAC [1] La détection de zones de parole conversationnelle peut être un moyen d'anticiper les difficultés auxquelles sont confrontés les systèmes de transcription automatique de la parole quand celle-ci est spontanée [10]. La segmentation du flux en zones d'interaction doit aussi permettre de détecter, aux frontières de ces zones, les entités nommées correspondant à l'identité de certains interlocuteurs ou des indicateurs sur les thèmes abordés. Enfin, les zones caractérisées comme des débats, pourront servir de base aux recherches relatives à l'étude des opinions.

2.2. Détection du rôle : état de l'art

Dans les documents audiovisuels, les comportements adoptés par les locuteurs dépendent souvent du rôle qui leur incombe. Une personne peut intervenir tout au long d'une séquence, courte ou longue, seule ou en interaction avec une ou plusieurs autres personnes, soit parce qu'elle anime la séquence, soit parce qu'elle y est interviewée ou invitée. L'interaction peut même être houleuse si il s'agit d'un débat animé. La reconnaissance des rôles est un domaine de recherche assez récent. En 1999, Stolcke [13] met en évidence le lien entre les changements de rôles et les changements de thèmes dans les émissions. En 2000, Barzilay [1] présente les premiers résultats d'une méthode de reconnaissance automatique des rôles. L'approche proposée utilise, entre autres, des caractéristiques lexicales des transcriptions, comme dans [4]. Le corpus de test se compose alors de 35 enregistrements d'un même programme radiophonique (17 heures). En 2006, Liu [12] utilise deux approches différentes pour identifier le rôle d'un locuteur parmi trois catégories : *« présentateur »*, *« journaliste »* et *« autre »*. La première approche est fondée sur les Modèles de Markov Cachés et la seconde sur le maximum d'entropie. Dans les deux cas, Liu utilise des modèles de langage N-gram pour modéliser les suites de mots caractéristiques des différents rôles ainsi que des tours de parole successifs. Les évaluations sont réalisées sur un corpus de 336 émissions d'information issues de différentes sources et atteignent 77% de taux de reconnaissance. Les auteurs

1. The EPAC Project. http ://epac.univ-lemans.fr/.

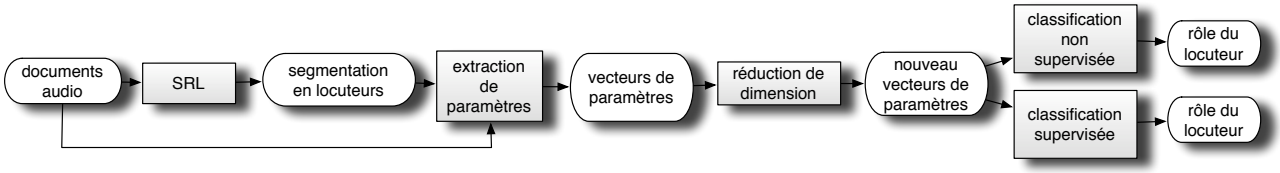


Figure 1: Système de détection de rôles.

observent que chaque méthode est plus performante pour un rôle plutôt qu'un autre. Plus récemment, Vinciarelli [14] propose deux méthodes de reconnaissance de rôles, l'une basée sur la distribution des durées des interventions des locuteurs, et l'autre sur l'analyse des réseaux sociaux dont l'objectif est de représenter les interactions entre locuteurs. Ces méthodes utilisent en entrée le résultat d'outils de segmentation et regroupement en locuteurs (SRL) et sont appliquées sur un corpus très homogène de 96 émissions d'une même source radiophonique (bulletins d'information d'une durée moyenne de 11 à 12 minutes). Les résultats obtenus atteignent 85% de temps total correctement étiqueté en terme de rôle. Favre poursuit ces travaux [7] et intègre une modélisation prédictive à base d'un modèle de langage et l'applique à des corpus plus diversifiés (*news, talk-shows*) et moins formels (*réunions*). Les performances obtenues illustrent la difficulté de détecter les rôles dans des documents moins structurés en utilisant cette approche.

2.3. Détection du rôle : contribution

Dans la plupart de ces méthodes les corpus sont homogènes. Or, traiter des flux audiovisuels, variés en terme de rôles, de durée et de structure, nécessite une méthode plus générique n'utilisant pas d'information *a priori*. Ceci permet d'agir en amont de la phase de transcription afin que les outils en charge de cette tâche puissent choisir les modèles les mieux adaptés au traitement de la parole conversationnelle. Notre méthode s'appuie sur les segmentations en locuteurs comme dans [14] et [7]. Nous extrayons d'abord un ensemble de descripteurs que nous pensons caractéristiques de chaque locuteur : la durée, la façon dont la parole est produite (plutôt lentement, de manière contrôlée, ou bien de manière spontanée, hésitante, avec de nombreux silences ou pauses...), ou l'intonation peuvent nous informer sur la nature de l'intervention. Nous avons choisi d'extraire des descripteurs à partir des segmentations temporelles, du signal audio ou de la prosodie. Chaque locuteur est alors représenté par un vecteur de paramètres. Après réduction de la dimension de ce vecteur, nous effectuons reconnaissance suivant 3 classes (rôles).

3. Méthode de détection des rôles

Les principales étapes de notre méthode sont reportées en figure 1 : suite à une SRL, différents paramètres sont extraits. Une classification en rôles s'effectue suite à une réduction de la dimensionalité.

3.1. Segmentation en locuteurs

Ce processus dit de « speaker diarization », consiste à détecter les tours de parole et à regrouper ceux appartenant au même locuteur. Généralement, les systèmes procèdent sans connaissance *a priori* sur le nombre de locuteurs présents dans le document audio. Ce processus est le plus souvent basé sur une première étape de segmentation qui consiste à partitionner le contenu audio en segments (chaque segment devant être le plus long possible et homogène au sens acoustique). La deuxième étape est un regroupement qui consiste à donner le même label aux segments contenant la parole du même locuteur. Idéalement, chaque groupe doit correspondre à un seul locuteur et inversement. Dans la pratique, les documents traités contiennent de multiples sources audio, telles que des morceaux de musique ou des jingles qui viennent parfois se superposer à la parole : le processus de « speaker diarization » est alors rendu plus difficile. Pour notre étude, nous avons utilisé les résultats fournis par la méthode développée par El Khoury [6]. La phase de segmentation est basée sur l'utilisation conjointe des critères GLR (Generalized Likelihood Ratio) et BIC (Bayesian Information Criterion) dont la mise en œuvre minimise au maximum l'ajustement des paramètres : ceci permet à cette méthode d'être robuste sur tout type de contenus audio, sans réglages spécifiques. Sur le corpus de test que nous utilisons, et qui sera décrit plus tard, cette méthode réalise la SRL avec un taux d'erreur (DER) de $11,35\%$ sur le temps total traité.

3.2. Paramétrisation

Pour chaque segmentation en locuteur nous calculons 34 paramètres. 5 paramètres sont basés sur des mesures temporelles [2]. Nous évaluons le temps de parole de chaque locuteur, son étendue, son taux d'inactivité, son nombre de segment et le rapport entre son temps de parole et la durée de l'émission où il intervient. Une autre partie des descripteurs est basée sur l'énergie du signal. En utilisant les fichiers sonores nous calculons des grandeurs caractérisant les zones de silence (leur durée, nombre, taux, énergie moyenne, minimum, maximum et variance). Nous extrayons également ces mesures sur les zones d'énergie élevée. Nous mesurons le rapport Signal sur Bruit et détectons les zones de parole téléphonique. Nous obtenons ainsi 25 paramètres et les 4 derniers sont obtenus à partir du pitch : le taux de zones voisées, la moyenne et la variance du pitch, les valeurs minimales et maximales de celui-ci. En fonction des documents traités, certains de ces paramètres peuvent être insignifiants ou fortement corrélés entre eux.

3.3. Réduction de dimensionalité

Nous utilisons deux méthodes classiques afin de réduire la dimensionnalité des vecteurs de paramètres. Nous appliquons une Analyse en Composantes Principales (ACP) [5] et nous conservons les composantes

qui représentent 95% de l'information initiale. La seconde méthode est une Analyse Factorielle Discriminante (AFD) [8]. Le nombre de dimensions est réduit à $(N-1)$ avec N le nombre de rôles.

3.4. Méthodes de classification

La classification non supervisée s'appuie sur l'hypothèse que des locuteurs jouant le même rôle partagent des paramètres similaires et peuvent être ainsi regroupés. Nous avons testé deux méthodes, l'algorithme des K-means et DBSCAN qui ne se sont pas montrés satisfaisantes compte tenu du nombre réduit de données disponibles. Les méthodes de classifications supervisées sont les Modèles de Mélanges de lois Gaussiennes (GMM), les Machines à Vecteur Support (SVM) et l'algorithme des K plus proches voisins (k-ppv).

4. Expériences

4.1. Corpus

Nous utilisons la partie DEV (environ 6 heures) et TEST (environ 7 heures) du corpus de la campagne ESTER2 [9]. Ceci correspond à 46 émissions radiophoniques, plutôt de type « information » (41 bulletins d'information, 3 émissions de société et 2 débats) enregistrées sur plusieurs sources de radios francophones et réparties suivant treize types d'émissions, offrant ainsi une certaine hétérogénéité (radios, horaires, durées, structures et nombres de locuteurs).

4.2. Annotation en rôles

L'annotation en rôles a été réalisée à partir des segmentations manuelles en locuteurs (issues de la campagne) et automatiques (issues de notre outil). Dans le but de nous comparer à l'état de l'art nous considérons uniquement 3 rôles génériques :

- Présentateurs : il s'agit du locuteur dominant de l'émission qui il anime le bulletin et introduit les autres locuteurs.
- Journalistes : ce sont des locuteurs professionnels travaillant pour la radio.
- Autres : ce sont tous les autres locuteurs : en général, des locuteurs non professionnels.

Le corpus ESTER2-DEV-ref (20 présentateurs, 149 journalistes et 117 autres) est utilisé pour les phases d'apprentissage puis de développement des méthodes de classification supervisées. L'évaluation est menée sur le corpus ESTER2-TEST (26 présentateurs, 143 journalistes et 128 autres). La reconnaissance est également menée après une ACP, ou après une AFD.

4.3. Résultats

Nous réalisons une évaluation à partir des résultats de segmentation en locuteurs manuelles et automatiques dans le but d'observer l'influence des erreurs de segmentation. Le tableau 1 présente les résultats pour les deux méthodes de réduction de dimensionalité, ainsi que pour les méthodes de classification supervisées. Les méthodes non supervisées donnent des résultats peu probants : ceci semble dû à une taille de corpus trop faible.

Table 1: Résultats de la reconnaissance de rôles.

		Accuracy(%)		
		GMM	k-ppv	SVM
TEST-ref	ACP	61,75	63,6	61,12
TEST-auto	ACP	57,64	64,04	60,01
TEST-ref	AFD	56,68	63,59	60,56
TEST-auto	AFD	53,69	63,05	58,96

La méthode des k-ppv donne les meilleurs résultats. Du fait du petit nombre d'échantillons d'apprentissage pour la classe des présentateurs (20 échantillons), la méthode GMM est réduite à une seule loi gaussienne. Ceci explique sans doute ces performances peu convaincantes. Les meilleurs scores pour le système SVM sont obtenus à partir d'un noyau gaussien. Les deux méthodes de réduction de dimensionalité donnent des résultats équivalents et l'utilisation de l'outil SRL (DER de 11,35%) ne provoque pas d'effondrement des résultats (pertes de 3 à 4% seulement).

Ensuite, nous avons distingué deux types de locuteurs présents dans le corpus : les locuteurs dit « ponctuels » et les « non ponctuels », les ponctuels ne comportant qu'un seul segment. Il n'y a pas de présentateur parmi les ponctuels et on compte 48 journalistes et 36 « autres » dans le corpus de test. Cette stratégie améliore la reconnaissance de manière significative (cf. table 2) : environ 6% avec les k-ppv et plus de 12% avec la modélisation gaussienne quand la reconnaissance est effectuée à la suite d'une AFD. Le meilleur score (70,92%) est atteint par la combinaison ACP/k-ppv.

Table 2: Résultats de la reconnaissance des rôles avec une distinction ponctuel/non ponctuel.

		Accuracy(%)		
		GMM	k-ppv	SVM
TEST-auto	ACP	60,50	70,92	64,15
TEST-auto	AFD	65,96	69.02	64,11

Les matrices de confusion (tables 3 et 4) détaillent les résultats obtenus pour la meilleure méthode de classification (k-ppv). Comme nous pouvons le voir, après une AFD, les présentateurs sont mieux détectés (100%). Tandis qu'après une ACP les journalistes et les autres sont mieux discriminés.

Une amélioration sensible des résultats peut être obtenue en appliquant une reconnaissance par une loi gaussienne pour les locuteurs ponctuels et un k-ppv pour les non ponctuels. Dans les deux cas après une ACP sur les vecteurs de paramètres, nous obtenons 71,92% de rôles bien détectés. Ce résultat correspond à 78,66% du temps traité correctement indexé en rôles.

5. Conclusion et perspectives

Dans cet article, nous présentons notre contribution en détection des rôles des locuteurs. Nous faisons l'hypothèse qu'il existe des indices sur les rôles dans des

Table 3: Matrice de confusion après une ACP.

	présentateur	journaliste	autre
présentateur	19	4	3
journaliste	4	61	29
autre	0	27	56

Table 4: Matrice de confusion après une AFD.

	présentateur	journaliste	autre
présentateur	26	0	0
journaliste	2	77	11
autre	6	45	36

descripteurs temporels et prosodiques. Des évaluations sont menées sur le corpus de la campagne ESTER2, constitué de 13 types d'émissions différentes (pour un volume de 13 heures). Notre méthode atteint 71,92% de locuteurs dont le rôle est bien reconnu ce qui représente 78,66% du temps traité bien annoté. Ces résultats, similaires à l'état de l'art, méritent d'être soulignés d'une part parce qu'ils sont obtenus à partir de segmentations automatiques et d'autre part parce que les données utilisées sont structurellement hétérogènes. De plus les locuteurs de la classe présentateur sont tous bien détectés, ce qui est plutôt encourageant dans une perspective de structuration des documents puisque ce rôle est considéré comme central que ce soit pour la classification de documents et la recherche d'information. Pour aller plus loin dans nos travaux et valider l'aspect générique de notre méthode, il sera nécessaire d'augmenter la taille du corpus ainsi que sa diversité, notamment en l'appliquant à des émissions de télévision par exemple. Cela nous conduira également à étendre notre ensemble de paramètres en prenant en compte l'évolution temporelle de certains d'entre eux dans et hors zones d'interaction et à reprendre les méthodes de classification laissées de côté.

6. Remerciements

Ces travaux se sont déroulés dans le cadre du projet ANR ANR-06-CIS6-MDCA-006.

Références

[1] Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker. The rules behind roles : Identifying speaker role in radio broadcasts. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 679–684. AAAI Press / The MIT Press, 2000.

[2] Benjamin Bigot, Isabelle Ferrané, and Zein Al Abidin Ibrahim. Towards the detection and the characterization of conversational speech zones in audiovisual documents. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 162–169. IEEE, 2008.

[3] R. Cai, L. Lu, and A. Hanjalic. Unsupervised content discovery in composite audio. In *MULTIMEDIA '05 : Proceedings of the 13th annual ACM international conference on Multimedia*, pages 628–637, 2005.

[4] Leonardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain.

[5] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition, 2000.

[6] E. El Khoury, C. Senac, and R. André-Obrecht. Speaker diarization : Towards a more robust and portable system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, Hawaii, USA*, pages 489–492, Honolulu, Hawaii, USA, April 2007. IEEE, IEEE.

[7] Sarah Favre, Alessandro Vinciarelli, and A. Dielmann. Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In *ACM International Conference on Multimedia*, Beijing, October 2009.

[8] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7 :179–188, 1936.

[9] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts. In *INTERSPEECH 2009*, pages 6–10, Brighton, UK, 2009.

[10] J-L. Gauvain, G. Adda, L. Lamel, F. Lefèvre, and H. Schwenk. Transcription de la parole conversationnelle. In *TAL*, volume 45, pages 35–47. Association pour le traitement automatique des langues, Paris, FRANCE(1993),(revue), 2004.

[11] Baoxin Li, James H. Errico, Hao Pan, and Ibrahim Sezan. Bridging the semantic gap in sports video retrieval and summarization. *Journal of Visual Communication and Image Representation*, 15(3) :393–424, March 2004.

[12] Yang Liu. Initial study on automatic identification of speaker role in broadcast news speech. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers*, pages 81–84, New York City, USA, 2006. Association for Computational Linguistics.

[13] Andreas Stolcke, Elizabeth Shriberg, Dilek Hakkani-Tür, Gökhan Tür, Ze'ev Rivlin, Andreas Stolcke Elizabeth Shriberg, Gokhan Tur, and Kemal Sönmez. Combining words and speech prosody for automatic topic segmentation. In *In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 61–64, 1999.

[14] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *Multimedia, IEEE Transactions on*, 9(6) :1215–1226, Oct. 2007.

[15] Rong Zhao and W. Grosky. Narrowing the semantic gap - improved text-based web document retrieval using visual features. *Multimedia, IEEE Transactions on*, 4(2) :189–200, 2002.

Evaluation d'une nouvelle méthode de suivi de formants sur un corpus Arabe

Imen JEMAA[1,2], Oussama REKHIS[1] ,Kais OUNI[1], Yves LAPRIE[2]

[1] Unité de Recherche Traitement du Signal, Traitement de l'Image et Reconnaissance de Formes

(99/UR/1119)

Ecole Nationale d'Ingénieurs de Tunis, BP.37, Le Belvédère 1002, Tunis, Tunisie

Imen.jemaa@loria.fr , oussamarekhis@gmail.com , kais.ouni@enit.rnu.tn

[2]Equipe Parole, LORIA-CNRS – BP 239 – 54506 Vandœuvre-lès-Nancy, France

Yves.Laprie@loria.fr

ABSTRACT

This paper develops a formant tracking technique based on Fourier ridges detection. In this method we have introduced a continuity constraint based on the computation of centre of gravity for a set of frequency formant candidates which leads to connect a frame of speech to its neighbours and thus to improve the robustness of track. The formant trajectories obtained by the algorithm proposed are compared to those of a hand edited formant Arabic database, created especially for this work, and those given by Praat with LPC data.

Keywords: Arabic database, speech processing, formantic labelling, formant tracking, continuity constraint.

1. INTRODUCTION

Etant donné que les formants sont des porteurs fondamentaux de l'information, le suivi de formant peut jouer un rôle important dans certaines disciplines du traitement automatique de la parole. Il peut être très utile dans l'identification phonétique, en particulier celles des voyelles [1] et autres sons vocaliques [2], le pilotage des synthétiseurs, la reconnaissance [1] et le codage de la parole. En raison de l'importance des résonances de l'appareil vocal, de nombreux travaux ont été consacrés à élaborer des méthodes automatiques de suivi des formants dont la plupart sont basées sur la détection des racines de LPC [3] comme estimation initiale des fréquences des formants. Les résultats de plusieurs de ces méthodes ont été utilisés dans des applications de traitement de la parole. Cependant, il y a un manque manifeste de bases de données qui sont nécessaires pour l'évaluation quantitative de ces méthodes, en particulier pour la langue arabe. D'où nous avons eu l'idée d'enregistrer et d'étiqueter en termes de formants un corpus en langue arabe standard.

Nous présentons dans ce papier notre nouvel algorithme de suivi automatique de formants basé sur la détection des crêtes de Fourier qui sont les maxima de spectrogramme. Cet algorithme utilise une contrainte de continuité en calculant le centre de gravité d'un ensemble des fréquences formantiques candidates. Ensuite, pour évaluer l'algorithme proposé on le compare à la méthode de suivi automatique de formants basée sur LPC mise en œuvre dans le logiciel Praat en utilisant la base de données étiquetée comme référence.

Ce papier est présenté comme suit. Dans la section 2, nous présentons l'algorithme de suivi de formants proposé, dans la section 3, description du corpus Arabe, dans la section 4, les différentes étapes du processus d'étiquetage manuel des formants, dans la section 5, les résultats de l'étude et l'évaluation de la méthode de suivi proposée. Enfin, nous donnons quelques perspectives dans la section 6.

2. ALGORITHME DE SUIVI DE FORMANTS BASÉ SUR LES CRÊTES DE FOURIER

Le diagramme de la Figure.1 décrit les principales étapes de l'algorithme proposé. Chaque étape du diagramme est décrite brièvement ci-dessous.

2.1. Prétraitement

Puisque nous nous sommes intéressés aux trois premiers formants, le signal est ré-échantillonné à 8 kHz afin de ne pas prendre en compte de formants candidats au-dessus de 4 kHz, et nous permettre d'utiliser une analyse d'ordre inférieur plus rapide. Ensuite, le signal de parole est préaccentué par un filtre de premier ordre pour accentuer les hautes fréquences.

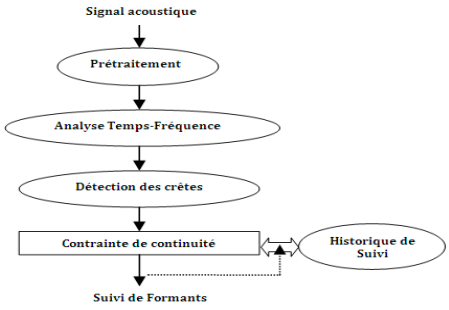


Figure 1 : Diagramme de l'algorithme de Suivi de Formants

2.2. Analyse Temps-Fréquence

Une analyse temps fréquence du signal est opérée par le module au carré de la transformée de Fourier fenêtrée pour obtenir le spectrogramme du signal. Le spectrogramme utilisée ici est un spectrogramme à large bande qui lisse l'enveloppe spectrale du signal et permet par conséquent de visualiser l'évolution temporelle des formants.

2.3. Détection des crêtes

L'algorithme de crêtes calcule les fréquences instantanées à partir des maxima locaux du spectrogramme.

Étant donné que les fréquences des formants varient lentement en fonction du temps, ils ont été assimilés aux fréquences instantanées du signal. Ainsi l'algorithme calcule toutes les fréquences instantanées du signal qui sont considérées comme des maxima du spectrogramme. Dans la suite, nous montrons comment la fréquence instantanée a été validée comme un maximum local du spectrogramme (pour preuve détaillée, voir [9]). Nous avons généré la famille de temps-fréquence des atomes, fonction fenêtrée, de la transformée de Fourier fenêtrée notée $g_{u,\xi}$ par des translations temporelles et des modulations fréquentielles d'une fenêtre g(t) réelle et symétrique de type Hamming et de taille 4 ms avec un chevauchement de 75%. Cet atome a comme fréquence centrale ξ et il est symétrique par rapport à u facteur de translation.

Il a été démontré dans [9] que la fréquence instantanée de *f* est reliée à la transformée de Fourier fenêtrée $Sf(u,\xi)$ si $\xi \geq 0$ par la relation suivante (Voir Eq.1.) :

$$Sf(u,\xi) = \frac{\sqrt{s}}{2} a(u) \exp^{i(\phi(u) - \xi(u))} \times \left[\hat{g}\left(s\left[\xi - \phi'(u)\right]\right) + \varepsilon(u,\xi) \right] \quad (1)$$

Avec *s* est une échelle appliquée sur la fenêtre de Fourier *g*, $\hat{g}$ est la transformée de Fourier de g et $\varepsilon(u,\zeta)$ est le terme correctif. Comme $|\hat{g}(\omega)|$ est maximum en $\omega = 0$, l'équation (1) montre que pour chaque *u*, le spectrogramme $|Sf(u,\xi)|^2$ est maximum en sa fréquence centrale $\xi(u) = \phi'(u)$. Les maxima $(u, \xi(u))$ forment donc les crêtes de Fourier du plan temps-fréquence. On détecte donc tous les maxima locaux du spectrogramme ; puis on procède à un seuillage selon trois largeurs de bande correspondantes aux trois premiers formants [10]. Ainsi on obtient pour chaque formant la combinaison de fréquences candidates.

2.4. Contrainte de continuité

Il est considéré que, en général, formants varient lentement en fonction du temps, ce qui conduit à imposer une contrainte de continuité dans le processus de sélection des fréquences des formants de l'ensemble des candidats. Pour les autres algorithmes basés sur les spectres LPC, la contrainte de la continuité utilisée pour chaque trajectoire des formants est la moyenne mobile des racines LPC sur sa bande de fréquences respectives [6] [11]. Dans cet algorithme, nous proposons le calcul du centre de gravité d'un ensemble de fréquences formantiques candidates, détectées par l'étape de la détection des crêtes, comme une contrainte de continuité entre les trames du signal. Puisque la détection de crêtes donne plusieurs candidats rapprochés pour un formant, l'idée est de calculer le centre de gravité de l'ensemble des candidats situés dans la bande de fréquence associée au formant en considération. La fréquence résultante f est donnée

$$\bar{f} = \frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i} \quad (2)$$

Avec f_i est la fréquence du $i^{ème}$ candidat et p_i est son énergie spectrale correspondante.

3. DESCRIPTION DU CORPUS

Pour élaborer notre corpus, nous avons utilisé une liste de phrases arabes phonétiquement équilibrées proposée par Boodraa et al. [4]. Ainsi, il couvre l'ensemble de réalisations phonétiques et phonologiques de la langue arabe standard. La plupart des phrases de cette base de données sont extraites du Coran et Hadith. Elle est constituée de 20 listes chacune constituée de 10 phrases courtes. Chaque liste est constituée de 104 CV (C: consonne et V: voyelle), c'est-à-dire 208 phonèmes [4].

Nous avons enregistré ce corpus dans une chambre sourde pour dix locuteurs tunisiens (cinq hommes et cinq femmes) dont leur âge varie entre 22 et 30 ans. Le signal est numérisé à une fréquence de 16 kHz. Ce corpus contient 2000 phrases (200 phrases prononcées par chaque locuteur) soit affirmative ou interrogative. De cette façon, la base de données présente une sélection équilibrée de locuteurs, de sexes et de phonèmes. Toutes les phrases du corpus sont riches en contextes phonétiques et sont ainsi une bonne collection de phénomènes phonétique-acoustique qui mettent en œuvre d'intéressantes trajectoires formantiques [5].

Ainsi, pour préparer notre base de données, nous avons phonétiquement annoté toutes les phrases du corpus à la main en utilisant le logiciel Winsnoori [6]. Une capture d'écran de cet outil est présentée ci-dessous dans la Figure.2 montrant l'annotation phonétique de la phrase «أَتُؤْذِيهَا بِآلامِهِمْ؟» (« 3atu3Di:ha: bi3a:la: mihim») qui signifie *(Est-ce que tu souhaites la blesser avec leurs douleurs*) prononcée par une locutrice. Une fois l'annotation terminée, le corpus a été révisé par des phonéticiens pour corriger toutes les erreurs qui ont été faites. Cette étape est très importante pour obtenir un bon suivi de formants comme référence.

4. ETIQUETAGE FORMANTIQUE MANUEL

Pour faciliter le processus d'étiquetage formantique de notre corpus, nous avons d'abord obtenu un ensemble de fréquences formantiques candidates fournies par les racines de LPC [3] à l'aide de logiciel Winsnoori [6]. Sur la base de ces valeurs candidates estimées, nous avons édité les trajectoires des formants à la main à l'aide du curseur de la souris. La Figure.2 montre un exemple d'une phrase prononcée par une locutrice illustrant le processus de l'étiquetage des formants et les résultats. Nous avons suivi et enregistré les trois premiers formants (F1, F2 et F3) toutes les 4 ms pour chaque phrase de la base de données. L'ordre de prédiction utilisé en LPC est 18 et la fenêtre d'analyse utilisée est de 16 ms. La durée temporelle de la fenêtre d'analyse spectrale est de 4 ms pour avoir un spectrogramme à large bande qui montre mieux l'évolution des trajectoires des formants. La plupart des difficultés se manifestent pour les cas, où il ya une faible énergie au niveau du spectrogramme ou lorsque les protubérances spectrales ne coïncident pas avec les prévisions des fréquences des résonances et particulièrement pour les segments consonantiques. Pour surmonter ces difficultés, nous avons prévu des valeurs nominales spécifiques aux voyelles ainsi que les consonnes [7][8]. Enfin, afin de s'assurer de l'exactitude des trajectoires des formants de chaque phrase, nous avons synthétisé le son avec les trois premiers formants utilisant le synthétiseur Klatt mis en œuvre dans Winsnoori [6] pour vérifier si le signal synthétisé correspond bien à l'original c'est-à-dire le signal d'entrée. L'évaluation a été tout d'abord subjective puisque les auteurs sont les seuls juges de la qualité des résultats lors de l'écoute du signal synthétisé ensuite on a passé à une évaluation objective en vérifiant la précision des trajectoires formantiques étiquetées avec des experts arabes en phonétique.

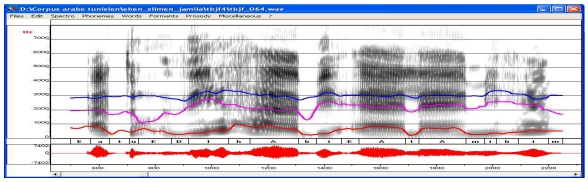


Figure 2 : Etiquetage phonétique et formantique de l'enregistrement «أَتُوذِيهَا"بِآلامِهِمْ؟» (« 3atu3Di:ha: bi3a:la:mihim ») prononcé par une locutrice.

5. ETUDE ET ÉVALUATIONS

Nous avons utilisé dans ce papier notre corpus étiqueté manuellement comme référence pour évaluer notre nouvelle méthode de suivi. Pour permettre d'abord une comparaison visuelle, les Figures 2 et 3 montrent le suivi automatique des formants (F1, F2 et F3). La Figure.2 correspond à l'étiquetage référence obtenu à la main et la Figure.3 correspond au suivi automatique de l'algorithme proposé. On peut noter que pour la plupart des segments vocaliques qui se manifestent par des parties denses en énergie au niveau du spectrogramme sont clairement identifiables, donc à ce niveau là on obtient bien un bon suivi.

Pour évaluer quantitativement la méthode de suivi automatique de Fourier, nous l'avons comparé avec la méthode LPC automatique mise en œuvre dans le logiciel Praat [11]. Nous avons utilisé notre corpus étiqueté comme référence en calculant la différence absolue moyenne (Eq.3) et l'écart-type normalisé par rapport aux valeurs de références (Eq.4) pour chaque trajectoire formantique (F1, F2 et F3). Nous avons donc étudié les résultats obtenus pour la voyelle courte / a / au sein de la syllabe CV. La Table.1 montre les résultats obtenus sur la voyelle / a / précédée d'une consonne de chaque classe phonétique et pour les trois formants (F1, F2 et F3). La collection des différentes CV a été prise des trois phrases suivantes prononcées par cinq différents locuteurs males: «عَرَفَ وَالِيًا وَقَائِد» (« ɛarafa wa:liyan wa qa:3idan ») qui signifie (Il connaissait un gouverneur et un commandant), « هِيَ هُنَا لَقَدْ آبَتْ» (« hiya Huna: laqad 3a:bat ») qui signifie (Elle est ici et elle était pieux) et « لَقَدْ كَانَ مُسَالِمًا وَقُتِلَ » ("Laqad ka:na musa:liman wa qutila" qui signifie "Il était un pacifiste et a été tué").

$$Diff = \frac{1}{N} \times \sum_{p=1}^{N} \left| F_r(p) - F_c(p) \right| \text{ Hz} \qquad (3)$$

Avec F_r est la fréquence de référence, F_c la fréquence calculée correspondant à la méthode LPC et N le nombre total de fréquences de chaque suivi de formant.

$$\sigma = \sqrt{\frac{1}{N} \sum_{p=1}^{N} \left(\frac{\left| F_r(p) - F_c(p) \right|}{F_r} \right)^2} \qquad (4)$$

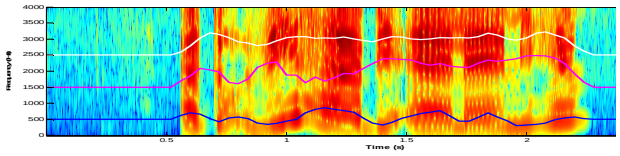


Figure 3 : La les trajectoires formantiques (F1, F2 et F3) par l'algorithme proposé de l'enregistrement (« 3atu3Di:ha: bi3a:la:mihim ») prononcé par une locutrice.

Table 1 : La moyenne des erreurs de suivi de formants mesurées pour cinq différents locuteurs males de la voyelle courte /a/.

		LPC	Fourier	LPC	Fourier	LPC	Fourier
		F1		F2		F3	
Occlusive voisée : /d /	Diff	118	52	110	82	197	125
	σ	46	14	38	19	63	32
Occlusive non voisée: /q/	Diff	67	62	84	109	113	66
	σ	18	15	27	28	31	15
Fricative voisée: /H/	Diff	52	39	87	100	92	87
	σ	9	6	18	16	18	18
Fricative non voisée: /f/	Diff	38	24	77	39	123	50
	σ	11	6	26	9	40	11
Nasale: /m/	Diff	70	90	60	172	192	110
	σ	17	21	16	49	52	27
Latérale: /l/	Diff	66	50	44	64	87	74
	σ	18	11	10	12	22	16
Trille: /r/	Diff	57	37	86	73	173	162
	σ	13	9	23	16	40	36
Semi-voyelle: /w/	Diff	75	85	85	46	154	105
	σ	19	19	32	10	42	25
Total	Diff	68	55	79	86	141	97
	σ	19	13	24	20	39	23

La comparaison des valeurs figurant dans la Table.1 montre que l'algorithme proposé présente des résultats proches de la référence et mieux que la méthode LPC e particulier dans certains cas lorsque la voyelle /a/ est précédée par une occlusive voisée et une fricative non voisée (sauf pour F1 et F2 lorsque /a : est précédée par une nasale). Pour les autres cas les deux méthodes donnent les bons résultats. Enfin nous constatons que, globalement, il n'y a pas une grande différence en termes d'erreurs entre les deux méthodes de suivi mais dans la plupart des cas l'algorithme de Fourier présente une grande différence par rapport à la référence au niveau de F3 probablement due aux hautes fréquences qui présentent de faibles énergies.

Les (Tables 2 et 3) présentent la moyenne des erreurs mesurées par la différence moyenne absolue et l'écart type normalisé par rapport aux valeurs de référence. La collection des différentes voyelles, c'est-à-dire les voyelles courtes (/ a /, / i / et / u /) et les voyelles longues (/ A /, / I / et / U /) a été prises depuis quatre phrases prononcées par cinq différents locuteurs males, respectivement cinq différentes locutrices. Les phrases des tests effectués sont: « "هِيَ هُنَا لَقَدْ أَبَتْ", " عَرَفَ وَالِيًا وَقَائِدً,« " آلنُوْذِيهَا بآلامِهِمْ؟" », citées ci-dessus et la dernière phrase est « " أَسَرُونا بِمُنْعَطَف » (« 3asaru:na: bimunɛatafin ») qui signifie (Ils nous ont capturés au niveau d'un virage).

Table 2 : La moyenne des erreurs de suivi de formants pour cinq différents locuteurs males mesurées pour chaque type de voyelle

		LPC	Fourier	LPC	Fourier	LPC	Fourier
		F1		F2		F3	
a	Diff	38	23	77	40	123	50
	σ	11	6	26	9	40	11
A	Diff	52	63	76	90	82	68
	σ	11	13	21	20	19	14
i	Diff	34	30	53	120	94	128
	σ	12	10	20	39	38	44
I	Diff	49	58	66	198	112	130
	σ	19	18	27	63	47	46
u	Diff	60	75	155	140	314	276
	σ	24	23	67	38	97	76
U	Diff	103	77	257	101	368	175
	σ	38	25	120	32	121	60
Total	Diff	56	54	114	115	182	138
	σ	19	16	47	33	60	42

Table 3 : La moyenne des erreurs de suivi de formants pour cinq différentes locutrices pour caque type de voyelle

		LPC	Fourier	LPC	Fourier	LPC	Fourier
		F1		F2		F3	
a	Diff	50	30	81	75	70	38
	σ	10	6	18	15	15	7
A	Diff	58	146	81	118	116	84
	σ	9	18	13	18	18	13
i	Diff	59	117	71	92	109	131
	σ	18	36	24	30	40	43
I	Diff	97	41	362	403	226	262
	σ	43	10	104	106	70	71
u	Diff	78	121	179	136	300	183
	σ	25	37	76	33	86	50
U	Diff	133	107	156	146	198	179
	σ	24	20	34	29	47	45
Total	Diff	79	94	155	162	170	146
	σ	21	21	45	39	46	38

La Table.2 montre que les résultats de la méthode de suivi automatique de Fourier sont bons en particulier pour les voyelles (/ a /, / A / et / i /) contrairement aux voyelles (/I /, / u / et / U /), probablement en raison de leur faible énergie. La Table.3 montre que les résultats sont bons pour les voyelles (/ a /, / i /) par rapport aux autres voyelles pour les deux méthodes de suivi. Cependant, les résultats ne sont pas très bons pour les autres voyelles.

6. CONCLUSION

Dans ce papier, le développement d'une base de données Arabe étiquetée formantiquement est présentée. En outre, nous rapportons dans cet article une utilisation exploratoire de la base de données pour évaluer quantitativement un nouvel algorithme de suivi automatique de formants basée sur la détection de crêtes de Fourier. Cet algorithme fournit un suivi bien précis des formants pour F1 et F2 et des résultats moins bons pour F3 dans certains cas. Nos futurs travaux viseront l'amélioration de cette méthode en particulier pour les formants de haute fréquence.

7. REMERCIMENTS

Ce travail est soutenu par le CMCU : Comité Mixte franco-tunisien de Coopération Universitaire (Projet de Recherche CMCU, code 07G 1112).

BIBLIOGRAPHIE

[1] Thibault, F., "Formant Trajectory Detection using Hidden Markov Models", In Proc. Of Sound Processing and Control Lab, Montreal, Canada, 2003.

[2] Ali, J. A. M., Spiegel, J. V. D. and P. Mueller, "Robust Auditory-based Processing using the Average Localized Synchrony Detection", In Proc. of IEEE Trans. Speech and Audio Proc, 2002.

[3] McCandless, S., "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans, 22:135-141, 1974

[4] Boudraa, M., Boudraa, B. and Guerin, B., "Twenty Lists of Ten Arabic Sentences for Assessment", Act of Communication ACUSTICA, 86:870-882, 2000.

[5] Deng, L.,"A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing", In Proc. of ICASSP, 2006.

[6] http://www.loria.fr/~laprie/WinSnoori/.

[7] Ghazeli, S., "Back consonants and backing coarticulation in Arabic", PhD dissertation, University of Texas, Austin, 1977.

[8] Braham, A., "An Acoustic study of temporal organization in Arabic specific to Tunisian speakers", PhD dissertation,(written in Arabic), university of Manouba, Tunis, 1997.

[9] Mallat, S.,"A Wavelet Tour of Signal Processing", Academic Press, 1999.

[10] Châari, S., Ouni, K. and Ellouze, N.,"Wavelet Ridge Track Interpretation in Terms of Formants", In Proc. of INTERSPEECH-ICSLP, 1017-1020. Pittsburgh, Pennsylvania, USA, 2006.

[11] http://www.praat.org/

L'équation de locus comme mesure de distinction sociale de *gender* en arabe koweïtien

Mohamed Embarki[1], Ammar Ahmad[2]
[1] *LASELDI EA 2281 Université de Franche-Comté, Besançon (France)*
[2] Université Paul-Valéry, Montpellier III (France)

ABSTRACT

This is an exploratory study on locus equation parameters used to measure gender distinction in Kuwaiti Arabic. Five male and five female speakers, aged 20 to 25, produced a word list disyllabic typed [CV-CV], where the first stressed syllable contained either the pharyngealised consonant /tˤ sˤ ðˤ/ or its non pharyngealised cognate /t s ð/, each followed by six vowels, the three short vowels /i a u/ and the three long ones /iː aː uː/. The results showed that locus equations were accurate to show social stratification based on gender distinction. Slopes of the equations for females were found to be less steep for non pharyngealized consonants and steep for pharyngealized cognates, compared to males. This reveals that women tend to produce a more light pharyngealisation than men.

Keywords: locus equation1, Kuwaiti Arabic2, pharyngealisation3, gender4, social stratification5.

1. INTRODUCTION

Lindblom [10] est le premier à avoir conceptualisé l'équation de locus (EL), telle qu'elle est utilisée actuellement en phonétique. EL est une régression linéaire de *F2onset* (fréquence de *F2* au début de la voyelle) sur *F2mid* (fréquence de *F2* au milieu de la voyelle) de plusieurs voyelles devant la même consonne - *F2onset* =k**F2mid*+c (où *k* et *c* sont la pente et l'ordonnée de la fonction de l'intersection *y*). Une pente relativement plate manifeste un minimum de coarticulation entre les deux segments, *F2onset* étant dans ce cas insensible à la nature de la voyelle qui suit, quelle que soit la cible fréquentielle vocalique à atteindre. Une pente relativement forte est indicatrice d'un maximum de coarticulation entre les deux segments, *F2onset* et *F2mid* ont la même fréquence, quelle que soit la cible à atteindre.

Cette régression linéaire a été utilisée dans l'indication de plusieurs contrastes phonétiques (cf. [19] pour une synthèse). Le lieu d'articulation de la consonne étant celui qui a le plus concentré l'intérêt des chercheurs. Sussman et *al*. [16, 17] ainsi que Krull [6, 7] ont montré que la valeur de la pente va décroissant pour les trois plosives : /g/ > /b/ > /d/. Ces auteurs ont trouvé que la consonne vélaire a une pente à peine plus forte que celle de la consonne labiale, l'intersection *y* est plus faible pour cette dernière ; la consonne dentale présente une valeur d'intersection *y* élevée mais une pente plus plate. Sussman et *al*. [18] ont montré la stabilité de cette hiérarchie, i.e. /g/ > /b/ > /d/, à travers le changement de style de parole, lecture *vs* discours spontané. Cette hiérarchie de lieu d'articulation révélée par les EL se maintient malgré l'intégration de traits phonologiques supplémentaires, comme l'aspiration. L'ordre vélaire > labial > dental est le même pour les plosives non aspirées et aspirées [12].

2. EQUATION DE LOCUS ET CONTRASTE DE PHARYNGALISATION EN ARABE

L'arabe standard et ses variétés dialectales possèdent le contraste de pharyngalisation. La corrélation en arabe standard, constituée des consonnes pharyngalisées /tˤ dˤ sˤ ðˤ/ d'un côté et des consonnes non pharyngalisées /t d ð s / de l'autre, peut varier d'un dialecte arabe à l'autre [2]. Parmi la série pharyngalisée, l'arabe marocain ne possède pas le /ðˤ/, l'arabe koweïtien, qui est visé par cette étude, ne possède pas le /dˤ/.

Sussman & *al*. [17] ont été les initiateurs de l'application de l'EL au contraste de pharyngalisation en arabe. Bien que leur étude n'ait concerné qu'une paire de consonnes /dˤ/ *vs* /d/ dans une variété dialectale, *i.e*. l'arabe égyptien du Caire, leurs résultats ont permis néanmoins de valider la pertinence de cette régression pour signaler le contraste de pharyngalisation. La première étude à appliquer l'EL aux consonnes pharyngalisées et non pharyngalisées de l'arabe standard [20] a montré que cette régression permettait une distinction très nette entre d'une part les consonnes non pharyngalisées /t d s ð/ et d'autre part les consonnes pharyngalisées /tˤ dˤ sˤ ðˤ/, les dernières émergent comme une classe distincte ayant les pentes les plus faibles.

Embarki et *al*. [3] ont exploité les différences de pharyngalisation signalées par les EL à des fins socio-dialectales. L'étude s'est appuyée sur la variation, exprimée par les EL, pour inférer différents degrés de constriction pharyngale, *i.e*. une pente élevée révèle une pharyngalisation faible *vs* une pente basse exprime une pharyngalisation forte. Les résultats ont révélé des variations régionales importantes pour la série /tˤ dˤ sˤ ðˤ/ entre d'une part les locuteurs jordaniens, koweïtiens et yéménites qui avaient des valeurs de pente faibles et d'autre part les locuteurs marocains qui avaient des valeurs de pente élevées, le degré de constriction pharyngale réalisé par ces derniers étant décrit comme léger.

3. GENDER ET PHARYNGALISATION

La littérature indique que la distinction homme *vs* femme, appelée *gender*, n'est pas une propriété intrinsèque de l'individu, mais plutôt relative à une manière d'être culturelle et sociale, et partant, une construction qui se développe au cours de la vie [cf. [15] pour une synthèse].

Il a été montré que le *gender* est corrélé à la manière qu'adoptent les sujets pour parler et s'exprimer [4]. Et depuis l'étude de Labov [8], la plupart des chercheurs s'accordent sur le fait que les femmes sont plus attachées que les hommes de leur communauté aux variables linguistiques les plus standard et les plus prestigieuses. Ce schéma du gender n'est pas transposable tel quel sur d'autres sociétés. Les études sur le gender dans le Monde arabe intègrent en plus d'autres paramètres, comme l'urbanisation. Les parlers urbains sont par exemple perçus plus féminisés et sophistiqués que les parlers ruraux [11]. On relève parmi les traits phonologiques caractéristiques des parlers urbains une tendance avérée à éviter la vélarisation trop marquée et à préférer la pharyngalisation légère [1].

La pharyngalisation en arabe est l'objet de variation multiple. Il ne serait donc pas surprenant de constater l'implication du *gender* dans cette variation. Aussi loin que peuvent remonter nos connaissances, aucune étude n'a encore porté spécifiquement sur le caractère prestigieux de la réalisation pharyngalisée dans le Monde arabe. En l'état actuel des choses, nous ne pouvons opter que pour une tendance générale en sociolinguistique arabe associant les parlers urbains à des parlers féminins et sophistiqués caractérisés globalement par une pharyngalisation légère.

Nous défendons l'hypothèse suivante : les hommes et les femmes se distinguent socialement par leur réalisation du trait de pharyngalisation des consonnes /tˁ sˁ ðˁ/ en arabe koweïtien, les femmes réaliseront une pharyngalisation légère, alors que les hommes réaliseront une pharyngalisation forte. Comme les EL traduisent l'influence coarticulatoire de la consonne sur la voyelle, nous nous attendons à ce que la distinction du *gender* dans l'articulation pharyngalisée se traduise par des EL différentes reflétant une hiérarchie entre les deux sexes. Les consonnes pharyngalisées de l'arabe koweïtien devraient donc se traduire par des EL élevées chez les femmes et basses chez les hommes. Afin de mieux mesurer l'amplitude du trait de pharyngalisation (de légère à forte), nous intégrerons à l'analyse les consonnes non pharyngalisées /t ð s/ correspondantes.

4. METHODOLOGIE

Dix locuteurs koweïtiens, cinq femmes et cinq hommes, ont participé à cette étude. Les dix sujets, âgés de 20 à 25 ans, ont tous l'arabe koweïtien comme langue maternelle, ils ont tous un niveau d'études supérieures homogène (niveau Master). Tous les sujets vivent depuis leur naissance au Koweït, aucun d'eux n'a vécu dans un autre pays arabe (il n'y a donc aucune influence dialectale possible sur leurs réalisations).

Le corpus est constitué de mots courants de structure bisyllabique [CV-CV] ; les consonnes /tˁ sˁ ðˁ/ et leurs correspondantes non pharyngalisées /t s ð/ apparaissant en position initiale accentuée. Chacune des six consonnes est suivie de 6 voyelles, trois brèves /i a u/ et trois longues /iː aː uː/, ce qui porte le nombre de mots à 36 (6 consonnes X 6 voyelles). Chaque mot a été inséré dans une phrase porteuse [ilju:m ingu:l … da:jma:n biʒumla:] (aujourd'hui on dit … toujours dans une phrase). La liste de mots a été répétée trois fois par chacun des dix locuteurs. Au total, notre analyse a porté sur 1080 séquences phonétiques ([(6 consonnes x 6 voyelles) x (10 locuteurs x 3 répétitions)]=1080). Les enregistrements ont été effectués dans les mêmes conditions pour tous les sujets.

Le corpus a été traité sous PRAAT, la fréquence de F2 de la voyelle a été extraite manuellement pour les deux trames nécessaires au calcul de l'EL, i.e. au début (*onset*) et au milieu de la voyelle (*midset*). Les valeurs de l'onset ont été relevées au début de la résonance vocalique (≈ 5ms du début) ; les valeurs du midset ont été relevées au milieu de la voyelle, et dans la mesure du possible sur la partie stable.

Quand il s'agit de mesures de fréquence impliquant les deux sexes, il est d'usage, à la suite de Peterson & Barney [14], de procéder à la normalisation des valeurs fréquentielles absolues en échelle auditive de Bark [13]. Nous ne procéderons pas ici à une telle conversion, car d'une part, l'équation de locus ne reflète que des transitions relatives entre la consonne et les voyelles adjacentes ; et d'autre, nous estimons, comme Johnson [5], que les attentes des auditeurs à propos des voix masculine et féminine ont un impact sur la perception de la parole. Par conséquent, un corpus qui a comme objectif de révéler des différences de *gender* ne nécessite pas de mesure de normalisation, qui elle, élimine au contraire cette différence. Car une telle conversion ne permet pas, comme le dit Labov [9], de ne neutraliser avec certitude que les effets dus aux différences de longueur du tractus vocal et de laisser intact les différences dues au *gender*. Par ailleurs, aucune mesure statistique n'a été effectuée, car les valeurs d'équation de locus sont limitées à une par consonne.

5. RESULTATS

Comme le montre le tableau n° 1, récapitulant les valeurs de pente, d'intersection *y* et de confiance (R^2) pour l'ensemble des dix locuteurs, l'équation de locus permet de distinguer clairement les consonnes pharyngalisées de leurs correspondantes non pharyngalisées. Les consonnes pharyngalisées de l'arabe koweïtien, à l'instar des autres variétés arabes [17, 20, 3], ont des valeurs de pente plus faibles que leurs correspondantes non pharyngalisées. Cependant, les valeurs de l'intersection *y* ne sont pas plus basses pour les pharyngalisées, comme c'est le cas de l'étude de Sussman et *al*. [16], et de Yeou [19].

Table 1 : équations de locus des consonnes pharyngalisées /tˁ sˁ ðˁ/ et non pharyngalisées /t s ð/ en arabe koweïtien (dix locuteurs)

C	t	tˁ	s	sˁ	ð	ðˁ
pente	0.71	0.44	0.65	0.52	0.60	0.55
Inter-y	530	630	520	545	680	472
R^2	0.73	0.69	0.77	0.75	0.81	0.82

Ces moyennes sont mieux visualisées dans la figure n° 1 ci-dessous. On y voit nettement l'apparition des valeurs faibles des consonnes pharyngalisées regroupées à gauche, et les valeurs fortes de leurs correspondantes regroupées à droite.

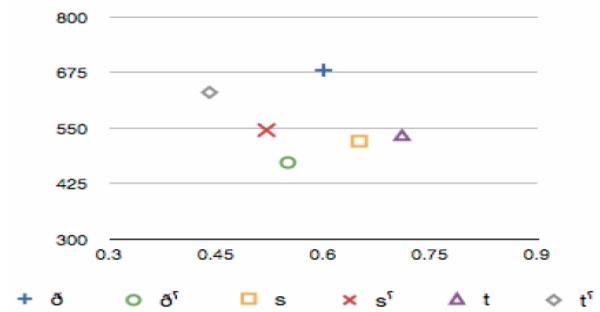


Figure n° 1 : Moyennes des valeurs de pente (ordonnées) et intersection *y* (abscisses) des consonnes pharyngalisées /tˤ sˤ ðˤ/ et non pharyngalisées /t s ð/ en arabe koweïtien (10 locuteurs).

Les EL exprimées en fonction du sexe du locuteur (cf. figure n° 2), permettent de révéler des distinctions de *gender* assez claires.

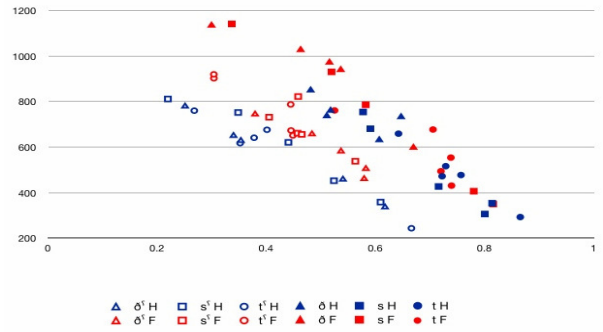


Figure n° 2 : valeurs de pente (ordonnées) et intersection *y* (abscisses) des consonnes pharyngalisées /tˤ sˤ ðˤ/ et non pharyngalisées /t s ð/ en fonction du sexe (5 hommes en bleu ; 5 femmes en rouge).

Tableau n° 2 : valeurs de pente, d'intersection *y* (inter-y) et R^2 en fonction du sexe pour les consonnes pharyngalisées /tˤ sˤ ðˤ/ et non pharyngalisées /t s ð/ en arabe koweïtien (5 hommes *vs* 5 femmes)

Hommes						
C	t	tˤ	s	sˤ	ð	ðˤ
pente	0.73	0.41	0.67	0.43	0.56	0.44
inter-y	495	580	495	587	688	542
R^2	0.90	0.73	0.89	0.73	0.85	0.77
Femmes						
C	t	tˤ	s	sˤ	ð	ðˤ
pente	0.71	0.46	0.62	0.48	0.53	0.51
inter-y	540	716	600	657	858	594
R^2	0.61	0.68	0.66	0.77	0.76	0.87

Comme le montre le tableau n° 2 (cf. supra), la moyenne des valeurs par sexe est différente. Les hommes présentent des valeurs de pente plutôt basses pour les consonnes pharyngalisées /tˤ sˤ ðˤ/ comparées à celles des femmes [0.41 contre 0.46 pour /tˤ/ (cf. figure n° 3) ; 0.43 contre 0.48 pour /sˤ/ (cf. figure n° 4) ; et 0.44 contre 0.51 pour /ðˤ/ (cf. figure n° 5)].

Le schéma est totalement inversé pour les consonnes non pharyngalisées, les hommes ayant les valeurs de pente les plus élevées, comparées à celles des femmes.

Ces résultats, quoique limités à dix locuteurs, permettent d'esquisser une certaine tendance : les valeurs de pente élevées chez les femmes pour les consonnes pharyngalisées traduisent une constriction pharyngale moins forte pour les consonnes /tˤ sˤ ðˤ/, comparées à celles des hommes. C'est probablement ce degré de pharyngalisation faible qui est associé socialement au Koweït au parler féminin, et les EL se révèlent ainsi un bon indicateur de distinction sociale.

Les valeurs de pente des consonnes non pharyngalisées sont plus élevées pour les hommes que pour les femmes. L'ensemble des valeurs de pente, consonnes pharyngalisées et non pharyngalisées, montre que les hommes présentent un contraste important entre les deux articulations, alors que les femmes présentent un contraste plus réduit, frôlant même le chevauchement entre /ðˤ/ (0.51) et /ð/ (0.53).

Le fait que les hommes occupent les deux extrémités pour les valeurs de pente (les valeurs les plus basses pour les consonnes pharyngalisées et les valeurs les plus élevées pour les consonnes non pharyngalisées) montre *a posteriori* que le calcul des EL intégrant des hommes et des femmes ne nécessite pas de conversion de données sur l'échelle de Bark, comme l'ont montré Johnson [5] et Labov [9].

Que ce soit pour les consonnes pharyngalisées ou non pharyngalisées, les valeurs de l'intersection *y* sont toujours plus élevées pour les femmes, ceci est dû manifestement aux différences de tractus vocal.

6. CONCLUSION

Les EL présentées ici révèlent des distinctions sociales de *gender*. Les résultats permettent de valider l'hypothèse générale du travail. Les valeurs de pente des consonnes pharyngalisées /tˤ sˤ ðˤ/ sont plus faibles chez les hommes que chez les femmes, tandis que celles des consonnes non pharyngalisées /t s ð/ sont plus élevées chez les hommes que chez les femmes.

Les résultats révèlent également que les hommes et les femmes ont des patrons coarticulatoires différents entre la consonne et les voyelles adjacentes. La distance séparant la pente de la consonne pharyngalisée de celle de sa correspondante non pharyngalisée est importante chez les hommes, comparée à la même distance chez les femmes. Cette différence révèle une probable tendance chez les femmes koweïtiennes à réaliser une articulation pharyngalisée moins forte que les hommes. Cette articulation, de toute vraisemblance motivée socialement, a des effets coarticulatoires moins forts sur la voyelle, et par conséquent les deux trames de la voyelle, *i.e.* l'onset et le midset, sont plus proches.

Les résultats de notre étude confirment qu'il n'est pas nécessaire de procéder à la normalisation des valeurs fréquentielles pour le calcul des EL. Sans normalisation, les différences de sexe sont bien révélées par les EL, et ce ne sont pas les femmes qui ont toujours les valeurs les plus élevées.

Si notre étude montre que les femmes ont tendance à réaliser des consonnes pharyngalisées avec une constriction plus légère, elle ne permet cependant pas de montrer si cette constriction légère est plus ou moins valorisée dans la société koweïtienne. Nos résultats, bien que prometteurs, doivent par conséquent être complétés par des enquêtes sociolinguistiques plus larges portant

sur le caractère prestigieux du degré de pharyngalisation et son utilisation spécifique par les deux sexes.

7. REFERENCES

[1] A. Boucherit & J. Lentin. Les dialectes féminins dans le Monde arabe : des dialectes minoritaires et leur évolution. E. Koskas & D. Leeman (eds.), Genre et Langage, Paris : Linx 21, 17-37.

[2] M. Embarki. Les dialectes arabes modernes : état et nouvelles perspectives pour la classification géo-sociologique. *Arabica*, 55:583-604, 2008.

[3] M. Embarki, M. Yeou, Ch. Guilleminot & S. Al Maqtari. An acoustic study of coarticulation in Modern Standard Arabic and Dialectal Arabic: pharyngealized *vs* non-pharyngealized articulation. *Proceeds. of 16th ICPhS*, Saarbrücken, Germany, 141-146, 2007.

[4] P. Foulkes & G. Docherty. The social life of phonetics and phonology. *Journ. of Phonetics*, 34: 409–438, 2006.

[5] K. Johnson. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journ. of Phonetics*, 34: 485–499, 2006.

[6] D. Krull. Acoustic properties as predictors of perceptual responses: a study of Swedish voiced stops. *PERILUS*, 7: 66-70, 1988.

[7] D. Krull. Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *PERILUS*, 10: 87-108, 1989.

[8] W. Labov, W. The social stratification of English in New York City. Washington, DC: Center for Applied Linguistics, 1966.

[9] W. Labov. A sociolinguistic perspective on sociophonetic research. *Journ. of Phonetics*, 34: 500–515, 2006.

