XXXe édition des Journées d'Etudes sur la Parole

Le Mans, 23 → 27 juin

2014

Actes de la conférence

lium
Laboratoire d'Informatique de l'Université du Maine

AFCP
Association Francophone de la Communication Parlée

lina
Laboratoire d'Informatique de Nantes Atlantique

avec le soutien de :

Le Mans Acoustique · atlanSTIC · Université du Maine · Inria informatiques mathématiques · Le Mans métropole communauté urbaine · IMMI · AIRBUS DEFENCE & SPACE · UNIVERSITÉ DE NANTES · La Ruche Numérique un service CCI LE MANS SARTHE · LM SARTHE · ina EXPERT Enseignement Formation Recherche · ANR Soda · Région PAYS DE LA LOIRE · Conseil général Sarthe

Message du président de l'AFCP

C'est toujours avec plaisir que je vois revenir « les années JEP » et ce rendez-vous crucial pour notre communauté, afin d'échanger autour des dernières avancées scientifiques mais également afin de maintenir et de renforcer les liens scientifiques et humains qui nous amènent à nous retrouver régulièrement, au-delà des disciplines et des générations.

Ces liens revêtent aujourd'hui une importance particulière face aux modifications profondes que subit le monde de la recherche académique dans la sphère francophone et au-delà : compétition accrue pour les financements sur projet, individualisation des parcours et course aux indicateurs de visibilité, et donc, à la publication internationale, incitation au transfert de technologie et à un impact sociétal immédiat, etc. Dans ce contexte, les « petites » conférences nationales souffrent trop souvent de ne pouvoir offrir une visibilité internationale immédiate ! Et pourtant, ces conférences, sans être aussi sélectives que les conférences internationales majeures, n'en sont que plus capitales. En particulier, les Journées d'étude sur la Parole demeurent une manifestation de grande qualité scientifique, à taille humaine et sans la barrière de la langue, dans laquelle se côtoient les différentes disciplines qui font la richesse et l'intérêt de notre domaine. Ces particularités uniques permettent la naissance de nouveaux courants de pensées, de nouveaux axes de recherche, de nouvelles collaborations, souvent pluridisciplinaires, de projets, etc. Autant d'éléments qui font la force de notre communauté et qui offrent à chacun des atouts pour se maintenir dans la compétition à laquelle nous ne pouvons échapper. Autant d'éléments qu'il faut chérir.

J'en suis sûr, la 30[e] édition, au Mans, ne dérogera pas aux règles de qualité scientifique et de convivialité qui sont celles des JEP depuis leur création. Quelques faits sur cette édition. Nous avons reçu un total de 111 soumissions : sans être un record, ce nombre montre l'attrait que continuent à exercer les JEP dans le contexte difficile dont je parlais en introduction. Au final, le programme comporte 79 communications, dont 22 à l'oral. Je rappelle à cette occasion que, contrairement à de nombreuses manifestations, le comité de programme n'établit aucune hiérarchie entre communication orale et communication affichée, le choix étant principalement guidé par les thématiques que nous avons souhaité mettre en avant dans les sessions orales. Cette année, outre la session « Transcription et traitement automatique », qui correspond à la sensibilité des organisateurs, nous avons privilégié des thématiques dans lesquels se mêlent Sciences humaines et sociales et Sciences et technologie de l'information. Ainsi, les trois sessions « Prosodie et synthèse », « Voix » et « Langues régionales ou peu dotées » regroupent chacune des communications de ces deux disciplines. Cette pluridisciplinarité, qui est une des caractéristiques fascinantes de notre domaine, se retrouve bien évidemment dans les sessions affichées, comme il est de coutume depuis 2006. Pour clore ce tour d'horizon du programme 2014, il me paraît important de souligner l'ouverture au grand public jeudi soir, avec des démonstrations et une conférence sur la voix : on ne peut que louer la volonté des organisateurs de faire connaître notre sujet d'étude et le résultat de nos travaux au plus grand nombre et je vous invite à y assister.

Je souhaite remercier vivement les membres du comité de programme et du comité scientifique pour le temps qu'ils ont consacré au processus de sélection des communications : il s'agit d'un travail prenant, peu valorisé, mais indispensable. Mais surtout, je tiens à remercier chaleureusement les organisateurs de cette édition, au Mans et à Nantes, pour l'énorme travail qu'ils ont accompli pour que cet événement soit à n'en pas douter un grand succès et un moment fort de la vie de notre communauté.

Je vous souhaite une conférence fructueuse sur les plans scientifique et humain, faite d'échanges, de discussions et de découvertes.

Guillaume Gravier,
Président du CA de l'Association Francophone de la Communication Parlée.

Message des organisateurs

Pour leur trentième édition riche d'une histoire débutée en 1970, c'est avec grand plaisir et une immense fierté que nous avons accueilli les Journées d'études sur la Parole au Mans.

L'organisation des JEP 2014 a été assurée par les équipes LST du LIUM (Laboratoire d'Informatique de l'Université du Maine) et TALN du LINA (Laboratoire d'Informatique de Nantes Atlantique). L'année précédente, ces deux équipes s'étaient également associées pour l'organisation de la conférence TALN 2013 aux Sables d'Olonne. Cette complicité grandissante entre ces équipes, toutes deux membres de la Fédération de Recherche CNRS AtlanSTIC, illustre le dynamisme de la thématique « Traitement automatique du langage et de la parole » dans la région des Pays de la Loire.

Environ 120 conférenciers provenant de 6 pays différents étaient présents cette année, pour un total de 79 communications scientifiques sélectionnées par le comité de programme sur 111 soumissions. En plus de ces communications, nous avons eu le plaisir d'assister aux interventions de quatre conférenciers invités, que nous remercions ici encore pour leur participation : Jérôme Bellegarda, Renato De Mori, Cécile Fougeron et Nathalie Henrich Bernardoni. Nous avons souhaité que le mot d'ordre des JEP 2014 soit le mot « échanges ». Tout d'abord échanges entre chercheurs bien entendu, et tout particulièrement entre chercheurs confirmés et jeunes chercheurs qui le deviendront ; mais aussi échanges entre recherche académique et monde industriel, et échanges entre recherche et grand public.

Les échanges entre chercheurs francophones sont la raison d'être des JEP, et ont trouvé comme d'habitude leur source dans les sessions poster et orales, ponctuées des repas de midi pris tous ensemble et des événements sociaux : accueil sarthois le lundi soir dans un magnifique salon de la Chambre de Commerce et d'Industrie du Mans et de la Sarthe, apéritif le mardi soir offert par la Mairie du Mans à l'Hôtel de Ville, banquet du mercredi dans le Musée des 24 Heures au cours duquel les amateurs de belles mécaniques ont été comblés et les autres pas oubliés. Le jeudi a été une journée riche en échanges : en plus des communications scientifiques, était organisée une session de démonstrations de projets académiques et industriels dont la dernière heure était ouverte à tous, suivie d'une nouvelle conférence de Nathalie Henrich Bernardoni à destination du grand public, co-organisée avec l'institut Le Mans Acoustique et l'association Maine Sciences. Merci à l'AFCP de nous avoir soutenus dans cette démarche de dissémination scientifique.

Pour fêter dignement cette 30^{e} édition des Journées d'études sur la Parole, nous avons souhaité faire un effort particulier sur les présentations orales : elles ont toutes été filmées, puis transcrites avec le système de reconnaissance automatique de la parole du LIUM. Le corpus constitué sera offert à la communauté qui pourra s'en emparer comme elle le désire.

Toute cette organisation n'aurait pas pu être possible sans tous les partenaires institutionnels et privés qui se sont joints à nous pour faire de cette 30^{e} conférence JEP une véritable réussite. Qu'ils en soient ici remerciés.

Yannick Estève et Emmanuel Morin,
pour le comité d'organisation des JEP 2014.

Liste des articles par auteur

A

Mohammad Abuoudeh

66 Équations de locus et temporalité de la parole : Une étude de l'impact de la durée vocalique sur les paramètres des équations de locus en Arabe Jordanien ..... p. 475

Audrey Acher

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Gilles Adda

36 Modélisation acoustico-phonétique de langues peu dotées : Études phonétiques et travaux de reconnaissance automatique en luxembourgois ..... p. 284

Martine Adda-Decker

36 Modélisation acoustico-phonétique de langues peu dotées : Études phonétiques et travaux de reconnaissance automatique en luxembourgois ..... p. 284
48 Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques ..... p. 366
56 Comment évaluer la qualité des transcriptions automatiques pour la détection d'entités nommées ? p. 430

Angélique Amelot

80 Etude de la production des voyelles nasales du français chez des apprenants espagnols et colombiens p. 576

Zeineb Ammar

101 A la recherche des traces dialectales dans l'arabe standard : production des voyelles et des fricatives inter-dentales par des locuteurs tunisiens et marocains. ..... p. 684

Régine André-Obrecht

21 Application du critère BIC pour la segmentation en tours de chant ..... p. 166

Vincent Arnaud

79 Les voyelles fermées tendues relâchées et allongées du français québécois : la contribution d'indices statiques/dynamiques et absolus/normalisés à la détermination de leur identité acoustique ..... p. 567
82 Portrait acoustique d'une variation régionale en français québécois contemporain : l'ouverture de la voyelle /ɛ/ en finale de mot ..... p. 585

Thierry Artières

62 Etude des HMMs paramétriques pour la reconnaissance de parole en environnement bruité . p. 466

Suzanne Assadi

95 L'effet des contextes segmentaux et prosodiques sur la glottalisation de la voyelle initiale de mot en persan ..... p. 640

Arnaud Attyé

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Jean-Julien Aucouturier

32 Analyse prosodique des affects sociaux dans l'interaction face à face en japonais ..... p. 257

Nicolas Audibert

30 Mesures de protrusion par capture optique de mouvements : quelle métrique est la plus représentative de l'opposition d'arrondissement en français ? ..... p. 239

Mathieu Avanzi

6 Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale p. 49

B

Pierre Badin

59 Comparaison des stratégies articulatoires d'un locuteur bilingue anglais-français : Données et modèles préliminaires ..... p. 448

Nicolas Ballier

76 Modélisation de l'intonation interlangue : le cas des questions ..... p. 540

Nelly Barbot

6 Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale p. 49
7 Évaluation d'un algorithme de chunking appliqué à la dictée ..... p. 58

Melissa Barkat-Defradas

89 iPhocomp : calcul automatique de l'indice de complexité phonétique de Jakielski ..... p. 622
90 Influence du développement phonético-phonologique sur les premières productions lexicales d'enfants arabophones : une étude longitudinale ..... p. 631

Claude Barras

70 Étude des voyelles et de la force de voix par analyse discriminante ..... p. 504

Loïc Barrault

97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

Giulia Barreca

39 Multidimensionnalité de la liaison variable et difficultés de classification. Le cas des adverbes monosyllabiques. ..... p. 311

Aurélia Bascoul

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive ..... p. 248

Richard Beaufort

54 Train&Align : un outil d'alignement phonétique automatique disponible en ligne ..... p. 412

Denis Beautemps

43 L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles ..... p. 329

Frédéric Béchet

73 Correction interactive de transcriptions de parole par fusion de phrases ..... p. 522
98 Détection et caractérisation d'erreurs dans des transcriptions automatiques pour des systèmes de traduction parole-parole ..... p. 667

Nathalie Bedoin

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive ..... p. 248

Lamia Hadrich Belguith

71 Phonétisation automatique du dialecte tunisien ..... p. 513

Hacène Bellemmouche

90 Influence du développement phonético-phonologique sur les premières productions lexicales d'enfants arabophones : une étude longitudinale ..... p. 631

Mohamed Ameur Ben Jannet

56 Comment évaluer la qualité des transcriptions automatiques pour la détection d'entités nommées ? p. 430

Laurent Besacier

12 Construction faiblement supervisée d'un phonétiseur pour la langue iban à partir de ressources en malais ..... p. 101

Georges Bettega

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Benjamin Bigot

107 Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques ..... p. 720

Eric Bilinski

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Hélène Bonneau-Maynard

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Xavier Bost

108 Catégorisation multi-thématique de dialogues téléphoniques ..... p. 728

Mohamed Bouallègue

109 Analyse factorielle pour une représentation vectorielle des états des Modèles de Markov Cachés p. 737

Mohamed Bouaziz

24 Décodage hybride dans les SRAP pour l'indexation automatique des documents multimédia p. 184

Abdessalam Bouchekif

58 Exploitation de la distribution des locuteurs pour la segmentation thématique de journaux télévisés p. 439

Victor J. Boucher

99 L'identification vocale : pour une quantification des effets de la familiarité ..... p. 676

Mounir Boudjerda

22 Estimation de la source glottique par décomposition modale empirique ..... p. 175

Jérôme Boudy

62 Etude des HMMs paramétriques pour la reconnaissance de parole en environnement bruité . p. 466

Fethi Bougares

71 Phonétisation automatique du dialecte tunisien ..... p. 513
97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

Philippe Boula de Mareüil

13 Une étude prosodique comparative des questions en français en contact avec l'occitan ..... p. 111

Véronique Boulenger

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive ..... p. 248

Souhir Bousselmi

60 Implémentation d'un nouvel algorithme de codage de la parole basé sur la transformation en paquets de framelettes ..... p. 457

Merouane Bouzid

9 Quantificateur vectoriel à divisions par classification : Application au codage des fréquences spectrales d'immitance en large-bande ..... p. 75

Muriel Brix

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Sandrine Brognaux

54 Train&Align : un outil d'alignement phonétique automatique disponible en ligne ..... p. 412

Sylvie Brunessaux

97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

David Alejandro Bustamante

80 Etude de la production des voyelles nasales du français chez des apprenants espagnols et colombiens p. 576

C

Nathalie Camelin

86 Boosting de bonzaïs pour la combinaison efficace de descripteurs : application à l'identification du rôle du locuteur ..... p. 603

Julie Cano-Chervel

50 Capacités d'apprentissage phonétique chez les sujets âgés francophones ..... p. 385

Brice Carlot

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Manon Carrissimo-Bertola

33 Labial-Coronal vs. Labial-Vélaire : Étude du phasage des gestes en Français ..... p. 266

Jean Carrive

8 Quand l'horloge parlante a beaucoup à raconter sur l'évolution des techniques d'archivage audiovisuel ..... p. 67

Diane Caussade

78 Coordination/synchronisation gestes-voix dans la démence de la maladie d'Alzheimer en voix parlée et chantée : une étude de cas ..... p. 558

Delphine Charlet

58 Exploitation de la distribution des locuteurs pour la segmentation thématique de journaux télévisés p. 439

Jonathan Chevelu

16 Roots : un outil pour manipuler facilement efficacement et avec cohérence des corpus annotés de séquences ..... p. 130

Ioana Chitoran

105 La réorganisation gestuelle interagit avec la phonotactique spécifique à chaque langue ..... p. 712
33 Labial-Coronal vs. Labial-Vélaire : Étude du phasage des gestes en Français ..... p. 266

Gérard Chollet

62 Etude des HMMs paramétriques pour la reconnaissance de parole en environnement bruité . p. 466

Chloé Clavel

48 Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques ..... p. 366

Jean-Marc Colletta

75 Production et planification de la parole spontanée en contexte narratif chez l'enfant âgé de 3 à 11 ans p. 531
78 Coordination/synchronisation gestes-voix dans la démence de la maladie d'Alzheimer en voix parlée et chantée : une étude de cas ..... p. 558

Léa Cottavoz

55 Aspects de la coordination pneumophonique dans la dysarthrie parkinsonienne ..... p. 421

Lise Crevier-Buchman

77 Contribution à l'étude physiologique des possibilités articulatoires dans la cavité pharyngale p. 549

Olivier Crouzet

66 Équations de locus et temporalité de la parole : Une étude de l'impact de la durée vocalique sur les paramètres des équations de locus en Arabe Jordanien ..... p. 475

D

Sawssan Dabbar

71 Phonétisation automatique du dialecte tunisien ..... p. 513

Géraldine Damnati

58 Exploitation de la distribution des locuteurs pour la segmentation thématique de journaux télévisés p. 439

Renato De Mori

108 Catégorisation multi-thématique de dialogues téléphoniques ..... p. 728
3 Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ? ..... p. 31

Elisabeth Delais-Roussarie

6 Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale p. 49
7 Évaluation d'un algorithme de chunking appliqué à la dictée ..... p. 58
76 Modélisation de l'intonation interlangue : le cas des questions ..... p. 540

Paul Deléglise

97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

Céline Delooze

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40

Véronique Delvaux

50 Capacités d'apprentissage phonétique chez les sujets âgés francophones ..... p. 385

Didier Demolin

55 Aspects de la coordination pneumophonique dans la dysarthrie parkinsonienne ..... p. 421

Myriam Desainte-Catherine

37 Caractérisation et classification automatique des modes phonatoires en voix chantée ..... p. 293

Julien Despres

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Ivana Didirkova

20 Etude préliminaire de l'influence des caractéristiques phonétiques sur le bégaiement : le cas du français et du slovaque ..... p. 157

Thomas Drugman

54 Train&Align : un outil d'alignement phonétique automatique disponible en ligne ..... p. 412

Richard Dufour

107 Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques ..... p. 720
3 Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ? ..... p. 31

Grégor Dupuy

42 Segmentation et regroupement en locuteur pour le traitement incrémental des collections volumineuses ..... p. 320

Camille Dutrey

48 Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques ..... p. 366

E

Layla El Asri

51 Régression ordinale pour la prédiction de la qualité d'interaction ..... p. 394

Marc El-Bèze

108 Catégorisation multi-thématique de dialogues téléphoniques ..... p. 728

Mariem Ellouze Khemakhem

71 Phonétisation automatique du dialecte tunisien ..... p. 513

John H. Esling

77 Contribution à l'étude physiologique des possibilités articulatoires dans la cavité pharyngale p. 549

Yannick Estève

24 Décodage hybride dans les SRAP pour l'indexation automatique des documents multimédia p. 184
42 Segmentation et regroupement en locuteur pour le traitement incrémental des collections volumineuses ..... p. 320
71 Phonétisation automatique du dialecte tunisien ..... p. 513
97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

F

Nora Fangel-Gustavson

88 Le contraste quantitatif en saami de Lule : un système à trois niveaux ..... p. 612

Camille Fauth

27 Perturbations du Voice Onset Time (V.O.T) et du Voice Termination Time (V.T.T.) dans des séquences VCV produites par des patients francophones souffrant de paralysies récurrentielles ..... p. 211

Benoît Favre

73 Correction interactive de transcriptions de parole par fusion de phrases ..... p. 522
98 Détection et caractérisation d'erreurs dans des transcriptions automatiques pour des systèmes de traduction parole-parole ..... p. 667

Emmanuel Ferragne

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive ..... p. 248

Gaëlle Ferré

11 Tension gestuelle en co-présence d'un accent d'intensité ..... p. 92

Dominique Fohr

18 Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique ..... p. 139

Lionel Fontan

103 Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit ..... p. 694

Cécile Fougeron

101 A la recherche des traces dialectales dans l'arabe standard : production des voyelles et des fricatives inter-dentales par des locuteurs tunisiens et marocains. ..... p. 684
35 L'aspiration du /s/ coda en andalou occidental est-elle fonction du type de /s/ ? ..... p. 275

Dominique Fourer

32 Analyse prosodique des affects sociaux dans l'interaction face à face en japonais ..... p. 257

Corinne Fredouille

107 Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques ..... p. 720

G

Pascal Gaillard

103 Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit ..... p. 694

Olivier Galibert

56 Comment évaluer la qualité des transcriptions automatiques pour la détection d'entités nommées ? p. 430

Philippe Gambette

89 iPhocomp : calcul automatique de l'indice de complexité phonétique de Jakielski ..... p. 622

Peggy Gatignol

15 Caractérisation acoustico-phonétique de parole provenant de patients atteints de paralysie faciale p. 121

Jean-Luc Gauvain

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Frédérique Gayraud

90 Influence du développement phonético-phonologique sur les premières productions lexicales d'enfants arabophones : une étude longitudinale ..... p. 631

Cedric Gendrot

25 Perception et Réalisation du /R/ standard français en finale de mot ..... p. 193

Laurianne Georgeton

30 Mesures de protrusion par capture optique de mouvements : quelle métrique est la plus représentative de l'opposition d'arrondissement en français ? ..... p. 239

Silvain Gerber

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220
55 Aspects de la coordination pneumophonique dans la dysarthrie parkinsonienne ..... p. 421

Alain Ghio

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40

Mark Gibson

35 L'aspiration du /s/ coda en andalou occidental est-elle fonction du type de /s/ ? ..... p. 275

Ludivine Glas

67 Lien entre acquisition des premiers mots chez des enfants francophones monolingues et leur environnement langagier ..... p. 484

Li Gong

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Sibylle Gonzalez-Monge

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive ..... p. 248

Arseniy Gorin

44 Modélisation de trajectoires et de classes de locuteurs pour la reconnaissance de voix d'enfants et d'adultes ..... p. 339

Cécile Grigoli

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40

David Guennec

10 Utilisation d'un algorithme A* pour l'analyse de la sélection d'unités en synthèse de la parole . p. 83

Marine Guerry

32 Analyse prosodique des affects sociaux dans l'interaction face à face en japonais .....p. 257

Camille Guinaudeau

46 Analyse du corpus MATRICE-INA : exploration et classification automatique d'archives audiovisuelles de 1930 à 2012 ..... p. 357

Hélène Guiraud

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive .....p. 248

Patrice Guyot

21 Application du critère BIC pour la segmentation en tours de chant ..... p. 166

H

Ali Hadian Cefidekhanie

75 Production et planification de la parole spontanée en contexte narratif chez l'enfant âgé de 3 à 11 ans p. 531

Rim Hamdi-Sultan

90 Influence du développement phonético-phonologique sur les premières productions lexicales d'enfants arabophones : une étude longitudinale ..... p. 631

Bernard Harmegnies

50 Capacités d'apprentissage phonétique chez les sujets âgés francophones .....p. 385

Nathalie Henrich Bernardoni

78 Coordination/synchronisation gestes-voix dans la démence de la maladie d'Alzheimer en voix parlée et chantée : une étude de cas ..... p. 558

Vania Herbillon

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive .....p. 248

Sophie Herment

76 Modélisation de l'intonation interlangue : le cas des questions ..... p. 540

Fabrice Hirsch

20 Etude préliminaire de l'influence des caractéristiques phonétiques sur le bégaiement : le cas du français et du slovaque ..... p. 157

Kathy Huet

50 Capacités d'apprentissage phonétique chez les sujets âgés francophones ..... p. 385

I

Irina Illina

18 Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique ..... p. 139

Léonidas Ioannidis

37 Caractérisation et classification automatique des modes phonatoires en voix chantée ..... p. 293

Fanny Ivent

13 Une étude prosodique comparative des questions en français en contact avec l'occitan ..... p. 111
35 L'aspiration du /s/ coda en andalou occidental est-elle fonction du type de /s/ ? ..... p. 275

J

Yvan Josse

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Denis Jouvet

104 Combinaison de mots et de syllabes pour transcrire la parole ..... p. 703
44 Modélisation de trajectoires et de classes de locuteurs pour la reconnaissance de voix d'enfants et d'adultes ..... p. 339

K

Abdellah Kacha

22 Estimation de la source glottique par décomposition modale empirique ..... p. 175

Juliette Kahn

56 Comment évaluer la qualité des transcriptions automatiques pour la détection d'entités nommées ? p. 430

Takeki Kamiyama

26 Opposition /y/-/u/ chez 7 apprenantes japonophones avec et sans l'aide de l'ultrason lingual : 1. analyse acoustique comparative voyelles isolées / logatomes ..... p. 202
96 Opposition /y/-/u/ chez sept apprenantes japonophones avec et sans aide de l'ultrason lingual : 2. identification perceptive des voyelles isolées par 16 francophones natifs ..... p. 649

Sophie Kern

67 Lien entre acquisition des premiers mots chez des enfants francophones monolingues et leur environnement langagier ..... p. 484

Rajesh Khatiwada

83 L'articulation des coronales occlusives en népalais ..... p. 594

Khaled Khelif

97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

Hatim Khouzaimi

51 Régression ordinale pour la prédiction de la qualité d'interaction ..... p. 394

Tanja Kocjančič Antolík

26 Opposition /y/-/u/ chez 7 apprenantes japonophones avec et sans l'aide de l'ultrason lingual : 1. analyse acoustique comparative voyelles isolées / logatomes ..... p. 202
96 Opposition /y/-/u/ chez sept apprenantes japonophones avec et sans aide de l'ultrason lingual : 2. identification perceptive des voyelles isolées par 16 francophones natifs ..... p. 649

Alexandre Krainik

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Sonia Krifi-Papoz

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive ..... p. 248

L

Laurent Lamalle

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220
59 Comparaison des stratégies articulatoires d'un locuteur bilingue anglais-français : Données et modèles préliminaires ..... p. 448

Lori Lamel

36 Modélisation acoustico-phonétique de langues peu dotées : Études phonétiques et travaux de reconnaissance automatique en luxembourgois ..... p. 284
49 Traduction de la parole dans le projet RAPMAT ..... p. 375
53 Développement d'un système de reconnaissance automatique de la parole en coréen avec peu de ressources annotées ..... p. 403

Romain Laroche

51 Régression ordinale pour la prédiction de la qualité d'interaction ..... p. 394

Antoine Laurent

24 Décodage hybride dans les SRAP pour l'indexation automatique des documents multimédia p. 184
46 Analyse du corpus MATRICE-INA : exploration et classification automatique d'archives audiovisuelles de 1930 à 2012 ..... p. 357
49 Traduction de la parole dans le projet RAPMAT ..... p. 375
53 Développement d'un système de reconnaissance automatique de la parole en coréen avec peu de ressources annotées ..... p. 403
86 Boosting de bonzaïs pour la combinaison efficace de descripteurs : application à l'identification du rôle du locuteur ..... p. 603

Việt Bắc Lê

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Maxime Le Coz

15 Caractérisation acoustico-phonétique de parole provenant de patients atteints de paralysie faciale p. 121

Sébastien Le Maguer

6 Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale p. 49
7 Évaluation d'un algorithme de chunking appliqué à la dictée ..... p. 58

Gwénolé Lecorvé

16 Roots : un outil pour manipuler facilement efficacement et avec cohérence des corpus annotés de séquences ..... p. 130

Benjamin Lecouteux

38 Reconnaissance automatique de la parole à l'aide de colonies de fourmis ..... p. 302

Hyeran Lee

89 iPhocomp : calcul automatique de l'indice de complexité phonétique de Jakielski ..... p. 622

Jérôme Lehuen

19 Le couple YASP/YADE – Yet Another Semantic Parser/Yet Another Dialogue Engine ..... p. 148

Jean-Sylvain Liénard

70 Étude des voyelles et de la force de voix par analyse discriminante ..... p. 504

Georges Linarès

107 Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques ..... p. 720
109 Analyse factorielle pour une représentation vectorielle des états des Modèles de Markov Cachés p. 737
18 Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique ..... p. 139

3 Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ? ..... p. 31

Damien Lolive

10 Utilisation d'un algorithme A* pour l'analyse de la sélection d'unités en synthèse de la parole . p. 83
16 Roots : un outil pour manipuler facilement efficacement et avec cohérence des corpus annotés de séquences ..... p. 130
6 Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale p. 49
7 Évaluation d'un algorithme de chunking appliqué à la dictée ..... p. 58

M

Cynthia Magnen

103 Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit ..... p. 694

Paolo Mairano

29 Structures syllabiques et caractéristiques du cycle mandibulaire : une étude articulatoire des asymétries ..... p. 230

Mathieu Manta

97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

Marielle Mas

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40

Abir Masmoudi

71 Phonétisation automatique du dialecte tunisien ..... p. 513

Driss Matrouf

109 Analyse factorielle pour une représentation vectorielle des états des Modèles de Markov Cachés p. 737

Julie Mauclair

15 Caractérisation acoustico-phonétique de parole provenant de patients atteints de paralysie faciale p. 121

Sylvain Meignier

42 Segmentation et regroupement en locuteur pour le traitement incrémental des collections volumineuses ..... p. 320

Céline Mercier

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40

Pierrick Milhorat

62 Etude des HMMs paramétriques pour la reconnaissance de parole en environnement bruité . p. 466

Kemiha Mina

22 Estimation de la source glottique par décomposition modale empirique ..... p. 175

Mohamed Morchid

3 Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ? ..... p. 31

N

Quang Thuấn Nguyễn

45 Étude des productions des consonnes du français par des apprenants vietnamiens du Français Langue Etrangère ..... p. 348

O

Luiza Orosanu

104 Combinaison de mots et de syllabes pour transcrire la parole .....p. 703

Kaïs Ouni

60 Implémentation d'un nouvel algorithme de codage de la parole basé sur la transformation en paquets de framelettes .....p. 457

P

Thomas Pellegrini

15 Caractérisation acoustico-phonétique de parole provenant de patients atteints de paralysie faciale p. 121

21 Application du critère BIC pour la segmentation en tours de chant ..... p. 166

Catherine Pelletq

75 Production et planification de la parole spontanée en contexte narratif chez l'enfant âgé de 3 à 11 ans p. 531

Pascal Perrier

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale .....p. 220

Myriam Piccaluga

50 Capacités d'apprentissage phonétique chez les sujets âgés francophones ..... p. 385

Olivier Pietquin

51 Régression ordinale pour la prédiction de la qualité d'interaction ..... p. 394

Claire Pillot-Loiseau

26 Opposition /y/-/u/ chez 7 apprenantes japonophones avec et sans l'aide de l'ultrason lingual : 1. analyse acoustique comparative voyelles isolées / logatomes ..... p. 202
80 Etude de la production des voyelles nasales du français chez des apprenants espagnols et colombiens p. 576
96 Opposition /y/-/u/ chez sept apprenantes japonophones avec et sans aide de l'ultrason lingual : 2. identification perceptive des voyelles isolées par 16 francophones natifs ..... p. 649

Julien Pinquier

21 Application du critère BIC pour la segmentation en tours de chant ..... p. 166

Julien Plante-Hébert

99 L'identification vocale : pour une quantification des effets de la familiarité ..... p. 676

R

Mathieu Radenen

62 Etude des HMMs paramétriques pour la reconnaissance de parole en environnement bruité . p. 466

Christian Raymond

86 Boosting de bonzaïs pour la combinaison efficace de descripteurs : application à l'identification du rôle du locuteur ..... p. 603

Rachid Ridouane

101 A la recherche des traces dialectales dans l'arabe standard : production des voyelles et des fricatives inter-dentales par des locuteurs tunisiens et marocains. ..... p. 684
88 Le contraste quantitatif en saami de Lule : un système à trois niveaux ..... p. 612

Christian Adrien Righini

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Albert Rilliard

13 Une étude prosodique comparative des questions en français en contact avec l'occitan ..... p. 111
32 Analyse prosodique des affects sociaux dans l'interaction face à face en japonais ..... p. 257

Josiane Riverin-Coutlée

82 Portrait acoustique d'une variation régionale en français québécois contemporain : l'ouverture de la voyelle /ɛ/ en finale de mot ..... p. 585

Danièle Robert

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40

Marina Robert

15 Caractérisation acoustico-phonétique de parole provenant de patients atteints de paralysie faciale p. 121

Jean-François Rodier

27 Perturbations du Voice Onset Time (V.O.T) et du Voice Termination Time (V.T.T.) dans des séquences VCV produites par des patients francophones souffrant de paralysies récurrentielles ..... p. 211

Sophie Roekhaut

54 Train&Align : un outil d'alignement phonétique automatique disponible en ligne ..... p. 412

Olivier Rosec

6 Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale p. 49

Solange Rossato

12 Construction faiblement supervisée d'un phonétiseur pour la langue iban à partir de ressources en malais ..... p. 101
29 Structures syllabiques et caractéristiques du cycle mandibulaire : une étude articulatoire des asymétries ..... p. 230

Sophie Rosset

48 Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques ..... p. 366
56 Comment évaluer la qualité des transcriptions automatiques pour la détection d'entités nommées ? p. 430

Jean-Luc Rouas

32 Analyse prosodique des affects sociaux dans l'interaction face à face en japonais ..... p. 257
37 Caractérisation et classification automatique des modes phonatoires en voix chantée ..... p. 293

Anthony Rousseau

97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

Mickael Rouvier

73 Correction interactive de transcriptions de parole par fusion de phrases ..... p. 522

Anindya Roy

46 Analyse du corpus MATRICE-INA : exploration et classification automatique d'archives audiovisuelles de 1930 à 2012 ..... p. 357

S

Sarah Samson Juan

12 Construction faiblement supervisée d'un phonétiseur pour la langue iban à partir de ressources en malais ..... p. 101

Marc Sato

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220
43 L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles ..... p. 329

Christophe Savariaux

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Thomas R. Sawallis

59 Comparaison des stratégies articulatoires d'un locuteur bilingue anglais-français : Données et modèles préliminaires ..... p. 448

Lucie Scarbel

43 L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles ..... p. 329

Sébastien Schmerber

43 L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles ..... p. 329

Didier Schwab

38 Reconnaissance automatique de la parole à l'aide de colonies de fourmis ..... p. 302

Jean-Luc Schwartz

43 L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles ..... p. 329

Holger Schwenk

97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658

Natalia Segal

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Grégory Senay

107 Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques ..... p. 720

Takaaki Shochi

32 Analyse prosodique des affects sociaux dans l'interaction face à face en japonais ..... p. 257

Caroline Sigouin

79 Les voyelles fermées tendues relâchées et allongées du français québécois : la contribution d'indices statiques/dynamiques et absolus/normalisés à la détermination de leur identité acoustique ..... p. 567

Rudolph Sock

27 Perturbations du Voice Onset Time (V.O.T) et du Voice Termination Time (V.T.T.) dans des séquences VCV produites par des patients francophones souffrant de paralysies récurrentielles ..... p. 211
69 Les occlusives de l'alsacien : une étude temporelle ..... p. 494

Lucie Steiblé

69 Les occlusives de l'alsacien : une étude temporelle ..... p. 494

T

Julien Tardieu

103 Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit ..... p. 694

Bernard Teston

55 Aspects de la coordination pneumophonique dans la dysarthrie parkinsonienne ..... p. 421

Marwa Thlithi

21 Application du critère BIC pour la segmentation en tours de chant ..... p. 166

Mark Tiede

105 La réorganisation gestuelle interagit avec la phonotactique spécifique à chaque langue ..... p. 712

Anne Tortel

76 Modélisation de l'intonation interlangue : le cas des questions ..... p. 540

Thi Thuy Hien Tran

29 Structures syllabiques et caractéristiques du cycle mandibulaire : une étude articulatoire des asymétries ..... p. 230
45 Étude des productions des consonnes du français par des apprenants vietnamiens du Français Langue Etrangère ..... p. 348

V

Nathalie Vallée

29 Structures syllabiques et caractéristiques du cycle mandibulaire : une étude articulatoire des asymétries ..... p. 230
33 Labial-Coronal vs. Labial-Vélaire : Étude du phasage des gestes en Français ..... p. 266
45 Étude des productions des consonnes du français par des apprenants vietnamiens du Français Langue Etrangère ..... p. 348
78 Coordination/synchronisation gestes-voix dans la démence de la maladie d'Alzheimer en voix parlée et chantée : une étude de cas ..... p. 558

Félicien Vallet

8 Quand l'horloge parlante a beaucoup à raconter sur l'évolution des techniques d'archivage audiovisuel ..... p. 67

Ioana Vasilescu

48 Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques ..... p. 366

Béatrice Vaxelaire

27 Perturbations du Voice Onset Time (V.O.T) et du Voice Termination Time (V.T.T.) dans des séquences VCV produites par des patients francophones souffrant de paralysies récurrentielles ..... p. 211

François Viallet

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40
55 Aspects de la coordination pneumophonique dans la dysarthrie parkinsonienne ..... p. 421

Coriandre Vilain

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220

Pierre-Philippe Volkmar

27 Perturbations du Voice Onset Time (V.O.T) et du Voice Termination Time (V.T.T.) dans des séquences VCV produites par des patients francophones souffrant de paralysies récurrentielles ..... p. 211

Y

Hiyon Yoo

7 Évaluation d'un algorithme de chunking appliqué à la dictée ..... p. 58

François Yvon

49 Traduction de la parole dans le projet RAPMAT ..... p. 375

Z

Chakir Zeroual

77 Contribution à l'étude physiologique des possibilités articulatoires dans la cavité pharyngale p. 549

Liste des articles par session

Session poster 1

7 Évaluation d'un algorithme de chunking appliqué à la dictée ..... p. 58
10 Utilisation d'un algorithme A* pour l'analyse de la sélection d'unités en synthèse de la parole . p. 83
18 Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique ..... p. 139
24 Décodage hybride dans les SRAP pour l'indexation automatique des documents multimédia p. 184
26 Opposition /y/-/u/ chez 7 apprenantes japonophones avec et sans l'aide de l'ultrason lingual : 1. analyse acoustique comparative voyelles isolées / logatomes ..... p. 202
30 Mesures de protrusion par capture optique de mouvements : quelle métrique est la plus représentative de l'opposition d'arrondissement en français ? ..... p. 239
33 Labial-Coronal vs. Labial-Vélaire : Étude du phasage des gestes en Français ..... p. 266
35 L'aspiration du /s/ coda en andalou occidental est-elle fonction du type de /s/ ? ..... p. 275
39 Multidimensionnalité de la liaison variable et difficultés de classification. Le cas des adverbes monosyllabiques. ..... p. 311
42 Segmentation et regroupement en locuteur pour le traitement incrémental des collections volumineuses ..... p. 320
46 Analyse du corpus MATRICE-INA : exploration et classification automatique d'archives audiovisuelles de 1930 à 2012 ..... p. 357
56 Comment évaluer la qualité des transcriptions automatiques pour la détection d'entités nommées ? p. 430
59 Comparaison des stratégies articulatoires d'un locuteur bilingue anglais-français : Données et modèles préliminaires ..... p. 448
96 Opposition /y/-/u/ chez sept apprenantes japonophones avec et sans aide de l'ultrason lingual : 2. identification perceptive des voyelles isolées par 16 francophones natifs ..... p. 649

Session orale 1 : Production et perception

43 L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles ..... p. 329
66 Équations de locus et temporalité de la parole : Une étude de l'impact de la durée vocalique sur les paramètres des équations de locus en Arabe Jordanien ..... p. 475
103 Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit ..... p. 694
105 La réorganisation gestuelle interagit avec la phonotactique spécifique à chaque langue ..... p. 712

Session orale 2 : Prosodie et synthèse

6 Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale p. 49
13 Une étude prosodique comparative des questions en français en contact avec l'occitan ..... p. 111
32 Analyse prosodique des affects sociaux dans l'interaction face à face en japonais ..... p. 257

Session poster 2

3 Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ? ..... p. 31
12 Construction faiblement supervisée d'un phonétiseur pour la langue iban à partir de ressources en malais ..... p. 101

28 Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale ..... p. 220
29 Structures syllabiques et caractéristiques du cycle mandibulaire : une étude articulatoire des asymétries ..... p. 230
44 Modélisation de trajectoires et de classes de locuteurs pour la reconnaissance de voix d'enfants et d'adultes ..... p. 339
49 Traduction de la parole dans le projet RAPMAT ..... p. 375
53 Développement d'un système de reconnaissance automatique de la parole en coréen avec peu de ressources annotées ..... p. 403
60 Implémentation d'un nouvel algorithme de codage de la parole basé sur la transformation en paquets de framelettes ..... p. 457
62 Etude des HMMs paramétriques pour la reconnaissance de parole en environnement bruité . p. 466
67 Lien entre acquisition des premiers mots chez des enfants francophones monolingues et leur environnement langagier ..... p. 484
73 Correction interactive de transcriptions de parole par fusion de phrases ..... p. 522
80 Etude de la production des voyelles nasales du français chez des apprenants espagnols et colombiens p. 576
82 Portrait acoustique d'une variation régionale en français québécois contemporain : l'ouverture de la voyelle /ɛ/ en finale de mot ..... p. 585
83 L'articulation des coronales occlusives en népalais ..... p. 594
86 Boosting de bonzaïs pour la combinaison efficace de descripteurs : application à l'identification du rôle du locuteur ..... p. 603
88 Le contraste quantitatif en saami de Lule : un système à trois niveaux ..... p. 612
89 iPhocomp : calcul automatique de l'indice de complexité phonétique de Jakielski ..... p. 622
90 Influence du développement phonético-phonologique sur les premières productions lexicales d'enfants arabophones : une étude longitudinale ..... p. 631
95 L'effet des contextes segmentaux et prosodiques sur la glottalisation de la voyelle initiale de mot en persan ..... p. 640
97 Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français ..... p. 658
99 L'identification vocale : pour une quantification des effets de la familiarité ..... p. 676
101 A la recherche des traces dialectales dans l'arabe standard : production des voyelles et des fricatives inter-dentales par des locuteurs tunisiens et marocains. ..... p. 684

Session orale 3 : Voix

4 Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques ..... p. 40
21 Application du critère BIC pour la segmentation en tours de chant ..... p. 166
37 Caractérisation et classification automatique des modes phonatoires en voix chantée ..... p. 293
70 Étude des voyelles et de la force de voix par analyse discriminante ..... p. 504

Session orale 4 : Développement et apprentissage

31 Perception de la parole rapide chez les enfants présentant une dysphasie expressive ..... p. 248
50 Capacités d'apprentissage phonétique chez les sujets âgés francophones ..... p. 385
75 Production et planification de la parole spontanée en contexte narratif chez l'enfant âgé de 3 à 11 ans p. 531

Session poster 3

9 Quantificateur vectoriel à divisions par classification : Application au codage des fréquences spectrales d'immitance en large-bande ..... p. 75
11 Tension gestuelle en co-présence d'un accent d'intensité ..... p. 92
15 Caractérisation acoustico-phonétique de parole provenant de patients atteints de paralysie faciale p. 121
16 Roots : un outil pour manipuler facilement efficacement et avec cohérence des corpus annotés de séquences ..... p. 130
19 Le couple YASP/YADE – Yet Another Semantic Parser/Yet Another Dialogue Engine ..... p. 148
20 Etude préliminaire de l'influence des caractéristiques phonétiques sur le bégaiement : le cas du français et du slovaque ..... p. 157
22 Estimation de la source glottique par décomposition modale empirique ..... p. 175
25 Perception et Réalisation du /R/ standard français en finale de mot ..... p. 193
27 Perturbations du Voice Onset Time (V.O.T) et du Voice Termination Time (V.T.T.) dans des séquences VCV produites par des patients francophones souffrant de paralysies récurrentielles ..... p. 211
45 Étude des productions des consonnes du français par des apprenants vietnamiens du Français Langue Etrangère ..... p. 348
48 Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques ..... p. 366
51 Régression ordinale pour la prédiction de la qualité d'interaction ..... p. 394
54 Train&Align : un outil d'alignement phonétique automatique disponible en ligne ..... p. 412
55 Aspects de la coordination pneumophonique dans la dysarthrie parkinsonienne ..... p. 421
76 Modélisation de l'intonation interlangue : le cas des questions ..... p. 540
77 Contribution à l'étude physiologique des possibilités articulatoires dans la cavité pharyngale p. 549
78 Coordination/synchronisation gestes-voix dans la démence de la maladie d'Alzheimer en voix parlée et chantée : une étude de cas ..... p. 558
104 Combinaison de mots et de syllabes pour transcrire la parole ..... p. 703
107 Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques ..... p. 720
108 Catégorisation multi-thématique de dialogues téléphoniques ..... p. 728
109 Analyse factorielle pour une représentation vectorielle des états des Modèles de Markov Cachés p. 737

Session orale 5 : Transcriptions et traitements automatiques

8 Quand l'horloge parlante a beaucoup à raconter sur l'évolution des techniques d'archivage audiovisuel ..... p. 67
38 Reconnaissance automatique de la parole à l'aide de colonies de fourmis ..... p. 302
58 Exploitation de la distribution des locuteurs pour la segmentation thématique de journaux télévisés p. 439
98 Détection et caractérisation d'erreurs dans des transcriptions automatiques pour des systèmes de traduction parole-parole ..... p. 667

Session orale 6 : Langues régionales ou peu dotées

36 Modélisation acoustico-phonétique de langues peu dotées : Études phonétiques et travaux de reconnaissance automatique en luxembourgois ..... p. 284
69 Les occlusives de l'alsacien : une étude temporelle ..... p. 494
71 Phonétisation automatique du dialecte tunisien ..... p. 513
79 Les voyelles fermées tendues relâchées et allongées du français québécois : la contribution d'indices statiques/dynamiques et absolus/normalisés à la détermination de leur identité acoustique ..... p. 567

Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ?

Mohamed Morchid[†], Richard Dufour[†], Georges Linarès[†],
Renato de Mori[†‡], *Fellow, IEEE* *
[†]Laboratoire Informatique d'Avignon (LIA), Université d'Avignon, France
[‡]McGill University, School of Computer Science, Montréal, Québec, Canada
prénom.nom@univ-avignon.fr, rdemori@cs.mcgill.ca

RÉSUMÉ

Les systèmes de reconnaissance automatique de la parole (RAP) sont désormais très performants. Néanmoins, la qualité de transcription est fortement dégradée dans des environnements très bruités, ce qui influe sur les performances des applications les utilisant, telles que les tâches de classification. Dans ce papier, nous proposons d'identifier les thèmes présent dans des services vocaux téléphoniques au moyen de l'approche classique à base de fréquences de mots (TF-IDF avec le critère de pureté Gini) et au moyen de l'approche à base d'espaces de thèmes (LDA). Ces deux représentations sont ensuite utilisées dans un processus de classification utilisant les SVM afin de retrouver le thème présent dans la conversation. Enfin, nous proposons de discuter autour de la qualité, en termes de taux d'erreur-mot, des mots identifiés comme discriminants et non-discriminants par les méthodes de représentation des dialogues étudiées dans cet article.

ABSTRACT

Classification of highly imperfect automatic transcriptions : Should we adapt the word error rate ?

Although the current transcription systems could achieve high recognition performance, they still have a lot of difficulties to transcribe speech in very noisy environments. The transcription quality has a direct impact on classification tasks using text features. In this paper, we propose to identify themes of telephone conversation services with the classical Term Frequency-Inverse Document Frequency using Gini purity criteria (TF-IDF-Gini) method and with a Latent Dirichlet Allocation (LDA) approach. These approaches are coupled with a Support Vector Machine (SVM) classification to resolve theme identification problem. Results show the effectiveness of the proposed LDA-based method compared to the classical TF-IDF-Gini approach in the context of highly imperfect automatic transcriptions. Finally, we discuss the impact of discriminative and non-discriminative words extracted by both methods in terms of transcription accuracy.

MOTS-CLÉS : classification automatique, TF-IDF, LDA, taux d'erreur-mot.

KEYWORDS: document indexing, TF-IDF, LDA, word error rate.

*. Ce travail a été réalisé dans le cadre des projets SUMACC et DECODA financés par l'Agence Nationale de la Recherche (ANR) sous les contrats ANR-10-CORD-007 et ANR-09-CORD-005.

1 Introduction

L'application étudiée dans cet article concerne l'analyse automatique de conversations téléphoniques entre un agent et un client dans le cadre du service vocal d'aide aux usagers de la RATP (*Régie Autonome des Transports Parisiens*). L'agent doit suivre un protocole particulier pendant la conversation afin de répondre aux requêtes ou plaintes des usagers. Ce papier présente un système permettant l'extraction automatique de thèmes à partir de conversations téléphoniques. Un unique thème est alors proposé pour chaque conversation. Dans cette tâche particulière, les conversations peuvent contenir des segments très bruités, ce qui a pour conséquence de fortement dégrader la qualité de leur transcription automatique au moyen d'un système de reconnaissance automatique de la parole (RAP).

Dans le domaine de la recherche d'information, la caractéristique principale utilisée est la *fréquence des mots*. Cette caractéristique spécifique permet d'obtenir un sous-ensemble de mots discriminants [1] pour une classe considérée (un *thème* dans cette étude). Au final, cet ensemble de mots discriminants doit permettre une représentation vectorielle des thèmes d'une conversation dans un espace sémantique.

Bien que la fréquence des mots soit une caractéristique performante dans le contexte de textes écrits ou transcrits manuellement, son application aux transcriptions automatiques semble être beaucoup plus difficile, et ce, à cause des erreurs de transcription. En effet, ces erreurs peuvent conduire à une représentation incorrecte des mots discriminants. Pour cette raison, la projection des mots transcrits automatiquement dans un espace de représentation plus abstrait pourrait améliorer la robustesse des applications face aux erreurs du système de RAP.

Nous proposons de comparer deux représentations non-supervisées des mots discriminants dans des transcriptions très imparfaites (*i.e.* taux d'erreur-mot très élevé) avec pour finalité d'identifier automatiquement les thèmes de conversations téléphoniques. La méthode classique à base de fréquences de mots (TF-IDF avec le critère de pureté Gini (Robertson, 2004)) est tout d'abord appliquée pour extraire les mots discriminants pour chaque thème à identifier. Nous proposons ensuite d'étudier une représentation des mots discriminants avec des espaces de thèmes au moyen de l'approche *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003). Chaque représentation est ensuite utilisée pour apprendre un classifieur de type *Support Vector Machine* (SVM) afin d'associer automatiquement un thème à une conversation. Enfin, nous proposons également dans cet article une discussion autour de l'impact sur les performances de classification des mots discriminants et non-discriminants choisis par les deux méthodes de représentation étudiées.

La partie suivante présente les travaux antérieurs réalisés dans ce domaine. La partie 3 introduit les méthodes de représentation des mots étudiées. Les résultats expérimentaux sont exposés dans la partie 4. Une discussion autour de la précision de la transcription des mots discriminants est proposée dans la partie 4.3, avant de conclure dans la partie 5.

2 Travaux antérieurs

Différents résumés de travaux récents ont été réalisés autour de l'analyse de conversations (Tur et De Mori, 2011), du traitement de la parole (Melamed et Gilbert, 2011), de l'identification (Hazen, 2011) et de la segmentation de thèmes (Purver, 2011). L'approche classique à base de fréquences de mots, *Term Frequency-Inverse Document Frequency* (TF-IDF) (Robertson, 2004), a

1. Le terme *discriminant* est associé à un mot s'il permet de choisir une classe par rapport à une autre.

été très largement utilisée afin d'extraire les mots discriminants contenus dans des textes. Des améliorations ont été constatées en lui associant le critère de pureté Gini (Dong *et al.*, 2011).

D'autres approches ont proposé de considérer le document comme une mixture de thèmes latents. Ces méthodes, telles que *Latent Semantic Analysis* (LSA) (Deerwester *et al.*, 1990; Bellegarda, 1997), *Probabilistic LSA* (PLSA) (Hofmann, 1999), ou *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003), permettent un niveau de représentation plus élevé des mots au moyen d'un espace de thèmes. Les documents sont alors considérés comme des sacs-de-mots (Salton, 1989) ; les performances de ces méthodes ont pu être démontrées sur de nombreuses tâches.

LDA est un modèle génératif qui considère qu'un thème est associé à chaque occurrence d'un mot composant le document, plutôt que d'associer un thème au document complet. Ainsi, un document peut changer de thèmes d'un mot à un autre. Cependant, les occurrences des mots sont connectées par une variable latente qui contrôle le respect global de la distribution des thèmes dans le document. Ces thèmes latents sont caractérisés par une distribution des probabilités des mots qui leur sont associées. Les modèles PLSA et LDA obtiennent généralement de meilleurs résultats que le modèle LSA sur les tâches de recherche d'information (Hofmann, 2001). De plus, LDA fournit une estimation directe de la pertinence d'un thème sachant un ensemble de mots.

Les machines à vecteurs de support, *Support Vector Machines* (SVM), sont un ensemble de techniques d'apprentissage supervisé. Sachant un échantillon, les SVM déterminent un plan séparateur entre les parties de l'échantillon appelé *vecteur de support*. Ensuite, un hyperplan séparateur maximisant la *marge* entre les vecteurs de support et l'hyperplan séparateur (Vapnik, 1963) est calculé. Les SVM ont été utilisés pour la première fois par (Boser *et al.*, 1992), pour des tâches de régression (Müller *et al.*, 1997) et de classification (Joachims, 1999). La popularité des SVM est due aux bons résultats atteints dans ces deux tâches et au faible nombre de paramètres nécessitant un ajustement.

Une approche à base de LDA, combinée avec une classification SVM, a été récemment étudiée dans de nombreux domaines, tels que la classification de textes (Zrigui *et al.*, 2012), la stylométrie (Arun *et al.*, 2009), la recherche d'information (Kim *et al.*, 2009), la détection d'évènements sociaux (Morchid *et al.*, 2013), ou la détection d'images (Tang *et al.*, 2009). À notre connaissance, une approche LDA-SVM n'a jamais été appliquée à la classification de thèmes au moyen de transcriptions automatiques très imparfaites, mais a été utilisée dans le contexte d'extraction de mots clés dans des transcriptions automatiques (Sheeba et Vivekanandan, 2012). La méthode TF-IDF couplée avec une classification SVM, qui constitue notre système de base, a été très largement étudiée dans le contexte de la classification de textes, comme par exemple dans (Lan *et al.*, 2005; Georgescul *et al.*, 2006).

3 Système d'identification de thèmes

Cette partie présente le système de classification de thèmes proposé utilisant les mots discriminants extraits à partir de transcriptions très imparfaites. Le système est composé de deux parties principales. La première crée une représentation vectorielle des mots au moyen de deux approches non-supervisées : un vecteur de fréquences de mots Okapi/BM25 (Robertson, 2004) avec la méthode TF-IDF-Gini (Dong *et al.*, 2011), et une représentation par espace de thèmes avec l'approche LDA (Blei *et al.*, 2003). La seconde partie utilise les vecteurs extraits afin d'apprendre des classifieurs SVM. La figure 1 présente l'architecture globale du système de classification proposé utilisant des transcriptions manuelles (MAN) et automatiques (RAP).

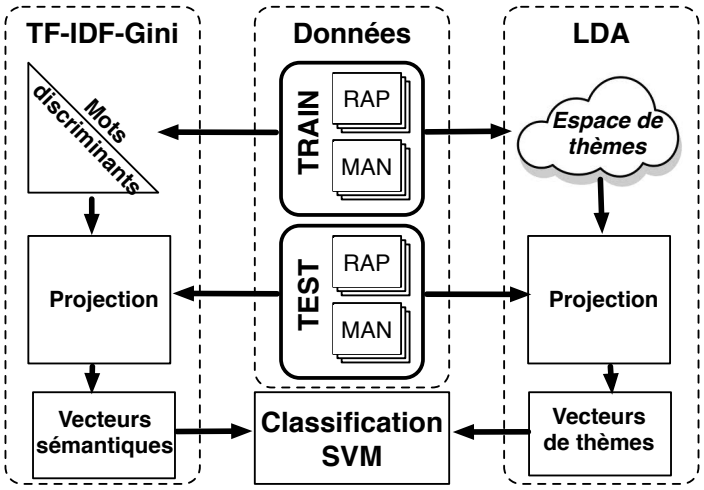


FIGURE 1 – Architecture globale du système de classification.

3.1 Représentation des dialogues

3.1.1 Représentation par fréquence des mots

Considérons un corpus D de dialogues d ayant un vocabulaire $\mathbf{V} = \{w_1, \ldots, w_N\}$ de taille N où d est vu comme un sac-de-mots (Salton, 1989). Un mot de $\mathbf{V}$ est choisi en fonction de son importance δ_t dans le thème t en calculant sa fréquence (TF), sa fréquence inverse (IDF) (Robertson, 2004), et le critère de pureté Gini (Dong *et al.*, 2011) commun pour tous les thèmes. Cet ensemble de scores δ compose le modèle de fréquence f :

$$\delta_t^w = tf_t(w) \times idf_t(w) \times gini_t(w) \tag{1}$$

Ensuite, les mots ayant les scores les plus élevés Δ pour tous les thèmes $\mathbf{T}$ sont extraits et constituent le sous-ensemble de mots discriminants $\mathbf{V}_\Delta$ (chaque thème $t \in \mathbf{T}$ possède son propre score δ_t) et sa propre fréquence γ dans le modèle f :

$$\gamma_f^t = \frac{\#d \in t}{\#d \in D} \tag{2}$$

Notons qu'un même mot w peut être présent dans différents thèmes, mais avec des scores différents (TF-IDF-Gini) en fonction de sa pertinence dans le thème :
$\Delta(w) = P(w|f) = \int_t P(w|t)P(t|f)\,\mathrm{dt}$
$= \sum_{t \in \mathbf{T}} P(w|t)P(t|f)$
$= \sum_{t \in \mathbf{T}} \delta_t^w \times \gamma_f^t$
$= \overrightarrow{\delta^w}\overrightarrow{\gamma^f}$
Pour chaque dialogue $d \in D$, un vecteur de caractéristiques sémantiques V_d^s est déterminé. La n^{ieme} $(1 \leq n \leq |\mathbf{V}_\Delta|)$ caractéristique $V_d^s[n]$ contient le nombre d'occurrences du mot w_n $(|w_n|)$ dans d, et le score Δ de w_n (voir équation 3.1.1) dans l'ensemble des mots discriminants $\mathbf{V}_\Delta$:

$$V_d^s[n] = |w_n| \times \Delta(w_n) \tag{3}$$

3.1.2 Représentation par espace de thèmes

La représentation par espace de thèmes est réalisée au moyen de l'approche LDA (voir partie 2). Un espace thématique m de n thèmes est alors obtenu avec, pour chaque thème z, la probabilité de chaque mot w de $\mathbf{V}$ sachant z ($P(w|z) = V_z^w$), et pour le modèle complet m, la probabilité de chaque thème z sachant le modèle m ($P(z|m) = V_m^z$).

Pour chaque dialogue d du corpus D, un premier paramètre θ est défini en fonction d'une loi de Dirichlet de paramètre α. Un second paramètre ϕ est défini en fonction de la même loi de Dirichlet de paramètre β. Ensuite, pour générer tous les mots w du document d, un thème latent z est défini à partir d'une distribution multinomiale sur θ. Sachant ce thème z, la distribution des mots est une multinomiale de paramètre ϕ. Le paramètre θ est défini pour tous les documents à partir du même paramètre initial α. Cela permet d'obtenir un paramètre reliant tous les documents ensemble (Blei *et al.*, 2003).

Projection des conversations/espace de thèmes : L'algorithme de Gibbs sampling (Griffiths et Steyvers, 2002) est utilisé pour inférer un dialogue d avec les n thèmes de l'espace thématique m. Cet algorithme s'appuie sur la méthode *Markov Chain Monte Carlo* (MCMC). Ainsi, le Gibbs sampling permet d'obtenir des échantillons des paramètres de distribution θ sachant un mot w d'un document de test et un thème donné z. Un vecteur de caractéristiques V_z^d de la représentation du thème de d est alors obtenu. La k^{ieme} caractéristique (où $1 \leq k \leq n$) est la probabilité du thème z_k sachant le dialogue d :

$$V_z^d[k] = P(z_k|d) \tag{4}$$

3.2 Classification à base de SVM

Durant cette étape, les classifieurs sont entrainés à partir de la représentation vectorielle des mots afin d'attribuer automatiquement le thème le plus pertinent à chaque conversation. Ce processus de classification nécessite un classifieur multi-classes. Pour ce problème multi-thème, T représente le nombre de thèmes et $t_i, i = 1, \ldots, T$ représente les thèmes. Un classifieur binaire (méthode *un-contre-un*) est utilisé avec un noyau linéaire (Yuan *et al.*, 2012) pour chaque paire de thèmes distinct : tous les classifieurs binaires $T(T-1)/2$ sont ensuite construits ensemble. Le classifieur binaire $C_{i,j}$ est entrainé, t_i étant une classe positive et t_j une classe négative ($i \neq j$). Une fois le vote de tous les classifieurs achevé, du thème ayant le plus grand nombre de votes est attribué au dialogue d.

4 Expériences

4.1 Protocole expérimental

Les expériences sur l'identification du thème d'une conversation sont menées sur le corpus du projet DECODA (Bechet *et al.*, 2012). Ce corpus est composé de 1 076 conversations téléphoniques découpées en un corpus d'apprentissage (740 dialogues) et un corpus de test (327 dialogues). Ces dialogues ont été manuellement annotés selon 8 thèmes : *problème d'itinéraire*, *objet perdu et trouvé*, *horaire*, *carte de transport*, *état du traffic*, *prix du ticket*, *infraction* et *offre spéciale*.

L'ensemble d'apprentissage est utilisé pour définir un sous-ensemble de mots discriminants (voir partie 3.1). Ce sous-ensemble permet d'élaborer un espace sémantique pour chaque conversation du corpus de test au moyen de la méthode basique TF-IDF-Gini. Dans ces expériences, le nombre de mots discriminants a été varié de 800 à 7 920 mots (nombre total de mots de l'apprentissage). Le corpus de test contient 3 806 mots (70,8 % étant contenu dans le corpus d'apprentissage).

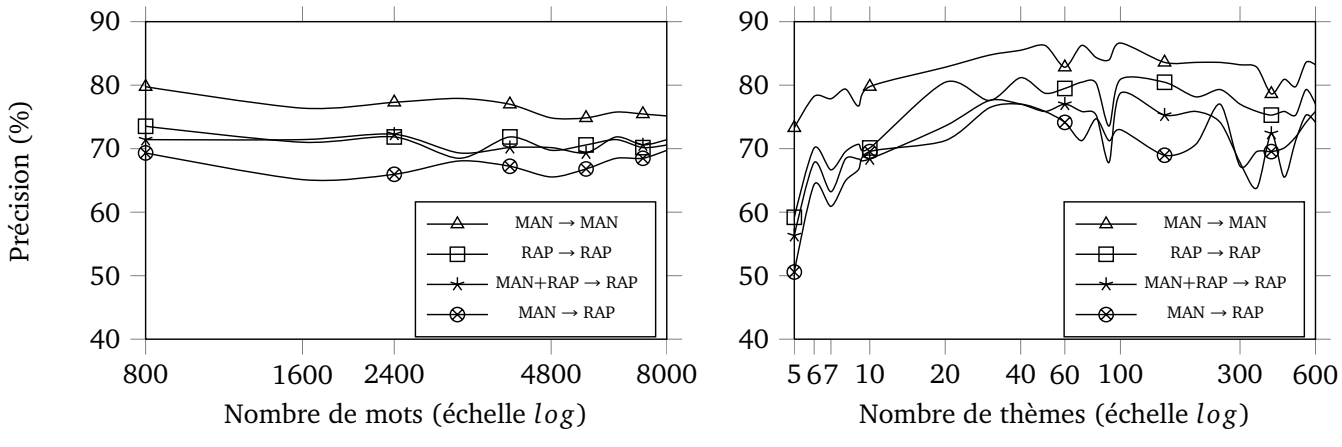
De la même manière, un vecteur de thèmes est calculé en projetant chaque dialogue du corpus de test dans chaque espace de thèmes. Un ensemble de 25 espaces de thèmes associé à un nombre de thèmes différents ($\{5, \ldots, 600\}$) est élaboré au moyen d'un modèle LDA construit à partir du corpus d'apprentissage avec l'outil Mallet (McCallum, 2002).

Ensuite, pour ces deux configurations, un classifieur SVM est entrainé au moyen de la librairie LIBSVM (Chang et Lin, 2011). Les paramètres sont optimisés par validation croisée sur le corpus d'apprentissage.

Le système de RAP Speeral (Linarès *et al.*, 2007) a été utilisé. Les paramètres des modèles acoustiques (230 000 gaussiennes / modélisation triphone) sont estimés au moyen d'une adaptation par maximum *a-posteriori* (MAP) à partir de 150 heures de parole (conditions téléphoniques). Un modèle de language tri-gramme a été obtenu en adaptant un modèle de language basique avec les transcriptions du corpus d'apprentissage de DECODA. Le vocabulaire contient 5 782 mots. Le taux d'erreur-mot (TEM) initial atteint 45,8 % sur le corpus d'apprentissage et 58,0 % sur le corpus de test. Ces TEM élevés sont principalement dus à la présence de nombreuses disfluences, et à des conditions acoustiques bruitées, quand, par exemple, les utilisateurs appellent à partir de gares avec un téléphone portable. Une liste de rejet de 126 mots [2] a été utilisée, ce qui donne au final un TEM de 33,8 % (apprentissage), et 49,5 % sur l'ensemble (test).

Les expériences sont menées au moyen des deux méthodes non-supervisées proposées (TF-IDF-Gini / LDA) sur les transcriptions manuelles (MAN) et les transcriptions automatiques (RAP) seules. Nous proposons également d'étudier la combinaison des transcriptions manuelles et automatiques (MAN+RAP) afin de voir si les erreurs de RAP peuvent être compensées par les mots corrects (*i.e.* issus de la référence).

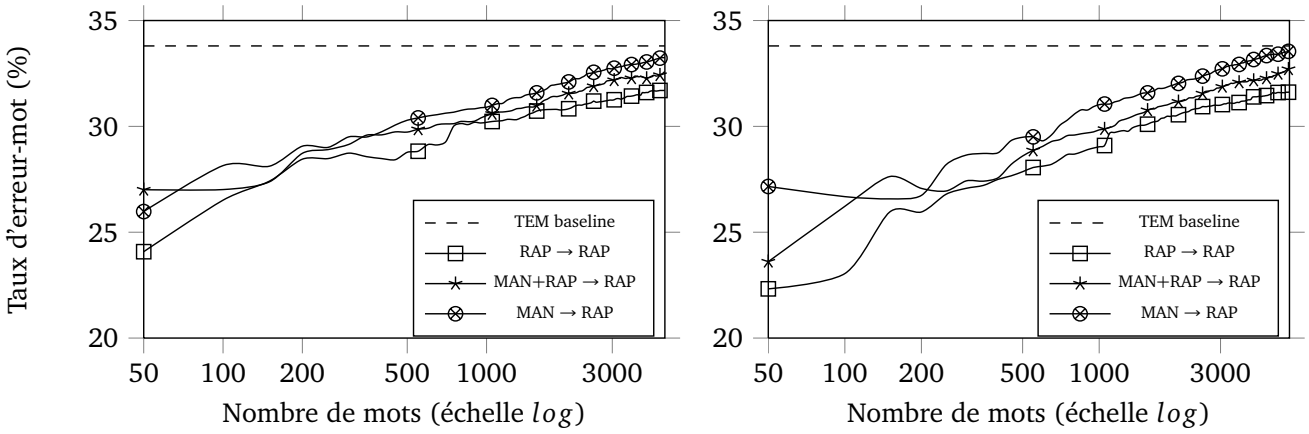
FIGURE 2 – Performance de la classification de thèmes en faisant varier le nombre de mots discriminants (a) et le nombre d'espaces de thèmes (b).



(a) **TF-IDF-Gini** (b) **LDA**

2. http ://code.google.com/p/stop-words/

FIGURE 3 – Taux d'erreur-mot des n mots discriminants extraits avec TF-IDF-Gini (a) et LDA (b).



(a) **TF-IDF-Gini** (b) **LDA**

4.2 Performance de l'identification de thèmes

La figure 2 présente les précisions de la classification de thèmes obtenues par les approches TF-IDF-Gini et LDA sur le corpus de test pour les différentes configurations étudiées (MAN / RAP) en faisant varier les conditions d'extraction des mots (nombre de mots discriminants et nombre de thèmes). Nous pouvons noter que la méthode LDA surpasse tous les résultats obtenus par l'approche TF-IDF-Gini (voir tableau 1).

TABLE 1 – Précision de la classification de thèmes.

Données		*Meilleure Précision (%)*			
Apprentissage	**Test**	**#mots**	**TF-IDF-Gini**	**#thèmes**	**LDA**
MAN	MAN	*800*	79,7	*100*	**86,6**
MAN	RAP	*8 000*	69,7	*40*	**77,0**
RAP	RAP	*800*	73,5	*60*	**81,4**
RAP+MAN	RAP	*2 400*	72,2	*100*	**78,7**

Comme attendu, la configuration *MAN → MAN* donne les meilleurs résultats de classification avec un gain de 6,9 points avec la méthode LDA. Si nous comparons les configurations du corpus d'apprentissage, nous notons également que les meilleurs performances sur le corpus de test RAP sont obtenues avec le corpus d'apprentissage RAP. Un gain de 10,9 points est constaté avec la méthode LDA en comparaison de l'approche TF-IDF-Gini sur les transcriptions automatiques. Il semble évident qu'utiliser des configurations d'apprentissage et de test comparables permet d'atteindre les meilleurs résultats de classification, et ce, peu importe que l'on traite des transcriptions manuelles ou automatiques.

Nous pouvons enfin noter que les performances obtenues avec l'approche LDA ont tendance à fluctuer lorsque le nombre de thèmes varie. Ceci peut s'expliquer par les taux d'erreur-mot (TEM) du corpus traité : en effet, les mots choisis comme *discriminants* dans des conditions particulières de l'espace de thèmes peuvent être mal transcrits dans des proportions élevées. Nous pouvons étayer cette remarque en analysant les résultats obtenus avec 90 thèmes (voir figure 2). Une baisse importante des performances est observée pour la condition d'apprentissage RAP (RAP → RAP et MAN → RAP) alors qu'une faible baisse est constatée lorsque les transcriptions de références sont ajoutées dans le processus d'apprentissage (RAP+MAN → RAP et MAN → MAN).