[10] B. Lindblom. On vowel reduction. *Q. Prog. Status Rep.*, Speech Transm. Lab., Royal Inst. Tech., Stockholm 29, 1963.

[11] C. Miller. Variation and changes in Arabic urban vernaculars. M. Haak, K. Versteegh & R. Dejong (eds.), Approaches to Arabic Dialects: Collection of Articles Presented to Manfred Woidich on the Occasion of his Sixtieth Birthday, Amsterdam: Brill, 177-206.

[11] G. Modarresi, H.M. Sussman, B. Lindblom and E. Burlingame. Locus equation encoding of stop place: revisiting the voicing/VOT issue. *Journ. of Phonetics*, 33: 101-113, 2005.

[12] T. Nearey. Phonetic feature system for vowels. University of Connecticut dissertation, 1977.

[13] G. E. Peterson & H. L. Barney. Control methods used in a study of the vowels. *JASA*, 24: 175–184, 1952.

[14] S. Romaine. Variation in language and gender. J. Holmes & M. Meyerhoff (eds.), The Handbook of Language and Gender, Oxford: Blackwell, 98-117, 2003.

[15] H.M. Sussman, H.A. McCaffrey and S.A.: Mathews. An investigation of locus equations as a source of relational invariance for stop place of articulation. JASA, 90: 1309-1325, 1991.

[16] H.M. Sussman, K. Hoemeke and F. Ahmed. A cross-linguistic investigation of locus equations as a relationally invariant descriptor of place of articulation. JASA, 94: 1256-1268, 1993.

[17] H.M. Sussman, E. Dalston and S. Gumbert. The effect of speaking style on a locus equation characterization of stop place of articulation. *Phonetica*, 55: 204–225, 1998.

[18] H. M. Sussman, D. Fruchter, J. Hilbert and J. Sirosh. Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, 21: 241–299, 1998.

[19] M. Yeou. Locus equations and the degree of coarticulation of Arabic consonants. *Phonetica*, 54: 187-202, 1997.

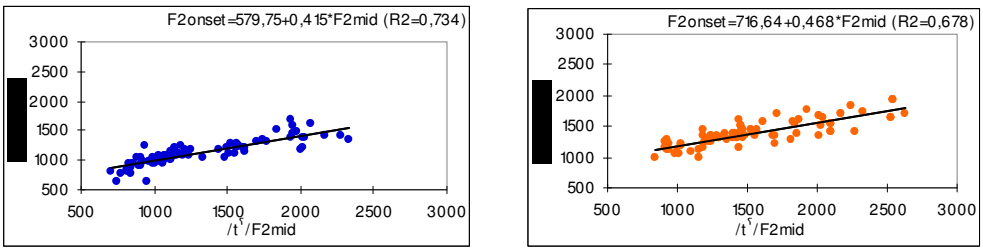


Figure 3 : équations de locus pour la consonne /tˤ/ pour les hommes (à gauche) et les femmes (à droite).

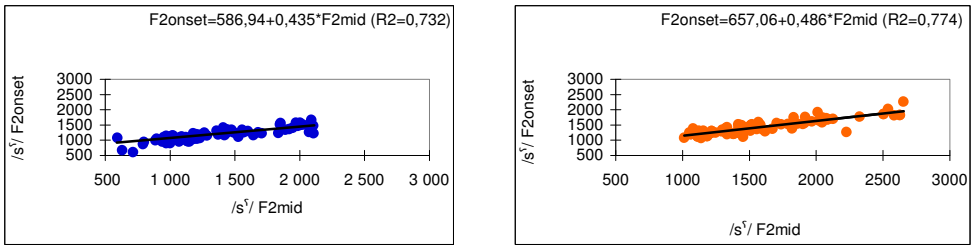


Figure 4 : équations de locus pour la consonne /sˤ/ pour les hommes (à gauche) et les femmes (à droite).

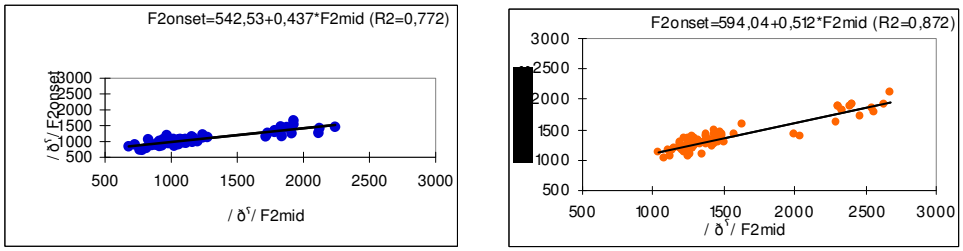


Figure 5 : équations de locus pour la consonne / ðˤ/ pour les hommes (à gauche) et les femmes (à droite).

Y a-t-il un impact de l'imitation sur la reconnaissance des mots parlés dans un accent régional non-natif?

Angèle Brunellière, Sophie Dufour et Noël Nguyen

Laboratoire Parole et Langage UMR6057 Université de Provence
5, avenue Pasteur, 13 604 Aix-en-Provence, France
angele.brunelliere@lpl-aix.fr, sophie.dufour@lpl-aix.fr, noel.nguyen@lpl-aix.fr

ABSTRACT

The plasticity of the representations for words in the mental lexicon under exposure to another speaker's voice has often been explored by observing imitative behaviours. In the study, we investigated if phonological convergence across speakers may facilitate spoken word recognition in non-native regional accent. Eighteen Northern French-speakers were exposed to CVC words recorded by a native speaker of Southern French. These speakers first realized either a shadowing task or a semantic categorization task that entailed no overt speech production and then all did a lexical decision task. The imitation of the non-native regional accent was found in shadowing but did not appear to facilitate later recognition of words.

Keywords: imitation, convergence, regional accent, spoken word recognition

1. INTRODUCTION

Lors d'une conversation entre deux interlocuteurs, il est souvent décrit qu'ils tendent à converger sur différents indices comme le débit de parole (Giles & al. [1]), l'intensité (Natale [2]) ou les constructions grammaticales (Branigan & al. [3]). Le phénomène d'imitation a été également étudié en *shadowing*, c'est-à-dire durant une tâche de répétition de mots présentés auditivement afin d'explorer la plasticité des représentations des mots en mémoire (Goldinger [4], Nye & al. [5]).

Récemment, il a été montré qu'une convergence phonétique lors d'une interaction conversationnelle puisse persister au-delà de la fin de l'interaction, suggérant que le fait d'imiter son interlocuteur puisse avoir des effets à long terme sur la manière de produire les mots (Pardo [6]). Des effets à long terme de convergence phonétique ont également été décrits en situation non conversationnelle (Delvaux & al. [7]). Ces observations suggèrent ainsi la possibilité d'ajustement dynamique des représentations des mots en mémoire (McQueen & al. [8]). Un tel ajustement lors de la production, comme en imitation, pourrait amener à une modification de la perception des mots parlés grâce à des boucles perceptivo-motrices. En effet, des liens entre perception et production ont déjà été proposés comme par exemple dans la théorie motrice (Liberman & al. [9]). D'ailleurs, récemment, Goldinger & al. [10] ont pu appuyer l'existence de liens entre perception et production en montrant que la production de mots en lecture pouvait être modifiée après différentes expositions auditives de ces mots. Dans cette étude, nous voulons tester l'inverse à savoir si une modification des prononciations de mots en production a un impact sur la reconnaissance des mots parlés. Plus exactement, nous explorons si le fait d'imiter un accent régional non-natif peut faciliter la reconnaissance des mots parlés dans cet accent. Plus particulièrement, nous examinons l'influence de l'imitation d'un locuteur méridional chez participants de type Français Standard.

Parmi les différences existantes dans l'inventaire phonémique du Français Standard et du Français dans les régions du Sud de la France, il y a le contraste /o/-/ɔ/ qui existe seulement en Français Standard en syllabe fermée (Fagyal & al. [11]). Ainsi, les mots paume /pom/ et pomme /pɔm/ sont prononcés de manière identique /pɔm/ dans les régions du Sud de la France. Les participants de type Français Standard étaient exposés au cours d'une première phase à des mots CVC produits par un locuteur méridional avec une voyelle /ɔ/ dont la moitié est habituellement produit /o/ en Français Standard. Durant cette phase, tandis que la moitié des participants devaient répéter les mots produits par le locuteur méridional, l'autre moitié réalisaient sur les mêmes items une tâche de catégorisation sémantique n'engendrant ainsi aucun mécanisme de production. Selon notre hypothèse, nous prédisons que les mots avec une voyelle /ɔ/ qui se produisent habituellement avec une voyelle /o/ en Français Standard seraient reconnus plus rapidement après la phase d'imitation comparé au groupe de sujets qui effectuaient la tâche de catégorisation sémantique.

2. MÉTHODES

2.1. Participants

Cette étude a été effectuée sur 18 participants francophones de type Français Standard (14 femmes et 4 hommes) d'un âge moyen de 25,6 ans. Tandis que 9 participants réalisaient une tâche de *shadowing*, les 9 autres effectuaient une tâche de catégorisation. Il a été

vérifié durant une tâche de lecture à haute voix que tous les participants produisaient bien distinctement le contraste /o/-/ɔ/ en utilisant des items différents de ceux présents dans la suite de l'expérience. L'ensemble des participants ne présentaient aucuns troubles auditifs ou troubles du langage. Les deux groupes de participants étaient appariés en âge et en sexe.

2.2. Stimuli

Tous les stimuli ont été produits par un locuteur francophone méridional. 200 mots monosyllabiques CVC étaient présentés aux participants durant la première phase. Au sein de cette liste de mots, 20 étaient des mots composés d'une voyelle réalisée en /o/ en Français Standard et en /ɔ/ dans les régions du Sud de la France comme rose (**Mots /o/**) alors que 20 autres mots possédaient une voyelle réalisée en /ɔ/ à la fois en Français Standard et dans les régions du Sud de la France comme robe (**Mots /ɔ/**). Il est à noter que les mots /o/ choisis n'existent pas en Français Standard avec la voyelle /ɔ/. Les mots /o/ et /ɔ/ étaient équilibrés en fréquence ($F(1,38) = 0.00002$, $p>0.2$) et en nombre de voisins phonologiques ($F(1,38) = 1.4$, $p>0.2$) à partir de la base Vocolex (Dufour et & [12]) et en durée ($F(1,38) = 0.22$, $p>0.2$). 160 autres mots servaient de remplisseurs. Durant la seconde phase, les 20 mots /o/ et /ɔ/ étaient à nouveau présentés aux participants ainsi que 160 autres mots CVC et 200 non-mots CVC. Un groupe de 20 items (**Fillers)** a été choisi parmi les 160 autres mots de manière à être équilibré avec les deux autres groupes d'items en fréquence, en nombre de voisins phonologiques et en durée. L'utilisation du groupe d'items mots /ɔ/ et Fillers nous permettait de vérifier qu'un effet de l'imitation était bien spécifique du groupe d'items /o/.

2.3. Procédure expérimentale

Phase 1

Au cours de la première phase, les 200 mots CVC étaient séparés en 4 blocs expérimentaux de 50 essais et chaque bloc était suivi d'une pause. L'intervalle de temps entre la fin d'un mot et le début d'un autre mot était de 2 s. La moitié des participants devaient répéter les mots tels qu'ils ont été produits immédiatement après leur écoute (**Groupe Test**). L'autre moitié devait indiquer le plus rapidement possible et le plus justement possible si les mots entendus étaient associés à une catégorie sémantique (**Groupe Contrôle**). Une catégorie sémantique était donnée à rechercher à chaque début de bloc expérimental (par exemple animaux).

Phase 2

Durant la seconde phase, les 200 mots et 200 non-mots étaient répartis en 2 blocs expérimentaux. Au sein de chaque bloc, un essai commençait par un point de fixation qui apparaît au centre de l'écran pendant 500 ms suivi d'un stimulus auditif. Les participants avaient pour tâche d'indiquer le plus rapidement possible et le plus justement possible en appuyant sur des touches spécifiques d'un boitier réponse si le stimulus auditif était un mot ou un non-mot en français. Leur main dominante était positionnée sur la touche qui permettait d'indiquer que le stimulus auditif était un mot. Au-delà de 1800 ms après la présentation du stimulus auditif, un nouvel essai débutait. L'intervalle de temps entre chaque essai était de 2 s.

Les deux phases débutaient par un entraînement de 10 essais et l'ensemble des stimuli étaient présentés aux participants grâce au logiciel d'expérimentation E-prime.

2.4. Analyses statistiques

Deux analyses de variance ANOVA ont été effectuées l'une par sujets (F1) et la seconde par items (F2) dans le but de tester l'influence de l'imitation (Groupe Test vs. Groupe Control) en fonction du type d'items (Mots /o/, Mots /ɔ/, Fillers) sur les taux d'erreur et sur les temps de réponse en tâche de décision lexicale. Les temps de réponse étaient mesurés en ms à partir du début du stimulus auditif.

3. Résultats

6 items ont été exclus de l'analyse des temps de réponse et des taux d'erreur car ils comportaient un taux d'erreur supérieur à 30%. L'analyse des temps de réponse a été effectuée sur les réponses correctes. Pour chaque participant, les temps de réponse supérieurs ou inférieurs à 2,5 écart-types de la moyenne des temps de réponse dans chaque condition ont été exclus des analyses. Les taux d'erreur sont présentés dans la Table 1 et les temps de réponse dans la Table 2. D'un point de vue visuel sur la Table 1, le groupe Test semble présenter de moins bonnes performances que le groupe Contrôle. Cependant, une différence significative entre le groupe Test et le groupe Contrôle apparaissait uniquement par items ($F2(1,53)=4.39$, $p<0.05$) et non par sujets ($F1(1,16)=2.22$, $p=0.15$). Aucuns autres facteurs ou interactions n'étaient significatifs.

Table 1 : Taux d'erreur en % pour chaque groupe de participants et par type d'items.

	Groupe Test	Groupe Contrôle
Mots /o/	3,9	2
Mots /ɔ/	3,9	1,1
Fillers	1,8	1,2

De même sur les temps de réponse, les participants ne montraient pas de facilité de traitement des mots /o/ après imitation (Table 2). Il apparaissait uniquement un effet tendanciel du type d'items par sujets ($F1(2,32)=2.84$, $p=0.07$; $F2(2,53)=0.7$, $p>0.5$) provoqué par des temps de réponse plus rapides pour les items mots /ɔ/ pour certains sujets.

Table 2 : Temps de réponse en ms pour chaque groupe de participants et par type d'items.

	Groupe Test	Groupe Contrôle
Mots /o/	871	896
Mots /ɔ/	850	867
Fillers	878	881

4. Discussion

D'après notre étude, l'imitation d'un accent régional non-natif ne semble pas faciliter la reconnaissance des mots parlés dans cet accent. Nos résultats pourraient s'expliquer par le fait que le groupe Test ne tend pas à imiter le locuteur méridional, c'est-à-dire ne réalisent pas la voyelle /ɔ/ pour les mots /o/. Cependant, de premières analyses acoustiques ont montré que sur les 9 participants dans le groupe test, 7 ont présenté une tendance à imiter la manière dont les mots /o/ avaient été prononcés par le locuteur méridional. L'imitation s'est produite de manière systématique dès le début de la tâche de shadowing pour 6 de ces participants, tandis que le 7ème a alterné entre [o] et [ɔ] dans la production des mots /o/. À titre d'exemple, la figure 1 contient les spectrogrammes relatifs aux mots *taupe* et *cause* tels qu'ils ont été prononcés, le premier dans le pré-test, et le second dans la tâche de shadowing, par l'un des participants. La figure montre que la voyelle transcrite *au* a été produite comme une voyelle mi-fermée dans le pré-test mais comme une voyelle mi-ouverte en shadowing.

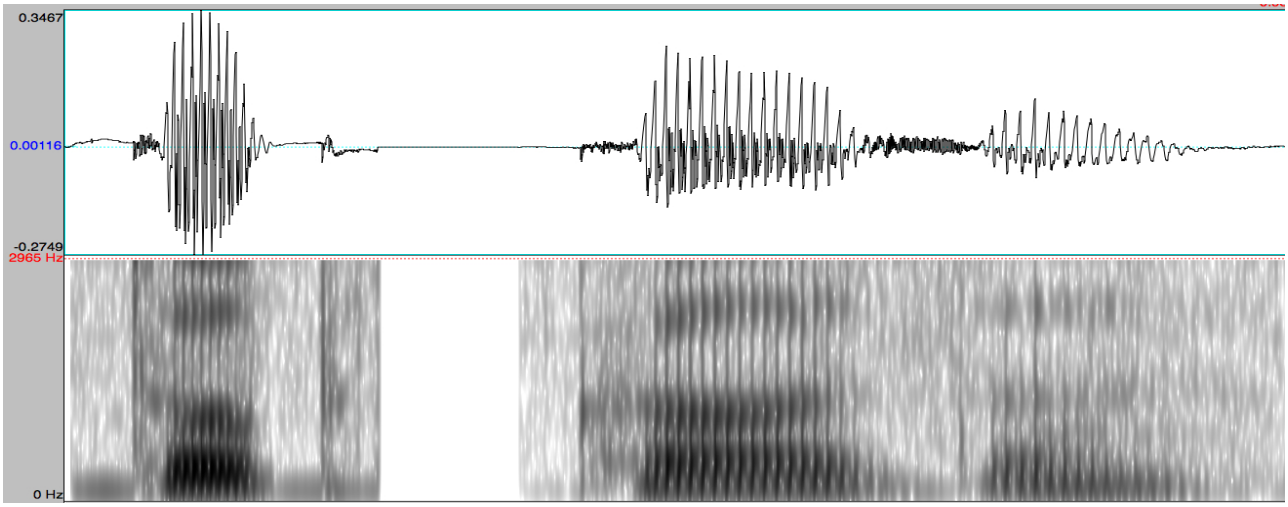


Figure 1: Formes prononcées du mot *taupe* [top] dans le pré-test (à gauche) et du mot *cause* [kɔzə] (à droite) dans la tâche de shadowing par un locuteur non-méridional.

D'autre part, il est intéressant de noter qu'en *shadowing* différé avec un délai de 3s, Goldinger [4] montrait que pour la plupart des mots répétés, l'imitation n'était pas détectée. L'absence d'imitation après un délai tardif semble indiquer l'existence de traces en mémoire épisodique des mots. En accord avec cette proposition, cet auteur a également trouvé qu'il est plus facile d'imiter un item lexical lorsque celui-ci est un mot rare. Selon un modèle en mémoire épisodique, l'influence à long terme de l'imitation dépend de la fréquence d'occurrence des traces en mémoire et augmente avec le nombre de répétitions des nouvelles traces. Par conséquent, l'absence d'influence de l'imitation dans notre étude pourrait être due à la demande d'une seule répétition pour tous les items ne permettant pas un encodage stable en mémoire.

Un autre aspect pour expliquer nos résultats serait qu'un ajustement des représentations des mots en mémoire s'est effectué suite à l'exposition de l'accent régional non-natif indépendamment de si la tâche engendrait un phénomène d'imitation. En accord avec cette hypothèse, Evans & al. [13] ont montré que des auditeurs anglais sont capables d'ajuster leurs réponses en tâche de catégorisation de voyelles en fonction de l'accent régional porté par les phrases. Cela suggère que la prononciation de variantes phonologiques n'apporte pas un avantage pour les modifications des représentations lexicales. En effet, seul un apprentissage perceptif comme lors de l'écoute des mots en tâche de catégorisation sémantique serait suffisant pour changer les représentations lexicales en mémoire. Afin de départager les différentes interprétations proposées ci-dessus, d'autres études devraient être menées par exemple en manipulant le nombre de répétions des mots /o/ produits /ɔ/ par des locuteurs méridionaux.

Pour finir, notre étude semble suggérer que des modifications dans les productions d'un locuteur n'engendrent pas de modifications dans la manière dont les mots sont perçus. De manière similaire dans l'étude d'Evans & al. [14] alors que des modifications dans les productions ont été observées chez des étudiants du Nord de l'Angleterre en contact avec des

locuteurs du Sud de l'Angleterre, aucune modification de perception en tâche de catégorisation de voyelles n'a été rapportée. De telles observations nous amènent à nous interroger sur les liens entre production et perception et semblent indiquer que les deux systèmes n'interagissent pas de façon étroite Notons qu'en accord avec cette idée, une étude récente a montré qu'un changement dans les représentations perceptives n'amène pas à des modifications de production (Kraljic & al. [15]).

5. Conclusion

Notre étude indique que la prononciation de mots dans un accent régional non-natif ne provoque pas une facilitation directe de la reconnaissance de ces mots. Ces résultats entrouvrent des interrogations sur la rapidité des phénomènes d'adaptation d'accents régionaux et sur les liens entre perception et production.

Bibliographie

[1] H. Giles, J. Coupland and N. Coupland. *Contexts of accommodations: Developments in Applied Sociolinguistics.* Cambridge University Press, Cambridge, UK, 2002.

[2] M. Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32:790-804, 1975.

[3] H.P. Branigan, M.J. Pickering and A.A. Cleland. Syntactic co-ordination in dialogue. Cognition, 75:B13-B25, 2000.

[4] S.D. Goldinger. Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105:251-279, 1998.

[5] W.P. Nye and C.A. Fowler. Shadowing latency and imitative: the effect of familiarity with the phonetic patterning of English. Journal of Phonetics, 31:63-79, 2003.

[6] J.S. Pardo. On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America, 119:2382-2393, 2006.

[7] V. Delvaux and A. Soquet. The inuence of ambient speech on adult speech productions through unintentional imitation. Phonetica, 64: 145-173, 2007.

[8] J.M. McQueen, D. Norris and A. Cutler. The dynamic nature of speech perception. Language and Speech, 49:101-112, 2006.

[9] A.M. Liberman and I.G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21:1-36,1985.

[10] S.D. Goldinger and T. Azuma. Episodic memory reflected in printed word naming. Psychonomic Bulletin and Review, 11:716-722, 2004.

[11] Z. Fagyal, D. Kibbee and F. Jenkins. French: A Linguistic Introduction. Cambridge University Press, Cambridge, UK. 2006.

[12] S. Dufour, R. Peereman, C. Pallier and M. Radeau. VOCOLEX : une base de données lexicales sur les similarités phonologiques entre les mots français. L'Année psychologique, 102:725-746, 2002.

[13] B.G. Evans and P. Iverson. Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. Journal of Acoustical Society of America, 115: 352-361, 2004..

[14] B.G. Evans and P. Iverson. Plasticity in vowel perception and production: A study of accent change in young adults. Journal of Acoustical Society of America, 121: 3814-3826, 2007.

[15] T. Kraljic, E.B. Brennan and A.G. Samuel. Accommodating variation: Dialects, idiolects, and speech processing. Cognition, 107:54-81, 2008.

Contribution à l'étude des consonnes labialisées de l'arabe marocain

Chakir Zeroual [1&2], Phil Hoole [3], and John H. Esling [4]

1. Faculté Polydisciplinaire de Taza, BP. 1223, Taza, Maroc.
2. Laboratoire de Phonétique et Phonologie, CNRS-UMR7018, Paris.
3. Institut für Phonetik, Munich. 4. Department of Linguistics, University of Victoria, Victoria, Canada.
chakirzeroual@yahoo.fr, hoole@phonetik.uni-muenchen.de, esling@uvic.ca

ABSTRACT

The relationship between labialization and velarization is examined during Moroccan Arabic (MA) "labialized" dorsals /K^w/ and labials /BB^w/ using EMA. Temporal and spatial positions of the tongue and the lower-lip are calculated based on velocity curves for pairs of words including /CK^w/, /K^{w}C/, /BB^w/ and pharyngalized /T, S/ contexts before /a, i/. Results: Geminate labials /BB^w/ are velarised; in /#K^{w}C/ and /#CK^w/ the labialization is aligned with the initial consonant release; /k/ is backed in /K^{w}Ca/ and /CK^{w}a/ but fronted in /KCa/ and /CKa/; /a/ is velarized in /CK^{w}a/ and in /K^{w}C2a/ when /C2/ is labial.

Keywords: Labialization, velarization, EMA, Arabic.

1. INTRODUCTION

L'articulation secondaire de labialisation ajoute un arrondissement des lèvres, c'est-à-dire "the action of bringing the corners of the lips towards one another so that the mouth opening is reduced" (Ladefoged [6]). Elle est transcrite par /^{w}/, qui est le symbole d'une articulation labiovélaire puisque : "in the great majority of cases where lip rounding is employed as a secondary articulation, there is also an accompanying raising of the back of the tongue, i.e. a velarization gesture. […] This double secondary articulation type is sometimes called labiovelarization" (Ladefoged et al. [7]). Même si les labiales peuvent être labialisées, ce sont les dorsales qui le sont le plus souvent ([Ladefoged et al. [7], UPSID [5]), probablement parce que l'opposition CV vs. CwV est optimale lorsque C est vélaire (Sun [11]). Pour des raisons auditives aussi, /C^w/ n'est pas attestée devant les voyelles non-arrondies, ce qui suggère que /^{w}/ est alignée avec le relâchement de /C/ (Ladefoged et al. [7]). Nous montrons que certains aspects de la labialisation en arabe marocain (AM) rejoignent ces tendances, d'autres s'en démarquent.

Puisque la labialisation en AM apparaît généralement dans des clusters initiaux #C1C2V, où C1 ou C2 est une dorsale [K] = [k g χ ʁ q] et [V] ≠ [u], plusieurs auteurs l'associent phonologiquement à ces consonnes (/#K^{w}C/ et /#CK^w/. Elle est souvent la trace d'un /u/ qui s'efface synchroniquement (AM, /kursi/ « chaise » → /k^{w}rasa/ « chaises ») ou diachroniquement (arabe standard : /ʁuraab/ « le corbeau » → AM : /ʁ^{w}rab/ « idem »). La labialisation en AM peut aussi apparaître par analogie dans des #C1C2, sans qu'un /u/ ne soit sous-jacent ni synchroniquement (AM, /qdim/ « ancien » → /q^{w}dam/ « anciens ») ni diachroniquement (AS **:** /χasarat/ « la perte » → AM : /χ^{w}sara/ « idem »).

Des réalisations phonétiques différentes sont prédites par les auteurs pour /#K^{w}C #CK^w/. Selon Heath [3], la labialisation dans ces suites est associée phonétiquement à la consonne initiale même dans /#CK^w/. Pour Mitchell [9], /C/ est aussi labialisée dans /#K^{w}C/ lorsqu'elle est une labiale /B/ = /b, m, f/ (son exemple : /χ^{w}f^{w}af/ « légers »). Elmedlaoui [1] suggère que, dans /#CK^{w}a #K^{w}Ba/, /K^w/ développe un affaissement de la partie médiane de la langue qui se propage jusqu'à /a/. Il ajoute que cet affaissement de la partie médiane de la langue durant /a/ se fait sans élévation ni abaissement de son dos ; alors qu'après une consonne emphatique, [a] est produite avec l'abaissement et de la partie médiane et du dos de la langue. La configuration articulatoire attribuée par Elmedlaoui à /a/ dans /#CK^{w}a/ et #K^{w}Ba/, ne correspond à aucune articulation secondaire canonique ; celle qu'il associe à /a/ emphatisée renvoie à la pharyngalisation.

Les labiales /b, m, f/ de l'AM (/B/) peuvent aussi s'accompagner d'une prononciation particulière, devant /a, i/, souvent considérée comme une labialisation. Elle est attribuée à l'absorption d'un /w/ sous-jacent par la labiale initiale (AM : /fwam/ « bouches » → [ff^{w}am]). /BB^w/ peuvent aussi apparaître, par analogie, même en l'absence d'un /w/ sous-jacent ([bb^{w}a] , « mon père »). Selon Mitchell [9], /BB^w/ sont « usually realized as emphatic » produites avec un « off-glide ». Pour Heath [3], /BB^w/ sont réalisées avec un « labialized release », il hésite à les caractériser comme des vélarisées ou des pharyngalisées. Selon El-medlaoui [1], /a/ possède les mêmes propriétés articulatoires après /BB^w/ et /#K^{w}B/.

Dans cette étude, nous adoptons les représentations phonologiques /#K^{w}C #CK^w BB^w/. Nous analysons les propriétés articulatoires des gestes de la lèvre inférieure et de la langue durant ces suites de consonnes de l'AM, pour mieux caractériser leur articulation secondaire et déterminer son domaine d'ancrage.

2. MÉTHODE ET MATÉRIEL

Un locuteur marocain (S1 male, 37 ans) a participé à une analyse par EMA tridimensionnelle (AG500, Carstens Medizinelektronik [4]). Grâce à des capteurs collés sur les articulateurs, nous avons enregistré (échantillonnage : 200Hz) les mouvements de la lèvre inférieure (LL), de la pointe (TT), la partie médiane (TM) et du dos de la langue (TB). Avec un programme élaboré initialement par M. Tiede (*Haskins Laboratories*), nous avons déterminé automatiquement les positions temporelles (Onset, Target, Maximum, Release, Offset) (Fig. 9) et spatiales de ces

gestes à partir de leurs courbes de vélocité (seuil : 20%). Les durées des phases de fermeture (Target-Onset), plateau (Release-Target) et d'ouverture (Offset-Release) de chaque geste ont été calculées (Fig. 2).

Des paires d'items contenant /CK^{w} K^{w}C BB^{w}/, les pharyngalisées /T, S/ et leurs correspondantes simples devant /a, i/, ont été prononcées (7 fois) par S1 dans [galha____hnaya] « il lui a dit ici ». Cette étude est limitée à ceux de la table 1. Nos données physiologiques et acoustiques ont été analysées par Matlab et traitées par Excel, et les tests statistiques (t-tests) par le logiciel R.

3. RÉSULTATS ET DISCUSSION

3.1. Données acoustiques

La fréquence du 1er pic du burst de /k^{w}/ (1063Hz) et /q^{w}/ (639Hz) est très basse comparée à /k/ (2357Hz) et /q/ (1389Hz). Dans /q^{w}lal/ et /qlal/, F2_Ons et F2_Med de /a/ sont statistiquement identiques (Fig. 1). Par contre, F2_Ons et F2_Med de /a/ dans /k^{w}bal, sk^{w}at, mm^{w}ana/ sont significativement inférieurs à leurs correspondantes dans /kbaʃ, sgat, mana/ ($p<0,001$). Ces deux résultats confirment que l'effet acoustique de /K^{w}/ dans /K^{w}Ca/ se propage à travers /b/ jusqu'à /a/, mais pas à travers /l/. Dans /Tab/, F2_Ons et F2_Med sont plus bas ($p<0,001$) et F1_Ons plus élevé ($p<0,001$) comparés à /tab/ ; ceci est attendu, puisque /T/ est pharyngalisée (voir aussi Ghazeli [2]). L'abaissement de F2_Ons dans /k^{w}bal, mm^{w}ana/ est plus important que dans /Tab/ ($p<0,001$), de même que /k^{w}bal, mm^{w}ana/ n'induisent pas d'élévation de F1. L'effet acoustique de /K^{w}b/ et /mm^{w}/ est donc différent de celui de la pharyngalisation. Durant /sk^{w}at/, et comparé à /sgat/, F2_Ons baisse mais F1_Ons s'élève légèrement; Une explication à ce comportement sera donnée en bas.

3.2. Données physiologiques

Puisqu'en générale les différences entre les suites labialisées et non labialisées par rapport aux positions spatiales de LL cessent vers /a/_Med, les mouvements de LL sont décrits jusqu'à /a/_Ons (Fig. 3-4-5-6). Parce que LL est bien réalisé dans /sk^{w}at, q^{w}lal/, ses phases d'ouverture, plateau et fermeture y sont analysées (Fig.2).

Durant /mm^{w}/, LL est légèrement (non significativement) plus avancé et plus élevé comparé à /m/ (Fig. 3), dû vraisemblablement à l'amplitude plus élevée de son mouvement induite par la gémination (Löfqvist [8]). Les positions spatiales de LL durant a_Ons et a_Med après /mm^{w}/ et /m/ sont quasi-identiques. Ces deux observations montrent que l'effet acoustique de /BB^{w}/ sur /a/ (baisse substancielle de F2, F1 non affecté) n'est pas dû à un arrondissement supplémentaire. Dans /k^{w}bal/ et /kbaʃ/ (Fig. 5), la position la plus élevée et la plus avancée de LL est enregistrée durant /b/, puisque c'est une labiale. LL est également légèrement plus élevé et surtout plus avancé dans /k^{w}b/ et a_Ons ($p<0,01$) de /k^{w}bal/ comparé à /kbaʃ/.

Durant /sgat/ et /sk^{w}at/ (Fig. 4), LL descend et recule progessivement entraîné par la rétraction de la mâchoire inférieure. LL est plus avancé et plus élevé durant /sk^{w}/ ($p<0,001$) et /a/_Ons ($p<0,01$) dans /sk^{w}at/ comparé à /sgat/ ; ces différences sont plus importantes entre /sk^{w}/ et /sg/. Dans /sk^{w}at/, la position maximale de LL est plus proche de la fin de /s/ ; son plateau commence durant la 2ème moitiée de /s/ et s'achève durant le début de /k^{w}/ (Fig. 2) : durant /sk^{w}/, la labialisation est aligné avec la fin du signal acoustique de /s/. LL est plus élevé et plus avancé durant /q^{w}l/ comparé à /ql/ ($p<0,001$) (Fig. 6); ces différences sont plus importantes entre /q^{w}/ et /q/. Le plateau de LL est pratiquement synchrone avec son relâchement acoustique (Fig. 2): dans q^{w}lal, la labialisation est donc alignée avec le relâchement de /q^{w}/.

Durant /T/ (Fig. 7) et /a/_Ons (Fig. 8) dans /Tab/, TT, TB et surtout TM sont très abaissés comparé à /tab/, ce qui est en accord avec les hypothèses d'Elmedlaoui [1]. Comparé à /mana/, /mm^{w}/ et /a/_Ons dans /mm^{w}ana/ sont associés à une position plus reculée ($p<0,01$) et surtout plus élevée ($p<0,01$) de TB, combinée à une baisse et à un recul de TT, sans aucun affaissement de TM (/m vs mm^{w}/ : $p=0,24$). Ceci ne confirme pas les prédictions d'Elmedlaoui [1] et Mitchell [9], et montre que /mm^{w}/ et /a/_Ons dans /mm^{w}ana/ sont vélarisées et non pharyngalisées, d'où F2 très bas et F1 non élevé. Pour déterminer la coordination temporelle du geste de TB durant /a/, nous avons aligné les positions temporelles et spatielles de son mouvement horizontal par rapport à /a/_Ons (Fig. 9). Ces mesures montrent que le recul maximal de TB durant /mm^{w}ana/ coïncide avec le début de /a/ (Fig. 9i), où il est également assez élevé (Fig. 9ii).

Durant /k/ dans /kbaʃ/, la portion de la langue (TMax) entre TM et TB semble très élevée (Fig. 7). Les positions spatiales de cette portion ont été déduites par (spline) interpolation en nous basant sur les orientations des capteurs. /k^{w}/ dans /k^{w}bal/ et /sk^{w}at/ sont produites sans élévation de TMax et avec un recul et un abaissement de TT et TM. Le lieu d'articulation semble donc plus avancé (au niveau de TMax) durant /k/, et plus reculé (au niveau de TB) durant /k^{w}/. Les différences spectrales au niveau du burst entre /k/ et /k^{w}/ sont dues principalement à ce recul surtout dans /sk^{w}/, où l'arrondissement est aligné avec /s/. Les positions les plus élevées et les plus avancées de TT, TM et TB sont enregistrées durant /a/_Ons de /sgat/ (Fig. 8), d'où son F2_Ons le plus élevé et son F1_Ons le plus bas. Le recul maximal de TB durant /k^{w}bal/ et /sk^{w}at/ est observé légèrement après /a/_Ons (Fig. 9i), où TB est aussi élevé mais plus reculé dans /k^{w}bal/ comparé /sk^{w}at/ (Fig. 9ii), d'où F2_Ons de /a/ plus bas dans /K^{w}bal/ comparé à /sk^{w}at/.

Durant /q^{w}/ et /q/, les positions spatiales de TB restent pratiquement identiques (Fig. 7), alors que TT et TM reculent et baissent légèrement. Durant /a/_Ons dans /q^{w}lal/ et /qlal/, les positions de TT, TM et TB sont aussi très similaires (Fig. 8). Dans /q^{w}lal/ et /qlal/, toutes les positions temporelles relevées durant le recul de TB sont presque synchrones (Fig. 9i,ii). Ces ressemblances des gestes de la langue entre /q^{w}lal/ et /qlal/ et l'alignement de

l'arrondissement avec le relâchement de /q^{w}/ expliquent les valeurs presque confondues de F1 et F2 de /a/_Ons dans ces deux items : /a/ non vélarisée dans /q^{w}lal/.

4. CONCLUSION

Contrairement à la tendance qui veut que la vélarisation s'associe souvent à la latérale [l] (UPSID, [5]), la prononciation spéciale des labiales géminées de l'AM est une vélarisation non accompagnée d'un arrondissement supplémentaire. Dans /K^{w}Ca/ et /CK^{w}a/ (K = dorsale), la labialisation est alignée phonétiquement avec le relâchement de la consonne initiale ce qui est en accord avec les prédictions de Heath [3]. /k^{w}/ est plus reculée que /k/ très probablement pour allonger encore plus la cavité antérieure et renforcer l'effet acoustique de l'arrondissement. La vélarisation des labiales est alignée avec le début de la voyelle suivante /a/ qui est elle-même vélarisée. De même que, dans /CK^{w}a/ et /K^{w}Ba/ (B = labiale), /a/ est principalement vélarisée. Ces résultats confirment partiellement la prédiction de Sproat & Fujimura [10] pour qui la vélaritaion est un geste vocalique qui est attiré par le noyau de la syllabe.

En arabe marocain, /k^{w}/ est non seulement labialisée mais aussi postériorisée ; elle pevent être caractérisée, en accord avec Ladefoged & Maddieson [7], comme labiovélarisée. Par contre, les labiales géminées de l'AM semblent, être simplement vélarisées.

Table 1 : Liste des items retenus pour l'expérience EMA avec leurs sens. N. : nom, V. : verbe, A. : adjectif.

Contexte non labialisé et non pharyngalisé	Contexte labialisé ou pharyngalisé
/kbaʃ/ : « N., moutons »	/k^{w}bal/ : « N., maïs »
/tab/ : « V., se repentir »	/Tab/ : « V., cuire »
/qlal/ : « V., devenir rares »	/q^{w}lal/ : «A., rares »
/sgat/ : « V., puiser de l'eau »	/sk^{w}at/ : « N., silence »
/mana/ : « pas moi »	/mm^{w}ana/ : « N., notre mère »

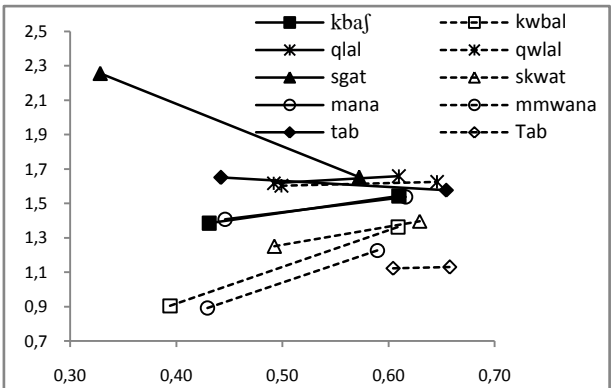


Figure 1 : Moyenne (7 répétitions) des valeurs de F1 (axe-x : kHz) et F2 (axe-y : kHz) au début (Ons) et au milieu (Med) de /a/ (1ère voyelle) des items de la table 1.

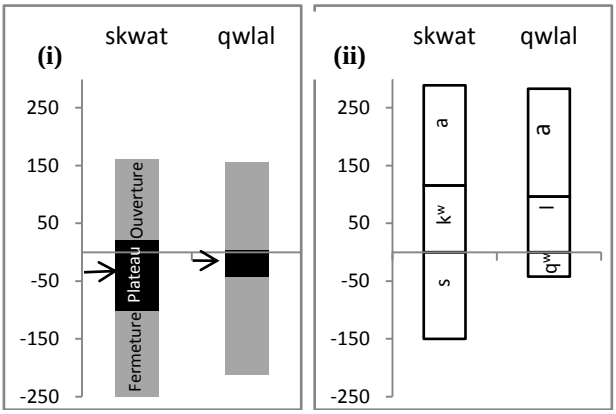


Figure 2 : Durée (ms) de la fermeture, plateau et l'ouverture, du geste de LL et la position de son amplitude maximale (flèche) durant sk^{w}, q^{w}l (i). Durée des trois 1ers segments à partir du début de /s/ et du relâchement de /q/ (ii). Ces graphiques sont alignés avec l'onset acoustique de /C2/.

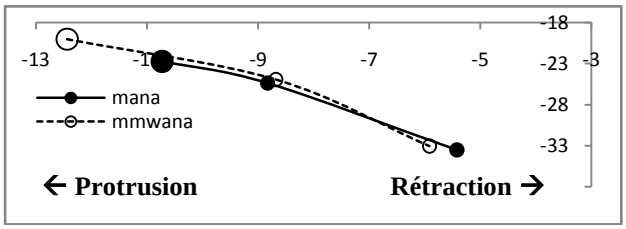


Figure 3 : Positions (en mm.) horizontale (axe-x) et verticales (axe-y) de LL durant successivement C1_Med (symbole agrandi), /a/_Ons et /a/_Med dans /mana/ et /mm^{w}ana/ (7 répétitions).

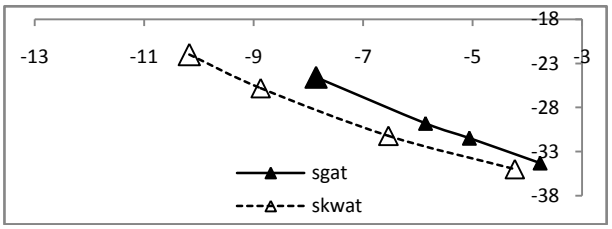


Figure 4 : Positions horizontale et verticales (en mm.) de LL durant C1_med (symbole agrandi), C2_Rel, /a/_Ons et /a/_Med dans /sgat/ et /sk^{w}at/

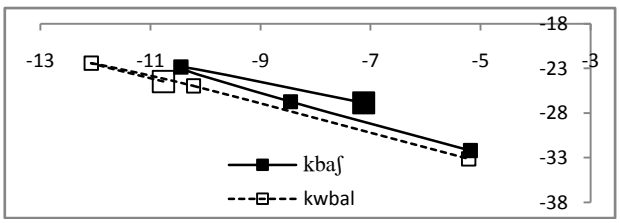


Figure 5 : Positions horizontale et verticales de LL durant C1_Rel (symbole agrandi), C2_Med, /a/_Ons et /a/_Med dans /kbaʃ/ et /k^{w}bal/.

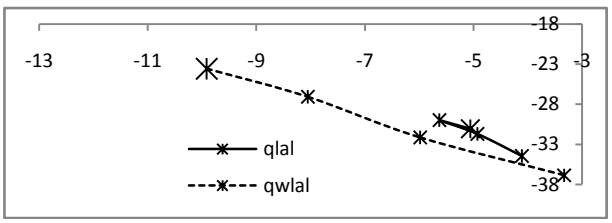


Figure 6 : Positions horizontale et verticales de LL durant C1_Rel (symbole agrandi), C2_Med., /a/_Ons et /a/_Med dans /qlal/ et /q^{w}lal/.

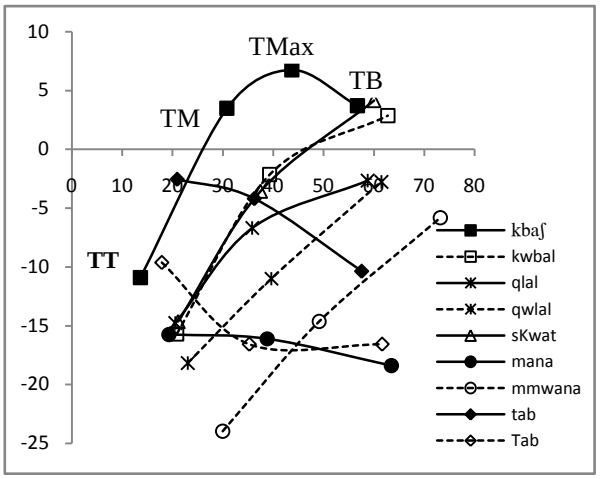


Figure 7 : Positions verticales (axe-y, en mm.) et horizontales (axe-x) de TT, TM et TB pendant le relâchement de /t T K K^w mm^w/ et au milieu de /m/ dans des items (Table1). TMax : position maximale déterminée par (spline) interpolation.

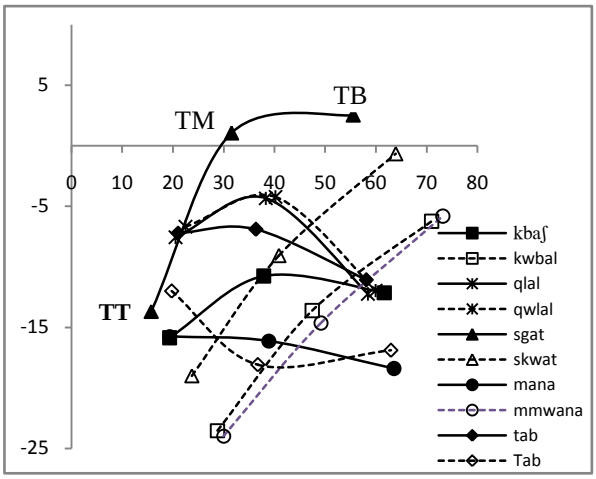


Figure 8 : Positions verticales (axe-y, en mm.) et horizontales (axe-x, en mm.) de TT, TM et TB durant /a/_Ons produite dans les items de la table 1.

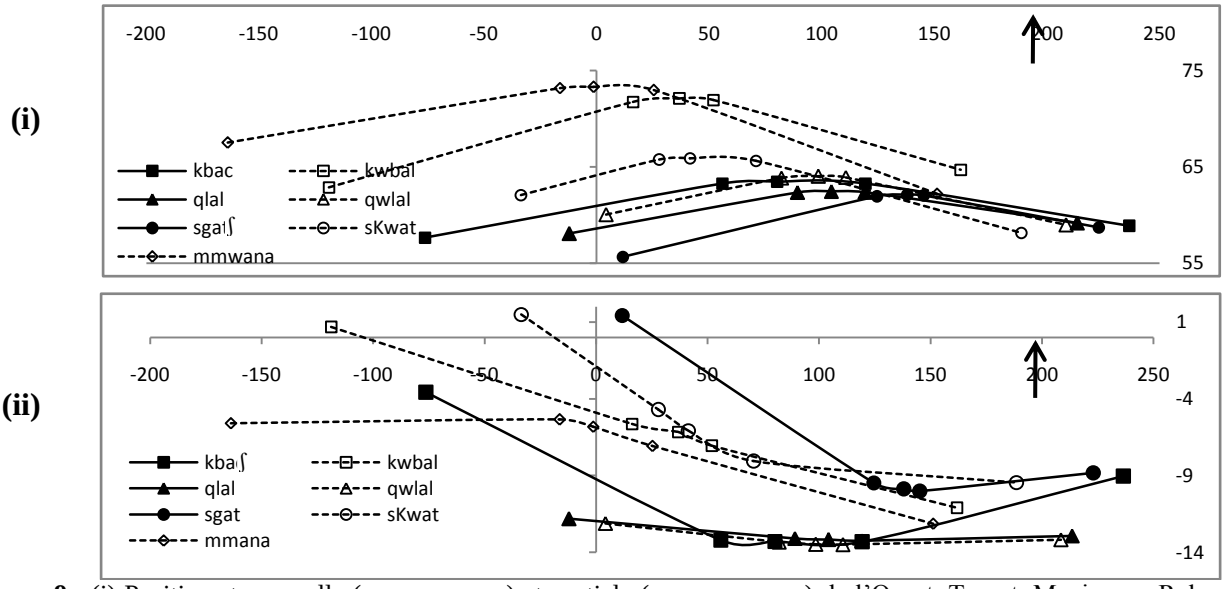


Figure 9 : (i) Positions temporelle (axe-x, en ms.) et spatiale (axe-y, en mm.) de l'Onset, Target, Maximum, Release et Offset du mouvement de recul horizontale de TB alignées avec /a/_Ons et relevées dans plusieurs items. (ii) Valeurs simultanées des positions verticales de TB. Les deux flèches correspondent à la durée moyenne (en ms.) de /a/.

BIBLIOGRAPHIE

[1] M. Elmedlaoui. *Aspects des représentations phonologiques dans certaines langues chamito-sémitiques*. Série: Thèses et mémoires, Faculté des Lettres et des Sciences Humaines, Rabat, 1995.

[2] S. Ghazeli. *Back Consonants and Backing Articulation in Arabic*. Ph.D. dissertation, University of Texas, Austin, 1977.

[3] J. Heath. *Ablaut and Ambiguity: Phonology of a Moroccan Arabic Dialect*. Albany, NY : State University of New York Press, 1987

[4] P. Hoole. Methodological considerations in the use of electromagnetic articulography in phonetic research. In *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, vol. 31: 43-64, 1993.

[5] http://web.phonetik.uni-frankfurt.de/upsid_find.html

[6] P. Ladefoged. *A course in phonetics*. (4th edition). Fort Worth: Harcourt College Publishers, 2001.

[7] P. Ladefoged. & I. Maddieson. *The Sounds of the World's Languages*. Blackwell Publishing, Cambridge USA & Oxford, UK, 1996.

[8] A. Löfqvist.. Lip kinematics in long and short stop and fricative consonants. J. Acoust. Soc. Am. 117(2): 858-878, 2005.

[9] T.F. Mitchell. *Pronouncing Arabic*. Clarendon Press, Oxford, 1993.

[10] R. Sproat and O. Fujimura. Allophonic variation in English /l/ and its implications for Phonetic implementation. Jouranl of Phonetics 21: 291-311.

[11] Y. Sun. "Consonant-labiovelar glide combinations in Spanish and Korean". Proceedings of the XVIth *International Congress of Phonetic Sciences*, Saarbrücken. 1736-1776, 2007.

Variabilité(s) et invariance dans la production des tons en thaï

Kanittarat Boottawong, Véronique Delvaux, Kathy Huet, Myriam Piccaluga, Bernard Harmegnies

Service Métrologie et Sciences du langage
Université de Mons, 20, place du Parc, B-7000 Mons, Belgique
Kanittarat.Boottawong@umons.ac.be

ABSTRACT

This study addresses invariance and variability in the production of the 5 Thai tones based on acoustic data. The F0 curves globally confirm those obtained previously in the literature, although from a larger corpus and number of speakers. The ratio of inter-tones to intra-tones variability was computed using F de Snedecor. Results show that F is maximal in the median portion of long vowels and the last portion of short vowels, which suggests that the mora might be the tone bearing unit (TBU) in standard Thai.

Keywords: thai, tone, tone production, mora, variability.

1. INTRODUCTION

Dans certaines langues asiatiques et africaines, on appelle « ton » l'une des propriétés distinctives capables de différencier des unités du lexique. Le ton est d'ordinaire considéré comme résultant de variations inter- (tons statiques) et/ou intra- (tons dynamiques) segmentales de la hauteur organisées de manière variable dans les langues tonales. Il est dès lors couramment lié aux variations du « pitch », le terme étant soit pris dans son acception perceptuelle, soit considéré comme synonyme de « fréquence fondamentale ». L'équation « ton=F0 » est cependant extrêmement hasardeuse et les questions liées aux processus perceptuels impliqués, ainsi qu'à leur éventuelle liaison avec d'autres éléments acoustiques que la F0 de l'unité supportant le ton, sont loin d'être triviales. L'expérience rapportée ici constitue la première étape d'une étude plus large de la problématique.

La langue investiguée est le thaï, qui présente l'avantage non seulement de comporter un nombre important de tons (5), mais également parmi ceux-ci, des représentants des deux catégories de tons (statiques et dynamiques). La variété standard, qui fait l'objet de cette étude, est couramment utilisée dans le centre de la Thaïlande. Il s'agit de la langue administrative enseignée dans toutes les écoles dès la maternelle. Elle comporte 21 phonèmes consonantiques /p, p^h, b, t, t^h, d, k, k^h, ʔ, m, n, ŋ, f, s, h, tɕ, $tɕ^h$, r, j, w, l/ et 21 phonèmes vocaliques : neuf voyelles simples longues /i, e, ɛ, a, ɤ, ɯ, u, o, ɔ/ et les neuf brèves correspondantes, ainsi que trois diphtongues, toujours longues /ia, ɯa, ua/ [13].

On distingue cinq tons: deux tons dynamiques, le ton montant et le ton descendant, et trois tons statiques, le ton bas, le ton moyen et le ton haut [1, 2, 3]. Dès 1962, Abramson [1] a proposé des profils de F0 pour les cinq tons, provenant d'une étude ne reposant que sur un seul locuteur et basé sur un corpus qui comprend 88 mots de type $C_1V(:)$ (C_2), avec une grande variété de C_1 et de C_2. Les travaux suivants d'Abramson sur la perception des tons en thaï [2,3] se basent sur ses données de production de [1].

De même, c'est en prenant comme référence l'allure des courbes de [1] qu'un grand nombre de chercheurs ont investigué différentes variables pouvant influencer la production des locuteurs natifs. En termes de variables inter-sujets, l'effet du facteur d'âge et, dans une moindre mesure, du facteur genre ont été mis en évidence par des études empiriques [4,5] ou par méta-analyse [6]. Les variables proprement linguistiques ont également été étudiés: l'environnement phonologique, et plus généralement les phénomènes de coarticulation [7,8], mais aussi l'effet du débit de parole [9], ainsi que l'effet de la durée de la voyelle (pour des monosyllabes isolés ou insérés dans des phrases, et pour des séquences bisyllabiques) [10].

Au plan méthodologique, toutes ces études ont suivi, dans les grandes lignes, l'étude pionnière d'Abramson [1]. En conséquence, elles ne s'intéressent qu'aux *moyennes* de F0 des productions et considèrent que la syllabe est l'unité porteuse de ton. Aucune d'entre elles n'a proposé de critères spécifiques permettant la distinction entre les différents types de tons, autre que l'allure générale de la courbe de F0 sur toute la syllabe. En 2006, Morén et Zsiga ont soulevé ce problème des critères de distinction entre les différents tons, et remarqué que les observations phonétiques montrent des courbes de F0 qui ne correspondent pas strictement à leurs étiquettes de ton « descendant », « bas », etc. De plus, la distinction entre les tons statiques et les tons dynamiques paraît peu opératoire parce que les tons statiques aussi bien que les tons dynamiques présentent une évolution temporelle de la F0. En fait, seul le ton moyen serait approximativement statique du point de vue phonétique [11:118]. Suite à ce constat, Zsiga et Nitisaroj [12] ont mené une étude de perception partant du principe que ces incohérences seraient liées au fait que l'unité porteuse du ton (TBU) ne serait pas la

syllabe mais bien la more. Cette hypothèse permettrait de délimiter un domaine plus restreint pour la réalisation phonétique du ton, plus approprié à l'établissement de critères de distinction entre les différents tons. Des limitations en termes de traitements statistiques ne permettent cependant pas de statuer sur l'hypothèse.

Ainsi, malgré l'abondance des travaux sur les tons du thaï, la littérature manque de paramètres précis, quantifiés et en nombre restreint pour distinguer entre tous les types de tons en thaï, sans doute parce que le domaine temporel d'analyse est large, à savoir la syllabe entière. Par ailleurs, dans ces études, le nombre de locuteurs [1] et/ou le corpus [4,5] sont souvent restreints. En fait, la plupart des études sur les tons en thaï ne s'intéressent qu'aux variabilités en fonction de facteurs externes aux cinq tons proprement dits. L'invariance, au sens de ce qui est commun, spécifique, voire suffisant, à la réalisation phonétique de chaque type de ton pour un grand nombre de locuteurs et de contextes phonétiques, ne semble pas constituer le principal moteur de la recherche dans le domaine.

La présente étude vise à fournir une description précise et quantifiée et ainsi à participer à la recherche de l'invariance dans la production des tons en thaï, avec pour objectif plus général de progresser dans notre connaissance sur le fonctionnement des tons chez les natifs du thaï. A partir des données issues de la production, notre objectif est d'aboutir à des hypothèses à tester ensuite concernant le traitement des tons en perception. En termes de méthodologie, nous proposons de découvrir l'invariance en mettant en jeu à la fois une grande diversité (en comparaison des autres travaux) au niveau inter-individuel (10 locuteurs), et une grande variété au niveau des paramètres linguistiques présidant à la constitution du corpus, afin de faire émerger de cette variabilité potentielle les caractéristiques communes à chaque type de ton. Par ailleurs, à la différence de nos prédécesseurs, nous nous intéressons à la variabilité des moyennes de F0, avec l'hypothèse qu'elle peut varier dans le temps de la production et doit être minimale là où l'information fréquentielle est la plus significative pour le contrôle du ton. Concrètement, afin de caractériser les 5 tons, nous calculons le rapport entre (i) la variabilité entre classes tonales (variances inter-tons) et (ii) la variabilité au sein de chaque classe tonale (variances intra-tons) en 11 instants durant la réalisation. Ce rapport nous permet de désigner le moment précis, au cours de l'évolution temporelle de la F0, où la discrimination des tons par la fréquence fondamentale est la plus importante en production. L'objectif ici est de progresser dans l'établissement de critères précis et quantifiés de distinction entre les tons, et de contribuer au débat sur la TBU.

2. MÉTHODOLOGIE

2.1. Caractéristiques des sujets

Notre échantillon comporte 10 locuteurs : 5 hommes et 5 femmes, âgés de 28 à 37 ans (moyenne= 32,5 σ=2,8). Ils ne font pas état de problèmes de vision, ni d'audition. Tous sont originaires de la région du centre de la Thaïlande et ont le thaï standard comme langue maternelle. Leur niveau d'études est universitaire.

2.2. Constitution du corpus

Le corpus est constitué de 108 monosyllabes de type $C_1V(:)$ (C_2), où C_1= /p b p^hf m w/, V= /aː a iaː/ et C_2= /k/. Les cinq tons sont associables à tous les types de syllabes sauf les syllabes se terminant par la consonne finale /k/ qui ne peuvent être phonotactiquement associées au ton moyen [13]. Il y a donc deux types de syllabes : syllabes ouvertes (voyelle obligatoirement longue, en l'occurrence /aː/ ou /iaː/) et syllabes fermées par /k/ (voyelle /a/ ou /aː/).

2.3. Récolte des données

Le corpus est induit via la lecture. Chaque monosyllabe est inséré dans une phrase porteuse et répété trois fois à la fin de cette phrase /bɔːkdaː […] daŋ daŋ […] […] […]/ qui veut dire : «Dis à Da […] fortement […] […] […] ». Les monosyllabes entourant le monosyllabe cible ont le ton moyen. La présentation en ordre aléatoire a été réalisée à l'aide du logiciel Microsoft Power Point. Le corpus a été enregistré par un enregistreur numérique de type DAT-TCD D7. Il y a eu au total deux séries d'enregistrements (appelées « répétitions »). Chaque enregistrement a duré un peu moins de 18 minutes. Pour l'ensemble des sujets, 8640 occurrences (2 répétitions x 4 essais en production x 108 monosyllabes x 10 sujets) ont été récoltées. Au stade actuel, notre analyse prend en compte uniquement le premier essai de chaque répétition. L'analyse effective comprend 2069 occurrences.

2.4. Traitement des données

Deux traitements nous ont permis d'objectiver l'analyse des données. Le premier est acoustique : il concerne l'évolution temporelle de la F0, mesurée sur la portion voisée de la production. A l'aide de l'algorithme « pitch analysis » de PRAAT (version 4.3.12), onze valeurs de F0 ont été extraites à intervalle temporel relatif constant (une valeur tous les 10% de la durée totale de la portion voisée). Le second est statistique : il se centre sur l'évolution temporelle du rapport de deux estimations indépendantes de la variance de la F0, l'une influencée par l'existence de différences entre classes tonales, et l'autre pas.. Pratiquement, ce rapport de variances (exprimé en F de Snedecor) a été obtenu par le biais de la routine

UNIANOVA de SPSS 14, pour laquelle l'unique facteur fixe était la classe tonale, et la variable dépendante, la fréquence fondamentale.

3. RÉSULTATS

3.1. La dynamique globale d'évolution de f_0

L'évolution au cours du temps de la moyenne de F0 (en Mels) est représentée sur la figure 1 pour chaque type de ton, toutes autres variables confondues. Comme remarqué précédemment dans la littérature [11], les allures générales des courbes de F0 ne correspondent que modérément aux images qu'évoquent les dénominations classiques des 5 tons. En particulier, les 3 tons statiques sont très similaires pour les premiers repères temporels et les tons dynamiques se caractérisent, en leur début, par une variation fréquentielle de sens opposé à celle escomptée. Seule la négligence de la première moitié des points de mesure permet de faire apparaître des allures de courbes plus conformes à ces attentes. Il y a donc lieu de s'interroger sur l'évolution de l'informativité de la F0 au cours du temps de la production.

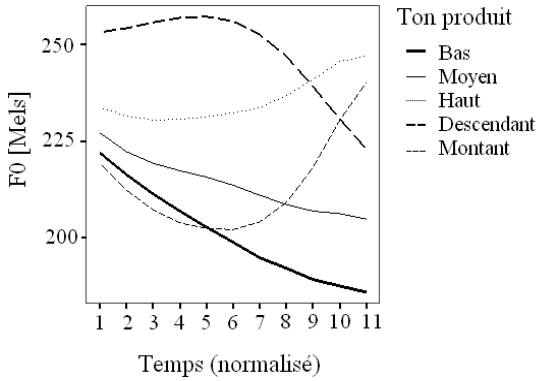


Figure 1 : Evolution temporelle (temps normalisé) de la F0 (en Mels) lors de la production des 5 tons en thaï, toutes variables confondues.

3.2. A la recherche d'instants discriminants

La figure 2 représente l'évolution temporelle du F de Snedecor (en temps normalisé) pour les quatre combinaisons Voyelle-C2 présentes dans le corpus, tous tons confondus.

Tout d'abord, les valeurs de F sont toutes élevées, et, étant donné les degrés de liberté concernés, les valeurs de p associées sont toutes significatives ($p<.001$), ce qui signifie que pour chaque repère temporel les 5 tons sont significativement différents en termes de F0. Ce qui nous intéresse ici n'est cependant pas la significativité au sens des critères usuels, mais bien la variation de cette dernière. Les degrés de liberté étant constants, nous pouvons cependant nous contenter d'observer l'évolution de la statistique F au cours du temps. On observe ainsi deux types de courbe. Pour les voyelles brèves (en haut de la figure 2), F augmente continuellement au cours du temps, ce qui signifie que la variabilité inter-tons par rapport à la variabilité intra-tons ne cesse d'augmenter tout au long de la voyelle. Pour les voyelles longues, on constate tout d'abord une augmentation de la valeur de F, pour atteindre un maximum autour du 6ème moment, suivie d'un plateau ou d'une diminution plus ou moins franche selon le type de syllabe étudié. Remarquons que le maximum de F est atteint à la fin des voyelles brèves (en moyenne: à 144 ms) alors qu'il est atteint dès le milieu des voyelles longues, qui sont approximativement deux fois plus longues que les brèves (moyenne: 326 ms) (voir table1).

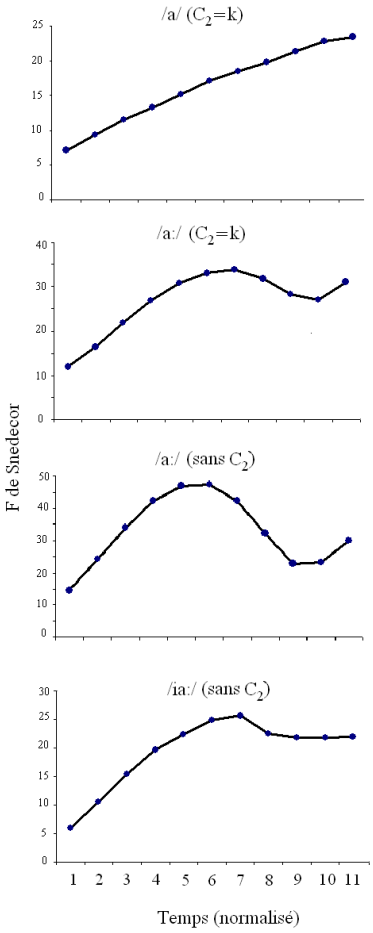


Figure 2. Evolution temporelle du F de Snedecor (en temps normalisé) pour les quatre combinaisons Voyelle-C2 présentes dans le corpus, tous tons confondus.

4. DISCUSSION

Le premier objectif de cette étude était de fournir une description quantifiée de la production des 5 tons en thaï. Globalement, nos résultats (figure 1) montrent une congruence avec les allures de courbes de F0 obtenues par nos prédécesseurs [1,5], qui sont donc ici confirmées sur un plus grand nombre de locuteurs avec un corpus linguistique étendu.

Un second objectif était d'examiner l'évolution temporelle de la variance inter-tons/intra-tons afin de déterminer le moment où la distinctivité entre les 5 tons est maximale, et de contribuer plus généralement au débat sur la TBU. Les résultats montrent que le F de Snedecor est maximal autour du 6^{e} moment des voyelles longues et du 11^{e} moment des voyelles brèves. Ceci suggère que la *portion médiane* des voyelles longues, et la *portion finale* des voyelles brèves, constituent le lieu de distinctivité maximale en termes de production. Au niveau articulatoire, ceci pourrait suggérer que cette portion est celle où les locuteurs exercent un contrôle articulatoire renforcé. Il s'agit d'un bon candidat dans notre recherche de l'invariance puisque, malgré le nombre de locuteurs ainsi que la variété des paramètres linguistiques de notre corpus, nous avons isolé ici un domaine temporel relativement restreint où la robustesse de la distinctivité est patente, et où l'on pourrait donc trouver ce qui est commun, spécifique, voire suffisant, à la réalisation phonétique de chaque type de ton.

Deux autres éléments suggèrent qu'une attention particulière doit être apportée au niveau de la portion médiane/finale. D'une part, nous observons comme [11] que c'est seulement à partir de leur deuxième tiers que les courbes de F0 correspondent phonétiquement aux étiquettes phonologiques de ton descendant, montant, moyen, haut et bas (figure 1). D'autre part, dans la grammaire prescriptive du thaï, des restrictions phonotactiques précises s'imposent lorsque la syllabe *se termine par* certaines consonnes, ce qui n'est pas le cas pour les consonnes initiales.

En ce qui concerne le débat plus large sur l'unité porteuse du ton, nous remarquons que le maximum de discriminativité est atteint approximativement au même moment, en termes de durée absolue, dans les brèves et dans les longues puisque la durée moyenne des voyelles brèves (moy=144ms, σ=31.5ms) vaut approximativement la moitié de la durée moyenne des voyelles longues (moy=326ms, σ=75.5ms). Ainsi, le maximum en termes de rapport de la variance inter-ton et intra-ton est atteint dans tous les cas *à la frontière de more*. Ce résultat est congruent avec la proposition de Morén & Zsiga [11] selon laquelle la more serait la TBU.

En perspective, nous menons actuellement des expériences de perception dont l'un des objectifs est de vérifier que la frontière de more constitue l'endroit du signal contenant les indices perceptuels les plus pertinents pour les natifs du thaï standard.

Cette recherche a bénéficié des financements de la convention ARC AUWB- 08/12-UMH 17 de la Communauté française de Belgique et du projet MCF/FRC - FRFC 2.4644.09 du Fonds National de la Recherche Scientifique belge

BIBLIOGRAPHIE

[1] A.S. Abramson. The vowels and tones of standard Thai: Acoustical measurements and experiments. Bloomington. *Folklore and Linguistics* Indiana University Research Center in Anthropology, 1-143, 1962

[2] A.S. Abramson. The tones of central Thai: Some perceptual experiments. In J.G. Harris & J.R. Chamberlain, (Eds.), Studies in Tai Linguistics in honor of William J. Geddney. *Central Institute of English Language*, Bangkok, 1-16, 1975

[3] A.S. Abramson. Static and dynamic acoustic cues in distinctive tones. *Language and Speech*, 21, 319-325, 1978.

[4] J. Gandour, S. Potisuk, S. Ponglorpisit & S. Dechongkit. Inter- and intraspeaker variability in fundamental frequency of Thai tones. *Speech Communication*, 10, 355-372, 1991.

[5] Ph. Teeranon. The change of Standard Thai high tone: an acoustic study and a perceptual experiment. *SKASE Journal of Theoretical Linguistics*, 4, 3, 1-17, 2007.

[6] P. Pittayaporn. Directionality of tone change. *Proceedings ICPHS*, Saarbrücken, 6-10 augustus, 1421-1424, 2007.

[7] A.S. Abramson. The coarticulation of tones: An acoustic study of Thai. In T. L. Thongkum, P. Kullavanijaya, V. Panupong, & K. Tingsabadh (Eds.). Studies in Tai and Mon-Khmer phonetics and phonology in honour of Eugenie J. A. Henderson. *Indigenous Languages of Thailand Research Project*, Bangkok, 1-9, 1979.

[8] J. Gandour, S. Potisuk & S. Dechongkit. Tonal coarticulation in Thai. *Journal of Phonetica*, 22, 477-492, 1994.

[9] J. Gandour, A. Tumtavitikul & N. Satthamnuwong. Effects of Speaking Rate on Thai Tones. *Phonetica*, 56, 123-134, 1999.

[10] R. Roengpitya. The Variations, Quantification, and Generalizations of Standard Thai Tones. In M.J Solé, P.S. Beddor & M. Ohala. (Eds.). *Experimental Approach to Phonology*. Oxford University Press, Baltimore, 270-301, 2007.

[11] Br. Morén & E. Zsiga. The lexical and post-lexical phonology of thai tone. *Natural Language & Linguistic Theory*, 24, 113–178, 2006.

[12] E. Zsiga & R. Nitisaroj. Tone Features, Tone Perception, and Peak Alignment in Thai. *Language and Speech*, 50, 3, 343-383, 2007.

[13] K. Naksakul. (Ed.) [กาญจนา นาคสกุล]. *บรรทัดฐานภาษาไทย เล่ม ๑* [*La base de la langue thaïe: volume 1*]. Suksapan Krasuang Suksathikarn [Ministère de l'éducation], Bangkok, 2002.

Etude articulatoire du mouvement des lèvres lors d'émotions et une attitude simulées.

Laurianne Georgeton

LPP Laboratoire de Phonétique et de Phonologie – CNRS/ UMR 7018
ILPGA, 19 rue des bernardins – 75005 Paris
laurianne.georgeton.1@univ-paris3.fr

ABSTRACT

Lip aperture, width and protrusion for the French vowels /a, i, u, y/ are measured for three female speakers reading 20 verses extracted from Phèdre, Racine. The verses were recorded in a neutral way and in a variety of expressive modes including anxiety, angry, disgust, joy, sadness, and one attitude: tenderness. Lips motion was captured using a Qualisys optical motion capture system. Compared to the articulation in the neutral condition, the vowels show: more stretching of lips (=larger lip width) in the joy, tenderness and angry conditions, a larger lip aperture in the same conditions and disgut and particularly greater lip protrusion in joy. Results show that lip width, aperture and protrusion vary during the vowel interval when comparing measurements at the onset, middle and offset of the vowel. We observe tendencies regrouping /a/-/i/ and /u/-/y/, the unrounded vowels /a,i/ tend to pattern together compared to the rounded vowels /u, y/.

1. INTRODUCTION

Fonagy [Fon83] fait une distinction éclairante entre émotions primaires et attitudes. Les émotions primaires sont motivées et sont exprimées par des changements physiologiques, à travers tout le corps et à tous les niveaux des appareils vocaux. Les attitudes sont un mélange de motivé (elles dérivent en partie des émotions primaires) et d'arbitraire (dépendant de la langue).
L'objectif de cette étude est tout d'abord d'étudier les mouvements d'ouverture, d'étirement et de protrusion des voyelles /a/, /i/, /u/, /y/ en contexte « neutre », au cours du temps, pour déterminer un point de mesure où ces mouvements sont maximaux. Ensuite, nous ferons une étude comparative de ces gestes en condition « émotions simulées » et en condition « neutre ».
L'articulateur étudié ici, les lèvres, a guidé le choix des émotions et de l'attitude choisies. Nous avons étudié 4 des 6 émotions de base d'Ekman [Ekm78] et leurs réalisations articulatoires aux niveaux des lèvres. La peur n'a pas été étudiée comme telle : nous avons désigné cette émotion par « anxiété » plus proche du « stress », étudié par Scherer [Sch84] [Sch01]. Nous avons choisi la tendresse, attitude particulièrement étudiée par Fonagy [Fon83], et qui est associée à un avancement des lèvres (marque d'une attitude positive envers l'interlocuteur). Or, un geste de protrusion, d'après Ohala, allonge le conduit vocal, et donc, abaisse les fréquences de résonance : des formants plus bas donnent alors l'illusion d'un plus grand conduit vocal, qui serait une composante de l'agressivité [Oha83] [Oha96].

2. MÉTHODE

2.1. Protocole

Pour cette étude articulatoire, nous avons utilisé le système Qualysis (http://www.qualisys.se/). Le principe des caméras 3D consiste à détecter (100images/sec) la position de marqueurs réfléchissants qui sont placés sur le sujet à des localisations biomécaniques cibles : deux aux commissures des lèvres, une sur l'arc de cupidon de la lèvre supérieure et une en regard sur la lèvre inférieure comme illustré sur la figure 1(gauche). Des capteurs de référence sont situés sur l'arête du nez (point de référence) et sur les tempes. Le rayonnement infra-rouge est capté par les caméras sous des angles différents qui permettent ensuite une reconstruction 3D du déplacement des marqueurs.

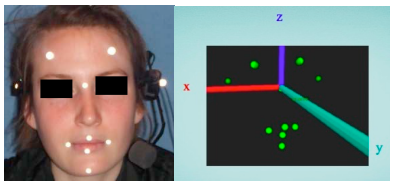


Figure 1 : A gauche, positions des marqueurs sur le visage d'un locuteur lors des enregistrements. A droite, visualisation du visage via le logiciel QTM, avec un axe de référence x, y, z.

Nous avons fait appel à trois femmes d'âge compris entre 23 et 35 ans, de langue maternelle française et ayant une expérience en expression scénique amateur ou semi professionnel. Nous avons voulu présenter un corpus qui permet aux locuteurs de se projeter et de favoriser la performance scénique. L'alexandrin que l'on retrouve dans le théâtre classique s'est alors imposé comme la forme la plus adéquate. Nous avons extrait notre corpus de la pièce de Phèdre de Racine (voir page 4).

2.2. Corpus

Condition « neutre » : Nous avons demandé aux locuteurs de lire deux fois *de façon la plus neutre possible* les alexandrins précédemment décrits. Cet enregistrement (deux répétitions du corpus) sera considéré dans la suite des travaux comme référence et sera désigné par « normal » sur les graphiques.
Condition « émotions simulées » : Ensuite, nous leur

avons demandé de lire le corpus une fois en simulant une par une les cinq émotions et l'attitude qui nous intéressent. Les données enregistrées sur le logiciel d'exploitation du Qualisys, ont été exportées sur Matlab, la segmentation du signal acoustique a été effectuée sur Praat.

2.3. Mesures

Nous avons étudié le même nombre et les mêmes voyelles pour chacune des productions. Nous avons effectué 105 mesures de la voyelle /a/, 87 mesures de la voyelle /i/, 27 mesures de /y/ et 42 mesures de /u/ à différents moments : au début, à la fin et au milieu acoustiques de la voyelle. Nous avons sélectionné les voyelles de telle manière que la première occurrence de la voyelle /a/ d'une condition « émotion simulée » soit comparable avec la première occurrence de /a/ de la condition « neutre ». Ceci nous permet donc de comparer les voyelles entre elles. Nous avons donc effectué les mesures suivantes :

Ouverture aux lèvres (HB)	Distance euclidienne entre réflecteurs des lèvres supérieure et inférieure.
Etirement des lèvres (DG)	Distance euclidienne entre réflecteurs des coins droit et gauche des lèvres.
Etirement des lèvres (DG)	Trajectoire sur un axe x, y, z du réflecteur gauche des lèvres par rapport à un point d'origine : l'arête du nez.

3. VOYELLES EN CONDITION NEUTRE

Afin de déterminer nos points de mesures pour la comparaison avec les émotions, nous nous sommes intéressée aux variations de l'articulation labiale au cours du temps. L'ensemble des locuteurs montre les mêmes variations.

3.1. Étirement DG :

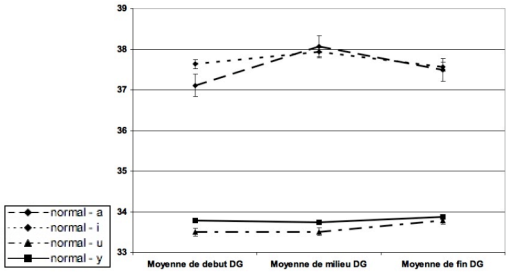


Figure 2: Variations des moyennes des distances DG au début, milieu et fin des voyelles /a/, /i/, /u/, /y/, en condition neutre.

A tous les instants, la distinction entre voyelles arrondies et voyelles non-arrondies apparaît : l'étirement est moindre pour les voyelles arrondies /y/ et /u/ que pour les voyelles non-arrondies /i,a/. La variation de l'étirement, au cours du temps, est également différente en fonction des deux classes de voyelles : les voyelles /a/ et /i/ ont une valeur DG maximale au milieu de la voyelle alors que les voyelles /u/ et /y/ ont une valeur maximale à la fin de la voyelle.

	Début	Milieu	Fin
Voyelle /a/	37,10	38,06	37,49
Voyelle /u/	33,50	33,50	33,79

Tableau 1: Valeurs moyennes de la distance DG au début, milieu et fin des voyelles /a/ et /u/, pour un locuteur (L1), en mm.

3.2. Ouverture aux lèvres HB :

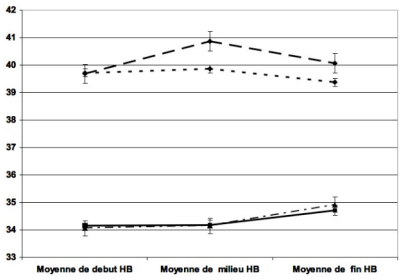


Figure 3: Variations des moyennes des distances DG au début, milieu et fin des voyelles /a/, /i/, /u/, /y/, en condition neutre.

Nous retrouvons les caractéristiques des voyelles fermées /u/ (HB= 34,15mm) et /y/ (HB= 34,17mm) et de la voyelle ouverte /a/ qui demande un degré d'aperture maximale (HB= 40,86mm). On observe, cependant, une valeur d'aperture pour la voyelle /i/ assez élevée (HB= 39,86 mm), ce qui nous laisse supposer que la voyelle /a/ est éloignée de sa cible articulatoire (et se rapproche ainsi de l'articulation des phonèmes qui l'entourent) pour des raisons d'effort articulatoire [Lin90]. Une autre explication, qui n'exclut pas la précédente, est que la voyelle /a/ est la voyelle la plus fréquente en français [Wio85]. Ces explications mises en corrélation avec les résultats de la distance DG montre en effet que la voyelle /a/ ne montre pas les caractéristiques attendues, mais est proche de la réalisation articulatoire de la voyelle /i/. Ensuite, le relevé plus précis des valeurs de la distance HB au début, milieu et fin de la voyelle, nous montre que les voyelles /a/ et /i/ ont une valeur maximale au milieu de leur réalisation alors que les voyelles /u/ et /y/ observent une valeur maximale à la fin de leur réalisation.

	Début	Milieu	Fin
Voyelle /a/	39,68	40,87	40,06
Voyelle /u/	34,05	34,15	34,92

Tableau 2: Valeurs moyennes de la distance HB au début, milieu et fin des voyelles /a/ et /u/, pour un locuteur (L1), en mm.

3.3. Protrusion Prot :

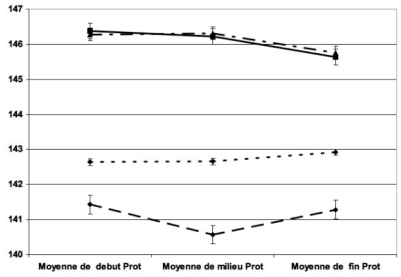


Figure 4: Variations des moyennes des distances DG au début, milieu et fin des voyelles /a/, /i/, /u/, /y/, en condition neutre.

Nous pouvons observer un avancement plus important du réflecteur pour les voyelles /u/ au milieu de la voyelle (Prot= 146,32mm) et /y/ (Prot= 146,22mm) comparé aux voyelles /i/ (Prot= 142,65mm) et /a/ (Prot= 140,57mm). La voyelle /u/ est légèrement plus protruse que la voyelle /y/ ce qui est en accord avec les travaux de Benguèrel [Ben74]. Les valeurs de protrusion de la voyelle /i/ sont remarquables (Prot= 142,65 mm). On peut supposer que la voyelle /i/, qui montrait des valeurs d'ouverture et d'étirement similaires à la voyelle /a/, se contraste de cette voyelle grâce à ce léger mouvement de protrusion, d'avancement des lèvres. De plus, le relevé précis des valeurs de protrusion au début, milieu et fin montre contrairement aux valeurs d'étirement et d'ouverture, une valeur maximale de protrusion au début de la réalisation des voyelles /a/ et /i/ et au milieu de la réalisation des voyelles /u/ et /y/ .

	Début	Milieu	Fin
Voyelle /a/	141,43	140,57	141,28
Voyelle /u/	146,28	146,32	145,76

Tableau 3: Valeurs moyennes de la trajectoire Prot au début, milieu et fin des voyelles /a/ et /u/, pour un locuteur (L1), en mm.

Le relevé précis des distances HB, DG et trajectoire Prot, nous permet de voir que le geste qui caractérise et qui distingue les voyelles entre elles, est maximal au milieu de celles-ci, excepté pour la voyelle /y/. Pour les voyelles /a/ et /i/, les gestes d'étirement et d'ouverture sont maximaux au milieu de la réalisation acoustique de la voyelle. Pour la voyelle /u/, le geste de protrusion est maximal au milieu de la voyelle.

4. VOYELLES DANS LES ÉMOTIONS SIMULÉES

Compte tenu des résultats précédents, nous avons voulu poursuivre notre étude sur le mouvement des lèvres en condition « émotions simulées » sur les données prises au milieu de la voyelle. Nous avons pour cela effectué un test Anova pour observer l'influence des émotions et attitude sur la réalisation des voyelles /a/, /i/, /u/ et /y/.

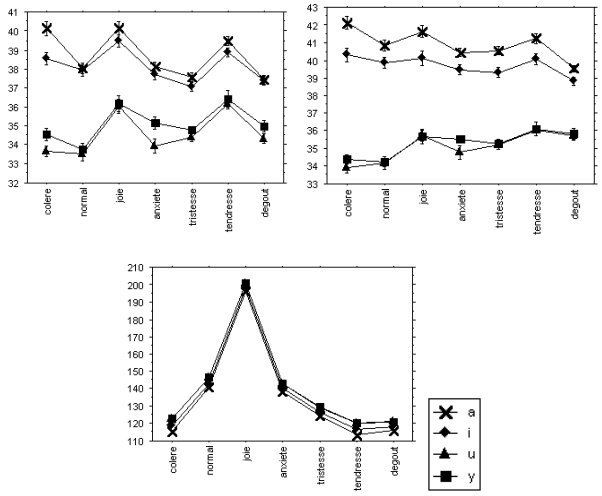


Figure 5: Courbes des interactions pour les distances DG (gauche), HB (droite) et de la protrusion (milieu), effet émotions*voyelles, pour un locuteur (L1).

4.1. Étirement DG :

La valeur F(6, 18)= 23 p< 0,0001 nous permet de conclure d'un effet global des émotions sur les valeurs de la distance DG, et donc sur l'étirement des lèvres. Pour les voyelles /a/ et /i/, nous observons une augmentation des valeurs DG pour la colère, la joie la tendresse. Pour les voyelles /u/ et /y/, Les valeurs de DG augmentent lors de la simulation de la joie et la tendresse. Contrairement aux propos de Fonagy, il semblerait que les locuteurs expriment la tendresse avec un sourire plutôt qu'un allongement des lèvres vers l'avant. Il existe un contraste significatif entre ces émotions et la condition normale, comme le montre le test PLSD de Fisher ci-dessous :

	Diff. moy	Diff. crit.	Valeur p
Neutre/Colère	1,154	0,436	<0,0001
Neutre/Joie	-2,002	0,437	<0,0001
Neutre/Tendresse	-1,595	0,436	<0,0001

Cependant, nous pouvons nous interroger sur la présence d'étirement dans chacune des émotions ou attitudes actées par rapport à la condition contrôle. Cette marque d'étirement peut-être un indice d'hyperarticulation des acteurs en condition de jeu théâtral qui pourrait être induit par la situation de jeu tout à fait anti-naturelle qui est celle de la chambre sourde, pour un acteur habitué à sa liberté de mouvement ce qui le pousserait à exagérer son expressivité orale.

4.2. Ouverture aux lèvres HB :

La valeur F(6, 18)= 4 p= 0,0001 nous permet de conclure d'un effet global des émotions sur les valeurs de la distance HB, et donc sur l'ouverture aux lèvres. Pour les voyelles /u/ et /y/, excepté pour la colère on observe une augmentation de la distance HB pour l'ensemble des émotions et attitude. Pour la voyelle /a/, nous observons une augmentation lors de la simulation de la colère, la

joie et la tendresse et une diminution de la distance pour le dégoût. Pour la voyelle /i/, nous observons également une diminution de la distance HB pour le dégoût. Il existe un contraste significatif entre ces émotions, attitude et la condition normale, comme le montre le test PLSD de Fisher ci-dessous :

	Diff. moy	Diff. crit.	Valeur p
Neutre/Colère	0,645	0,464	0,0065
Neutre/Joie	-0,777	0,465	0,0011
Neutre/Tendresse	-0,731	0,465	0,0021
Neutre/Dégoût	0,465	0,464	0,0495

4.3. Protrusion Prot :

La valeur F(6, 18)= 17463 p< 0,0001 nous permet de conclure à un effet global des émotions sur les valeurs d'avancement des lèvres ou de protrusion. Pour chaque émotion ou attitude, nous obtenons des valeurs assez similaires entre les voyelles malgré des propriétés intrinsèques différentes. L'expression des émotions et attitude semble donc avoir une influence plus importante que les qualités intrinsèques des voyelles. On observe une augmentation remarquable de la protrusion pour l'ensemble des voyelles lors de la simulation de la joie. Un test PLSD de Fisher nous montre qu'il existe un contraste significatif entre la condition « neutre » et l'ensemble des émotions et attitude étudiées.

5. CONCLUSION

Cette étude articulatoire du mouvement des lèvres, nous a d'abord permis d'étudier les mouvements d'étirement, d'ouverture et de protrusion lors d'émotions et attitude simulées. Les résultats du 1er enregistrement (correspondant à une réalisation « la plus neutre possible ») nous permet de mieux comprendre les variations des voyelles /a/, /i/, /u/ et /y/ au cours du temps. Pour l'ensemble des mesures, on observe un regroupement des voyelles /a/-/i/ et /u/-/y/. Si le regroupement des voyelles /u/-/y/ était attendu, le rapprochement des voyelles /a/-/i/ est remarquable puisqu'elles ne possèdent pas intrinsèquement les mêmes propriétés articulatoires. Les valeurs articulatoires peuvent êtres justifiées par les hypothèses d'effort articulatoire [Lin90], ou de fréquence de la voyelle en français [Wio85] et dans le texte. Pour les voyelles /a/ et /i/, les gestes d'étirement et d'ouverture sont maximaux au milieu de la réalisation acoustique de la voyelle. Pour la voyelle /u/, le geste de protrusion est maximal au milieu de la voyelle. Cette étude nous permet également d'observer que l'expression des émotions et attitude semble avoir une influence plus importante que les qualités intrinsèques des voyelles. L'étirement DG est maximal pour la joie, la colère et la tendresse. Ce dernier résultat ne va pas dans le sens de Fonagy, qui associait la tendresse à un avancement des lèvres. Ici les locuteurs semblent associer un geste proche du sourire à cette attitude. Cette remarque souligne toute la difficulté de l'étude des attitudes dans une langue, mélange de motivé et d'arbitraire, elles restent extrêmement sensibles à la personnalité ou caractéristiques du locuteur. En regard du nombre de locuteur, nous ne pouvons conclure et associer un geste particulier à la tendresse. L'ouverture aux lèvres est remarquable lors de la colère, tendresse, dégoût et la joie. Les valeurs de protrusion sont maximales pour la joie. Enfin, il nous faut nuancer nos propos quant au caractère dit « normal » du premier enregistrement, car il est difficile de penser que l'alexandrin est une forme proche de la parole spontanée d'aujourd'hui.

REMERCIEMENTS

Nous remercions particulièrement, Mme J. Vaissière, Mme C. Fougeron et Mme A. Amelot pour leurs corrections et leurs conseils avisés. Nous remercions également les locuteurs (Cécile, Claire, Zoé et Laure) qui ont permis la bonne réalisation de cette expérience.

RÉFÉRENCES

[Ben74] Benguèrel A.P. and Cowan H.A. (1974), Coarticulation of Upper Lip Protrusion in French, Phonetica 30, pp. 41-55.

[Ekm78] Ekman P. & Friesen W. V. (1978): Facial action coding system: A technique for the measurement of facial movement, *Consulting Psychologists Press,* Palo Alto, California.

[Fon83] Fonagy I. (1983): *La vive voix*, Paris, Payot.

[Lin90] Lindblom B. (1990) "Explaining phonetic variation : a sketch of the hyper- and hypospeech theory", Speech Production and Speech Modelling, Hardcastle and Marchal (eds.), Kluwer Academic Publishers, 403-439.

[Oha83] Ohala J. J. (1983). Cross-language use of pitch: an ethological view. Phonetica, 40, pp. 1-18.

[Oha96] Ohala J. J. (1996). Ethological theory and the expression of emotion in the voice. 4th International Conference on Spoken Language Processing, Philadelphia, USA. Vol. 3, pp. 1812-1815.

[Sch84] Scherer K. R. (1984). Emotion as multicomponent process : A model and some cross-cultural data. Review of Personality and Social Psychology (5), pp. 37-63.

[Sch01] Scherer K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K Scherer, A Schorr, & T. Johnstone (Eds.). Appraisal processes in emotion: Theory, Methods, Research, Oxford University Press, pp. 92-120

[Wio85] Wioland F. (1985): Les structures syllabiques du français: fréquence et distribution des phonemes consonantiques, contraintes idiomatiques sans les sequences consonantiques, Genève, Slatkine; Paris, Champion.

vers 280 à 283	*« Je lui bâtis un temple, et pris soin de l'orner,* *De victimes moi-même à toute heure entourée,* *Je cherchais dans leurs flancs ma raison égarée. »*
vers 1597 à 1600	*« Mais, Madame, il est mort, prenez votre victime,* *Jouissez de sa perte, injuste ou légitime.* *Je consens que mes yeux soient toujours abusés* *Je le crois criminel, puisque vous l'accusez. »*
vers 1574 à 1586	*« La timide Aricie est alors arrivée.* *Elle venait, Seigneur, fuyant votre courroux,* *À la face des Dieux l'accepter pour époux.* *Elle approche : elle voit l'herbe rouge et fumante ;* *Elle voit (quel objet pour les yeux d'une amante !)* *Hippolyte étendu, sans forme et sans couleur.* *Elle veut quelque temps douter de son malheur ;* *Et ne connaissant plus ce héros qu'elle adore,* *Elle voit Hippolyte, et le demande encore.* *Mais trop sûre à la fin qu'il est devant ses yeux,* *Par un triste regard, elle accuse les Dieux ;* *Et froide, gémissante, et presque inanimée,* *Aux pieds de son amant, elle tombe pâmée. »*

Etude acoustique de la parole après hémiglossectomie et reconstruction par lambeau infra-hyoïdien

Audrey Acher, Cécile Fougeron

Laboratoire de Phonétique et Phonologie, UMR7018, CNRS-Paris3/Sorbonne Nouvelle

19 rue des Bernardins, 75005 Paris

Courriel : audrey.acher@etud.sorbonne-nouvelle.fr ; cecile.fougeron@univ-paris3.fr

ABSTRACT

The aim of this paper is to assess speech functional consequences after hemiglossectomy with flap reconstruction. The speech of 2 patients before and after surgery (1 month and 3 months after) was evaluated by acoustic analysis and articulation was examined with ultrasound. This paper presents acoustic analysis of velar plosive consonants /k/ and /g/ and constrictive consonants /s/ and /z/. Results show impaired production depending on vowel context and delai after surgery especially for /k/ and /g/ with /i/ and /a/ ; for /s/ and /z/, deviations are seen in context /u/.

Keywords : hemiglossectomy, tongue cancer, ultrasound, acoustic analysis.

1. INTRODUCTION

Le cancer de la langue peut être traité par chirurgie. Lorsque la tumeur linguale nécessite une exérèse de la moitié de la langue dans le sens longitudinal, l'opération réalisée s'appelle une hémiglossectomie. Lors de cette intervention, le chirurgien peut choisir de combler la perte de substance par un lambeau. Il semble important de connaître les conséquences fonctionnelles de ce type d'intervention sur l'articulation de la parole, d'une part pour pouvoir évaluer les différents types de reconstruction et d'autre part pour mieux prendre en charge ces patients en rééducation orthophonique.

Pour autant les troubles articulatoires liés à la glossectomie sont peu connus, cette population étant peu représentée dans la littérature en phonétique clinique. Il a pourtant été relevé qu'après hémiglossectomie, des troubles articulatoires peuvent exister. L'altération du mouvement lingual occasionnée par ce type d'intervention a été associé à des difficultés de prononciation de /s/, /ʃ/ et /t/ en japonais (Imai & al. 1992 [4], étude perceptive et articulatoire EPG), /k/ est trouvé altéré en finnois (Korpijaakko-Huuhka & al. 1998 [5], étude perceptive). Savariaux et al. [7] ont mis en évidence un bruit important lors de l'articulation des occlusives /t, d, k, g/ en français lors d'une étude acoustique, les chercheurs se sont interrogés sur un phénomène de compensation. Sun et al. [8] ont montré qu'en chinois si la pointe de langue est préservée lors d'hémiglossectomie, les résultats sont meilleurs. Bressmann et al. [2] [3] ont réalisé une étude avec l'échographie, ils ont montré qu'en post-opératoire, l'antériorité, la convexité de la langue et son asymétrie avaient des valeurs plus importantes qu'en pré-opératoire.

Dans une étude précédente (Acher 09 [1]), visant à apprécier les conséquences fonctionnelles articulatoires d'une hémiglossectomie chez deux patients, nous avons procédé à une évaluation perceptive d'un échantillon de leurs productions (syllabes VCV) par 5 juges. Les consonnes jugées les plus altérées étaient les consonnes vélaires /k, g/ et les fricatives antérieures /s, z/. Nous avons également observé les contours linguaux dans la production de ces consonnes altérées sur des enregistrement échographiques. Les observations principales étaient une importante asymétrie du profil de la langue dans les coupes coronales..Toutefois, si l'échographie permet de décrire la hauteur, le recul et la forme de la langue pendant l'articulation, elle ne permet pas d'évaluer sa position relative au palais (qui ne se voit pas sur l'image). Pour les occlusives, il n'est donc pas possible de dire avec certitude si l'occlusion est réalisée et pour les fricatives de voir le chenal entre la pointe de la langue et le palais. Le travail présenté ici s'inscrit donc en continuité de ces travaux en proposant une évaluation acoustique des productions des patients. Il sera question d'évaluer le nombre de productions de consonnes acoustiquement 'déviantes', en fonction du type de consonnes, en fonction du contexte vocalique adjacent et en fonction du délai post-opératoire. Ces données seront dans le futur mises en regard avec les données articulatoires.

2. PROTOCOLE EXPERIMENTAL

2.1. Matériel linguistique

Un corpus de logatomes a été constitué afin d'évaluer la capacité des locuteurs à atteindre les cibles articulatoires consonantiques du français. Il est composé de logatomes de type CVCVC où C = /t d k g s z ʃ ʒ l j/ et V = /i u a/ pour l'étude de la variation des consonnes en fonction du contexte vocalique. Cette liste comprend 3 répétitions de chaque logatome présenté en ordre aléatoire.

2.2. Locuteurs

Nous avons réalisé une étude sur un échantillon de 2 patients ayant subi une hémiglossectomie droite avec reconstruction par lambeau infra-hyoïdien. Le patient 1 est un homme de 28 ans et le patient 2 est une femme de 62 ans. Les patients ont été enregistrés en pré-opératoire, en post-opératoire 1 mois puis en post-opératoire 3 mois. Nous avons également enregistré sur le même matériel 3 sujets contrôles : 1 femme de 53 ans, et deux hommes de 31 et 33 ans.

2.3. Recueil et analyse des données

Données acoustiques

Nous avons analysé la parole des deux patients de l'étude avec le logiciel d'analyse acoustique Praat. Le signal sonore analysé a été enregistré avec le logiciel d'acquisition et de traitement des données Articulate Assistant Advanced (AAA) version 2.09 d'Articulate Instrument [9] lors de la prise de données articulatoires. Nous avons analysé les réalisations acoustiques des six répétitions (3 sagittal, 3 coronal) des consonnes /k/, /g/, /s/ et /z/ à chaque temps. La caractérisation acoustique est effectuée par une observation visuelle des productions (sur le signal et le spectrogramme) et une caractérisation des productions en fonction de la présence/absence des indices acoustiques suivants :

- Pour les occlusives /k/ et /g/, nous nous sommes intéressées à l'absence de relâchement (barre d'explosion et signal apériodique) et à la présence d'énergie dans les moyennes et hautes fréquences, interprétés comme des indices de désocclusion, lénition
- Pour les fricatives /s/ et /z/, nous avons regardé la présence de formants dans les hautes fréquences. et la distribution du bruit

Nous avons étudié les données des locuteurs hémiglossectomisés en pré-opératoire et en post-opératoire 1 mois et 3 mois.

Données articulatoires

Nous avons enregistré le mouvement de la langue lors de l'articulation des logatomes avec un échographe portable 2D Mindray DP-6600 équipé d'une sonde convexe 35C20EA de fréquence 3,5 MHz permettant l'acquisition de données articulatoires linguales. Les données linguales ont été recueillies dans les plans coronal et sagittal. Nous avons utilisé le module d'acquisition de données WinEPG pour synchroniser le signal sonore et l'échographie. L'échographe et le module d'acquisition étaient reliés via les ports vidéo et audio à un PC équipé du logiciel AAA.

3. RÉSULTATS

3.1. Analyse acoustique

Le tableau 1 présente la distribution des productions en fonction des indices examinés par inspection visuelle. En pré-opératoire, environ 10% des productions présentent un caractère déviant tout patient confondu.

Le patient #1 présente globalement plus d'altérations acoustiques que le patient #2 tant en post-opératoire 1 mois (patient #1: 39 et 50%, patient #2: 26 et 17%) et en post-opératoire 3 mois (patient #1: 36 et 29%, patient #2: 21 et 8%). Pour le patient #1, en post-opératoire 1 mois, une occlusive sur 2 est altérée ; les fricatives sont aussi altérées mais dans une moindre mesure. A 3 mois, fricatives et occlusives sont altérées dans des proportions similaires. La comparaison entre post-opératoire 1 mois et 3 mois pour ce patient montre une amélioration uniquement pour les occlusives. Pour le patient #2, les altérations touchent plus les fricatives que les occlusives en post-opératoire 1 mois et 3 mois. Comme pour le patient #1, le patient #2 montre une amélioration sur les occlusives avec l'augmentation du délai opératoire.

Les altérations observées dépendent du contexte vocalique. Les altérations sont ainsi plus fréquentes quand un grand mouvement articulatoire est requis entre la voyelle et la consonne. Ainsi, il existe plus d'altération sur les vélaires en contexte /a/ et /i/ (24% vs. 15% pour /u/) et sur les fricatives apico-alvéolaires en contexte /u/ (44% vs. 13% pour /a/ /i/).

Séquences		Types de déviations	Patient #1			Patient #2		
			Pre-op	Post-op 1m	Post-op 3m	Pre-op	Post-op 1m	Post-op 3m
/s/	contexte a	formants	0	0	0	0	0	0
		bruit bas et diffus	0	0	0	0	33	0
	contexte i	formants	0	0	0	0	0	0
		bruit bas et diffus	0	50	0	17	50	17
	contexte u	formants	0	0	0	0	0	0
		bruit bas et diffus	0	100	100	17	83	100
/z/	contexte a	formants	0	33	0	0	33	0
		bruit bas et diffus	0	17	0	0	0	0
	contexte i	formants	50	67	100	0	0	0
		bruit bas et diffus	0	67	33	0	33	17
	contexte u	formants	50	33	100	50	0	17
		bruit bas et diffus	17	100	100	0	83	100
Total			**10**	**39**	**36**	**7**	**26**	**21**
/k/	contexte a	formants/bruit	0	0	0	0	0	0
		explosion pas nette	0	0	0	0	17	0
	contexte i	formants/bruit	0	100	0	0	0	50
		explosion pas nette	0	100	0	0	33	0
	contexte u	formants/bruit	0	0	33	0	0	0
		explosion pas nette	17	0	33	33	0	0
/g/	contexte a	formants/bruit	17	83	33	17	33	0
		explosion pas nette	0	100	67	0	50	0
	contexte i	formants/bruit	0	100	50	33	0	50
		explosion pas nette	0	100	17	33	67	0
	contexte u	formants/bruit	33	0	33	0	0	0
		explosion pas nette	67	17	83	0	0	0
Total			**11**	**50**	**29**	**10**	**17**	**8**

Tableau 1 : Nombre de cas exprimés en pourcentage calculés sur un nombre de répétitions (N=6) présentant des déviations en fonction des consonnes /s,z,k,g/; du contexte vocalique, du délai opératoire, par patient, avec moyenne

Afin d'illustrer ces tendances générales, nous allons décrire les types de déviances observées.

Concernant les occlusives, en pré-opératoire chez les 2 patients, la barre d'explosion et le silence dans les moyennes et hautes fréquences sont souvent présents sur

le spectrogramme (à part pour [ugu] chez le patient #1 où dans 67% des cas, la barre d'explosion est peu visible). La figure 1 montre l'articulation en pré-opératoire d'une séquence [aga] canonique chez le patient #1.

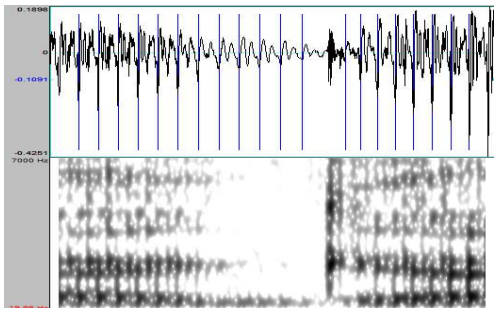


Figure 1 : Signal et spectrogramme produits par le patient #1 durant la séquence [aga] en pré-opératoire

Lors de l'articulation de [g] en post-opératoire 1 mois en contexte [a] et [i], la barre d'explosion n'est pas claire et il existe des formants pour le patient #1 en contexte [a] dans 83 % des productions (figure 2), du bruit en contexte [i] pour toutes les réalisations (figure 3).

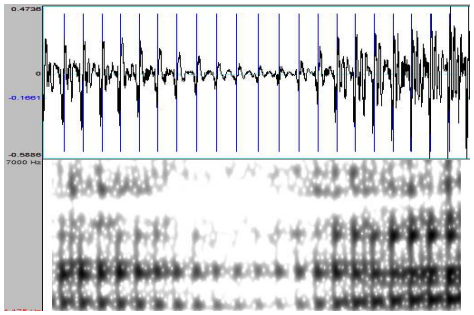


Figure 2 : Signal et spectrogramme produits par le patient #1 durant la séquence [aga] en post-opératoire 1mois

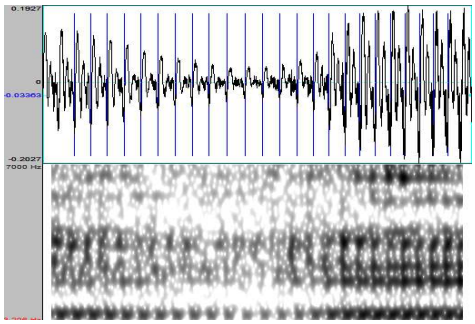


Figure 3 : Signal et spectrogramme produits par le patient #1 durant la séquence [igi] en post-opératoire 1 mois

Lors de l'articulation de la séquence [iki] nous trouvons une explosion peu claire et du bruit chez le patient #1 en post opératoire 1 mois dans les 6 répétitions comme le montre la figure 4.

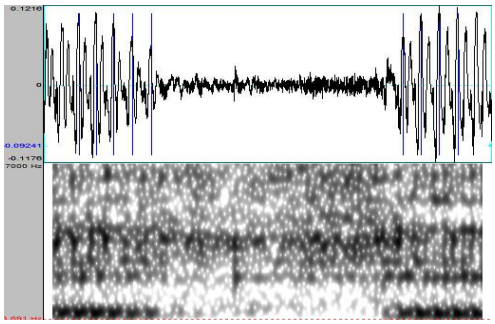


Figure 4 : Signal et spectrogramme produits par le patient #1 durant la séquence [iki] en post-opératoire 1 mois

Pour les séquences [uku] et [ugu], nous n'avons pas relevé d'indices pertinents sur le spectrogramme qui pourraient mettre en évidence une éventuelle déviation, les occlusives vélaires en contexte [u] apparaissent ainsi non altérées.
Ainsi, la présence de formants et/ou de bruit de friction témoigne d'une occlusion linguale incomplète lors de la réalisation des consonnes occlusives vélaires /k/ et /g/. Il semblerait y avoir un effet du contexte vocalique car lorsque les consonnes /k/ et /g/ sont en contexte /u/, le profil acoutique est plus canonique c'est justement dans ce contexte que l'amplitude linguale requise pour le mouvemnt VCV est la plus faible.

Concernant les fricatives /s/ et /z/, nous remarquons globalement un abaissement du centre de gravité du bruit de friction et/ou un bruit plus diffus entre les temps pré-opératoire et post-opératoire. Cette tendance s'observe en particulier en contexte /u/ pour les deux patients comme l'illustre la figure 5.

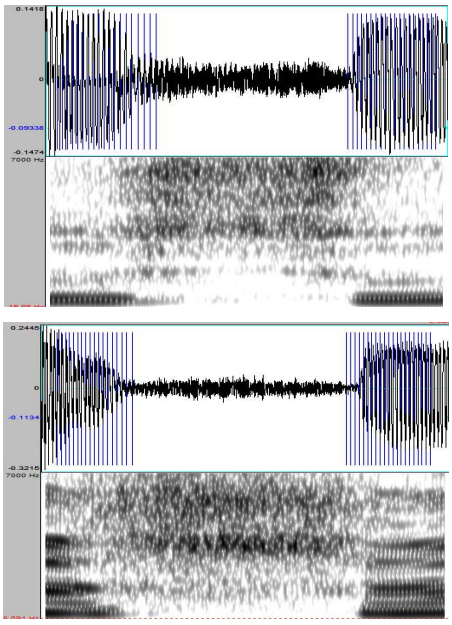


Figure 5 : Signal et spectrogramme produits par le patient #2 durant la séquence [usu] en pré-opératoire (en haut) et en post-opératoire 1 mois (en bas)

4. DISCUSSION ET CONCLUSION

L'étude de ces deux patients par inspection visuelle du spectrogramme et du signal témoigne donc d'une difficulté à produire les consonnes occlusives vélaires [k] et [g] et les fricatives [s] et [z] 1 mois après l'intervention avec une amélioration de la production des occlusives plus importante que celle des fricatives entre les temps post opératoire 1 mois et 3 mois.

Il serait nécessaire de confronter les données acoustiques avec les données articulatoires. Les données acoustiques apportent des informations sur l'occlusion de la langue contre le palais et les données articulatoires pourraient permettre d'interpréter les déviances acoustiques observées en ayant une idée de la forme de la langue lors de la tenue de l'occlusion.

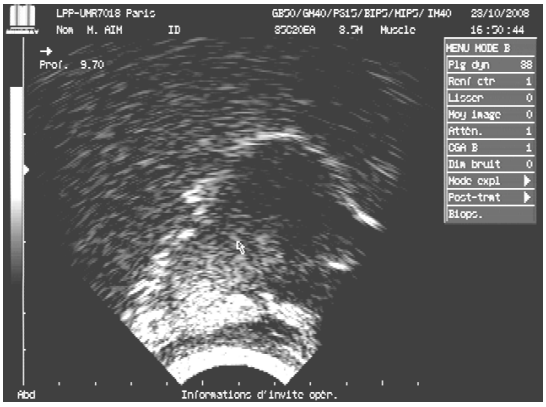


Figure 6 : Forme de la langue lors de la tenue d'un /g/ en contexte /iCi/ en pré-opératoire chez le patient #1

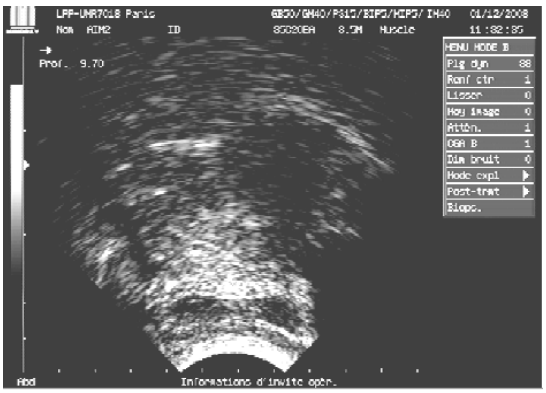


Figure 7 : Forme de la langue lors de la tenue d'un /g/ en contexte /iCi/ en post-opératoire 1 mois chez le patient #1

Ainsi, sur la figure 6, en pré-opératoire, nous remarquons que lors de l'articulation d'un [g] en contexte [i], la langue est arquée, l'occlusion se fait à l'avant du conduit vocal, la pointe de la langue est basse. En post-opératoire 1 mois (figure 7), lors de l'articulation de la même séquence chez le même patient, la masse de la langue est plutôt centrale, la langue n'est pas assez bombée sur sa partie antérieure pour la réalisation de l'occlusion.

L'hémiglossectomie génère une forte gêne pour la production des consonnes vélaires comme l'ont déjà montré Savariaux, Perrier et al. [6] ainsi que pour la production des fricatives apico-alvéolaires (comme montré par [4]). Ces dernières restent altérées 3 mois après l'intervention.. L'échographie permet d'obtenir des données intéressantes quant à la forme de la langue lors de l'articulation des phonèmes occlusifs et fricatifs que nous analyserons à l'avenir en parallèle des données acoustiques afin de mieux cerner les difficultés des patients hémiglossectomisés.

BIBLIOGRAPHIE

[1] A. Acher. Etude perceptive et articulatoire de la parole à partir de données échographiques en 2D : comparaison de la parole normale et de la parole pathologique de patient hémiglossectomisés ; M2, Université Paris III Sorbonne Nouvelle, 2009.

[2] T. Bressmann, P. Thind and al. Quantitative three-dimensional ultrasound analysis of tongue protrusion, grooving and symmetry: data from 12 normal speakers and a partial glossectomee. In *Clinical Linguistics and Phonetics*, volume 19 (6-7), pages 573-588, 2005.

[3] T. Bressmann, E. Ackloo and al. Quantitative three-dimensional ultrasound imaging of partially resected tongues. In *Otolaryngolohy Head & Neck Surgery*, volume 136 (5), pages 799-805, 2007

[4] S. Imai and K.-I. Michi. Articulatory Function After Resection of the Tongue and Floor of the Mouth : Palatometric and Perceptual Evaluation. In *Journal of Speech and Hearing Research*, volume 35 (1), pages 68-78, 1992.

[5] A.-M. Korpijaakko-Huuhka and A.-M. Söderholm. Long-lasting speech and oral-motor deficiencies following oral cancer surgery : a retrospective study. In *Logopedics Phoniatrics Vocology*, volume 24 (3), pages 97-106, 1998.

[6] C. Savariaux, P. Perrier and al. Production de parole après traitement de la cavité endobuccale. In *Proceedings of JEP 2000,* France : Aussois

[7] C. Savariaux, P. Perrier and al. Speech production after glossectomy and reconstructive lingual surgery: a longitudinal study. In *Proceedings of the 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Italia : Firenze, 2001.

[8] J. Sun, Y. Weng. Analysis of determinants on speech function after glossectomy. In *Journal of Oral and Maxillofacial Surgery*, volume 65 (10), pages 1944-1950, 2007.

[9] A. Wrench. Articulate Assistant Advanced : ultrasound module. In *Oral presentation at Ultrafest IV*, USA : New York, 2007.

La résistivité de la gémination en tarifit

[1]*Fayssal Bouarourou* [1]*Béatrice Vaxelaire* [2] *Rachid Ridouane* [1]*Fabrice Hirsch* [1]*Rudolph Sock*
[1]Institut de Phonétique de Strasbourg & U.R.. 1339 – LiLPa -Equipe Parole et Cognition – Université de Strasbourg
[2]Laboratoire de Phonétique et Phonologie (CNRS – Sorbonne Nouvelle)
fayssalbouarourou@noos.fr

ABSTRACT

This investigation reports on gemination in Tarifit Berber. It presents results of an acoustic investigation of singleton and geminate voiceless stops, produced in intervocalic position at a normal and fast speaking rate. They show that geminates are systematically produced with longer closure durations compared to singletons. Gemination, however, does not affect the duration of adjacent vowels, neither does it affect the duration of VOT. In addition, this contrast is shown to be resistant to increased speaking rate. Results are discussed in the light of X-ray data obtained in a previous companion study.

Keywords: Berber, gemination, acoustics, speech rate

1. INTRODUCTION

L'objectif de cette investigation est d'étudier les occlusives simples et géminées produites par 2 locuteurs natifs (F et K) du tarifit, à partir de données acoustiques. La vitesse d'élocution est variée, afin de tester la résistivité du phénomène phonologique. Notre étude s'insère dans le cadre de recherches programmatiques, articulatoires, acoustiques et perceptives, plus larges sur la gémination en tarifit. Les consonnes géminées ont donné lieu à différentes études acoustiques. Une constante a pu être tirée, à partir de ces recherches, sur ces segments : les consonnes géminées sont plus longues que leurs homologues simples (voir, par ex., Ridouane [6] et [7]). Nous tenterons dans ce travail de lier nos résultats acoustiques à des considérations articulatoires.

Löfqvist et Gracco [5] ont analysé des événements moteurs durant la production d'occlusives bilabiales, à l'aide de données cinématiques couplées avec des informations portant sur la pression d'air orale, et la force de contact labial. Leurs résultats suggèrent que la vélocité du mouvement des lèvres était plus élevée durant la tenue consonantique, avec une cible pour les lèvres qui pouvait correspondre à une région d'ouverture labiale. Une telle stratégie de contrôle laisse penser que les lèvres formeraient un ensemble hermétique indépendamment de toute variabilité contextuelle. L'étude cinéradiographique de Vaxelaire [9] sur des consonnes simples et doubles du français montre que la zone de contact palais/langue est plus importante pour les occlusives longues que pour les occlusives brèves. En vitesse d'élocution rapide, les différences au niveau de l'étendue de contact s'accroissent, puisqu'une augmentation nette de l'aire du contact, entre la langue et le palais, est observée, notamment pour les occlusives longues.

2. MÉTHODE

Le corpus entier est composé de 54 phrases courtes de 6 syllabes en moyenne, comprenant 27 paires minimales d'occlusives et de fricatives simples et géminées. Dans ce travail, seules 3 paires minimales ont été prises en considération, et cela en position intervocalique. Quant aux séquences-cibles, elles étaient du type V1C(C)V2 Les occlusives examinés étaient: [t, k, q] *vs*. [tt, kk, qq]. Exemple du corpus : [a**k**as] *vs*. [a**kk**as]. Tous les mots étaient insérés dans la même phrase porteuse : [Ini---- iʒ umar], signifiant « Dis ___ une fois ». Les deux sujets (F et K) étaient assis à l'aise, à environ 20 cm du microphone. Les phrases ont été enregistrées d'abord en vitesse d'élocution normale, puis en vitesse d'élocution rapide. Plusieurs événements temporels ont été relevés sur le signal audio : la durée de V1, la durée d'occlusion, la durée du VOT et la durée de V2. La durée de V1 et V2 a été mesurée comme l'intervalle temporel entre l'apparition et la disparition de la structure formantique stable de ces voyelles. La durée d'occlusion correspond à l'intervalle entre la fin de V1 et le début du relâchement, et le VOT comme l'intervalle entre le relâchement et l'apparition de la structure formantique stable de V2. Rappelons que l'augmentation de la vitesse d'élocution est un facteur perturbateur crucial et naturel du système linguistique. Il s'agit surtout d'évaluer la résistivité du trait phonologique de la gémination en observant si, indépendamment des changements éventuels des valeurs absolues, les relations entre les variables principales de la gémination (durées consonantiques brèves et longues) resteraient inchangées.