4.3 Précision de la transcription des mots discriminants

Alors que l'approche TF-IDF-Gini est clairement meilleure sur les transcriptions manuelles (tableau 1), les performances sont quasiment identiques sur les transcriptions manuelles et automatiques avec la méthode LDA (avec une précision respective de 86,6 % et 81,4 %). Nous pensons que l'approche LDA doit mieux gérer les erreurs contenues dans les transcriptions automatiques en choisissant les mots discriminants les mieux transcrits (*i.e.* ayant le plus faible TEM). La figure 3 permet de comparer les taux d'erreur-mot (TEM) des n mots discriminants extraits au moyen des méthodes TF-IDF-Gini et LDA sur toutes les configurations (MAN / RAP). Le score $s(w)$ est utilisé pour trouver les mots les plus pertinents de l'approche LDA selon la formule :

$$\mathrm{s(w)} = \mathrm{P(w|m)} = \int_z P(w|z)P(z|m)\,\mathrm{dz}$$
$$= \sum_{z \in m} P(w|z)P(z|m)$$
$$= \sum_{z \in m} V_z^w \times V_m^z$$
$$= \overrightarrow{V^w}\overrightarrow{V^m}$$

où $\overrightarrow{V^w}$ est la représentation vectorielle d'un mot w dans tous les thèmes z de l'espace de thèmes m, $\overrightarrow{V^m}$ est la représentation vectorielle de tous les thèmes z dans m et $\cdot\cdot$ est le produit scalaire. Le TEM est ensuite calculé sur les n mots les plus discriminants (poids de 1 pour chaque mot).

Si nous comparons tout d'abord les différentes configurations (MAN / RAP), nous pouvons noter que plus la précision de la classification est élevée (tableau 1), moins le TEM l'est. Ce constat est observé pour les deux méthodes. De plus, nous pouvons voir que le TEM obtenu avec l'approche LDA est légèrement plus bas que celui obtenu avec la méthode TF-IDF-Gini, peu importe la configuration considérée. Cela signifie que les mots discriminants extraits avec l'approche LDA sont mieux transcrits en comparaison de ceux obtenus avec la méthode TF-IDF-Gini, ce qui peut expliquer les meilleures performances de classification obtenues avec l'approche LDA.

5 Conclusions

Dans cet article, nous avons présenté une architecture permettant d'identifier le thème d'une conversation en utilisant des transcriptions très imparfaites. Deux méthodes non-supervisées de représentation des conversations (TF-IDF-Gini et LDA) ont été comparées. Nous avons montré que la représentation par espace de thèmes obtenue avec la méthode LDA surpasse les résultats de classification obtenus avec la représentation classique TF-IDF-Gini. La précision de la classification atteint 86,6 % sur les transcriptions manuelles et 81,4 % sur les transcriptions automatiques, avec un gain respectif de 6,9 et 10,9 points.

Nous avons également discuté du lien possible entre performance de classification et précision de la transcription. L'analyse proposée a montré que les meilleurs résultats de classification sont obtenus avec des configurations extrayant les mots discriminants ayant les taux d'erreur-mot les plus faibles. Ces résultats prometteurs conduiront à une analyse qualitative plus détaillée dans des travaux futurs. En effet, cette étude préliminaire pourrait être fortement étendue avec de nouvelles analyses, en prenant par exemple en compte le poids des mots discriminants dans l'évaluation de la précision des transcriptions. Une perspective générale serait de proposer une solution pour estimer les performances de classification selon la qualité des transcriptions. Dans un contexte lié aux métriques d'évaluation, il serait intéressant de trouver une façon d'estimer la précision des transcriptions automatiques sur des tâches spécifiques, le taux d'erreur-mot n'étant pas un bon indicateur de la qualité d'une transcription dans un cadre applicatif.

Références

Arun, R., Saradha, R., Suresh, V., Narasimha Murty, M. et Veni Madhavan, C. E. (2009). Stopwords and Stylometry : A Latent Dirichlet Allocation Approach. *In NIPS Workshop on Applications for Topic Models : Text and Beyond*, Canada.

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Beze, M., De Mori, R. et Arbillot, E. (2012). Decoda : a call-centre human-human spoken conversation corpus. LREC'12.

Bellegarda, J. (1997). A latent semantic analysis framework for large-span language modeling. *In Fifth European Conference on Speech Communication and Technology*.

Blei, D., Ng, A. et Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Boser, B., Guyon, I. et Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *In 5th annual workshop on Computational learning theory*, pages 144–152.

Chang, C.-C. et Lin, C.-J. (2011). Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Dong, T., Shang, W. et Zhu, H. (2011). An improved algorithm of bayesian text categorization. *Journal of Software*, 6(9):1837–1843.

Georgescul, M., Clark, A. et Armstrong, S. (2006). Word distributions for thematic segmentation in a support vector machine approach. *In Conference on Computational Natural Language Learning*.

Griffiths, T. et Steyvers, M. (2002). A probabilistic approach to semantic representation. *In 24th annual conference of the cognitive science society*, pages 381–386. Citeseer.

Hazen, T. (2011). Topic identification. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pages 319–356.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *In Proc. of Uncertainty in Artificial Intelligence*, page 21.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *In Machine learning-international workshop then conference*, pages 200–209.

Kim, S., Narayanan, S. et Sundaram, S. (2009). Acoustic topic model for audio information retrieval. *In Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 37–40.

Lan, M., Tan, C.-L., Low, H.-B. et Sung, S.-Y. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. *In International Conference on World Wide Web*, pages 1032–1033.

Linarès, G., Nocéra, P., Massonie, D. et Matrouf, D. (2007). The lia speech recognition system : from 10xrt to 1xrt. *In Text, Speech and Dialogue*, pages 302–308. Springer.

McCallum, A. (2002). Mallet : A machine learning for language toolkit.

Melamed, I. et Gilbert, M. (2011). Speech analytics. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pages 397–416.

Morchid, M., Dufour, R. et Linarès, G. (2013). Event detection from image hosting services by slightly-supervised multi-span context models. *In CBMI'13*.

Müller, K., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J. et Vapnik, V. (1997). Predicting time series with support vector machines. *ICANN'97*, pages 999–1004.

Purver, M. (2011). Topic segmentation. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pages 291–317.

Robertson, S. (2004). Understanding inverse document frequency : on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520.

Salton, G. (1989). Automatic text processing : the transformation. *Analysis and Retrieval of Information by Computer*.

Sheeba, J. I. et Vivekanandan, K. (2012). Article : Improved keyword and keyphrase extraction from meeting transcripts. *International Journal of Computer Applications*, 52(13):11–15.

Tang, S., Li, J., Zhang, Y., Xie, C., Li, M., Liu, Y., Hua, X., Zheng, Y.-T., Tang, J. et Chua, T.-S. (2009). Pornprobe : an lda-svm based pornography detection system. *In International Conference on Multimedia*.

Tur, G. et De Mori, R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*.

Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.

Yuan, G.-X., Ho, C.-H. et Lin, C.-J. (2012). Recent advances of large-scale linear classification. 100(9):2584–2603.

Zrigui, M., Ayadi, R., Mars, M. et Maraoui, M. (2012). Arabic text classification framework based on latent dirichlet allocation. *CIT*, 20(2):125–140.

Contribution à la modélisation de la phonation parkinsonienne : contraintes physiologiques et linguistiques

Alain Ghio[1], Marielle Mas[2], Cécile Grigoli[2], Céline Delooze[1], Danièle Robert [1, 2], Céline Mercier[3], François Viallet[1,3]

(1) Université d'Aix-Marseille, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(2) Service ORL, CHU de la Timone, Marseille, France
(3) Service de neurologie, CH du Pays d'Aix, Aix-en-Provence, France

alain.ghio@lpl-aix.fr

RESUME

Des caractéristiques importantes de la parole dans la maladie de Parkinson (MDP) sont un trouble phonatoire et/ou une déstructuration mélodique. La fréquence fondamentale (F0) apparait donc comme un paramètre important à étudier dans ce cadre. Or, la littérature laisse apparaître des résultats contradictoires : élévation, baisse ou pas de différence de F0 dans la MDP par rapport à une population de sujets contrôle. Ces contradictions sont liées à la non prise en compte de modèle de phonation dans la MDP. L'objet de ce travail est donc de fournir un cadre conceptuel et méthodologique permettant d'expliquer les résultats obtenus sur la F0 de 44 locuteurs (29 patients MDP et 15 sujets témoins) notamment en contrôlant des contraintes physiologiques (sexe, âge, état pharmacologique des patients) et linguistiques (tâches de production de parole).

ABSTRACT

An important characteristic of speech in Parkinson's disease (PD) is a vocal disorder and/or a melodic dysfunction. The fundamental frequency (F0) thus appears as an important parameter to study in this context. However, the literature reveals conflicting results: increased, decreased or no difference of F0 in PD speakers compared to healthy speakers. These contradictions are related to a lack of model in PD dysarthria. The purpose of this work is to provide a conceptual and methodological framework to explain the results of 44 speakers (29 patients and 15 control subjects) including physiological constraints (gender, age, pharmacological status of patients) and language constraints (speaking tasks).

MOTS-CLES : dysarthrie, dysphonie, parole parkinsonienne, modèle de production de parole

KEYWORDS: dysarthria, dysphonia, Parkinson's Disease Speech, speech production model

1 Introduction

Le processus de communication parlée est soumis à des contraintes linguistiques et physiologiques aussi bien chez le locuteur que chez l'auditeur. Il est difficile d'appréhender globalement de façon naturelle et spontanée le jeu de contraintes qui opère. L'analyse de la parole en situation dysfonctionnelle est un cadre intéressant dans la mesure où, la plupart du temps, les aspects physiologiques sont perturbés par la maladie et parfois rétablis par compensation spontanée chez le patient ou par des moyens thérapeutiques. Dans cette perspective, le cas de la maladie de Parkinson (MDP) est un cadre privilégié dans la mesure où (1) la maladie, son avancement et ses symptômes sont bien décrits ; (2) il existe des moyens thérapeutiques comme la prise de L-dopa pour atténuer les symptômes ; (3) les atteintes au niveau de la parole la laissent relativement intelligible et autorise ainsi l'analyse, sauf dans les cas très avancés de dysarthrie ; (4) il est relativement facile d'enregistrer de tels patients dans notre cas (Service de neurologie du Dr Viallet, CH du Pays d'Aix, Aix-en-Provence)

2 La dysarthrie parkinsonienne

Dans la maladie de Parkinson (MDP), il est généralement admis que les symptômes prédominants au niveau de la parole sont la dysphonie (trouble phonatoire) et la dysprosodie (trouble mélodique ou rythmique) (Pinto et al., 2010). L'hypophonie, qui consiste en une diminution de l'intensité vocale, est la caractéristique la plus souvent citée et l'intensité SPL apparaît donc comme un paramètre important à mesurer. Mais le trouble phonatoire et/ou la déstructuration prosodique, notamment mélodique, peut surtout être analysé à travers la fréquence fondamentale (F0). Le modèle physiopathologique de Goberman & Coelho (2002) met en relation l'anatomie, les déficits moteurs, les corrélats physiologiques et les mesures observables dans la maladie de Parkinson. Au niveau phonatoire (Figure 1), il prédit notamment une élévation de la F0 conséquente aux tensions laryngées liées à la rigidité parkinsonienne.

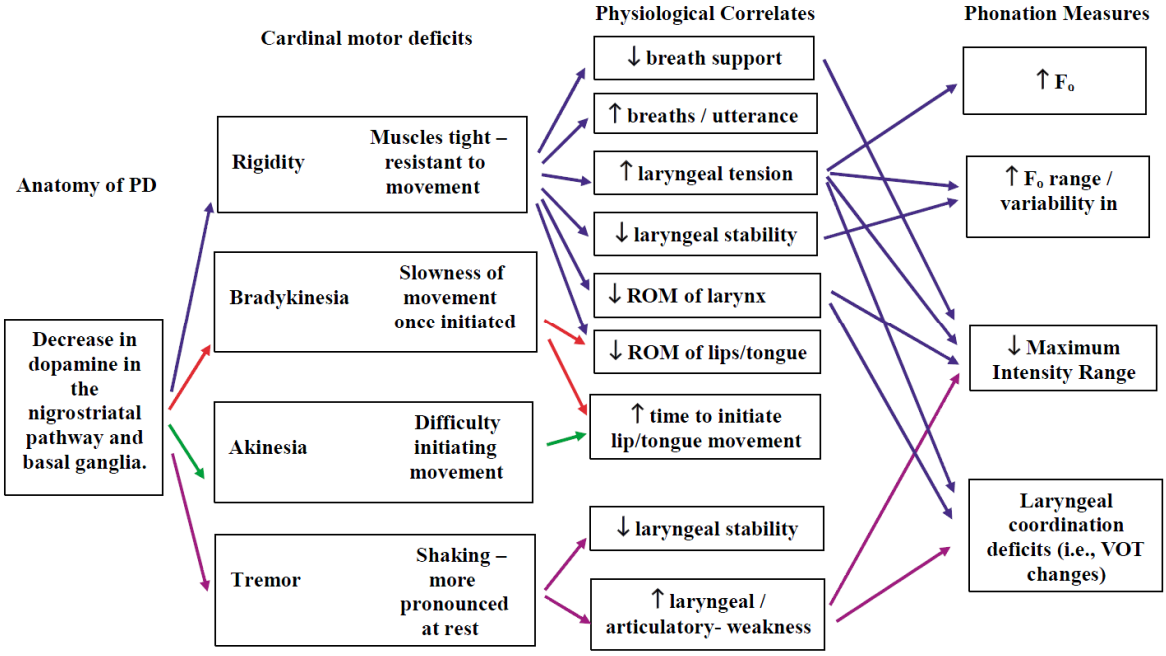


FIGURE 1 – Modèle de phonation parkinsonnienne d'après Goberman & Coelho (2002)

Or, une revue de la littérature sur cette thématique (Robert & Spezza, 2005) complétée par une analyse de publications récentes (Mas & Grigoli, 2012) laisse apparaître des résultats contradictoires sur les changements de la hauteur de voix liés à la MDP : élévation, baisse ou pas de différence de F0 par rapport à une population de sujets contrôle. Plusieurs facteurs peuvent expliquer ces résultats contradictoires:

1. le sexe des locuteurs (il n'est pas certain que les symptômes vocaux liés à la maladie soient identiques chez les hommes et les femmes)
2. l'âge des locuteurs : effet du vieillissement sur la voix et notamment sur la F0 (Fig.2)
3. l'état pharmacologique du patient (avec ou sans traitement à la L-dopa)
4. la tâche de production de parole (voyelle tenue vs lecture vs parole spontanée)

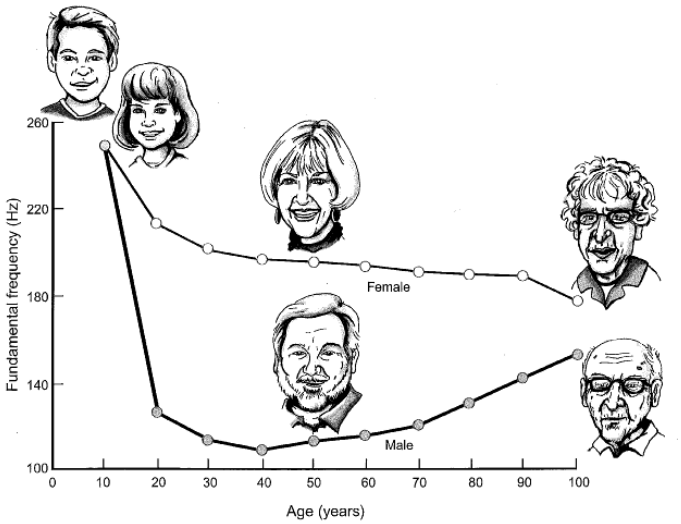


FIGURE 2 – Evolution de la F0 selon l'âge et le sexe (Hixon et al., 2008)

Pour ne prendre en compte que le facteur 'sexe' et 'âge', la Figure 2 illustre parfaitement le fait que mesurer la F0 sur des hommes sains d'environ 40 ans et des locuteurs MDP d'environ 75 ans inclut deux facteurs confondus (tranche d'âge et maladie) pouvant rendre ininterprétables les analyses statistiques : une différence significative entre les groupes serait-elle due à l'âge ou à la maladie ? De même que des contraintes physiologiques peuvent influencer le registre vocal du locuteur, diverses études montrent aussi que le style ou la tâche de parole peut avoir un effet sur le registre moyen. Par exemple, Swertz et al. (1996) ont observé qu'en lecture, la F0 est généralement plus élevée que sur de la parole spontanée.

3 Hypothèses et protocole expérimental

Notre étude a consisté à analyser la F0 moyenne de locuteurs sains et MDP en maitrisant les 4 facteurs précédemment cités. Deux d'entre eux ont été neutralisés en focalisant l'étude sur une population uniquement masculine (prédominante dans la MDP) et en appariant nos patients à une population 'contrôle' d'âge non significativement différent. Nous avons pu ainsi étudier l'effet de la maladie, du traitement dopaminergique (contraintes physiologiques) et des styles de parole (contrainte linguistique) sur la fréquence fondamentale. En effet, les locuteurs ont été enregistrés dans différentes tâches de parole, telles que la production de voyelles tenues, la répétition de phrases, la lecture de texte et la description d'image. Ces tâches, classiques dans le cadre de l'évaluation en milieu clinique, se distinguent par leur aspect préparée (voyelles tenues, répétition, lecture) ou non préparée (description d'image).

Nos hypothèses étaient les suivantes :

H1. La F0 dans la MDP sera plus élevée que chez les témoins (modèle de Goberman)

H2. La F0 en situation ON sera moins élevée qu'en situation OFF (le médicament attenue les symptômes, en l'occurrence la rigidité =>la tension laryngée => la montée de F0)
H3. La F0 dans les tâches de parole préparée (/a/ tenu, répétition, lecture) sera plus élevée que dans les tâches non préparée (effet de performance sur les tâches préparées)

4 Corpus et méthode

4.1 Locuteurs

Notre corpus est constitué de 44 locuteurs (29 patients parkinsoniens + 15 sujets témoins) enregistrés dans le Service de Neurologie du Centre Hospitalier du Pays d'Aix à Aix-en-Provence (Dr Viallet). Les locuteurs sont des hommes âgés en moyenne de 70.6 ans ($\boldsymbol{\sigma}=9.6$) pour les patients parkinsoniens et de 70.3 ans pour les témoins ($\boldsymbol{\sigma}=11.8$). Il n'y a aucune différence significative au niveau de l'âge entre les deux groupes. Les patients parkinsoniens ont une durée moyenne de maladie de 9 ans ($\boldsymbol{\sigma}=5$). Leur score UPDRS (Unified Parkinson's disease rating scale) de déficit moteur en situation OFF (sevrage médicamenteux depuis au moins 12h) est de 30.6 ($\boldsymbol{\sigma}=10.2$), ce qui correspond à une maladie installée. Une heure après la prise de L-dopa (situation ON), le score UPDRS moyen est de 13.6 ($\boldsymbol{\sigma}=7.3$), ce qui correspond à une amélioration de plus de 50% par rapport à la situation OFF, amélioration notable qui valide la dépendance à la dopa des patients sélectionnés.

4.2 Enoncés

Tous les locuteurs ont été enregistrés sur :

- la production d'un /a/ tenu d'environ 3 secondes (aaa),
- la production d'un /a/ tenu le plus longtemps possible (a_max),
- la répétition de la phrase « papa ne m'a pas parlé de beau papa » (phrase_papa)
- la lecture du 1er paragraphe de la chèvre de M. Seguin (A. Daudet) à vitesse normale (lect_normale)
- idem à vitesse rapide (lect_rapide)
- la description spontanée d'une image (issue de la batterie d'évaluation HDAE)

4.3 Mesures

Les mesures de F0 ont été effectuées avec le logiciel Praat (www.fon.hum.uva.nl). Pour limiter la prise en compte de valeurs aberrantes de F0, nous avons utilisé l'algorithme proposé par De Looze et al. (2012) dans lequel la détection se fait en deux passes. Dans une première étape, les valeurs autorisées de F0 sont fixées à l'intervalle 60-600 Hz. Dans un second temps, cet intervalle de détection est ajusté à chaque locuteur en fonction des mesures calculées lors de la 1ère passe. La F0 minimale est ainsi ajustée à *q15*0,83* (où *q15* est le 15^{ième} percentile de la distribution). De même, la F0 maximale est ajsutée à *q65**1,92 (où *q65* est le 65 ^{ième} percentile de la distribution). Ces formules sont issues de résultats de corpus plus importants et permettent de réduire les erreurs.

Se pose ensuite la question du choix de la valeur synthétique de l'observation et notamment la préférence à la moyenne ou la médiane. Dans nos mesures, la médiane est significativement plus basse que la moyenne, preuve d'une asymétrie positive des distributions de F0. Mais il existe une forte corrélation entre les deux mesures *(R^2=0.97)*. Par la suite, nous utiliserons la moyenne calculée sur l'ensemble de l'énoncé choisi.

5 Résultats

Afin de réaliser les statistiques de notre étude, nous avons utilisé le logiciel 'R' version 2.14. Le modèle statistique utilisé est un modèle mixte à deux facteurs fixes :

- état_clinique à 3 états : OFF, ON, CTRL
- tache_de_parole : a_max, aaa, phrase_papa, lect_rapide, lect_normale, description

auquel s'ajoute un facteur aléatoire qui est le locuteur.

Nous avons considéré que les résultats étaient significatifs quand $p_{MCMC} < 0.05$.

Nous n'avons pas observé d'interaction significative entre les deux facteurs fixes. Nous pouvons donc décomposer les résultats par facteur.

5.1 Effet de la maladie et effet médicamenteux

Nous n'avons pas observé de différences significatives de la F0 moyenne chez les sujets parkinsoniens en sevrage médicamenteux (OFF ; $F0_{moy}$ =125 Hz) par rapport à la population contrôle (CTRL ; $F0_{moy}$ =130.1 Hz), ni entre les sujets parkinsoniens avec médicament (ON ; $F0_{moy}$ =134.7 Hz) et la population contrôle (Figure 3). En revanche, il existe une différence significative de F0 moyenne de 9.7 Hz entre la situation OFF et ON, ce qui correspond à une augmentation de 7.8% en ON, c'est à dire 1.3 demi-ton.

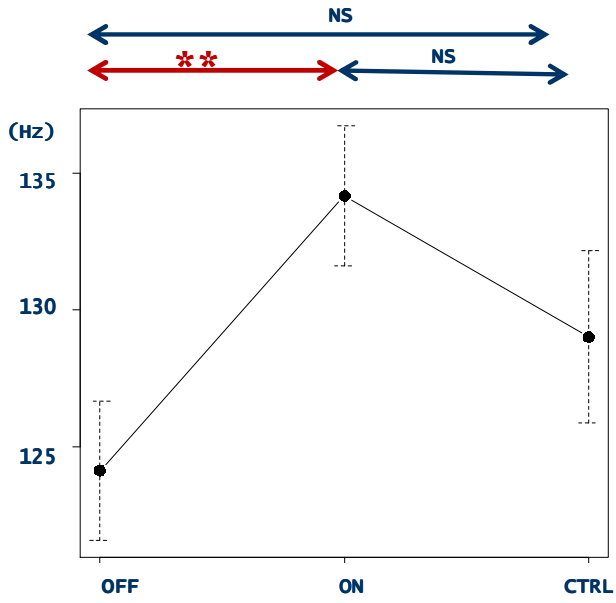


FIGURE 3 – F0 moyenne en Hz (avec intervalle de confiance à 95%) pour les patients MDP sevrés (OFF), avec médicament (ON) et les sujets contrôles (CTRL)

5.2 Effet de la tâche de parole

Il n'existe aucune différence significative de F0 moyenne entre les tâches de production de voyelles tenues (a_max, aaa), de répétition de phrase ou de lecture (rapide ou normale). En revanche, la tâche de description d'image se distingue de toutes les autres de façon significative avec un abaissement notable de la F0. Cette caractéristique s'observe quel que soit le groupe des locuteurs (pas d'interaction).

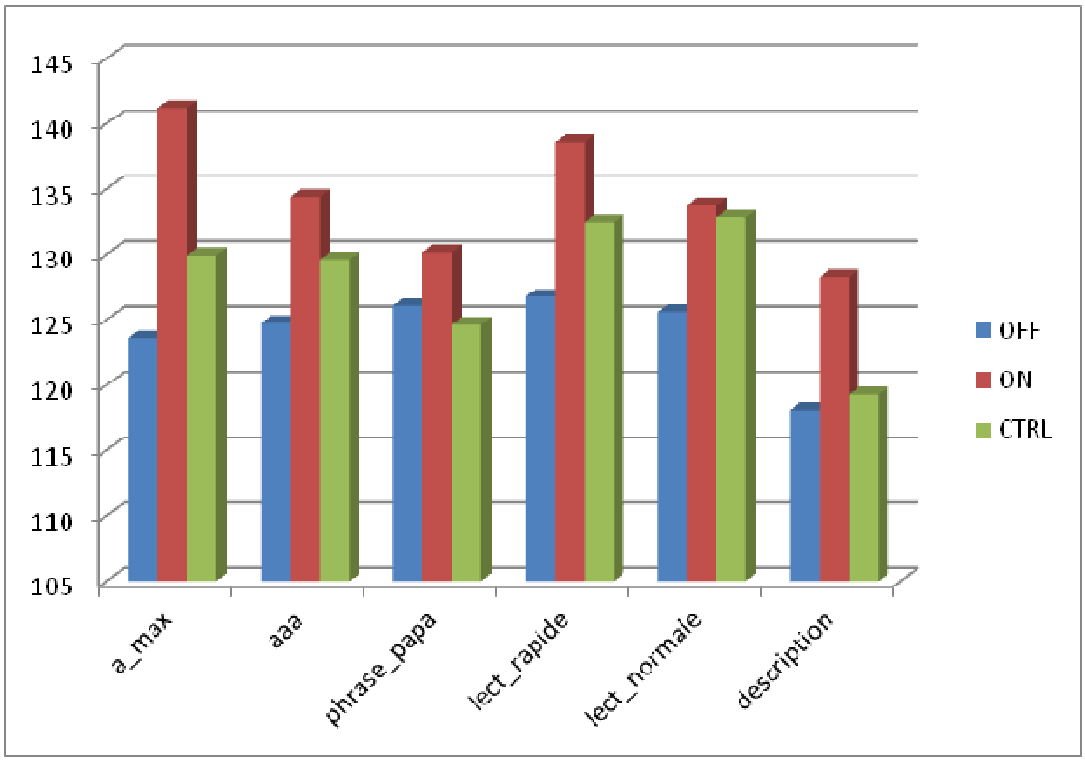


FIGURE 4 – F0 moyenne (en Hz) en fonction de la tâche de parole pour les patients MDP sevrés (OFF), avec médicament (ON) et les sujets contrôles (CTRL)

6 Discussion

6.1 Effet de la maladie et effet médicamenteux

Les résultats obtenus contredisent nos hypothèses fondées sur le modèle de Goberman & Coelho (2002). Celui-ci prédit une F0 plus élevée chez les patients MDP liée à la rigidité laryngée. Nous n'observons pas cette élévation en situation OFF. D'après le même modèle, l'effet thérapeutique lié à la prise de L-dopa réduisant la rigidité devrait s'observer par une baisse de F0 par rapport à la situation de sevrage. Nous observons exactement le contraire.

Ces résultats peuvent s'expliquer par l'absence de prise en compte des mécanismes pneumophonatoires dans le modèle. En effet, Sarr et al. (2009a) ont montré que la forme hypokinétique de la dysarthrie parkinsonienne induit une baisse significative de la pression sous-glottique (PSG) lors de la phonation. Or, diverses études dont celle d'Atkinson (1978) ont mis en relation les liens étroits entre PSG et F0, avec notamment une forte corrélation positive. On peut donc prédire une baisse de F0 d'origine pneumophonatoire dans la MDP et proposer ainsi un complément au modèle de Goberman (Figure 5).

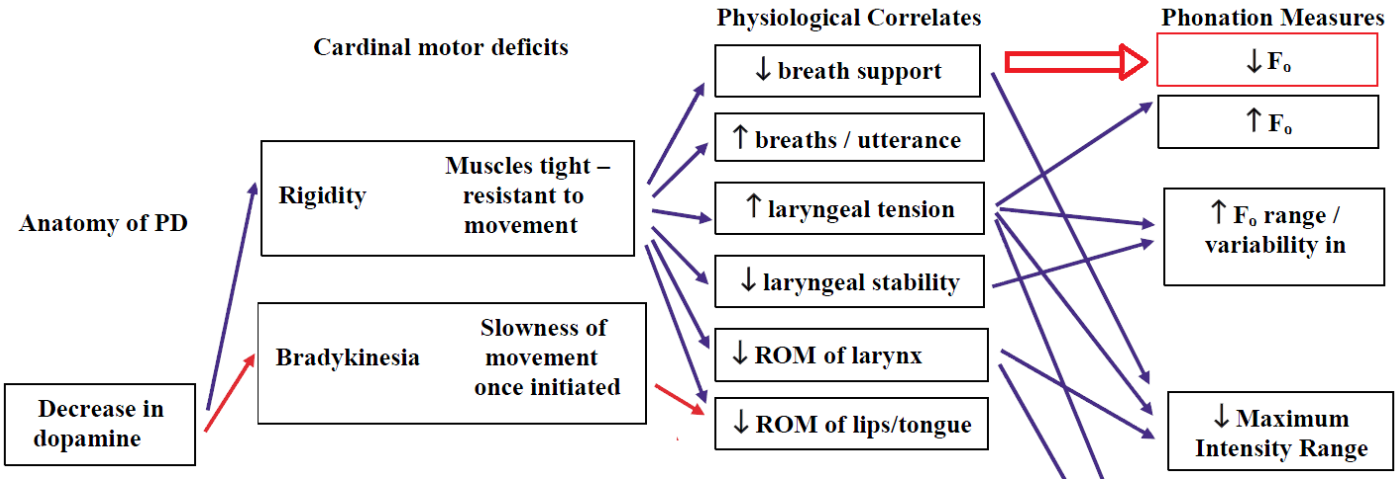


FIGURE 5 – Compléments au modèle de phonation parkinsonnienne

Il serait intéressant d'étudier, à ce sujet, si la baisse de pression sous-glottique dans la phonation parkinsonienne est plutôt liée à la rigidité (comme suggéré par Goberman dans son modèle) ou à l'hypokinésie comme suggéré dans Sarr et al. (2009). Nos données ne permettent pas de répondre à cette question. En revanche, nous pouvons maintenant expliquer l'absence d'augmentation de F0 chez les patients parkinsoniens OFF dans la mesure où interviennent deux mécanismes antagonistes d'abaissement (par la pression) et d'élévation (par la tension) qui se soldent par un effet globalement nul. De même, l'élévation de F0 observée en situation médicamenteuse (ON) peut être prédite dans la mesure où la prise de L-dopa permet aux patients de « normaliser » (augmenter) leur pression sous-glottique (Sarr et al.,2009b), ce qui se traduit par une hausse de F0. Cela signifierait aussi que la prise de L-dopa ne permettrait pas de réduire significativement les tensions laryngées, d'où un mécanisme global d'élévation de F0 comme prévu dans le modèle. On peut aussi supposer que les stratégies compensatrices habituellement utilisées par le patient pour augmenter son efficacité vocale et entraînant une augmentation de la F0 se maintiennent avec la prise de L-dopa.

6.2 Effet de la tâche de parole

En ce qui concerne l'influence de la tâche de parole sur la F0 moyenne, nos résultats mettent en évidence la particularité de la description d'image, qui est une tâche de parole non préparée et qui se distingue des autres tâches de parole (notamment la lecture et la voyelle tenue), avec une F0 plus basse pour la description. Ces résultats sont cohérents avec ceux de l'étude de Van Lancker Sidtis et al. (2004) qui ont observé sur des populations parkinsoniennes et aphasiques un effet significatif de la tâche de parole sur la F0, et notamment une moyenne plus basse pour la conversation par rapport à la lecture et la voyelle tenue. Il en est de même dans le travail de Swertz et al. (1996) en parole « normale » où les auteurs ont relevé un registre plus élevé en lecture qu'en parole spontanée. La voyelle tenue et la lecture étant des situations de parole plus guidées que la description d'image, on peut supposer qu'elles induisent pour tous les locuteurs un comportement de performance vocale qui fait augmenter la F0. Cette notion de performance, à replacer dans un cadre clinique d'examen, est particulièrement observable sur la Figure 4 dans laquelle les deux tâches d'effort (temps maximal de phonation et lecture rapide) ont induit les valeurs les plus élevées chez les patients en situation ON.

7 Conclusion

L'étude des dysfonctionnements de la parole nécessite un cadre conceptuel et méthodologique qui permet d'avancer des explications cohérentes relatives aux observations. Les résultats de ce travail montrent la complexité des contraintes qui opère dans la production de la parole notamment chez des patients atteints de la maladie de Parkinson. Les résultats contradictoires que l'on retrouve dans la littérature pour les effets de la maladie sur le registre vocal des patients peuvent s'expliquer par divers facteurs confondus dont les effets peuvent s'annuler ou s'additionner conjointement (sexe, âge, conditions thérapeutiques…). Si l'on considère tous les points développés dans ce travail, la situation dans laquelle on observerait un registre vocal le plus bas serait la condition OFF en parole spontanée. A l'inverse, une demande de performance telle qu'un temps maximal de phonation ou une lecture rapide en condition ON entrainerait un registre vocal haut. Des registres intermédiaires s'obtiennent sur du ON spontané ou du OFF avec performance (lecture rapide). Cela illustre ainsi l'intrication des contraintes physiologiques et linguistiques qui interagissent dans la production de parole.

D'autre part, ces résultats laissent apparaitre un point sensible d'un point de vue méthodologique : des situations de performance avec de la parole préparée (voyelle tenue, lecture) ne masquent-elles pas des effets observables dans des tâches de production de parole plus spontanée (description libre d'image) et qui passent plus inaperçus dans des situations où le locuteur se sent en examen induisant une compensation ad hoc, notamment une augmentation de F0 par mécanisme hyperfonctionnel; le corolaire de cette hypothèse serait ainsi un mécanisme plus hypofonctionnel en situation plus spontanée.

Ce travail illustre enfin la complexité des mécanismes de compensation mis en jeu dans la production de parole. En effet, la logique voudrait que la condition ON soit un état intermédiaire entre la condition OFF et la normale. Or, on observe dans notre travail un phénomène de dépassement, d'overshoot de production sous l'effet médicamenteux. Seul le rapprochement de modèles physiopathologiques comme celui de Goberman et al. (2002) et de modèles décrivant les aspects hypo/hyperfonctionnels de la parole (Lindblom, 1990) peuvent nous permettre de mieux comprendre les mécanismes mis en jeu.