Ce travail se propose de vérifier quatre *hypothèses* : Premièrement, et comme cela est attesté dans la littérature, nous pensons que la durée de la tenue consonantique serait plus longue pour les géminées par rapport aux consonnes simples. Comme deuxième hypothèse, il est probable que la durée des voyelles situées avant et après une géminée soit affectée par la présence de cette dernière : les voyelles seraient plus brèves dans ce type d'environnement (Lehiste *et al.*, [4]), dans le cas de syllabes isochrones. Troisièmement, il est judicieux de croire que le VOT serait plus long pour les géminées, étant donné que leur phase d'occlusion est généralement plus longue, ce qui retarderait le début du voisement, de par une pression intra-orale plus élevée (Ridouane, [6]), engendrant ainsi une durée de relâchement plus longue.

Notre quatrième hypothèse porte sur le paradigme de la variation de la vitesse d'élocution. Nous pensons, suivant les résultats habituellement attestés dans la littérature sur les oppositions de quantité (voir, par ex., Abry *et al.*, [1] ou Sock , [8]), que malgré la compression que risque de subir les paramètres mesurés, face à l'augmentation de la vitesse d'élocution, la différence de durée de la tenue consonantique (le paramètre de prédilection du contraste phonologique) sera maintenue. Etant donné la grande plasticité du signal de parole (Gaitenby, [3]) suivant les locuteurs, la vitesse d'élocution, les divers contextes…, les différences de *durée absolue* entre entités simples et géminées seront normalisées. Nous calculerons ainsi la proportion prise par la tenue consonantique dans la syllabe CV2. En effet, il a été démontré pour les langues à quantité consonantique (voir, entre autres, Lehiste *et al*. [4]) que c'est dans ce domaine CV que les contrastes temporelles se manifestaient au mieux. *In fine*, l'analyse minutieuse de nos données et les enseignements que nous en tirerons se feront, en conséquence, à partir de ces *valeurs relatives*.

3. RÉSULTATS ET DISCUSSION

3.1. Remarques générales

Les résultats présentés dans cette investigation reposent sur des mesures de : 6 occlusives non voisées (3 paires de simples *vs*. géminées), prononcées par 2 locuteurs (F & K), dans deux conditions de vitesse d'élocution (normale *vs*. rapide) à 10 reprises (retenues sur les douze répétitions). Des analyses de variance (ANOVA) ont été effectuées pour toutes les *variables* (V1, tenue consonantique, VOT et V2), afin de déterminer s'il existait des effets de *gémination*, de *vitesse* d'*élocution* et de *type consonantique*. Deux *effets principaux se sont* révélés statistiquement significatifs pour la variable *occlusion consonantique* : *gémination* [(df=1,238, F=1426.27, p<0.0000)] et *vitesse* d'*élocution* [(df=1,238, F=2940.624, p<0.0000)] Par conséquent, nous n'avons procédé à des comparaisons *post-hoc*, paire-par-paire (*h.s.d. de Tukey*) de la moyenne des valeurs absolues et relatives que pour cette variable.

3.2. Résultats en valeurs absolues

De manière générale, les résultats, en *valeurs absolues*, montrent que la durée d'occlusion des géminées est sensiblement plus longue que celle de leurs homologues simples. La Figure 1 est une illustration typique de ce fait. Ce résultat est donc conforme à l'hypothèse 1. On remarque, comme on pouvait s'y attendre suivant les résultats statistiques cités *supra*, que la gémination n'a pas d'influence sur la durée de V1 qui est de 76 ms (écart-type = 5 ms) avant les géminées, et de 86 ms (écart-type = 10 ms) avant les simples. Aucune répercussion n'est visible non plus sur la durée de V2, puisque cette voyelle a été quantifié à 165 ms (écart-type = 9 ms) lorsqu'elle est précédée d'une géminée, et à 169 ms (écart-type = 12 ms) avant une simple. Cela signifierait que la deuxième hypothèse qui avait été formulée n'a pas été vérifiée. Il en va de même de la troisième hypothèse, puisque les valeurs du VOT sont similaires pour les deux catégories de consonnes étudiées. En revanche, la quatrième hypothèse a été confirmée, étant donné que la différence de durées consonantiques entre les consonnes simples et géminées est maintenue en vitesse d'élocution rapide. Cela suggère que la gémination est résistante, malgré la compression des durées segmentales (de 119 ms [écart-type = 4 ms] en condition normale à 72 ms [écart-type = 4 ms] en rapide pour les brèves, et de 246 ms [écart-type = 4 ms] en normale à 169 ms [écart-type = 3 ms] en rapide pour les longues). .

3.3. Résultats en valeurs relatives

Il convient toutefois de vérifier la robustesse de ces données, en se focalisant, comme il se doit, sur l'analyse des *valeurs relatives* du paramètre de prédilection de la gémination, à savoir la tenue consonantique. En effet, les résultats montrent que, nonobstant la compression de la durée des tenues consonantiques, la proportion de temps prise par les géminées dans la syllabe CV2 est toujours supérieure à celle des simples [(df=1,238, F=108.2184, p<0.0000)], quelle que soit la vitesse d'élocution [(df=1,238, F=228.5323, p<0.0004)]. La Figure 2 illustre ce résultat. On voit que la géminée /tt/ occupe 67% (écart-type = 4%) de la syllabe CV2 en vitesse d'élocution normale, alors que la simple /t/ n'en prend que 44% (écart-type = 3%). Les proportions restent relativement stables en vitesse d'élocution rapide, puisque ce sont des proportions comparables que l'on retrouve dans cette condition d'élocution, à savoir 60% (écart-type = 5%) pour les géminées *vs*. 41% (écart-type = 4%) pour les simples. Signalons, cependant, que le contraste était plus difficilement maintenu dans le contexte uvulaire (/q/ *vs*. /qq/), en vitesse d'élocution rapide, pour le locuteur K (p=*ns*). Nous renvoyons à Ridouane [6] pour le comportement particulier des uvulaires en berbere tashlhiyt, aussi bien du point de vue phonétique que phonologique. Il arrive parfois que le pourcentage de la tenue consonantique dans la syllabe CV2 augmente (*sic*) avec l'augmentation de la vitesse d'élocution. C'est ce que nous avons observé pour nos deux locuteurs dans les contextes vélaire et uvulaire. La Figure 3 montre que la vélaire simple /k/ passe de 50% (écart-type = 2%) de la syllabe en vitesse d'élocution normale à 58% (écart-type = 5%) en vitesse d'élocution rapide. Corollairement, son homologue géminée augmente sa proportion de 62% (écart-type = 3%) en vitesse d'élocution normale à 69% (écart-type = 7%) en rapide. Il s'agit ici, comme l'a déjà montré Sock [8] pour diverses langues à quantité, d'une stratégie de préservation de l'opposition de quantité. En effet, la consonne simple semble obéir à une contrainte de non-réduction de la proportion de la phase consonantique, sous peine d'augmentation du taux de voisement provenant, soit d'une valeur intrinsèque trop réduite, soit de l'effet des voyelles adjacentes. Elle avait le « choix »

de maintenir sa tenue relativement stable, ou de l'augmenter. C'est cette dernière option qui sera adoptée. Cependant une telle trajectoire a le risque potentiel de provoquer une confusion entre la classe des simples et celle des géminées. Cette dernière classe augmente, en conséquence, sa proportion dans la syllabe, ce qui non seulement ne représente aucun danger idiosyncrasique pour son identité, mais a aussi l'effet de préserver l'opposition phonologique. Cette stratégie est visible aussi en contexte uvulaire (Figure 4), où la simple passe de 50% (écart-type = 6%) de la syllabe en vitesse d'élocution normale, à 63% (écart-type = 8%) en vitesse rapide. Conséquemment, la géminée augmente la proportion de sa tenue consonantique, allant de 63% de la syllabe à 80% de celle-ci.

3.4. Relations articulatori-acoustiques

Si nous tentons de relier ces résultats acoustiques aux données articulatoires attestées dans la littérature, et à nos propres résultats obtenus par cinéradiographie, nous pourrons faire les constats suivants : 1) Les durées consonantiques plus longues pour les géminées semblent correspondre aux étendues de contact plus importantes que nous avions trouvées (Bouarourou *et al.*, [2]) pour cette catégorie de consonnes (et *vice versa* pour les simples) ; cela laisse penser qu'il devrait y avoir une certaine relation entre ces deux paramètres. 2) En effet, les résultats trouvés par Löfqvist et Gracco [5], semblent corroborer nos hypothèses articulatori-acoustiques ; ils montrent que la vélocité de la langue est plus élevée pour les géminées par rapport aux simples, lors de la tenue consonantique. 3) L'augmentation de la proportion de la phase consonantique acoustique avec l'augmentation de la vitesse d'élocution pourrait s'expliquer sur le plan articulatoire suivant les résultats obtenus par Vaxelaire [10], et les hypothèses qu'elle en formule. En vitesse d'élocution rapide, afin de maintenir les différences linguistiques de durée, le locuteur doit diminuer les temps de transition, ce qui provoquerait une augmentation de la force d'exécution du geste, le résultat étant, parfois, une augmentation de l'écrasement de la langue contre le palais. De telles hypothèses appellent, bien sûr, des études supplémentaires, portant notamment sur la cinématique de la langue et sur la pression d'air orale durant la production des occlusives analysées dans cette étude.

4. CONCLUSIONS

Des données acoustiques pour les occlusives non voisées simples et géminées du tarifit ont été analysées et ont confirmé une mesure pertinente pour la gémination : la durée de la tenue consonantique. Les autres paramètres acoustiques retenus n'ont pas révélé de comportement différent selon que l'on ait affaire aux simples ou aux géminées. L'augmentation de la vitesse d'élocution a provoqué la compression des segments acoustiques mesurés, y compris celle du paramètre pertinent de la gémination : la tenue consonantique. Malgré le réaménagement temporel de ce paramètre critique, la distinction des deux classes reste possible par la tenue consonantique, aussi bien en termes absolus que relatifs. Cela semble démontrer la résistivité de la gémination en tarifit. Nos avons tenté d'établir des liens articulatori-acoustiques en suggérant des corrélations entre la tenue consonantique acoustique et l'étendue de contact articulatoire, un autre paramètre robuste, puisqu'il nous avait permis (Bouarourou *et al*, [2]) de distinguer les deux catégories linguistiques à travers différents contextes consonantiques, et dans différentes positions à l'intérieur du mot, et cela chez nos deux sujets. Certaines de nos hypothèses appellent, bien entendu, des études cinématiques supplémentaires. Des tests de perception devraient également être conduits afin d'évaluer l'impact des différences de durées consonantiques sur la discrimination des deux classes phonologiques.

BIBLIOGRAPHIE

[1] C. Abry, J-P. Orliaguet & R. Sock. Patterns of Speech Phasing. Their Robustness in the Production of a Timed Linguistic Task : Single vs. Double (Abutted) Consonants in French. *European Bulletin of Cognitive Psychology* volume 10, pages 269-288, 1990.

[2] F. Bouarourou, B. Vaxelaire, R. Ridouane, F. Hirsch and R. Sock. Gemination in Tarifit Berber: X-ray and acoustic data. *In Proceedings of the 8th International Seminar on Speech Production*, pages 117-120, 2008.

[3] J. Gaitenby. The elastic word. *Haskins Laboratories, Status Report, Speech Research* 2, 1-12, 1965.

[4] I.Lehiste,M. Tatham & K.Morton. . An instrumental study of consonant gemination. *Journal of Phonetics* 1, 131-148, 1973.

[5] A. Lövqvist, and V. Gracco. Lip and jaw kinematics in bilabial stop consonant production, *Journal of Speech and Language Hearing Research* 40, 877-893, 1997.

[6] R. Ridouane. Suite des consonnes en berbère: phonétique et phonologie. *Thèse de Doctorat, Université Sorbonne-Nouvelle*, 2003.

[7] R. Ridouane. Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *Journal of the International Phonetic Association* 37 (2), 119-142, 2007.

[8] R. SOCK (1999) Organisation temporelle en production de la parole. Emergence de catégories sensori-motrices phonétiques. *Presses Universitaires du Septentrion. Villeneuve d'Ascq*, 479 pages, 1999.

[9] B. Vaxelaire, 1995. Single *vs*. double (abutted) consonants across speech rate. X-ray and acoustic data for French. *Proceedings of the 13th International Conference on Phonetic Sciences* volume 1, pages 384 – 387, 1995.

[10] B.Vaxelaire. Le geste et la production de la parole. Résultats et implications d'études quantitatives cinéradiographiques. Habilitation *à Diriger des Recherches (HDR)*. Université Marc Bloch – Strasbourg 2, 133 pages, 2007.

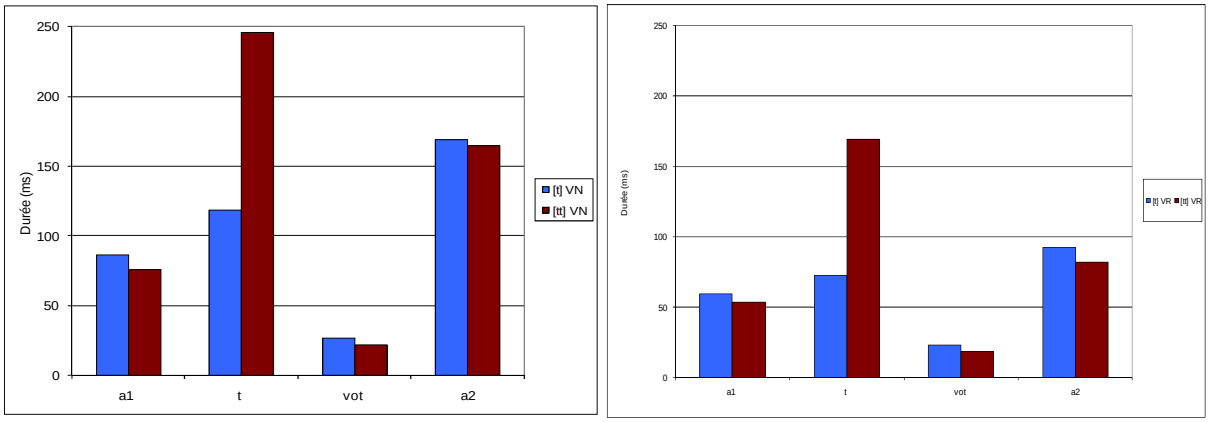


Figure 1 : A gauche, valeurs moyennes (ms), en vitesse d'élocution normale, indiquant les effets de la gémination sur les paramètres acoustiques étudiés (a1 = durée de la voyelle précédant la consonne, t = durée de la consonne alvéolaire /t/, vot = durée du délai d'établissement du voisement, a2 = durée vocalique de la voyelle suivant la consonne). A droite, mêmes indications, en vitesse d'élocution rapide. Locuteur F.

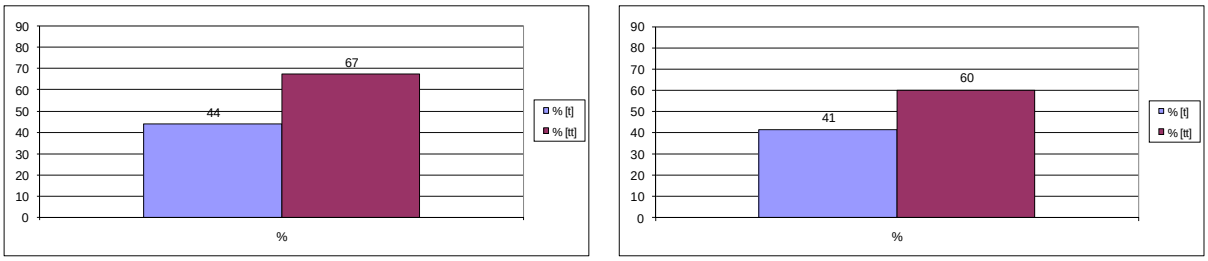


Figure 2 : A gauche, valeurs moyennes (%), en vitesse d'élocution normale, indiquant les effets de la gémination sur le paramètre acoustique principal : la durée d'occlusion alvéolaire /t/ *vs.* /tt/. A droite, mêmes indications, en vitesse d'élocution rapide. Locuteur F.

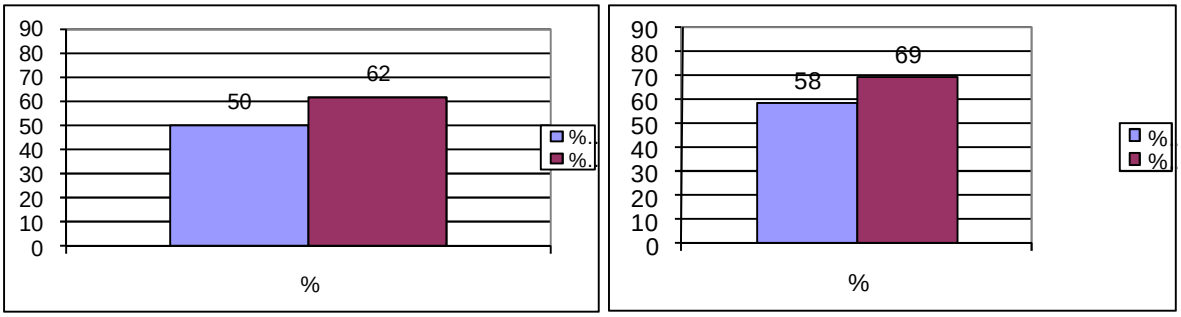


Figure 3 : A gauche, valeurs moyennes (%), en vitesse d'élocution normale, indiquant les effets de la gémination sur le paramètre acoustique principal : la durée d'occlusion vélaire /k/ *vs.* /kk/. A droite, mêmes indications, en vitesse d'élocution rapide. Locuteur K.

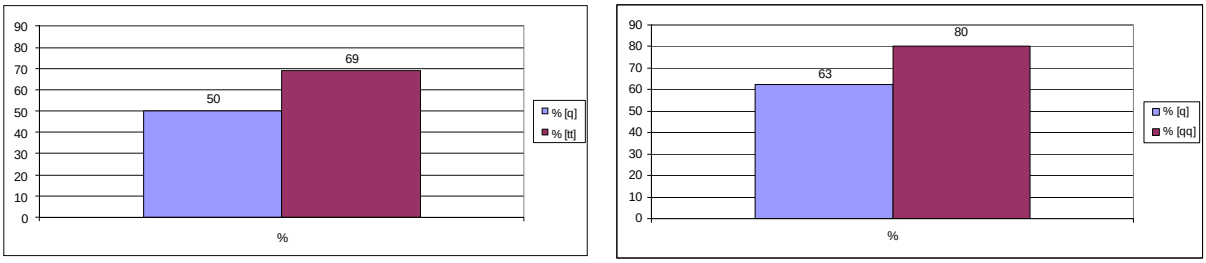


Figure 4 : A gauche, valeurs moyennes (%), en vitesse d'élocution normale, indiquant les effets de la gémination sur le paramètre acoustique principal : la durée d'occlusion uvulaire /q/ *vs.* /qq/. A droite, mêmes indications, en vitesse d'élocution rapide. Locuteur F.

Remerciements : Ce travail de recherche a été soutenu par une subvention de l'ANR (ANR-07-CORP-018-01, DOCVACIM), 2007 - 2011 et un programme MISHA (Perturbations et réajustements : parole normale *vs.* parole pathologique), 2008 – 2012.

Réordonnancement automatique d'hypothèses pour l'assistance à la transcription de la parole

Antoine Laurent[†§], *Sylvain Meignier*[†], *Paul Deléglise*[†]

LIUM[†] – Laboratoire d'Informatique de l'Université du Maine – Le Mans
prenom.nom@lium.univ-lemans.fr

Spécinov[§] – Trélazé
a.laurent@specinov.fr

ABSTRACT

Large vocabulary automatic speech recognition (ASR) technologies perform well in known, controlled contexts. However, some mistakes still have to be corrected. Human intervention is necessary to check and correct the results of such systems in order to make the output of ASR understandable. We propose a method for computer-assisted transcription of speech, based on automatic reordering confusion networks. It allows to significantly reduce the number of actions needed to correct the ASR outputs. WER computed before and after every network reordering shows an absolute gain of about 3.4%.

Keywords: Speech recognition, Automatic correction, Cache models, Confusion network

1. Introduction

Cet article présente une méthode d'assistance à la transcription automatique de la parole. Le transcripteur humain dispose de la meilleure hypothèse fournie par un système de reconnaissance automatique de la parole (SRAP). A chaque correction de sa part, le système propose une nouvelle transcription prenant en compte cette correction. Cette méthode permet au système et au correcteur de collaborer pour converger plus rapidement vers une transcription correcte (sans erreur).

Dans la littérature, peu d'articles sont consacrés à cette tâche. Dans l'article [9], les auteurs proposent de relancer le processus de décodage après chaque correction de l'utilisateur en se basant sur celle-ci. L'inconvénient majeur de cette méthode est qu'elle risque d'avoir un impact négatif sur le temps de réaction de l'interface homme machine.

D'autres travaux [3, 10, 6] ont été réalisés sur la traduction assistée par ordinateur (Computer-Assisted Translation - CAT). Ces types de systèmes proposent une traduction qui est lue par l'utilisateur. Dès qu'un mot erroné est corrigé, le système propose une traduction alternative. Certains travaux [1, 11, 2] présentent des méthodes de traduction assistées utilisant en entrée, non pas du texte, mais du langage parlé. L'idée générale consiste à utiliser un modèle de langage combinant un modèle de langage n-gramme avec les probabilités de traduction de chaque mot.

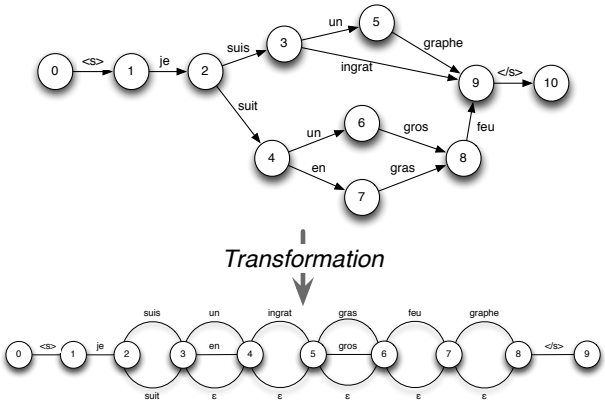


Fig. 1: Obtention d'un réseau de confusion

2. Méthode

La méthode proposée s'inspire des méthodes développées pour la tâche de CAT. Elle consiste à réévaluer la meilleure hypothèse d'un réseau de confusion en fonction des corrections apportées par l'utilisateur, sans nécessiter un décodage complet.

2.1. Réseau de confusion

Le réseau de confusion est obtenu à partir des treillis de mots construits lors du décodage. Les treillis de mots représentent, après élagage, tous les chemins hypothèses développés. Chaque état correspond à un instant dans l'enregistrement à transcrire, chaque lien représente un mot auquel est associé une probabilité.

Ce graphe est transformé en réseau de confusion (figure 1). La transformation consiste à fusionner les mots localement identiques, à regrouper sur des ensembles de confusion communs les mots temporellement proches, et à supprimer les chemins d'hypothèse trop faibles [8]. Chaque mot obtient un nouveau score qui est sa probabilité *a posteriori* (obtenue à partir du treillis de mots) divisée par la somme des probabilités *a posteriori* des mots en concurrence avec lui. Le symbole ϵ représente une transition vide (absence de mot).

Les réseaux de confusion contiennent, en plus de l'ensemble des mots en concurrence, des temps approximatifs obtenus à partir du graphe de mot. Cet ajout est une extension par rapport à [8]. L'instant de départ d'un état du réseau de confusion correspond au plus petit instant des mots associés à cet état ; le

temps de fin correspond à l'instant de fin du dernier de ces mots. La meilleure hypothèse de reconnaissance qui sera donnée à corriger au transcripteur est la séquence de mots qui maximise chacune des probabilités *a posteriori* du réseau de confusion.

2.2. Principe

A partir d'une séquence d'observations acoustiques X, l'objectif du SRAP est de trouver la séquence de mots $\hat{W}$ la plus probable parmi l'ensemble des séquences possibles W. La recherche de $\hat{W}$ maximisant la probabilité d'émission de W sachant X correspond à l'équation suivante après application du théorème de Bayes et simplification :

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \quad (1)$$

$P(W)$ est fourni par le modèle de langage, la probabilité $P(X|W)$ correspond à la probabilité attribuée par le modèle acoustique. Dans notre cas, la séquence de mots possibles W est séparée en deux : un préfixe p qui a été validé et/ou corrigé par l'utilisateur et un suffixe s à déterminer en fonction du préfixe p. Nous cherchons donc la séquence de mots $\hat{s}$, parmi toutes les séquences de suffixes possibles s maximisant l'équation suivante :

$$\hat{s} = \arg\max_{s} P(X|s,p)P(s|p) \quad (2)$$

La méthode présentée ne remet pas en cause l'acoustique, $P(X|s,p)$ est constant. $P(s|p)$ sera calculé à partir de la combinaison linéaire entre un modèle de langage quadrigramme P_{4G} et un modèle cache P_{cache} [4]. Nous cherchons donc, parmi tous les suffixes candidats $\tilde{s}$ dans le réseau de confusion, le suffixe $\hat{s}$ maximisant la probabilité suivante :

$$\hat{s} = \arg\max_{\tilde{s}}((1-\lambda)P_{4G}(s|p) + \lambda P_{cache}(s|p)) \quad (3)$$

Le modèle cache est construit à partir des mots contenus dans le préfixe p. Il permet de renforcer la probabilité des mots récemment rencontrés, on suppose ici qu'ils ont une plus forte chance d'apparaître dans le futur (dans s).

Plusieurs types de modèles caches sont proposés dans la littérature. Dans l'application proposée, les meilleurs résultats en terme de perplexité ont été obtenus à partir de la méthode proposée dans [4]. La probabilité d'apparition du mot w_i est exponentiellement proportionnelle à la distance entre la position actuelle et les apparitions précédentes du mot w_i dans l'historique h_i :

$$P_{cache}(w_i|h_i) = \beta \sum_{j=1}^{i-1} I_{\{w_i=w_j\}} e^{-\alpha(i-j)} \quad (4)$$

avec α le coût du décalage dans le cache, $I_{w_i=w_j} = 1$ si $w_i = w_j$ et 0 sinon, et β est une constante de normalisation calculée comme suit :

$$\beta = \frac{1}{\sum_{j=1}^{i-1} e^{-\alpha j}} \quad (5)$$

2.3. Application

A partir du réseau de confusion et du préfixe corrigé et/ou validé, la recherche des suffixes candidats dans le réseau de confusion est effectuée de la manière suivante. Soit t l'instant de fin du dernier mot validé. Le principe va être de rechercher, parmi tous les états suivants du réseau de confusion ceux ayant un instant de début supérieur ou égal à t. La recherche est récursive. Pour chacun des états concurrents, nous recherchons à nouveau les états pouvant lui succéder et ainsi de suite. L'utilisation de ces temps, bien qu'approximatifs, permet d'éviter de choisir des séquences de mots dans lesquelles certains mots se chevaucheraient. La méthode proposée ne se déclenche que lorsque l'utilisateur remplace un mot par un autre (erreur de substitution).

Le réordonnancement automatique s'arrête dès que le dernier mot proposé correspond à un mot qui était présent dans l'hypothèse initiale. Si l'utilisateur fait le choix de supprimer un mot (mot incorrect glissé entre deux mots corrects) ou d'en insérer un (mot manquant entre deux mots corrects) plutôt que de faire une correction (substitution), nous avons fait l'hypothèse que le mot suivant était juste.

La première étape va consister à rechercher à quel état du réseau de confusion le mot venant d'être corrigé peut être rattaché. Si ce mot est présent dans le graphe à l'endroit de la correction, il sera rattaché à l'état correspondant, s'il n'est pas présent, le mot sera ajouté à l'état du mot substitué. Cette première étape réalisée, toutes les séquences possibles de mots pouvant succéder à l'état sélectionné du graphe sont recherchées, en respectant les indices temporels.

Le score calculé grâce à l'équation (3) permet de départager les différentes séquences possibles. Si deux séquences de mots ont la même probabilité (cas exceptionnel), la probabilité *a posteriori* moyenne de la séquence de mots devient l'élément discriminant. Cette stratégie d'auto-réordonnancement s'arrête dès que le dernier mot proposé automatiquement est identique à celui qui était dans l'hypothèse précédente (figure 2).

2.4. Cas des mots hors vocabulaire

La liste de mots pouvant être proposée de façon automatique est limitée aux mots se trouvant dans le réseau de confusion. Si l'utilisateur saisit un mot inconnu du SRAP dans le préfixe, la recherche du suffixe parmi les différents suffixes candidats fera intervenir la probabilité du mot inconnu. Ce mot sera ajouté au modèle cache, mais il n'est pas possible, en l'état actuel, que ce mot nouveau réapparaisse de façon automatique dans la suite des corrections.

3. Expériences

Les expériences menées simulent le comportement d'un transcripteur corrigeant la meilleure hypothèse du réseau de confusion généré par le SRAP.

3.1. Corpus & SRAP

L'optimisation des coefficients du modèle cache a été réalisée sur le corpus de test d'ESTER 1. Les expériences ont été effectuées sur le corpus de test d'ESTER 2 [7]. Il s'agit d'émissions radiophoniques francophones, complétées par des articles provenant du journal "Le Monde". Les réseaux de confusion sont créés à partir du graphe d'hypothèse des mots généré par le SRAP du LIUM développé lors de la campagne ESTER 2. Le décodage s'effectue en 5 passes détaillées dans l'article [5]. Sans l'ajout de traitements particuliers pour les segments provenant de la radio africaine, le taux d'erreur mot du système sur le corpus de test de la campagne d'ESTER 2 est de 19,2%.

3.2. Métriques

La méthode est évaluée selon deux métriques. Le taux d'erreur mot classique (WER) et le KSR (*Keystroke Saving Rate*) [12].

Le KSR a été mis en place dans les systèmes de communication assistés destinés aux handicapés. Il se calcule de la façon suivante :

$$KSR = (1 - \frac{k_p}{k_a}) \times 100 \qquad (6)$$

Où k_p est le nombre d'appuis effectivement réalisés par l'utilisateur lors de la saisie d'un message et k_a le nombre d'appuis qui auraient été nécessaires sans aide à la composition de mots. Ces appuis peuvent être des appuis sur un clavier ou sur un dispositif particulier mis en place pour la gestion de l'handicap de l'utilisateur : joystick, clignement d'un oeil, etc.

Dans notre cas, k_p représentera le nombre d'actions réalisées par l'utilisateur pour corriger l'hypothèse du SRAP, en utilisant un clavier, et k_a le nombre d'actions qui auraient été nécessaires en partant d'une hypothèse vide (ne contenant pas de mots).

Pour calculer le KSR, il est supposé que l'utilisateur appuiera sur le moins de touches possible pour obtenir la transcription corrigée. Deux stratégies sont retenues pour minimiser le nombre d'actions : soit tous les mots de la zone erronée sont supprimés puis remplacés par les mots corrects, soit le maximum de lettres de l'hypothèse sont conservées et les actions ne portent que sur les lettres erronées.

Un alignement entre l'hypothèse générée par le système de reconnaissance automatique de la parole et la référence de transcription est effectué. Cet alignement sera réalisé au niveau des mots et au niveau des lettres correspondant aux zones en erreur.

Les coûts des actions sont les suivants :
- Les coûts de déplacement à l'intérieur du texte de mot en mot ne sont pas pris en compte. L'application d'aide à la correction, comme celle de [9], présente les mots un à un lors du procédé d'aide à la transcription, et chaque mot est corrigé lors de son apparition.
- La suppression d'un mot coûte 1 (raccourci clavier permettant de supprimer un mot entier).
- L'appui sur une touche du clavier coûte une action.

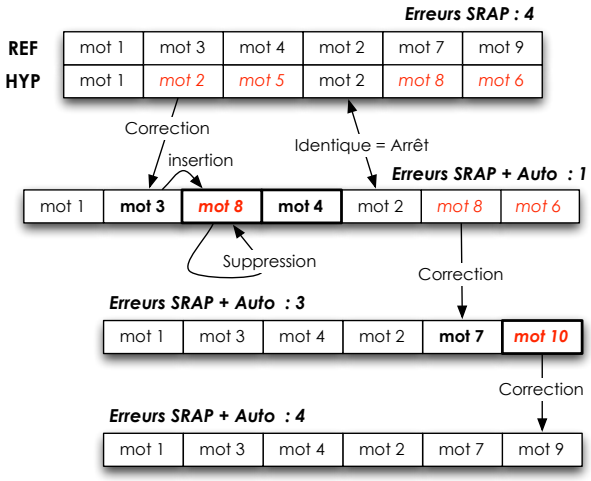


Fig. 2: Calcul du taux d'erreur mot de la méthode de réordonnancement automatique

Cette méthode de calcul étant relativement subjective, puisqu'elle ne prend pas en compte la pénibilité de chaque action pour l'utilisateur, la méthode a également été évaluée en terme de WER. Le nombre d'erreurs (Insertion + Substitution + Suppression) a été calculé pour chaque segment avant et après chaque procédure de correction automatique. Quand la méthode automatique ne se déclenche pas, son coût est considéré identique à celui de la méthode manuelle (voir figure 2).

La figure 2 présente un cas où la méthode de réordonnancement ne permet pas d'observer de gain en terme de taux d'erreur mot.

3.3. Résultats

Dans un premier temps, le nombre d'actions à réaliser par l'utilisateur a été calculé en ne lui proposant que la sortie du système de reconnaissance automatique à corriger, sans autre aide. La possibilité de remplacer un mot par un autre par simple sélection d'un mot concurrent dans une liste de mots a ensuite été évaluée. La liste est constituée des mots présents dans les réseaux de confusion à cet instant. Nous supposons que la liste des concurrents est toujours visible à l'écran, un coût de 1 est attribué pour le remplacement d'un mot de cette manière.

Le tableau 1 présente un résumé des résultats obtenus dans différentes configurations. La référence est composée de 435 005 lettres (caractères espace compris). Il faut donc appuyer 435 005 fois sur les touches du clavier pour la saisir dans sa totalité (Manuelle dans le tableau). En utilisant les sorties du système de reconnaissance automatique de la parole (ligne SRAP), ce nombre d'actions est de 53 492. L'utilisation du système de reconnaissance automatique de la parole a donc permis de réaliser un gain en terme de *KSR* de 87,7%, pour un taux d'erreur mot de 19,2%. Avec les listes déroulantes (ligne SRAP+liste), le nombre d'actions à effectuer par l'utilisateur diminue et passe à 52 003, soit un *KSR* de 88%. Les sorties du SRAP associées à la mise en oeuvre de la méthode de réordonnancement automatique permet d'observer un *KSR* de 89,2% (46 795 actions), pour un taux d'er-

reur mot de 17% (ligne SRAP+auto). L'ajout de la sélection des mots dans une liste déroulante permet de diminuer ce nombre d'actions à 44 732, soit un *KSR* de 89,7% (SRAP+auto+liste). Enfin, le système complet (SRAP+auto+liste+cache) utilisant le modèle cache, la technique de réordonnancement automatique et la sélection des mots dans la liste déroulante permet d'observer un *KSR* de 90,3% (41 992 actions), et un WER de 15,8%.

Tab. 1: *KSR et WER sur le corpus de test ESTER 2*

Méthode	Nb actions	KSR	WER
Manuelle	435005	0%	–
SRAP	53492	87,7%	19,2%
SRAP+liste	52003	88,0%	19,2%
SRAP+auto	46795	89,2%	17,0%
SRAP+auto+liste	44732	89,7%	17,0%
SRAP+auto+liste+cache	41992	90,3%	15,8%

En terme de taux d'erreur mot, la méthode proposée permet d'obtenir un gain d'environ 3,4% sur le corpus de test d'ESTER 2.

Il est à noter que lorsque le WER diminue, le nombre d'actions à réaliser suit la même tendance. En effet, le WER passe de 19,2% à 15,8%, soit un gain relatif de 17,7%. Le nombre d'actions, quant à lui, chute de 53 492 à 41 992, soit un gain relatif de 21,5%. Lorsque l'on compare le nombre d'actions nécessaires à la correction de la sortie du SRAP seul, avec l'utilisation du SRAP complété de la méthode de réordonnancement automatique, là encore, le nombre d'actions et le WER diminuent tous les deux : 12,5% de gain relatif en terme de nombre d'actions et 11,4% de gain relatif en terme de WER.

4. Conclusion

Cet article présente une technique de réordonnancement automatique des hypothèses du SRAP. Cette technique permet d'observer un gain de 3,4% absolu en terme de WER et de diminuer le nombre d'actions de 21,5% (relatif) par rapport à l'utilisation seule des sorties du système de reconnaissance automatique de la parole. Certaines améliorations peuvent encore être apportées à cette méthode, puisqu'elle ne permet pas, pour l'instant, l'ajout automatique de nouveaux mots (*ie* de mots qui ne seraient pas présents dans le réseau de confusion). La méthode pourrait également être améliorée en propageant les corrections apportées par l'utilisateur. Pour l'instant, la correction d'un mot déclenche le réordonnancement automatique des hypothèses du réseau de confusion pour la fin du segment en cours de correction. Cette correction pourrait avoir un impact sur d'autres segments se trouvant plus éloignés dans la transcription.

Références

[1] J. C. Amengual, J. M. Benedí, A. Castaño, A. Castellanos, V. M. Jiménez, D. Llorens, A. Marzal, M. Pastor, F. Prat, E. Vidal, and J.M. Vilar. The EuTrans-I speech translation system. *Machine Translation*, 14(3) :941–951, 2000.

[2] F. Casacuberta, E. Vidal, A. Sanchis, and J.M. Vilar. Pattern recognition approaches for speech-to-speech translation. *Cybernetic and Systems : an International Journal*, 35(1) :3–17, 2004.

[3] J. Civera, J.M. Vilar, E. Cubel, A.L. Lagarda, S. Barrachina, F. Casacuberta, and E. Vidal. A novel approach to computer assisted translation based on finite-state transducters. *Proceedings of Finite-State Methods and Natural Language Processing (FSMNLP)*, 4002 :32–42, 2006.

[4] P.R. Clarkson and A. J. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 97)*, pages 799–802, 1997.

[5] P. Deléglise, Y. Estève, and S. Meignier. Improvements to the lium french asr system based on cmu sphinx : what helps to significantly reduce the word error rate ? In *Proceedings of International Conference on Spoken Language Processing (ISCA, Interspeech 2009)*, Brighton, UK, 2009.

[6] G. Foster. *Text Prediction for Translators*. PhD thesis, Université de Montréal, 2002.

[7] S. Galliano, G. Gravier, and L. Chaubard. The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *Proceedings of International Conference on Spoken Language Processing (ISCA, Interspeech 2009)*, Brighton, UK, September 2009.

[8] H. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition : Word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4) :373–400, 2000.

[9] L. Rodríguez, F. Casacuberta, and E. Vidal. Computer Assisted Transcription of Speech. *Lecture Notes In Computer Science, Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I*, 4477 :241–248, 2007.

[10] J. Tomás and F. Casacuberta. Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of Coling/Association for Computational Linguistics*, pages 835–841, Sydney, Australia, 2006.

[11] E. Vidal, F. Casacuberta, L. Rodríduez, J. Civera, and C. Martnez. Computer-assisted translation using speech recognition. *IEEE Transaction on Audio, Speech and Language Processing*, 14(3) :941–951, 2006.

[12] M. Wood and E. Lewis. Windmill - the use of a parsing algorithm to produce predictions for disabled persons. In *Proceedings of the 1996 Autumn Conference on Speech and Hearing*, pages 315–322, 1996.

Simulation du processus de croyance mutuelle de la compréhension dans le dialogue (*grounding process*) à l'aide des réseaux bayésiens

Stéphane Rossignol, Olivier Pietquin, Michel Ianotto

Équipe IMS
Supélec – Campus de Metz, 2, rue Édouard Belin ; 57070 Metz, France
stephane.rossignol@supelec.fr
http ://ims.metz.supelec.fr/

ABSTRACT

User simulation has become an important trend of research in the field of spoken dialogue systems because collecting and annotating real interactions with users is often expensive and time consuming. Yet, such data are generally required for designing and assessing efficient dialogue systems. The general problem of user simulation is thus to produce as many as necessary natural, various and consistent interactions from as few data as possible. In this paper, we propose a user simulation method based on Bayesian Networks (BN) that is able to produce consistent interactions in terms of user goal and dialogue history but also to simulate the grounding process that often appears in human-human interactions. The BN is trained on a database of 1234 human-machine dialogues in the TownInfo domain (a tourist information application). Experiments with a state-of-the-art dialogue system (REALL-DUDE/DIPPER/OAA) have been realized and promising results are presented.
Keywords: Bayesian network, spoken dialogue systems, grounding process

1. Introduction

Les systèmes de dialogue oral sont à présent très répandus et sont utilisés dans un grand nombre de domaines (de la réservation de vols à l'aide en ligne à la réparation). Mettre en place de telles interfaces basées sur la parole est habituellement un processus itératif qui comprend plusieurs cycles de prototypage, de tests et de validation. Les tests et les validations requièrent des interactions entre le système courant et des utilisateurs humains, ce qui rend ces phases chères et très gourmandes en temps. Pour cette raison, la simulation d'utilisateurs est devenue, depuis une dizaine d'années, un très important domaine de recherche. La simulation d'utilisateurs peut dès lors être utilisée pour tester les performances d'un système et pour optimiser celui-ci [4, 7]. Il peut s'agir par exemple d'optimiser la politique du gestionnaire de dialogue (GD) à l'aide de méthodes d'apprentissage par renforcement. La simulation d'utilisateurs ne doit pas être confondue avec la modélisation d'utilisateurs. Dans un système de dialogue, cette dernière est utilisée en général dans des buts internes, comme par exemple pour représenter la connaissance qu'a l'utilisateur de l'état d'avancement du dialogue, ou pour inférer le but de l'utilisateur [2, 5], ou encore pour simuler le comportement du module de compréhension [8] de la parole. Par contre, l'objectif de la simulation d'utilisateurs est la création d'un grand nombre d'interactions simulées avec un système de dialogue, et de ce fait l'utilisateur peut être considéré comme étant extérieur au système.

L'utilisateur simulé (US) présenté dans cet article est basé sur les réseaux bayésiens (RB). Ce modèle a été choisi pour plusieurs raisons. Premièrement, les RB sont des modèles génératifs et peuvent donc être utilisés aussi bien pour inférer que pour générer des données, ce qui est bien sûr requis pour la simulation. Deuxièmement, il s'agit d'un paradigme statistique, qui peut donc générer une grande variété de dialogues cohérents statistiquement. Troisièmement, les paramètres des RB peuvent être soit fixés par des experts, soit entraînés à partir d'une base de données. Étant donné que la collecte de données se révèle souvent un processus difficile, l'apport d'un expert peut être très utile. Pour finir, il existe, pour effectuer l'inférence dans les RB et l'entraînement des RB, un grand nombre d'outils libres.

Cet article se fonde sur des travaux précédents [7] où les RB sont utilisés pour simuler des utilisateurs, mais il met l'accent sur deux contributions nouvelles. Tout d'abord, le modèle a été modifié pour générer des problèmes de différence de croyance mutuelle de la compréhension dans le dialogue (*grounding process*). Ce processus de *grounding* est considéré ici comme étant le processus utilisé par deux locuteurs d'une même conversation pour s'assurer qu'ils partagent la même connaissance de l'état d'avancement du dialogue. En pratique, ceci veut dire que l'utilisateur simulé réagira automatiquement si un problème de transmission de l'information est détecté, en donnant de nouveau une information correcte au système. Le but principal de ce travail est d'entraîner des politiques de GD qui prennent en compte de tels problèmes de *grounding* [6]. Ensuite, le modèle est entraîné avec des données réelles concernant des interactions homme-machine ; puis il est testé en combinaison avec un système de dialogue de l'état de l'art (l'environnement REALL-DUDE/DIPPER/OAA [3, 1]).

La tâche considérée se place dans le domaine du TownInfo. Elle concerne l'aide aux touristes. La tâche consiste à fournir des informations concernant des restaurants dans une ville donnée. Ceci peut être considéré comme étant une tâche de remplissage d'un formulaire à trous, où les trois différents attributs requérant d'être informés sont : le type de nourriture,

la gamme de prix et l'emplacement dans la ville. Ces attributs peuvent respectivement prendre 3, 3 et 5 valeurs.

2. Description du modèle

2.1. *Les réseaux bayésiens*

Le cœur du réseau bayésien utilisé ici est décrit dans [7]. Les lecteurs sont invités à se référer à ce travail pour des compléments théoriques. À la base, ce RB s'inspire de l'idée que la réponse de l'utilisateur à une action du système est influencée par le type du message reçu (AS), par le but de l'utilisateur (GOAL) et par sa connaissance (KNW) concernant l'état d'avancement du dialogue. La réponse de l'utilisateur peut être de différents types : INFORM, CONFIRM, et une action spéciale consistant à vouloir clore le dialogue. La figure 1 montre en détails la structure de l'US. Pour des raisons de portabilité et pour être à même d'intégrer cet US dans le système de dialogue visé, il a été implanté en C++, en utilisant la librairie Smile (http ://genie.sis.pitt.edu/about.html#smile). Smile est une librairie de classes C++ qui mettent en place des modèles de graphes probabilistes et de théorie de la décision, comme les RB, etc.

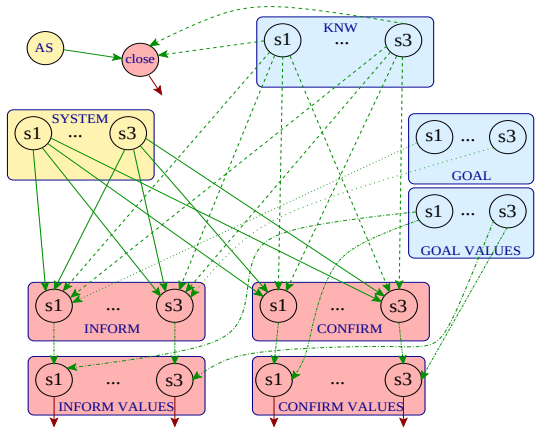


Fig. 1: Utilisateur simulé à base de réseaux bayésiens

Le nœud AS correspond à l'action du système (INFORM, CONFIRM, REQUEST...). Les nœuds SYSTEM correspondent aux attributs concernés par l'action du système. Ces nœuds et les attributs qui leur sont associés sont des variables indépendantes ; de ce fait, ces nœuds n'ont pas de parents. Ils sont, de manière équivalente, les variables de sortie du gestionnaire de dialogue, ou les variables d'entrée de l'US.

Les nœuds KNW correspondent à la connaissance qu'a l'US concernant l'historique du dialogue. Une valeur pour cette connaissance est établie pour chacun des attributs dans la tâche (3 dans notre cas). Trois niveaux de connaissance (soit 3 valeurs possibles) sont considérés : bas, moyen et élevé. Ces valeurs représentent la connaissance qu'a l'utilisateur concernant le fait que l'information à propos de l'attribut correspondant a été donné au système. Par exemple, si l'utilisateur a donné une fois l'information concernant le type de nourriture qu'il souhaite, la connaissance attribuée à cet attribut passe de *bas* à *moyen*. Si cette information a été donnée plusieurs fois, la connaissance devient *élevée*. Il peut être considéré que la connaissance correspond à une estimation de l'état d'avancement du dialogue, du point de vue de l'US.

Les nœuds GOAL correspondent au but de l'utilisateur. Ils indiquent quels attributs sont présents dans son but. Les nœuds GOAL VALUES indiquent la valeur attribuée à chaque attribut présent dans le but. Ceci assure que le comportement de l'US sera cohérent, relativement à un but donné.

Les nœuds INFORM et INFORM VALUES correspondent à l'action "inform" de l'utilisateur et à ses attributs. Les nœuds INFORM correspondent aux attributs transmis dans le message de l'utilisateur, et les nœuds INFORM VALUES correspondent aux valeurs associées à ces attributs transmis. Les nœuds CONFIRM et CONFIRM VALUES correspondent à l'action "confirm" ou "negate" de l'utilisateur. Les nœuds CONFIRM correspondent aux attributs transmis dans le message de l'utilisateur, et les nœuds CONFIRM VALUES correspondent aux valeurs (c'est-à-dire "oui" ou "non") associées aux attributs transmis. Le nœud CLOSE indique si l'US décide ou non de clore le dialogue. Ces nœuds sont les variables de sortie de l'US.

Il faut remarquer que si le GD réclame (action REQUEST) des informations pour un attribut pour lequel l'US a déjà une connaissance moyenne ou élevée (c'est-à-dire un attribut pour lequel l'US a déjà donné de l'information), il est probable qu'un problème de *grounding* soit survenu. L'US décrit jusque à présent est mis en place de telle sorte qu'il envoie une valeur pour cet attribut, avec une certaine probabilité, qui peut être petite (ou, possiblement, l'US désirera clore le dialogue). Cependant, la valeur de la connaissance pourrait être utilisée pour inférer la survenue d'un problème de *grounding*. Ceci est fait en ajoutant la composante de *grounding* montrée sur la figure 2, où les nœuds KNW et GOAL sont les variables internes de l'US. Plus de détails sont fournis dans la prochaine section.

L'ensemble des probabilités dans les tables de probabilités conditionelles (CPT) ne peuvent pas aisément être apprises, puisque que cela concerne des milliers d'entre elles (2151, plus précisément, pour le RB décrit dans cet article). Cependant, un ensemble de 25 probabilités sont supposées bien généraliser le système et ont pu être apprises à partir d'une base de données comprenant plus de 1000 dialogues (voir la Section 3.2).

2.2. *Les nœuds de grounding*

La composante de *grounding* est décrite sur la figure 2. Une valeur de *grounding* est obtenue pour chaque attribut i en suivant le processus suivant. Les états et les valeurs des nœuds KNW et GOAL pour l'attribut i sont récursivement copiés dans les nœuds KNW GROUNDING et GOAL GROUNDING. La machine d'inférence est utilisée ensuite pour inférer la valeur pour le nœud de *grounding*, en utilisant l'action du système pour compléter l'évidence. Si un problème de *grounding* est détecté pour l'attribut i, l'US est forcé de

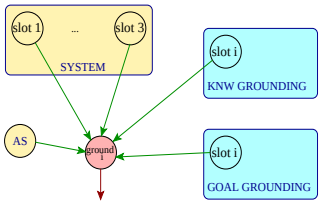


Fig. 2: La composante de *grounding*

donner l'information concernant cet attribut.

3. Expérimentations

3.1. Interaction avec REALL-DUDE/DIPPER/OAA

L'utilisateur simulé a été interfacé avec le système de dialogue oral inclus dans l'environnement REALL-DUDE/DIPPER/OAA (voir [3] et [1]). Cet environnement, originellement, a pour but l'entraînement de politiques de gestion de dialogue via des méthodes d'apprentissage par renforcement. Jusqu'à présent, la politique utilisée pour nos expérimentations a été entraînée indépendamment de l'US présenté ici et est utilisée seulement pour des tests. L'US accompagne chaque hypothèse qu'il envoie au GD d'un score de confiance simulé, qui serait donné, entre autres, par le reconnaisseur de parole (ASR). Si cette confiance est trop petite, le GD peut décider de demander à l'utilisateur de confirmer.

Dans cette section, sont présentées des statistiques calculées sur des dialogues obtenus en interfaçant l'US presenté dans cet article et l'environnement REALL-DUDE/DIPPER/OAA. Comme la tâche comprend trois attributs et comme l'US est configuré pour ne pas informer le GD au sujet de plus d'un attribut par tour, le nombre minimum de tours nécessaires pour atteindre la fin du dialogue est quatre : un par attribut, plus le tour "Close". Il faut noter qu'un tour est défini ici comme étant un couple <syst act>/<user act> (excepté pour l'action "Close", puisque c'est le GD qui a la possibilité de stopper les dialogues).

En considérant seulement les dialogues où un problème de *grounding* est survenu, et pas la totalité des dialogues, les résultats de la table 1 sont obtenus. Ils se révèlent prometteurs. Le taux de dialogues plus long que 5 tours diminue de 30.6 %.

3.2. Statistiques sur les dialogues

Dans cette section, sont présentées des statistiques calculées sur les dialogues obtenus en interfaçant deux versions de l'US présenté dans cet article et l'environnement REALL-DUDE/DIPPER/OAA. La première version de l'US est telle que les paramètres du RB ont des valeurs heuristiquement déterminées par un expert. La seconde version est telle que ces paramètres sont appris. La base de données utilisée pour l'entraînement contient 1234 dialogues. Elle comprend des dialogues beaucoup plus compliqués que ceux obtenus pour la tâche considérée ici. Douze attributs en tout sont présents. Les valeurs pour certains de ces attributs sont requis par le GD, comme "type de nourriture", etc., réclamant donc une action "Request" du système et une action "Inform" de l'utilisateur ; d'autres sont requis par l'utilisateur, comme "adresse", etc. ; la "gamme de prix" peut être requise par les deux protagonistes. De plus, plus de dix actions différentes, du système et de l'utilisateur, sont considérées (voir [10] pour une liste exhaustive). Finalement, plus d'un attribut et plus d'une action peuvent être présentés pendant chaque tour, aussi bien considérant les messages du GD que ceux de l'utilisateur. Ainsi, la base de données a nécessité d'être labellée avec soin, dans l'objectif de fournir des données utilisables pour cet article. La base de données a été décrite avec plus de détails dans [9], où elle a été utilisée pour entraîner des stratégies du GD.

Tab. 1: Nombre moyen de tours requis pour atteindre la fin des dialogues ; pourcentage de dialogues pour lesquels la fin est atteinte en plus de 5 tours

	moyenne	>5 tours
sans composante *grounding*	5.254	29.22 %
avec composante *grounding*	5.069	20.29 %

Tab. 2: Nombre moyen de tours requis pour atteindre la fin des dialogues ; nombre max de tours ; nombre min de tours ; pourcentage de dialogues pour lesquels la fin est atteinte en 4 tours ; pourcentage de dialogues pour lesquels la fin est atteinte en moins de 9 tours

	moyenne	max	min	4 tours	<9 tours
h-RB	4.969	21	4	93.20 %	93.40 %
e-RB	4.577	9	4	58.00 %	99.90 %

Dans la table 2, sont présentés les résultats obtenus en utilisant le RB heuristique (h-RB) et le RB entraîné (e-RB). Mille dialogues ont été simulés pour chaque RB. Sur la figure 3, est présenté l'histogramme du nombre de tours requis pour atteindre la fin des dialogues. Considérant le h-RB, on peut remarquer que la plupart des dialogues sont, en effet, complétés en quatre tours, comme attendu (93.4 %). Ceci est dû au fait qu'il a été spécialement mis en place pour obtenir des dialogues aussi courts que possible.

L'e-RB donne des dialogues plutôt plus longs que ceux donnés par le h-RB. Ceci ne peut pas être remarqué si l'on considère le nombre moyen de tours ; il faut considérer le pourcentage de dialogues qui ont nécessité exactement quatre tours pour atteindre leur fin : ce pourcentage passe de 93.2 % à 58.0 %. Cependant, de façon très prometteuse, le nombre moyen de tours et le pourcentage de dialogues qui ont nécessité moins de neuf tours pour atteindre leur fin sont plutôt meilleurs quand on utilise l'e-RB ; de plus il faut remarquer que les très longs dialogues (plus de neuf tours), indiquant une profonde incompréhension entre le GD et l'US, ont complètement disparus. Ceci indique que les dialogues obtenus avec l'e-RB sont plus naturels, au moins du point de vue du GD, que les

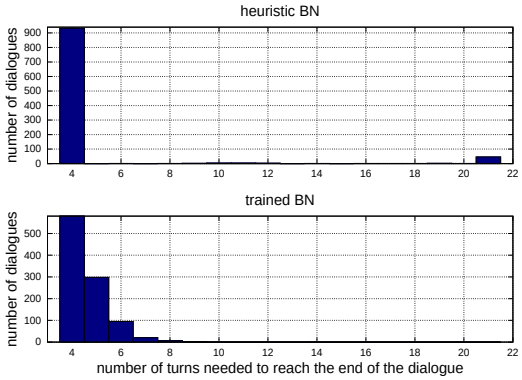


Fig. 3: Histogramme du nombre de tours requis pour atteindre la fin du dialogue – haut : le RB heuristique est utilisé – bas : le RB entraîné est utilisé

Tab. 3: Nombre de tours requis pour chaque attribut ; h-RB, plus longs dialogues conservés ; h-RB, plus longs dialogues éliminés ; e-RB ; et base de données

h-RB	h-RB sans longs dialogues	e-RB	base de données
1.6563	1.3986	1.5257	1.5661

dialogues obtenus avec le h-RB. La distribution des actions est en fait plus en accord avec les données, ce qui montre de nouveau que les dialogues simulés sont plus naturels avec l'e-RB.

La table 3 présente le nombre de tours requis par attribut, respectivement quand le h-RB est utilisé les plus longs dialogues étant conservés, quand le h-RB est utilisé les plus longs dialogues n'étant pas conservés (ils reflètent seulement la condition d'arrêt incluse dans le GD), quand l'e-RB est utilisé, et considérant le base de données. Clairement, l'e-RB donne des dialogues plus réalistes, en terme de nombre de tours requis pour chaque attribut.

4. Conclusion et perspectives

Dans cet article, un modèle d'utilisateur simulé basé sur les réseaux bayésiens est proposé pour simuler des dialogues homme-machine (au niveau de l'intention), réalistes, et incluant des comportements de *grounding*. Notre but était de montrer l'intérêt de simuler le processus de *grounding*, qui survient souvent dans les dialogues humain-humain. Ceci est fait en comparant le nombre de tours requis pour atteindre la fin d'un dialogue quand différentes configurations du modèle proposé sont utilisées. Plusieurs perspectives sont de plus envisagées.

Premièrement, cet Utilisateur Simulé sera utilisé pour l'entraînement des politiques du GD, ce considérant le paradigme de l'apprentissage par renforcement. Deuxièmement, des résultats préliminaires concernant l'interaction de l'US présenté ici avec un GD indépendamment développé sont montrés. Il est prévu à court terme d'interagir avec le système de dialogue inclus dans l'environnement REALL-DUDE/DIPPER/OAA d'une manière beaucoup plus intensive et systématique. Ceci nous permettra de comparer le nombre de tours obtenus avec les Utilisateurs Simulés basés sur les RB et le nombre de tours obtenus avec des utilisateurs humains, d'analyser le taux de complétion de la tâche correspondant, etc. Ceci nous permettra aussi d'entraîner la politique utilisée dans les POMDP implantés dans l'environnement REALL-DUDE/DIPPER/OAA. Troisièmement nous souhaiterions utiliser la capacité des RB d'apprendre en ligne leurs paramètres afin d'améliorer le caractère naturel des dialogues simulés, alors que des utilisateurs réels sont en train d'interagir avec le système ; et nous souhaiterions ré-entraîner des politiques du GD à partir de ces interactions réelles.

Références

[1] J. Bos, E. Klein, O. Lemon, and T. Oka. DIPPER : Description and Formalisation of an Information-State Update Dialogue System Architecture. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124, 2003.

[2] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The Lumiere Project : Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In *Proceedings of the 14th Conference on Uncertainty in Artifical Intelligence*, 1998.

[3] O. Lemon, X. Liu, D. Shapiro, and C. Tollander. Hierarchical Reinforcement Learning of Dialogue Policies in a Development Environment for Dialogue Systems : REALL-DUDE. In *10th SemDial Workshop on the Semantics and Pragmatics of Dialogue ; BRANDIAL*, 2006.

[4] E. Levin, R. Pieraccini, and W. Eckert. A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies. In *IEEE Transactions on Speech and Audio Processing*, volume 8, pages 11–23, 2000.

[5] H. Meng, C. Wai, and R. Pieracinni. The Use of Belief Networks for Mixed-Initiative Dialog Modeling. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, 2000.

[6] O. Pietquin. Learning to Ground in Spoken Dialogue Systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 165–168, 2007.

[7] O. Pietquin and T. Dutoit. A Probabilistic Framework for Dialog Simulation and Optimal Strategy Learning. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 14, pages 589–599, 2006.

[8] O. Pietquin and T. Dutoit. Dynamic Bayesian Networks for NLU Simulation with Applications to Dialog Optimal Strategy Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[9] J. D. Williams and S. Young. Partially Observable Markov Decision Processes for Spoken Dialog Systems. In *Computer Speech and Language*, volume 21, pages 231–422, 2007.

[10] S. Young. CUED Standard Dialogue Acts. Technical report, Cambridge University Engineering Dept, 2007.

Evaluation d'un alignement automatique sur la parole dysarthrique

Nicolas Audibert[1], Cécile Fougeron[2], Corinne Fredouille[1], Christine Meunier[3], Olavo Panseri[2]

[1] Université d'Avignon, CERI/LIA, Avignon, France
[2] Laboratoire de Phonétique et Phonologie, UMR 7018 CNRS-Paris3/Sorbonne Nouvelle, Paris, France
[3] Laboratoire Parole et Langage, CNRS, Aix-Marseille Université, France
{prenom.nom}@univ-avignon.fr, cecile.fougeron@univ-paris3.fr, christine.meunier@lpl-aix.fr, olavo.panseri@gmail.com

ABSTRACT

Phonetico-acoustic analysis of pathological speech requires a reliable phonetic alignment. Since manual labeling is highly time-consuming, automatic alignment may be necessary for analyzing large databases. This paper evaluates the reliability of automatic alignment for dysarthric speech. Results on read speech samples of 4 dysarthric speakers compared to 2 normophonic speakers show that alignment performance depends on the severity of dysarthria. Specific patterns for different phonetic classes and directions for filtering reliable parts are discussed.

Keywords: dysarthric speech, phonetico-acoustic study, automatic phonetic alignment, evaluation.

1. INTRODUCTION

Depuis une quinzaine d'années, la phonétique clinique, dédiée à l'étude des troubles de la parole, de la voix et du langage, est devenue un domaine multidisciplinaire où se côtoient cliniciens et chercheurs des domaines des sciences du langage et du traitement automatique de la parole.

La dysarthrie est un trouble de la parole d'origine neurologique qui se manifeste par une déficience motrice. Elle a fait l'objet de nombreuses études dans la littérature, portant notamment sur sa caractérisation dans le domaine acoustique ou donnant lieu à diverses classifications. Si un ensemble de paramètres jugés pertinents dans la différentiation de patients touchés par des types de dysarthries différents a pu être défini, la dysarthrie nécessite encore des études approfondies. Notamment, une description phonétique rigoureuse permettrait de mieux appréhender et prendre en compte la grande diversité des phénomènes observée dans la parole des patients. De telles études nécessitent le traitement de grands corpus de données, comportant suffisamment de patients, de types de dysarthrie (voire de maladies) et d'échantillons de parole pour envisager une analyse fine, rigoureuse et robuste. Dans ce contexte, une analyse manuelle seule n'est pas envisageable puisque la tâche de segmentation en phonèmes d'un signal de parole, étape préalable requise pour toute analyse phonétique, est à elle seule fastidieuse et surtout excessivement consommatrice en termes de ressources humaines. L'utilisation de systèmes issus du traitement automatique de la parole est, par conséquent, une solution à considérer dans ce contexte très particulier.

Il existe différents types de systèmes d'alignement automatique de la parole (le lecteur pourra se référer à Nefti [1] pour une revue complète). Nous nous intéresserons ici aux systèmes d'alignement contraint par le texte, qui, à partir d'une transcription orthographique du texte prononcé dans l'échantillon de parole associée à un lexique phonétisé, permettent de déterminer automatiquement les frontières des phonèmes présents. Nefti souligne que les performances de systèmes de ce type varient, dans la littérature, entre 80 et 90% de concordance avec une segmentation manuelle de référence. Par ailleurs, différents travaux ont été menés afin d'évaluer la pertinence de l'utilisation de tels systèmes dans le cadre d'analyse en phonétique et phonologie comme par exemple l'étude des voyelles [2], ou encore du schwa [3]. Cette pertinence de l'approche automatique est, dans la plupart des études, démontrée bien qu'elle soit clairement accompagnée de précautions à prendre dans l'analyse des résultats.

L'objectif de ce papier est d'étudier cette même pertinence dès lors que le système automatique est appliqué sur des échantillons de parole dysarthrique, présentant des degrés de sévérité variables. Le résultat de cette étude devrait conduire à l'élaboration de recommandations d'utilisation de ces systèmes dans ce contexte très particulier. Pour ce faire, une description du système d'alignement automatique ainsi que des procédures de corrections manuelles est présentée. Les différentes procédures et comparaisons mises en place pour mesurer la concordance entre alignement manuel et automatique sont ensuite détaillées. Enfin, l'évaluation des performances de l'alignement automatique (comparées à un alignement manuel) est faite en regard des points suivants : les types d'erreurs, les profils de locuteurs et les types de phonèmes.

2. METHODES

2.1. Procédure d'alignement automatique

L'alignement automatique utilisé dans ce travail est dit contraint par le texte dans le sens où il consiste à faire correspondre une chaîne phonétique à un signal de parole en identifiant les phonèmes produits et en segmentant les parties du signal leur correspondant (émission de frontières de début et de fin pour chaque phonème).

Développé par le Laboratoire Informatique d'Avignon (LIA), ce système repose sur l'utilisation classique de modèles de Markov Cachés (HMM) associés à un algorithme de décodage de type Viterbi (le lecteur pourra se référer à Rabiner *et al.* [4] pour une revue détaillée sur les HMM et l'algorithme Viterbi, et à Brugnara *et al.* [5] pour un descriptif du processus d'alignement). Pour accomplir sa tâche de segmentation du signal en phonèmes, le système requiert différentes ressources linguistiques : (1) une transcription orthographique « fidèle » du message linguistique véhiculé par l'échantillon de parole (prise en compte des insertions, substitutions et suppressions de mots ou sons dans le texte et des disfluences), (2) un lexique phonétisé défini à partir de la transcription orthographique, pouvant comporter, pour chaque entrée lexicale, un ensemble de variantes phonologiques, (3) un ensemble de modèles HMM représentant les différentes formes acoustiques des phonèmes estimé sur un corpus d'apprentissage. Ici, 38 modèles HMM indépendants du contexte appris sur le corpus d'émissions radiophoniques ESTER [6] sont utilisés. Pour notre étude, le lexique du système a été réduit aux lexèmes contenus dans le texte étudié. Les variantes de prononciation de ces lexèmes ont également été filtrées en fonction de ce dernier.

2.2. Procédure de correction manuelle de l'alignement automatique

La segmentation manuelle qui nous servira de référence dans cette étude a été réalisée par des experts humains à partir de l'alignement automatique. Elle consiste en une correction/vérification manuelle des étiquettes phonémiques apposées par l'aligneur automatique, et de la localisation sur le signal de leurs frontières de début et de fin. Les corrections incluent donc des ajouts, suppressions, substitutions (modifications) d'étiquettes phonémiques dans les cas où elles ne correspondent pas à ce qui a effectivement été réalisé, et des décalages temporels des frontières vers la droite ou la gauche dans les cas où les frontières placées ne correspondent pas aux critères de l'expert. S'il est difficile de définir avec certitude les frontières de phonèmes dans le continuum de parole, les phonéticiens utilisent des critères assez robustes de segmentation à partir de l'examen du signal de parole et du spectrogramme : l'apparition et la disparition du 2^{e} formant pour la segmentation des voyelles, du bruit dans les hautes et moyennes fréquences pour les fricatives, d'une tenue voisée ou silencieuse pour les occlusives (la tenue des occlusives sourdes n'est pas identifiable lorsqu'elles suivent une pause), d'un bruit correspondant au relâchement des occlusives, etc. Les aspects les plus délicats de la segmentation sont les suites de consonnes et de voyelles caractérisées par une structure de formants. En effet, le passage continu d'une articulation à une autre ne correspond pas toujours à une adresse temporelle précise.

Si la qualité essentielle d'une segmentation manuelle réside dans la consistance à toujours appliquer les mêmes critères de segmentation, il n'en reste pas moins vrai que les critères choisis sont fonction des représentations phonétiques que l'expert a du signal de parole, et peuvent donc varier d'un expert à l'autre [7]. Nous avons donc comparé les segmentations de deux experts phonéticiens sur une partie du corpus (voir plus loin pour les détails).

Dans le cas de la parole pathologique, les problèmes de segmentation sont amplifiés du fait des réalisations phonétiques souvent très perturbées des patients et de la présence de continua acoustiques quasi insegmentables. Dans ces cas, les deux experts ont eu des stratégies différentes. L'expert 1 a préféré noter ces continua comme insegmentables, tandis que l'expert 2 a tenté de distinguer les différents segments lorsque cela lui semblait possible.

2.3. Corpus

Les productions de parole dysarthrique utilisées dans cette étude ont été mises à notre disposition par l'Hôpital de la Pitié-Salpétrière. Le corpus est composé d'enregistrements de 4 patients atteints de maladies rares de surcharges lipidiques, présentant différents degrés de sévérité de dysarthrie, suivant le degré d'évolution de leur maladie. Les codes des locuteurs indiquent dans l'ordre leur genre (M ou F), le degré de sévérité de la dysarthrie (0=sujet contrôle; 1=dysarthrie légère ; 2=dysarthrie sévère), et l'initiale de leur prénom. La population, équilibrée en genre, est composée de deux patients à dysarthrie sévère (M2V et F2S) et deux patients à dysarthrie légère (M1A et F1C). Deux sujets contrôles non-dysarthriques d'âges similaires (M0A et F0D) font également partie des analyses. Nous nous sommes focalisés dans cette étude sur la lecture du texte 'Tic Tac' de la batterie de C. Chevrie-Müller. Les enregistrements ont été effectués dans un environnement calme mais non contrôlé. Les patients pouvant présenter des états de fatigue très variables, les patients dysarthriques les plus sévères ne sont pas arrivés jusqu'au bout du texte.

Les productions de ces locuteurs ont été transcrites orthographiquement selon une convention stricte et les signaux de parole ont été alignés par le système automatique. Le premier expert humain a corrigé les alignements automatiques (AA) des 4 patients, le second a corrigé les alignements des contrôles et de deux patients (M1A et F2S). Les corrections du premier expert et du second sont notées respectivement AM1 et AM2.

3. COMPARAISON DES ALIGNEMENTS AUTOMATIQUE VS. MANUEL

Les étiquettes et frontières phonémiques définies par l'AA ont été comparées à celles placées manuellement par les experts au moyen d'une procédure semi-automatique. Un premier filtrage a été nécessaire pour ne conserver que les segments comparables, excluant notamment les continua jugés insegmentables. Afin de tenir compte du décalage induit par les insertions, suppressions ou substitutions de phonèmes sur les

frontières des segments voisins, et évaluer les performances de l'AA indépendamment des erreurs issues de la transcription, les segments voisins ont également été exclus.

3.1. Divergences d'étiquettes phonémiques

Les divergences entre étiquettes proviennent en majorité des continua jugés insegmentables, dont le nombre est compris entre 40 pour M2V et 89 pour M1A dans AM1, et de 28 pour F2S et 43 pour M1A dans AM2. Les insertions, suppressions et substitutions de phonèmes sont en nombre plus restreint. Quel que soit l'alignement manuel considéré, leur nombre total est en effet compris entre 13 pour F1C et 29 pour F2S. Le nombre de segments analysés après élimination des segments non-comparables et la proportion du nombre total de segments sont présentés dans la table 1 pour chaque comparaison entre alignements.

3.2. Décalage temporel

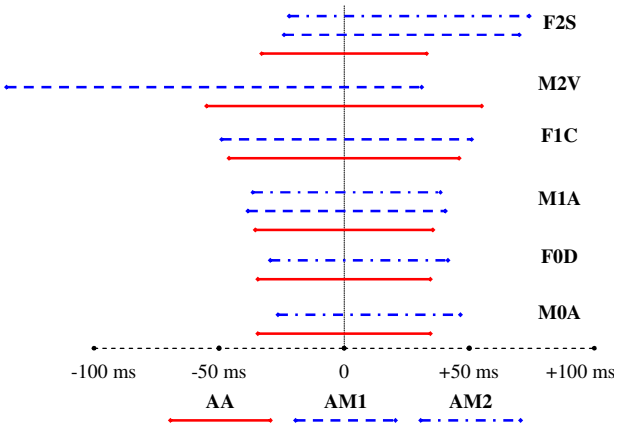


Figure 1 : Segments moyens pour chaque locuteur et chaque alignement, avec leurs décalages relatifs. Les lignes représentant les segments sont alignées sur le point central de l'AA (traits pleins rouges).

La déviation temporelle entre alignements est analysée en termes de décalage temporel entre les onsets (DecOn), points centraux (DecCtr) et offsets des unités segmentées (DecOff). Ce dernier critère cherche à vérifier si les phonèmes sont correctement alignés au signal au niveau de leur point central, même si les frontières de début et de fin sont décalées. Les valeurs négatives indiquent que le point considéré apparaît plus tôt dans le signal. La figure 1 présente les segments moyens issus des 3 alignements et leurs décalages relatifs pour chacun des 6 locuteurs. Les valeurs représentées pour les locuteurs M1A et F2S correspondent aux segments communs aux comparaisons AA vs. AM1 et AA vs. AM2. Pour chaque comparaison entre alignements, l'effet du locuteur sur les mesures de décalage a été évalué par des tests ANOVA. Les résultats de ces tests sont présentés dans la table 1. Une comparaison directe des points initiaux, centraux et finaux, dont les résultats ne sont pas détaillés ici, confirme un résultat prévisible : les alignements sont en effet tous significativement différents deux à deux.

a) Patients vs. contrôles

Il ressort de la comparaison entre AA et AM2 que la proportion de segments pour lesquels le décalage du point central est supérieur à 20 ms est peu élevé pour les locuteurs M0A (17%), F0D (14%) et M1A (21%). En revanche cette proportion est de 52% pour la locutrice F2S. Un effet significatif du locuteur sur DecCtr et DecOff est observé, les valeurs de DecOn étant en revanche comparables entre locuteurs. Les décalages sont significativement plus importants pour F2S que pour M1A et les deux sujets contrôles.

b) Décalages en fonction de la nature du segment

La répartition des décalages temporels entre AA et AM2 a ensuite été examinée par type de segment. N'ont été retenues pour l'analyse que les classes contenant au moins 10 exemplaires par locuteur. 10 classes acoustiques sont comparées : Fricatives sourdes et voisées, Occlusives sourdes et voisées, /l/, /ʁ/, Consonnes nasales, Voyelles orales, Voyelles nasales et Semi-voyelles. Les valeurs de décalage ont été examinées pour les valeurs supérieures à 20 ms, soit 2 trames pour le système automatique, sur DecCtr ou sur les différences de durée déduites de DecOn et DecOff. Cet examen révèle des décalages entre AA et AM particulièrement importants pour les occlusives sourdes, y compris pour les sujets contrôles. En effet, bien que les décalages moyens observés sur l'onset de la tenue et l'offset du burst soient généralement très faibles pour ces locuteurs (-2 à -4 ms, à l'exception de l'offset du burst pour M0A, décalé de +20 ms), indiquant un placement peu dégradé des frontières de l'occlusive considérée dans sa globalité, l'explosion est décalée de manière récurrente (-29 ms). Le locuteur M1A présente le même pattern, avec de plus l'onset de la tenue décalé de +31 ms en moyenne. La valeur moyenne de DecOn sur les semi-voyelles de M0A est de -19 ms, et celle de DecOff sur les voyelles nasales de F0D de -17 ms, pour des différences de durée respectives de +20 ms et -26 ms. Pour la locutrice F2S, des décalages importants sont observés pour l'ensemble des types de segments.

c) Décalages en fonction de la sévérité de la dysarthrie

La comparaison entre AA et AM1, qui porte sur les 4 patients, indique que la proportion de segments dont le décalage du point central est supérieur à 20 ms semble dépendre du degré de sévérité de la dysarthrie : elle est en effet de 20% pour M1A et 26% pour F1C, contre 53% pour F2S et 66% pour M2V. Un effet significatif du locuteur sur les trois mesures de décalage est mesuré, les décalages étant significativement plus importants pour M2V que pour les trois autres patients.

d) Variabilité des alignements manuels

Enfin, la comparaison entre AM1 et AM2 indique que seuls 3% des segments de M1A présentent une valeur de DecCtr supérieure à 20 ms, et 7% de ceux de F2S, tandis que pour ces 2 locuteurs le désaccord entre experts mesuré au niveau du point central est inférieur à 1 ms

dans plus de 75% des cas. Les écarts observés sont dus en grande partie à quelques segments précédés ou suivis d'un bruit de durée importante, considérés comme faisant partie du segment dans AM1 et comme distincts de ce segment dans AM2. Un effet significatif du locuteur sur les valeurs de DecDtr et DecOff est observé, mais pas sur les valeurs de DecOn.

4. DISCUSSION

L'analyse menée met en évidence une variabilité des performances du système d'alignement en fonction de la sévérité de la dysarthrie, mais également une différence entre experts humains qui, si elle n'est pas de même ampleur, n'est pas pour autant négligeable.

Une part des décalages les plus importants mesurés (jusqu'à 300 ms pour les patients les plus dysarthriques) peut s'expliquer par le fonctionnement du système d'alignement automatique. En effet, lorsque le système ne parvient pas à apparier les trames du signal avec les modèles des phonèmes supposés présents, il opère une resynchronisation de l'alignement au niveau de la portion de faible énergie suivante (modèle de silence). La majorité des trames qui n'ont pu être appariées sont alors intégrées à ce segment de faible énergie, les segments supposés présents d'après le texte à aligner étant placés au début de la portion de signal non reconnue, avec une durée de 30 ms (seuil minimal fixé par le système).

Les observations relatives au placement de l'explosion des occlusives sourdes indiquent que des mesures directement dépendantes de ces frontières comme celles de VOT ne peuvent être extraites de façon fiable à partir de l'alignement automatique. Toutefois, les décalages en cascade de segments adjacents induits par les erreurs de l'alignement automatique liés à la resynchronisation des portions de signal mal reconnues limitent la portée des autres analyses des décalages en fonction de la nature des segments.

Afin de ne pas prendre en compte dans de futures analyses acoustiques menées à partir de l'alignement automatique des segments présentant un tel décalage, la prochaine étape de notre travail sera de les filtrer automatiquement en tirant parti de mesures de confiance estimées durant le processus d'alignement automatique comme utilisé dans Clément *et al*. [8].

Remerciements : *Ce travail a été réalisé dans le cadre du projet « DesPho Apady » (ANR-08-BLAN-0125), financé par l'Agence Nationale de la Recherche (ANR).*

Nos remerciements vont à Georges Linares pour son aide sur l'utilisation du système d'alignement ainsi qu'à Nathalie Lévêque et Frédéric Sedel pour la mise à notre disposition du corpus de parole dysarthrique.

BIBLIOGRAPHIE

[1] S. Nefti. *Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance*. Thèse de doctorat, Université de Rennes 1, 2004.

[2] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde and S. Rauzy. Le Cid - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, 49 (3):105-134, 2008.

[3] A. Bürki, C. Gendrot, G. Gravier, G. Linares and C. Fougeron. Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa. *Traitement Automatique des Langues*, 49(3):165-197, 2008.

[4] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, 1993.

[5] F. Brugnara, D. Falavigna and M. Omologo. Automatic segmentation and labeling of speech based on Hidden Markov Models, *Speech Communication*, 12:357-370, 1993.

[6] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa and K. Choukri. Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proceedings of LREC'06,* 2006.

[7] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling and W. Raymond. The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. *Speech Communication*, 45:89-95, 2005.

[8] P. Clément, C. Fredouille and N. Lévêque. Méthodes objectives appliquées à la dysarthrie, In *Actes 3èmes Journées de Phonétique Clinique*, 2009.

Table 1 : Résultats des tests ANOVA réalisés pour chaque comparaison sur les variables decOn, decCtr et decOff. Les différences entre groupes résultant des comparaisons multiples (test HSD de Tukey) sont indiquées pour p<.05. N seg : nombre et proportion de segments analysés ; N loc : nombre de locuteurs pris en compte dans la comparaison.

Comparaison	N seg	N loc	Variable	F	p	Comparaisons multiples
AA vs AM1	1319 (77%)	4	decOn decCtr decOff	11.5 7.3 4.5	<.001 <.001 .004	M2V≠(F2S, F1C, M1A) M2V≠(F2S, F1C, M1A) M2V≠(F2S, F1C, M1A)
AA vs. AM2	1917 (86%)	4	decOn decCtr decOff	2.0 35.4 9.2	.111 <.001 <.001	(F2S, M0A, F0D, M1A) F2S≠(M0A, F0D, M1A) F2S≠(M0A, F0D, M1A)
AM1 vs. AM2	716 (81%)	2	decOn decCtr decOff	0.9 10.8 20.6	.355 .001 <.001	(M1A, F2S) M1A≠F2S M1A≠F2S

Déficit de compréhension de la parole dans le bruit chez le dyslexique adulte et lien avec le système efférent auditif

Véronique Boulenger[1], Michel Hoen[2], Claire Grataloup[1], Evelyne Veuillet[3], Lionel Collet[3], Fanny Meunier[1]

[1] Laboratoire Dynamique du Langage, UMR 5596 CNRS - Université Lyon 2, Lyon, France
[2] Institut de Recherche sur les Cellules Souches et le Cerveau, INSERM U846 - Université Lyon 1, Bron, France
[3] Laboratoire Neurosciences et Systèmes Sensoriels, UMR 5020 CNRS - Université Lyon 1, Lyon, France
Veronique.Boulenger@ish-lyon.cnrs.fr

ABSTRACT

This paper investigates speech-in-noise comprehension in adult dyslexics. We report results showing that dyslexics experience difficulties in identifying speech in multi-talker babble but that their performance is influenced by lexical variables in the same way as for normal readers. We then assess the link between speech-in-noise comprehension deficits in dyslexics and the functionality of the auditory efferent system involved in the descending control from cortical areas to the cochlea. Results reveal that despite normal functioning of the efferent system in attenuating acoustic otoemissions, dyslexics show a lack of asymmetry of this system compared to normal readers. This specific lateralization pattern could reflect a lack of asymmetry of central language areas and contribute to the speech-in-noise comprehension deficit in dyslexia.

Keywords: developmental dyslexia, speech-in-speech, auditory efferent system

1. INTRODUCTION

La dyslexie développementale est définie comme un trouble 'spécifique' de l'apprentissage de la lecture et de l'écriture, en l'absence de déficit intellectuel ou handicap sensoriel associé. Si les enfants dyslexiques se caractérisent d'abord par leurs difficultés d'apprentissage de la langue (lecture extrêmement lente et laborieuse, difficultés à décoder des pseudo-mots ou à épeler des mots), d'autres traits symptomatiques, tels que des troubles de production et de compréhension de la parole, sont fréquemment observés. Au niveau comportemental, on note également des difficultés scolaires et des troubles attentionnels, causés par les difficultés de compréhension orale, qui les rapprochent souvent d'enfants atteints de troubles du traitement auditif (TTA). La dyslexie est par ailleurs souvent considérée comme une pathologie infantile alors que les difficultés persistent à l'âge adulte [1]. Il est établi depuis longtemps que la dyslexie comporte une dimension acoustique ou auditive, mais, du fait de son ancrage dans les pathologies du langage, cet aspect a toujours été considéré sur le versant langagier. L'une des théories explicatives principales de la dyslexie, la théorie phonologique, suppose que la dyslexie résulte d'un déficit de représentation de l'information phonologique ou de l'accès à cette représentation lors de la compréhension de la parole ou de la lecture [2]. De récentes études montrent que ces difficultés d'accès ne se manifestent pas de manière systématique mais seulement lorsque l'accès est rendu difficile, notamment par l'ajout de bruit ou lorsque le signal de parole est dégradé *per se*. Ziegler et al. [3] rapportent ainsi chez des enfants dyslexiques des difficultés à percevoir la parole dans le bruit qui sont généralisables à différents indices phonologiques (voisement, lieu d'articulation) et que le bruit soit extrinsèque ou intrinsèque.

Le but de la présente étude était de déterminer si cette sensibilité exacerbée à la présence de bruit lors de la compréhension de la parole persiste chez les dyslexiques adultes, et le cas échéant, si les dyslexiques utilisent des stratégies de compensation lors de l'accès lexical par rapport à des adultes normo-lecteurs. Nous avons mesuré les performances de dyslexiques et de normo-lecteurs lors de l'identification de mots et de pseudo-mots cibles dans un bruit parolier concurrent et alors que la fréquence lexicale des mots cibles était manipulée.

Nous nous sommes ensuite intéressés au fonctionnement du système auditif de nos populations. Le système auditif humain est constitué de voies ascendantes (de la périphérie vers le cortex) et de voies descendantes (du cortex vers la cochlée). Dans cette étude, nous avons évalué les capacités d'encodage périphérique de nos participants grâce à des tests audiométriques et nous avons exploré le fonctionnement d'une partie des voies efférentes, le système olivocochléaire médian (SEOCM). Des travaux réalisés chez l'humain et sur l'animal suggèrent en effet que cette voie pourrait être impliquée lors de l'écoute en contexte bruité [4-5]. En outre, une précédente étude de notre groupe a mis en évidence une corrélation élevée entre les mesures de SEOCM et les performances d'intelligibilité [6]. Sans postuler un lien direct entre cette voie et les performances d'intelligibilité de la parole dans le bruit, ce résultat nous a conduits à poursuivre son exploration.

2. INTELLIGIBILITE DE LA PAROLE DANS LA PAROLE CHEZ LES DYSLEXIQUES

2.1. Méthode

Participants

Quarante adultes normo-lecteurs (20 femmes) âgés de 18 à 25 ans (M = 21.49 ans) et 49 adultes dyslexiques diagnostiqués (25 femmes) âgés de 16 à 40 ans (M = 24.73) ont participé à cette étude. Tous les participants avaient une vue normale ou corrigée et ne présentaient aucun trouble auditif. Les participants ont signé un formulaire de consentement et ont été dédommagés pour leur participation.

Stimuli

Mots cibles

Cent-vingt mots ont été sélectionnés dans la base de données Lexique 2 [7]. Tous les mots étaient des noms communs de la langue française, monosyllabiques et composés de trois phonèmes. La moitié des mots était de haute fréquence (50.1 < f+ < 149.23 occurrences par million, M = 66.39), l'autre de basse fréquence (1 < f- < 4.94 occurrences par million, M = 2.54). Les 120 mots ont été enregistrés (22 kHz, 16 bits, mono) dans un caisson insonorisé par une locutrice de langue maternelle française. Les enregistrements ont ensuite été normalisés à une intensité de -3dB puis segmentés en 120 fichiers indépendants contenant chacun un mot cible.

Pseudo-mots cibles

Cent-vingt pseudo-mots monosyllabiques ont été créés en recombinant les phonèmes des mots cibles. Tous les pseudo-mots respectaient les contraintes phonotactiques du français. Les 120 pseudo-mots ont été enregistrés, normalisés et segmentés selon la même procédure que celle utilisée pour les mots.

Bruits de fond *cocktail party*

Les bruits de fond étaient constitués de bruits *cocktail party* enregistrés par 8 locuteurs de langue maternelle française (4 femmes) dans un caisson insonorisé (44 kHz, 16 bits, stéréo). Les locuteurs devaient lire des listes de mots de fréquence lexicale moyenne dans la langue. Les enregistrements individuels ont été vérifiés et modifiés selon le protocole suivant : (i) suppression des pauses excédant 1 seconde et des mots contenant des erreurs de prononciation, (ii) réduction du bruit optimisée pour les sons de parole et (iii) calibration de l'intensité en dB-A et normalisation de chaque source à 80dB-A. Les bruits *cocktail party* ont été créés en mixant les enregistrements individuels, en équilibrant le genre des locuteurs (50% femmes, 50% hommes) et en variant le nombre de locuteurs dans le cocktail (4, 6 ou 8). Au total, 3 bruits *cocktail party* mixtes ont donc été générés (C4, C6 et C8).

Stimuli expérimentaux

Les stimuli expérimentaux ont été créés en mixant les 120 mots et 120 pseudo-mots cibles avec des extraits de *cocktail party* de 4 secondes choisis aléatoirement. Les mots et pseudo-mots cibles étaient systématiquement insérés 2.5 s après le début de l'extrait sonore. Au travers des participants, chaque cible apparaissait dans chaque type de cocktail et n'était entendue qu'une seule fois. Une normalisation en intensité aléatoire sur une gamme de ± 3dB par pas de 1dB a été appliquée aux stimuli de sorte que leur intensité globale ne prédise pas la condition expérimentale.

Procédure

Les participants étaient assis face à un écran d'ordinateur et devaient écouter attentivement les stimuli présentés en mode binaural grâce un casque audio à un niveau d'écoute confortable. Les participants normo-lecteurs devaient retranscrire au clavier le mot ou le pseudo-mot cible entendu alors que les dyslexiques devaient répéter l'item cible entendu. L'expérience débutait par une phase d'entraînement de 12 items (différents des stimuli expérimentaux) non pris en compte dans les analyses. Tous les bruits, listes de mots et conditions ont été distribués pseudo-aléatoirement entre les participants. Les stimuli composés de mots cibles et de pseudo-mots cibles étaient présentés dans deux sessions différentes, chacune durant 30 min. Les transcriptions des participants ont été codées en valeurs numériques puis analysées en termes de pourcentage d'items correctement reproduits.

2.2. Résultats

Les performances des deux groupes de participants ont été comparées grâce à une analyse de variance (ANOVA) à mesures répétées incluant les facteurs Groupe (normo-lecteurs vs. dyslexiques), Lexicalité (mots vs. pseudo-mots) et Nombre de Voix dans le cocktail (C4 vs. C6 vs. C8). En moyenne, les normo-lecteurs ont obtenu des performances de restitution des items cibles meilleures (49.96%, ET = 6.29) que les dyslexiques (43.45%, ET = 8.07; $F(1, 87) = 22.81$, $p < .001$). Cette baisse des performances chez les dyslexiques était observée à la fois pour les mots (-6.18%) et les pseudo-mots (-6.83%). Un effet significatif du facteur Lexicalité a également émergé ($F(1, 39) = 717.16$, $p < .001$), les mots étant mieux restitués (57.41%, ET = 6.59) que les pseudo-mots (36%, ET = 7.76). Cet avantage lexical était observé aussi bien chez les normo-lecteurs (mots : 60.5%, ET = 5.46 vs. pseudo-mots : 39.42%, ET = 7.02 ; $F(1, 39) = 394.87$, $p < .001$) que chez les dyslexiques (mots : 54.32%, ET = 7.63 vs. pseudo-mots : 32.59%, ET = 8.51 ; $F(1, 48) = 498.17$, $p < .001$; Figure 1). La magnitude de l'avantage lexical (différence entre les performances d'intelligibilité pour les mots et les performances pour les pseudo-mots) était en outre comparable chez les deux groupes de participants (normo-lecteurs : 21.08% ; dyslexiques : 21.73%). Un effet significatif du Nombre de Voix dans le cocktail a enfin été observé ($F(2, 78) = 5.34$, $p < .01$), la restitution des items cibles étant moins bonne dans un cocktail à 8 voix (C4 : 47.04%, ET = 8.56 ; C6 : 47.95%, ET = 7.97; C8 = 45.12%, ET = 7.56). Aucune interaction significative entre les facteurs testés n'a été observée.

Une analyse des performances des participants pour identifier les mots cibles a par ailleurs mis en évidence un effet significatif de la Fréquence Lexicale des mots sur les scores de restitution ($F(1, 39) = 1251.35$, $p < .001$). Les mots fréquents (f+ : 68.03%, ET = 10.81) étaient mieux

restitués que les mots peu fréquents (f- : 46.18%, ET = 11.9), tant chez les normo-lecteurs (f+ : 71.08%, ET = 6.39 vs. f- : 49.92%, ET = 6.73 ; F (1, 39) = 368.72, p < .001) que chez les dyslexiques (f+ : 65.54%, ET = 7.5 vs. f- : 43.1% , ET = 9.25 ; F (1, 48) = 488.11, p < .001 ; Figure 1). Cette supériorité des mots fréquents sur les mots peu fréquents était d'ailleurs comparable chez les deux groupes de participants (normo-lecteurs : 21.17% ; dyslexiques : 22.40%). En outre, le déficit des dyslexiques pour comprendre les mots dans le bruit était identique quelle que soit leur fréquence lexicale (f+ : -5.54% ; f- : -6.79%).

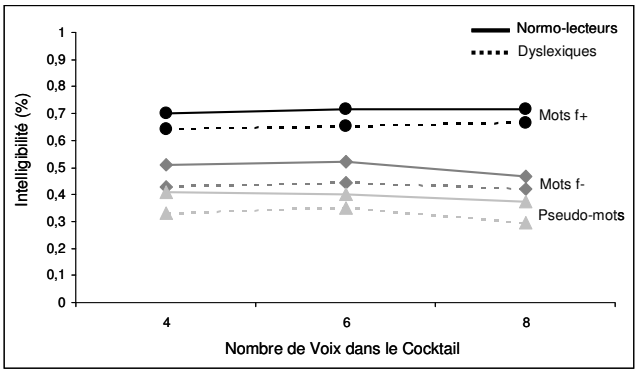


Figure 1 : Performances d'intelligibilité (%) des mots cibles fréquents (f+) et peu fréquents (f-), et des pseudo-mots cibles pour les normo-lecteurs et les dyslexiques en fonction du nombre de voix dans le cocktail.

2.3. Discussion

Les résultats montrent un déficit de compréhension de la parole dans la parole chez les dyslexiques adultes, à la fois lorsque le signal cible est un mot et un pseudo-mot. Si leurs performances sont globalement inférieures à celles des normo-lecteurs, les dyslexiques sont néanmoins sensibles aux propriétés psycholinguistiques des mots tout autant que les normo-lecteurs. Les performances des deux populations en situation de *cocktail party* sont ainsi influencées par la lexicalité du signal cible (les mots sont plus faciles à identifier que les pseudo-mots) et par la fréquence des mots cibles (les mots fréquents sont plus faciles à identifier que les mots peu fréquents). La taille de ces effets lexicaux est par ailleurs remarquablement similaire entre les deux groupes de participants. Dans l'ensemble, ces résultats suggèrent que les dyslexiques adultes ne s'appuient pas sur l'information lexicale de manière plus importante que les normo-lecteurs lors de l'identification de la parole dans la parole. Les dyslexiques n'utiliseraient donc pas de stratégies lexicales de compensation spécifiques pour reconnaître les mots en situation bruitée.

3. EVALUATION DU SYSTEME AUDITIF

3.1. Méthode

Participants

Dix-huit dyslexiques ont été sélectionnés parmi ceux ayant participé aux tests comportementaux et 18 participants normo-lecteurs leur ont été apparié en genre (10 femmes), âge (M = 25 ans) et latéralité manuelle (droitiers). Tous les participants étaient volontaires et ont été dédommagés pour leur participation.

Tests auditifs

Les tests auditifs ont été pratiqués dans le service Explorations Fonctionnelles de l'Hôpital Édouard Herriot à Lyon. L'ensemble des tests durait entre 1h et 1h30.

Les mesures auditives comprenaient la vérification du fonctionnement du système auditif périphérique par une audiométrie tonale dans le silence et une tympanométrie. Les mesures audiométriques tonales permettent de vérifier l'intégrité de l'oreille interne en établissant une mesure quantitative du seuil de détection de sons purs pour chaque oreille. La tympanométrie est un examen permettant de mesurer la souplesse du tympan et la présence du réflexe stapédien (contraction du muscle de l'étrier de l'oreille moyenne) qui protège l'oreille interne en cas de sur-stimulation acoustique (à partir de 75 dB en moyenne).

La fonctionnalité du SEOCM a été évaluée de façon non invasive en mesurant son effet suppresseur sur les otoémissions acoustiques provoquées (OEAP), sons produits par les cellules ciliées externes de la cochlée [8]. Nous avons ainsi pu calculer l'atténuation équivalente pour chaque oreille (AE) et un quotient de latéralité (QL) pour chaque participant, qui correspond à la différence d'AE entre les deux oreilles [9]. Une valeur positive indique une prévalence de l'oreille gauche, une valeur négative indique que le système efférent droit est plus inhibiteur que le gauche.

3.2. Résultats

Audiométrie tonale

Tous les participants dyslexiques ainsi que leurs témoins avaient une audition normale, c'est-à-dire des seuils inférieurs ou égaux à 20dB-HL dans la gamme de fréquences [125Hz - 8KHz]. Ils avaient tous une tympanométrie normale et un réflexe stapédien présent avec un seuil supérieur ou égal à 75dB.

Système efférent

Pour les dyslexiques, l'AE de l'oreille droite ne différait pas de celle de l'oreille gauche (AE_{droite} = -3.12, ET = 2.52 ; AE_{gauche} = - 3.26, ET = 2.11 ; F (1, 34) < 1, n.s.). En moyenne, le système auditif des participants dyslexiques était très faiblement latéralisé à gauche avec un quotient de latéralité 'oreille droite – oreille gauche' de 0.13 (ET = 1.47). En revanche, pour les normo-lecteurs, l'AE de l'oreille droite tendait à être plus importante que celle de l'oreille gauche (AE_{droite} = - 2.68, ET = 1.54 ; AE_{gauche} = - 1.85, ET = 0.98 ; F (1, 34) = 3.69, p = .06). En moyenne, le système auditif des témoins était latéralisé à droite avec un quotient de latéralité de - 0.83 (ET = 1.3).

Une analyse inférentielle a permis de mettre en évidence un QL significativement différent entre les deux groupes de participants ($F(1, 34) = 4.29$, $p < .05$).

3.3. Discussion

Les tests d'audiométrie tonale montrent que les participants des deux groupes ont tous des seuils d'audition normaux. Les mesures de système efférent révèlent chez les participants dyslexiques une symétrie des voies auditives descendantes, alors que les normo-lecteurs présentent une latéralisation de ces voies en faveur du côté droit. D'après les travaux de Khalfa et al. [10], un SEOCM efficace est caractérisé par la présence d'AE dans les deux oreilles mais surtout par une asymétrie fonctionnelle d'ailleurs liée à la latéralité manuelle. Étant donné que tous nos participants (dyslexiques et normo-lecteurs) étaient droitiers, nous aurions dû observer une asymétrie du SEOCM en faveur du côté droit. Le fait que ce résultat n'ait été observé que chez le groupe témoin suggère clairement une anomalie de fonctionnement du SEOCM chez notre population dyslexique. Il est cependant notable que cette anomalie ne vient pas d'une faiblesse des AE (AE supérieure pour la population dyslexique) mais bien d'un problème de latéralisation.

4. DISCUSSION GENERALE ET CONCLUSION

Cette étude révèle l'existence d'un déficit de compréhension de la parole dans le bruit chez le dyslexique adulte, confirmant les résultats observés chez les enfants [3], et la persistance de ces difficultés malgré une expertise langagière plus forte et la mise en place possible de stratégies compensatoires chez l'adulte dyslexique. Nos résultats comportementaux ne révèlent aucune particularité stratégique chez les dyslexiques (influence lexicale plus forte ou modulation de l'effet de fréquence lexicale par exemple), et montrent, au contraire, un pattern de sensibilité au bruit et aux paramètres psycholinguistiques parfaitement identique à celui des normo-lecteurs. Ces résultats suggèrent donc un trouble de bas niveau, d'ordre sensoriel, pré-perceptuel.

En accord avec cette interprétation, nous avons observé un déficit d'asymétrie du SEOCM chez le dyslexique adulte, suggérant une latéralisation moins marquée de l'ensemble du système auditif. Ce résultat est en accord avec ceux de Kumar et Vanaja [4] montrant une corrélation entre fonctionnalité du SEOCM et compréhension de la parole dans le bruit pour certains rapports signal/bruit. Il s'accorde également avec les travaux de Veuillet et al. [9] rapportant un défaut d'asymétrie du SEOCM chez des enfants dyslexiques.

Notre étude permet ainsi de mettre en évidence un lien, chez l'adulte dyslexique, entre défaut fonctionnel du système auditif périphérique et déficit comportemental général dans une tâche de compréhension de la parole dans le bruit. La nature de ce lien reste toutefois à élucider. Le SEOCM pourrait en effet être directement impliqué dans le 'débruitage' de sources de parole concurrentes, ce qui resterait à établir directement. Mais il se pourrait également, comme suggéré par Grataloup et al. [11], que le déficit de symétrisation observé au niveau du SEOCM reflète d'autres asymétries fonctionnelles atypiques observées au niveau central chez les dyslexiques.

5. REMERCIEMENTS

Cette étude a été financée par la région Rhône Alpes et par l'European Research Council (ERC; Projet SpiN attribué à F. Meunier).

6. BIBLIOGRAPHIE

[1] P.N. Sanchez and D. Coppel. Adult outcomes of verbal learning disability. *Seminar in Clinical Neuropsychiatry*, 5(3):205-209, 2000.

[2] F. Ramus and G. Szenkovits. What phonological deficit? *Quarterly Journal of Experimental Psychology*, 61(1):129-141, 2008.

[3] J.C. Ziegler, C. Pech-Georgel, F. George, and C. Lorenzi. Speech-perception-in-noise deficits in dyslexia. *Developmental Science,* 12(5):732-745, 2009.

[4] A. Kumar and C.S. Vanaja. Functioning of olivocochlear bundle and speech perception in noise. *Ear and Hearing*, 25:142-146, 2004.

[5] J. de Boer and A.D. Thornton. Neural correlates of perceptual learning in the auditory brainstem: efferent activity predicts and reflects improvement at a speech-in-noise discrimination task. *Journal of Neuroscience*, 28:4929-4937, 2008.

[6] C. Grataloup. *La reconstruction cognitive de la parole dégradée: Etude de l'intelligibilité comme indice d'une capacité cognitive humaine.* Thèse de Doctorat, Université Lyon 2, 2007.

[7] B. New, C. Pallier, M. Brysbaert and L. Ferrand. Lexique 2: A New French Lexical Database. *Behavior Research Methods, Instruments and Computers,* 36(3):516-24, 2004.

[8] D.T. Kemp. Evoked acoustic emissions from human auditory system. *Journal of Acoustical Society of America*, 64:1386-1391, 1978.

[9] E. Veuillet, F. Bazin and L. Collet. Objective evidence of peripheral auditory disorders in learning-impaired children. *Journal of Audiology and Medicine*, 8:18-29, 1999.

[10] S. Khalfa, E. Veuillet and L. Collet. Influence of handedness on peripheral auditory asymmetry. *European Journal of Neuroscience*, 10(8):2731-2737, 1998.

[11] C. Grataloup, M. Hoen, E. Veuillet, L. Collet, F. Pellegrino and F. Meunier. Speech restoration: An interactive process. *Journal of Speech Language and Hearing Research*, 52(4):827-838.

Déficit de perception catégorielle chez les enfants dysphasiques

Catherine Zobouyan, Josiane Bertoncini & Willy Serniclaes
Laboratoire Psychologie de la Perception, CNRS & Université Paris Descartes

ABSTRACT

Developmental dysphasia is defined by severe specific and delayed oral language development. The aim of this study is to evidence a possible deficit in the categorical perception of speech sounds by dysphasic children. Identification and discrimination responses to stimuli generated by selective modifications of natural /aga/ and /aka/ logotemes were collected in a group of dysphasic children and a group of control children of the same chronological age. The results confirm the predictions and suggest that dysphasic children display an allophonic mode of speech perception similar to the one previously evidenced in dyslexic children.

.

Keywords: Specific language impairment (SLI), speech perception, categorical perception.

1. INTRODUCTION

Les dysphasies du développement se définissent par un trouble sévère, spécifique et primitif du développement du langage oral. Ce déficit comprend différentes formes selon qu'il atteint le langage expressif (dysphasie de type expressif) ou à la fois l'expression et la perception (dysphasie de type mixte). Leurs symptômes communs sont une atteinte de la perception au niveau de la discrimination des sons et de l'expression phonologique et syntaxique [1].

Les enfants dysphasiques présentent des déficits de discrimination des traits « phonologiques » (les traits propres à une langue donnée) dans le bruit, révélant ainsi une organisation atypique des capacités perceptives de l'enfant dysphasique [2]. Ces déficits de perception dans le bruit sont d'ailleurs également présents chez les enfants dyslexiques [3]. Des recherches plus récentes sur les enfants *dyslexiques* ont mis des déficits plus subtils de perception catégorielle pouvant engendrer une mauvaise représentation des catégories phonologiques et par conséquent des troubles d'apprentissage de la lecture [4, 5]. Cette sensibilité aux différences intra-phonémiques provient d'une meilleure discriminabilité des distinctions allophoniques, celles qui correspondent à des traits psychoacoustiques élémentaires et qui peuvent être éventuellement phonémiques dans d'autres langues.

La question soulevée dans la présente étude est de savoir si les enfants *dysphasiques* présentent également un déficit de perception catégorielle et une perception allophonique. Afin de répondre à ces objectifs, nous avons mis en place une expérience de perception auditive de logatomes, visant à tester trois hypothèses :

- comparés aux enfants contrôles, les enfants dysphasiques auraient une perception catégorielle altérée, la perception catégorielle se caractérisant par le fait que les stimuli ne sont discriminables que s'ils qui appartiennent à des catégories différentes, mises en évidence par les résultats d'identification [6].

- d'après l'hypothèse d'un déficit de couplage entre traits « psychoacoustiques » (ceux mis en évidence chez l'enfant pre-linguistique et qui correspondent aux « basic cuts » dans l'espace sonore évoqués par Kuhl [7]), les enfants dysphasiques devraient montrer des performances de discrimination supérieures aux enfants contrôles face à des stimuli dans lesquels les traits psychoacoustiques sont en conflit.

- enfin, par analogie avec les dyslexiques, les enfants dysphasiques pourraient également présenter un déficit de *précision* catégorielle, caractérisé par une pente d'identification plus faible [4].

Dans ce travail, les traits que l'on a mis en conflit sont le VOT positif et l'intervalle de silence à la fin de l'occlusion (IS). Ces deux traits contribuent à la perception du trait phonologique de voisement en français [8]. Nous les considérons comme des traits psychoacoustiques plutôt que comme des indices acoustiques car ils sont traités de manière dichotomique en fonction de seuils auditifs (pour le VOT : seuil localisé à environ 30 ms : discrimination chez l'enfant pre-linguistique : [9] ; potentiels évoqués chez l'adulte : [10]).

2. METHODE

Participants. Le groupe d'enfants dysphasiques comprenait 16 participants (11 garçons) recrutés à l'hôpital de Garches (Dr Picard) et dans le centre Hospitalier du Kremlin Bicêtre (Dr Billard), francophones, sans handicaps sensori-moteur ou troubles psychologiques, âgés de 9;2 ans en moyenne (ET =1;58). Le groupe contrôle comprenait 36 enfants contrôles, francophones, sans handicap sensori-moteurs ou psychologiques appariés en âge avec les enfants dysphasiques (M= 9;4 ans, ET=1;29).

Stimuli. (Figure 1). Les stimuli auditifs ont été réalisés par modification sélective de la parole naturelle à partir de 2 logatomes de type VCV (voyelle consonne voyelle) /aga/ et /aka/ prononcés par une locutrice francophone et extraits du corpus utilisé dans [2]. Un continuum de voisement a été réalisé à partir des énoncés originaux /aga/ et /aka/. A l'aide d'un éditeur de signal (Praat), les vibrations périodiques présentes durant l'occlusion et après la détente de l'occlusion de l'occlusive voisée /g/ ont été progressivement remplacées par des segments non-périodiques prélevés dans l'occlusive non-voisée /k/, en partant de la détente. L'IS et le VOT positif sont nuls pour le logatome /aga/ choisi, et respectivement de 215 et 39 ms pour /aka/.

Le continuum utilisé comporte 4 stimuli, dont l'un des points extrêmes (S1) correspond au logatome /aga/ non modifié et l'autre à ce même logatome après insertion de l'IS et VOT du logatome /aka/. Les 2 valeurs intermédiaires (S2 et S3) ont des valeurs d'IS et de VOT correspondant au tiers (33% de 215 et 39 ms, soit 72 et 13 ms) et aux deux tiers (67 % de 215 et 39 ms, soit 143 et 25 ms) de leurs valeurs dans /aka/.

Deux stimuli supplémentaires (S5 et S6) ont été construits en mettant l'IS et le VOT en conflit. S5 est construit avec une valeur d'IS de 215 ms (celui de /aka/) et un VOT de 0 ms (celui de /aga/). Le S6 est construit de manière inverse : l'IS est de 0 ms comme dans /aga/) et le VOT est de 39 ms (celui de /aka/). Les deux segments vocaliques, initial et final, sont ceux de /aga/ tant pour S5 que pour S6.

Procédure L'expérience élaborée à partir du programme Superlab Pro. Elle comporte une tâche d'identification et une tâche de discrimination catégorielle Pour l'ensemble des tâches, l'enfant est placé en face de l'écran et doit répondre en appuyant sur un des deux boutons reliés au clavier Azerty de l'ordinateur. Des étiquettes près de chaque bouton rappellent à quel stimulus ils correspondent.

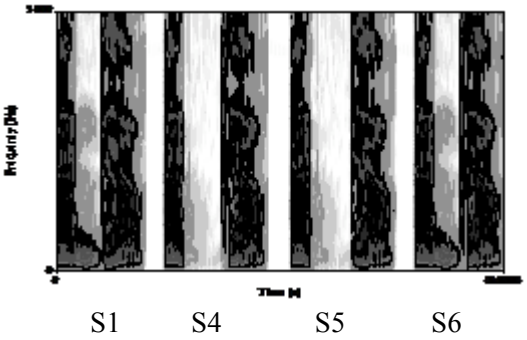


Figure 1 : Spectrogrammes des stimuli VCV : S1, S4 (extrema du continuum IS+VOT) et S5, S6 (IS et VOT en conflit).

Lors de la tâche d'identification, l'enfant doit identifier des sons en choisissant entre 2 réponses /aga/ ou /aka/. 6 stimuli sont présentés 6 fois en série aléatoire, soit 36 présentations. La tâche de discrimination utilise 4 paires de stimuli différents (S1S2, S2S3, S3S4, S5S6), 4 paires différentes dans l'ordre inverse (S2S1, S3S2, S4S3, S6S5) et 6 paires identiques (S1S1, S2S2, S3S3, S4S4, S5S5, S6S6). Pour cette tâche, les 14 paires de stimuli seront présentées 4 fois en série aléatoire, soit 56 présentations au total.

Traitement des données. Les pentes des fonctions d'identification ont été estimées séparément pour chaque participant par Régression Logistique.

3. RESULTATS

Les fonctions d'identification relatives au continuum obtenu par allongement progressif de l'IS et du VOT sont présentées dans la Fig. 2 pour chaque groupe d'enfants. Par rapport aux enfants contrôles (CTL dans la suite du texte), la pente de la fonction d'identification des enfants dysphasiques (DYS dans la suite de ce texte) est plus faible et les réponses recueillies aux stimuli extrêmes plafonnent à des valeurs inférieures à 100 % de réponses /aga/ ou /aka/. Cependant, l'effet du groupe sur les pentes des fonctions, n'est pas significatif ($F(1,50)=1,39$; $p=0,24$). Toutefois, la différence de plafonnement entre groupes, testée par ANOVA à mesures répétées Stimulus (2 extrema du continuum: S1, S4) x Groupe (CTL et DYS) avec les scores d'identification (après transformation en arcsinus racine) comme variable indépendante, est marginalement significative (interaction Stimulus x Groupe: $F(1,50) = 3.93$, $p=.05$).

Les scores de discrimination observés et ceux attendus à partir des scores d'identification sont présentés dans les Figs. 3 et 4 pour les trois paires du continuum IS-VOT (S1S2, S2S3, S3S4) et la paire S5S6 avec l'IS et le VOT en conflit. La comparaison entre scores observés et attendus permet de tester la perception catégorielle.

L'examen de la Figure 3 montre un écart entre les scores observés et prédits pour les enfants dysphasiques, en particulier pour la paire avec l'IS et le VOT en conflit (S5S6), mais non pour les enfants contrôles. Une ANOVA à mesures répétées Paire (4 valeurs: S1S2, S2S3, S3S4, S5S6) x Tâche (Discrimination vs. Identification) x Groupe (CTL vs. DYS) met en évidence des effets significatifs de la Paire ($F(3,150) = 117$, $p<.001$), de la Tâche ($F(1,50) = 10.3$, $p<.01$), mais non du Groupe ($F<1$).

Les interactions Tâche x Paire et Tâche x Groupe sont également significatives (respectivement: $F(3,150) = 3.16$, $p<.05$; $F(1,50) = 6.00$, $p<.05$), les autres interactions étant non significatives ($F<1$).

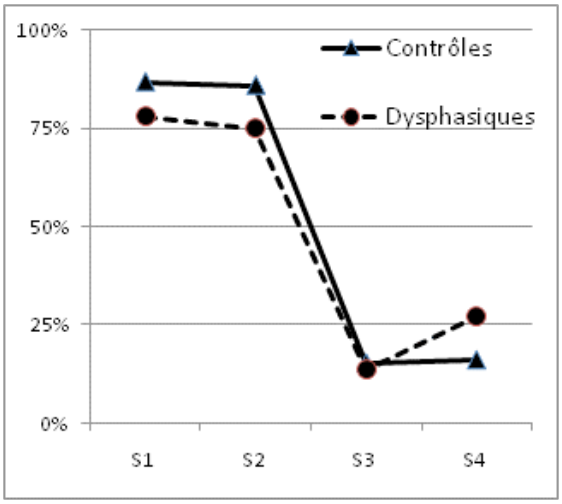


Figure 2: Pourcentage de réponses d'identification « aga » en fonction de l'allongement de l'IS et du VOT.

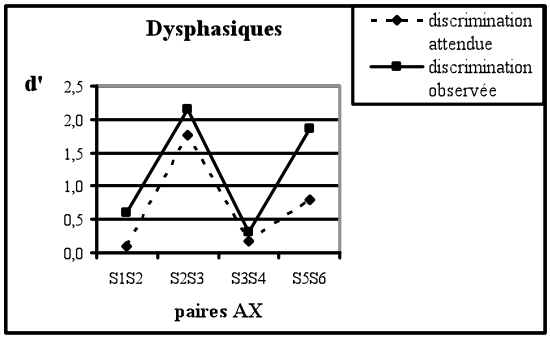
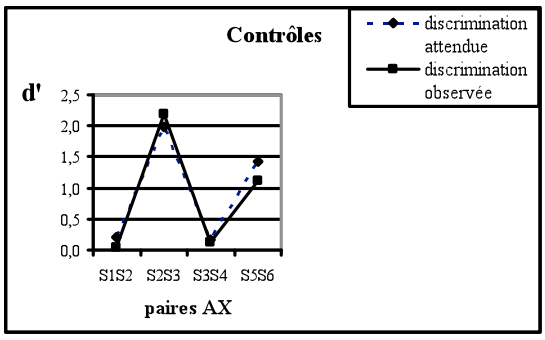


Figure 3



Figures 3 et Figure 4 : *Scores de discrimination Même-différent (transformées d')* pour les enfants dysphasiques (Fig.3) et les enfants contrôles (Fig.4).

4. Discussion

Les enfants dysphasiques présentent à la fois un déficit de précision catégorielle et de perception catégorielle par rapport aux contrôles. Mais chacun de ces déficits s'est manifesté de manière peu classique.

Les résultats d'identification des stimuli du continuum généré par co-variation de lS et du VOT ne montrent pas de différence de pente d'identification, cette dernière constituant un critère classique de précision catégorielle. Cependant, un autre indicateur de précision catégorielle, le plafonnement plus important des scores d'identification, met en évidence un déficit de précision chez les enfants dysphasiques. Quoiqu'encore peu répandu à l'heure actuelle (voir cependant [11]) cet indicateur de précision est mieux approprié à la forme « hypersigmoïdale » des fonctions d'identification de stimuli complexes. Par rapport à la fonction sigmoïdale simple, la fonction hypersigmoïdale se caractérise par une pente accrue au voisinage de la frontière et des décélérations de pente plus marquées aux extrema du continuum [11].

D'autre part, nos résultats vont dans le sens d'une perception moins « catégorielle » des enfants dysphasiques comparés aux témoins comme le montre la différence entre scores d'identification observés et prédits (interaction Tâche x Groupe significative). Cependant, ce déficit de perception catégorielle se manifeste également de manière assez inattendue. Alors que nous nous attendions à une meilleure performance pour les seules paires intra catégorielles, elle se manifeste également pour la paire inter catégorielle (S2S3). Ces résultats mettent donc en évidence un déficit de perception catégorielle chez l'enfant dysphasique, semblable aux enfants dyslexiques : en effet, la discrimination intra-catégorielle est meilleure chez les dysphasiques. Mais contrairement à ce qui avait été constaté chez les dyslexiques [5], la discrimination inter catégorielle est également plus élevée chez les dysphasiques.

Tant le déficit de précision catégorielle que celui de perception catégorielle vont à l'appui de nos hypothèses. De plus, certains aspects des résultats vont en faveur d'une perception "allophonique". Le déficit de perception catégorielle des enfants dysphasiques tend en effet à être particulièrement élevé lorsque des indices correspondants à deux traits différents sont en conflit (Figs. 3 et 4 : l'écart entre scores de discrimination observé et prédit est plus important pour la paire S5S6), bien que cette tendance ne soit pas significative (interaction Tâche x Paire x Groupe non significative).

5. Conclusion

Les points essentiels de nos résultats sont les suivants :
- Les enfants dysphasiques présentent une frontière d'identification légèrement décalée et un plafonnement plus important de la fonction d'identification aux extrêmes du continuum. Ce plafonnement des réponses met en évidence un déficit de précision catégorielle.