Remerciements

Ce travail a été réalisé dans le cadre de l'ANR-12-BSH2-0003 « TYPALOC » et du Labex BLRI portant la référence ANR-11-LABX-0036 qui a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du projet Investissements d'Avenir A*MIDEX portant la référence n° ANR-11-IDEX-0001-02.

Références

ATKINSON JE, (1978), Correlation analysis of the physiological factors controlling fundamental voice frequency. J. Acoust. Soc. Am.; 63(1), 211–222.

DE LOOZE, C., GHIO, A., SCHERER, S., POUCHOULIN, G., VIALLET, F. (2012) Automatic analysis of prosodic variations in Parkinsonian speech, Proc. Speech Prosody 2012, Shanghai, China. 4 p.

GOBERMAN, A., COELHO, C. (2002). Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation, 17*, 237-246.

HIXON T.J, WEISMER G., HOIT J. D. (2008). *Preclinical Speech Science: Anatomy, Physiology, Acoustics and Perception*, Plural Publishing, 642 p.

LINDBLOM B, (1990) Explaining Phonetic Variation: A Sketch of the H&H Theory. Speech Production and Speech Modelling, Springer Netherlands, 403–439.

MAS M., GRIGOLI C., (2012), Influence de la L-dopa et des styles de parole sur la f0 chez le locuteur parkinsonien, Certificat de capacité d'orthophonie, Marseille, 102 p.

PINTO, S.; GHIO, A.; TESTON, B.; VIALLET, F (2010). La dysarthrie au cours de la Maladie de Parkinson. Histoire naturelle de ses composantes: dysphonie, dysprosodie et dysarthrie. *Revue Neurologique*, vol. 166, no. 10. 2010, p. 800-810.

ROBERT, D., SPEZZA, C. (2005), La dysphonie parkinsonienne. *Les troubles de la parole et de la déglutition dans la maladie de Parkinson*, Solal., 131–143.

SARR, M.; PINTO, S.; JANKOWSKI, L.; TESTON, B.; PURSON, A.; GHIO, A.; RÉGIS, J.; PERAGUT, JC; VIALLET, F (2009). Contribution de la mesure de la pression intra-orale pour la compréhension des troubles de la coordination pneumophonique dans la dysarthrie parkinsonienne. *Revue Neurologique*, vol. 165, no. 12. 2009, p. 1055-1061.

SARR M, PINTO S, JANKOWSKI L, PURSON A, GHIO A, ESPESSER R, TESTON B, VIALLET F, (2009), L-dopa and STN stimulation effects on pneumophonic coordination in Parkinsonian dysarthria: intra-oral pressure measurements. In: Proc. International Congress of Parkinson's Disease and Movement Disorders, Paris, France; pp. Th–240.

SWERTS, M.; STRENGERT, E. ; HELDNER, M. (1996), F0 Declination in Read-aloud and Spontaneous Speech, 4th ICSLP, Philadelphia, USA.

VAN LANCKER SIDTIS D, HANSON W, JACKSON C, LANTO A, KEMPLER D, METTER EJ, (2004), Fundamental Frequency (F0) Measures Comparing Speech Tasks in Aphasia and Parkinson Disease. Journal of Medical Speech - Language Pathology, 12(4), 207-212

Algorithme de découpage en groupes prosodiques pour la dictée par l'usage de synthèse vocale

Sébastien Le Maguer[1] *Elisabeth Delais-Roussarie*[2] *Nelly Barbot*[1]
Mathieu Avanzi[2] *Olivier Rosec*[3] *Damien Lolive*[1]
[1]IRISA, Université de Rennes 1, Lannion, France
[2] UMR 7110-Laboratoire de Linguistique Formelle,
Université Paris-Diderot, France [3] VOXYGEN, Lannion, France
Sebastien.Le_maguer@irisa.fr, Elisabeth.roussarie@wanadoo.fr,
Nelly.Barbot@irisa.fr, mathieu.avanzi@gmail.com, olivier.rosec@voxygen.fr,
Damien.Lolive@irisa.fr

RÉSUMÉ

L'objectif de cet article est de présenter un algorithme de segmentation automatique d'un texte en une séquence de groupes prosodiques dans le cadre de la génération de dictée. Ce découpage vise à respecter les règles et procédures utilisées dans un cadre de dictée pour des élèves de primaire. Une meilleure compréhension et modélisation de ces règles est primordiale pour développer des outils automatiques robustes ayant recours à la synthèse vocale ; ceux-ci pourront alors être utilisés par les enfants de manière autonome et leur permettre d'améliorer leur compétences à l'écrit par l'usage de synthèse vocale comme feedback audio. Les différentes étapes de dérivation des groupes prosodiques sont expliquées à travers des exemples concrets. À partir de l'analyse d'un corpus de 10 dictées en langue française données à des enfants dans des écoles primaires françaises et canadiennes, nous proposons un algorithme qui génère automatiquement les groupes prosodiques à partir du texte.

ABSTRACT

Prosodic chunking algorithm for dictation with the use of speech synthesis

The aim of this paper is to present an algorithm that automatically segment a text in prosodic chunks for dictation by conforming to the rules and procedures used in real settings to dictate a text to primary school children. A better understanding and modeling of these rules and procedures is crucial to develop robust automatic tools that could be used in autonomy by children to improve their spelling skills through dictation with the use of speech synthesis. The different steps used to derive the prosodic chunks from a given text will be explained through concrete examples. The proposal made here relies on the analysis of a corpus of 10 dictations given to children in French and French Canadian elementary schools. The phrasing observed in the data is described. It is thus simplified in order to develop an algorithm that automatically generates prosodic chunks from texts.

MOTS-CLÉS : Synthèse de la parole, prosodie, segmentation automatique.

KEYWORDS: Speech synthesis, prosodic phrasing, automatic parsing.

1 Introduction

L'utilisation de logiciels dans le cadre de l'enseignement des langues offre plusieurs avantages. S'adapter au besoin de l'apprenant et lui proposer un environnement au sein duquel il peut évoluer en toute autonomie font partie de ces avantages. Toutefois, à l'heure actuelle, il n'existe que très peu de logiciels faisant appel à de la synthèse de la parole parmi ceux utilisés en école primaire pour enseigner la lecture et l'écriture du Français (cf. toutefois (Lectramini, 2013) et PLATON (Beaufort et Roekhaut, 2011)).

Un des objectifs du project ANR Phorevox (`http://www.phorevox.fr/`) est de développer des outils de ce type en portant une attention particulière à l'acquisition de l'écrit. Le logiciel devra proposer des exercices qui permettent à l'enfant d'améliorer ses compétences en écriture. Dans ce contexte, des dictées leur seront proposées afin qu'ils travaillent sur des difficultés orthographiques ou grammaticales spécifiques.

L'utilisation de procédures automatiques et de la synthèse de la parole dans le cadre d'une dictée impose plusieurs contraintes. En effet, il est nécessaire de (i) proposer une synthèse intelligible afin que l'apprenant puisse entendre chaque mot en vue de son écriture ; (ii) proposer des groupes qui, tout en ayant une taille raisonnable, permettent de gérer les difficultés orthographiques et grammaticales (règles d'accord, etc.) ; (iii) proposer un environnement agréable à l'enfant en tenant compte, notamment, de sa vitesse de frappe. Dans le cadre de cet article, nous nous focalisons sur la génération automatique de groupes à partir du texte de la dictée. Afin de proposer une procédure automatique, nous avons tout d'abord analysé un ensemble de dictées réalisées en primaire. Les résultats de ces observations ont servi à définir et formaliser les règles utilisées par l'algorithme de génération de dictée.

Cet article s'articule comme suit. Dans la section 2, les caractéristiques prosodiques du français et les règles utilisées pour segmenter le flux sonore en groupes prosodiques sont brièvement présentées. La section 3 présente les données et la méthodologie utilisées pour étudier les découpages prosodiques observés lors des exercices de dictée en français. Les règles principales sont listées dans la section 4. Enfin, la section 5 présente l'algorithme, qui repose sur les règles sélectionnées, et qui permet de générer une dictée de manière automatique à partir de n'importe quel texte.

2 Contexte

Les études sur la prosodie du français mettent traditionnellement en avant le fait que l'accentuation, le phrasé et l'intonation sont fortement corrélés (cf., entre autre, (Di Cristo, 1998), (Post, 2000) et (Martin, 2009)). En français, l'absence d'accent lexical implique un syncrétisme entre l'intonation et l'accentuation. Dans les études consacrées au phrasé prosodique, deux ou trois niveaux hiérarchiques au delà du mot sont généralement retenus. Le niveau le plus bas correspond au syntagme mineur (MiP) - également appelé groupe accentuel (cf., par exemple, (Jun et Fougeron, 2000) et (Delais-Roussarie *et al.*, 2013)), groupe phonologique (cf., entre autre, (Post, 2000)) ou groupe rythmique ((Verluyten, 1982) par exemple). Cette unité se caractérise par la réalisation d'un accent sur sa dernière syllabe, indiquant ainsi la limite droite du syntagme. Dans la littérature, il y a un large consensus concernant la définition de cette unité : elle correspond globalement à un mot lexical et à l'ensemble des mots fonctionnels qui lui sont associés (cf.,

entre autre (Post, 2000), (Jun et Fougeron, 2000), (Delais-Roussarie *et al.*, 2013) et (Verluyten, 1982)). À titre d'exemple, les phrases sous (1) sont segmentées en syntagmes mineurs comme illustré sous (2).

(1) a. *Les enfants sont venus dans l'après-midi.*
b. *Bernard est rentré de son voyage en Asie.*

(2) a. (Les enfants)$_{MiP}$ (sont venus)$_{MiP}$ (dans l'après-midi)$_{MiP}$.
b. (Bernard)$_{MiP}$ (est rentré)$_{MiP}$ (de son voyage)$_{MiP}$ (en Asie)$_{MiP}$.

En plus des syntagmes mineurs, deux niveaux de groupement prosodique sont souvent identifiés : le syntagme intermédiaire (ou ip) (Delais-Roussarie *et al.*, 2013; Michelas, 2011), appelé également syntagme majeur (Selkirk, 1986) ou syntagme phonologique restructuré (comme dans (Post, 2000)) ; et le syntagme intonatif (ou IP) (Di Cristo, 1998; Post, 2000; Jun et Fougeron, 2000; Delais-Roussarie *et al.*, 2013) qui est également appelé *groupe de souffle*. Parmi ces deux niveaux, seul le syntagme intonatif est retenu dans tous les modèles. Il correspond à l'unité prosodique la plus grande et se caractérise par la présence d'un contour intonatif sur sa frontière droite, le plus fort degré d'allongement final de syntagme, et il est souvent suivi d'une pause. Dans le cas d'une séquence de propositions, chaque proposition est souvent traitée comme un groupe intonatif indépendant.

Dans la section 4, les découpages prosodiques observés et les informations nécessaires à la construction des ces groupes seront exposés en détail. Cela nous permettra d'établir quelles sont les différences entre les découpages prosodiques standard et les découpages spécifiques aux dictées.

3 Corpus et méthodologie

3.1 Corpus

Afin d'étudier les découpages prosodiques et l'intonation dans le cadre de la lecture de dictée, un ensemble de dictées lues à des enfants allant du CP au CE2 en France et au Québec (Canada) a été collecté. Ces données proviennent de trois sources :

– quatre dictées courtes sont issues de l'association canadienne "Fondation Paul Gerin-Lajoie" (Gerin-Lajoie, 2013), en particulier, les dictées du premier niveau (équivalent CP) ;
– quatre dictées sont issues du site français Ladictéee.fr (Ladictée, 2013) qui offre un ensemble varié de dictées et d'exercices grammaticaux pour les écoliers ;
– deux dictées ont été enregistrées en situation réelle par des chercheurs du projet Phorevox.

Le choix des dictées a été effectué afin de couvrir les différents niveaux scolaires visés pour le logiciel. Bien que le nombre de dictées étudié soit limité, il est important de noter que les variations portent plus sur la façon de répéter les groupes et d'en introduire de nouveaux, que sur la formation des groupes à proprement parler. Ce point est essentiel, dans la mesure où c'est surtout la formation des groupes qui nous importe ici.

3.2 Méthodologie

Les dictées ont été annotées par deux des auteurs en ajoutant, pour chaque texte, deux types d'informations : les découpages prosodiques (i.e. la manière dont les textes ont été segmentés

en groupes durant la lecture) ; et la forme du contour intonatif observé à la fin des différents groupes prosodiques. L'annotation a été effectuée au moyen d'une analyse perceptive et d'une étude instrumentale des données. L'analyse perceptive a consisté à écouter attentivement le signal sonore ce qui a permis de définir les groupes et la forme des contours intonatifs à la fin des groupes. L'analyse acoustique, réalisée en utilisant le logiciel Praat (Boersma et Weenink, 2014), a pour but de valider ce qui a été perçu. Une attention particulière a été portée aux contours intonatifs et aux pauses, qui indiquent en partie les découpages.

4 Analyse : définition des règles de segmentation

Avant de décrire en détail les règles utilisées pour obtenir les groupes, ainsi que les procédures mises en place pour effectuer la répétition et introduire de nouveaux groupes, nous souhaitons faire mention de trois points qui ont été observés dans l'ensemble des dictées analysées. Tout d'abord, le titre et le texte sont lus intégralement une fois à un débit relativement lent avant la lecture de la dictée proprement dite. Ensuite, durant la phase de dictée, la phrase complète est répétée une fois dès que tous les groupes la composant ont été lus. Enfin, les marques de ponctuation sont prononcées aux positions auxquelles elles apparaissent dans le texte afin que l'élève puisse les écrire, comme cela est illustré sous (3) et (4), où la transcription orthographique de ce qui a été prononcé est indiquée.

(3) *avec Papa point je marche dans la nature avec Papa point*
(4) *à l'école virgule je travaille toujours avec lui point*

Néanmoins, des variations apparaissent dans la façon de nommer les marques de ponctuation : il n'est pas nécessaire de nommer explicitement une virgule ou tout autre signe de ponctuation faible lorsque la phrase est répétée pour la dernière fois, après avoir été dictée par bribes ; en revanche, il se peut que les points ne soient nommés que lors de cette ultime répétition. Puisque ces derniers éléments ne sont pas systématiques, ils n'ont pas été pris en compte lors de la conception de l'algorithme présenté dans la section 5 : une marque de ponctuation est toujours dictée à la position à laquelle elle apparait dans le texte.

4.1 Découpages prosodiques observés

Dans les données observées, les groupes utilisés durant la dictée correspondent, dans plus de 95% des cas, à des syntagmes mineurs (voir section 2). Pour un groupe prépositionnel, par exemple, le nom est toujours dicté avec la préposition et le déterminant, comme indiqué dans (5a), et de même, un auxiliaire est dicté avec le verbe comme dans (5b).

(5a) *Je marche dans la nature avec Papa* → (je marche)$_{MiP}$ (dans la nature)$_{MiP}$ (avec Papa)$_{MiP}$
(5b) *Il se demande si sa maman a trouvé les bons médicaments* → (il se demande)$_{MiP}$ (si sa maman)$_{MiP}$ (a trouvé)$_{MiP}$ (les bons médicaments)$_{MiP}$

Il faut toutefois noter que deux syntagmes mineurs, obtenus grâce à l'information morphosyntaxique, peuvent être regroupés en un seul syntagme lorsque leur taille est inférieure à deux syllabes. Dans beaucoup de cas, par exemple, l'auxiliaire *être* est regroupé avec l'attribut comme cela est montré dans (6). Néanmoins, il arrive que cette règle de regroupement ne soit pas respectée, notamment si cela donne naissance à un groupe relativement long, comme sous (7).

(6) *Le lac est bleu* → [le lac] [est bleu]
(7) *C'est mon meilleur ami* → [c'est] [mon meilleur ami]

Chaque groupe prosodique mineur est réalisé comme un IP (c'est-à-dire suivi d'une pause), qu'il soit isolé comme sous (8) ou bien intégré à une phrase comme sous (9).

(8) *Avec mon ami.* → [avec mon ami]$_{IP}$ [point]$_{IP}$
(9) *Je joue dans l'eau avec mon ami* → [je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

Cela signifie qu'en dictée - vu ici comme un style de parole particulier - un groupe prosodique mineur est réalisé avec les caractéristiques prosodiques des IPs, en particulier avec un allongement final important et la présence d'une pause. Cette prosodie a été décrite comme complètement appropriée en français par (Verluyten, 1982). Il résulte de ce que (Verluyten, 1982) appelle *l'élasticité prosodique* c'est-à-dire le fait qu'un MiP peut être réalisé comme un IP sans restructuration supplémentaire.

La procédure de segmentation utilisée pour dicter le texte repose ainsi sur une analyse qui introduit une coupure majeure après les noms, verbes, adjectifs et adverbes s'ils ne sont pas des modifieurs du mot suivant. Ce dernier principe doit permettre de prononcer dans un même IP un adjectif prénominal et un nom, comme dans « le petit garçon », un adverbe et un adjectif comme dans « très ennuyeux », ou un auxiliaire et le participe passé comme dans « est arrivé ».

4.2 Procédures de répétition

En plus des règles de segmentation et de prononciation des groupes, il existe des procédures pour introduire les nouveaux groupes. Elles sont décrites dans les paragraphes suivants. Comme une certaine variabilité existe, nous avons inféré de l'observation des données trois procédures distinctes.

4.2.1 Procédure IP par IP

Cette procédure consiste à prononcer une phrase de la façon suivante. La phrase est d'abord prononcée en entier, mais chaque syntagme mineur MiP est séparé du suivant par une pause dont la durée varie entre 500 ms et 1 seconde. Puis chaque MiP est produit en isolation (voire parfois répété), une pause de 2 secondes minimum est alors réalisée entre un MiP et celui qui suit (ou la répétition de ce même MiP). Pour finir, une fois que tous les MiP ont été produits en isolation, la phrase entière est à nouveau répétée, chaque MiP étant alors séparé du suivant par une pause de 500 ms à 1 seconde. Généralement, les mouvements mélodiques réalisés à la fin des MiP sont montants. Pour (10), cette procédure amène au découpage et à la réalisation présentée en (11). Chaque retour à la ligne indique que le groupe est prononcé de manière séparée des autres et suivi d'une longue pause. Cette procédure a principalement été utilisée pour les dictées faites devant les plus jeunes élèves (1re année d'école élémentaire).

(10) Je joue dans l'eau avec mon ami.
(11) [je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$
[dans l'eau]$_{IP}$
[avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

4.2.2 Procédure par chaînage des IP

La procédure par chaînage d'IP repose sur une segmentation en MiP, et sur l'introduction progressive des MiP, lors d'une répétition. Comme pour la procédure présentée section 4.2.1, chaque IP correspond, pour ce qui est de son extension, à un MiP. Une phrase comme (12) est segmentée comme indiqué sous (13). Les sauts de ligne rendent compte de la présence de pause très importante, d'une durée supérieure à 3 secondes. L'IP [le chien], qui est comparable à un mot prosodique, est d'abord produit en isolation et suivi d'une longue pause. Il est alors répété, et suivi de l'IP (ou MiP) qui le suit. Une pause d'une durée courte (200 ms en moyenne) est réalisée entre les deux IP [le chien] et [s'étire]. L'IP introduit est alors réalisé en isolation et suivi d'une longue pause, et ainsi de suite. Une fois que la phrase a été produite, elle est répétée, la segmentation étant clairement marquée par la réalisation de pauses d'une durée de 200 à 500 ms.

(12) Le chien s'étire sur le tapis.

(13) $[\text{le chien}]_{IP}$
$[\text{le chien}]_{IP}\ [\text{s'étire}]_{IP}$
$[\text{s'étire}]_{IP}$
$[\text{s'étire}]_{IP}\ [\text{sur le tapis}]_{IP}$
$[\text{sur le tapis}]_{IP}\ [\text{point.}]_{IP}$
$[\text{le chien}]_{IP}\ [\text{s'étire}]_{IP}\ [\text{sur le tapis}]_{IP}\ [\text{point}]_{IP}$

Même si cette procédure a été utilisée dans approximativement 25% des cas dans nos données, elle possède de nombreuses faiblesses, en particulier, dans le cas de phrases complexes. Par exemple, lorsque le complément s'intercale entre le sujet et le verbe, l'accord ne peut être déduit de manière directe, ce qui peut induire des erreurs.

4.2.3 Procédure phrase par phrase

Cette dernière procédure consiste à dicter le texte phrase par phrase. Chaque phrase est prononcée une ou deux fois en fonction de sa taille, à un débit relativement lent, et elle est suivie d'une pause longue. Pour ce qui est de la segmentation interne, chaque MiP est séparé du MiP suivant par une pause dont la durée varie entre 200 et 600 ms. La présence de la pause permet de considérer les MiP comme équivalents à des IP.

Dans les cas complexes ou dans les cas de phrases longues, la segmentation peut être réalisée proposition par proposition. Une proposition réfère à plusieurs types d'éléments : (i) un syntagme séparé par une virgule, dans le cas d'un ajout en périphérie suivi d'une proposition comme la séquence soulignée sous (14) ; (ii) des propositions subordonnées ou coordonnées comme sous (15). Lorsqu'une telle procédure de découpage est utilisée, la phrase complète est prononcée une fois de plus après que toutes les parties ont été dictées.

(14) <u>À l'école</u>, je travaille toujours avec lui.

(15) Mon père rentre très tard à la maison <u>parce qu'il est musicien</u>.

5 Adaptation des règles et procédures pour la synthèse de parole

L'implémentation du découpage en groupes pour la dictée automatique avec un synthétiseur de parole a été réalisée en respectant les propriétés mentionnées au début de la section 4. Pour étayer les explications, nous allons considérer l'énoncé suivant comme exemple :

(16) *En réalité, c'est un hélicoptère. Avec une cheminée qui crache de la fumée, comme une locomotive à vapeur.*

5.1 Génération des groupes

5.1.1 Algorithme

La génération des groupes peut être considérée comme un dérivé des techniques de type "chink & chunk" (Liberman et Church, 1992) en ce sens qu'il repose sur l'arbre syntaxique associé au texte. Celui-ci est obtenu grâce à l'outil Synapse pos tagger (Synapse, 2011). Dans l'arbre syntaxique, nous considérons les nœuds comme des groupes syntaxiques et les feuilles sont associées à chaque mot.

Étant donné un nœud N_0, notons les nœuds enfants de N_0 par $(N_1 \dots N_n)$. Chaque nœud N_j représente un groupe de mots. Ainsi, N_j est défini par $N_j = (s_j, e_j)$ où s_j est l'indice du premier mot du groupe et e_j est l'indice du dernier mot du groupe. L'arbre syntaxique possède les propriétés suivantes qui impliquent un ordre sur les nœuds fils : $s_0 = s_1$; $e_0 = e_n$ et $s_j = e_{j-1} + 1$.

L'objectif de l'algorithme est de trouver la séquence des groupes syntaxiques pouvant servir de base à la formation des groupes. Pour cela, nous supposons un paramètre w défini par l'utilisateur représentant le nombre idéal de mots que doit contenir un groupe et associons le coût $C(N_j) = |(e_j - s_j + 1) - w|$ à chaque nœud N_j. Le principe de cette fonction de coût est de déterminer si le découpage d'un groupe N_0 en groupes $N_1, \dots, N_n$ de taille plus faible permet d'approcher la taille idéale. L'algorithme est le suivant :

1. si N_0 est une feuille, le groupe est défini comme étant le mot contenu par N_0,
2. si $(C(N_0) < \sum_{j=1}^{n} C(N_j))$ et N_0 n'est pas une feuille, deux cas sont considérés :
   – si la séquence $(N_1, \dots, N_n)$ est constituée uniquement de feuilles, le groupe retenu est N_0,

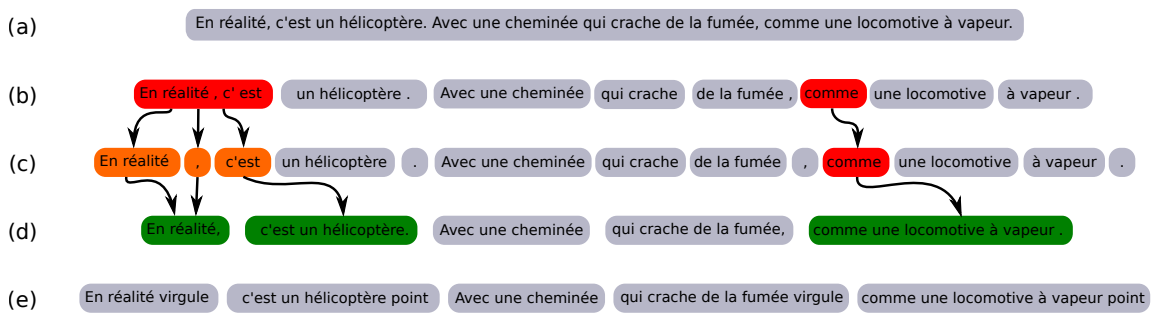


FIGURE 1 – Procédure de découpage pour (18)

– sinon, $(N_1 \dots N_n)$ est recomposée en éliminant les feuilles. Pour réaliser cela, chaque feuille, si possible, est regroupée avec le groupe précédent de manière incrémentale. On note $(N'_1 \dots N'_{n'})$ la séquence de nœuds ainsi obtenue. Si $(C(N_0) \geq \sum_{j=1}^{n} C(N'_j))$ alors l'algorithme est appliqué en considérant $N_0 = N'_j$, pour chaque $j \in \{1, \dots, n'\}$. Dans le cas contraire, N_0 est conservé.

3. dans les autres cas, l'algorithme est appliqué en considérant $N_0 = N_j$ pour tout $j \in \{1, \dots, n\}$

Cet algorithme appliqué sur (16) produit la segmentation (b) décrite dans la figure 1 ($w = 3$).

5.1.2 Post-traitement

L'étape précédente donne naissance à des découpages prosodiques qui ne sont pas toujours satisfaisants. Sur l'exemple, nous pouvons observer que certains groupes sont composés d'un seul mot (« comme »), d'autres contiennent des ponctuations au milieu, ce qui n'est pas adapté au contexte de la dictée (« en réalité, c'est »). En conséquence, nous avons défini une étape de post-traitement qui repose sur une approche par règles. Tout d'abord, une étape de découpage est appliquée afin d'isoler les ponctuations en utilisant des informations liées aux PoS. Les ponctuations doivent être isolées en raison de leur traitement spécial dans les dictées. De plus, une frontière de groupe est également ajoutée avant les conjonctions de coordination. À l'issue de cette étape, la segmentation présentée dans la figure 1 (c) est obtenue.

Une étape de regroupement est ensuite appliquée afin de satisfaire deux contraintes. La première est qu'il ne doit pas y avoir de groupe constitué d'un seul mot ou commençant par un caractère non-alphabétique. La deuxième contrainte est de s'assurer d'un nombre minimal de syllabes dans chaque groupe. Pour estimer le nombre de syllabes présentes, nous utilisons le nombre de voyelles non consécutives dans le groupe. Le nombre minimal de voyelles est un paramètre défini par l'utilisateur. Le résultat de cette étape est présenté dans la figure 1 (d).

5.2 Procédure de dictée complète

Une fois le texte segmenté en groupes par la procédure présentée précédemment, il est possible de le dicter de manière automatique. Chaque groupe est ensuite annoté de sorte qu'il soit prononcé de manière isolée (suivi par une pause très longue), les marques de ponctuation sont rendues explicites. Cela conduit, pour le texte (16), à la segmentation finale fournie dans la figure 1 (e). Le texte annoté est ensuite fourni en entrée du système de synthèse qui peut alors le traiter, en utilisant une voix de synthèse conçue de manière spécifique pour la dictée.

6 Conclusion et perspectives

Cet article présente un algorithme de découpage en groupes prosodiques qui permet des découpages en groupes comparables à ceux observés dans les dictées. D'autres travaux sont actuellement conduits dans le but de (i) déterminer quelle procédure est la plus appropriée en fonction du niveau des élèves ; (ii) fournir une analyse plus précise des contours intonatifs utilisés à la fin des différents IP non finaux.

Remerciements

Le travail présenté ici a été effectué dans le cadre du projet ANR-CONTINT 2011 *PHOREVOX* financé par l'ANR.

Références

BEAUFORT, R. et ROEKHAUT, S. (2011). Automation of dictation exercises. a working combination of call and nlp. *Computational Linguistics in the Netherlands Journal*, 1:1–20.

BOERSMA, P. et WEENINK, D. (2014). Praat (version 5.3). *Amsterdam : Institute of Phonetic Sciences*. Version 5.3.51, retrieved 2 June 2013 from `http://www.praat.org`.

DELAIS-ROUSSARIE, E., POST, B., AVANZI, M., BUTHKE, C., DI CRISTO, A., FELDHAUSEN, I., JUN, S.-A., MARTIN, P., MEISENBURG, T., RIALLAND, A., SICHEL-BAZIN, R. et YOO, H.-Y. (2013). *Developing a ToBI system for French (Accepted)*, chapitre 3. Oxford University Press.

DI CRISTO, A. (1998). Intonation in french. *Intonation systems : A survey of twenty languages*, pages 195–218.

GERIN-LAJOIE, F. P. (2013). La dictée p.g.l. `http://fondationpgl.ca/audio/`.

JUN, S.-A. et FOUGERON, C. (2000). A phonological model of french intonation. *In* BOTINIS, A., éditeur : *Intonation : Analysis, Modelling and Technology*, pages 209–242. Kluwer Academic Publishers.

LADICTÉE (2013). `http://www.ladictee.fr/`.

Lectramini (2013). Lectramini. `http://www.lectramini.com/`.

LIBERMAN, M. Y. et CHURCH, K. W. (1992). Text analysis and word pronunciation in text-to-speech synthesis. *Advances in speech signal processing*, pages 791–831.

MARTIN, P. (2009). *Intonation du français*. A. Colin.

MICHELAS, A. (2011). *Caractérisation phonétique et phonologique du syntagme intermédiaire en français : de la production à la perception*. Thèse de doctorat, Université de Provence-Aix-Marseille I.

POST, B. (2000). *Tonal and phrasal structures in French intonation*, volume 34. Thesus.

SELKIRK, E. (1986). On derived domains in sentence phonology. *Phonology yearbook*, 3(1986): 371–405.

SYNAPSE (2011). Cordial analyseur. `http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm`.

VERLUYTEN, P. (1982). *Recherches sur la prosodie et la métrique du français*. Thèse de doctorat, Université d'Angers.

Évaluation d'un algorithme de chunking appliqué à la dictée

Hiyon Yoo[1] *Sébastien Le Maguer*[2] *Elisabeth Delais-Roussarie*[1] *Nelly Barbot*[2] *Damien Lolive*[2]

[1] UMR 7110-Laboratoire de Linguistique Formelle, Université Paris-Diderot, France ; [2]IRISA, Université de Rennes 1, Lannion, France

yoo@linguist.jussieu.fr, Sebastien.Le_maguer@irisa.fr,

Elisabeth.roussarie@wanadoo.fr, Nelly.Barbot@irisa.fr, Damien.Lolive@irisa.fr

RÉSUMÉ

Dans cet article, nous présentons une évaluation subjective concernant l'utilisation de la synthèse de la parole pour la production automatique de dictées destinées à des élèves d'école primaire. L'évaluation consiste en des questionnaires permettant d'obtenir des informations qualitatives sur trois aspects de l'algorithme et du système de synthèse : (i) la pertinence de la segmentation, en particulier en ce qui concerne à la position des frontières prosodiques, (ii) la taille des groupes, et (iii) la pertinence de l'utilisation d'une voix de synthèse, une attention particulière étant portée au débit de parole et aux patrons intonatifs. Afin d'obtenir des retours de la part des utilisateurs potentiels de l'outil pédagogique que nous développons, nous avons élaboré deux questionnaires distincts pour les deux populations (élèves et enseignants).

ABSTRACT

Evaluation of a prosodic chunking algorithm : application to dictation

In this paper, we proposed a subjective evaluation concerning the use of speech synthesis for generating dictation to primary school pupils. The evaluation test consists of questionnaires that allow getting qualitative information on three distinct points concerning the algorithm and the speech synthesis system : (i) accuracy of the segmentation procedure, in particular regarding the location of the prosodic boundaries, (ii) size of the automatically calculated chunks, and (iii) accuracy and intelligibility of the speech synthesis system used, particular attention being given to the speech rate and the intonation patterns. Since we wanted to get feedback from potential users of the pedagogical tool we are currently developing, we designed a distinct questionnaire for two different populations (pupils and teachers).

MOTS-CLÉS : Synthèse de la parole, découpage prosodique, dictée, évaluation.

KEYWORDS: Speech synthesis, prosodic chunking, dictation, evaluation.

1 Introduction

Bien que les logiciels et autres outils informatiques offrent de nombreux avantages, notamment pour l'enseignement des langues, ils sont rarement utilisés pour enseigner l'entrée dans l'écrit en

langue première à des enfants d'école élémentaire (e.g. Lectramini (Lectramini, 2013) et Platon (Beaufort et Roekhaut, 2011)). Pourtant, les technologies vocales, et plus particulièrement la synthèse de parole peuvent s'avérer intéressantes dans une perspective d'apprentissage. Dans le cadre du projet ANR Phorevox (http ://www.phorevox.fr), nous travaillons au développement d'un logiciel d'aide à l'apprentissage de l'écrit auprès d'élèves de 6 à 9 ans, la synthèse de parole jouant alors souvent un rôle de feedback audio. Le développement d'un tel outil nécessite d'adapter la synthèse de manière à ce qu'elle soit hautement intelligible et adaptée à la prononciation de textes de genres variés (lecture de contes, de dictées, de poèmes, mais aussi prononciation de consignes, etc.). Ce travail d'adaptation de la synthèse passe par l'élaboration de différents modules (algorithme de segmentation en vue de dictées, cf. (Le Maguer *et al.*, 2014)), par la création de voix parlant plus lentement, mais il est aussi important d'avoir l'avis d'utilisateurs potentiels (élèves, enseignants, voire orthophonistes). L'objectif de cet article est de présenter les résultats d'une étude pilote effectuée auprès d'enseignants et d'élèves et consistant en une évaluation subjective de dictées produites de manières automatiques. Cette évaluation porte sur trois points : la segmentation du textes en groupes de mots en vue de la dictée (emplacement des frontières et taille des groupes) ; le débit de la voix synthétisée utilisée pour les dictées ; et les patrons prosodiques associés aux groupes dictés. L'article est organisé de la manière suivante : la section 2 présente l'algorithme utilisé afin de segmenter les textes pour la dictée. La section 3 décrit la méthodologie employée pour réaliser l'évaluation. Les résultats sont ensuite exposés et discutés dans la section 4.

2 Algorithme de découpage en groupes

Au cours d'un exercice de dictée, l'enseignant a tendance à segmenter le texte en groupes de mots, qu'il lit souvent de façon très lente. De plus, il répète fréquemment certains groupes, et il laisse suffisamment de temps aux élèves pour les écrire. Dans le logiciel d'apprentissage que nous développons, nous avons implanté une procédure automatique qui repose sur trois briques technologiques distinctes : un algorithme de découpage du texte en groupes de mots, une voix de synthèse construite spécifiquement pour la dictée, et un outil de suivi de frappe indiquant quel groupe est actuellement en cours d'écriture par l'apprenant et pour lequel des répétitions peuvent être nécessaires. Étant donné que l'un des objectifs du test est d'évaluer la pertinence du découpage automatique du texte, l'algorithme associé est brièvement présenté.

La conception de l'algorithme a été précédée par l'analyse d'un ensemble de dictées énoncées par des enseignants dans des conditions réelles. Cela a permis de distinguer les éléments linguistiques déterminants dans la construction des groupes. Il a ainsi été observé que les coupures syntaxiques importantes et les marques de ponctuation jouent un rôle crucial dans la segmentation du texte en groupes prosodiques. De plus, des contraintes rythmiques sont prises en compte afin d'éviter des groupes trop courts ou trop longs en termes de nombre de syllabes (ce type de contrainte est également valable pour le français standard, cf. (Martin, 1987; Delais-Roussarie, 1996; Martin, 2009)). En conséquence, l'algorithme utilise ces critères, pouvant être paramétrés par l'utilisateur, afin de fournir une segmentation en groupes appropriée.

L'algorithme se décompose en trois étapes. Tout d'abord, une analyse syntaxique du texte est réalisée à l'aide de l'outil Synapse (Synapse, 2011). Les mots du texte correspondant à des feuilles de l'arbre syntaxique sont regroupés selon leur proximité dans l'arbre, et en tenant compte des

contraintes de taille minimale (en termes de nombre de mots par groupe). En raison des erreurs possibles d'analyse et des contraintes de taille, le système permet de générer des groupes réduits à un seul mot comme « comme » sous (1) et des groupes possédant une ponctuation au milieu comme « En réalité, c'est » sous (2).

(1) *Avec une cheminée qui crache de la fumée, comme une locomotive à vapeur.*
→ (avec une cheminée) (qui crache) (de la fumée) (comme) (une locomotive) (à vapeur)
(2) *En réalité, c'est un hélicoptère* → (en réalité, c'est) (un hélicoptère)

Afin d'éviter cela, les marques de ponctuation sont considérées comme des frontières prosodiques. Ainsi, l'énoncé (2) est divisé commme suit : (en réalité,) (c'est un hélicoptère).

De plus, les groupes ne contenant qu'un seul mot sont fusionnés avec un groupe adjacent. Ainsi, « comme » est regroupé avec « une locomotive ». Le mécanisme de regroupement repose sur le nombre de syllabes par groupe comme paramètre principal, et non sur le seul nombre de mots. De plus, le nombre de voyelles graphiques non consécutives est pris en compte comme estimateur du nombre de syllabes. Pour le test, chaque groupe contient au moins 5 voyelles non consécutives. À l'issue du découpage, une indication de contour mélodique est ajoutée : un contour montant à la fin d'un groupe et avant une ponctuation ; un contour descendant avec un point. Enfin, des pauses courtes sont ajoutées entre un mot et une ponctuation tandis que des pauses longues (environ 3s) sont insérées après chaque groupe.

3 Méthodologie d'évaluation

L'emploi de questionnaires en linguistique est assez courant dans les domaines de la typologie, de la sociolinguistique (Labov, 1984), de la syntaxe (Buchstaller et Corrigan, 2011) ou de la variété des langues (Honeybone, 2010). Dans le domaine de la parole (cf. (Labov, 1968) et (Martinet, 1971)), (Wells, 2003) a montré que les questionnaires écrits permettaient de tester des aspects phonologiques, mais n'étaient d'aucun secours pour les points relevant de la phonétique. Pour ce qui est de la prosodie, (Hardison, 2005) a montré que les questionnaires pouvaient être utilisés pour évaluer les effets de certains paramètres dans l'acquisition de la prosodie. Dans cet article, le questionnaire est apparu comme la méthode la plus appropriée pour connaître le jugement des utilisateurs potentiels de l'application quant aux découpages prosodiques générés automatiquement. Dans cette section, nous présentons les stimuli à partir desquels les questionnaires sont construits (un pour chaque population), les participants, ainsi que la procédure suivie.

3.1 Stimuli utilisés

Les stimuli présentés sous (3) et (4) sont deux textes extraits d'un livre d'exercices pour enfants (extrait de Calle et Ahrweiller (2012)). Comme mentionné auparavant, plusieurs découpages peuvent être générés par l'algorithme en fonction des paramètrages retenus. Pour le test, qui porte sur la tâche de dictée, nous avons retenu les découpages donnés sous (3) et (4) et indiqués par le symbole "|".

(3) Lila marche sur la plage. | Elle voit un crabe qui | se cache sous une algue. | Elle soulève l'algue avec | un bout de bois. | Mais le crabe n'est plus | là, | il a disparu dans le sable. |

(4) Quand Marinette joue | de la trompette, | Marion joue du violon, | Martin tape sur un tambourin, | Mario gratte son banjo, | Manon nous chante une chanson, | et moi, | je frappe dans les mains. |

Ces stimuli ont été choisis pour deux raisons principales. La première est que le découpage généré présente des erreurs qui apparaissent fréquemment et qui peuvent s'avérer problématiques pour la tâche de dictée (certaines frontières, par exemple, ne vont pas dans le sens des découpages grammaticaux, comme « l'algue avec »). La deuxième raison est que nous souhaitions évaluer des groupes, créés avec des paramètres de taille que nous jugions à priori trop longs, notamment pour le public enfant.

Pour chaque dictée, le texte complet est lu une première fois avec un débit normal, la dictée commence ensuite. La voix de synthèse spécifique à la dictée produit des pauses entre chaque groupe. La durée des pauses peut être fixée par l'utilisateur.

3.2 Participants

L'outil que nous développons est dédié à un public d'enfants d'écoles primaires en cours d'acquisition de l'écriture et de la lecture (cycle 2, CP et CE1), et il doit offrir des exercices en lien avec les difficultés des apprenants à l'écrit. Il était donc important de savoir ce que les utilisateurs potentiels, les élèves et les enseignants d'école primaire, pensent de la procédure de dictée.

Un groupe de 7 filles ayant entre 6 ans 1 mois et 8 ans 2 mois et vivant en région parisienne a participé au test. Parmi ces élèves, une était en CE2 (8 ans 2 mois lors de la passation du test), une était en CP (6 ans 1 mois), et 5 en CE 1 (l'âge moyen étant de 7 ans 2 mois). Les élèves ont été choisies par les enseignants dans la mesure où elles n'avaient aucune difficulté scolaire, et présentaient même une certaine aisance dans l'acquisition de la langue écrite.

Le groupe des enseignants est quant à lui composé de 14 personnes (6 vivant en Région Parisienne, 6 en Bretagne, et 2 en Pays de Loire) qui ont accepté de participer au test de manière gratuite. Ils sont tous en activité comme enseignant dans des écoles primaires, à l'exception d'un participant qui est orthophoniste. Parmi eux, 10 sur 14 enseignent en cycle 2 (CP-CE1), et certains sont très intéressés par l'usage de nouvelles technologies dans leur classe même si seuls deux participants ont déjà utilisé des outils pédagogiques incluant de la synthèse de parole.

3.3 Procédure d'évaluation

La procédure d'évaluation est adaptée à chaque groupe. En conséquence, deux questionnaires ont été réalisés afin de collecter des informations sur le découpage et sur la tâche elle-même.

L'évaluation par les enfants est effectuée de la manière suivante. Les enfants doivent réaliser la dictée. Pour des raisons de place, nous ne présenterons ici que les résultats pour la première dictée (voir (3)). Nous avons utilisé un ordinateur portable équipé de haut-parleurs afin de lire la dictée produite par le système de synthèse. Les groupes ont été produits comme indiqué dans les sections 2 et 3.1. Aucune limite de temps n'a été imposée pour la réalisation de la dictée, et chaque groupe pouvait être répété autant de fois que nécessaire. Pendant la dictée, l'interviewer a pris note des emplacements pour lesquels la compréhension était difficile ainsi que du nombre de répétitions de chaque groupe. Après la dictée, les enfants devaient répondre oralement au

questionnaire (en majorité des questions de type oui/non et quelques questions libres). Ces questions concernaient (i) l'intelligibilité de la parole synthétique, en particulier lorsque des groupes ont été oubliés par les enfants ; (ii) les tailles des groupes, en particulier dans le cas des répétitions ; (iii) l'usage d'une voix de synthèse pour la dictée.

L'évaluation par le groupe des enseignants consiste en un test en ligne dans lequel les participants écoutent les stimuli (sans aucune contrainte), répondent aux questions, et peuvent donner leur opinion sur différents aspects de l'outil. Le questionnaire est divisé en trois parties : la première concerne le stimulus (3), la seconde concerne le stimulus (4) et la troisième présente des questions sur le débit de parole, et sur l'adéquation des patrons mélodiques dans une séquence particulière. Les participants de ce groupe écoutent tout d'abord la dictée, puis lisent le texte de la dictée présenté à l'écran. Ils répondent alors à des questions qui adressent les points suivants : la localisation des frontières de groupes, en particulier lorsqu'une erreur de découpage apparaît, la taille des groupes et l'intelligibilité de certains groupes. Afin d'être en mesure d'interpréter les résultats, la plupart des items du questionnaire attendent des réponses fermées (de type oui/non, trop lent/ normal/ trop rapide, etc.). Les résultats bruts de l'évaluation (et non en pourcentage en raison du nombre réduit de participants) sont présentés et discutés dans la section suivante.

4 Résultats

4.1 Évaluation par les enfants

Le tableau 1 donne les résultats pour chaque item du questionnaire concernant la voix de synthèse pour la dictée. On peut observer en premier lieu que les enfants peuvent réellement être des utilisateurs potentiels de l'application puisqu'ils trouvent cet outil utile dans le cas des dictées. La voix de synthèse est jugée naturelle malgré le fait que trois élèves trouvent la manière de prononcer étrange. La plupart des enfants ont éprouvé des difficultés à comprendre les mots « algue » (pour la deuxième occurrence lorsqu'il est groupé avec « avec »), « crabe » (à chaque fois) ainsi que le prénom « Lila ». De plus, 6 enfants sur 7 trouvent le débit trop élevé pour la dictée même si aucun des élèves n'a eu de difficultés à comprendre les phrases dictées.

Concernant les frontières, les élèves ne semblent pas gênés par le fait que « crabe » soit regroupé avec le pronom « qui ». Seulement deux enfants ont trouvé l'absence de frontière à cet endroit gênant. Cependant, regrouper « l'algue » avec la préposition « avec » a été perçu comme étrange, mais cela est probablement dû à l'intelligibilité du mot « algue » à cet endroit.

Enfin, les élèves trouvent que les groupes constitués d'un énoncé complet comme « Lila marche sur la plage » ou « il a disparu dans le sable » sont trop longs. Chaque groupe a été répété 7,7 fois en moyenne. En considérant les différents groupes, nous observons que la longueur des groupes n'est pas liée au nombre de répétitions (sauf pour le groupe 7 qui se résume au monosyllabique « là »). Nous notons également que le nombre de répétitions ne reflète pas les difficultés des enfants puisque la plupart ont demandé des répétitions dans le but de vérifier ce qu'ils avaient écrit. Les erreurs commises par les élèves sont cohérentes avec les mots perçus difficiles à comprendre. Ainsi, les groupes incluant les mots « crabe », « algue » et « Lila » sont les groupes qui possèdent le plus d'erreurs (pour 4 enfants sur 7).

Pour résumer, les enfants considèrent les groupes assez longs mais ne semblent pas gênés par

l'absence de correspondance entre frontières syntaxiques et prosodiques. Ils adoptent plutôt une stratégie de dictée mot par mot. Par exemple, dans le cas de « là », tous les enfants ont dans un premier temps écrit « la » sans placer d'accent, puis quatre d'entre eux ont corrigé le mot.

	Oui	Non
(a) J'aime les dictées	6	1
(b) Je voudrais utiliser un outil comme celui-ci pour faire des dictées à la maison ou en classe	7	0
(c) La voix utilisée dans la dictée ressemble à celle d'un robot	1	6
(d) La voix est agréable	7	0
(e) La manière de lire la dictée est étrange	4	3
(f) Le sens de la phrase est compris [1]	7	0
(g) La dictée était facile	3	4
(h) La manière de lire était trop rapide	6	1
(i) Tous les mots étaient faciles à comprendre	2	5
(j) C'est bizarre de regrouper « crabe qui »	2	5
(k) Le groupe « Lila marche sur la plage » est trop long	7	0
(l) C'est bizarre de regrouper « l'algue avec »	7	0
(m) Le groupe « il a disparu dans le sable » est trop long	6	1
(n) Mots difficiles à comprendre : algue (6), crabe (4), Lila (4), bout de bois (1)		

TABLE 1 – Résultats du questionnaire enfants. Scores bruts (de (a) à (m)) et réponses libres pour la question (n). Le nombre d'enfants ayant mentionné le mot est donné entre parenthèses.

4.2 Évaluation par le groupe d'enseignants

Le tableau 2 [2] donne les résultats du questionnaire concernant la première partie de l'évaluation. Contrairement aux enfants, le groupe des adultes semble plus attaché au respect des frontières liées aux constructions syntaxiques. 10 participants parmi 14 considèrent qu'un pronom relatif doit être regroupé avec la subordonnée qu'il introduit, et qu'une préposition comme « avec » doit être regroupée avec le groupe nominal qu'elle introduit. Les découpages proposés par les enseignants montrent que les déterminants doivent être regroupés avec leur nom (à l'exception de 2 réponses parmi 13, pour lesquelles une frontière est proposée entre « un » et « tambourin »). De manière générale, les enseignants considèrent que le découpage doit respecter les frontières morpho-syntaxiques ; dans le cas contraire, cela peut induire les élèves en erreur. L'évaluation par les enfants a montré que c'était bien le cas pour la préposition « avec » mais pas pour le pronom relatif. Des études plus systématiques sur ce point sont nécessaires afin de voir si la position des enseignants se fonde sur des résultats effectifs d'élèves, ou si elle repose sur une conception normative de la langue. Dans le cas de la séquence « mais le crabe n'est plus là », la plupart des participants ont proposé une frontière juste après le groupe sujet (11 cas sur 14, et pour un cas seulement, une frontière après l'adverbe « mais »). Cependant, l'évaluation de la seconde dictée (cf. tableau 3) montre que les enseignants acceptent de regrouper le sujet uniquement avec le verbe (9 cas sur 14) et non avec le complément du verbe, même si du point de vue de la syntaxe et de la prosodie en français, on fait souvent l'hypothèse d'une frontière forte entre

2. Par manque de place, les réponses libres sur les découpages des deux séquences sont données dans le texte.

	Oui	Non
(a) Trouvez-vous gênant que le pronom relatif « qui » soit regroupé avec son nom référent et non sa subordonnée ?	10	4
(b) Pensez-vous que ce groupe peut induire les enfants en erreur ?	10	4
(c) Pensez-vous que de tels groupes sont un obstacle à l'usage du logiciel ?	10	4
(d) Trouvez-vous gênant que la préposition « avec » soit détachée de son complément ?	10	4
(e) Si vous avez répondu Oui en (4), pensez-vous que la préposition doit toujours être regroupée avec son complément ?	10	0
(f) Trouvez-vous trop long le groupe « Lila marche sur la plage » ?	8	6
(g) Trouvez-vous trop long le groupe « il a disparu dans le sable point » ?	9	5
(h) La séquence « mais le crabe n'est plus là » est scindée en deux groupes. Auriez-vous fait la même chose pendant une dictée ?	5	9
(i) Pensez-vous que le groupe « là virgule » est trop court ?	10	4

TABLE 2 – Résultats pour le questionnaire enseignants, première dictée (3).

le sujet et le verbe (Delais-Roussarie, 1995; Di Cristo, 1998) [3]. Seulement 50% des participants auraient préféré une frontière après le sujet, et 8 participants sur 14 ne pensent pas que le sujet doive toujours être regroupé avec le verbe. L'examen des groupes proposés par les enseignants confirme cette remarque lorsque le sujet est un nom : 6 participants sur 8 ont formé le groupe (Lila marche) [4] et 10 sur 13 le groupe (Marion tape). Ce regroupement est systématique lorsque le sujet est un pronom (comme dans « il a disparu »). La séquence « elle soulève l'algue » est toujours préférée en un seul groupe, mais les enseignants peuvent être influencés par ce qu'ils ont entendu auparavant lorsqu'ils recommandent fortement de séparer le complément direct du complément indirect introduit par la préposition « avec ». Enfin, les découpages proposés par les enseignants montrent que les séquences courtes comme « mais », « et moi » et « là » peuvent former des groupes indépendants.

Concernant la taille des groupes, il semble que différentes stratégies puissent être adoptées. Ainsi, 10 participants sur 14 trouvent la séquence « là virgule » trop courte, mais « et moi » est accepté de manière unanime. Dans les autres cas, les réponses sont moins homogènes. Pour « Martin tape sur un tambourin », la séquence est perçue comme trop longue pour former un seul groupe (13 sur 14) mais les groupes « Lila marche sur la plage » et « il a disparu dans le sable » (ils ont respectivement 2 syllabes et 1 syllabe de plus que le premier groupe) sont acceptés comme pouvant former un seul groupe (resp. 6 et 5 réponses). Il est difficile de dire si c'est le nombre de syllabes, ou bien si le tri-syllabique « tambourin » qui sont en cause. Ce dernier peut être considéré comme difficile et ainsi devrait être présenté dans un groupe plus court.

La dernière partie du questionnaire concerne le débit de parole et les patrons prosodiques utilisés dans des séquences considérées comme problématiques. Dans le cas du débit de parole, pour toutes les questions, une nette préférence apparaît pour l'emploi d'un débit adapté (13 sur 14) plutôt qu'un débit normal jugé trop rapide par les participants (12 sur 14). Rappelons que les enfants ont également trouvé le débit trop élevé. Le contour prosodique (montant ou descendant) ne semble pas avoir d'influence positive sur la séquence. Pour la séquence « Quand Marinette

3. Cependant plusieurs études ont montré que le sujet doit être prononcé avec le verbe pour satisfaire l'eurythmie (cf. Martin (1987) et Dell (1984)).

4. Seulement 8 participants ont donné une réponse à cette question.

joue », 12 participants n'ont pas de préférence quant au patron prosodique utilisé. Pour des séquences problématiques, comme « elle soulève l'algue avec », nous obtenons des réponses assez aléatoires avec 5 pour un contour montant, 3 pour un contour plat et 6 pour un contour descendant. Il apparaît donc que le choix du contour intonatif utilisé n'améliore pas la voix de synthèse dans cette séquence particulière. Il semblerait même que certains cas ne suivent aucune règle en vigueur en français standard.

En résumé, comparées à celles de élèves, les réponses des enseignants soulignent l'importance de la syntaxe dans la détermination des frontières de groupe, au moins au niveau syntaxique le plus fin (i.e. les prépositions doivent être regroupées avec leur nom, les pronoms relatifs avec leur proposition, les déterminants avec leur nom, etc.). Cependant, pour les frontières après le groupe sujet, les réponses semblent être plus aléatoires et refléter des stratégies personnelles mises en oeuvre pendant la dictée. C'est également le cas pour la taille des groupes, où non seulement le nombre de syllabes est pris en compte mais aussi d'autres facteurs tels que les difficultés introduites par le lexique. Le débit de parole utilisé pendant la dictée est généralement jugé satisfaisant par les enseignants et le type de contour prosodique ne semble pas améliorer l'intelligibilité des séquences problématiques.

	Oui	Non
(a) Trouvez-vous étrange de regrouper sujet et verbe dans « Marinette joue »	5	9
(b) Auriez-vous préféré deux groupes avec une frontière après le sujet comme dans (Quand Marinette) et (joue de la trompette) ?	7	7
(c) Pensez-vous que le sujet doit toujours est regroupé avec le verbe ?	6	8
(d) Si vous avez répondu Oui à la question (3), pensez-vous que des propositions subordonnées ou indépendantes doivent être prononcées en un seul groupe, même s'il est long ?	1	5
(e) Trouvez-vous le groupe « Martin tape sur un tambourin » trop long ?	13	1
(f) Trouvez-vous le groupe « et moi » trop court ?	0	14
(g) Dans une situation réelle, auriez-vous prononcé « Et moi, je frappe dans les mains » en un seul groupe ?	1	13

TABLE 3 – Résultats pour le questionnaire enseignants, deuxième dictée (4).

5 Conclusion

Dans cet article, nous avons présenté une méthodologie qui a permis d'évaluer la pertinence d'un algorithme de découpage en groupes prosodiques pour la dictée. Les questionnaires, adaptés aux élèves et aux enseignants, nous ont permis de mettre en avant les paramètres qui doivent être pris en compte pour améliorer le programme que nous développons, en particulier pour ce qui est du placement des frontières et la taille des groupes. Les résultats montrent des différences de jugement entre les élèves et les enseignants. Ils montrent également qu'un contrôle de la procédure par l'utilisateur semble nécessaire. Nous allons poursuivre cette évaluation auprès d'un nombre plus important d'enfants et d'adultes, en dissociant enseignants et orthophonistes. Nous tenterons également d'apporter des améliorations aux points jugés problématiques.

Remerciements

Le travail présenté ici a été effectué dans le cadre du projet ANR-CONTINT 2011 *PHOREVOX* financé par l'ANR.

Références

BEAUFORT, R. et ROEKHAUT, S. (2011). Automation of dictation exercises. a working combination of call and nlp. *Computational Linguistics in the Netherlands Journal*, 1:1–20.

BUCHSTALLER, I. et CORRIGAN, K. (2011). How to make intuitions succeed : Testing methods for analysing syntactic microvariation. *Analysing variation in English*, pages 30–48.

CALLE, V. et AHRWEILLER, L. (2012). *CE1 – Dictée*. collection 'Je comprends tout'.

DELAIS-ROUSSARIE, E. (1995). *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.

DELAIS-ROUSSARIE, E. (1996). Phonological phrasing and accentuation in french. *Dam Phonology : HIL Phonology, Holland Academic Graphics, La Haye*, pages 1–38.

DELL, F. (1984). L'accentuation dans les phrases en français. *Forme sonore du langage*, pages 65–122.

DI CRISTO, A. (1998). Intonation in french. *Intonation systems : A survey of twenty languages*, pages 195–218.

HARDISON, D. M. (2005). Contextualized computer-based l2 prosody training : Evaluating the effects of discourse context and video input. *Calico Journal*, 22(2):175–190.

HONEYBONE, P. (2010). Using questionnaires to investigate nonstandard dialects, 'northern english and scots, phonology and syntax'. website, http ://www.lel.ed.ac.uk/dialects/nesps.html.

LABOV, W. (1968). *The social stratification of English in New York city*. Center for Applied Linguistics.

LABOV, W. (1984). *Language in use : Readings in sociolinguistics*, chapitre Research methods of the project on linguistic change and variation, pages pp. 28–53. Englewood Cliffs : Prentice Hall.

LE MAGUER, S., DELAIS-ROUSSARIE, E., BARBOT, N., AVANZI, M., ROSEC, O. et LOLIVE, D. (2014). Prosodic chunking algorithm for dictation with the use of speech synthesis. *In Proceedings of Speech Prosody*.

Lectramini (2013). Lectramini. `http://www.lectramini.com/`.

MARTIN, P. (1987). Prosodic and rhythmic structures in french. *Linguistics*, 25(5):925–950.

MARTIN, P. (2009). *Intonation du français*. A. Colin.

MARTINET, A. (1971). *La prononciation du français contemporain*. Droz.

SYNAPSE (2011). Documentation technique : Composant d'étiquetage et lemmatisation.

WELLS, J. C. (2003). Pronunciation research by written questionnaire. *In Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pages 215–218. Citeseer.

Quand l'horloge parlante a beaucoup à raconter sur l'évolution des techniques d'archivage audiovisuel

Félicien Vallet Jean Carrive
Institut national de l'audiovisuel
4 avenue de l'Europe
94366 Bry-sur-Marne cedex, France
fvallet@ina.fr - jcarrive@ina.fr

RÉSUMÉ

Depuis sa création en 1974, l'Institut national de l'audiovisuel (Ina [1]) est chargé de rassembler et conserver les images et les sons qui fondent la mémoire collective française, leur donner du sens et les partager avec le plus grand nombre. En coïncidence des innovations technologiques, le métier d'archiviste a évolué au cours des années pour répondre à l'émergence de nouveaux défis : évolution des supports physiques, apparition de l'ordinateur, stockage massif de données, numérisation des fonds, multiplication des chaînes de télévision et de radio à collecter, *etc.*
Dans cet article nous revenons sur un des choix technologiques faits par les diffuseurs de programmes télévisés : celui d'utiliser un enregistrement de l'horloge parlante pour indexer les programmes de télévision sauvegardés. Cette manipulation d'horodatage, qui peut être considérée comme l'ancêtre du *timecodage*, consiste à utiliser la stéréophonie pour enregistrer le signal audio utile sur un canal et l'heure courante donnée par l'horloge parlante sur l'autre (le son TV étant quasi-exclusivement mono à cette époque). De la fin des années 1970 au milieu des années 1990, environ 70 000 bandes d'archives ayant subi ce traitement ont ainsi été collectées par l'Ina. Cependant, à cette époque, aucune méthode de traitement automatisé n'avait été imaginée...

ABSTRACT

When the speaking clock has a lot to say about audiovisual archiving

Since 1974, the French National Audiovisual Institute is responsible for gathering, storing and sharing sound and images belonging to the French memory. The job of archivist has greatly evolved over time with the appearance of the computer, the creation of new storage devices, the archive digitization or the increase of the number of radio and television channels to be collected. In this paper we comment the technical choice made during the 1970s to archive and index television programs by recording the speaking clock. This processing, which can be seen as an ancestor of the timecoding, is based on the use of the stereophonic sound to record the speaking clock on one channel and the useful TV signal on the other one (since TV sound was almost exclusively mono at that time). Between the end of the 1970s and the beginning of the 1990s about 70 000 tapes of programs where recorded that way. However, at that time, no automatic treatment had ever been designed...

MOTS-CLÉS : Horloge parlante, Transcription automatique de la parole, Archivage audiovisuel.

KEYWORDS: Speaking clock, Automatic speech transcription, Audiovisual archiving.

1. Institut national de l'audiovisuel : www.ina.fr

1 Introduction

Avec le foisonnement de documents audiovisuels issus des productions télévisuelles et radiophoniques mais également d'amateurs (via les sites internet de type *YouTube* ou *Dailymotion*), les archivistes professionnels comme l'Institut national de l'audiovisuel (Ina) ont de plus en plus besoin d'outils de structuration efficaces. En effet, un fonds documentaire n'est pas seulement un stock ou un ensemble de documents, mais une mémoire vivante en constante évolution. Cet aspect repose sur l'instrumentation du fonds, c'est-à-dire son **indexation** (Bachimont, 1998). Les missions des archivistes étant de conserver mais également de documenter et d'exposer les contenus, les besoins de méthodes automatiques d'indexation se font ressentir chaque jour avec un peu plus d'acuité (les équipes de documentalistes étant limitées). Si actuellement le problème principal n'est plus le stockage, l'apparition de « cimetières de données » au sein desquels des documents sont laissés à l'abandon et la connaissance et l'information qu'ils contiennent sont perdues présente un risque grandissant. Par conséquent, afin d'éviter que l'archive ne s'étiole jusqu'à ne devenir qu'un stock, l'utilisation de méthodes d'**indexation automatique** est devenue inévitable.

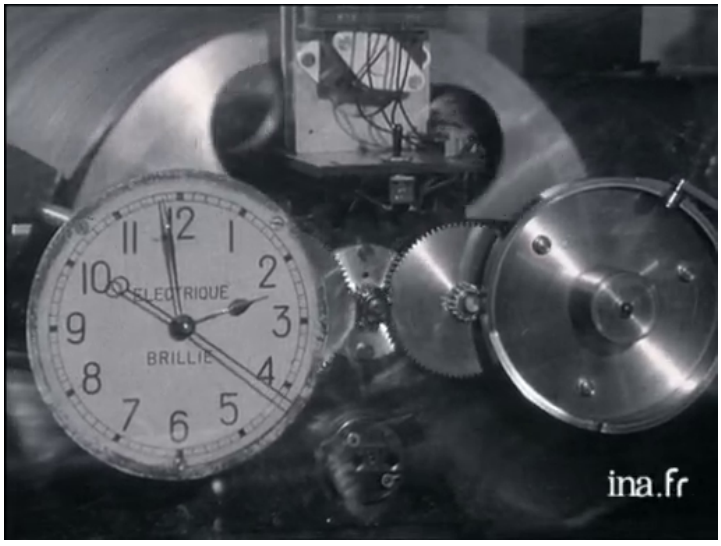


FIGURE 1: Première horloge parlante, système Brillié (1933-1965).

En 1933, l'Observatoire de Paris inaugure la première horloge parlante au monde (voir Figure 1). Ce dispositif annonce l'heure courante de manière automatisée et sous forme vocale par le biais d'un service téléphonique [2]. Par la suite, deux autres types d'horloges lui succèdent en 1965 et 1991 (cette dernière étant totalement électronique). Ce service, toujours en vigueur de nos jours, est accessible au 3699. En pratique, les heures sont énoncées de 0 à 23 puis les minutes de 0 à 59 en enfin les secondes à 10, 20, 30 et 40. La phrase « *au quatrième top, il sera exactement* » est ensuite prononcée pour annoncer le passage à la minute suivante. Chaque intervention est suivie d'un bip marquant le temps exact (heure atomique). Le changement de minute s'effectue à la suite de quatre bips consécutifs.

Afin de faciliter l'indexation des fonds, les diffuseurs de programmes télévisés décident dans les années 1970 de tirer profit du fait que le son TV est quasi-exclusivement mono pour effectuer une opération d'horodatage des contenus vidéo. Il s'agit d'enregistrer en « parallèle antenne », en plus du programme télévisé, l'heure courante donnée par l'horloge parlante. Avec l'aide de

2. voir « Bureau international de l'heure - l'horloge parlante » (vidéo Ina) : www.ina.fr/video/CAF93053168

la stéréophonie, le signal audio utile est donc sauvegardé sur un canal tandis que sur l'autre on enregistre l'horloge parlante. Il est ainsi possible de connaître avec exactitude l'heure de diffusion d'une émission et de faciliter son indexation en ayant recours à des grilles de programmes. Ce procédé autorise de plus à enregistrer d'affilée pour une même journée et sur un même support plusieurs programmes disjoints dans le temps et ainsi d'identifier rapidement à l'oreille les différents blocs temporels. Jusqu'aux années 1990 (et l'apparition de la piste TC sur les supports Béta SP), environ 70 000 bandes sont enregistrées de cette manière par les diffuseurs sur cassettes trois quarts de pouce (3/4") Umatic, BVU puis Béta TL avant d'être remises à l'Institut (qui les a numérisées depuis). Cette manipulation a été effectuée par d'autres gestionnaires de contenus à cette époque comme la British Broadcasting Company (BBC). Cependant, si l'enregistrement de l'horloge parlante permet de faciliter l'indexation d'un document, il peut également se révéler une mauvaise surprise pour les auditeurs. En effet, les acquéreurs d'archives ne sont pas spécialement désireux d'entendre la voix nasillarde de l'horloge. Il s'agit donc d'identifier dans les fonds de l'Ina les 70 000 bandes pour lesquelles un enregistrement de l'horloge parlante a été réalisé ce qui, en pratique, se révèle irréalisable manuellement.

Les avancées dans les domaines du traitement automatique de la parole et du signal rendent désormais possible la mise en place de procédures automatisées pour ces archives. Cet article nous permet donc, par le biais d'un problème d'ingénierie original d'éclairer le lecteur sur l'histoire et l'évolution des pratiques d'indexation documentaire et de le familiariser aux questions auxquelles les archivistes doivent répondre dans leur travail quotidien. Nous proposons dans la section 2 une description du mécanisme de détection de l'horloge parlante. Ensuite, une fois ce canal identifié (s'il y en a un), nous proposons une transcription automatique de la parole sous contrainte. Celle-ci permet d'identifier dans la section 4, les ruptures d'enregistrements à l'intérieur d'un même support et une correction de la transcription peut ensuite être effectuée par morceaux comme cela est décrit dans la section 5. La section 6 expose les résultats expérimentaux et apporte des éléments de discussion avant qu'une conclusion à l'étude ne soit proposée.

2 Détection de l'horloge parlante

Un des problèmes liés à l'enregistrement de l'horloge parlante dans les fonds de l'Ina tient à ce que les contenus ayant subi cette manipulation ne sont pas clairement répertoriés et que le canal affecté n'est pas toujours le même (il semble cependant s'agir majoritairement du canal droit). Par conséquent, la première étape d'un traitement massif consiste à détecter la présence d'un enregistrement de l'horloge parlante sur une des deux pistes audio.

Une visualisation du signal audio nous indique que dans ce cas les canaux présentent deux comportements différents (voir les formes d'ondes de la Figure 2). En particulier, le canal contenant l'horloge parlante présente un profil particulier avec des régularités (bips), des énonciations courtes (« *dix-huit heures trente-sept minutes vingt secondes* ») et de longues plages de silence.

Empiriquement, un détecteur de silence convient parfaitement pour distinguer de façon efficace lequel des deux canaux, s'il y en a un, contient un enregistrement de l'horloge parlante. En effet, si l'horloge parlante n'est pas enregistrée, le signal utile est dupliqué sur les canaux droit et gauche. Il s'agit donc de rechercher une forte disparité entre les proportions de silence présentes sur chacune des pistes audio. La méthode choisie pour la détection de silence est celle proposée dans (Saha *et al.*, 2005). En faisant l'hypothèse que les échantillons appartenant aux parties

non-voisées (*i.e.* bruit/silence) suivent une distribution gaussienne, il est possible après estimation de la moyenne μ et de la variance σ^2 de cette distribution de calculer la distance de Mahalanobis à une dimension pour une variable aléatoire x (dans notre cas un échantillon audio calculé pour 10 ms de signal) :

$$dist_{Mahalanobis} = \frac{|x - \mu|}{\sigma} \quad (1)$$

Les propriétés de la distribution gaussienne assurent en effet que l'échantillon considéré x a une probabilité de 99.7% d'être non-voisé si la distance de Mahalanobis est inférieure à 3.