- Chez les enfants dysphasiques, les scores de discrimination observés sont plus élevés que ceux prédits par l'identification contrairement à ce que l'on observe chez les contrôles pour lesquels les deux types de scores sont très proches. Ceci dénote un déficit de perception catégorielle chez les dysphasiques.

- Ce déficit tend à être plus marqué lorsque les traits qui contribuent à la même distinction phonologique sont en conflit, ce qui va à l'appui d'une perception « allophonique ». Toutefois cette tendance n'est pas significative et demande confirmation.

Le développement de la précision catégorielle des traits phonologiques est un processus long et progressif qui s'étend jusqu'au début de l'adolescence [12]. Le déficit de précision des enfants dysphasiques correspond à un retard développemental plutôt qu'à une déviance. Par contre, le développement de la perception catégorielle est plus précoce et semble être acquis vers l'âge d'un an chez les enfants typiques [12]. Le déficit de perception catégorielle est donc le plus préoccupant car son acquisition pourrait être beaucoup plus difficile en dehors d'une période sensible relativement précoce. Enfin, comme ce déficit affecte l'économie des processus de discrimination des sons de la parole, en ne « filtrant » pas suffisamment les différences phonologiquement non pertinentes, il pourrait entraver le développement du langage et être à l'origine de certains symptômes de la dysphasie.

BIBLIOGRAPHIE

[1] D.V.M. Bishop. Using mismatch negativity to study central auditory processing in developmental language and literacy impairments : Where are we, and where should we be going. *Psychological Bulletin*, Vol. 133, 651-672, 2007.

[2] J.C Ziegler., C. Pech-Georgel., F. George, F.X Alario and C. Lorenzi. Deficits in speech perception predict language learning impairment. *PNAS (Proceedings of the National Academy of Sciences of the United States of America)*, *39*: 14110-14115, 2005.

[3] J. Ziegler., C. Pech-Georgel., F. George and C. Lorenzi. Speech perception in noise deficits in dyslexia. *Developmental Science*, 12 : 732-745, 2009.

[4] W. Serniclaes., S. Van Heghe., Ph. Mousty., R. Carré and L. Sprenger-Charolles. Allophonic mode of speech perception in dyslexia. *Journal of experimental Child Psychology, 87*: 336-361, 2004.

[5] O. Dufor., W. Serniclaes., L. Sprenger-Charolles and J.F. Demonet. Left premotor cortex and allophonic speech perception in dyslexia : a PET study. *Neuroimage,* 46: 1, 241-248, 2009.

[6] A.M. Liberman., K.S. Harris., H.S. Hoffman and B.C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psych.* 54 : 358-368, 1957.

[7] P.K. Kuhl. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience, 5,* 831–843, 2004.

[8] W. Serniclaes. *Etude expérimentale de la perception du trait de voisement des occlusives du français.* Thèse de doctorat en Sciences psychologiques. Unpublished doctoral thesis. Université Libre de Bruxelles, 1987. http://lpp.psycho.univ-paris5.fr/person.php?name=WillyS

[9] I. Hoonhorst., C. Colin., E. Markessis., M. Radeau., P. Deltenre and W. Serniclaes. French native speakers in the making: from language-general to language-specific voicing boundaries. *Journal of Experimental Child Psychology, 104,* 353-366, 2009.

[10] I. Hoonhorst., W. Serniclaes., G. Collet., C. Colin., E. Markessis., M. Radeau and P. Deltenre. The acoustic correlates of voicing perception in French. *Clinical Neurophysiology*, *120,* 897-903, 2009.

[11] M. Treisman. There are two types of psychometric function: A theory of cue combination in the processing of complex stimuli with implications for categorical perception. *Journal of Experimental Psychology (General), 128,* 517-546, 1999.

[12] M. Medina., I. Hoonhorst and C. Bogliotti and W.Serniclaes,W. (accepted). Development of voicing perception in French: Comparing adults, adolescents and children. *Journal of Phonetics*.

Comparaison d'analyses phonétiques de parole dysarthrique basées sur un alignement manuel et un alignement automatique

Cécile Fougeron[1], Nicolas Audibert[2], Corinne Fredouille[2], Christine Meunier[3], Cédric Gendrot[1], Olavo Panseri[1]

[1] Lab. de Phonétique et Phonologie, UMR 7018 CNRS-Paris3/Sorbonne Nouvelle, Paris, France
[2] Université d'Avignon, CERI/LIA, Avignon, France
[3] Laboratoire Parole et Langage, CNRS, Université Aix-Marseille, France
cecile.fougeron@univ-paris3.fr, corinne.fredouille@univ-avignon.fr, christine.meunier@lpl-aix.fr

ABSTRACT

The reliability of an automatic speech alignment procedure for the phonetic description of dysarthric speech is assessed through the comparison of durational and spectral measurements obtained from an automatic and a manual alignment of the production of 4 dysarthric speakers varying in severity. Results show that formant values computed in the middle of the vowel intervals and center of gravity of fricative noise computed over the consonant intervals, are reliable when based on automatic alignments. However, the analysis of pause occurrences and absolute segmental duration require manual corrections of the automatic outputs.

Keywords: dysarthric speech, phonetico-acoustic study, automatic vs. manual alignment.

1. INTRODUCTION

La dysarthrie est un terme générique définissant un trouble de la parole d'origine motrice consécutif à une atteinte du système nerveux central et/ou périphérique. En fonction de la localisation de l'atteinte dans le cerveau, de la sévérité de la maladie associée, ou de particularités propres au locuteur, les caractéristiques de la dysarthrie varient. Ainsi, si toutes les dimensions de la parole (phonation, articulation, timing, prosodie, fluence…) peuvent être altérées, elles le sont à des degrés variables en fonction des patients. De plus, en fonction de la nature du trouble moteur, les altérations sur ces dimensions n'ont pas la même forme : un mauvais contrôle temporel des mouvements affectera la durée et les transitions entre segments, alors qu'une perte de force musculaire affectera plutôt l'amplitude de mouvement et l'atteinte des cibles articulatoires. Au vu de cette variabilité des formes de dysarthrie, nos recherches ont pour but de décrire finement les profils dysarthriques sur la base de leurs caractéristiques phonético-acoustiques, en isolant des critères robustes, quantifiables et surtout plus objectifs que des critères perceptifs dont se servent les classifications antérieures [1].

Les travaux présentés ici visent à établir une procédure d'analyse optimale en termes de temps et d'expertise humaine permettant l'analyse phonético-acoustique d'un grand nombre d'échantillons de parole dysarthrique pour faire face à la variabilité inter- et intra-locuteurs importante dans ce type de populations. La segmentation manuelle d'un continuum de parole est extrêmement coûteuse en temps et en expertise, et les altérations phonétiques dans la dysarthrie rendent ce travail encore plus ardu. Le recours à une segmentation automatique des productions apparaît donc comme une alternative des plus intéressantes. Pour autant, on sait que les systèmes d'alignement automatique peuvent générer des erreurs dans la localisation des frontières de phonèmes sur de la parole normale. Il est donc nécessaire d'une part, d'évaluer ces erreurs par rapport à une segmentation manuelle (voir Audibert et al. [2]), et d'autre part d'évaluer l'adéquation d'une telle approche en regard de la validité des analyses phonético-acoustiques qu'elle permet dans le contexte spécifique de la parole dysarthrique.

Notre objectif ici est donc d'évaluer la validité d'une analyse phonétique basée sur un alignement automatique de productions dysarthriques, en la comparant à une analyse basée sur un alignement manuel de référence. Il s'agira de savoir si les analyses phonétiques faites à partir de l'alignement automatique sont suffisamment proches de celles effectuées à partir d'un alignement manuel pour décrire les mêmes tendances. En fonction du type de critère phonétique étudié, nous chercherons à déterminer si l'on peut reposer l'analyse sur un alignement tout-automatique ou si celui-ci nécessite une phase préalable de vérification et correction manuelle.

2. METHODE

2.1. Corpus et locuteurs

Le corpus est constitué d'une partie du texte 'Tic Tac' de la batterie de C. Chevrie-Müller lu par quatre patients dysarthiques atteints de maladies rares de surcharges lysosomales. Ces enregistrements font partie d'une étude en collaboration avec F. Sédel et N. Lévêque (Hôpital de la Pitié Salpêtrière). Les locuteurs, 2 hommes et 2 femmes, présentent des dysarthries de type mixte à des degrés de sévérité différents. Les patients M1A, F1C (avec M pour homme et F pour femme) sont légèrement dysarthriques ; les patients M2V, F2S ont des dysarthries sévères, marquées plus particulièrement par des altérations de la qualité vocale

et un débit articulatoire ralenti chez M2V, et des altérations d'articulation consonantique et vocalique chez F2S.

2.2. Alignements automatique et manuel

Dans ce travail, le système automatique d'alignement contraint par le texte, développé par le Laboratoire Informatique d'Avignon (LIA), est utilisé. Ce système repose sur l'utilisation classique de modèles de Markov Cachés (38 modèles HMM indépendants du contexte estimés sur le corpus d'émissions radiophoniques ESTER, [3]), associés à un algorithme de décodage de type Viterbi [4]; une description plus détaillée du système est donnée dans Audibert *et al.* [2].

Les échantillons de parole étudiés ont été retranscrits manuellement sous une forme orthographique de façon à inclure toutes les insertions, suppressions, substitutions et répétitions produites par les patients par rapport au texte d'origine. Le lexique phonétisé, utilisé en entrée du système automatique, a été restreint aux seules entrées lexicales du texte lu, puis adapté dynamiquement à chaque transcription orthographique pour prendre d'éventuelles entrées manquantes (dues à des substitutions ou faux départs par exemple). Finalement, les variantes de prononciation de chaque entrée lexicale ont été vérifiées de façon à n'inclure que celles possibles dans le texte. Sur la base de la transcription orthographique fournie et du lexique phonétisé, le système automatique va analyser le signal de parole et identifier les frontières phonémiques (sous la forme d'étiquettes de début et de fin) de la séquence de phonèmes attendue.

Un alignement manuel (AM) a ensuite été réalisé par un expert humain sur la base de l'alignement automatique. Sa tâche consistait, d'une part, à vérifier la réalisation des phonèmes transcrits (donc éventuellement à changer, insérer ou supprimer des étiquettes de phonèmes) et, d'autre part, à déplacer les étiquettes de début et de fin lorsqu'il le jugeait nécessaire par rapport au signal produit. Le placement des frontières a été fait sur la base de critères de segmentation couramment utilisés : l'apparition et la disparition du 2^{e} formant pour les voyelles, le bruit caractéristique des fricatives, la tenue voisée ou silencieuse pour les plosives, le bruit correspondant au relâchement des occlusives, etc. Les difficultés majeures pour l'expert ont résidé dans le placement d'une frontière au sein de suites de voyelles et de consonnes de type vocalique ou sonant (comme dans "horloge", par exemple). Dans les cas les plus difficiles, l'expert a codé les portions de signal comme 'insegmentables'.

2.3. Procédure de comparaison

Afin de comparer les procédures manuelle et automatique sur les mêmes portions de signal, seuls les phonèmes segmentés dans les deux alignements, avec une étiquette phonémique similaire, ont été conservés. Les segments insérés ou supprimés dans l'un ou l'autre des alignements ont donc été exclus, ainsi que les parties du signal que l'expert humain a jugé comme 'insegmentables'. Les segments retenus ont été regroupés en grandes classes phonétiques par locuteur, et les classes présentant moins de 10 exemplaires par locuteur ont été éliminées (consonnes et voyelles nasales, semi-voyelles). Au final, 901 segments ont été retenus et leur distribution par classe phonétique est indiquée dans le tableau 1.

Tableau I : Nombre d'occurrences et distribution des segments comparés par locuteur et classe phonétique. F : fricative, O : occlusive, (b) : burst, (t) : tenue, Vo : voyelle orale, S : sourde, V : voisée, # : pause

Loc	FS	FV	OSb	OSt	OV	/R/	/l/	Vo	#
M1A	11	15	24	18	14	13	15	67	10
F1C	12	17	23	22	15	17	18	85	25
F2S	14	18	19	19	18	20	16	89	30
M2V	12	17	26	25	16	18	15	79	29

2.4. Critères phonétiques étudiés/comparés

Afin d'évaluer si l'utilisation d'un alignement automatique est adaptée pour l'étude de propriétés phonetico-acoustiques de la parole dysarthrique, différentes mesures acoustiques obtenues à partir de l'alignement manuel (AM) servent de référence. Elles sont comparées aux mesures obtenues à partir de l'alignement automatique (AA). Pour chaque mesure, l'effet du type d'alignement (AA vs. AM), et les interactions avec les facteurs 'locuteur' et 'classe phonétique' sont testés à l'aide d'ANOVAs.

Les comparaisons se basent sur quatre types d'analyses : mesures de la durée et du nombre de pauses, des durées segmentales, des fréquences des formants F1 et F2 des voyelles, du centre de gravité spectral (CoG) du bruit des fricatives. Le choix de ces mesures répond aux critères suivants. Premièrement, ces mesures peuvent caractériser différents aspects de la parole pouvant être altérés dans la dysarthrie : la fluence (pauses), la prosodie (durée et pauses), l'articulation des voyelles (durée, formants) et des consonnes (durée, CoG pour les fricatives). Ces mesures ont ainsi pu être utilisées dans la littérature dans le cadre de la description de parole pathologique (voir les revues dans [5] et [6]). Deuxièmement, elles touchent à des dimensions acoustiques différentes : spectrales pour les formants et le CoG, temporelles pour les durées. Un décalage temporel entre les étiquettes de l'AA et celles de l'AM peut avoir des répercussions sur des mesures de durée absolue des segments mais permettre la comparaison entre phonèmes ou entre locuteurs si les décalages sont systématiques. Les deux mesures spectrales sont calculées sur des empans différents : une mesure locale pour les formants pris au centre des voyelles, une mesure globale pour le CoG mesurée sur la fenêtre temporelle des fricatives.

3. RESULTATS

3.1. Nombre et durée des pauses

La comparaison de la durée des 94 intervalles segmentés comme des pauses (silences) par l'AA et l'AM ne montre pas d'effet du type d'alignement ($F(1,180)=1.59$, $p=.2$) : les durées issues des deux alignements sont similaires et ceci pour tous les locuteurs (interaction $F(3,180)=.82$, $p=.48$).

Pour autant, l'adéquation de l'alignement automatique pour l'analyse des pauses est illusoire. En effet, si toutes les pauses relevées par l'AM ont été détectées par l'AA, l'inverse n'est pas vrai. L'AA insère des pauses qui ne sont pas notées par l'expert humain et leur proportion n'est pas négligeable (64 insertions erronées sur le total des 4 locuteurs, pour 94 pauses réelles). Ces pauses apparaissent généralement dès lors que l'AA rencontre des difficultés pour aligner la séquence de phonèmes attendue sur le signal (soit parce que les phonèmes sont très dégradés, soit parce qu'ils sont trop longs et, par conséquent, qu'ils ne correspondent plus aux modèles de phonèmes du système). La présence de pauses optionnelles après chaque mot dans le lexique phonétisé offre la possibilité au système d'insérer une pause pour résoudre les incohérences temporelles qu'il rencontre. L'avantage de ce procédé est d'éviter de répercuter des erreurs d'alignement au-delà du mot. Néanmoins, ce dernier se fait au prix d'insertions erronées de multiples pauses dans l'alignement qu'il est donc nécessaire de corriger.

3.2. Durées segmentales

La comparaison effectuée sur les durées de 901 phonèmes extraites à partir des segmentations de l'AA et de l'AM montre un effet significatif du type d'alignement ($F(1,1550)=94.2$; $p<.0001$) : les durées de l'AA sont globalement plus courtes. Pour autant, cet effet varie en fonction du locuteur (interaction aligneur*locuteur $F(3, 1550)=20.15$, $p<.0001$) et de la classe de phonèmes (interaction aligneur*classe $F(7,1550)=18.28$, $p<.0001$).

En ce qui concerne les locuteurs, un effet du type d'alignement est trouvé chez les deux patients ayant une dysarthrie sévère (F2S et M2V) mais aussi chez la patiente à dysarthrie légère (F1C). Chez ces trois locuteurs, il y a une interaction avec la classe de phonèmes. Chez le locuteur H1A, les sorties de l'AA sont comparables à celles de l'AM.

Les résultats des analyses concernant les différentes classes de phonèmes sont illustrés sur la figure 1. Pour l'ensemble des fricatives l'AA donne des durées significativement plus courtes. Mais cet effet varie en fonction du locuteur. Comme illustré sur la figure 1, l'effet est notable pour les deux patients les plus dysarthriques (F2S et M2V) pour toutes les fricatives, et uniquement pour les fricatives voisées pour F1C. Pour les occlusives voisées et pour les voyelles, les durées de l'AA sont également significativement plus courtes et ceci pour tous les locuteurs. La tenue des occlusives est significativement plus longue dans l'AA avec une différence notable chez les deux patients les moins dysarthriques (H1A et F1C). L'explosion des occlusives est, quant à elle, plus courte dans l'AA chez tous les patients (mais avec un degré variable). Seules les durées des liquides /l/ et /ʁ/ ne diffèrent pas entre l'AA et l'AM. Ceci est particulièrement intéressant car ces segments sont souvent difficiles à délimiter sur le signal, même par des experts humains.

3.3. Formants des voyelles

Les occurrences de voyelles dans la partie du texte produite par nos patients ne sont pas nombreuses. Pour notre comparaison, nous avons élargi notre critère d'inclusion à au moins 9 exemplaires par type de voyelles pour chaque locuteur, de façon à pouvoir comparer 81 /a/, 43 /i/, 71 /e,ɛ/, 45 /o, ɔ/.

La comparaison des fréquences formantiques obtenues pour F1et F2 au centre de la voyelle ne montre pas d'effet du type d'alignement, ni d'interaction avec les facteurs 'locuteur' et 'type de voyelles'. Dans [2] nous montrons que le centre de la voyelle de l'AA et le centre de la voyelle de l'AM peuvent être décalés. Afin de déterminer quelles seraient les mesures optimales à appliquer sur la segmentation de l'AA, nous avons comparé les mesures formantiques prises au centre de l'AM (comme référence) à celles prises aux 1/3, 1/2 et 2/3 de l'intervalle vocalique segmenté par l'AA, et à une valeur moyenne calculée sur les 3 points. Pour chaque comparaison, une bonne corrélation est obtenue ($r=.9$ pour les mesures à 1/2, $r=.7$ pour 1/3 et $r=.7$ pour 2/3, $r=.9$ pour la moyenne), mais il semble toutefois plus prudent de prendre des mesures formantiques au centre de la voyelle ou une moyenne sur trois points avec un alignement automatique.

3.4. Centre de Gravité du bruit des fricatives

Nous avons comparé les mesures de CoG issues des deux alignements pour les fricatives en les regroupant selon des contrastes de lieu. On sait que le bruit des dentales est plus aigu que celui des fricatives labiales et post-alvéolaires. Afin d'avoir un nombre suffisant d'occurrences (au moins 10 par locuteur dans chaque catégorie), nous avons distingué articulation dentale ($N=92$) et articulation non-dentale ($N=140$). La comparaison des valeurs de CoG ne montre pas d'effet du type d'alignement. Les fricatives dentales ont un CoG plus haut pour tous les locuteurs et dans les deux alignements comme illustré sur la figure 2.

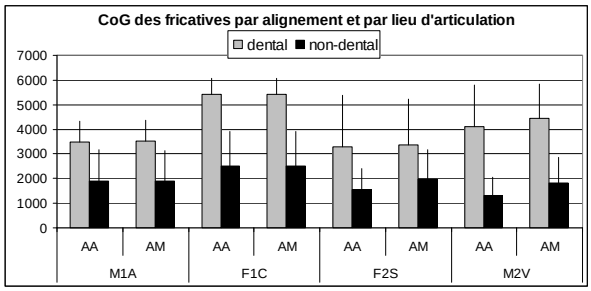


Figure 2 : Centre de gravité spectral (CoG en Hertz) des fricatives selon leur lieu d'articulation pour chaque patient et selon le type d'alignement.

4. DISCUSSION ET CONCLUSION

Il ressort de nos résultats que l'application d'une procédure d'alignement automatique pour l'analyse de critères phonético-acoustiques sur la parole dysarthrique est tout à fait envisageable, même si, en fonction des critères étudiés, elle nécessitera une vérification manuelle. Pour l'étude des pauses, nous avons vu que les segmentations automatiques doivent être vérifiées pour éliminer les insertions erronées de pauses. Pour l'étude de la durée des segments, la fiabilité d'un alignement automatique doit être appréciée en regard du type de segment à étudier et en fonction de la précision temporelle requise pour l'objet d'étude. Si l'on peut se reposer sur l'AA pour l'étude de la durée des liquides, une correction/vérification manuelle sera requise pour les autres segments. Par contre l'utilité d'une segmentation automatique n'est pas à exclure si l'étude des durées segmentales sert à examiner des contrastes entre types de phonèmes ou entre locuteurs. En effet, le contraste de durée entre fricatives sourdes et sonores apparaît aussi bien avec l'AM qu'avec l'AA (les fricatives sourdes sont plus longues que les sonores dans les deux alignements). De même, l'allongement des durées segmentales particulièrement important chez le patient M2V ressort dans les deux alignements. En ce qui concerne les mesures spectrales, les analyses basées sur une segmentation automatique semblent fiables aussi bien pour les formants des voyelles que pour les mesures de CoG sur le bruit des fricatives, segments dont la durée est sous-estimée par l'AA. Ces résultats sont particulièrement encourageants car ils montrent que ces analyses spectrales peuvent être effectuées directement sur l'AA, sans correction manuelle. Par ailleurs, la fiabilité de l'alignement automatique semble dépendre de la sévérité de la dysarthrie des locuteurs. En effet, les différences observées entre AA et AM sont plus fréquentes et importantes chez les patients les plus dysarthriques H2V et F2S. Toutefois, il faut noter que les experts humains sont également en difficulté face au signal de parole particulièrement altéré chez ce type de patients. Le nombre de séquences jugées insegmentables par les experts en témoigne.

Remerciements : *Ce travail est financé par l'ANR-08-BLAN-0125 et l'association Vaincre les Maladie Lysosomales. Nous remercions Georges Linares du LIA pour son aide sur le système d'alignement automatique.*

BIBLIOGRAPHIE

[1] F. L. Darley, A. E. Aronson and J. R. Brown. Clusters of Deviant Speech Dimensions in the Dysarthrias. *JSHR*, 12: 462-496, 1969.

[2] N. Audibert, C. Fougeron, C. Fredouille, C. Meunier and O. Panseri. Evaluation d'un alignement automatique sur la parole dysarthrique. In *Actes XXVIIIèmes JEP*, 2010.

[3] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa and K. Choukri. Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proc. LREC'06*, 2006.

[4] F. Brugnara, D. Falavigna and M. Omologo. Automatic segmentation and labelig of speech based on hidden Markov models, *Speech Communication*, 12: 357-370, 1993.

[5] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian and J. R. Duffy. Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of Com. Disorders*, 32/3:141-186, 1999.

[6] B. E. Murdoch. *Dysarthria*: *A physiological approach to assessment and treatment*. Stanley Thornes, Cheltenham, UK, 1998.

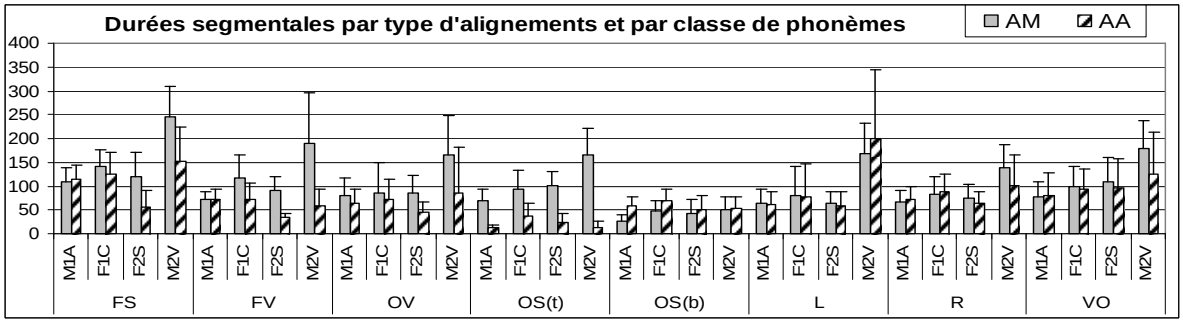


Figure 1 : durées (en ms) selon le type d'alignement pour chaque patient et pour chaque classe de phonèmes

Débit de parole dans les dysarthries de la maladie de Wilson

- Etude de l'influence des troubles attentionnels et dysexécutifs en condition de double tâche -

Michaela Pernon(1,2), Jean-Marc Trocello(1), Jacqueline Vaissière(2), Cécile Fougeron(2), Alice de Tassigny(1), Catherine Cousin(1), Gérard Chevaillier(1), Pascal Rémy(1), France Woimant(1)

(1) CNR Wilson, Service de Neurologie, Hôpital Lariboisière, 2 rue Ambroise Paré, 75475 Paris Cedex 10.
(2) Lab. de Phonétique et de Phonologie, UMR 7018, CNRS-Université Paris 3, 19 rue des Bernardins, 75005 Paris.
michaela.pernon@lrb.aphp.fr

ABSTRACT

The control of speech rate by dysarthric patients with Wilson's disease is examined in this paper, as well as their abilities in dual task conditions. Different patterns of speech rate are found according to the different neurological profiles of the patients. All patients show a slower speech rate compared to control subjects. Akinetic-rigid and ataxic patients present a deterioration of their acceleration abilities. In dual task, control subjects and dystonic patients significantly increase their speech rate. These results allow to consider speech rate modulation and executive processes for the evaluation and rehabilitation of patients with various dysarthric profiles.
Keywords: dysarthria, speech rate, Wilson's disease, attention, dual task.

1. INTRODUCTION

La maladie de Wilson est une affection génétique rare, liée à des mutations du gène ATP7B entraînant des troubles du métabolisme du cuivre dans de nombreux organes, dont le cerveau (noyaux gris centraux, noyaux du cervelet, thalamus et tronc cérébral) (Trocello et al. [12], Woimant [14]). Les principaux tableaux neurologiques associent dystonie (contractions inadaptées par atteinte des noyaux gris centraux), tremblement-ataxie (syndrome cérébelleux, par atteinte du cervelet) et/ou syndrome akinéto-rigide (syndrome parkinsonien par atteinte nigro-striatale). Les patients atteints de la maladie de Wilson peuvent ainsi présenter des dysarthries mixtes, à prédominance akinéto-rigide, ataxique ou dystonique (Pernon et al. [10]). Elles figurent parmi les manifestations les plus précoces et les plus fréquentes de la forme neurologique de cette maladie (Berry et al. [1], Ghika et al. [6]), qui intègre également des déficits attentionnels et dysexécutifs (Brewer [2]). Les données de la littérature se rapportant aux troubles de la parole dans la maladie de Wilson sont peu nombreuses, du fait sans doute de la rareté de cette pathologie (Berry et al. [1], Dordain et Chevrie-Muller [4], Hefter et al. [7], Volkmann et al. [13]). La plupart des travaux s'attachent à décrire le débit de parole, se révélant d'un patient à l'autre, soit normal, soit accéléré, soit ralenti. Certaines études proposent une classification des dysarthries observées en fonction du débit de parole. Cependant, ces données ne prennent pas en compte la présentation neurologique (akinéto-rigide, ataxique ou dystonique) des patients wiloniens.

De nombreux patients dysarthriques rapportent que leur parole, perturbée, nécessite une attention accrue, notamment lorsqu'ils essaient de parler de la manière la plus intelligible possible ou lorsqu'ils doivent effectuer deux actions simultanément. Leur dysarthrie, entraînant une réduction de l'automaticité de la production de leur parole, rend ainsi plus difficiles les situations de double tâche rencontrées dans la vie quotidienne, contribuant à une augmentation du recours à leurs capacités d'attention divisée. Or, il a été bien établi, dans le cadre d'études portant sur la marche, que les troubles attentionnels et dysexécutifs, observés dans des pathologies impliquant les noyaux gris centraux, telles la maladies de Parkinson (Rochester et al. [11], Yogev et al. [16]), de Huntington (Delval et al. [3]), avaient un impact sur cette fonction motrice lors de paradigmes de double tâche.

Cette étude décrit le débit de parole des patients présentant une maladie de Wilson. Elle a pour objectif de rechercher l'existence d'éventuelles perturbations de ce paramètre prosodique pour chaque groupe de patients, répartis en fonction de leur présentation neurologique et comparés à leurs sujets contrôles respectifs. Le maintien, les capacités d'accélération et de ralentissement du débit de parole sont étudiées dans ce cadre. Cette étude cherche également à établir un éventuel classements des dysarthries wilsoniennes en fonction du débit de parole. En dernier lieu, l'effet de la double tâche sur le débit de parole est examiné dans le cadre d'un paradigme de double tâche. Les résultats obtenus en condition isolée sont ici comparés à ceux obtenus en condition de double tâche.

2. MÉTHODOLOGIE

2.1. Sujets

Dix-sept patients dysarthriques présentant une maladie de Wilson ont été répartis en 3 groupes en fonction de leur présentation neurologique (4 patients ataxiques, 5 patients akinéto-rigides, 8 patients dystoniques). Ils ont été recrutés dans le cadre du Centre National de Référence pour la maladie de Wilson coordonné par le Dr F. Woimant (Service de Neurologie, Hôpital Lariboisière, Paris). Chaque patient a été apparié en âge (+/- 4 ans) et en sexe à 2 sujets contrôles (N=34). Les moyennes d'âge des échantillons constitués sont relativement homogènes (Moyennes d'âge : patients ataxiques 35,7 ± 0,26 ; patients akinéto-rigides : 35,3 ± 1,7 ; patients dystoniques : 33,5 ± 1,7 ; sujets contrôles : 35,2 ± 3,1). La sévérité de la dysarthrie, de l'atteinte motrice et cognitive ainsi que les données d'IRM cérébrale ont été décrites pour chaque échantillon de patients (Pernon [9]). Ont été inclus les sujets dysarthriques présentant une forme neurologique de la maladie de Wilson, ayant pris connaissance de l'étude et ayant accepté d'y participer. Ont été exclus les patients en période d'exclusion d'une autre étude, dont la langue maternelle n'était pas le français et qui présentaient des perturbations majeures de l'intelligibilité et du caractère naturel de la parole.

2.2. Protocole et corpus

Dans un premier temps, les sujets ont répété, en condition isolée, une phrase composée principalement d'occlusives sourdes (« le coquin pépito papotait tout à coup ») durant 20 secondes, à 3 débits de parole différents (normal, voulu rapide, voulu lent). Les occlusives sourdes de cette phrase viennent faciliter, dans ce contexte de parole pathologique et de modulations du débit de parole, la détection des intervalles intervocaliques lors de l'alignement automatique, puis lors de la correction effectuée manuellement, et ce, grâce aux silences de cette catégorie de sons. Dans un second temps, les sujets ont réalisé 3 épreuves mettant en jeu différents processus attentionnels et exécutifs (tâches de barrage de cibles). Enfin, dans le cadre d'un paradigme de double tâche, la même phrase a été répétée à débit de parole normal durant 20 secondes, simultanément à la réalisation de chacune des épreuves attentionnelles. Il était précisé au sujet de ne jamais cesser de traiter les deux tâches à la fois, aucune priorité ne devant être donnée à l'une des deux tâches.

2.3. Analyse acoustique et mesures

Nos données acoustiques ont été segmentées sous Praat de manière semi-automatique au moyen du logiciel EasyAlign. Les données des groupes de sujets dysarthriques et des sujets contrôles ont fait l'objet de comparaison pour le nombre de syllabes produites par seconde, les durées moyennes des intervalles vocaliques et intervocaliques et leurs écarts type moyens, la durée moyenne des pauses interphrases et leur écart type moyen. Nous avons également comparé la production de la phrase à débit de parole normal en condition isolée et en double tâche pour mesurer l'effet de la double tâche. Les mesures d'écarts types moyens des durées ont permis de déterminer les capacités de maintien du débit de parole. Le débit de parole normal constituait ici une ligne de base pour les mesures d'accélération et de ralentissement, en le comparant aux débits de parole voulu rapide et voulu lent. Les pourcentages d'accélération et de ralentissement ont ainsi été également calculés à partir des ratios du nombre de syllabes produites par seconde, des durées moyennes des intervalles vocaliques et inter-vocaliques. Nous ne détaillerons ici que les données relevant du nombre de syllabes produites par seconde. Compte tenu du faible nombre de sujets au sein de chacun de nos échantillons, les données ont fait l'objet d'analyses statistiques au moyen de tests non paramétriques (Mann-Whitney, Kruskal-Wallis, Wilcoxon).

3. RÉSULTATS

Les débits de parole des patients s'avèrent plus lents que ceux des sujets contrôles et ce, de manière significative pour les débits de parole normal et voulu rapide ($p \leq 0,05$) (table 1).

Le groupe des sujets contrôles présente des capacités effectives d'accélération lors de la comparaison entre le débit de parole voulu rapide et le débit de parole normal (pourcentage d'accélération : 46,75 % ; $Z = -5,1$; $p < 0,0001$) et de ralentissement lors de la comparaison entre le débit de parole voulu lent et le débit de parole normal (pourcentage de ralentissement : 25,08 % ; $Z = -4,7$; $p \leq 0,027$). Parmi les groupes de patients, seuls les patients dystoniques possèdent des capacités d'accélération préservées (pourcentage d'accélération : 40,87 % ; $T = 36$; $p < 0,012$) (patients akinéto-rigides : pourcentage d'accélération : 15,95 % ; $T = 9$; $p = 0,144$) (patients ataxiques : pourcentage d'accélération : 36,62 % ; $T = 10$; $p = 0,068$). Les capacités de ralentissement sont altérées pour l'ensemble des groupes de patients (patients akinéto-rigides : pourcentage de ralentissement : - 10,19 % ; $T = 2$; $p = 0,273$) (patients ataxiques : - 10,93 % ; $T = 0$; $p = 0,063$) (patients dystoniques : pourcentage de ralentissement : - 17,41% ; $T = 3$; $p = 0,063$).

Globalement, les patients ataxiques présentent les débits de parole (normal, voulu rapide et voulu lent) les plus lents. Les patients dystoniques présentent les débits de parole les plus rapides. Les patients akinéto-rigides réalisent des débits de parole intermédiaires (table 1). Toutefois, aucune des analyses concernant les débits de parole normal, voulu rapide, voulu lent et les capacités d'accélération et de ralentissement voulus, ne

retrouve de différences significatives entre les trois groupes de patients (akinéto-rigides, ataxiques, dystoniques). Une tendance à la significativité des différences entre les trois groupes de patients est cependant notée pour le nombre de syllabes produites par seconde en débit de parole normal ($H = 5,41$; $p = 0,067$) et les capacités d'accélération voulue ($H = 1,66$; $p = 0,049$).

Les groupes de patients ne présentent pas en condition de double tâche les mêmes comportements de débit de parole. Les locuteurs akinéto-rigides et ataxiques ralentissent leur débit de parole en condition de double tâche, alors que les sujets contrôles et dystoniques l'accélèrent. Ces résultats n'atteignent pas la significativité chez les patients (table 2).

table 1 : Comparaison du nombre de syllabes produites par seconde pour chaque débit de parole, entre chaque groupe de patients et leurs groupes de sujets contrôles.

	débit de parole normal				**débit de parole voulu rapide**				**débit de parole voulu lent**			
	Moy.	E.T.	U	p	Moy.	E.T.	U	p	Moy.	E.T.	U	p
patients akinéto-rigides	3,51	0,76	**8,5**	**0,053**	4,03	0,81	**0**	**0,002**	3,22	1,06	18,5	0,426
sujets contrôles	5,01	0,75			6,89	0,73			3,56	0,63		
patients ataxiques	2,76	0,33	**0**	**0,007**	3,99	1,12	**0**	**0,007**	2,46	0,34	6	0,089
sujets contrôles	5,01	0,75			6,94	0,75			3,31	0,76		
patients dystoniques	3,94	1,27	**27**	**0,024**	5,52	1,71	**29**	**0,032**	3,29	1,52	46	0,270
sujets contrôles	4,82	0,72			6,76	0,87			3,57	0,93		

table 2 : Etude de l'effet de la double tâche (I : condition isolée ; DB : condition de double tâche) pour le nombre de syllabes produites par seconde lors de chaque épreuve attentionnelle et exécutive (épreuves 1, 2, 3), pour chaque groupe de sujets (AKIN : patients akinéto-rigides ; ATAX : patients ataxiques ; DYST : patients dystoniques ; CONTR : sujets contrôles).

épreuves réalisées concomitamment au débit de parole		épreuve attentionnelle et exécutive 1								épreuve attentionnelle et exécutive 2								épreuve attentionnelle et exécutive 3							
groupes de patients		AKIN		ATAX		DYST		CONTR		AKIN		ATAX		DYST		CONTR		AKIN		ATAX		DYST		CONTR	
conditions de production du débit de parole normal		I	DB	I	DB	I	DB	I	DB	I	DB	I	DB	I	DB	I	DB	I	DB	I	DB	I	DB	I	DB
nombre de syllabes par seconde	**Moy.**	3,51	3,3	2,76	2,65	3,94	4,29	4,75	5,2	3,51	3,24	2,76	2,45	3,94	4,22	4,75	5,18	3,51	3,21	2,76	2,65	3,94	4,44	4,75	5,27
	E.T.	0,74	0,71	0,33	0,96	1,27	1,41	0,75	0,68	0,74	0,85	0,33	0,94	1,27	1,71	0,75	0,75	0,74	1,22	0,33	0,96	1,27	1,41	0,75	0,76
	T ouz	0		4		28		**-4,72**		1		3		28		**-4,36**		3		4		27,5		**-4,8**	
	p	0,083		0,715		0,161		**<0,0001**		0,080		0,465		0,161		**<0,0001**		0,225		0,715		0,183		**<0,0001**	

4. DISCUSSION ET CONCLUSION

L'étude des capacités de modulation du débit de parole et de l'effet de la double tâche a permis de définir différents profils de patients.

Le débit de parole des patients est plus lent que celui des sujets contrôles et ce, de manière significative pour les débits de parole normal et rapide. Les patients ataxiques produisent le débit de parole le plus lent, les patients dystoniques le débit de parole le plus rapide. Ceci est concordant avec les caractéristiques perceptives de la dysarthrie wilsonienne décrite par Berry et al. [1], les troubles dysarthriques étant marqués par « un ralentissement du débit de parole », « des silences inappropriés », « un allongement des sons » et « un allongement des pauses ». Ces données vont également dans le sens du ralentissement objectivé sur le plan acoustique par Volkmann et al. [13]. Nous ne retrouvons pas la forme dysarthrique au « débit de parole rapide » retrouvée par Dordain et Chevrie-Muller [4]. Cette discordance pourrait s'expliquer par l'absence de prise en compte de la symptomatologie neurologique prédominante, la dysarthrie étant considérée comme une entité à part entière, sans répartition des patients en sous-groupes. D'autre part, les patients sont comparés les uns aux autres, et non pas à des sujets contrôles. En effet, si nous considérons uniquement nos groupes de patients, nous pouvons noter que le débit de parole du groupe des patients

dystoniques est effectivement plus rapide que celui des patients akinéto-rigides et des ataxiques.
Les capacités d'accélération voulue du débit de parole sont préservées pour les patients dystoniques et altérées pour les patients akinéto-rigides et ataxiques. Les capacités de ralentissement sont perturbées pour l'ensemble des patients. Ces données rejoignent celles de la littérature. Dordain et Chevrie-Muller |4] relevaient que si le débit de parole normal était ralenti, les patients ne pouvaient l'accélérer. Hefter et al. [7] retrouvaient une réduction du débit de parole maximal.
En double tâche, les contrôles et les sujets dystoniques accélèrent significativement leur débit de parole ; les patients akinéto-rigides et ataxiques le ralentissent de manière non significative. L'accélération des sujets contrôles et dystoniques pourrait être liée à un effet d'attracteur de la vitesse adoptée lors du traitement des épreuves attentionnelles, en rapport avec un modèle d'interférence rythmique (Ebersbach et al. [5]) ou articulatoire (Yardley et al. [15]). Le ralentissement observé chez les patients ataxiques et akinéto-rigides, sujets ayant par ailleurs montré des capacités d'accélération perturbées en condition isolée, pourrait s'expliquer par le phénomène de « capacity sharing » (moindre performance pour 2 tâches simultanées) (Pashler [8]).

Le débit de parole, son contrôle et l'influence des épreuves de double tâche nous permet ainsi de distinguer différents profils de dysarthries chez les patients wilsoniens. La prédominance de la dysarthrie est en relation avec la symptomatologie neurologique principale, ce qui n'avait pas été étudié jusqu'à présent. Il serait intéressant de réitérer ces analyses sur des échantillons de plus grande taille afin de confirmer la tendance à la significativité de certains résultats et de préciser cet outil diagnostic et thérapeutique. En effet, l'intégration des aspects : modulation du débit de parole et condition de double tâche pourrait être alors envisagée dans le cadre de l'évaluation clinique des dysarthrie et des stratégies rééducatives orthophoniques.

BIBLIOGRAPHIE

[1] W.R. Berry, F.L. Darley, A.E. Aronson and N.P. Goldstein. Dysarthria in Wilson's Disease. *Journal of Speech and Hearing Research*, 17 : 169-83, 1974a.

[2] G.J. Brewer. (2005) Behavioral Abnormalities in Wilson's Disease. *Advances in Neurology*, 96 : 262-74, 2005.

[3] A. Delval, P. Krystkowiak, M. Delliaux, K. Dujardin, J.-L. Blatt, A. Destée, P. Derambure and L. Defebvre. Role of attentional resources on gait performance in Huntington's disease. *Movement disorders* : 23 : 684-89, 2008.

[4] M. Dordain and C. Chevrie-Muller. Voice and Speech in Wilson's Disease. *Folia Phoniatrica*, 29 : 217-232, 1977.

[5] G. Ebersbach, MR. Dimitrijevic and W. Poewe. Influence of concurrent tasks on gait : a dual-task approach. *Perceptual and Motor Skills*, 81 : 107-13, 1995.

[6] J. Ghika, F. Vingerhoets, P. Maeder, F.-X. Borruat and J. Bogousslavsky. Maladie de Wilson. *EMC-Neurologie* 1: 481-511, 2004.

[7] H. Hefter, G. Arendt, W. Stremmel and H.-J. Freund. Motor impairment in Wilson's disease, II : slowness of speech». *Acta Neurologica Scandinavica*, 87 : 148-60, 1993b.

[8] H. Pashler. Dual-task interference in simple tasks : data and theory. *Psychological Bulletin*, 116 : 220-44, 1994.

[9] M. Pernon. *Débit de parole dans les dysarthries de la maladie de Wilson, Etude de l'influence des troubles attentionnels et dysexécutifs en condition de double tâche.* Mémoire de Master 2 de Phonétique, LPP, Université de Paris 3, 2009a.

[10] M. Pernon, JM. Trocello, C. Cousin, T. Peron-Magnan, A. de Tassigny, G. Chevaillier, P. Rémy, F. Woimant. Spécificité de la pratique orthophonique auprès de patients atteints de la maladie de Wilson : expérience du Centre National de Référence pour la maladie de Wilson. *L'Orthophoniste*, 292 : 19-26, 2009b

[11] L. Rochester, V. Hetherington, D. Jones, A. Nieuwboer, A.-M. Willems, G. Kwakkel and E. Van Wegen. Attending to the Task : Interference Effects of Functional Tasks on Walking in Parkinson's Disease and the Roles of Cognition, Depression, Fatigue, and Balance. *Arch. of Physical Med. and Rehab.*, 85 : 1578–85, 2004.

[12] JM. Trocello, P. Chappuis, P. Chaine, P. Rémy, D. Debray, JC. Duclos-Vallée and F. Woimant. Maladie de Wilson. *La Presse Médicale*, 38 : 1089-98, 2009.

[13] J. Volkmann, H. Hefter, H.W. Lange and H.J. Freund. Impairment of Temporal Organization of Speech in Basal Ganglia Diseases. *Brain and Language* : 43 : 386-99, 1992.

[14] F. Woimant, P. Chaine, P. Favrole, J. Mikol and P. Chappuis. Mise au point : La maladie de Wilson. *Revue Neurologique* : 162 : 773-8, 2006.

[15] L. Yardley, M. Gardner, A. Leadbetter and N. Lavie. Effect of articulatory and mental tasks on postural control. *Neuroreport*, 10 : 215-19, 1999.

[16] G. Yogev, N. Giladi, C. Peretz, S. Springer, E.S. Simon and J.M. Hausdorff. Dual tasking, gait rythmicity, and Parkinson's disease. *European Journal of Neuroscience*, 22 : 1248-56, 2005.

Approche acoustico-statistique de la fluence chez les PQB

Audrey Leclercq, Myriam Piccaluga, Kathy Huet, Bernard Harmegnies

Laboratoire des sciences de la parole de l'AUWB
UMONS, 18 place du parc, B-7000 Belgique
Audrey.leclercq@umons.ac.be

http://w3.umh.ac.be/~compa/

ABSTRACT

A non-stuttering person as well as two persons who stutter with contrasted stuttering profiles have undergone 4 conditions of Delayed Auditory Feedback (NAF, DAF 80, DAF 120 and DAF 160) while performing a map task. The expected behavioural variations have been confirmed by means of objective as well as subjective processes. The speech productions are analyzed by means of the Inter-Syllabic Interval (ISI), which exhibits good performance in discriminating subjects and conditions.

Keywords: stuttering, ISI, fluency, prosody, syllable.

1. INTRODUCTION

Les difficultés variées que rencontrent les Personnes Qui Bégaient (PQB) dans l'actualisation phonique des messages qu'elles ont l'intention d'émettre ont pour conséquence des perturbations très diversifiées du flux de parole (répétitions non souhaitées d'événements phoniques, blocages soudains, tics verbaux, perturbations du rythme, etc.). Malgré cette objective diversité phénoménologique, cette parole est souvent, pour l'auditeur, à l'origine d'une sensation subjective unitaire de disfluence. L'étiologie précise et la nature des processus impliqués dans le bégaiement demeurent aujourd'hui encore largement débattues. Pourtant, le phénomène a été très largement étudié, mais la plupart du temps par des personnes qu'animaient un -légitime- souci d'aide directe aux PQB ou la nécessité de faire face à des exigences en termes d'évaluation ou de diagnostic. Souvent, les descriptions proposées demeurent très holistiques, basées sur des variables macroscopiques (par exemple, décompte du nombre de mots ou de syllabes affectés par le bégaiement [1]) et leur objectif principal se limite, dans de nombreux cas, à l'appréciation du degré de sévérité du bégaiement [2]. Ces études sont fréquemment basées sur des analyses subjectives, parfois appuyées sur des modèles d'évaluation destinés à l'objectivation du jugement expert [3], voire sur l'auto-évaluation des PQB [4]. En marge de cette littérature à caractère orthophonique, médical, voire psychologique, très peu de travaux visant à finement appréhender les caractéristiques phoniques de la parole bégayée ont vu le jour. Ceux-ci, hormis de rares exceptions [5,6,7], ont privilégié la parole des PQB anglophones. Or, vu les différences de structures notamment prosodiques, il n'est pas raisonnable de penser que ces résultats soient, sans autre forme de procès, directement transférables au français. La disponibilité de processus raffinés pour l'étude de la parole des PQB francophones apparaît donc très souhaitable, tant en vue de la mise en œuvre de processus objectifs d'évaluation, que pour une meilleure compréhension des mécanismes à l'origine du bégaiement. Cette communication se présente comme une contribution exploratoire en ce sens, visant à investiguer le potentiel d'un indice statistique à base acoustique (l'Ecart Inter Syllabique (EIS), initialement destiné à d'autres fins) [8,9] pour la mesure des variations de la fluence du discours émis par les PQB.

2. DISPOSITIF EXPERIMENTAL

2.1. Méthodologie

Nous cherchons à questionner le potentiel informatif d'un indice (l'EIS). Il est donc nécessaire de déterminer s'il varie sous l'effet des caractéristiques du phénomène qui nous intéresse : le(s) bégaiement(s). Il s'impose donc de mettre en place un dispositif susceptible de maximiser la variabilité des aspects du phénomène. A cet effet, d'une part, nous agirons sur la variabilité inter sujets (à côté de 2 PQB à profils de bégaiement contrastés, nous étudierons une Personne Non Bègue, PNB) et d'autre part, nous diversifierons les conditions intra-sujet par le recours à des types de feedback auditif dont la littérature a montré le potentiel d'influence sur la parole des PQB [1,3,4,5]. Avant de procéder à l'analyse de la réactivité de l'EIS, nous tenterons de vérifier si le dispositif expérimental a effectivement produit des modifications du comportement des locuteurs, ceci tant via une analyse objective des phénomènes phoniques observés que par le recours à l'évaluation en aveugle d'experts en orthophonie.

2.2. Sujets

Nos sujets sont francophones monolingues, de sexe masculin, de nationalité belge et ont le français pour langue maternelle. Mis à part PNB, ils ont fait l'objet d'un diagnostic de bégaiement en bas âge. Du point de vue clinique, PQB1 présente un profil de type plutôt « clonique » (haute fréquence de répétitions d'éléments divers de la chaîne parlée), alors que PQB2 se caractérise plutôt par un profil de type « tonique », les blocages soudains étant dominants dans ses productions.

2.3. Tâche

Nous avons recouru à la *tâche de la carte* [5] ; sur base d'une carte routière, le sujet doit expliquer à un

interlocuteur (dans ce cas, francophone, lointain et silencieux) comment se rendre d'une ville à une autre. Les sujets ont la possibilité, avant de l'expliquer, de préparer le trajet, qui doit tenir compte d'un certain nombre de contraintes (par exemple, évitement de grandes villes, préférences pour certaines zones touristiques, etc). Le but ainsi poursuivi est de les amener à produire une quantité de matériau phonique importante dans le cadre d'une situation réellement communicative, sans intervention de processus de lecture et minimisant tant l'effet de la dynamique interactionnelle avec un interlocuteur que le risque de contamination de la situation par l'émergence de thématiques à haut degré d'implication affective . Chaque locuteur a effectué 4 tâches de ce type, toutes similaires mais basées sur des parcours différents. Trois des quatre tâches se caractérisaient par l'application d'un dispositif de feedback à retard (DAF), le délai étant respectivement fixé à 80 ms, 120 ms et 160 ms (comme en [5]). Une tâche a été effectuée sans retard du feedback (NAF).

2.4. Traitement acoustique

Le retard de feedback a été obtenu par le recours à la routine DAF de la station Kay SCL 4300. Tous les enregistrements ont été effectués dans la chambre anéchoïque du Laboratoire de Phonétique de l'Université de Mons, au moyen d'un microphone Neumann U87P48, connecté à un codeur PCM 501ES Sony, dont les signaux de sortie étaient stockés sur un magnétoscope VHS Panasonic. Les analyses acoustiques ont été effectuées au moyen du logiciel Praat.

2.5. Evénements phoniques

L'événement élémentaire sur lequel nous nous centrons est le résultat acoustique (segment phonique ou groupe de segments phoniques) de la réalisation d'une *syllabe* dans le sens classique que la phonologie a donné à ce terme (chacun pouvant être associé à tout ou partie d'une unité lexicale du français). Néanmoins, en parole spontanée (et spécialement dans la parole spontanée des PQB), des événements phoniques non susceptibles d'être couverts par cette définition apparaissent fréquemment. Ces « pronunciation units » [11, p 338], correspondent au concept assez largement accepté de « syllabe », cette fois au sens phonétique du terme ; nous les prendrons également en considération. L'ensemble des événements que nous étudions peut ainsi être divisé en deux sous-groupes : d'une part, les événements que nous étiquetterons **α** (sons ou groupes de sons correspondant à la prononciation normative d'une syllabe d'un mot appartenant à l'inventaire lexical du français), d'autre part, les événements **β** (tout autre son isolé ou groupe de sons constituant une unité de prononciation et ne contribuant pas à la réalisation normative d'un mot appartenant à l'inventaire lexical du français : voyelle isolée, syllabe isolée, onset syllabique, pause pleine, etc). Prenons par exemple un extrait d'une production de PQB1 sous NAF. Le message est *ouais puis, là, je vais continuer*, mais la réalisation est [p p p p p p p p p p p p p p p wɛ p p pɥi la ɔː ʒə ve k k kɔ̃ ti nɥe]. Tous les éléments de la salve des 15 [p] initiaux sont étiquetés β, de même que, les 2 [p] qui suivent [wɛ] (*ouais*), la pause pleine et les 2 [k] précédant [kɔ̃]. Par contre, [wɛ], [pɥi], [la], [ʒə], [ve], [kɔ̃], [ti] et [nɥe], sont étiquetés α.

3. VALIDATION DU DISPOSITIF

3.1. Distribution des événements phoniques

La figure 1 présente l'évolution de la contribution des événements β au nombre total d'événements par minute (quotient du débit des événements β au débit global des événements) pour les 2 PQB (PNB n'ayant produit que 0,2% d'événements dans la seule condition DAF 160 n'est pas ici pris en considération).

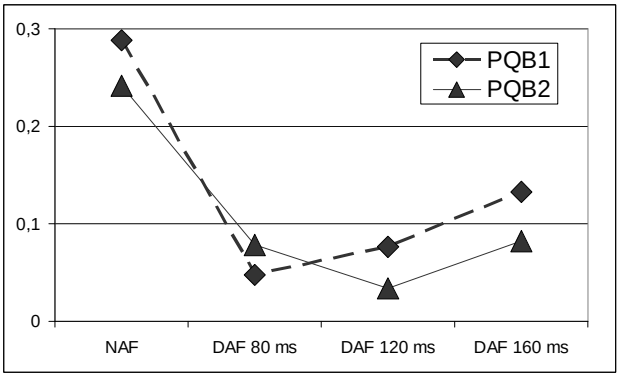


Fig. 1 : Débit relatif des événements β (ordonnée), par sujet et condition

Il apparaît que les deux PQB émettent, lorsqu'ils s'expriment en condition normale, d'importantes quantités de matériau phonétique non pertinent d'un point de vue linguistique (entre 25 et 30% des événements phoniques); sous condition DAF, cependant, une diminution spectaculaire du rapport se fait jour, les valeurs pouvant chuter jusque .05 (PQB1) ou .03 (PQB2).

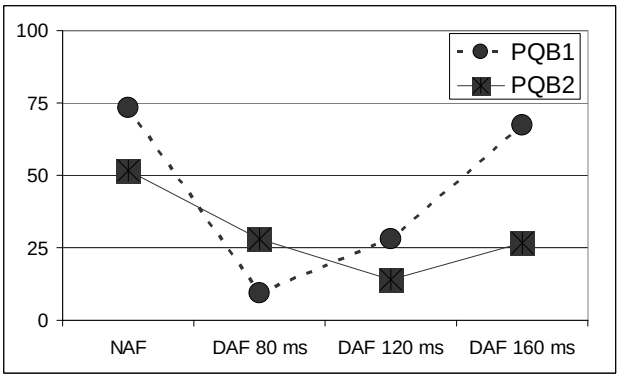


Fig. 2 : Taux de sévérité estimé (ordonnée), par sujet et condition

On note, par ailleurs, que les effets maximaux sont observés pour les valeurs de DAF les plus basses (80 ms pour PQB1 et 120 ms PQB2) ; sous la condition DAF 160, les valeurs du rapport remontent chez les deux

locuteurs, tout en demeurant cependant nettement inférieures à celles observées sous condition NAF.

3.2. Contrôle par experts

Deux orthophonistes chevronnées ont été exposées aux huit productions recueillies (2 sujets * 4 conditions). Elles ont été priées d'émettre deux évaluations indépendantes (l'une en ignorant l'origine des productions, l'autre en étant informées du fait qu'elles émanaient de sujets soumis à des conditions de production différentes). Leurs évaluations se sont appuyées sur une adaptation du Test de sévérité du bégaiement de Boey (sélection des items auditifs) [10]. Comme le montre la figure 2, qui moyenne les deux évaluations (exprimées ici sur une échelle de gravité allant de 0 à 100), les tendances générales révélées par dénombrement objectif des types d'événements (fig. 1) sont particulièrement similaires à ces évaluations subjectives puisque sont conservées les relations d'ordre entre d'une part les sujets et d'autre part les conditions, même si PQB1 se voit attribuer, sous DAF 160, un taux de gravité pratiquement similaire à celui observé en NAF, alors que le nombre d'événements β est demeuré bas.

4. ANALYSE ACOUSTIQUE

Nos constatations précédentes confirment que les variables indépendantes contrôlées (sujet et condition) ont bien un effet sur le comportement locutoire de nos sujets. Nous sommes donc fondés à interroger la sensibilité des paramètres acoustiques aux variations comportementales attestées.

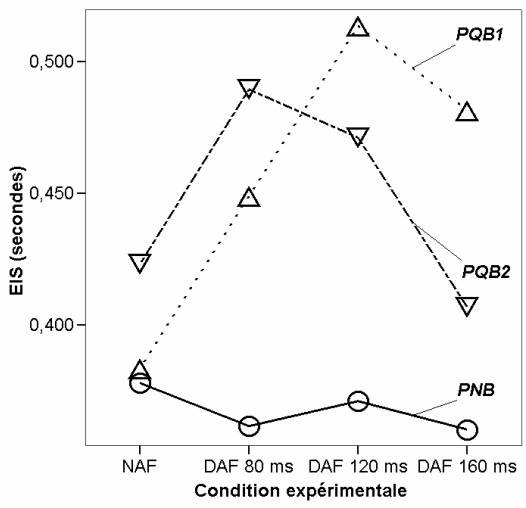


Fig. 3: EIS moyen par sujet et condition (événements α et β)

Comme indiqué plus haut, nous nous centrons ici sur l'EIS, un indice basé sur la différenciation temporelle des pics d'intensité. Pour chaque pic, l'intervalle temporel le séparant du pic précédent est calculé. Pour un corpus de n événements, une liste chronologiquement ordonnée de n-1 délais entre événements peut ainsi être obtenue.

Une analyse descriptive d'ensemble des valeurs de l'EIS suggère sa sensibilité aux effets de la variation des conditions expérimentales (fig. 3). Dans la condition NAF, les trois sujets ont des moyennes d'EIS assez proches les unes des autres (entre 378 ms et 423 ms). Sous l'effet du DAF, PNB se caractérise par des valeurs très stables (de 360 ms à 371 ms) ; à l'opposé, les 2 PQB montrent des tendances nettes à l'accroissement de ces valeurs. Les EIS les plus importants atteignent ainsi 514 ms chez PQB1 sous DAF 120 et 489 ms chez PQB2 sous DAF 80. Pour les PQB, une diminution sensible est observée sous la condition DAF 160, par rapport à DAF 120. L'analyse de variance confirme ces observations globales, révélant un effet significatif du sujet ($F_{2,3354}$ = 14.788, p < 0,01), un effet significatif de l'interaction sujet * condition ($F_{6,3354}$ = 2,300, p = .032) et un effet de la condition faiblement significatif ($F_{(3,3354)}$ = 2.470, p = .060). En bref, il apparaît que les sujets ne réagissent pas à l'action du DAF de la même manière, que les PQB sont fortement affectés et que l'effet, chez eux, consiste en un accroissement de l'EIS sous l'effet du DAF 80 et du DAF 120. Ce traitement inférentiel devrait néanmoins être considéré avec circonspection, dans la mesure où les distributions d'EIS tendent à être assez dissymétriques (avec une concentration dans la portion de l'axe correspondant aux plus petites valeurs et aussi bon nombre de valeurs extrêmes dans la partie supérieure de l'axe). Si, au lieu des moyennes, on prend en considération les médianes, l'observation d'un accroissement graduel de l'EIS de la condition NAF à la condition DAF 120 demeure vraie pour PQB1 (test de la médiane : χ^2= 35.347, DL= 2, p < .001) mais pas pour PQB2 (χ^2= 1.699, DL= 2, p = .428). De plus, un effet significatif est observé chez PNB (χ^2= 20.338, DL= 2, p < .001). Néanmoins, l'amplitude de l'augmentation de l'EIS est beaucoup plus importante chez PQB1 (médiane de 241 ms jusqu'à 372 ms) que chez PNB (médiane de 222 ms jusqu'à 294 ms).

Par ailleurs, un examen minutieux des distributions (non illustré ici) suggère que le DAF agit de manière plus sensible sur les événements β, pour lesquels l'accroissement de l'EIS est spectaculairement plus grand que pour les événements α. Il est dès lors important d'étudier la distribution des EIS pour les événements β. La figure 4 présente ces distributions pour les seuls PQB. Il apparaît ici que, sous condition NAF, les PQB se caractérisent par une importante concentration dans la zone correspondant aux très basses valeurs d'EIS. Celles-ci correspondent à des phénomènes brefs tels que des tics verbaux et des répétitions cloniques de segments ; on notera d'ailleurs que, pour PQB1 (au profil à dominance clonique), le mode local, approximativement situé à 200 ms, est inférieur à celui qui caractérise PQB2 (au profil à dominance tonique), localisé environ à 300 ms, ce qui suggère la présence d'une quantité plus importante de répétitions cloniques. La figure montre que lorsque les sujets sont exposés aux deux premiers délais de DAF, une sensible raréfaction des faibles valeurs d'EIS se fait jour. Le phénomène est particulièrement important pour PQB1 dans la condition DAF 80 et pour PQB2 dans la condition DAF 120.

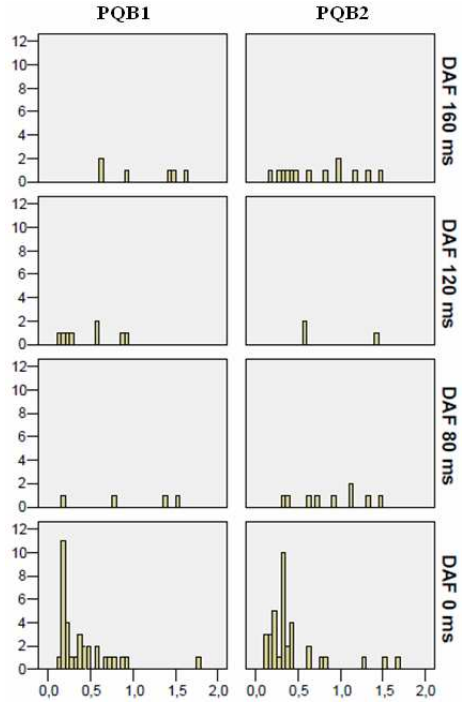


Fig. 4 : Nombre d'EIS (ordonnée), par durée en s (abscisse) des événements β selon le sujet et la condition

Un important facteur de l'accroissement des moyennes d'EIS chez les PQB sous les conditions de feedback retardé apparaît donc être la disparition des événements β de durée brève, ceux-ci étant fréquents chez les deux PQB en condition normale et absents des productions de PNB.

5. CONCLUSION

Procédant par contrôle tant des conditions de rétroaction auditive (NAF et 3 délais de DAF) que du profil des sujets (PNB et PQBs à profils contrastés : tonique *vs.* clonique), la présente étude a produit des manifestations comportementales du bégaiement dont la diversité a été attestée par la convergence de mesures objectives et d'évaluations d'experts. L'écart inter-syllabique (EIS) s'y est révélé sensible. Notre étude nous a amenés à souligner le caractère non gaussien des distributions observées. Ceci engage certes à la circonspection dans l'utilisation de modèles de traitement statistique. Mais plus profondément, cette observation suggère l'existence d'une structure sous-jacente de causalités complexes, certains groupes des valeurs d'EIS pouvant être liés à certains types de phénomènes (par exemple les très basses valeurs correspondant au comportement clonique). Probablement différents niveaux de traitement et/ou processus de production du langage sont-ils associables à l'une ou l'autre manifestation acoustique. Les expériences à conduire dans le futur pourraient ainsi se centrer sur deux objectifs : sans doute celui d'une meilleure description et d'une meilleure évaluation des performances des PQB (le but initial de notre recherche), mais également une meilleure compréhension des processus mis en œuvre par les locuteurs pour faire face aux conséquences des altérations du retour auditif, procédures qui ne sont pas nécessairement les mêmes chez les PNB et les PQB et, de surcroît, ne sont probablement pas identiques chez les PQB de divers types. Ces efforts devraient permettre d'œuvrer à l'établissement d'une sémiologie croisée des évaluations d'experts et des analyses à base acoustique.

6. REFERENCES

[1] J. Armson, & A. Stuart. Effect of extended exposure to frequency altered feedback on stuttering during reading and monologue. *JSLHR*, 41, 479-490, 1988.

[2] G. Sparks, D. Grant, K. Millay, D. Walker-Baston & L. Hynan. The effect of fast speech rate on stuttering frequency during delayed auditory feedback, *Journal of Fluency Disorders*, 27, 187-201, 2002.

[3] A. Stuart & J. Kalinowski. The perception of speech naturalness of post-therapeutic and altered auditory feedback speech of adults with mild and severe stuttering. *Folia Phoniatrica et Logopaedica*, 56, 347-357, 2004.

[4] J. Kalinowski, V. K. Guntupalli, A. Stuart & T. Saltuklaroglu. Self-reported efficacy of an ear-level prosthetic device that delivers altered auditory feedback for the management of stuttering. *Int. J. of Rehabilitation Research*, 27, 167-170, 2004.

[5] B. Harmegnies, M. Bruyninckx & D. Poch. Delayed Auditory Feedback effects on stutterers' voice qualities and vowels systems. *Proc. of the 13th ICPhS*, 1, 400-403, 1995.

[6] F. Hirsch. *Le bégaiement. Perturbation de l'organisation temporelle de la parole et conséquences spectrales.* Thèse doctorale, Université Marc Bloch, Strasbourg 2, 2007.

[7] F. Hirsch, M. C. Monfrais-Pfauwadel, R. Sock, & B. Vaxelaire. Etude de la structure formantique des voyelles dans la parole bègue en vitesses d'élocution normale et rapide. *Revue de laryngologie, d'otologie et de rhinologie*, vol. 130, 17-22, 2009.

[8] M. Piccaluga. *Approches psycholinguistiques de l'interprétation*. Thèse doctorale, Université de Mons-Hainaut, 2004.

[9] M. Piccaluga, J.-L. Nespoulous & B. Harmegnies. Disfluency surface markers and cognitive processing ; the case of simultaneous interpreting. *Proc. of the 16th ICPhS*, 1317-1320, 2007.

[10] R. Boey. *Test voor Stotterernst-Lezers (TvS-L). Test voor Stotterernst Niet-Lezers (TvS-NL).* Garant, Leuven, Belgique, 2001.

[11] Crystal, D., *A dictionary of Linguistics and Phonetics.* Basil Blackwell,Cambridge Ma, 1992

Corrélats acoustico-perceptifs des consonnes non relâchées du vietnamien

Thi-Thuy-Hien Tran, Nathalie Vallée

Département Parole et Cognition, GIPSA-lab
1180, avenue Centrale, BP 25, 38040 Grenoble Cedex 9, France
thi-thuy-hien.tran@gipsa-lab.grenoble-inp.fr, nathalie.vallee@gipsa-lab.grenoble-inp.fr

ABSTRACT

One of the difficulties faced by Vietnamese subjects upon learning French is the pronunciation of consonant clusters. Those clusters are often mispronounced, even if specific consonant combinations are found in both languages. In Vietnamese, consonant sequences are found only at syllable boundaries and final stops are often unreleased which is not the case in French. Aiming to discover the responsible factors which prevent Vietnamese subjects from correctly pronouncing French clusters, the main objective of our study was to investigate the perception of Vietnamese syllable-final stops in order to examine if acoustic cues in stop-consonant could affect the perception of French consonant clusters. More specifically, this paper deals with acoustic cues which allow the identification of unreleased final stops. Data from an identification task provide evidence that specific acoustic characteristics, and also the lexical frequency of final consonants, have led the Vietnamese speakers in their responses.

Keywords : Vietnamese, unreleased stops, consonant cluster, acoustic and perceptive cues

1. Introduction

La relation entre perception et production des sons des langues secondes a été largement étudiée (par ex. Best [1]) avec un objectif général, celui de comprendre comment la perception des non-natifs influence l'apprentissage d'une langue seconde et notamment ses aspects phonétiques et phonologiques. En début d'apprentissage, un apprenant confronté au système phonético-phonologique d'une autre langue éprouve régulièrement des difficultés avec les unités sonores de cette langue qui n'existent pas dans sa langue maternelle. Ces unités manquantes sont plus ou moins bien perçues suivant les cas et plus ou moins bien reproduites. À ce sujet, il a souvent été avancé que pour qu'un son soit correctement (re)produit, il faut d'abord qu'il soit bien perçu, c'est-à-dire reconnu et identifié par rapport aux autres [2].

Les apprenants vietnamiens éprouvent des difficultés récurrentes à réaliser certains sons du français, parmi lesquels figurent les groupes de consonnes (clusters). Ceux-ci n'existent pas en vietnamien, langue tonale et isolante où les mots sont invariables et indécomposables. Par conséquent, les clusters du français sont souvent réalisés déformés par rapport à la cible (ex. « psychique » prononcé [siʃik], « insecte » réalisé [ɛ̃sɛkˀtə]). Cette difficulté perdure même après plusieurs années de pratique, même chez des étudiants de niveau déjà confirmé [6]. En quoi consistent exactement les erreurs de réalisation et quelles sont leurs implications dans l'acquisition des percepts phonétiques ? Quelles caractéristiques de la langue maternelle gênent la réalisation des clusters consonantiques du français ? L'objectif général de ce travail est de répertorier et de comprendre les facteurs responsables de cette difficulté rencontrée par les apprenants vietnamiens, l'hypothèse étant que l'étude de la perception de la parole par des non-natifs peut rendre compte des difficultés que les apprenants tardifs rencontrent avec certains contrastes et segments de la L2.

Vietnamien et français sont deux langues pour lesquelles de nombreuses dissemblances peuvent être relevées à propos des gabarits lexicaux et patrons syllabiques. Le vietnamien est une langue monosyllabique sur le plan phonologique mais polysyllabique en partie sur le plan lexical [5]. La différence entre mot simple et mot composé n'existe que par le nombre de syllabes. Le mot simple est monosyllabique. Les composés à deux syllabes sont les plus nombreux et il n'existe pas de composés de plus de quatre syllabes [10]. Doan [4] représente les patrons syllabiques du vietnamien par $C_1(w)V(C_2)$, les parenthèses indiquant les constituants optionnels. La diversité des patrons syllabiques est beaucoup plus présente en français, l'attaque et la coda pouvant être occupées par des clusters $(C_1)(C_2)(C_3)V(C_4)(C_5)(C_6)(C_7)$ [7]. Ce cas n'est jamais rencontré en vietnamien où les séquences de consonnes ne se rencontrent qu'à la frontière de mot ou à la frontière syllabique à l'intérieur d'un mot composé. De fait, les séquences de consonnes appartiennent à deux syllabes différentes. Cependant, 96% des structures syllabiques du français entrent dans le patron $(C_1)(C_2)V(C_3)$ qui se rapproche de la structure syllabique du vietnamien $C_1(w)V(C_2)$.

Une autre particularité de la structure syllabique du vietnamien par rapport à celle du français concerne la réalisation des consonnes finales. En vietnamien, dans la plupart des cas, la tenue de l'occlusion des plosives /p t k/ en coda, quelle que soit leur condition de réalisation, n'est pas suivie d'un bruit caractéristique d'explosion rapide et audible [4]. Au contraire, en français, les plosives finales, accompagnées ou non d'un voisement, sont généralement suivies d'un relâchement audible. Tran [9] dans une étude récente de données acoustiques montre, en fait, que les consonnes finales du vietnamien comportent des caractéristiques différentes en fonction du type de frontière syllabique qu'elles précèdent : d'une part, les plosives et nasales sont significativement plus longues devant frontière de mot que devant frontière de syllabe à l'intérieur d'un mot composé lexical ; d'autre part, les plosives sont plus souvent non relâchées, sinon avec un burst significativement plus bref et de plus faible amplitude, en finale de

syllabe à l'intérieur de mot.

Dans le prolongement de ce travail, l'étude que nous présentons ici a pour objectif d'identifier les corrélats acoustico-perceptifs des consonnes finales du vietnamien en lien avec un éventuel effet de l'influence du type de frontière syllabique sur la perception des consonnes. En d'autres termes, il s'agit de répondre aux questions suivantes : quels indices permettent d'identifier le lieu d'articulation d'une consonne du vietnamien lorsque celle-ci est non relâchée ? Existe-il des différences dans la perception des consonnes en fonction de leur position dans la syllabe ? Dans le mot ?

2. Méthodologie

2.1. Stimuli

Un ensemble de 55 items sonores naturels de type CVC correspondant soit à un mot monosyllabique, soit à la première syllabe d'un mot composé ont été sélectionnés pour l'étude. Ils comportent l'une des 6 occlusives / p t k m n ŋ / en contexte de la voyelle la plus ouverte /a/ et de même contexte tonal (ton montant D1). Plus concrètement, dans cet ensemble d'items lexicaux, ces six consonnes se trouvent :

- Soit en position finale C_2 des mots monosyllabiques [C_1a**C**$_2$] (exemples : [tap] « táp » *happer*, [kak] « các » *marque pluriel*, [maŋ] « máng » *mangeoire*) ;
- Soit en position finale C_2 de première syllabe [C_1a**C**$_2$] de mots composés lexicaux dissyllabiques [C_1a**C**$_2$.C_3VC_4] (exemples : [mat] dans [mat.za] « mát dạ » *content*, [fap] dans [fap.leŋ] « pháp lệnh » *ordonnance*, [saŋ] dans [saŋ.taw] « sáng tạo » *créer*)

Ces stimuli sont extraits des phrases porteuses d'un corpus lu par un sujet natif masculin, âgé de 27 ans, originaire du Nord du Vietnam. Les stimuli ont été choisis de manière à ce que leur entourage dans la phrase porteuse soit le plus neutre possible du point de vue articulatoire pour faciliter l'extraction de la syllabe : toute consonne sourde en coda est suivie d'une consonne sonore, et inversement. Au total, 29 items auditifs CVC terminés par une plosive et 26 items terminés par une nasale ont été retenus pour le test.

Vingt sujets (10 hommes et 10 femmes) originaires du Nord du Vietnam, âgés de 24 à 29 ans, ont participé au test. Aucun d'entre eux ne possédait de formation universitaire en phonétique-phonologie. La plupart avaient eu connaissance de notions non approfondies de phonétique lors de leur apprentissage en français langue étrangère au Vietnam.

2.2. Déroulement

Nous avons choisi de tester les plosives à part des nasales. Le programme utilisé pour le test est PerceptA de R. Carré. Les sujets avaient pour consigne d'écouter, puis identifier la consonne finale du stimulus présenté parmi 3 solutions proposées (choix forcé) correspondant soit aux 3 plosives / p t k /, soit aux 3 nasales / m n ŋ / en cliquant la réponse à l'écran avec l'aide de la souris.

Les deux tests étaient constitués de 3 répétitions des items présentés dans un ordre aléatoire pour chaque sujet. Le test s'est déroulé dans une salle calme du laboratoire en utilisant un casque de haute qualité binaural Sennheiser HD-25-13.

3. Résultats

Les données ont été ensuite analysées en utilisant le logiciel SPSS (*Statistical Package for the Social Sciences*). Des tests d'analyses de variance (ANOVA) ont été effectués.

Les analyses montrent que pour les plosives en coda, le meilleur score d'identification est obtenu par /t/ et /k/ (cf. Tab. 1). La différence de score de bonnes réponses est significative entre /p/ et /t/ (p = 0) ; entre /p/ et /k/ (p = 0), mais pas entre /t/ et /k/ (p = 0,79).

Table 1: Score d'identification des plosives.

Consonnes	Choix p	Choix t	Choix k
p	**74**%	19%	7%
t	8%	**84**%	9%
k	11%	6%	**83**%

Pour les nasales, la vélaire /ŋ/ obtient le meilleur score d'identification par rapport à /m/ et /n/ (cf. Tab. 2). La différence de score de bonnes réponses est significative entre /m/ et /ŋ/ (p = 0) ainsi que entre /n/ et /ŋ/ (p = 0), mais pas entre /m/ et /n/ (p = 0,34).

Table 2: Score d'identification des nasales.

Consonnes	Choix m	Choix n	Choix ŋ
m	**85**%	10%	6%
n	9%	**86**%	4%
ŋ	1%	5%	**95**%

Regroupés par lieu d'articulation, les vélaires obtiennent un meilleur score par rapport aux labiales et coronales (cf. Fig. 1). Par mode d'articulation, les nasales engendrent moins de confusion de la part des sujets par rapport aux plosives (89% contre 81%).

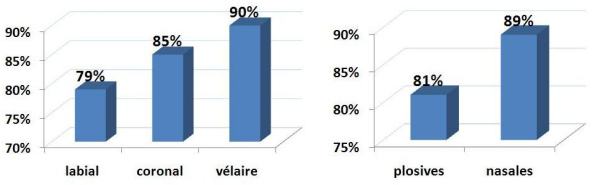


Figure 1: Score d'identification des consonnes finales par lieu et mode d'articulation.

Concernant l'influence du type de frontière syllabique sur le score de bonnes réponses, la figure 2 montre davantage de confusion pour les plosives et les nasales lorsque la coda C_2 précède une frontière intra-mots (devant C_3 de mots composés), et cette différence est significative dans les deux cas (respectivement p = 0,011 ; p = 0,008).

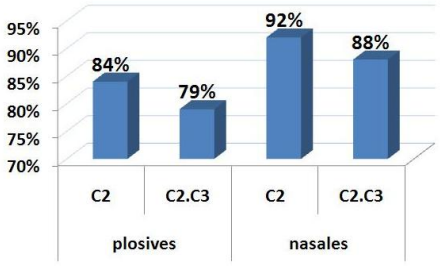


Figure 2: Score d'identification des consonnes finales selon leur position dans la syllabe.

4. Discussion

Quels sont les indices qui permettent d'identifier le lieu d'articulation d'une consonne lorsque celle-ci n'est pas relâchée ? Pourquoi les nasales ont été mieux perçues en coda par rapport aux plosives ? Est-ce que ces résultats sont dus à l'effet de la fréquence de ces consonnes dans le lexique ? Des indices acoustiques du corpus d'où ont été extraits les stimuli jouent-ils un rôle d'indicateur ?

4.1. Fréquence lexicale

Nous avons travaillé avec un lexique vietnamien de 5000 entrées que nous avons intégré à la base de données ULSID (UCLA Lexical and Segmental Inventory Database), développée au sein de notre laboratoire [11]. Nous observons que /t/ et /ŋ/ qui sont les mieux identifiés par les sujets, possèdent les meilleures fréquences d'occurrences dans leur catégorie (plosives et nasales) toutes distributions confondues (cf. Fig. 3).

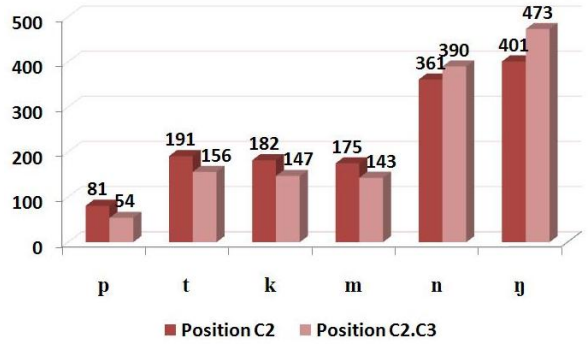


Figure 3: Fréquence lexicale des consonnes en coda.

Les nasales sont particulièrement présentes en position finale de syllabe (77%) par rapport à la position initiale (23%). Ce qui pourrait expliquer un meilleur résultat d'identification des nasales par rapport aux plosives en position de coda.

Les nasales /m/ et /n/ très fréquentes en coda par rapport à la position d'attaque obtiennent des scores d'identification (respectivement 85% et 86%) proches de ceux des plosives /t/ et /k/ (respectivement 84% et 83%) qui sont cependant plus fréquentes en position d'attaque que de coda.

Une tendance entre le score de bonnes réponses et la fréquence d'occurrences des consonnes en position de coda dans le lexique a été effectivement observée. Les coefficients calculés (R^2) indiquent une corrélation importante entre fréquence d'occurrences et score d'identification des plosives comme des nasales, respectivement 0,76 et 0,89. Cependant, cette corrélation n'est plus observée dans le contexte /a/ - D1 pour les plosives ($R^2 = 0{,}28$). Par conséquent, la fréquence lexicale n'explique pas tout. Quels autres facteurs ont pu guider les sujets dans leur choix de réponse ?

4.2. Indice acoustique - le burst

Malgré la caractéristique non relâchée des plosives finales en vietnamien, nous avons constaté la présence de bursts dans 34,5% des réalisations. Il est intéressant de noter que dans ces cas, les bursts sont de durée brève (en moyenne 4 ms) et d'intensité plus faible (60 dB en moyenne) comparés à ceux produits en position initiale (10 ms et 68 dB en moyenne).

En considérant la présence / absence de burst dans la réalisation de la consonne en coda des stimuli, les scores montrent qu'ils sont significativement meilleurs lorsque le burst est présent et ce quelle que soit la frontière qui suit la consonne (frontière de mot ou de syllabe intra-mots) (cf. Fig. 4). La tendance est identique lorsque sont considérés les lieux d'articulation.

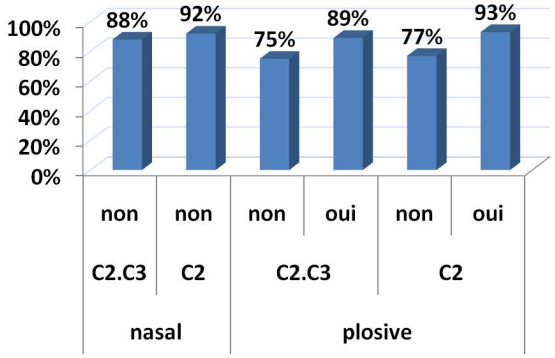


Figure 4: Score de bonnes réponses en fonction de la présence/absence du burst (oui/non) et du type de frontière : intra-mot ($C_2.C_3$), inter-mot (C_2).

Ces résultats montrent clairement que le burst facilite l'identification du lieu d'articulation de la consonne. Pourtant, le nombre de plosives finales /p/ /t/ /k/ relâchées (avec présence de burst) dans les stimuli du test est beaucoup plus restreint (respectivement 25%, 50% et 22%) par rapport au nombre de plosives non relâchées. Le score d'identification des plosives non relâchées est au-dessus du seuil du hasard. Ces plosives comportent vraisemblablement d'autres indices acoustiques porteurs d'informations sur leurs lieux d'articulation.

4.3. Indice acoustique - transition formantique

L'importance de la transition formantique pour la perception des consonnes dans des séquences bi-segmentales CV a été mise en évidence dans la littérature depuis Delattre [3]. Pour la transition VC des plosives de notre étude, nous avons observé l'évolution des 3 premiers formants F1, F2, F3, de la fréquence fondamentale et de l'intensité, estimée à partir de la différence des valeurs mesurées à 50% (partie la moins influencée par les consonnes de l'entourage) et 90% (partie de transition entre voyelle et consonne finale) de la durée de la voyelle. Des analyses statistiques ont été effectuées sous SPSS pour repérer d'éventuels effets d'interaction entre lieux d'articulation et paramètres acoustiques dans cette partie de la transition voyelle - consonne.

Les ANOVA montrent globalement un effet non significatif de la transition de F0 entre les plosives [$F(2,28) = 0{,}63$; $p = 0{,}54$]. Que le lieu soit labial, coronal ou vélaire, les analyses ont également montré un effet non significatif des valeurs de delta F3 pour les plosives finales [$F(2,28) = 1{,}94$; $p = 0{,}16$]. Les mêmes tests effectués pour les valeurs de delta F1 [$F(2,28) = 3{,}8$; $p = 0{,}04$], delta F2 [$F(2,28) = 4{,}55$; $p = 0{,}02$] et delta Intensité [$F(2,28) = 7{,}83$; $p = 0{,}00$] montrent des différences significatives de transition en fonction du lieu d'articulation des plosives. Ces résultats sont en conformité avec les études antérieures [8] sur le rôle des transitions formantiques dans la perception du lieu d'articulation des consonnes.

4.4. *Indice acoustique - MFCC*

Nous avons calculé des coefficients MFCC (*Mel-Frequency Cepstrum Coefficients*) et leurs dérivées sur une fenêtre glissante pour caractériser les particularités spectrales dans la partie de transition située à partir des 90% de la durée de la voyelle. L'idée principale était d'extraire, au niveau spectral, des informations permettant une classification des voyelles [a] en fonction du lieu d'articulation des consonnes qui la suivent. L'analyse en composantes principales (PCA) a été effectuée à partir de ces données (cf. Fig. 5).

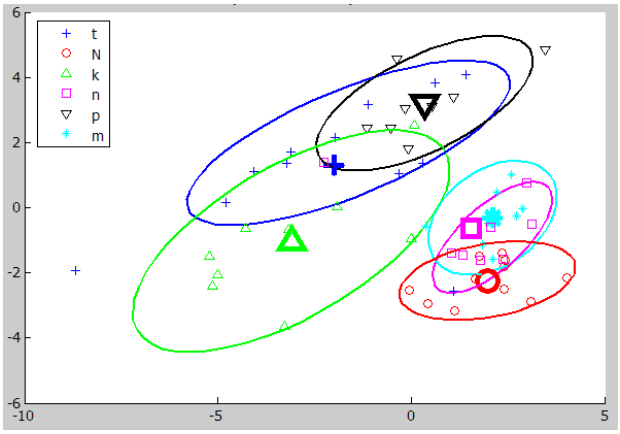


Figure 5: Analyse en composantes principales des coefficients MFCC-PCA de la transition à 90% de [a] dans les contextes consonantiques [t ŋ k n p m].

Cette analyse montre qu'à 90%, la voyelle contient des informations du lieu d'articulation de la consonne finale qui diffèrent par les caractéristiques spectrales mises en place pour les former : les zones de dispersion sont plus à l'écart pour les vélaires [k ŋ] qui obtiennent les meilleurs scores d'identification ; les zones des coronales situées entre celles des vélaires et des labiales expliquent probablement les confusions de réponses dans l'identification de /p/ et /t/ par les sujets vietnamiens. Cette analyse ne nous permet pas cependant d'expliquer pourquoi les sujets confondent davantage /k/ avec /p/ (11%) qu'avec [t] (6%). L'explication pour cette confusion reste à discuter.

5. Conclusion

Cette étude montre que des sujets vietnamiens, dans leur langue maternelle, identifient mieux les nasales que les plosives en position de coda. Bien que les plosives soient généralement non relâchées dans cette position, nos résultats montrent que la présence d'un burst même faible augmente le taux d'identification correcte. Dans le cas des plosives non relâchées, les scores restent supérieurs à 70%. Les analyses acoustiques ont montré qu'en l'absence de burst, les consonnes contenaient, dans leur transition avec la voyelle précédente, les informations sur leur lieu d'articulation, lesquels ont sans doute guidé les sujets dans leur choix de réponse.

Les résultats présentés ici constituent une étape dans la constitution d'un paradigme expérimental rendant compte de la perception des séquences de consonnes que l'on peut rencontrer en français et en vietnamien. Ces résultats sont actuellement complétés par une série d'expériences portant sur la perception et production des consonnes du français par des sujets vietnamiens. Les résultats seront mis en relation avec des données en production obtenues avec un articulographe électromagnétique EMA® (Carstens) pour des consonnes du français et du vietnamien produites par quatre Vietnamiens natifs.

Cette recherche a bénéficié d'un financement de l'Agence Universitaire de la Francophonie (PC - 411/2460).

Un grand merci à René Carré et Lionel Granjon pour l'aide matérielle qu'ils ont chacun apportée à ce projet et pour les échanges constructifs et discussions très fructueuses.

Références

[1] C. T. Best et al. Nonnative and second-language speech perception : Commonalities and complementarities. In *Second language speech learning : The role of language experience in speech perception and production*, pages 13–34. John Benjamins, Amsterdam, 2006.

[2] A. Borrell. Les rapports entre perception et (re)production dans l'acquisition des langues secondes et/ou étrangères. In *Cahiers du Centre Interdiscipline des Science du langage*, pages 29–41. Université de Toulouse, 1992.

[3] P. C. Delattre et al. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27 :769–773, 1995.

[4] T. T. Doan. *Ngu am tieng Viet (Tr. La phonétique du vietnamien)*. Edition de l'Université Nationale, Hanoi, 1999.

[5] A. Michaud. Final consonants and glottalization : New perspectives from hanoi vietnamese. *Phonetica*, 61 :119–146, 2004.

[6] T. B. M. Nguyen. *Regards sur l'enseignement de la phonétique dans la formation des étudiants en FLE à l'Université Pédagogique de Ho Chi Minh ville*. PhD thesis, Université de Rouen, France, 2000.

[7] I. Rousset. *Structures syllabiques et lexicales des langues du monde. Données, typologiques, tendances universelles et contraintes substantielles*. PhD thesis, Université Stendhal, Grenoble, France, 2004.

[8] W. Serniclaes. *Etude Expérimentale de la perception du trait de voisement des occlusives du français*. PhD thesis, Université Libre de Bruxelles, Belges, 2004.

[9] T. T. H. Tran and N. Vallee. An acoustic study of interword consonant sequences in vietnamese. *Journal of the Southeast Asian Linguistics Society*, pages 231–249, 2009.

[10] V. C. Truong. *Structure de la langue vietnamienne*. Centre universitaire des langues orientales vivantes, Paris, 1970.

[11] N. Vallee, S. Rossato, and I. Rousset. Favoured syllabic patterns in the world's languages and sensorimotor constraints. In F. Pellegrino, E. Marsico, I. Chitoran, and C. Coupé, editors, *Approaches to Phonological Complexity*. Mouton de Gruyter, Berlin, 2009.

Analyse acoustique d'un contraste dérivé en anglais d'Écosse

Emmanuel Ferragne[1], Joana Afonso-Santiago[1], François Pellegrino[2]

[1] CLILLAC-ARP – EA 3967/Université Paris 7
10 rue Charles V – 75004 PARIS
[2]Laboratoire Dynamique Du Langage – UMR 5596/CNRS – Université Lyon 2
14 avenue Berthelot – 69007 LYON
emmanuel.ferragne@univ-paris-diderot.fr

ABSTRACT

Phonological length is generally thought not to be relevant in Scottish English. However, some vowels are appreciably longer when followed by the past morpheme /d/ (instead of a tautomorphemic /d/), resulting in so-called 'derived contrasts'. Do derived contrasts qualify as full-fledged phonemic oppositions or should they be regarded as allophonic variation ? In the current paper, which is a preliminary to perceptual experiments addressing this question, we quantify the differences in duration and spectral variation brought about by appending the dental suffix of the past morpheme to /u/, /i/, and /ai/. We show that consistent lengthening occurs for /u/ and /ai/, as well as stronger diphthongization in /ai/, which suggests that duration should perhaps be included in the vowel system of Scottish English.
Keywords: derived contrasts, Scottish Vowel Length Rule, Gradient Phonemicity Hypothesis, vowel length, diphthongization

1. Introduction

L'analyse phonologique traditionnelle pose qu'une différence entre deux sons de parole reflète soit un contraste phonémique, soit une variation allophonique, à l'exclusion de toute autre option. Cependant, de nombreux auteurs mettent en doute cette vision strictement binaire [5, 7, 2]. En effet, pour diverses raisons – rendement fonctionnel faible, distribution lacunaire, les langues du monde présentent dans leur inventaire phonologique établi selon la méthode classique certains phonèmes qui sont moins typiquement phonémiques que d'autres. C'est par exemple le cas de /ð/ *vs* /θ/ en anglais britannique standard : la sonore n'apparaît en initiale de mot que dans des items grammaticaux, et le rendement fonctionnel de cette opposition est très limité (*thy/thigh*, et, dans certaines variétés, *either/ether*). À l'inverse, certains sons traditionnellement analysés comme des allophones d'un même phonème peuvent donner lieu à des contrastes ; c'est le cas des contrastes dérivés [4] : des contrastes inexistants dans des items mono-morphémiques émergent lors de dérivations morphologiques. Il s'ensuit donc que certaines entités analysées comme allophones sont moins typiquement allophoniques que d'autres. En suivant les auteurs qui ont bien compris le défi que représentent ces phénomènes pour l'analyse phonologique (voir références précédemment citées), nous formulerons l'hypothèse de la phonémicité gradiente (HPG), selon laquelle il existe plusieurs degrés d'allophonie et de phonémicité. De plus, HPG postule que ces statuts linguistiques intermédiaires donnent lieu à des corrélats cognitifs variables, différents de ceux élicités par des différences phonémiques ou allophoniques typiques, et identifiables par le biais d'expériences de perception.

Cet article présente des données de production préliminaires à l'analyse expérimentale d'un contraste dérivé typique de l'anglais d'Écosse, et notamment de Glasgow. Il s'agit d'une analyse phonétique acoustique d'un allongement vocalique conditionné par la présence d'une frontière morphémique. Cette particularité, rangée avec d'autres phénomènes sous l'appellation *Scottish Vowel Length Rule*, se manifeste lorsque les voyelles /i/, /u/ et /ai/ sont allongées (avec variation de timbre pour /ai/) sous l'effet du suffixe du passé /d/ [9, 4, 6, 3]. Ainsi, *need/kneed*, *brood/brewed* et *side/sighed* sont susceptibles de constituer autant de paires minimales au sens strict du terme. Cette question, du point de vue de l'analyse de la durée, a certes été abondamment traitée dans la littérature, mais l'analyse que nous proposons ici diffère des précédentes sur plusieurs points :

- ces données constitueront le matériel de base d'expériences de perception, il convient donc de :
  - quantifier précisément la variation sur les paramètres pertinents,
  - s'assurer que tous les locuteurs produisent ce contraste de façon cohérente.
- certains mots tests sont peu fréquents, voire non-attestés ; il s'agira donc d'observer dans quelle mesure ce facteur intervient,
- la variation de timbre pour /ai/ sera quantifiée acoustiquement.

2. Méthode

Une liste de 12 paires minimales potentielles a été conçue (Table 1), puis augmentée de 6 distracteurs. Les mots-cibles ont été présentés dans la phrase porteuse *He said the word ... and I didn't know how to spell it.* Quinze étudiants de l'Université de Glasgow ont pris part aux enregistrements, qui ont eu lieu dans un studio dédié. Chaque participant a été invité à lire l'intégralité de la liste à trois reprises de manière totalement autonome par le biais d'une interface en Tcl/Tk conçue par le troisième auteur. La liste était présentée dans un ordre aléatoire différent d'un sujet à l'autre. Le signal a été directement converti au

format numérique PCM mono avec un taux d'échantillonnage de 44,1 kHz et une résolution de 16 bits.

Table 1: Liste des 12 paires minimales employées dans l'étude.

Numéro	Paire
1	bide-bye'd
2	brood-brewed
3	could-cooed
4	crude-crewed
5	need-kneed
6	mood-moo'd
7	pride-pried
8	ride-rye'd
9	rude-rued
10	side-sighed
11	tide-tied
12	would-wooed

2.1. Analyse auditive préliminaire

Une analyse auditive préliminaire, conduite par le premier auteur, fait apparaître que les productions vont d'une absence totale de distinction audible (locutrice *gla09*) à un contraste systématique entre les deux membres de chacune des 12 paires (locuteur *gla14*). Sur les 15 (locuteurs) × 12 (paires) = 180 contrastes potentiels, 114 ont été perçus comme distincts, 32 comme non distincts, et dans les 34 cas restants, aucune conclusion fiable n'a pu être émise en se basant sur l'analyse auditive. Ces résultats sont reproduits dans la Table 2. Pour le reste de l'analyse, les locuteurs produisant un contraste perceptible pour au moins 6 paires ont été conservés. Ainsi, les locuteurs *gla02, gla09* et *gla15* ont été écartés. En outre, on constate dans la Table 2 que la paire 5 (*need-kneed*) ne s'est révélée contrastive que pour un seul locuteur (*gla14*) ; par conséquent, elle n'apparaît pas dans le reste de l'analyse.

Table 2: Résultats de l'analyse auditive : perçu comme différent (+), identique (-), doute (x).

Locuteur	1	2	3	4	5	6	7	8	9	10	11	12
gla01	+	+	+	+	-	+	+	+	+	+	+	+
gla02	-	-	-	-	-	-	-	-	-	+	+	-
gla03	+	+	+	x	-	+	x	+	+	+	+	+
gla04	+	+	+	+	x	+	x	-	+	+	x	+
gla05	+	+	+	+	x	+	x	+	+	+	x	+
gla06	+	-	+	x	-	x	+	+	x	+	+	x
gla07	+	+	+	+	x	+	x	+	x	+	+	+
gla08	+	+	+	+	-	+	+	+	+	+	+	+
gla09	-	-	x	x	-	-	-	-	x	-	-	-
gla10	-	x	+	x	x	-	+	+	+	+	+	+
gla11	+	+	+	+	-	+	+	+	+	+	+	+
gla12	+	+	x	+	-	x	+	+	x	+	+	+
gla13	x	x	+	x	-	x	+	+	+	+	+	x
gla14	+	+	+	+	+	+	+	+	+	+	+	+
gla15	+	x	x	x	-	x	+	+	x	+	+	-

2.2. Analyse de la durée et du timbre

Les voyelles, les mots-tests et les pauses ont été segmentés manuellement. La durée de chaque voyelle a ensuite été calculée. Pour ce qui est de la variation de timbre des voyelles, l'analyse cepstrale a été préférée à la méthode plus classique des formants en raison du manque de fiabilité de l'estimation automatique de formants. Douze coefficients MFCC et l'énergie ont été calculés avec Praat (options par défaut). Le paramètre choisi pour définir le degré de diphtongaison est la distance euclidienne dans l'espace MFCC entre la première et la dernière trame de chaque voyelle. Deux ANOVA à deux facteurs (SUFFIXATION et TYPE DE VOYELLE) ont été conduites sur la durée (Section 3.1) et le degré de diphtongaison (Section 3.2) indépendamment. Puis, puisque le facteur SUJET n'est pas inclus explicitement dans ces deux analyses afin de ne pas amplifier la complexité des deux modèles, la variation inter-individuelle sera examinée dans la Section 3.3 au moyen de t-tests.

3. Résultats

3.1. Durée

Les 792 valeurs de durée – 12 locuteurs × 22 stimuli (11 paires) × 3 répétitions – ont été soumises à une analyse de la variance à deux facteurs : SUFFIXATION (oui *vs* non) et TYPE DE VOYELLE (/u/ *vs* /ai/). Les résultats montrent un effet significatif de la SUFFIXATION ($F(1,791) = 1205\ ; p < 0,001$), du TYPE DE VOYELLE ($F(1,791) = 622,79\ ; p < 0,001$) ainsi qu'une interaction significative entre les deux facteurs ($F(1,791) = 49,73\ ; p < 0,001$). La durée objective des voyelles est donc influencée par la présence (longue) ou non (courte) d'une frontière morphémique avec le /d/. De plus, le TYPE DE VOYELLE (/u/ *vs* /ai/) a un effet sur la durée objective. Enfin, l'interaction montre que l'effet de la SUFFIXATION est différent selon qu'il s'agit de /u/ ou /ai/. Les comparaisons multiples *post-hoc* sont présentées dans la Table 3, qui récapitule la différence de durée moyenne estimée entre les quatre groupes de voyelles définis par les facteurs SUFFIXATION et TYPE DE VOYELLE. Les moyennes estimées sont de 204 ms, 96 ms, 250 ms et 179 ms pour /ai/ non suffixé, /u/ non suffixé, /ai/ suffixé et /u/ suffixé, respectivement. Dans la Table 3, on relève que l'allongement moyen estimé de /u/ sous l'effet de la SUFFIXATION (83 ms) est presque deux fois supérieur à celui observé pour /ai/ (46 ms). Non seulement le /ai/ non suffixé est beaucoup plus long que le /u/ non suffixé (108 ms), mais sa durée est également supérieure à celle du /u/ suffixé (25 ms). En termes de pourcentages, la SUFFIXATION provoque un allongement de 23 % pour /ai/ et de 86 % pour /u/.

3.2. Timbre

Les 792 valeurs de timbre ont fait l'objet d'une ANOVA à deux facteurs, sur le même modèle que celle utilisée dans la Section 3.1. On observe un effet significatif du TYPE DE VOYELLE ($F(1,791) = 1890,12\ ; p < 0,001$), de la SUFFIXATION ($F(1,791) = 30,47\ ; p < 0,001$) et une interaction marginalement significative ($F(1,791) = 5,85\ ; p = 0,016$). Les comparaisons *post*

Table 3: Différence de durée moyenne (ms) estimée entre les 4 catégories de voyelles définies par les facteurs SUFFIXATION (NS : non suffixé, S : suffixé) et TYPE DE VOYELLE (/ai/ *vs* /u/).

	/ai/ NS	/u/ NS	/ai/ S	/u/ S
/ai/ NS	0	108	-46	25
/u/ NS		0	-154	-83
/ai/ S			0	71
/u/ S				0

hoc montrent que le /u/ suffixé n'est pas significativement plus diphtongué que le /u/ non suffixé. En revanche, non seulement les deux types de /ai/ sont plus diphtongués que les /u/, mais /ai/ suffixé présente un degré de diphtongaison plus élevé que /ai/ non suffixé.

3.3. Variation inter-individuelle : durée et timbre

Les deux analyses de la variance présentées *supra* (Section 3.1 et 3.2) offrent une vision globale de la variation de durée et de timbre en fonction de la SUFFIXATION et du TYPE DE VOYELLE. Cependant, comme le montre la Figure 1, la dispersion dans les dimensions de la durée et de la distance dans l'espace MFCC des 4 catégories de voyelles définies par les facteurs SUFFIXATION et TYPE DE VOYELLE présente des chevauchements importants. Il est probable que ce résultat soit en partie imputable à des productions individuelles variables. En d'autres termes, les 12 locuteurs n'emploient pas tous la durée et la variation de timbre de la même manière. Quatre t-tests ont été effectués pour chaque locuteur afin de comparer individuellement la différence de durée et la différence de diphtongaison pour /ai/ suffixé *vs* /ai/ non suffixé et /u/ suffixé *vs* /u/ non suffixé. Les résultats sont rapportés dans la Table 4. On y constate que, très vraisemblablement, seule la durée entre en ligne de compte dans la réalisation du contraste /u/ suffixé *vs* /u/ non suffixé chez tous les locuteurs. Pour ce qui est de l'effet de la suffixation sur /ai/, tous les locuteurs (à l'exception du locuteur *gla06*) semblent réaliser l'opposition au moyen de la durée. Quatre d'entre eux produisent en outre une différence de degré de diphtongaison.

4. Discussion

Dans leur ensemble, les résultats sont en accord avec d'autres descriptions de la *Scottish Vowel Length Rule* [9, 6] : /u/ est affecté par l'ajout du morphème à dentale du passé ; ce phénomène semble se traduire exclusivement par l'allongement de la voyelle, et ceci, chez les 12 locuteurs analysés. En revanche, la suffixation après /ai/ engendre non seulement une augmentation – cependant moindre – de la durée de façon quasi-unanime, mais elle génère également une différence dans le degré de diphtongaison. Ce dernier paramètre ne semble être attesté que chez 4 locuteurs. La modification des voyelles /u/ et /ai/ sous l'effet de l'ajout du morphème du passé se traduit donc principalement

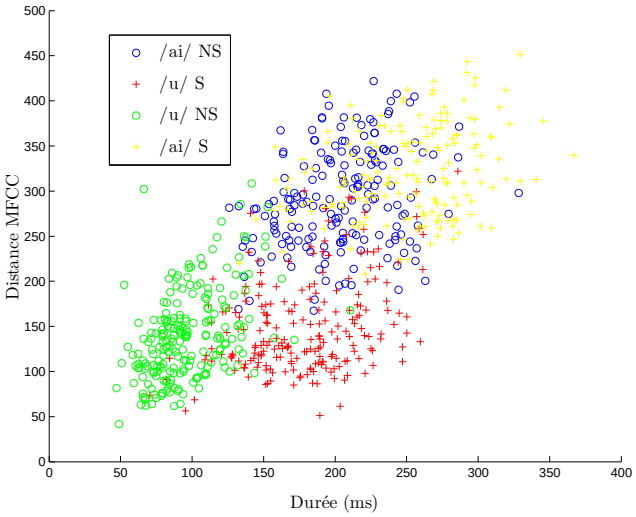


Figure 1: Dispersion des 4 types de voyelles en fonction de la durée et de la distance dans l'espace MFCC.

Table 4: Résultats des t-tests pour chaque individu. 1 : on rejette l'hypothèse nulle d'égalité des moyennes entre la version suffixée et non suffixée du timbre en question (/u/ ou /ai/) sur le paramètre concerné (durée ou variation de timbre). 0 : on ne peut pas rejeter l'hypothèse nulle. Alpha est fixé à $\alpha = 0,01$.

Sujet	/u/		/ai/	
	durée	timbre	durée	timbre
gla01	1	0	1	0
gla03	1	0	1	1
gla04	1	0	1	1
gla05	1	0	1	1
gla06	1	0	0	0
gla07	1	0	1	0
gla08	1	0	1	0
gla10	1	0	1	0
gla11	1	0	1	0
gla12	1	0	1	1
gla13	1	0	1	0
gla14	1	0	1	0

par un allongement. Cependant, il reste à définir si le paramètre que nous avons utilisé (distance euclidienne dans l'espace MFCC entre première et dernière trame temporelle) représente le degré de diphtongaison de façon adéquate. En effet, cette mesure ne prend en compte que le déplacement (au sens cinématique) et non pas la distance parcourue. Il se pourrait donc que notre mesure sous-estime le degré de diphtongaison. De plus, la distance que nous calculons ne rend pas compte des différences de valeur absolue du timbre au niveau du point de départ de la voyelle et de son point d'arrivée. Or les descriptions antérieures notent des timbres différents selon que /ai/ est suffixé – [ae] – ou pas – [ʌi] [8]. Il conviendra donc de prendre en compte ces différences dans la suite de nos travaux.

Parmi les autres améliorations possibles, la prise en compte du débit de parole fera l'objet d'une étude ultérieure. Ensuite, l'influence respective de la durée

et de la diphtongaison pour le contraste /ai/ suffixé *vs* /ai/ non suffixé devra être déterminée par le biais d'une expérience de perception qui constitue un prolongement nécessaire de l'étude rapportée ici.

La réalisation effective du contraste dérivé dans certains mot-tests totalement obsolètes – e.g. les verbes *bye* et *rye*, d'après la 2e édition du *Oxford English Dictionary* – nous laisse penser qu'il s'agit d'un phénomène productif chez nos 12 sujets. Ceci implique que le contraste dérivé étudié ici reste disponible pour de nouvelles formations *ad hoc* et, éventuellement, pour une diffusion plus vaste dans le lexique. Toutefois, la productivité étant influencée par la fréquence de type d'un phénomène [1], le faible nombre de paires minimales pour ce contraste pourrait constituer un frein à sa diffusion. Sur un plan méthodologique, notons que l'utilisation de l'apostrophe avant le <d> dans la graphie des stimuli a probablement fortement influencé les sujets en faveur d'un découpage en plusieurs morphèmes. C'est l'explication que nous retenons pour rendre compte du fait que si *rye'd* a engendré la réalisation attendue, *kneed* n'a pas donné lieu à l'allongement escompté.

Comme nous l'avons remarqué dans l'Introduction, les contrastes dérivés ont un statut ambigu en linguistique. Si l'on se fie au test des paires minimales, le contraste que nous étudions ici mériterait le statut de phonème. Il convient cependant de noter que son rendement fonctionnel – qui se résume virtuellement aux paires présentées dans cet article – est très faible, ce qui implique que ce contraste est moins typiquement phonémique que d'autres. Le critère de la prédictibilité – lorsqu'un choix entre deux sons n'est pas déterminé par leur environnement phonétique, ces deux sons sont en contraste phonémique – n'est pas réellement applicable ici puisque, certes, la réalisation du /u/ ou du /ai/ est prévisible, mais seulement à partir de la morphologie. Le critère de la proximité phonétique – deux allophones ont tendance à être proches, deux phonèmes, éloignés – n'est guère applicable, car une telle notion est difficilement quantifiable et probablement erronée. Divers auteurs ont proposé des appellations variées pour faire référence au statut intermédiaire de certaines oppositions, et notamment des contrastes dérivées, rompant ainsi avec le choix binaire de l'analyse classique selon laquelle une différence entre deux sons de parole ne peut être que exclusivement phonémique ou exclusivement allophonique. Qu'il s'agisse de contrastes *quasi-contrastive*, *quasi-allophonic* ou encore de *fuzzy contrasts* (voir [2] pour une liste des dénominations), ces cas-limites de l'analyse phonologique seront testés expérimentalement dans une étude ultérieure dans le cadre de l'Hypothèse de la Phonémicité Gradiente, selon laquelle un statut cognitif intermédiaire existe entre l'allophonie et la phonémicité.

5. Conclusion

Cette étude avait pour but de quantifier la variation de durée induite par l'ajout du morphème du passé /d/ aux voyelles /i/, /u/ et /ai/. Il s'agissait également de quantifier la différence de timbre pour la voyelle /ai/ dans les mêmes conditions. Le contraste potentiel pour la voyelle /i/ n'a pas pu être élicité, probablement en raison d'un mauvais choix dans la graphie des stimuli. En ce qui concerne /u/, la suffixation génère un allongement moyen de 86 % par rapport à la condition où la voyelle est suivi d'un /d/ appartenant au même morphème. Aucune différence de timbre n'a pu être mise en évidence. Pour ce qui est de /ai/, la suffixation engendre un allongement moyen de 23 % par rapport à la condition mono-morphémique, parfois accompagnée d'une variation de timbre. Ces productions constitueront le matériel de base d'expériences de perception visant à préciser le statut cognitif de ce contraste dérivé.

6. Remerciements

Ce travail est soutenu par une Subvention de Recherche de la Fondation Fyssen. Nous tenons à remercier Jane Stuart-Smith pour son aide lors de la collecte des données.

Références

[1] J. Bybee. *Phonology and Language Use*. Cambridge University Press, Cambridge, 2001.

[2] K. Currie Hall. *A Probabilistic Model of Phonological Relationships from Contrast to Allophony*. Ph.D dissertation, Ohio State University, 2009.

[3] E. Ferragne and F. Pellegrino. Formant frequencies in 13 accents of the British Isles. *Journal of the International Phonetic Association*, sous presse.

[4] J. Harris. Derived phonological contrasts. In S. Ramsaran, editor, *Studies in the Pronunciation of English : A Commemorative Volume in Honour of A.C. Gimson*, pages 87–105. Routlegde, Londres, 1990.

[5] W. Labov. *Principles of Linguistic Change : Internal Factors*. Blackwell, Cambridge, [Mass.], 1994.

[6] J.M. Scobbie, N. Hewlett, and A. Turk. Standard English in Edinburgh and Glasgow : The Scottish Vowel Length Rule revealed. In P. Foulkes and G. Docherty, editors, *Urban Voices : Accent Studies in the British Isles*, pages 230–245. Arnold, Londres, 1999.

[7] J.M. Scobbie and J. Stuart-Smith. Quasi-phonemic contrast and the fuzzy inventory : examples from Scottish English. In P. Avery, B. Dresher, and K. Elan Rice, editors, *Contrast in Phonology Theory, Perception, Acquisition*, pages 87–114. Mouton de Gruyter, Berlin, 2008.

[8] J. Stuart-Smith. Glasgow : Accent and voice quality. In P. Foulkes and G. Docherty, editors, *Urban Voices : Accent Studies in the British Isles*, pages 203–222. Arnold, Londres, 1999.

[9] J.C. Wells. *Accents of English : The British Isles*, volume 2. Cambridge University Press, Cambridge, 1982.

Comparaison du timing inter-gestuel des voyelles nasales en français de Marseille et de Tournai.

Kathy Huet[*], *Sandrine Clairet*[*], *Gilles Delmée*[*], *Véronique Delvaux*[*+], *Myriam Piccaluga*[*], *Bernard Harmegnies*[*]

*Laboratoire des sciences de la parole de l'Académie Universitaire Wallonie-Bruxelles, [+] FNRS
UMONS, 18 place du parc, B-7000 Belgique
kathy.huet@umons.ac.be

ABSTRACT

This paper provides a comparative gestural account of nasal vowels in Southern French (Marseille) and in Belgian French (Tournai). Timing measurements based on acoustic data confirm a different temporal organisation of the gestures between the two dialects. We also report on statistical comparisons between phonetic events (such as duration of the SF 'nasal appendix' vs. duration of a typical nasal consonant, duration of the oral consonant following an oral vs. a nasal vowel...) that help to confront competing hypotheses concerning the underlying phonological representation of SF nasal vowels.

Keywords: Southern French, Belgian French, nasal vowels, inter-gestural timing, underlying representation.

1. INTRODUCTION

En français comme dans la plupart des langues du monde, les voyelles nasales distinctives /ṽ/ proviennent de séquences homosyllabiques /VN/, suite à une évolution phonologique en deux étapes successives: (1) nasalisation vocalique par coarticulation ; (2) chute de la consonne coarticulante [1]. Certains phonologues soutiennent qu'en français contemporain la représentation sous-jacente des voyelles nasales demeure /VN/ [2], en particulier dans le cas du français méridional (FM) [3,4], où la réalisation phonétique des voyelles nasales distinctives est habituellement décrite comme une séquence d'une portion vocalique oralisée, d'une portion vocalique nasalisée et d'un appendice consonantique nasal, soit /ṽ/ = [$V^{\tilde{V}N}$] ([5,6,7]; pour une analyse alternative en /ṽ/ = [V^{N}]~[$\tilde{V}^{N}$]~[ṽ], voir [8]).

A notre connaissance, une seule étude a abouti à une *quantification précise* de ces trois parties dans la réalisation phonétique des voyelles nasales du FM, en l'occurrence sur base de mesures aérodynamiques collectées auprès de trois locuteurs [7]. Par ailleurs, toutes les études phonétiques et phonologiques mentionnées ci-dessus décrivent les voyelles nasales du FM en termes de séquences d'événements acoustiques (segments ou pseudo-segments), sans investiguer les constellations de gestes qui ont fait émerger lesdits événements acoustiques. Pour notre part, nous estimons qu'une évaluation phonétique du phénomène en termes de *gestes articulatoires* - même si elle est dérivée de données acoustiques - devrait permettre de mieux contribuer au débat concernant la représentation phonologique sous-jacente des voyelles nasales en FM.

Le but de cet article est double. Le premier objectif est descriptif : nous proposons une description acoustique de la réalisation des voyelles nasales du FM (Marseille), en la comparant au français de Belgique (FB), plus spécifiquement dans sa variété de Tournai, réputée proche du français standard [9]. Nous mesurons à cette fin la durée de différents événements acoustiques, segments (V, N, etc.) ou pseudo-segments (partie vocalique orale V, partie vocalique nasalisée $^{\tilde{V}}$, appendice nasal ^{N}).

Le second objectif est de contribuer au débat concernant la représentation phonologique sous-jacente des voyelles nasales en FM. Cette contribution est fondée sur une analyse de l'étendue des gestes articulatoires (inférée à partir des données acoustiques) présidant à l'émergence des (pseudo-)segments acoustiques.

Nous proposons de considérer que si la représentation profonde des voyelles nasales du FM est /ṽ/ (hypothèse 1 ou H1), alors l'implémentation phonétique de celles-ci dans des items $C_1\tilde{V}.C_2V$ consiste exclusivement en une *désynchronisation temporelle* entre les gestes de vibration des cordes vocales et d'occlusion/constriction buccale (pour C_2) d'une part et le geste d'abaissement du voile d'autre part, *l'étendue temporelle* des gestes vélique et buccal étant similaire à ce qui est observé en français de Tournai (voir figure 1). De même, le phasage entre les gestes glottique et vélique devrait être préservé, de sorte que la durée totale de C_2 serait finalement réduite.

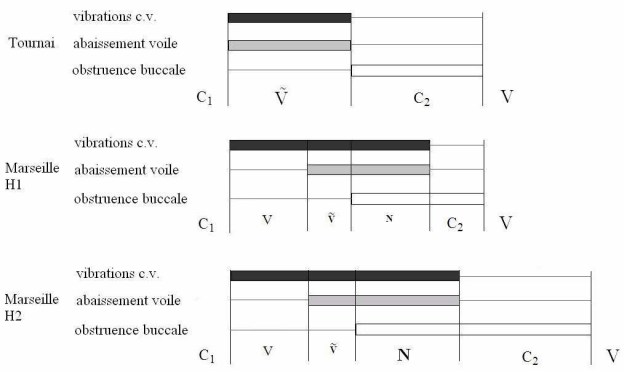


Figure 1. Schématisation de l'organisation temporelle des gestes prédite pour une séquence CṼ.CV pour Tournai (haut), et pour Marseille dans le cas de 2 hypothèses concurrentes sur la représentation phonologique profonde H1 (milieu) et H2 (bas).

Alternativement, si la représentation profonde des voyelles nasales du FM est /VN/ (H2), nous faisons l'hypothèse de l'émergence d'un véritable segment consonantique nasal N impliquant un allongement du geste d'abaissement vélique par rapport au français de Tournai, ainsi qu'un allongement du geste d'obstruence buccale préservant la durée complète de C_2 (dont l'attaque est ainsi retardée par rapport à la voyelle) (voir figure 1).