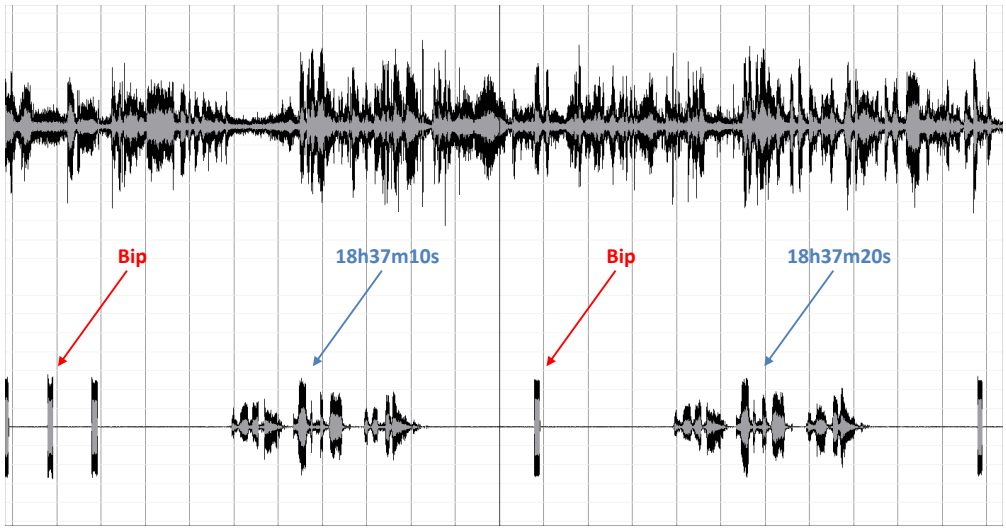


FIGURE 2: Formes d'ondes pour un document audiovisuel contenant un enregistrement d'horloge parlante sur la piste droite (en bas) et le signal audio utile sur la piste de gauche (en haut).

Grâce à cette méthode, il est donc possible de quantifier la proportion de silence sur chaque canal et de les comparer. Avec un seuillage adapté, un très bon taux de détection peut être obtenu (voir section 6). De plus, il est intéressant de noter que les erreurs de détection ne surviennent que pour les rares canaux muets ou défectueux.

3 Transcription automatique de la parole sous contrainte

Une fois le canal contenant un enregistrement de l'horloge parlante détecté, il s'agit d'exploiter les technologies de traitement de la parole pour extraire l'information utile à l'indexation du document. Pour cela une transcription automatique (Walker *et al.*, 2004) est réalisée à l'aide du logiciel Sphinx-4 [3]. Le modèle acoustique proposé par le Laboratoire d'Informatique de l'Université du Maine, LIUM [4], est employé pour permettre une transcription en français (Deléglise *et al.*, 2005).

Pour améliorer la qualité de la transcription, une adaptation de ce modèle acoustique est réalisée. Celle-ci est effectuée en ajoutant plusieurs centaines d'énonciations (phrases et mots) de l'horloge. Le vocabulaire utilisé étant un ensemble fini de 69 mots, il est très aisé de constituer le dictionnaire phonétique mettant en correspondance les mots et phonèmes. Enfin, on définit un modèle de

3. Sphinx-4 : www.cmusphinx.sourceforge.net
4. LIUM (outils et ressources) : www-lium.univ-lemans.fr

langage spécifique aux phrases prononcées par l'horloge parlante (en l'occurrence une grammaire) comme suit :

$$grammaire = ([au\ quatrième\ top,\ il\ sera\ exactement]\ x\ heures\ y\ minutes\ [z\ secondes])^* \quad (2)$$

avec x, y et z des entiers naturels satisfaisant : $x \in [0;23]$, $y \in [0;59]$ et $z \in \{10,20,30,40\}$. Ce système de transcription sous contrainte doit être suffisamment robuste pour traiter les horloges parlantes de 1965 et de 1991. Pour cette dernière, une alternance de deux voix (une féminine et une masculine) étant observée, des échantillons de chacune ont été ajoutés à ceux collectés pour l'horloge de 1965.

Les premiers résultats obtenus ont cependant montré que la retranscription des heures était délicate pour l'horloge parlante de 1965 (qui est celle majoritaire dans les enregistrements de l'Ina). En effet, il semble à l'écoute que les bandes de films parlants utilisées par l'Observatoire de Paris pour énoncer les heures sont moins intelligibles que celles pour les minutes. En particulier, les attaques sur les consonnes dentales (le [d] de « *dix-sept heures* » par exemple) sont très peu marquées. Une explication qui peut être avancée est que, sur une journée, la bande audio utilisée pour les heures passe 2.5 fois plus que celle des minutes, ce qui expliquerait une usure plus rapide. En effet, chaque minute est énoncée 5 fois par heure (au cours de la minute courante) et cela pour les 24 heures de la journée (120 énonciations/jour). Chaque heure est elle énoncée 5 fois par minute et ce pour chaque minute de l'heure courante (300 énonciations/jour).

Afin d'améliorer les scores de reconnaissance, 24 nouveaux mots, composés de la concaténation du chiffre et de l'heure (*zero_heure*, *une_heure*, *etc.*), ont été créés. La grammaire et le dictionnaire acoustique ont été eux aussi adaptés. Les taux de reconnaissance ont ainsi pu être réhaussés. Malheureusement, les résultats obtenus restent bien souvent sujets à contenir des erreurs et incomplets (puisque certaines énonciations ne permettent pas d'obtenir de transcription avec un niveau de confiance suffisant). En effet, les pistes audio contenant l'horloge parlante sont très souvent bruitées, rendant la transcription délicate. Il s'agit alors d'utiliser le caractère déterministe de l'horloge parlante pour corriger les omissions et erreurs de transcription.

4 Détection de ruptures

Afin de corriger la transcription, la première étape consiste à identifier les ruptures temporelles dans le cas où plusieurs programmes auraient été enregistrés consécutivement (bout-à-bout). Pour ce faire, nous exploitons la **cohérence temporelle**. En effet, à l'intérieur d'un bloc temporel cohérent les énonciations « *au quatrième top* » sont espacées d'exactement 60 secondes. Ainsi, la continuité temporelle est assurée si pour ε petit (0.25 secondes dans notre cas) on a :

$$|t_i^{au\ quatrième\ top} mod 60 \ - \ t_{i+1}^{au\ quatrième\ top} mod 60| \leq \varepsilon \quad (3)$$

Avec t le temps de l'énonciation en secondes, i un entier naturel tel que $i \in [1; N-1]$ et N le nombre de fois que la phrase « *au quatrième top* » est transcrite pour un enregistrement donné. On fait ainsi l'hypothèse que, pour deux blocs temporels consécutifs, il est impossible d'avoir deux enregistrements de l'horloge exactement synchrones.

Si l'équation 3 n'est pas vérifiée, une rupture est détectée et un nouveau bloc temporel est déclaré. En pratique, il s'agit de s'assurer de la robustesse de la détection de rupture dans le cas où la

phrase de référence « *au quatrième top* » n'a pas été transcrite correctement. Cela peut-être réalisé relativement facilement à l'aide d'hypothèses simples. Il est par exemple possible de stipuler qu'un bloc temporel a une durée minimale de 3 minutes, ce qui se vérifie empiriquement dans les programmes archivés par l'Ina sur ces types de supports.

5 Correction de la transcription

Une fois obtenus une transcription brute des énonciations de l'horloge parlante et un découpage en blocs temporels cohérents, il est possible de proposer une correction du flux de parole transcrit. En effet, comme cela a été évoqué à la section 3, l'élocution de l'horloge parlante présente des particularités spécifiques : le vocabulaire utilisé est particulièrement réduit et, à l'intérieur d'un bloc temporel, le discours est parfaitement déterministe.

Il s'agit donc d'effectuer une optimisation sous contrainte en exploitant la **cohérence syntaxique**. Les contraintes à satisfaire sont les données de transcription et l'enchaînement chronologique des séquences au sein d'un bloc temporel. Par conséquent, il faut trouver pour les heures et les minutes la séquence la plus probable. Ces deux séquences sont liées. Ainsi, un changement d'heure implique un passage de 59 à 0 minute et *vice versa*. Enfin, pour ce qui est des secondes nous faisons le choix de replacer celles-ci *a posteriori* dans la transcription corrigée (la précision attendue étant de l'ordre de la minute).

La Figure 3 propose un schéma du processus de décodage de l'horloge parlante comme il a été décrit jusqu'ici. De plus, un exemple d'alignement de séquence la plus probable est donné pour une transcription automatique bruitée et incomplète.

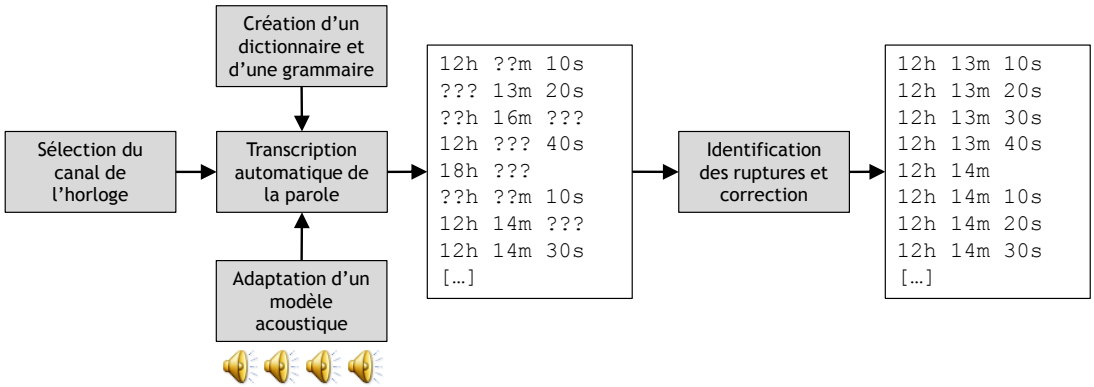


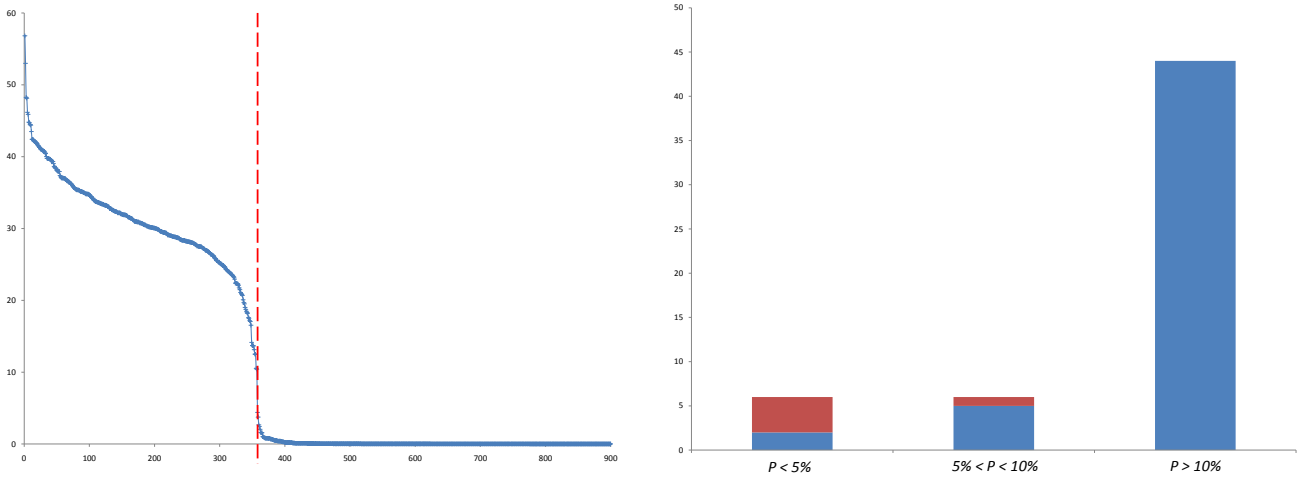
FIGURE 3: Schéma récapitulatif du processus de décodage de l'horloge parlante sur les documents Ina. La correction est effectuée par optimisation des contraintes imposées par la transcription.

L'avantage de forcer un alignement de la séquence déterministe la plus probable est que celle-ci permet le très simple calcul d'une probabilité d'obtenir cette séquence. Il est ainsi aisé de produire pour chaque bloc l'équivalent d'une valeur de confiance sur les heures et sur les minutes proposées. La probabilité d'avoir la séquence des heures corrigée dans le bloc temporel i se calcule donc comme suit (et de façon équivalente pour les minutes) :

$$P_i^{heure} = \frac{\text{Nb de fois que l'heure transcrite est identique à celle proposée dans la correction}}{\text{Nb total d'énonciations d'heures attendues dans le bloc } i} \quad (4)$$

6 Résultats expérimentaux

Comme cela a été présenté dans la section 2, la détection du canal sur lequel l'horloge parlante est enregistrée se fait par évaluation d'une différence de proportion de silence entre les canaux gauche et droit d'un enregistrement stéréo. L'analyse d'un corpus de 900 émissions, dont il est connu qu'entre un tiers et la moitié contiennent un enregistrement de l'horloge, présente un saut très marqué (en pointillés rouges sur la Figure 4a).



(a) Tracé de la valeur absolue de la différence de proportion de silence entre les canaux audio pour chaque émission analysée.

(b) Histogramme de la répartition des probabilités P pour les minutes. En bleu, les séquences correctes et en rouge celles erronées.

FIGURE 4: Identification de valeurs de seuils pour la détection et le décodage de l'horloge parlante.

Par conséquent, un seuillage pour une différence inter-canaux de 5% est effectué et évalué sur 200 fichiers annotés. Le taux de détection sur cet ensemble est de 97%. Les 3% restant étant constitués d'enregistrements comprenant des canaux muets ou abîmés avec beaucoup de silence.

L'évaluation de la correction de la transcription de l'horloge parlante a été effectuée sur un sous-corpus de 26 fichiers (22 avec l'horloge de 1965 et 4 avec celle de 1991). La Table 1 regroupe les les résultats obtenus. On peut remarquer que les ruptures temporelles sont toutes correctement détectées et que les taux d'erreurs sur les heures et les minutes sont de l'ordre de 8%.

entité détectée	**erreurs**
ruptures	0/30 (0.0%)
heures	4/56 (7.1%)
minutes	5/56 (8.9%)

TABLE 1: Résultats de détection de ruptures (30 au total) ainsi que de correction de transcription des séquences pour les heures et minutes à l'intérieur de chaque bloc temporel (56 au total).

Ces derniers taux s'avèrent trop élevés pour le traitement de plusieurs dizaines de milliers d'heures de programmes. Il est donc crucial de calculer des probabilités d'exactitude des séquences d'heures et de minutes obtenues pour chaque bloc temporel (voir section 5). La Figure 4b montre un histogramme des probabilités obtenues sur les 56 blocs temporels des 26 émissions traitées pour les séquences de minutes. Il apparaît clairement que la totalité des erreurs présente une

probabilité d'obtenir la séquence inférieure à 10% (et inférieure à 5% pour la grande majorité). Ainsi, il est possible de trouver un seuil au dessus duquel la décision est quasi-certaine. A l'issu d'un traitement massif, un opérateur pourra par la suite effectuer une validation manuelle des blocs temporels pour lesquels les probabilités de séquences sont faibles.

7 Conclusion et perspectives

Cet article propose une méthode simple utilisant des techniques de traitement de la parole et du signal pour la détection et le décodage d'enregistrements de l'horloge parlante. Il s'agit là d'un enjeu de taille pour les gestionnaires d'archives comme l'Institut national de l'audiovisuel (Ina) puisque de la fin des années 1970 jusqu'au début des années 1990 un horodatage des programmes archivés a été effectué par un enregistrement de l'heure courante. Ainsi, en filigrane, le lecteur peut appréhender l'évolution des pratiques d'indexation au cours des dernières décennies.
Les résultats de l'étude montrent que de très bonnes performances peuvent être atteintes ce qui va permettre la mise en œuvre prochaine d'une chaîne de traitement pour les dizaines de milliers d'heures de programmes concernées par l'enregistrement de l'horloge parlante. Il sera à cette occasion intéressant d'évaluer l'évolution des taux d'erreur en fonction de l'année d'enregistrement pour proposer, le cas échéant, une correction adaptée. Enfin, une fois obtenues plusieurs versions des énonciations des heures et minutes constituant l'intégralité d'une journée, des méthodes de corrélation par *audio template matching* (Gaudard *et al.*, 2007) pourront éventuellement être testées, améliorant ainsi les coûts calculatoires.

Remerciements

Les auteurs souhaitent remercier Pascal Flard, Jean-Michel Rodes et Jean Varra de l'Institut national de l'audiovisuel pour leur aide précieuse.

Références

BACHIMONT, B. (1998). Bibliothèques numériques audiovisuelles. Des enjeux scientifiques et techniques. *Document numérique*, 2:219 – 242.

DELÉGLISE, P., ESTÈVE, Y., MEIGNIER, S. et MERLIN, T. (2005). The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. *In International Speech Communication Association*, Lisbon, Portugal.

GAUDARD, C., ARADILLA, G. et BOURLARD, H. (2007). Speech recognition based on template matching and phone posterior probabilities. Rapport technique, IDIAP.

SAHA, G., CHAKROBORTY, S. et SENAPATI, S. (2005). A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. *In National Conference on Communications*, Kharagpur, India.

WALKER, W., LAMERE, P., KWOK, P., RAJ, B., SINGH, R., GOUVEA, E., WOLF, P. et WOELFEL, J. (2004). Sphinx-4 : A flexible open source framework for speech recognition. Rapport technique TR-2004-139, Sun Microsystems.

Quantificateur vectoriel à divisions par classification : Application au codage des fréquences spectrales d'immitance en large-bande

Bouzid Merouane
Laboratoire Communication Parlée et Traitement du Signal (LCPTS),
Université USTHB, BP 32, El-Alia, Bab-Ezzouar, ALGER, 16111, ALGERIE.
mbouzid@usthb.dz

RESUME

Dans cet article, nous présentons un schéma de quantification vectorielle à divisions par classification développé pour le codage efficace des paramètres fréquences spectrales d'immitance ISF (Immittance Spectral Frequencies) de parole en large-bande. Ce système de codage a été conçu à base d'un quantificateur vectoriel classifié (CVQ) combiné avec un quantificateur vectoriel divisé (SVQ), selon des suppositions de transmission à travers un canal non bruité.

ABSTRACT

Classified split vector quantizer for wideband immittance spectral frequencies coding

In this paper, we present a classified split vector quantization scheme developed for efficient coding of wideband Immittance Spectral Frequencies (ISF) parameters under noiseless channel conditions. It was designed based on the classified vector quantization (CVQ) structure combined with the split vector quantization (SVQ).

MOTS-CLES : Codage de source, quantification classifiée, paramètres ISF, parole large-bande.
KEYWORDS: Source coding, classified quantization, ISF parameters, wideband speech.

1 Introduction

Dans la plupart des systèmes de codage de la parole modernes, l'enveloppe spectrale à court terme d'un signal de parole est souvent modélisée par la réponse fréquentielle d'un filtre tout-pôle dont la fonction de transfert est donnée par $H(z) = 1/A(z)$, avec $A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p}$ (Kleijn et Paliwal, 1995). Les coefficients de ce filtre $\{a_i\}_{i=1,2,\ldots p}$, connus aussi sous le nom de coefficients de prédiction linéaire (LPC), sont dérivés du signal d'entrée par une analyse par prédiction linéaire d'ordre p sur chaque trame du signal de parole. La quantification efficace des coefficients LPC joue un rôle important dans la préservation de l'intelligibilité et de la qualité naturelle du signal de parole codé sur toute la largeur de bande. En pratique, ces coefficients ne sont pas quantifiés directement 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. Les fréquences de raie spectrales LSF (Line Spectral Frequencies) et les fréquences spectrales d'immitance ISF (Immittance Spectral Frequencies) se sont avérées les plus efficaces pour représenter les coefficients LPC des codeurs de parole modernes.

Dans ce travail, on s'est intéressé particulièrement à la quantification efficace des paramètres ISF des codeurs de parole en large-bande comme le codeur large-bande à débit multiple adaptatif AMR-WB (Adaptive Multi-Rate Wide-Band) norme ITU-T G.722.2 (Bessette et al., 2002). Notons que les paramètres LSF et ISF sont des représentations équivalentes qui se déduisent mathématiquement les unes des autres par de simples fonctions de conversion. Les paramètres ISF (ISFs), qui ont été introduits par Bistritz et al. (Bistritz et al., 1989), sont définies comme étant les pôles et les zéros d'une fonction d'immitance à la glotte. Comparé aux codeurs de parole en bande étroite (300–3400 Hz), les codeurs en large-bande ont amélioré le naturel et l'intelligibilité de la parole décodée en élargissant la bande passante utilisée pour la transmission du signal de parole (50–7000 Hz). Cependant, ils exigent un nombre plus élevé de paramètres LPC, typiquement 16, pour représenter l'enveloppe spectrale de parole. Ainsi, le quantificateur vectoriel (VQ) conventionnel doit fonctionner à des débits plus élevés et sur des vecteurs de dimensions plus grandes si on veut l'utiliser pour le codage des ISFs. Ces exigences pratiques vont conduire à une augmentation excessive de la complexité des calculs et de la taille mémoire.

Dans le passé, divers schémas de quantificateurs vectoriels (VQs) structurés ont été développés pour le codage des paramètres ISF/LSF en large-bande. Ces schémas peuvent réduire considérablement la complexité des calculs avec une perte modérée des performances de quantification. L'un des schémas de codage les plus présentés dans la littérature est sans doute le quantificateur vectoriel divisé SVQ (Split Vector Quantizer), qui a été d'abord développé par Paliwal et Atal (Paliwal et Atal, 1993) pour les codeurs de parole en bande étroite et exploré par la suite dans le codage de la parole en large-bande (Chen et Wang, 1996). Dans (Chang et al., 1997), Chang et al. ont conçu un VQ classifié (CVQ) avec division à classes dépendantes pour la quantification des LSFs en bande étroite. Dans leur travail, ils ont proposé une fonction d'importance relative des LSFs afin de mettre en œuvre l'étage du classificateur. Dans (Guibè et al., 2001), Guibè et al. ont développé plusieurs quantificateurs vectoriels prédictifs et sans-mémoire pour la quantification des LSFs en large-bande. Ils ont montré que la famille des schémas de codage SNVQ (Safety-Net Vector Quantization), qui combine les approches prédictive et sans-mémoire, peut améliorer les performances de codage des LSFs en large-bande dans le cas des transmissions à travers des canaux bruités et non bruités. D'autres quantificateurs LSF/ISF récents ont été développés par So et Paliwal (So et Paliwal, 2007). Il s'agit du quantificateur vectoriel à divisions commutées et le quantificateur multi-trame à base de GMM.

Dans cet article, nous présentons un schéma de quantification vectorielle à divisions par classification que nous avons développé pour le codage efficace des paramètres ISF en large-bande selon des suppositions de transmission à travers un canal non bruité. Il s'agit d'un schéma de codage hybride conçu à base d'un CVQ combiné avec plusieurs quantificateurs vectoriels divisés (SVQs). Dans la conception de notre système, appelé "codeur ISF-CSVQ", nous avons utilisé un classifieur phonétique qui classifie les paramètres ISF de chaque trame selon une classification phonétique spécifique de la parole (voisée/non-voisée).

2 Quantificateur vectoriel à divisions par classification

Le quantificateur vectoriel à divisions par classification CSVQ (Classified Split Vector Quantizer) est un schéma de codage hybride conçu à base d'un VQ classifié combiné avec plusieurs SVQs. Tout d'abord, rappelons brièvement les principes de base du SVQ et du CVQ.

Un SVQ de dimension k et de N parties (noté N-SVQ) est composé de N quantificateurs VQ classiques de tailles et de dimensions plus petites (Paliwal et Atal, 1993). Le principe de base d'un N-SVQ consiste à partitionner l'ensemble des 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 (noté ici LBG-VQ) (Linde et al., 1980; Gersho et Gray, 1992). Un N-SVQ est donc constitué de N dictionnaires VQ de tailles et de dimensions plus petites comparé à un VQ conventionnel.

D'autre part, le CVQ est constitué d'un classifieur et de dictionnaires VQ séparés pour chaque classe. La figure 1 présente le schéma bloc d'un CVQ conventionnel de M classes. Au début, la classe d'appartenance du vecteur d'entrée x est déterminée parmi un nombre prédéterminé de classes. Chaque classe j ($j = 1,..., M$) est représentée par un dictionnaire VQ correspondant C_j . Par la suite, une recherche totale est effectuée à travers le dictionnaire VQ de la classe sélectionnée afin de trouver le meilleur vecteur-code pour le vecteur x suivant le critère du plus proche voisin. Le mot-code i transmit par le codeur CVQ est composé de l'indice j spécifiant le dictionnaire sélectionné C_j et de $\log_2 || C_j ||$ bits spécifiant le plus proche vecteur-code sélectionné de C_j. Notons que dans la version originale du CVQ, les dictionnaires des classes C_j de taille N_j sont considérés comme des sous-dictionnaires d'un dictionnaire optimal global C de taille N tels que $\sum_{j=1}^{M} C_j = C$ et $\sum_{j=1}^{M} N_i = N$ (Gersho et Gray, 1992).

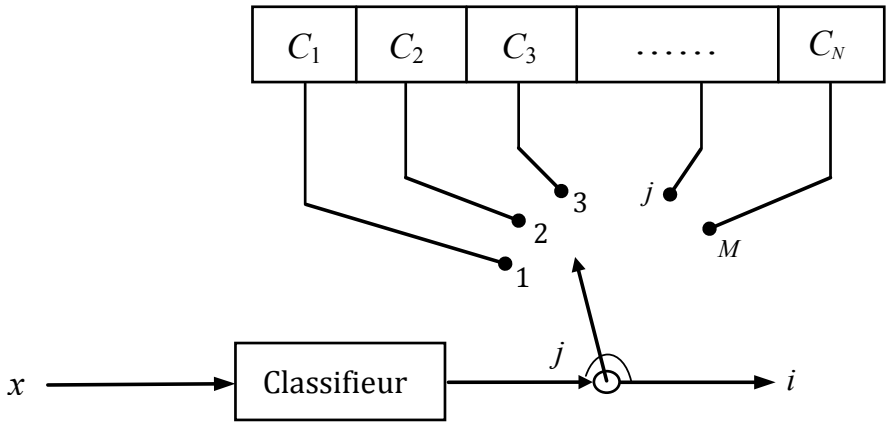


FIGURE 1 – Schéma-bloc d'un quantificateur CVQ.

Le CSVQ est une sorte de VQ à code produit où l'espace des vecteurs de la base d'apprentissage est divisé en plusieurs classes disjointes et un SVQ correspondant est conçu séparément pour chaque classe (Chang et al., 1997). La conception d'un classifieur précis est le problème principal dans tout système CVQ. Dans le cadre de la quantification des ISFs en large-bande, l'objectif est d'affecter les vecteurs ISF dans des classes ayant un comportement statistique particulier afin d'améliorer l'efficacité de codage. Dans la conception de notre codeur CVSQ, nous avons adopté un classifieur de type phonétique qui classifie les paramètres ISF en voisés et non-voisés. Il s'agit d'une approche de classification vocale qui distingue entre les ISFs des trames de parole voisée (V) et non voisée (NV). En effet, R. Hagen et al. ont montré dans (Hagen et al., 1999) qu'un gain significatif peut être obtenu par la conception de dictionnaires VQ séparément pour la parole voisée et non voisée, plutôt que

d'utiliser un seul dictionnaire VQ.

3 Codage efficace des paramètres ISF en large-bande

En utilisant la technique CSVQ, un schéma de codage efficace a été développé pour quantifier les paramètres spectraux ISF en large-bande. L'objectif de ce système, appelé "codeur ISF-CSVQ", est de réaliser une quantification de qualité transparente des ISFs pour des transmissions à travers un canal idéal non bruité.

Le codeur ISF-CSVQ a été conçu à base d'un CSVQ phonétique où chaque classe V/NV est représentée par un *N*-SVQ correspondant. Ainsi, les vecteurs ISF de la base d'apprentissage sont classifiés en deux classes V/NV à l'aide d'un classificateur phonétique. Un *N*-SVQ correspondant séparé est ensuite conçu pour chaque classe afin de préserver la caractéristique V/NV associée à chaque région. Dans notre travail, nous avons utilisé le détecteur d'activité vocale intégrée VAD (Bessette et al., 2002; ITU-T Rec. G.722.2, 2003) du G.722.2 pour effectuer la détection vocale au niveau du classifieur. Il s'agit d'utiliser l'indicateur VAD booléen qui indique la présence ("1") ou non ("0") de trame voisée.

La figure 2 présente le schéma-bloc du principe de codage par ISF-CSVQ. Chaque vecteur ISF à quantifier est d'abord classifié à un vecteur issu d'une trame de parole V ou NV. Par la suite, le vecteur sera quantifié par le *N*-SVQ correspondant sélectionné par le classifieur V/NV. Le codeur CSVQ transmet au décodeur un indice i composé de $N + 1$ indices binaires concaténés. Le premier indice i_m de 1 bit/trame désigne le mode V/NV sélectionné. Les N indices i_n ($n = 1,\ldots, N$) restants sont fournit par le *N*-SVQ correspondant à la direction i_m.

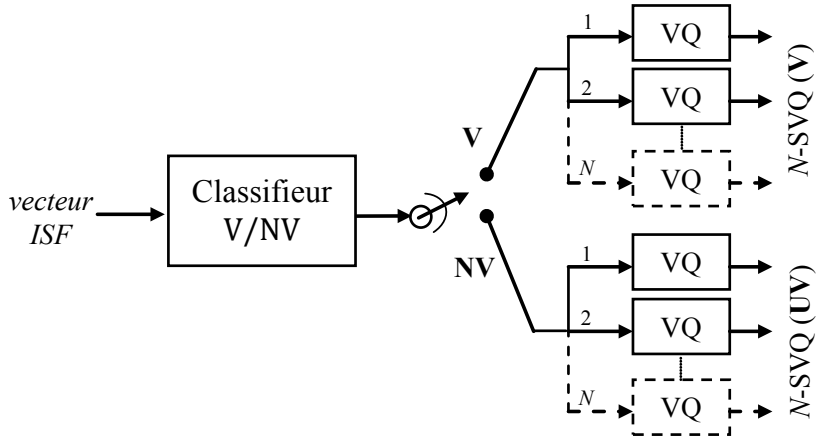


Figure 2 – Codage ISF-CSVQ.

Pour un ISF-CSVQ opérant à un débit de b bits/trame, on a besoin de $b_m = 1$ bit pour identifier la classe V/NV. Le débit restant de $(b - b_m)$ bits sera partagé en N débits partiels pour coder les N sous-vecteurs ISF par le *N*-SVQ correspondant. Ces débits partiels sont notés par b_j ($j = 1\ldots N$). Le décodeur CSVQ, qui possède les mêmes dictionnaires que ceux du codeur, reçoit donc l'indice envoyé $i = (i_m\ i_n)$. Il utilise le premier indice i_m pour repérer la classe sélectionnée. Ensuite, il construit le vecteur ISF décodé en concaténant les sous vecteurs-code d'indices i_n ($n = 1,\ldots, N$) délivrés par le *N*-SVQ de la classe sélectionnée.

Afin d'améliorer davantage les performances de nos codeurs ISF-CSVQ et obtenir une quantification transparente à des débits plus bas, nous avons utilisé une mesure de distance

euclidienne pondérée plus appropriée (Paliwal et Atal, 1993; Bouzid et al., 2005) :

$$d(f,\hat{f}) = \sum_{i=1}^{16} [w_i (f_i - \hat{f}_i)]^2 , \quad (1)$$

où f_i et $\hat{f}_i$ sont respectivement les $i^{ème}$ coefficients des ISFs original f et quantifié $\hat{f}$ et w_i représente le poids spectral (Paliwal et Atal, 1993) assigné au $i^{ème}$ coefficient du vecteur ISF.

Les performances de nos codeurs des ISFs sont évaluées par la distorsion spectrale SD (Spectral Distortion) moyenne. L'expression de la SD pour une trame i est donnée en décibels par (Paliwal et Atal, 1993; Bouzid et al., 2005; So et Paliwal 2007):

$$SD_i = \sqrt{\frac{1}{n_1 - n_0} \sum_{n=n_0}^{n_1-1} \left[10 \log_{10} \frac{S(e^{j2\pi n/N})}{\hat{S}(e^{j2\pi n/N})} \right]^2} , \quad (2)$$

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 (Paliwal et Atal, 1993) : 1)- la SD moyenne est d'environ 1 dB, 2)- aucune trame "Outliers" ne doit avoir une SD qui dépasse 4 dB et 3)- le pourcentage des trames "Outliers" ayant une SD entre 2 et 4 dB doit être inferieur à 2%. Suivant Guibé et al. (Guibé et al., 2001), les tests d'écoute ont montré que ces conditions de quantification transparente, qui sont souvent utilisés dans le cas du codage de la parole en bande étroite, sont aussi valables dans le cas du codage en large-bande. D'autre part, des expériences réalisées par Hagen et al. (Hagen et al., 1999), ont montré que ces critères de transparence ne sont pas vraiment valides dans le cas de la quantification LPC pour des trames non voisées. Ainsi, ces critères doivent être définis individuellement pour les classes voisées et non voisées. Les nouvelles conditions de codage transparent pour les trames NV sont donc :

- La SD moyenne ne doit pas dépasser 2 dB.
- Le pourcentage des trames ayant une SD supérieure à 4 dB doit être inferieur à 1%.

Aussi, une des conditions du codage transparent a été légèrement changée pour la quantification LPC des trames voisées (Hagen et al., 1999). En effet, le pourcentage des trames "Outliers" ayant une SD entre 2 et 4 dB doit être moins de 1%.

La base de données parole utilisée dans ce travail se compose d'environ 85 minutes de parole prise de la base de données internationale TIMIT, avec une fréquence d'échantillonnage de 16 kHz (Garofolo et al., 1988). Pour construire la base des vecteurs ISF, nous avons utilisé la même fonction d'analyse LPC du standard AMR-WB G.722.2 (Bessette et al., 2002) où une analyse LPC d'ordre 16, par la méthode d'autocorrélation, est effectuée sur chaque trame d'analyse de 20 ms. Une partie de la base de données ISF (208363 vecteurs ISF) est utilisée pour l'apprentissage et la partie restante, de 48606 vecteurs ISF (différente de la base d'apprentissage), est utilisée pour les tests. Les bases de données d'apprentissage et de test ont été séparées en sous bases voisée et non voisée en utilisant la procédure VAD du G.722.2. Dans la base d'apprentissage, 39% des trames ont été classées comme voisées et

61% ont été classées comme non voisées alors que pour la base de test, 32% des trames ont été classées comme voisées et 68% ont été classées comme non voisées.