2. MÉTHODE

2.1. Corpus et locuteurs

Un corpus contrôlé de parole lue constitué de mots bi-syllabiques ordonnés en triplets de type $C_1V.C_2V$ $C_1\tilde{V}.C_2V$ $C_1V.NV$ (e.g. /tate/ /tãte/ /tane/) a été construit en fonction de deux critères : (i) existence de paires minimales : CV.CV ~ C$\tilde{V}$.CV et CV.CV ~ CV.NV (ii) homorganicité entre l'appendice nasal de la voyelle nasale dans C$\tilde{V}$.CV et la consonne nasale dans CV.NV. Notre corpus est ainsi composé de douze triplets, quatre par paire nasale~orale /ɛ̃~ɛ ; ɑ̃~a ; ɔ̃~ɔ/, où C_1 est une occlusive ou une fricative non voisée, C_2 est une occlusive non voisée et la voyelle finale est /e/. La voyelle /œ̃/ a été exclue vu son instabilité en français contemporain. L'ensemble du corpus a été répété trois fois par chaque locuteur.

Dix locuteurs de Marseille (FM) et 10 locuteurs de Tournai (FB) (H et F) ont été enregistrés. Tous sont nés, travaillent et ont toujours habité dans la ville choisie. Ils ont entre 30 et 60 ans, et ont été sélectionnés sur base de leur appartenance à la classe socio-professionnelle moyenne.

2.2. Segmentation

Une segmentation manuelle des phonèmes et, plus finement, des trois parties de la réalisation des voyelles nasales, a été réalisée sous PRAAT 5.0.47 [10] par deux experts selon des critères de segmentation communs, puis a été suivie d'une vérification croisée. Aux traditionnels critères de segmentation en phonèmes s'ajoutent donc ceux utilisés pour déceler à quel moment commence la nasalité dans la voyelle nasale (afin d'inférer le début du geste d'abaissement du velum, cf. [11]).

Il est établi que la nasalisation d'une voyelle introduit des pôles et des zéros spectraux dont le principal effet est un affaiblissement dans les basses fréquences (zone de F1), mais également un affaiblissement dans la zone de F3, parfois appelé *'œil nasal'* (notion qui désigne la réduction importante de l'énergie entre 2000Hz et 3000Hz particulièrement saillante pour les postérieures du français [12, 13]), et plus globalement une atténuation du niveau général d'énergie (entre autres [12, 14]). Notre segmentation s'est appuyée sur l'examen de ces différents paramètres, évalués à l'aide d'un spectrogramme (formants), de la courbe d'intensité (énergie globale), et de la forme d'onde (changements de timbre et d'intensité).

La figure 2 présente un exemple de segmentation d'un triplet en FM.

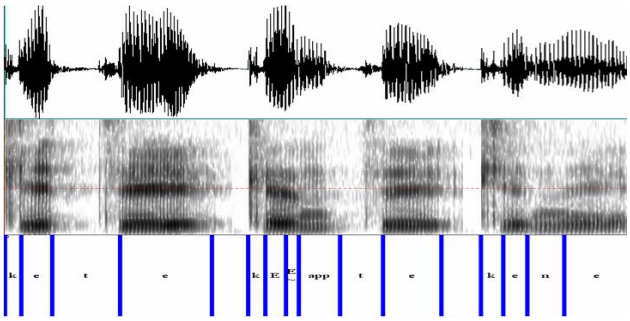


Figure 2 : Exemple de segmentation d'un triplet en FM**.** (L'intervalle de silence entre les 3 mots a été réduit ici).

3. MESURES ET ANALYSE DES RÉSULTATS

Les résultats présentés ici portent sur la mesure des durées de chaque événement acoustique (segments et/ou parties de la voyelle nasale) de l'ensemble des productions des dix locuteurs de Marseille et de quatre locuteurs de Tournai. Dans un premier temps, une description comparative de l'organisation temporelle des événements acoustiques est réalisée entre les deux parlers. Dans un second temps, des analyses comparatives de la durée d'événements particuliers ont été menées afin d'investiguer les hypothèses présentées à la figure 1.

3.1. Organisation temporelle des événements

Dans les mots bi-syllabiques

Les graphes de la figure 3 permettent de comparer l'organisation temporelle des événements composant les mots bi-syllabiques produits par les locuteurs de Tournai (FB) et les locuteurs de Marseille (FM). Ils présentent les durées relatives moyennes (en %) de chaque événement par rapport au mot auquel il appartient (de haut en bas : C_1VC_2V, $C_1\tilde{V}C_2V$ et C_1VNV).

Une première analyse de ces graphes permet de constater que les patrons temporels des locuteurs de Marseille sont très similaires à ceux des locuteurs de Tournai dans le cas des mots ne contenant pas de voyelle nasale, à savoir pour les mots C_1VC_2V (fig.3a) et C_1VNV (fig.3c). Les traitements statistiques de ces données ne montrent aucune différence significative entre les deux parlers excepté pour les voyelles finales (fig.3a : t=-2.569, dl=502, p=0.01 et fig.3c : t=-2.411, dl=502, p=0.016) dont les durées relatives sont en moyenne significativement plus courtes pour Marseille (FM).

Par contre, les durées relatives des événements produits lors de la lecture des mots contenant une voyelle nasale $C_1\tilde{V}C_2V$ (fig3b) présentent des différences significatives conduisant à des patrons temporels très distincts entre les marseillais et les tournaisiens. Les différences les plus importantes se situent au niveau de la partie orale de la voyelle nasale (durée relative moyenne: 11% en FM et 4.1% en FB) et au niveau de la partie nasalisée (5.6% en

FM contre 20.2% en FB). Par ailleurs, les tournaisiens ne présentent pas d'appendice nasal alors que pour les marseillais, celui-ci constitue en moyenne 18% de la durée totale du mot. On observe également que la durée relative moyenne de C_2 est moins importante (20.6%) en FM qu'en FB (29.7%) ainsi que la voyelle finale (FM: 30.1% et FB: 32.6%). Seules les durées relatives moyennes de la consonne initiale ne sont pas significativement différentes entre les deux parlers.

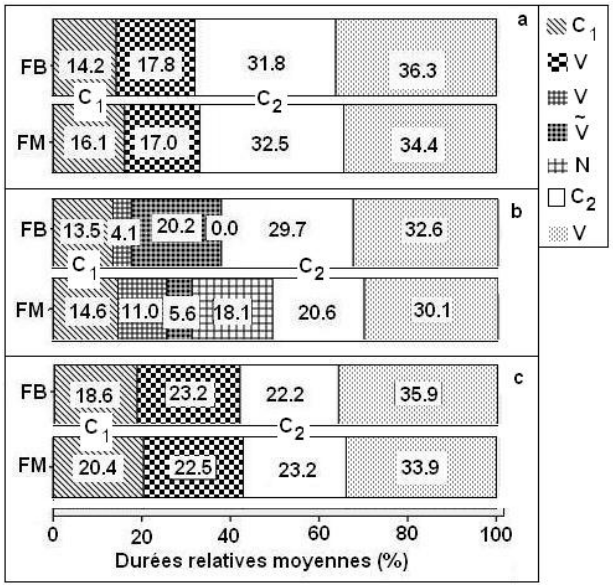


Figure 3 : Durées relatives moyennes des événements des productions pour Tournai (FB) et Marseille (FM), pour C_1VC_2V (a), $C_1\tilde{V}C_2V$ (b) et C_1VNV (c).

Dans la voyelle nasale

L'organisation temporelle des événements constituant la voyelle nasale est différente dans les deux parlers : la partie orale ne constitue que 17.1% de la voyelle nasale en FB alors qu'elle constitue près d'un tiers (31.7%) de $\tilde{v}$ en FM. A contrario, la partie nasalisée constitue 82.9% de la voyelle nasale en FB contre seulement 16.3% en FM. Par ailleurs l'appendice nasal, absent des réalisations en FB, constitue la moitié (52.0%) de $\tilde{v}$ en FM, ce qui contribue à l'allonger. En effet, une comparaison des durées absolues démontre que les voyelles nasales produites par les locuteurs marseillais (185ms) sont en moyenne significativement plus longues ($t=16.567$, $dl=502$, $p<0.001$) que celles produites par les locuteurs tournaisiens (138ms). Le rapport est de 1.3.

L'ensemble de ces constatations est illustré par la figure 4.

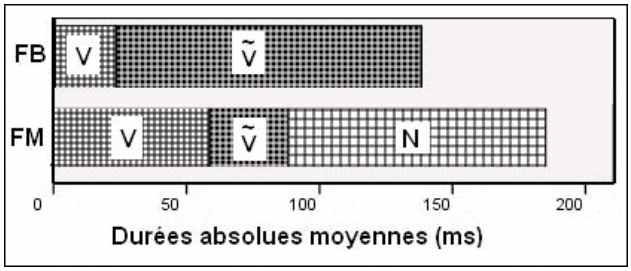


Figure 4 : Durées moyennes (ms) des événements constituant la voyelle nasale pour Tournai (FB) et pour Marseille (FM).

3.2. Analyse comparative des durées

Dans cette section, les durées absolues de certains événements des productions de nos locuteurs sont comparées. Les événements ont été sélectionnés afin de confronter les prédictions des hypothèses concurrentes H1 et H2 développées dans l'introduction (p.ex. durée absolue de C_2 après voyelle orale vs. voyelle nasale).

Voyelle nasale versus voyelle orale

Bien que les deux parlers présentent des voyelles nasales (FB : 138ms, FM : 185ms) plus longues que les voyelles orales (FB : 103ms, FM : 85ms), on n'observe pas le même rapport de durées entre voyelles nasales et voyelles orales : les voyelles nasales sont en moyenne 2.22 fois plus longues que les voyelles orales en FM alors qu'en FB elles ne le sont que de 1.35 fois.

Consonnes C2

La figure 5a compare la durée moyenne de la consonne C_2 dans les productions de C_1VC_2V et de $C_1\tilde{V}C_2V$ pour les locuteurs de Marseille de manière à pouvoir déterminer si la durée de C_2 est plus petite derrière une voyelle nasale (H1). L'analyse statistique de ces données indique que la différence entre ces durées est significativement différente, C_2 étant effectivement plus courte lorsqu'elle est produite à la suite d'une voyelle nasale en FM.

Appendice nasal versus consonne nasale

La figure 5b, quant à elle, présente, pour les locuteurs de Marseille, la durée moyenne de l'appendice nasal de $\tilde{v}$ dans la production $C_1\tilde{V}C_2V$ ainsi que celle de la consonne nasale dans C_1VNV. Une analyse statistique indique que ces deux durées ne sont pas significativement différentes en FM (H2).

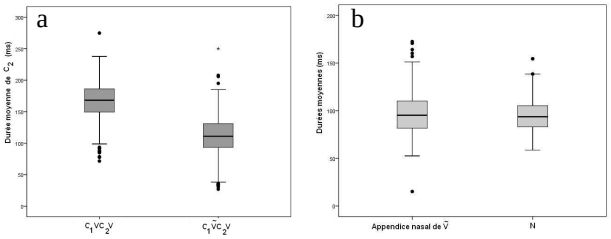


Figure 5: Durée moyenne (ms) en FM de (a) C2 dans C_1VC_2V vs. $C_1\tilde{V}C_2V$ (b) l'appendice ^{N} dans $C_1\tilde{V}C_2V$ vs. la consonne nasale N dans C_1VNV.

4. DISCUSSION

Le premier objectif de cet article était de proposer une première description quantifiée à grande échelle de la production des voyelles nasales en français méridional de Marseille, et de la comparer à un français régional proche du français standard, le français de Tournai. Notre étude a d'abord permis d'établir que la durée relative des phonèmes dans les items CV.CV et CV.NV est similaire dans les deux parlers, à l'exception des voyelles finales

qui sont typiquement allongées en Belgique [9]. Ensuite, la comparaison des productions des items CṼ.CV confirme l'existence de patrons temporels de nasalisation vocalique distincts en FM et en FB. Bien que les voyelles nasales soient intrinsèquement plus longues que les orales dans les deux régiolectes, le rapport de durée entre nasales et orales est nettement plus important en FM.

En effet, les voyelles nasales sont plus longues et plus complexes en FM qu'en FB. A Tournai, la voyelle nasale est produite avec un léger retard de nasalisation, celle-ci débutant approximativement à 15% de la durée totale de la voyelle, ce qui confirme de précédents résultats obtenus en français bruxellois sur base de mesures aérodynamiques [15]. A Marseille, les voyelles nasales sont réalisées comme une séquence de trois événements acoustiques : une portion vocalique orale, une courte portion vocalique nasalisée, et un appendice consonantique nasal d'une durée globalement équivalente aux deux parties précédentes. Même si les proportions relatives de ces trois parties ne sont pas exactement similaires à celles relevées précédemment par Clairet sur trois locuteurs [7], nos deux études établissent qu'en FM la nasalisation débute tardivement, au premier tiers de la nasale, et n'est vocalique que très brièvement, puis se poursuit longuement alors que l'occlusion pour la consonne orale suivante est en place.

En termes de gestes, ceci indique que l'abaissement du voile est retardé en FM par rapport au FB, à tel point qu'il chevauche le geste subséquent d'obstruence buccale. Si l'on évalue la durée du geste d'abaissement vélique à partir des durées des événements acoustiques nasalisés dans les deux régiolectes, on constate que le geste nasal est plus long en FM qu'en FB en termes de durées relatives (fig.3b), mais équivalent en termes de durées absolues (fig.4), ce qui ne permet pas vraiment de faire le départ entre les deux hypothèses concurrentes concernant la représentation phonologique sous-jacente des voyelles nasales en FM.

Le fait que l'appendice consonantique nasal est d'une durée respectable en FM, d'ailleurs non significativement différente d'une consonne nasale pleine (dans les items CV.NV), plaide en faveur de H2, à savoir que les voyelles nasales du FM correspondent à VN en représentation profonde. Mais nous considérons comme au moins aussi significatif le résultat selon lequel la durée de C2 dans les items CṼ.CV est significativement plus courte que dans les items CV.CV. Selon nous, ce résultat donne un avantage – peut-être décisif – à H1. En effet, on peut inférer des patrons temporels des événements acoustiques (fig.3b) qu'en FM le geste nasal est principalement *retardé* par rapport à l'attaque de la voyelle, mais qu'alors il en vient à chevaucher le geste d'obstruence buccale qui, lui, reste en phase (voire, est anticipé) par rapport à la voyelle. La coproduction des gestes d'abaissement vélique et d'obstruence buccale aboutit ainsi : (i) à l'émergence dudit appendice consonantique nasal et (ii) à une réduction de la durée de C2.

Bien entendu, ces premiers résultats appellent à être confirmés sur un plus grand nombre de locuteurs encore, de même que dans d'autres sous-régiolectes du français méridional, ce qui constitue l'une des voies de nos futurs travaux.

Cette recherche a bénéficié des financements de la convention ARC AUWB- 08/12-UMH 17 de la Communauté française de Belgique et du projet MCF/FRC - FRFC 2.4644.09 du Fonds National de la Recherche Scientifique belge.

BIBLIOGRAPHIE

[1] J. Hajek. *Universals of sound change in nasalization*, Philological Society, 31, Blackwell Publishers, 1997.

[2] S. A.. Schane. *French Phonology and Morphology*. Cambridge Massachusetts, The M.I.T. Press, 1968.

[3] J. Durand. Les phénomènes de nasalité en français du midi: phonologie de la dépendance et sous spécification, *Nouvelles Phonologies, Recherches Linguistiques de Vincennes, 17*, Paris VII, pages 29-54, 1988.

[4] J. P. Watbled. Segmental and suprasegmental structure in Southern French, J.C. Smith et Maiden M. (eds) *Linguistic theory and the romance language 122*, Amsterdam, John Benjamins, pages 181-200, 1995.

[5] D. Demolin, B. Teston. Aspects aérodynamiques et articulatoires des voyelles nasales du français, *TIPA*, vol. 18, pages 50-59, 1998.

[6] A. Violin-Wigent. Southeastern French Nasal Vowels: Perceptual and Acoustic Elements, *La revue Canadienne de Linguistique* , 51(1), pages 15-43, 2006.

[7] S. Clairet. Une étude aérodynamqiue de la nasalité vocalique en français méridional, *27ièmes JEP Avignon*, pages 297-300, 2008.

[8] P. Boula de Mareüil, M. Adda-Decker & C. Woehrling. Analysis of oral and nasal vowel realisation in northern and southern French varieties, *16ièmes ICPHS* Saarbrucken, pages 2017-2020, 2007

[9] D. Blampain, A. Goosse, J.-M. Klinkenberg et M. Wilmet. *Le français en Belgique.*, Duculot, Louvain-La-Neuve, 1997.

[10] P. Boersma & D. Weenink. Praat, a system for doing phonetics by computer, *Glot International* 5(9/10), pages 341-345, 2001.

[11] P. S. Beddor. Nasals and nasalization: the relation between segmental and coarticulatory timing, *16ièmes ICPhS*, Saarbrucken, Allemagne, pages 249-254, 2007.

[12] S. Maeda. Acoustics of vowel nasalization and articulatory shifts in French nasal vowels, In *Huffman M.K. et Krakow R.A. (eds.)*, San Diego, CA: Academic Press, pages 147-167, 1993.

[13] J. Vaissière. Aspects aérodynamiques et acoustiques de la nasalité. Ecole thématique CNRS 'Dynamique de la nasalité', Porquerolles, 2008.

[14] G. Feng et E. Castelli. Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization, *JASA 99* (6), pages 3694-3706, 1996.

[15] V. Delvaux, V., D. Demolin, B Harmegnies, B., A. Soquet. The aerodynamics of nasalization in French. *Journal of Phonetics 36, 4,* pages 578-606, 2008.

Une surdité persistante au contraste /e/-/ɛ/ : le cas des méridionaux

Sophie Dufour[1], Noël Nguyen[1], Ulrich Hans Frauenfelder[2]

[1] Laboratoire Parole et Langage, CNRS et Université d'Aix-Marseille, Aix-en-Provence, France
5, Avenue Pasteur, 13604 Aix-en-Provence
[2] Laboratoire de Psycholinguistique Expérimentale FPSE, Université de Genève
40, Bd du Pont d'Arve, CH 1205 Genève, Suisse.
sophie.dufour@lpl-aix.fr, noël.nguyen@lpl-aix.fr, Ulrich.Frauenfelder@unige.ch

ABSTRACT

In this study, we trained Southern French speakers to exploit the /e/-/ɛ/ contrast by asking them to associate new minimal pairs of "words" with visual shapes. Although, the performance in the training session was relatively high, the learning did not transfer to task known to reflect lexical access. This research underlines the difficulty for adult listeners to acquire new phonemic categories and highlights the importance to use a wide variety of task including on-line word recognition to assess training on a specific contrast.
Keywords: Speech perception, regional variety, phonological deafness.

1. INTRODUCTION

Nous savons depuis plusieurs décennies que les auditeurs adultes ont des difficultés dans la discrimination de contrastes étrangers. L'exemple le plus typique est celui des Japonais qui ont des difficultés à discriminer entre le /r/ et le /l/. Outre une difficulté pour l'auditeur à discriminer des contrastes non natifs, certaines études montrent également une difficulté dans la discrimination de contrastes natifs appartenant à une variété régionale autre que celle de l'auditeur [1, 2]. Dans une étude récente [2], nous avons examiné la manière dont des auditeurs méridionaux traitaient le contraste /e/-/ɛ/ en position finale de mot, qui se rencontre en français standard mais pas en français méridional. Cette étude nous a permis de montrer que les auditeurs méridionaux traitent les mots /epe/ (*épée* en français standard) et /epɛ/ (*épais* en français standard) comme étant homophones. Cela nous indique que les mots /epe/ et /epɛ/ sont associés à la même représentation phonologique /epe/ chez les auditeurs du Sud de la France et qu'à une étape précoce de traitement, les phonèmes /e/ et /ɛ/ sont assimilés comme étant le même phonème /e/.

Bien que les difficultés de discrimination s'avèrent robustes puisqu'elles sont identifiables chez des bilingues [3], des améliorations dans la perception de contrastes étrangers ont pu être montrées en laboratoire au travers de procédures d'apprentissage contrôlées. Même si la performance restait en dessous de celle des anglais natifs, Bradlow et ses collaborateurs [4] ont par exemple montré que les performances d'identification du /l/ et du /r/ par des auditeurs japonais augmentaient de façon significative après un apprentissage intensif faisant intervenir des stimuli enregistrés par différents locuteurs. L'apprentissage s'est généralisé à des nouveaux stimuli produits par des locuteurs différents de ceux utilisés dans la phase d'entraînement et se maintenait trois mois après la session d'entraînement [5]. Suite à ces travaux, nous avons dans cette recherche mis en place une procédure d'apprentissage dans laquelle des participants méridionaux devaient apprendre des « nouveaux-mots » se finissant par /e/ ou /ɛ/ et formant des paires minimales. L'intérêt d'une telle procédure est quelle nous permettait de focaliser l'attention des participants méridionaux sur le contraste /e/-/ɛ/ et de vérifier si oui ou non, ils étaient capables de l'exploiter. A notre connaissance, les études portant sur l'apprentissage de contrastes phonémiques ont examiné le transfert de l'apprentissage à des nouveaux jeux de stimuli à l'aide d'une tâche qui était exactement la même que celle utilisée dans l'entraînement et qui consistait à comparer les deux membres de paires minimales. Cependant, aucune étude n'a examiné le transfert de l'apprentissage à des tâches connues comme reflétant des processus en temps réel et où l'attention des participants est désengagée du contraste critique en présentant séparément les membres des paires minimales. C'est précisément le but de cette étude.

L'expérience s'est déroulée en trois phases : une phase de pré-test, une phase d'entraînement et une phase de post-test. Durant le pré-test, le paradigme d'amorçage de répétition combiné avec une tâche de décision lexicale dans laquelle les participants devaient juger si les stimuli présentés constituaient ou non des mots de la langue française a été utilisé. Cela nous permettait de répliquer l'effet d'amorçage de répétition observé sur des paires minimales de type /epe/ - /epɛ/ [2] et de remettre en évidence la surdité phonologique pour le contraste /e/-/ɛ/ chez des auditeurs méridionaux. Rappelons que l'effet d'amorçage de répétition réfère au fait qu'un mot est reconnu plus rapidement lorsqu'il

est rencontré pour la seconde fois. Ainsi les auditeurs du Sud de la France montrent une diminution des temps de réponse à la fois sur /epe/ et /epɛ/ quand /epe/ est présenté en première occurrence [2]. Durant la seconde phase, les participants devaient apprendre des nouvelles paires minimales basées sur le contraste /e/-/ɛ/ en les associant à des figures visuelles (Voir [6] pour la même procédure). Durant le post-test, la même procédure que celle du pré-test a été utilisée, ce qui nous permettait de tester les changements dans la perception du contraste /e/-/ɛ/ liés à l'apprentissage. Dans le but d'évaluer la persistance de l'apprentissage, le post-test a été administré à trois reprises, immédiatement après l'apprentissage, un jour après, et une semaine après. Notre hypothèse était que si après l'entraînement, les participants méridionaux exploitent automatiquement la catégorie phonémique /ɛ/ en position finale de mots, il ne devrait plus y avoir un appariement exact entre les informations extraites du signal de parole et les représentations phonologiques des mots stockés en mémoire, engendrant ainsi une diminution de l'effet d'amorçage de répétition sur les paires minimales. Notons que l'utilisation d'une tâche différente entre l'entraînement et le pré et post test nous permettait d'évaluer la généralisation de l'apprentissage à une autre tâche et garantissait ainsi qu'une éventuelle amélioration entre le pré et le post test ne soit pas simplement due à une simple habituation à la tâche.

2. EXPÉRIENCE

2.1. Méthode

2.1.1. Participants

24 volontaires de l'Université de Provence, tous originaires du Sud de la France ont participé à l'expérience.

2.1.2. Matériel

Les stimuli utilisés dans les phases de pré- et de post-test ont été repris de notre première étude [2]. Ils étaient constitués de 32 paires minimales basées sur le contraste /e/-/ɛ/. Les 32 paires ont été divisées en 2 groupes, l'un des groupes servant pour le pré-test et l'autre pour le post-test. Les 2 groupes ont été appariés le plus possible sur les variables fréquence, nombre de phonèmes, point d'unicité phonologique et durée connues pour affecter le temps de reconnaissance des mots. Pour les besoins de la tâche de décision lexicale, 64 non-mots bisyllabiques également divisés en 2 groupes et formant 32 minimales paires basées sur le contraste /e/-/ɛ/ ont été utilisés. A l'intérieur de chaque groupe 4 listes contrebalancées ont été créées de sorte à ce que chaque membre d'une paire minimale soit répété ou suivi de l'autre membre. Pour finir 66 mots et 66 non-mots additionnels ont servi de remplisseurs dans chacune des listes. Les listes ont été construites de sortes à ce que de 8 à 11 items soient présentés entre la répétition du mot ou l'autre membre de la paire minimale. L'ordre de présentation de chacun des groupes de mots a été contrebalancé au travers les participants de sorte à ce que les 2 groupes de paires minimales soient vus en pré- et en post-test.

Pour la phase d'entraînement, 12 non-mots bisyllabiques formant 6 paires minimales /e/-/ɛ/ ont été utilisés. 12 formes visuelles ne renvoyant à aucun objet réel ont été également construites et assignées de façon aléatoire à chacun des non-mots. Des exemples de formes visuelles et les non-mots qui leur sont associés sont fournis dans la figure 1.

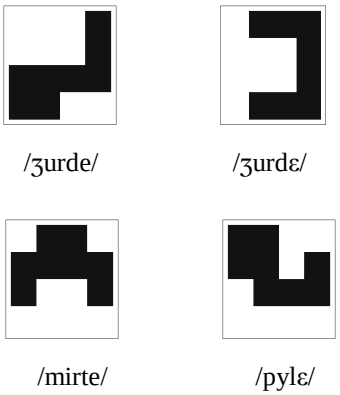


Figure 1: Exemples de formes visuelles avec leur nom correspondant.

2.1.3. Procédure

Les stimuli ont été enregistrés par une locutrice de langue maternelle française prononçant distinctement les voyelles /e/ et /ɛ/. Durant le pré- et le post- test, les participants devaient pour chaque stimulus indiquer le plus rapidement et le plus précisément possible si il constituait un mot ou non de la langue française, et devait fournir la réponse mot avec leur main dominante. Les temps de réponse (TRs) étaient enregistrés à partir du début des stimuli. Pour chaque stimulus, La réponse du participant et le début de présentation du stimulus suivant étaient séparés par un délai de deux secondes. Si les participants échouaient à répondre à l'intérieur des 1800 ms, aucune réponse n'était enregistrée et le stimulus suivant était présenté. Les participants ont été testés sur une seule liste expérimentale et ont commencé l'expérience avec 6 essais d'entraînement.

La session d'entraînement consistait en 5 blocs avec feed-back dans lesquels une information concernant la réponse correcte était donnée aux participants, et en 1 bloc sans feed-back nous permettant d'évaluer l'apprentissage. La structure de l'entraînement était la suivante : Un point de fixation apparaissait dans un premier temps à l'écran pendant 1000 millisecondes. Puis 4 formes étaient présentées à l'écran (figure 1) et

les participants entendaient l'un des 12 non-mots. Ils avaient pour tâche d'aller cliquer sur la forme qu'ils pensaient correspondre au non-mot. Durant les 5 premiers blocs, les 3 formes distractrices disparaissaient à la réponse des sujets et seul le référent correct restait à l'écran, et le nom de la forme était répété. Durant le dernier bloc, les 4 formes disparaissaient à la réponse des participants et le point de fixation annonçant l'essai suivant était présenté.

Chaque bloc d'entraînement était constitué de 60 essais. A l'intérieur de chaque bloc, chacun des 12 non-mots était présenté 5 fois. Pour chaque essai, l'une des 3 formes distractrices était la forme associée au non-mot correspondant à l'autre membre de la paire minimale et les deux autres étaient sélectionnées de façon aléatoire parmi les 10 restantes de sorte à ce que chaque forme apparaisse le même nombre de fois par bloc. La position des formes à l'écran était aléatoire.

2.2. Résultats et Discussion

Les pourcentage de réponses correctes obtenus dans les 6 blocs d'apprentissage sont présentés dans le Tableau 1. Comme nous pouvons le remarquer, le pourcentage de réponses correctes à la fin de l'entrainement (bloc sans feed-back) atteignait les 80% pour les non-mots se terminant par la voyelle /e/ et 84% pour les non-mots se terminant par la voyelle /ɛ/. La performance est relativement haute et montre une certaine capacité de la part des auditeurs du sud de la France à exploiter le contraste /e/ - /ɛ/, au moins dans le but d'associer les membres d'une paire minimale à la forme visuelle qui leur correspond.

Table 1 *:* Pourcentage de réponses correctes dans la session d'entraînement.

Bloc	Total	/e/ final	/ɛ/ final
1	38	33	43
2	55	50	60
3	68	63	73
4	75	72	78
5	80	77	83
sans feedback	82	80	84

Quel a été l'impact de cet entrainement sur le processus de reconnaissance des mots ? Comme les résultats au niveau des 3 post-tests sont similaires, nous présenterons uniquement les résultats obtenus dans le post-test immédiat ceci afin de ne pas alourdir le manuscrit. Les temps de réaction moyens obtenus en tâche de décision lexicale dans le pré-et le post-test immédiats sont respectivement présentés dans les Figures 1 et 2. Des analyses de variance (ANOVAs) par sujets (F_1) et par items (F_2) ont été conduites avec le type de paire (même, minimale) et l'occurrence ($1^{ère}$, $2^{ième}$) comme variables. Très peu d'erreurs ayant été faites, les anlayses ont été effectuées uniquement sur les temps de réaction.

Pré-test: L'effet simple du type de paire n'était pas significatif [*F*s < 1]. L'effet simple de l'occurrence était significatif [$F1(1,23)=35.21$, $p<.0001$; $F2(1,31)=44.06$, $p<.0001$]. Les temps moyens de réaction étaient plus rapides lorsque les mots cibles étaient rencontrés pour la seconde fois. L'interaction entre le type de paire et l'occurrence n'était pas significative [*F*s < 1], indiquant que l'amplitude de l'effet d'amorçage restait inchangé quel que soit le type de paire. Un tel résultat réplique nos premières observations et indique que les auditeurs du Sud de la France traitent le second membre de paires minimales basées sur le contraste /e/ - /ɛ/ comme étant une répétition du premier.

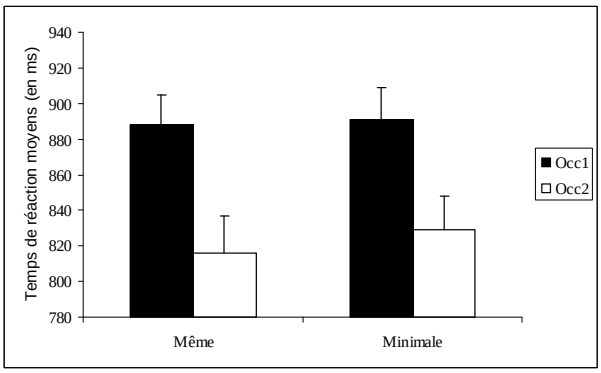


Figure 1 : Temps de réaction moyens (en ms) obtenus pour la première et la seconde occurrence en fonction du type de paire dans le pré-test.

Post-test immédiat : L'effet simple du type de paire n'était pas significatif [*F*s < 1]. L'effet simple de l'occurrence était significatif [$F1(1,23)=45.67$, $p<.0001$; $F2(1,31)=37.28$, $p<.0001$]. Les temps moyens de réaction étaient plus rapides lorsque les mots cibles étaient rencontrés pour la seconde fois. L'interaction entre le type de paire et l'occurrence n'était pas significative [*F*s < 1]. De façon cruciale, une analyse combinée des résultats obtenus dans le pré et le post-test immédiat ne montrait aucune interaction entre la session (pré, post) et l'occurrence ($1^{ère}$, $2^{ième}$) pour les paires minimales [*F*s < 1]. Il apparaît ainsi que l'entrainement n'a eu aucun aucun impact sur la taille de l'effet d'amorçage obtenu pour les paires minimales. Les résultats obetnus dans les post-tests à un jour et à une semaine montrent exactement les mêmes effets.

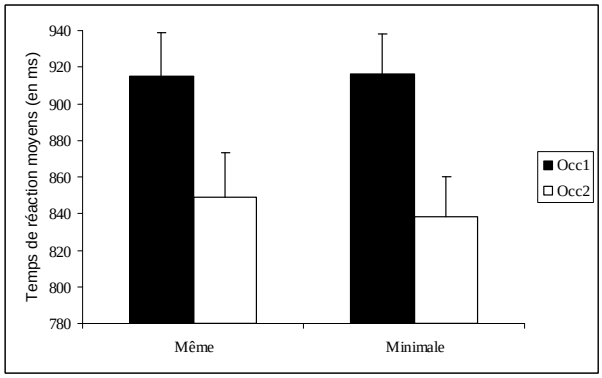


Figure 2 : Temps de réaction moyens (en ms) obtenus pour la première et la seconde occurrence en fonction du type de paire dans le post-test immédiat.

3. Conclusion

En accord avec d'autres études [4, 5], nous avons montré qu'un apprentissage à exploiter un contraste particulier est possible. Cependant, il apparaît au vu de cette étude que l'apprentissage a un impact limité puisqu'il ne s'étend pas à la reconnaissance de mots contenant le contraste critique. Les auditeurs méridionaux sont en effet capables de discriminer le /e/ du /ɛ/ à la condition de focaliser leur attention sur ce contraste. Cependant, ils n'exploitent pas le contraste, ni leurs connaissances quant à ce contraste lorsqu'ils sont mis dans des situations de reconnaissance automatique de mots puisque lorsque leur attention est désengagée des voyelles /e/ et /ɛ/, les auditeurs méridionaux traitent toujours les mots /epe/ et /epɛ/ comme étant des homophones. Sur le plan méthodologique, nos résultats sont importants puisqu'ils montrent que les tâches de discrimination généralement utilisées pour évaluer les changements dans la perception de contrastes peuvent sous-estimer les difficultés de traitement. Comme l'ont déjà souligné Dupoux et ses collaborateurs [7], il semble important de tester les capacités des participants avec un contraste particulier au travers une variété de tâches expérimentales allant de la simple discrimination à des tâches comme la décision lexicale connues pour refléter des processus en temps réels tels que ceux impliqués dans l'accès lexical.

Dans le but d'inciter des auditeurs méridionaux à exploiter le contraste /e/-/ɛ/ en position finale de mots, nous avons utilisé un paradigme d'apprentissage de « mots-nouveaux ». Les résultats montrent que seulement 5 blocs d'apprentissage dans lesquels chaque mot nouveau était présenté seulement 5 fois sont suffisants à une bonne exploitation du contraste en fin d'apprentissage. Bien qu'il nous est impossible de conclure quant à l'établissement d'une entrée lexicale associée à chacun des mots nouveaux, une telle observation semble indiquer en accord avec Gaskell et Dumay [8], une acquisition rapide de l'information phonologique. Plus d'études sont néanmoins nécessaires de façon à examiner la manière dont les mots-nouveaux ont été représentés dans le lexique mental et leurs effets sur le traitement des mots existants.

BIBLIOGRAPHIE

[1] B. Conrey, G. Potts and N. Niedzielski. Effects of dialect on merger perception: ERP and behavioral correlates. *Brain and Language,* 95: 435-449, 2005.

[2] S. Dufour, N. Nguyen and U.H. Frauenfelder. The perception of phonemic contrasts in a non-native dialect. *Journal of Acoustical Society of America*, 121: EL131-EL136, 2007.

[3] C. Pallier, A. Colomé, A. and N. Sebastián-Gallés. The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, 12: 445–449, 2001.

[4] A.R. Bradlow, D.B. Pisoni, R. Akahane-Yamada and Y. Tohkura. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101: 2299-2310, 1997.

[5] A. R. Bradlow, R. Akahane-Yamada, D. B. Pisoni and Y. Tohkura, Y. Training Japanese listeners to identify English /r/ and / l/: Long-term retention of learning in speech perception and production. *Perception & Psychophysics*, 61: 977-985, 1999.

[6] J. S. Magnuson, M. K. Tanenhaus, R. N. Aslin, and D. Dahan. The time course of spoken word recognition and learning: Studies with artificial lexicons. *Journal of Experimental Psychology: General,* 132: 202-227, 2003.

[7] E. Dupoux, N. Sebastian-Galles, E. Navarete and S. Peperkamp. Persistent stress `deafness': the case of French learners of Spanish. *Cognition*, 106: 682-706, 2008.

[8] M.G., Gaskell and N.Dumay. Lexical competition and the acquisition of novel words. *Cognition*, 89: 105-132, 2003.

Intonation des questions totales en français langue étrangère : suffit-il d'enseigner et apprendre la montée finale ?

Takeki Kamiyama et Megumi Sakamoto***

*Laboratoire de Phonétique et Phonologie (UMR7018) CNRS / Sorbonne-Nouvelle Paris 3
19 rue des Bernardins, 75 005 Paris, France
**Université Sophia, Tokyo
takeki.kamiyama@univ-paris3.fr
http://lpp.univ-paris3.fr/equipe/takeki_kamiyama.htm

ABSTRACT

It is generally taught that yes-no questions in French are formed with a rising intonation at the end. A comparison of 5 short yes-no questions read by 8 Japanese learners and 3 French speakers shows that the learners tended: 1) to realise a smaller final rise than the native speakers, 2) not to realise the deletion or suppression of F0 declination tendency often observed with native speakers. A perception test was conducted with 25 native French listeners using Mbrola (to imitate the prosody produced by the Japanese learners). The results show that the subjects judged only 50% of the sentences with Japanese learners' duration and F0 patterns as questions. These findings suggest that the deletion of F0 declination plays an important role, and that it is worth being taught explicitly to learners.

Keywords: intonation, total question, French as a foreign language, Japanese-speaking learners, F0 declination

1. INTRODUCTION

De nombreux manuels de français langue étrangère décrivent la question totale sans inversion de sujet-verbe ni « *est-ce que* » comme étant accompagnée d'une montée finale de la mélodie. Un survol de quelques ouvrages publiés au Japon révèle cette tendance : certains montrent une flèche montante (Hisatomi [7], Kokubu [10], Seto & Seto [14]), tandis qu'ils semblent moins nombreux à faire figurer un plateau suivi d'une montée finale (Kurakata [11]), ce qui correspond à la suppression ou diminution de la ligne de déclinaison observée dans de nombreuses langues (Vaissière [17]), dont le français (Vaissière [18]). C'est aussi le cas pour les manuels de prononciation publiés en France (Charliac & Morton [4], entre autres), mais il existe ceux qui montrent des courbes mélodiques schématisées représentant un plateau (Abry & Chalaron [2], Léon [12]).

En japonais, une montée finale est observée dans les questions totales et partielles. S'il y a un mot accentué lexicalement dans l'énoncé, la montée finale est précédée par une descente locale de la fréquence fondamentale (F0) due à l'accent lexical. (Abe [1], entre autres).

Concernant le mouvement global de F0, les questions totales seraient similaires aux déclaratives sauf sur la dernière syllabe. Shôchi *et al.* [15] ont montré dans une expérience de *gating* que des auditeurs japonais étaient incapables de distinguer les questions des déclaratives à moins d'entendre la dernière more de la phrase.

Quant à la continuation, une montée peut être observée, mais cette tendance est limitée en général à certains locuteurs et certains styles de parole (jeunes femmes dans un style familier, en particulier) ; voir Inoue [9] pour ce que l'on appelle « *shiriagari intonêshon* » (intonation avec une « montée à la queue »), qui se réalise comme un contour montant-descendant.

En ce qui concerne les questions totales en français, les observations relevées dans la littérature concordent sur l'importance de la montée finale. En revanche, on trouve des descriptions différentes concernant la partie précédente de la phrase : 1) présence d'un plateau, c'est-à-dire suppression ou diminution de la tendance à décliner (Delattre [5], Léon et Léon [13], Vaissière [18]) ; 2) contour similaire à celui des déclaratives (avec ou sans montée de continuation), sauf la montée finale (Di Cristo [6]). Dans ce dernier cas, la seule différence majeure qui se trouve entre les continuatives et les questions totales serait celle de l'ampleur de la montée finale.

Qu'est-ce que cela signifie concernant les apprenants japonophones ? Si la montée finale est traitée davantage que la suppression de la ligne de déclinaison dans les manuels japonais de français langue étrangère, cela signifie-t-il que les apprenants japonophones suppriment ou diminuent la ligne de déclinaison avec une facilité relative et qu'il est superflu de l'enseigner explicitement (question de recherche 1) ? Les locuteurs francophones suppriment-ils ou diminuent-ils véritablement la ligne de déclinaison (question 2) ? Perçoivent-ils des questions totales prononcées par des apprenants japonophones comme des questions (question 3) ?

Trois expériences ont été réalisée afin de répondre à ces 3 questions. Dans l'expérience 1, 5 courtes questions totales lues par 8 apprenants japonophones ont été analysées. Dans l'expérience 2, les mêmes phrases lues par 3 locutrices natives du français ont été étudiées. Dans l'expérience 3, des phrases synthétisées avec la durée segmentale et la F0 d'apprenants et de locutrices natives

ont été soumises à une expérience de perception (identification de modalité : question ou continuative) auprès de 25 auditeurs natifs du français.

2. EXPÉRIENCE 1 (PRODUCTION D'APPRENANTS JAPONOPHONES)

L'objectif de cette expérience était d'observer les contours mélodiques des questions totales produites par des apprenants japonophones.

2.1. Méthode

Cinq questions totales (Table 1) ont été lues 3 fois par 8 japonophones natifs (4 femmes et 4 hommes) apprenant le français langue étrangère au Japon. Les apprenants étaient des étudiants de première année en études françaises à l'Université Sophia (Tokyo, Japon). Ils avaient suivi 100 à 120 heures de cours de français langue étrangère.

Table 1 : Phrases lues par 8 apprenants japonophones.

Il aime les rats ?	Il aime la villa ?
Il aime le cinéma ?	Il aime le panorama ?
Il aime le dessin animé ?	

L'enregistrement a été réalisé dans le studio d'enregistrement du laboratoire de phonétique de l'Université Sophia. Les données ont été enregistrées à la fréquence d'échantillonnage de 22.050 Hz (16 bits) et analysées sous Praat (Boersma et Weenink [3]).

2.2. Mesures

Les valeurs relatives suivantes de F0 ont été mesurées : 1) différence entre F0 minimum et maximum au cours de la dernière syllabe de la phrase, divisée par la moyenne du reste de la phrase (« fin ») ; 2) différence entre F0 minimum et maximum au cours de la phrase jusqu'à l'avant-dernière syllabe, divisée par la moyenne de la même partie de la phrase (« pre »). Les valeurs instables à l'attaque vocalique au début de phrase ont été exclues.

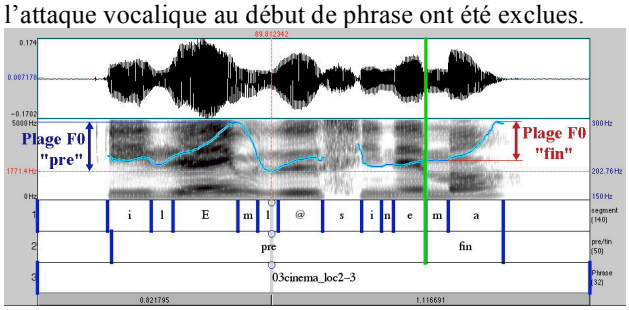


Figure 1 : Mesure de F0 : différence entre les valeurs minimum et maximum au cours de la phrase jusqu'à l'avant-dernière syllabe (« pre ») et au cours de la dernière syllabe (« fin »), divisée respectivement par la moyenne de la phrase jusqu'à l'avant-dernière syllabe. « *Il aime le cinéma ?* » prononcée par l'apprenante jp2.

2.3. Résultats

La Figure 2 représente les résultats. Trois apprenants (jp5, 6, 7) ont présenté une différence de F0 plus grande que les autres locuteurs pour la dernière syllabe (« fin » : 70% environ). Trois locuteurs (jp4, 5, 6) ont réalisé une différence de F0 plus petite que les autres pour le reste de la phrase (« pre »), ce qui suggère qu'ils ont produit un plateau. En revanche, 4 autres apprenants (jp1, 2, 3, 7 : la Figure 1 illustre la production de jp2) ont produit des valeurs relativement grandes pour « pre ». Cela suggère qu'ils n'ont pas supprimé ou diminué la ligne de déclinaison, dont la suppression, selon Vaissière [18], caractérise les questions totales et permet de les distinguer des continuatives. Par conséquent, les questions totales prononcées par ces apprenants pourraient être perçues comme continuatives par les auditeurs natifs du français.

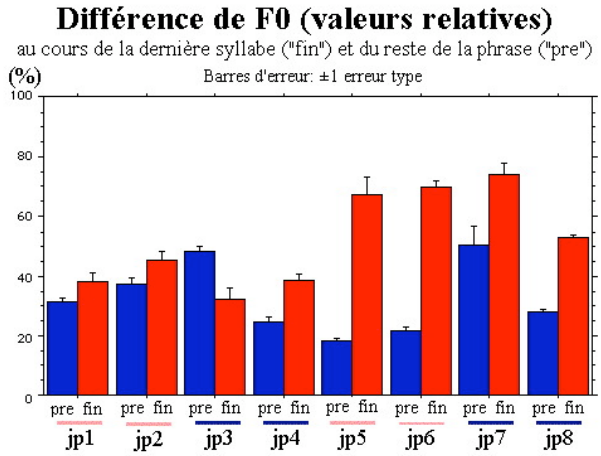


Figure 2 : Différence de F0 (valeurs relatives) au cours de la dernière syllabe (« fin ») et du reste de la phrase (« pre » : voir section 2.2., Figure 1). Valeurs moyennes de 3 répétitions et de 5 phrases pour chaque apprenant. Jp1, 2, 5, 6 sont des femmes, les autres sont des hommes.

3. EXPÉRIENCE 2 (PRODUCTION DE LOCUTRICES FRANCOPHONES NATIVES)

Une autre expérience de production a été effectuée afin de vérifier la réalisation des questions totales et des continuatives chez les locuteurs francophones natifs et de la comparer avec les questions totales produites par les apprenants japonophones dans l'expérience précédente.

3.1. Méthode

Quatre des 5 phrases dans la Table 1 (sauf « … *le dessin animé* »), ainsi que les continuatives correspondantes, ont été lues 3 fois par 3 locutrices francophones natives. Les continuatives ont été mises dans des phrases plus longues (ex. « *Il aime les rats, et elle aime les chats.* »). Les questions totales ont été lues avec et sans pause entre le verbe « *aime* » et le complément d'objet direct. Les phrases ont été présentées orthographiquement, avec une barre oblique (/) insérée à la position où une pause était demandée. Les mêmes mesures de F0 que dans l'expérience précédente ont été effectuées.

3.2. Résultats

Les 3 locutrices ont montré des tendances différentes pour les questions totales : la montée finale était précédée respectivement par 1) une légère montée, 2) un plateau, et 3) une montée et une descente comme dans les continuatives (Figure 3). Les deux premières tendances correspondent aux descriptions de Delattre [5], Léon et Léon [13], et Vaissière [18] ; la dernière correspond à celle de Di Cristo [6]. La mesure de différence de F0 (Figure 4) semble refléter ces différences. Les locutrices natives ont réalisé une montée plus ample que la plupart des apprenants (sauf jp 5, 6, 7 : Figure 2).

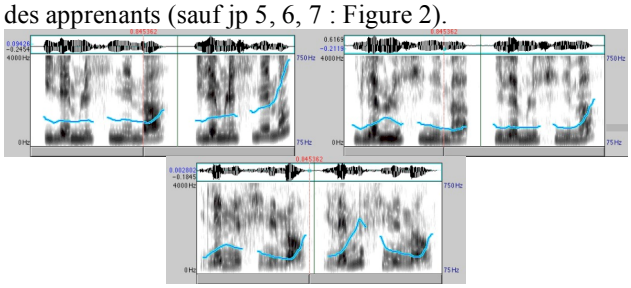


Figure 3 : Courbes de F0 de continuative (à gauche de chaque paire) et de question totale (à droite). « *Il aime le cinéma* » prononcé par 3 locutrices francophones natives (fr1 en haut à gauche, fr2 en haut à droite, fr3 en bas).

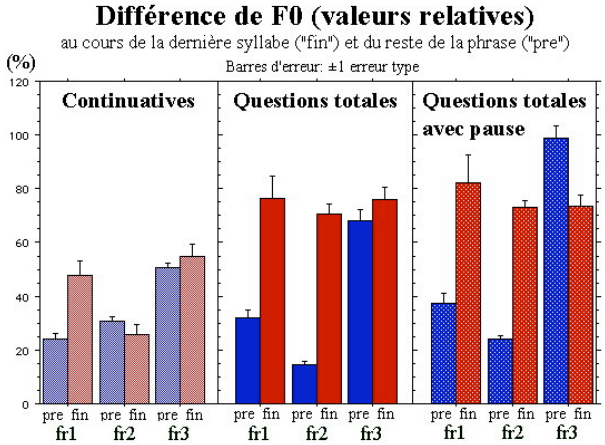


Figure 4 : Différence de F0 (valeurs relatives) au cours de la dernière syllabe (« fin ») et du reste de la phrase (« pre »). Moyennes de 3 répétitions et de 4 phrases pour chacune des 3 locutrices francophones natives (fr1, 2, 3).

4. EXPÉRIENCE 3 (PERCEPTION D'AUDITEURS FRANCOPHONES NATIFS)

En utilisant les données de production obtenues dans les deux études précédentes, une expérience de perception a été réalisée afin d'examiner si la prosodie (F0 et durée segmentale) des questions totales prononcées par des apprenants japonophones est perçue comme question ou continuative par des auditeurs francophones natifs.

4.1. Méthode

Les stimuli ont été synthétisés en utilisant Mbrola (Dutoit *et al.* [7]). Les séries de diphones Mbrola utilisées sont « FR2 » (homme francophone) et « FR3 » (femme francophone). Deux des 3 répétitions des 4 phrases (sauf « … *dessin animé* ») lues par les 8 apprenants japonophones, et une des 3 répétitions des phrases (4 phrases x 3 types) lues par les locutrices natives fr1 et fr2 dans les expériences précédentes ont été retenues (rappelons que ces locutrices ont supprimé la ligne de déclinaison). La F0 et la durée segmentale de ces phrases ont été calculées. Ces valeurs ont été mises dans des fichiers Mbrola (la pause insérée par la locutrice native au milieu des questions totales a été supprimée). La durée totale de phrase a été ajustée pour que la même phrase ait une durée approximativement identique.

Les stimuli synthétisés ainsi ont été présentés aux auditeurs dans un ordre semi-aléatoire à travers des haut-parleurs. Les tâches des auditeurs consistaient à écouter le stimulus, à indiquer s'ils avaient entendu une question (comme si le locuteur demandait une réponse) ou une continuation (comme si le locuteur comptait garder son tour de parole), et à fournir un degré de certitude de leur choix (1 : pas du tout sûr – 5 : tout à fait sûr). Vingt-cinq auditeurs francophones natifs inscrits à l'Université Lille 3 ont participé à l'expérience.

4.2. Résultats

La Figure 5 montre les résultats : pourcentage de réponses pour « question » (gauche), score (+125 : tous les auditeurs sont tout à fait sûrs que c'est une question ; -125 : ils sont tout à fait sûrs que c'est une continuation) calculé en multipliant le nombre de réponses d'identification par le degré de certitude (droite). Premièrement, les stimuli avec F0 et durée des questions totales (avec et sans pause) des locutrices francophones natives (fr1, 2) ont été identifiés comme questions dans 85% des cas, et ceux des continuatives comme telles dans 80% des cas (c'est-à-dire, 20% de réponses pour « question »). Deuxièmement, une comparaison de ces données des locutrices natives avec celles des stimuli de chaque apprenant permet de classer les apprenants dans 3 catégories : ceux qui ont produit des valeurs de F0 et de durée 1) proches de celles des continuatives des locutrices francophones (jp1) ; 2) proches de celles des questions totales des francophones (jp4, 5, 6) ; 3) entre les deux catégories (jp2, 3, 7, 8). On ne trouve pas de différence significative de pourcentage de réponses entre les continuatives des francophones et jp1, ni entre les questions totales des francophones et jp4, 5, 6, alors que jp2, 3, 7, 8 sont significativement différents des continuatives et des questions totales des francophones (PLSD de Fisher, seuil : 5%. ANOVA: $F_{(10, 77)} = 10{,}87$, $p < 0{,}0001$). Une tendance similaire est observée dans les données de score fondé sur l'identification (question ou continuation) et le degré de certitude.

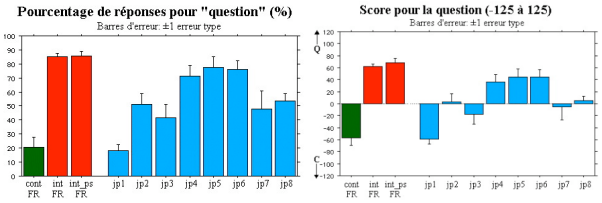


Figure 5 : Pourcentage de réponses « question » (gauche) et score calculé sur le nombre de réponses et le degré de certitude (droite). Chaque colonne représente les valeurs de 8 stimuli x 25 réponses (25 auditeurs x 1 réponse).

Ces résultats pourraient être attribués aux tendances observées dans l'Expérience 1 : le pattern de F0 et de durée des apprenantes jp5 et 6, qui ont présenté une petite différence de F0 pour « pre » et une grande différence pour « fin », correspond à des pourcentages élevés de perception de la question.

5. DISCUSSION ET CONCLUSION

La présente série d'expériences montre les résultats suivants : 1) dans les questions totales, certains apprenants japonophones réalisent une montée finale moins ample que les locuteurs natifs, et d'autres ne suppriment ni diminuent la déclinaison de F0 ; 2) certains locuteurs natifs produisent des questions totales avec un plateau, d'autres (une locutrice sur 3 dans l'Expérience 2) avec une déclinaison de F0 ; 3) les patterns de F0 et de durée des questions totales produites par les apprenants japonophones peuvent être perçus comme question, continuative, ou entre les deux, en fonction de la production de l'apprenant.

Cependant, il reste encore des phénomènes à explorer.

Concernant la perception des questions totales chez les auditeurs francophones, le rôle respectif de la montée finale et de la suppression ou diminution de la ligne de déclinaison reste encore à examiner dans une étude avec des patterns de F0 contrôlés plus précisément.

Umeda [16] a montré que la déclinaison de F0 n'est pas obligatoire dans les phrases déclaratives en anglais. Nous pouvons supposer que la suppression ou diminution de déclinaison n'est pas obligatoire non plus dans les questions totales en français. Il sera intéressant de réaliser des études sur des locuteurs et des auditeurs de divers groupes (sur le plan régional, social, *etc*.) et sur différents styles de parole (lecture, parole spontanée, *etc*.).

Nous avons traité, dans cette étude, de courtes questions totales et continuatives en contexte isolé *in vitro*, et comme s'il s'agissait de deux catégories binaires. Dans la communication *in vivo*, les phrases sont produites en contexte, ce qui peut influencer l'interprétation de la phrase, comme le signale Di Cristo [6]. Différents types de connotations impliquées dans les questions devraient être également considérés.

Malgré ces limitations, les résultats de la présente étude suggèrent l'importance de la suppression ou diminution de la ligne de déclinaison dans les questions totales en français. Il serait utile d'intégrer cet aspect prosodique dans la pratique de l'enseignement du français.

REMERCIEMENTS

Les auteurs remercient Jacqueline Vaissière et Antonia Colazo-Simon ainsi que les deux relecteurs anonymes pour leurs commentaires et suggestions sur les versions antérieures de cet article.

BIBLIOGRAPHIE

[1] I. Abe. Intonation in Japanese. In D. Hirst, A. Di Cristo, *Intonation systems*. Cambridge University Press, Cambridge, UK, 360-375, 1998.

[2] D. Abry and M.L. Chalaron. *Phonétique: 350 exercices*. Hachette, Paris, 1994.

[3] P. Boersma and D. Weenink. *Praat: doing phonetics by computer* (logiciel).

[4] L. Charliac and A.C. Motron. *Phonétique progressive du français*. CLE intl., Paris, 1998.

[5] P. Delattre. Les Dix Intonations de base du français. *The French Review* 40(1): 1-14, 1966.

[6] A. Di Cristo. Intonation in French. In D. Hirst and A. Di Cristo, *Intonation systems*. Cambridge University Press, Cambridge, UK, 195-218, 1998.

[7] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken. The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proc. Intl. Conf. on Spoken Language Processing-96*, 1393-1396, 1996.

[8] K. Hisatomi. *Tableau Grammaire*. Éditions Asahi, Tokyo, 2004.

[9] F. Inoue. Intonêshon no shakaisei [Aspect social de l'intonation]. In M. Sugitô, T. Kunihiro, H. Hirose, and M. Kôno, *Nihongo onsei 2: akusento, intonêshon, rizumu to pôzu*. Sanseidô, Tokyo, 143-168, 1997.

[10] T. Kokubu. *Manuel pratique de grammaire française*. Daisan shobô, Tokyo, 2006.

[11] H. Kurakata. *C'est la vie*. Sôbi, Tokyo, 2004.

[12] M. Léon. *Exercices systématiques de prononciation française*. Hachette, Paris, 2003.

[13] M. Léon and P. Léon. *La Prononciation du français*. Nathan Université, Paris, 1997.

[14] S. Seto and K. Seto. *Bonjour français*. Éditions Asahi, Tokyo, 2004.

[15] T. Shôchi, V. Aubergé, and A. Rilliard. Nihongo bogo-washa wa hatsuwa no ichibu o kikudakede taido o chikaku dekirunoka? – Gating paradaimu niyoru jikken o tôshite [Peut-on percevoir les attitudes japonaises avant la fin de phrase ? – test perceptif selon le paradigme de "*gating*"]. *Furansu Nihongo Kyouiku*, 4: 65-75, 2009.

[16] N. Umeda. "F0 declination" is situation dependent. *Journal of Phonetics*, 10: 279-290, 1982.

[17] J. Vaissière. Language-independent prosodic features. In A. Cutler, R. Ladd, *Prosody: models and measurements*. Springer, New York, 53-65, 1983.

[18] J. Vaissière. *La Phonétique*. PUF, Paris, 2006.

Index des auteurs

XXVIIIèmes Journées d'Etude sur la Parole

JEP 2010

Laboratoire des Sciences de la Parole de l'Académie Universitaire Wallonie-Bruxelles :
Laboratoire de phonétique (UMONS)
Laboratoire de phonologie expérimentale (ULB)
Laboratoire TCTS (UMONS)

Association Francophone de la Communication Parlée

L'AFCP est un 'special interest group' de l'International Speech Communication Association