Dans la conception du codeur ISF-CSVQ, nous avons utilisé des *N*-SVQs à différents débits pour le codage des classes V/NV où des *N*-SVQs à débits plus faibles ont été utilisés pour la classe NV. Le codeur ISF-CSVQ est donc caractérisé par son débit moyen, en fonction de la proportion des vecteurs ISF quantifiés par les *N*-SVQs des classes NV et V, respectivement (Guibé et al., 2001). Le nombre moyen de bits par trame est donc calculé par:

$$b_{Moy} = 1 + b_V N_V + b_{NV} N_{NV} \,, \tag{3}$$

où b_V et b_{NV} sont respectivement les débits des *N*-SVQs des classes voisées et non voisées, et N_V *et* N_{NV} sont les pourcentages des trames voisées et non voisées, respectivement.

Pour différents débits de codage, les performances des codeurs ISF-CSVQ pour les trames voisées et non voisées sont données respectivement dans les Tables 1 and 2. Notons que dans ces exemples de simulation, nous avons utilisé un SVQ de 5 parties (5–SVQ) pour chaque classe où les vecteurs ISF large-bande de dimension 16 sont divisés en 5 parties suivant la division (3 – 3 – 3 – 3 – 4). Concernant l'allocation des bits, ces derniers sont uniformément alloués dans chaque classe; ceci bien sûr dans la mesure du possible.

TABLE 1 –Performances des codeurs ISF-CSVQ pour des trames voisées

Débit - Bits/trame b ($b_1 + \ldots + b_5$)	SD Moy. (dB)	Trames "Outliers" (en %)	
		2-4 dB	> 4 dB
49 (10+10+10+10+9)	1.084	0.534	0.00
48 (10+10+10+9+9)	1.120	0.610	0.00
47 (10+10+9+9+9)	1.159	0.922	0.00
46 (10+9+9+9+9)	1.201	1.284	0.00

TABLE 2 –Performances des codeurs ISF-CSVQ pour des trames non voisées

Débit - Bits/trame b ($b_1 + \ldots + b_5$)	SD Moy. (dB)	Trames "Outliers" > 4 dB (en %)
28 (6+6+6+5+5)	1.973	0.596
27 (6+6+5+5+5)	2.064	1.113
26 (6+5+5+5+5)	2.168	1.703
25 (5+5+5+5+5)	2.228	1.876

Pour des transmissions à travers un canal non bruité, ces résultats de simulation montrent que le codeur large-bande ISF-CSVQ peut réaliser une quantification de qualité transparente à 49 bits/trame et 28 bits/trame pour des trames voisées et non voisées, respectivement. Ainsi, le codeur ISF-CSVQ assure la quantification transparente à un débit moyen b_{Moy} de 1 + 0.32 (49) + 0.68 (28) = 35.72 bits/trame.

Pour une évaluation comparative entre le codeur ISF-CSVQ et le codeur non classifié ISF-SVQ conçu par le SVQ conventionnel, nous présentons ci-dessous les performances du codeur ISF-SVQ de 5 parties opérant dans les mêmes conditions précédentes. Notons aussi que les codeurs ISF-SVQ conventionnels de 5 parties ont été conçus avec la même base d'apprentissage complète de 208363 vecteurs ISF non-classifiés où chaque vecteur ISF est divisé également suivant la division (3 - 3 - 3 - 3 - 4).

TABLE 3 –Performances des codeurs conventionnels ISF-SVQ de 5 parties

Débit - Bits/trame b (b_1 +...+ b_5)	SD Moy. (dB)	Trames "Outliers" (en %)	
		2-4 dB	> 4 dB
46 (10+9+9+9+9)	1.08	1.113	0.00
45 (9+9+9+9+9)	1.107	1.158	0.00
44 (9+9+9+9+8)	1.156	1.818	0.00
43 (9+9+9+8+8)	1.198	2.501	0.00
42 (9+9+8+8+8)	1.248	3.489	0.00

En comparant ces résultats avec ceux données dans les Tables 1 et 2, on remarque clairement que le CSVQ apporte une amélioration significative aux performances d'encodage des paramètres ISF comparé au SVQ conventionnel. En effet, le codeur ISF-SVQ de 5 parties a besoin de 46 bits/trame pour réaliser une quantification de qualité transparente alors qu'un débit moyen d'environ 36 bits/trame est suffisant pour le ISF-CSVQ afin d'assurer le codage transparent. Ce gain en performance SD apporté par le CSVQ est le résultat de l'utilisation d'une classification V/NV des vecteurs ISF d'entrée avant de les coder par un *N*-SVQ correspondant plutôt que d'utiliser un seul *N*-SVQ non-classifié.

4 Conclusion

Dans ce travail, un système de quantification vectorielle à divisions par classification V/NV a été développé pour le codage efficace des paramètres ISF de parole en large-bande. Comparé au codeur ISF-SVQ conventionnel, le ISF-CSVQ peut fournir de meilleures performances, tout en économisant plusieurs bits par trame.

Le comportement du codeur ISF-CSVQ en présence des erreurs de canal reste à être étudié. En outre une étude soigneuse de la robustesse du codeur, en tenant compte des changements dans les conditions d'enregistrement des signaux parole, est aussi nécessaire.

Références

BESSETTE, B., SALAMI, R., LEFEVRE, R., JELINEK, M., ROTOLA-PUKKILA, J., VAINIO, J., MIKKOLA, H. and JARVINEN, K. (2002). The adaptive multirate wideband speech codec (AMR-WB), *IEEE Transactions on Speech and Audio Processing*, 10(8):620-636.

BISTRITZ, Y., LEV-ARI, H. and KAILATH, T. (1989). Immittance domain levinson algorithms. *IEEE transactions on information theory*, 35:675–682.

BOUZID, M., DJERADI, A. and BOUDRAA, B. (2005). Optimized Trellis Coded Vector Quantization of LSF Parameters: Application to the 4.8 Kbps FS1016 Speech Coder, *Signal Processing Elsevier*, 85(9):1675-1694.

CHANG, D., ANN, S. and LEE, C. W. (1997). A classified vector quantization of LSF parameters, *Signal Processing*, 59:267-273.

CHEN, J. H. and WANG, D. (1996). Transform predictive coding of wideband speech signals. *In Proc. of the ICASSP'96*, Atlanta, USA, pages 275–278.

GAROFOLO, J. S. et al. (1998). DARPA TIMIT Acoustic-phonetic Continuous Speech Database, National Institute of Standards and Technology (NIST), Gaithersburg.

GERSHO, A. and GRAY, R. M. (1992). *Vector quantization and Signal compression*, Kluwer Academic Publishers, USA.

GUIBE, G. H., HOW, T. and HANZO, L. (2001). Speech spectral quantizers for wideband speech coding, *European Transactions on Telecommunications*, 12(6):535–545.

HAGEN, R., PAKSOY, E. and GERSHO, A. (1999). Voicing-Specific LPC Quantization for Variable-RateSpeech Coding, *IEEE Transactions on Speech Audio Processing*, 7(5):485-494.

ITU-T Recommendation G.722.2. (2003). Wideband coding of speech at around 16 kb/s using Adaptive Multi-rate Wideband (AMR-WB).

KLEIJN, W. B. and PALIWAL, K. K. (1995). *Speech coding and synthesis*. Elsevier Science B.V.

LINDE, Y., BUZO, A. and GRAY, R. M. (1980). An Algorithm for Vector Quantization Design, *IEEE Transactions on Communications*, 28:84-95.

PALIWAL, K. K. and ATAL, B. S. (1993). Efficient vector quantization of LPC parameters at 24 bits/frame, *IEEE Transactions on Speech and Audio Processing*, 1(1):3-14.

SO, S. and PALIWAL, K. K. (2007). Comparative study of LPC parameter representations and quantisation schemes for wideband speech coding, *Digital Signal Processing*, 17:114-137.

Utilisation d'un algorithme A* pour l'analyse de la sélection d'unités en synthèse de la parole

David Guennec[1] Damien Lolive[1]
(1) IRISA - University de Rennes 1, Lannion, France
david.guennec@irisa.fr, damien.lolive@irisa.fr

RÉSUMÉ

Les systèmes de synthèse de la parole par corpus utilisent presque tous un algorithme de Viterbi pour réaliser l'étape de sélection d'unités. Dans cet article, nous proposons un système de synthèse par sélection reposant sur l'algorithme A^* et basé sur une stratégie généraliste de recherche de chemins dans un graphe (et non un treillis). Notre implémentation emploie des techniques tirées de l'état de l'art. Elle est pensée en tant que plateforme expérimentale permettant d'obtenir aisément les N meilleurs chemins et d'évaluer de nouvelles stratégies de sélection. Une évaluation perceptuelle notant la qualité globale du système et le classement des chemins trouvés par la sélection d'unités a, en outre, été conduite.

ABSTRACT

Using a A* algorithm for unit selection analysis in TTS

Common corpus-based speech synthesis systems use the Viterbi algorithm to drive the unit selection step. In this paper, we propose a system relying on the A^* algorithm, a general pathfinding strategy developing a graph (rather than a lattice). The implementation we propose uses state of the art techniques and is built as an experimental platform providing easy obtainment of the N-best paths as well as evaluation of new selection strategies. A perceptual evaluation assessing the overall quality of the system output and the ranking of the paths returned by unit selection has also been conducted.

MOTS-CLÉS : Synthèse de la parole par sélection d'unités, Évaluation, Précision des fonctions de coût, Coût de concaténation, Pondération, Utilisation du corpus.

KEYWORDS: Corpus-based Unit Selection TTS, Evaluation, Cost function accuracy, Concatenation cost, Cost weighting, Corpus usage.

1 Introduction

Il existe aujourd'hui deux grandes techniques dans le champ de la synthèse de la parole. L'une d'entre elles est l'approche paramétrique pour laquelle HTS (Yamagishi *et al.*, 2008) est le principal système. La seconde est l'approche historique par concaténation étendue par (Sagisaka, 1988) et largement utilisée dans de nombreux systèmes (Black et Taylor, 1994; Hunt et Black, 1996; Taylor *et al.*, 1998; Breen et Jackson, 1998; Clark *et al.*, 2007). Même si la majorité des travaux de recherche actuels se concentrent autour d'HTS, l'approche par sélection d'unités de taille variable

demeure la méthode privilégiée aujourd'hui dans les systèmes commerciaux.

Dans le cadre de la synthèse par sélection d'unités, la question fondamentale est comment trouver de manière efficace la séquence d'unités la plus appropriée dans un corpus de parole annoté ? Cette séquence d'unités la plus appropriée peut être définie par deux critères : elle doit être la plus proche possible de la séquence à produire et minimiser les risques de générer des artéfacts de concaténation audibles (minimiser le nombre de concaténations). De nombreux critères ont été proposés (Yi, 1998), mais généralement la fonction de coût à optimiser est de la forme (Black et Taylor, 1994) :

$$U^* = \underset{U}{argmin} \left(\sum_{n=1}^{card(U)} C_t(u_n) + \sum_{n=2}^{card(U)} C_c(u_{n-1}, u_n) \right) \quad (1)$$

où U^* est la meilleure séquence d'unités selon la fonction de coût et u_n est l'unité candidate correspondant à la n^e unité cible dans la séquence candidate U. Le sous-coût $C_t(u_n)$ (coût cible, ou *target cost*) représente la distance entre le candidat et la cible correspondante. Le second sous-coût, $C_c(u_{n-1}, u_n)$ (coût de concatenation, ou *concatenation cost*), correspond à la distance entre l'unité candidate courante u_n et la précédente u_{n-1} sur le chemin considéré. Il permet de minimiser les artéfacts dans les zones de concaténation.

En raison de la complexité du problème (plusieurs millions d'unités candidates), un élagage du graphe exploré est plus que souhaitable si l'on souhaite obtenir un résultat dans un temps acceptable (en temps réel pour la plupart des applications industrielles). Pour la recherche de la meilleure séquence d'unités, l'algorithme de Viterbi (Viterbi, 1967) est presque toujours l'alternative utilisée (dans (Hunt et Black, 1996; Conkie *et al.*, 2000; Clark *et al.*, 2007) par exemple). L'algorithme de Viterbi est optimal et possède une complexité de $\mathcal{O}(N * K^2)$, K étant le maximum d'unités candidates pour un phonème cible donné. Ce n'est cependant pas le seul algorithme disponible. Une tentative d'utiliser un algorithme génétique par (Kumar, 2004) peut par exemple être citée. Cependant, certains algorithmes tels que A^* sont biens plus appropriés à la mise en place d'heuristiques pour accélérer l'étape de sélection d'unités.

Par ailleurs, la principale problématique de recherche est le contrôle de l'expressivité durant la synthèse (Govind et Prasanna, 2013). Ainsi, pour examiner le problème de la synthèse expressive par corpus, il est nécessaire d'explorer en profondeur le comportement du système de synthèse et tout particulièrement celui des mécanismes de notation et de sélection des unités.

Dans cet article, un nouveau système de synthèse de la parole par sélection d'unités, basé sur un algorithme A^* et bâti de manière à faciliter l'analyse du comportement des fonctions de coût et des stratégies d'exploration, est décrit en section 2. L'architecture générale du système est explicitée en section 3. Le corpus employé pour les expérimentations est décrit en section 4 et le protocole expérimental, visant à montrer les capacités du système, est décrit en section 5. Enfin, nos résultats sont présentés en section 6.

2 Recherche de la meilleure séquence d'unités

Dans cette section, nous présentons l'algorithme A^* dans le cadre d'une recherche de séquencement optimal d'unités candidates d'après une fonction de coût donnée. Certains détails concernant son implémentation sont ensuite décrits.

2.1 L'algorithme A*

Contrairement à l'algorithme de Viterbi, lequel calcule un treillis contenant tous les nœuds candidats (ou au moins M nœuds à chaque instant), A^* développe un graphe. L'avantage de développer un graphe est que l'on peut employer des mécanismes plus généraux et des heuristiques plus globales tout en relachant les contraintes liées aux treillis. Cet algorithme, à l'origine proposé dans (Hart *et al.*, 1968), recherche un meilleur chemin dans un graphe orienté, dont les nœuds ont un nombre fini de successeurs, entre un nœud de départ et un nœud cible. De manière à éviter tout choix arbitraire du nœud de départ, nous introduisons un nœud de départ s ayant comme successeurs les premières unités candidates ; un choix similaire est fait en introduisant une cible unique t. L'algorithme utilise une fonction de coût de la forme $f(n) = g(n) + h(n)$ où $g(n)$ est le coût du sous chemin entre s et le nœud courant n ainsi que $h(n)$ qui est le coût estimé entre n et t.

À chaque étape, A^* prend le nœud le plus prometteur selon $f(n)$ et recherche ses successeurs (calculant ainsi $f(n)$ pour ce nœud) jusqu'à ce que t soit atteint. $h(n)$ est une heuristique permettant d'accélérer l'algorithme en accordant un meilleur coût aux nœuds qui semblent être sur un chemin optimal que ceux qui possèdent un meilleur $g(n)$ mais risquent de mener à des coût plus élevés dans le futur (Nilsson, 1982). Si l'on considère un nœud cible unique, l'un des principaux avantages de A^* est que l'algorithme délivre une solution optimale si l'heuristique est admissible, c'est-à-dire si $h(n) \leq h^*(n)$, $h^*(n)$ étant la valeur de la distance minimale réelle jusqu'au nœud cible. Il convient de noter que l'algorithme est optimal dans le cas trivial $h(n) = 0$, ie. s'il n'y a pas d'heuristique, et devient alors équivalent à un algorithme de Dijkstra.

L'avantage d'une exploration de type A^* est que l'on développe généralement bien moins de noeuds qu'un parcours en profondeur type Viterbi. En effet, dans le cas d'A^*, la politique est d'étendre les chemins les plus prometteurs, ce qui en soi, constitue une heuristique permettant de tendre au plus vite vers le noeud cible. Dans le cas d'un Viterbi, la recherche implique de calculer entièrement le treillis, et donc tous les noeuds, avant de pouvoir trouver le chemin optimal. Au pire, A^* développe tout les noeuds du graphe, dont la taille n'excède pas celle du treillis du Viterbi, et beaucoup moins de noeuds dans la majorité des cas.

D'autre part, si Viterbi peut être accéléré par une recherche en faisceau, ou *beam search* (au prix de l'optimalité), le même mécanisme est envisageable avec A^*, via un bridage de la liste open (il suffit de passer directement les noeuds les moins prometteurs dans la liste close).

2.2 Adaptation à la sélection d'unités

En considérant l'algorithme présenté ci-dessus, les principaux aménagements à réaliser sont (1) le calcul de la fonction de coût et (2) la fonction de recherche des successeurs. Dans ces travaux, nous ne considérons que la partie $g(n)$ de la fonction de coût, positionnant ainsi $h(n)$ à 0 ce qui assure l'optimalité de l'algorithme. La fonction de coût est détaillée dans la section 3.2. L'inclusion d'heuristiques est planifiée pour les expérimentations suivantes.

En ce qui concerne la fonction de recherche des successeurs, il s'agit d'y ajouter les connaissances liées au domaine. Au cours du processus de recherche, chaque phonème de la séquence cible est considéré comme le début d'une unité potentielle permettant de développer le graphe. En outre, pour une unité particulière, tous les candidats au sein d'une fenêtre de longueur $l = 5$ sont

explorés. La taille de la fenêtre a été choisie arbitrairement de manière à garantir un temps de recherche satisfaisant.

De plus, de manière à améliorer les performances algorithmiques, la liste OPEN est implémentée sous la forme d'un tas binaire trié en fonction du coût d'un nœud et d'une table de hachage afin de réaliser rapidement des requêtes d'appartenance. Tous les nœuds du graphe ne sont pas calculés : seuls ceux découverts lors de la recherche des successeurs sont réellement créés.

Pour explorer le fonctionnement des fonctions de coût, nous avons modifié l'algorithme de manière à pouvoir obtenir, dans l'ordre, les N meilleurs chemins, ce qui est très simple, il suffit de ne pas s'arrêter immédiatement après avoir trouvé le chemin optimal. Nous pouvons également obtenir facilement les N meilleurs chemins entre un coût minimum et un coût maximum.

3 Architecture du système

Le système est interfacé avec le toolkit ROOTS décrit dans (Boeffard *et al.*, 2012). Pour la synthèse, l'extraction des informations linguistiques est faite depuis des fichiers ROOTS contenant la séquence cible avec les annotations voulues. L'étape de sélection d'unités est ensuite réalisée en utilisant l'algorithme A^* présenté en section 2. Pour cette étape, on peut renseigner une fonction de coût particulière et des paramètres utilisateurs (par exemple la demande du meilleur chemin ou bien celle des N meilleurs). Le processus de sélection effectue des demandes au corpus via des filtres de préselection comme indiqué en section 3.1. La génération du signal de parole à proprement parler se fait par le mélange des unités de gauche et de droite sur un horizon de deux périodes de pitch en tout, en utilisant des fenêtres de Hann. Elle peut fonctionner à partir d'une liste de chemins, la rendant parfaite pour une exploration des résultats des fonctions de coût.

3.1 Filtres de pré-sélection

Comme le problème de la recherche d'unités de taille variable dans un corpus est particulièrement coûteux calculatoirement, des tables de hachage et des filtres de pré-sélection ont été implémentés de façon à accélérer le processus de sélection (Conkie *et al.*, 2000). Pour cela, chaque phone dans le corpus possède une clé binaire permettant de prendre ou rejeter une unité avant même son intégration dans l'A^*. La clé contient des informations relatives aux niveaux phonétiques, linguistiques et prosodiques. Des masques binaires sont utilisés pour obtenir uniquement les informations désirées à chaque instant du calcul. Les filtres de pré-sélection sont intégrés aux fonctions de hachage utilisées pour accéder aux unités du corpus de manière à réduire le nombre de candidats ajoutés à la liste OPEN. L'ensemble complet des filtres utilisés est le suivant :

1. Le phone est-il un «*Non Speech Sound*» (NSS) ?
2. Le phone est-il dans l'onset de la syllabe ?
3. Le phone est-il dans la coda de la syllabe ?
4. Le phone est-il dans la dernière syllabe de son groupe de souffle ?
5. La syllabe courante est-elle en fin de mot ?
6. La syllabe courante est-elle en début de mot ?

L'accélération de la recherche n'est que l'une des fonctions des filtres. Leur deuxième fonction est de réduire l'ensemble des unités disponibles en ne gardant que celles qui s'avèrent les plus proches (caractéristiques linguistiques, phonétiques, contexte, ...) de l'unité cible. Les filtres de pré-sélection sont assimilables à des coûts cibles binaires.

La volonté d'appliquer des filtres caractérisant la position du candidat dans la syllabe est motivée par le fait qu'un même phonème selon qu'il est au début ou à la fin d'une syllabe (ie. avant ou après le noyau) se trouve altérée différement (au niveau du spectre notamment). Les filtres sur les mots sont basés sur le même genre de considérations. Enfin, la prosodie de fin de groupe de souffle est généralement différente du reste de la phrase, caractérisée principalement par une chute du F0. Il est donc plus que souhaitable d'utiliser une ou des unités elles mêmes en fin de groupe de souffle pour générer cette syllabe.

Si aucune unité correspondant à l'ensemble des contraintes n'est trouvée, les filtres de pré-sélection sont relâchés un par un, en commençant par la fin de la liste ci-dessus. Un désavantage de cette technique est que l'on peut explorer des candidats dont les propriétés sont éloignées de celles de l'unité cible, risquant aussi de produire des artéfacts de concaténation mais ce mécanisme de retour en arrière nous assure de trouver une unité et donc de produire une solution.

Les unités retenues (pour un même niveau de filtre) sont équivalentes du point de vue de la distance à l'unité cible correspondante. Nous considérons donc qu'étant donné un nombre satisfaisant de filtres de pré-sélection, l'apport du coût cible dans la fonction de coût (cf. eq. 1) ne se justifie plus. Avoir plus de filtres modélisant le coût cible plutôt qu'une fonction de coût (cible) procure un énorme avantage : au lieu d'avoir un grand nombre de nœuds à inclure et noter dans le graphe du problème (lourde charge calculatoire), seuls les meilleurs nœuds cibles (accès à la table de hachage via les filtres en $\mathcal{O}(1)$) sont classés en fonction du coût de concaténation.

3.2 La fonction de coût

Étant donné le point précédent sur le coût cible, notre fonction de coût $g(n)$ d'une unité contient uniquement un coût de concaténation. Une pondération des différents sous-coûts est introduite de manière à compenser les différences d'échelles (en d'autres termes, une importance équivalente est attribuée à chaque sous-coût). Trois sous coûts, couramment dans l'état de l'art, sont considérés ici pour l'évaluation de la qualité des concaténations : les distances calculées entre unité gauche et droite sur les MFCC, l'amplitude et le f0. Nous avons ainsi le coût de concaténation suivant :

$$C_c(u,v) = W_{mfcc}(u,v)C_{mfcc}(u,v) + W_{amp}(u,v)C_{amp}(u,v) + W_{F0}(u,v)C_{F0}(u,v)$$

où u et v sont les deux unités comparées, $W_{mfcc}(u,v)$ est le cœfficient de normalisation pour le coût des MFCC, $C_{mfcc}(u,v)$, $W_{amp}(u,v)$ est le cœfficient de normalisation pour le coût d'amplitude $C_{amp}(u,v)$ and $W_{F0}(u,v)$ est le cœfficient de normalisation pour le coût du F0 $C_{F0}(u,v)$.

L'apprentissage des poids est un problème à part entière, et de nombreuses méthodes ont été expérimentées. (Alías *et al.*, 2011) est une bonne revue des techniques les plus courantes. En fait, en l'absence de méthodes objectives suffisamment corrélées avec les évaluations perceptives, les méthodes subjectives de réglage des poids restent très utilisées dans les systèmes de synthèse. Nous considérons ici que les poids sont liés au phonème à concaténer. Ils sont donc calculés via des moyennes et écarts-types pour chaque phonème dans le corpus.

4 Analyse du corpus

Pour chaque expérience, c'est un corpus de voix française masculine et expressive, provenant d'un audiobook, qui est utilisé. Le corpus n'a jamais été utilisé antérieurement en synthèse. Le corpus est annoté automatiquement en utilisant le processus global décrit dans (Boeffard *et al.*, 2012). Il utilise la représentation du toolkit ROOTS (Chevelu *et al.*, 2014). Les signaux de parole sont échantillonnés à 44.1 kHz et la valeur de F0 moyenne, en considérant uniquement les segments de parole voisés, est de 87 Hz. Comme il s'agit d'un audiobook, le contenu du corpus comme son expressivité est incontrôlé, ce qui peut présenter des problèmes. Par exemple, comme la voix n'a pas été enregistrée à des fins de synthèse, la prosodie est parfois exagérée. Un tel corpus est toutefois intéressant en raison de la variabilité présente comme pour la haute qualité des signaux de parole.

En ce qui concerne la taille du corpus complet, il est constitué de 3 339 énoncés, totalisant 419 742 segments (388 251 phonèmes & 31 491 *Non Speech Sounds*). La durée totale du corpus est de 10 h 45 min. Les diphonèmes (et autres *Non Speech Sounds*) sont couverts à 78%. Les diphones restants sont la plupart du temps soit inutilisés soit inexistants en français. Le corpus complet à été divisé en trois parties : apprentissage (3 139 énoncés), test (100 énoncés) et validation (100 énoncés).

5 Expériences

5.1 Protocole expérimental

Les expériences que nous avons effectué visent à :

— Démontrer la capacité du système à extraire les N meilleurs chemins facilement.
— Évaluer les performances générales du système en terme de qualité globale. Il s'agit de déterminer si l'approche A^* est viable.
— Vérifier la stabilité de la fonction de coût présentée ci-dessus en analysant la variabilité et la qualité du classement pour les 100 meilleurs chemins trouvés par notre système.

La figure 1 montre l'évolution du coût global pour les 100 premiers chemins. La phrase cible est «Car ce n'est pas le chagrin qui la fit partir». En raison des différences entre les unités sélectionnées dans les différents chemins, les coûts (axe des y) sont reportés à chaque phonème de la séquence (axe des x). À première vue, on observe peu de variabilité au sein des chemins. La plus grande partie des changements semblent se produire sur les premières/dernières unités (des *Non Speech Sounds* en réalité). Par manque de place, nous ne pouvons pas illustrer le fait que les sous-coûts se compensent entre eux.

Pour synthétiser le corpus de test, il a fallu utiliser au total 14 014 phones du corpus, dont 12 607 distincts. En outre, l'unité la plus utilisée l'a été 123 fois (inspiration puis pause) tandis que 11 620 phones n'ont été sélectionnées qu'une seule fois. En tout, seulement 7 phones ont été sélectionnés plus de 5 fois, ce qui tend à montrer qu'il existe très peu d'unités à «forte valeur ajoutée» dans le corpus. Ce résultat est étonnant et pose l'interrogation suivante, à laquelle de prochains travaux devront répondre : le taux de réutilisation des phones est-il si faible en raison du fonctionnement de l'A^* ou est-il simplement lié à notre corpus ? Concernant la taille des segments contigus de phones sélectionnés, la moyenne pour les phrases générées est de 3,8

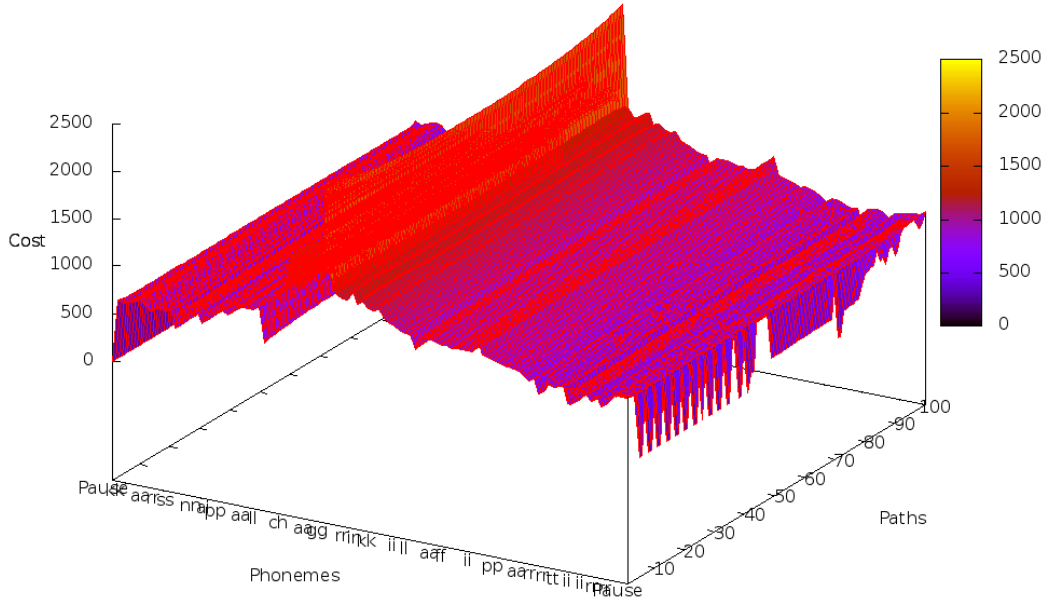


FIGURE 1 – Évolution du coût global pour les 100 premiers chemins (phrase cible «Car ce n'est pas le chagrin qui la fit partir»).

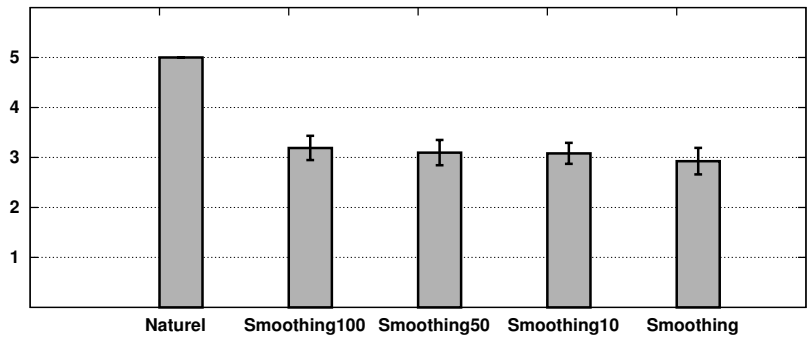


FIGURE 2 – Résultats DMOS pour les 1[er], 10[ème], 50[ème] et 100[ème] chemins pour la fonction de coût *Smoothing*.

phones, ce qui est satisfaisant.

En outre, des tests d'écoute ont également été effectués. Pour chacun d'entre eux, nous avons suivi la recommandation ITU-T P.800, avec 8 testeurs experts, les questions/réponses étant celles décrites dans la recommandation.

6 Résultats et discussion

Un test MOS visant à évaluer les performances globales du système avec la fonction de coût décrite en section 3.2 (*Smoothing* dans la suite) a été réalisé. Chaque testeur évalue 20 échantillons pour chaque système. Au total, notre système obtient un score MOS de 3.29 (±0.18).

Étant donné ces résultats, nous avons décidé d'effectuer un test de type DMOS visant à comparer plusieurs choix différents du A* pour une même phrase. À chaque fois, le signal naturel est confronté à un signal de synthèse correspondant soit au 1[er], 10[ème], 50[ème] ou 100[ème] chemin. Chaque auditeur entend le même nombre de phrases pour chaque système. La figure 2 montre les

résultats du test. Les scores pour tous les chemins sont approximativement de 3, traduisant une dégradation légèrement ennuyeuse. Étant donné le recouvrement des intervalles de confiance, aucune préférence n'apparait entre les différents chemins. Aucune corrélation entre le classement de la fonction et celui des testeurs n'est observée et la variabilité est minimale. Ce résultat apporte plusieurs interrogations. Premièrement, les coûts de concaténation ou les distances utilisées pourraient ne pas être adéquats. Deuxièmement, comme d'autres travaux ont monté une faible variabilité sur les premiers chemins, il serait peut-être plus judicieux de regarder plus loin, jusqu'aux 1 000 premiers chemins par exemple. Un bon point est que ces informations pourraient mener à une bonne heuristique pour le A^*.

En ce qui concerne les deux tests, des discussions avec les testeurs ont souligné le fait que leur attention était surtout retenue par les incohérences prosodiques bien plus que par les artéfacts de concaténation, ceux-ci étant considérés par la plupart comme étant rares. Nos travaux suivants vont donc se concentrer sur l'introduction de nouveaux filtres pour durcir la pré-sélection.

7 Conclusion

Dans cet article, nous avons présenté un nouveau système de synthèse de la parole par corpus avec un étage de sélection d'unités particulier puisqu'implémentant un algorithme A^*. Le système, conçu comme une plateforme expérimentale visant à explorer le comportement de la sélection d'unités en profondeur, implémente des mécanismes issus de l'état de l'art pour effectuer une sélection d'unités. Utiliser un A^* rend plus aisée l'obtention des N meilleurs chemins dans un problème de sélection d'unités. Une bonne heuristique pourrait même drastiquement améliorer le temps nécessaire pour résoudre le problème sans sacrifier l'optimalité du résultat. Des évaluations subjectives ont été effectuées et ont montré qu'un système basé A^* est efficace et donne des résultats satisfaisants en termes de qualité de la synthèse. Les remarques des testeurs suggérant que la qualité des concaténations était élevée, nos prochains travaux viseront à améliorer les filtres de pré-sélection en y incluant des coûts prosodiques et à inclure et tester des heuristiques. D'autre part, la prochaine étape dans nos travaux consiste à comparer A^* à Viterbi en terme de qualité, souplesse et efficacité calculatoire.

Quelques échantillons de synthèses sont disponibles en ligne à l'adresse suivante : http ://www-expression.irisa.fr/fr/demos/.

Références

ALÍAS, F., FORMIGA, L. et LLORÁ, X. (2011). Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms : A proof-of-concept. volume 53(5) de *Speech Communication*, pages 786–800.

BLACK, A. W. et TAYLOR, P. (1994). Chatr : a generic speech synthesis system. volume 2 de *Proc. of the* 15^{th} *conference on Computational linguistics*, pages 983–986.

BOEFFARD, O., CHARONNAT, L., MAGUER, S. L., LOLIVE, D. et VIDAL, G. (2012). Towards fully automatic annotation of audiobooks for tts. Proc. of LREC.

BREEN, A. P. et JACKSON, P. (1998). Non-uniform unit selection and the similarity metric within bt's laureate tts system. Proc. of the ESCA Workshop on Speech Synthesis, pages 373–376.

CHEVELU, J., LECORVÉ, G. et LOLIVE, D. (2014). Roots : a toolkit for easy, fast and consistent processing of large sequential annotated data collections. Proc. of LREC (accepted).

CLARK, R. A., RICHMOND, K. et KING, S. (2007). Multisyn : Open-domain unit selection for the festival speech synthesis system. volume 49 de *Speech Communication*, pages 317–330.

CONKIE, A., BEUTNAGEL, M. C., SYRDAL, A. K. et PHILIP, E. (2000). Preselection of candidate units in a unit selection-based text-to-speech synthesis system. volume 3 de *In ICSLP*, pages 314–317.

GOVIND, D. et PRASANNA, S. M. (2013). Expressive speech synthesis : a review. *International Journal of Speech Technology*, pages 1–24.

HART, P. E., NILSSON, N. J. et RAPHAEL, B. (1968). A formal basis for the heuristic determination of minimum cost paths. volume 4 de *IEEE Transactions on System Science and Cybernetics*, pages 100–107.

HUNT, A. J. et BLACK, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. volume 1 de *Proc. of IEEE ICASSP*, pages 373–376.

KUMAR, R. (2004). A genetic algorithm for unit selection based speech synthesis. Proc. of INTERSPEECH.

NILSSON, N. J. (1982). *Principles of artificial intelligence*. Springer-Verlag editions.

SAGISAKA, Y. (1988). Speech synthesis by rule using an optimal selection of non-uniform synthesis units. Proc. of IEEE ICASSP, pages 679–682.

TAYLOR, P., BLACK, A. W. et CALEY, R. (1998). The architecture of the festival speech synthesis system. Proc. of the ESCA Workshop in Speech Synthesis, pages 147–151.

VITERBI, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Tr. on Information Theory, pages 260–269.

YAMAGISHI, J., LING, Z. et KING, S. (2008). Robustness of hmm-based speech synthesis. Proc. of the international conference on speech prosody, pages 581–584.

YI, J. R. W. (1998). Natural-sounding speech synthesis using variable-length units. Mémoire de D.E.A.

Tension gestuelle en co-présence d'un accent d'intensité

Gaëlle Ferré
Université de Nantes – LLING
Chemin de la Censive du Tertre
Nantes cedex 3
Gaelle.Ferre@univ-nantes.fr

RESUME

Selon McNeill (1992, 2005), gestes co-verbaux et parole émergent d'une même source cognitive et forment un ensemble indivisible dans la communication linguistique. Cette proposition, déjà relativement ancienne, a trouvé un écho dans des travaux plus récents de sciences cognitives, dans les courants de cognition dite « située » (Barsalou, 2008) et « incarnée » (Iverson & Thelen, 1999) pour lesquels l'activité cognitive se nourrit de l'environnement (le contexte) et d'informations proprioceptives. La communication linguistique relève à la fois des modalités vocales et gestuelles, tant sur le plan segmental que sur le plan supra-segmental, et il est donc naturel de trouver des caractéristiques communes d'une activité linguistique dans les deux modalités. Dans cet article, nous étudions les liens entre un trait supra-segmental vocal (l'accent d'intensité) et un trait supra-segmental gestuel (la tension de la main) lorsque geste et accent d'intensité sont produits simultanément.

ABSTRACT

Gesture tension and emphatic prosodic stress

According to McNeill (1992, 2005), co-verbal gesture and speech stem from a unique cognitive unit and emerge as an indivisible whole in linguistic communication. This proposition has found an echo in later works in "grounded cognition" (Barsalou, 2008) and "embodied cognition" (Iversen & Thelen, 1999). In the two theories, cognition is based both on the environment (context) and on proprioceptive information. Since linguistic communication involves a vocal and a gestural modality, on the segmental and supra-segmental levels respectively, it is only natural to find common traces of linguistic activity in both modalities. In this paper, we therefore examine the links which exist between a vocal supra-segmental feature (emphatic stress) and a gestural supra-segmental feature (tension of hand shape) when co-speech gesture and emphatic stress are co-present in linguistic communication.

MOTS-CLES : Accent d'intensité, tension gestuelle, communication multimodale, cognition.
KEYWORDS: Emphatic stress, gesture tension, multimodal communication, cognition.

1 Introduction

Bolinger (1983), dans un article qui étudiait les liens entre gestes co-verbaux (manuels et faciaux) et intonation, proposait une correspondance entre la direction de réalisation du geste et la direction du mouvement intonatif : à un mouvement intonatif montant correspondait un mouvement vers le haut pour le geste co-verbal et à un mouvement intonatif descendant correspondait un mouvement vers le bas du geste co-verbal. Ainsi, il associait entre autres les haussements de sourcils à des montées intonatives. Aussi séduisante que puisse paraître une telle correspondance, les études portant sur des corpus de parole spontanée n'ont malheureusement pas pu la confirmer (Loehr, 2004 ; Ferré, 2014). Pour autant, on ne doit pas exclure l'idée du lien entre geste et parole tel qu'il est proposé par McNeill (1992, 2005), selon qui gestes co-verbaux et parole émergent d'une même source cognitive et forment un ensemble indivisible dans la communication linguistique. Ce lien a été montré par exemple par M. Dohen et ses collègues (Dohen & Lœvenbruck, 2004; Dohen & Loevenbruck, 2004, 2005; Dohen, Loevenbruck, Cathiard, & Schwartz, 2004; Dohen, Loevenbruck, & Hill, 2006) qui ont observé des correspondances récurrentes entre l'accent d'intensité et certains traits articulatoires faciaux dans un cadre expérimental. Dans la parole spontanée, Ferré (2014) a également montré qu'il existe des correspondances entre certains types de mise en relief et certains gestes co-verbaux : si la mise en relief syntaxique d'un élément de discours par antéposition est accompagnée de gestes dits métaphoriques (cf. la typologie de McNeill, 2005), l'accent d'intensité est en revanche généralement accompagné de battements de main et de tête. Cependant, au-delà du type de geste co-verbal trouvé en présence de l'accent d'intensité dans les corpus, on peut s'interroger sur la récurrence de traits articulatoires non-faciaux accompagnant ce phénomène prosodique. Ainsi, Ferré (Soumis) montre que l'amplitude des gestes co-verbaux manuels participe non seulement à la perception de l'accent d'intensité, mais est produite de manière distinctive en co-occurrence avec l'accent d'intensité par les locuteurs d'un corpus de parole spontanée.

L'article présenté ici propose une suite à cette démarche. Il s'inscrit dans l'approche psycholinguistique de McNeill (1992, 2005) et s'inspire de la théorie de la cognition « située » (Barsalou, 2008) et « incarnée » [1] (Iverson & Thelen, 1999; Wilson, 2002). Le postulat de base est qu'il existe un lien fort entre la modalité vocale et la modalité gestuelle dans la communication linguistique et que ce lien se traduit par une concordance de traits articulatoires, ici notamment le trait de tension. L'étude repose sur le *Corpus of Interactional Data* (Bertrand, et al., 2008), dont l'annotation multimodale permet la mise en correspondance de phénomènes prosodiques et gestuels.

2 Cadre théorique

En multimodalité, de nombreux chercheurs postulent un lien étroit entre la parole et certains gestes, dits co-verbaux, qui participent à la communication linguistique. C'est notamment la position de McNeill (1992, 2005), pour qui la gestualité co-verbale n'est pas là pour apporter un complément d'information à la parole, support du message, mais

[1] « Grounded cognition » et « embodied cognition » dans les travaux de langue anglaise.

pour qui le message linguistique est le résultat d'une imagerie cérébrale complexe qui intègre des informations gestuelles et langagières, ainsi que des informations environnementales. Cette position est largement soutenue en sciences cognitives par les travaux en cognition « située » (Barsalou, 2008) et « incarnée » (Iverson & Thelen, 1999 ; Wilson, 2002). Ainsi, comme l'écrivent Iverson & Thelen : « we believe that this linkage of the vocal expression of language and the arm movements produced with it are a manifestation of the embodiment of thought: that human mental activities arise through bodily interactions with the world. (…) Both hand and mouth are tightly coupled in the mutual cognitive activity of language." (1999 : 19-20). Cette position est confirmée dans le travail de doctorat de B. Roustan (2012) qui trouve une corrélation temporelle entre les cibles articulatoires faciales impliquées dans la production de l'accent d'intensité et les cibles articulatoires des gestes manuels concomitants. La communication linguistique s'opère donc par un couplage des modalités vocale et gestuelle, comme le montre la Figure 1.

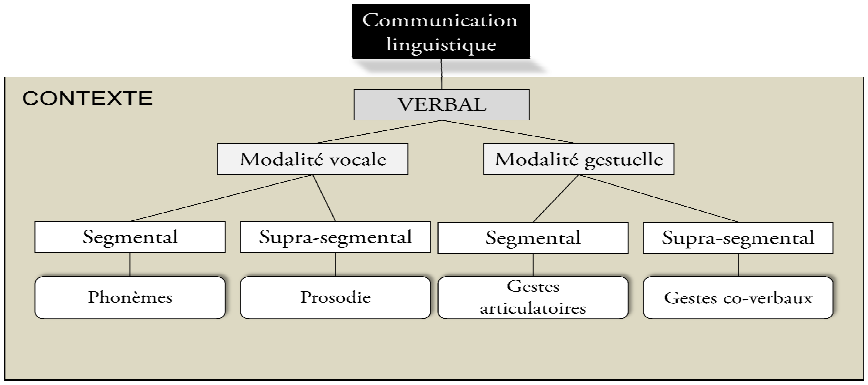


FIGURE 1 – Multimodalité de la communication linguistique

Dans le modèle de communication linguistique multimodale représenté dans la Figure 1, il apparaît que la communication linguistique prend appui sur un système verbal (la « langue » dans l'opposition saussurienne) qui est actualisé dans les modalités vocale et gestuelle (actualisation qui ne saurait donc correspondre à « parole », terme trop restrictif). Il existe une légère asymétrie entre les deux modalités car s'il est nécessaire de produire des gestes articulatoires pour parler, il n'est pas nécessaire de parler pour produire des gestes communicatifs (par exemple, l'interlocuteur peut produire un feedback gestuel uniquement, ou bien un geste peut se substituer à une ellipse de la parole). En revanche, les deux modalités fonctionnent en symétrie en ce qu'elles comportent toutes deux un niveau segmental (gestes articulatoires qui produisent des phonèmes) et un niveau supra-segmental (intonation et gestes co-verbaux faciaux, manuels…) qui agissent au niveau de l'énoncé ou du discours, même si les gestes co-verbaux ne sont pas obligatoires dans la communication linguistique. Enfin, toute communication linguistique s'inscrit dans un contexte au sens large (contexte discursif et situationnel).

Ce qui nous intéresse ici consiste à examiner les liens qui peuvent exister entre des réalisations supra-segmentales dans les modalités vocale et gestuelle. L'une des manières permettant de mettre en relief un élément du discours sur le plan de la prosodie consiste à produire un accent d'insistance, ou accent emphatique, qui implique une plus grande tension articulatoire ou tension des cordes vocales et qui se traduit par une attaque syllabique forte, une hausse de la fréquence fondamentale et/ou de l'intensité. Sur le

plan de la gestualité, les locuteurs peuvent également mettre en relief un élément du discours par certaines caractéristiques : « A strongly presented discursive object can be signaled (…) via added tension, finger curvature, added beats, etc. » (McNeill, 2005 : 50) et « tenser and faster is more salient » (Cienki, 2013 : 58). Nous voudrions aller plus loin en proposant qu'une emphase produite dans la modalité vocale doit trouver une corrélation dans la modalité gestuelle si le locuteur choisit d'associer les deux niveaux supra-segmentaux (comme il apparaît dans les travaux de Krahmer & Swerts, 2007), en nous concentrant ici sur la production manuelle uniquement pour la modalité gestuelle (les autres types de geste ayant fait l'objet de l'étude parue dans Ferré, 2014).

Ainsi, dans Ferré (soumis), nous avons pu montrer le lien fort qui existe entre la production et la perception des accents emphatiques et l'amplitude des gestes manuels qui les accompagnent. Le lien est si fort qu'il crée même un effet McGurk (McGurk & MacDonald, 1976) sur le plan de la perception, puisque les locuteurs qui voient un geste co-verbal large perçoivent un accent emphatique dans la parole, y compris dans des énoncés qui n'en contiennent pas. La seule présence d'un geste ne suffit pas à créer cet effet, car si l'énoncé est accompagné d'un geste moins ample, les auditeurs ne perçoivent pas d'accent emphatique dans les énoncés qui n'en contiennent pas. Plus important pour notre travail sur la parole conversationnelle spontanée est que les locuteurs produisent des gestes plus amples en concomitance avec l'emphase prosodique.

S'il est impossible de tester le lien entre vélocité gestuelle et emphase prosodique dans un corpus de parole spontanée (car cela nécessiterait un appareillage de mesure de la vélocité qui rendrait impossible la spontanéité de l'interaction), il est en revanche possible de tester le trait de tension articulatoire de la main. On sait que le type de parole a un impact sur la réalisation des phonèmes : dans un débit de parole rapide (ce qui est souvent le cas dans la parole conversationnelle), certains phonèmes vont se trouver réduits. Meunier & Floccia (1997) montrent même que certains phonèmes sont plus susceptibles de subir une réduction que d'autres, comme c'est le cas de /a/ en français, par exemple. On peut établir un lien entre la réduction acoustique observée et la réduction articulatoire qui en est à l'origine. De la même manière, on peut supposer que la majorité des gestes co-verbaux produits en parole conversationnelle présenteront une certaine « réduction » par rapport à des gestes produits dans d'autres contextes, qu'ils seront en quelque sorte moins bien « articulés », du moins en ce qui concerne la configuration manuelle. C'est ce qui apparaît dans la Figure 2 ci-dessous.

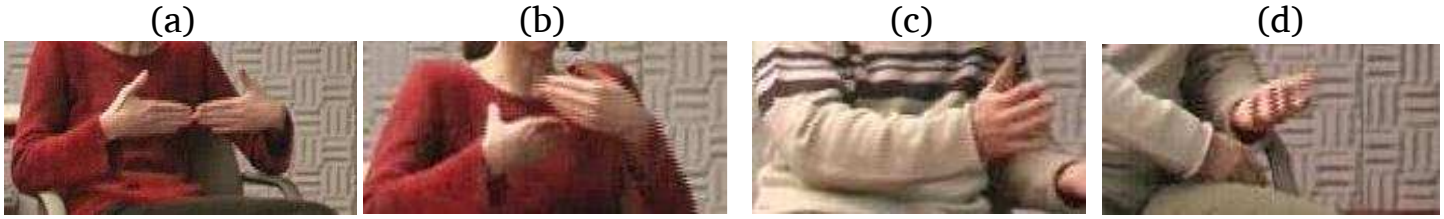


FIGURE 2 – Configurations manuelles tendues [(a) et (c)] et relâchées [(b) et (d)] pour 4 gestes statiques réalisés par 2 locuteurs. Chaque image correspond à l'apogée gestuelle

Ainsi qu'il apparaît dans les figures (2a) et (2b), la locutrice produit deux gestes statiques (à peu de temps d'intervalle et qui portent sur le même objet de discours). En (2a), les deux mains sont tendues dans la configuration prototypique « B-spread » (voir la description des configurations manuelles dans la section 3). En (2b), les mains

s'approchent de la configuration « B-spread » plus que de toute autre configuration, mais les mains ne sont pas aussi tendues qu'en (2a) au moment de l'apogée gestuelle, soit au point d'extension maximale du geste. On peut donc dire que le geste produit en (2a) est tendu, alors que celui qui est produit en (2b) est relâché. De la même manière, le locuteur en (2c) produit un geste statique dans la configuration manuelle « 5 » où les 5 doigts de la main sont tendus et écartés les uns des autres. En (2d), la main est ouverte et les doigts sont écartés comme dans la configuration prototypique, mais les doigts ne sont pas tendus.

L'hypothèse qui est testée dans cet article est que dans les contextes d'emphase prosodique qui servent à mettre en relief un élément de discours (quelle que soit la fonction discursive de cette mise en relief), les gestes co-verbaux manuels produits de manière simultanée à l'emphase vocale seront réalisés de manière plus tendue que les gestes produits hors contexte d'emphase vocale, et que ceci résulte du couplage des modalités dans la communication linguistique.

3 Corpus et annotations

Le corpus utilisé pour cet article est une sous-section du corpus CID (*Corpus of Interactional Data*, Bertrand, et al., 2008) qui comporte 3 enregistrements dialogiques d'une demi-heure entre 6 locuteurs. Entre autres annotations, les syllabes comportant un accent d'intensité ont été repérées manuellement sous Praat (Boersma & Weenink, 2009) et initialement codées selon leurs différentes fonctions en discours (fonctions qui n'ont pas été distinguées ici). Un accent d'intensité a été noté lorsque la syllabe était prononcée avec une attaque forte, éventuellement accompagnée d'une hausse de F0 et/ou d'intensité, et qu'elle montrait un allongement perceptible. On en dénombre 388 occurrences.

Par ailleurs, parmi les nombreuses annotations multimodales réalisées sous Anvil (Kipp, 2001), nous avons ici utilisé les phrases gestuelles (Kendon, 2004) : délimitation des unités gestuelles, que nous avons ensuite catégorisées selon leur fonction dans le discours en utilisant les dimensions proposées par McNeill (2005). Lorsqu'un geste a été codé comme relevant de plusieurs dimensions, c'est la dimension principale qui a été retenue ici. Pour chaque unité gestuelle, nous avons également noté la configuration manuelle au moment de l'apogée du geste, en nous basant sur le système de codage de la Langue des Signes Américaine. Dans la production des gestes bi-manuels asymétriques (dans lesquels les deux mains ne présentent pas la même configuration), les deux configurations ont été notées dans deux pistes distinctes d'Anvil. Cette étude ne porte donc pas réellement sur le nombre d'unités gestuelles, mais plutôt sur le nombre de configurations manuelles produites conjointement avec un accent d'intensité (305 occurrences). Le tableau ci-dessous représente les configurations manuelles prototypiques utilisées dans la description du CID (le chiffre donné entre parenthèses indique le nombre d'occurrences de chaque configuration dans l'ensemble du corpus) :

Enfin, pour chaque configuration manuelle, l'annotation indique si le geste est tendu ou relâché, c'est-à-dire si la configuration manuelle est clairement identifiable car elle correspond à une forme prototypique, ou si au contraire, bien que se rapprochant d'une forme prototypique, elle n'en présente pas toutes les caractéristiques (par exemple si les doigts de la main ne son pas tendus comme ils devraient l'être, ou ne sont pas suffisamment écartés les uns des autres).

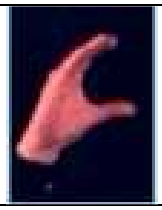
A (94)	B (11)	B-spread (482)	C (86)	D (1)	F (8)
					
G (124)	H (2)	L (86)	Bent-L (45)	O (199)	Baby-O (38)
Tapered-O (26)	S (16)	T (27)	X1 (50)	3 (17)	Bent-3 (11)
4 (2)	5 (480)	Bent-5 (101)	7 (4)		
				Total annotations : 1880	

TABLE 1 – Configurations manuelles et nombre d'occurrences de chaque configuration dans le corpus CID (d'après les images publiées dans Neidle, 2007)

4 Résultats

Nous comptons ici la simultanéité de production temporelle entre gestes manuels et accent d'intensité, même si l'empan des deux types de production n'est pas équivalent puisqu'un geste est plus long qu'une syllabe accentuée. Dès lors qu'une syllabe comportant un accent d'intensité est réalisée en co-présence d'un geste manuel, ceci devrait avoir un impact sur la tension du geste.

Les résultats en ce qui concerne la relation entre la tension des gestes manuels co-verbaux et l'accent d'intensité ne sont pas aussi marqués que ceux exprimant le lien entre amplitude et emphase prosodique. Pour 4 locuteurs sur 6, il existe réellement une correspondance : le test de proportion réalisé montre une différence significative de tension entre les gestes manuels réalisés en co-présence d'un accent d'intensité et ceux qui ne le sont pas ($X^2 = 4.8$, $df = 1$, $p = 0.02$), la proportion indiquant que les gestes manuels, quel que soit leur type, sont plus tendus en présence d'un accent d'intensité. En revanche, pour les deux autres locuteurs, le test de proportion n'est pas significatif ($X^2 = 0.6$, $df = 1$, $p = 0.4$). Cette absence de significativité chez ces deux locuteurs ne peut s'expliquer par le nombre d'accents d'intensité produit, par le type de geste, la configuration manuelle, ni même par une éventuelle compensation par l'amplitude des

gestes. Ces locuteurs ne diffèrent en rien des autres, si ce n'est qu'ils ne marquent pas la tension prosodique par une plus grande tension gestuelle.
En ce qui concerne maintenant le type de geste, on note un résultat très intéressant. Pour un premier groupe de gestes (Groupe 1 : adaptateurs, battements, butterworths, déictiques, emblèmes), on note une plus grande tension lorsque le geste est produit en présence d'un accent d'intensité que lorsqu'il ne l'est pas (Test de proportion : $X^2 = 4.1$, $df = 1$, $p = 0.04$). En revanche, un deuxième groupe de gestes (Groupe 2 : iconiques, métaphoriques) montrent un schéma inverse, c'est-à-dire qu'ils sont significativement plus tendus hors présence d'accent d'intensité ($X^2 = 6.53$, $df = 1$, $p = 0.01$).

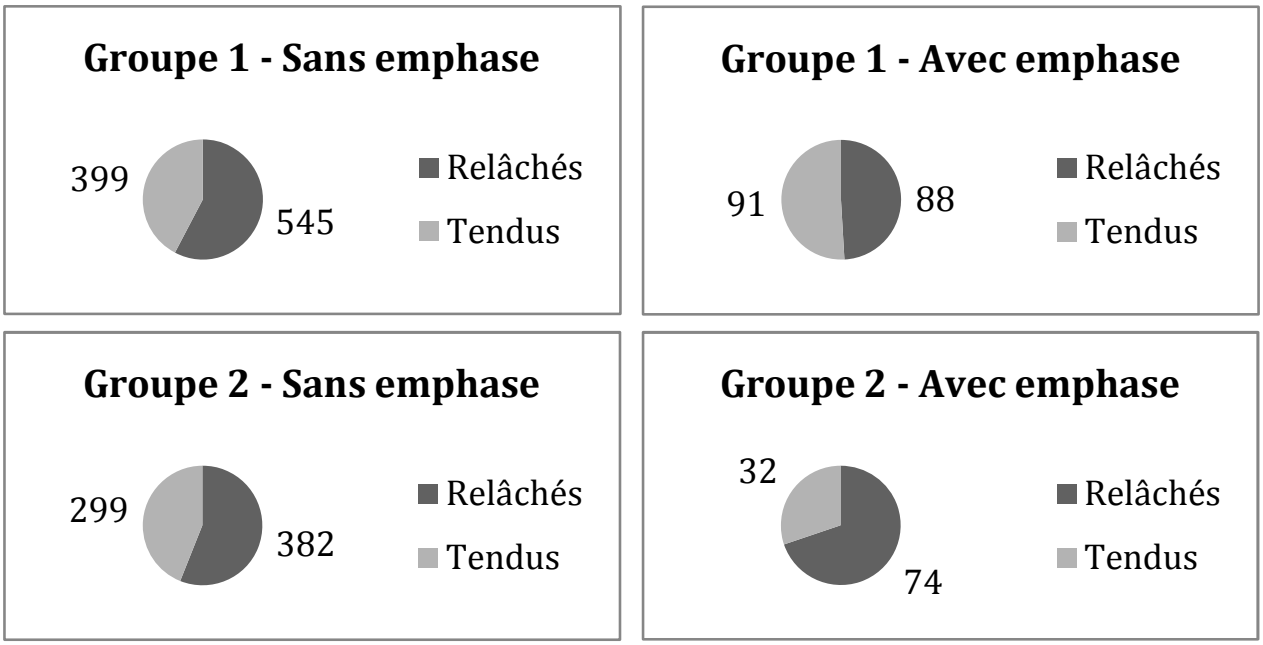


FIGURE 3 – Proportion de gestes tendus et relâchés pour les deux groupes de geste, avec ou sans emphase prosodique

Dans la Figure 3, il apparaît clairement que si les deux groupes de geste ont un comportement similaire en dehors de la présence d'un accent d'intensité, ils montrent un comportement opposé en présence de ce dernier. Ce qui est particulièrement intéressant ici est que dans l'étude qui porte sur l'accompagnement gestuel des accents d'intensité (Ferré, 2014), il apparaît clairement que les deux types de gestes qui n'accompagnent pas l'emphase prosodique dans ce corpus de parole spontané sont précisément les gestes iconiques et les métaphoriques.

5 Discussion et conclusion

Les modèles de cognition « incarnée » et « située » proposés en sciences cognitives montrent qu'il existe un lien au niveau de la cognition entre l'environnement et la communication linguistique d'une part, et entre les différentes activités corporelles impliquées dans la communication linguistique d'autre part. Si ce lien existe au niveau pré-verbal, il devrait être possible d'en trouver des traces dans le message verbal.
C'est précisément ce que nous avons étudié dans cet article, qui porte sur le lien existant entre les niveaux supra-segmentaux des modalités vocales et gestuelles dans un corpus de français spontané. Dans un article précédent (Ferré, Soumis), nous avons montré qu'il

existe un lien très fort tant sur le plan de la production que de la perception entre la production d'un accent d'intensité en prosodie et l'amplitude des gestes manuels qui accompagnent le discours oral. D'autres traits gestuels sont également susceptibles d'être en lien avec l'accent d'intensité, parmi lesquels se trouve la tension articulatoire du geste. Un geste manuel co-verbal peut effectivement être tendu, comme c'est le cas lorsqu'il correspond à une configuration manuelle prototypique, ou relâché lorsqu'il présente certaines caractéristiques d'une configuration prototypique mais pas toutes (par exemple lorsque les doigts de la main ne sont pas totalement tendus).
Dans l'extrait du corpus CID étudié, figurant 6 locuteurs, on peut dire que si la majorité des locuteurs associent une plus grande tension gestuelle à un accent d'intensité, ce trait ne constitue pas une caractéristique aussi liée à la prosodie que l'amplitude gestuelle. En revanche, des correspondances sont établies en ce qui concerne les relations entre gestes co-verbaux et phénomènes prosodiques : dans Ferré (2014), nous avions montré que les gestes iconiques et métaphoriques ne sont pas liés à la mise en relief prosodique. Or, ces deux types de gestes affichent un comportement différent des autres types en cela qu'ils ne présentent pas une tension plus importante en présence d'un accent d'intensité, comme si le locuteur qui, pour une raison discursive autre que la mise en relief prosodique, est amené à produire un de ces gestes en présence d'un accent d'intensité, s'attachait à en réduire l'impact sur le plan de l'intégration prosodique geste-voix.

Références

BARSALOU, L. W. (2008). Grounded Cognition. *The Annual Review of Psychology, 59*, 617-645.

BERTRAND, R., BLACHE, P., ESPESSER, R., FERRÉ, G., MEUNIER, C., PRIEGO-VALVERDE, B., et al. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues, 49(3)*, 105-133.

BOERSMA, P., & WEENINK, D. (2009). Praat: doing phonetics by computer (Version 5.1.05) [Computer program].

BOLINGER, D. (1983). Intonation and Gesture. *American Speech, 58(2)*, 156-174.

CIENKI, A. (2013). Conceptual Metaphor Theory in Light of Research on Speakers' Gestures. *Journal of Cognitive Semiotics, V(1-2)*, 349-366.

DOHEN, M., & LŒVENBRUCK, H. (2004). *La focalisation contrastive est-elle visible ? Une étude perceptive visuelle de la focalisation contrastive en français*. Paper presented at the XXVèmes Journées d'Étude sur la Parole. 189-192.

DOHEN, M., & LOEVENBRUCK, H. (2004). *Pre-focal Rephrasing, Focal Enhancement and Post-focal Deaccentuation in French*. Paper presented at the ICSLP 2004. 1313-1316.

DOHEN, M., & LOEVENBRUCK, H. (2005). *Audiovisual Production and Perception of Contrastive Focus in French: a Multispeaker Study*. Paper presented at the Interspeech/Eurospeech. 2413-2416.

DOHEN, M., LOEVENBRUCK, H., CATHIARD, M.-A., & SCHWARTZ, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication, 44*,

155-172.

DOHEN, M., LOEVENBRUCK, H., & HILL, H. (2006). *Visual Correlates of Prosodic Contrastive Focus in French: Description and Inter-Speaker Variability*. Paper presented at Speech Prosody. 1-4.

FERRÉ, G. (2014). A Multimodal Approach to Markedness in Spoken French. *Speech Communication, 57* (Special Issue on Gesture and Speech in Interaction), 268-282.

FERRÉ, G. (Soumis). How Far do Current Theoretical Models on Gesture-Speech Production Comply with Data from Spontaneous Interactions? *Gesture*.

IVERSON, J. M., & THELEN, E. (1999). Hand, Mouth and Brain. The Dynamic Emergence of Speech and Gesture. *Journal of Consciousness Studies, 6(11-12)*, 19-40.

KENDON, A. (2004). *Gesture. Visible Action as Utterance*. Cambridge, Cambridge University Press.

KIPP, M. (2001). *Anvil - A Generic Annotation Tool for Multimodal Dialogue*. Paper presented at the 7th European Conference on Speech Communication and Technology (Eurospeech). 1367-1370.

KRAHMER, E., SWERTS, M., 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57, 396–414.

LOEHR, D. (2004). *Gesture and Intonation*. Georgetown University, Georgetown, PhD thesis.

MCGURK, H., & MACDONALD, J. (1976). Hearing Lips and Seeing Voices. *Nature, 264*, 746-748.

MCNEILL, D. (1992). *Hand and Mind : What Gestures Reveal about Thought*. Chicago and London, The University of Chicago Press.

MCNEILL, D. (2005). *Gesture & Thought*. Chicago and London, The University of Chicago Press.

MEUNIER C., FLOCCIA, C. (1997). La spécificité des types de parole pour la perception de la voyelle, Actes des Journées d'Etudes Linguistiques « La voyelle dans tous ses états », Nantes, 158-163.

NEIDLE, C. (2007). SignStream™ Annotation: Addendum to Conventions used for the American Sign Language Linguistic Research Project. http://www.bu.edu/asllrp/cslgr/pages/handshape-palette.html

ROUSTAN, B. (2012). Étude de la coordination gestes manuels / parole dans le cadre de la désignation. Thèse de doctorat, gipsa-lab et École Doctorale Ingéniérie pour la Santé et la Cognition et l'Environnement, Grenoble.

WILSON, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9(4)*, 625-636.

Construction faiblement supervisée d'un phonétiseur pour la langue iban à partir de ressources en malais

Sarah Samson Juan[1], Laurent Besacier[1], Solange Rossato[1]
(1) Grenoble Informatics Laboratory (LIG)
University Grenoble-Alpes
Grenoble, France
(sarah.samson-juan@imag.fr, laurent.besacier@imag.fr, solange.rossato@imag.fr

RÉSUMÉ

Cet article décrit notre collecte de ressources pour la langue iban (parlée notamment sur l'île de Bornéo), dans l'objectif de construire un système de reconnaissance automatique de la parole pour cette langue. Nous nous sommes plus particulièrement focalisés sur une méthodologie d'amorçage du lexique phonétisé à partir d'une langue proche (le malais). Les performances des premiers systèmes de reconnaissance automatique de la parole construits pour l'iban (< 20% WER) montrent que l'utilisation d'un phonétiseur déjà disponible dans une langue proche (le malais) est une option tout à fait viable pour amorcer le développement d'un système de RAP dans une nouvelle langue très peu dotée. Une première analyse des erreurs fait ressortir des problèmes bien connus pour les langues peu dotées : problèmes de normalisation de l'orthographe, erreurs liées à la morphologie (séparation ou non des affixes de la racine).

ABSTRACT

Weakly supervised G2P training for Iban language using Malay Resources

This paper describes our experiments and results on using a local dominant language in Malaysia (Malay), to bootstrap automatic speech recognition (ASR) for a very under-resourced language : iban (also spoken in Malaysia on the Borneo Island part). Resources in iban for building a speech recognition were nonexistent. For this, we tried to take advantage of a language from the same family with several similarities. First, to deal with the pronunciation dictionary, we proposed a bootstrapping strategy to develop an iban pronunciation lexicon from a Malay one. A hybrid version, mix of Malay and iban pronunciations, was also built and evaluated. Following this, we experimented with three iban ASRs ; each depended on either one of the three different pronunciation dictionaries : Malay, iban or hybrid.

MOTS-CLÉS : langues peu-dotées, reconnaissance automatique de la parole, conversion graphème-phonème, iban, malais, amorçage, Kaldi.

KEYWORDS: under-resourced languages, speech recognition, G2P conversion, iban language, Malay language, bootstrapping, Kaldi.

1 Introduction

Le dictionnaire de prononciation est une ressource cruciale pour certaines technologies vocales telles que la reconnaissance automatique de la parole (RAP) ou la synthèse de la parole (TTS). Construire un lexique phonétisé pour une nouvelle langue et pour plusieurs dizaines de milliers de mots peut prendre un temps considérable. On peut bien sûr envisager de créer un système

automatique de conversion graphème-phonème (que nous nommerons, dans la suite de cette article G2P pour *Grapheme-to-Phoneme Conversion*). Cependant, la construction d'un système G2P nécessite des règles de phonétisation formalisées (méthode à base de règles) ou un corpus d'apprentissage contenant des mots déjà phonétisés (méthode empirique). Cet article propose d'utiliser des ressources déjà disponibles dans une langue proche assez bien dotée (le malais), pour amorcer la construction d'un système de conversion graphème-phonème pour une langue très peu dotée (l'iban).

Un amorçage de système G2P a déjà été proposé pour aider à la création de lexiques de prononciation pour les langues telles que l'afrikaans et le népalais ((Davel et Martirosian, 2009), (Maskey *et al.*, 2004)). Dans ces approches, un petit lexique disponible dans la langue visée était utilisé pour construire un premier système de conversion graphème-phonème. Ce système était ensuite utilisé pour générer les prononciations d'un plus gros vocabulaire, puis ses sorties étaient vérifiées et (éventuellement) corrigées par un locuteur natif de la langue cible. Le nouveau dictionnaire de prononciation, de plus grande taille, était ensuite utilisé pour construire le système G2P final. Cependant, pour de nombreuses langues pauvrement dotées (Besacier *et al.*, 2014), même un petit lexique initial est difficile (voir impossible) à obtenir. Nous proposons donc, dans ce travail, d'utiliser un lexique issu d'une langue proche et mieux dotée, pour amorcer la construction d'un système G2P pour une nouvelle langue.

En particulier, nous proposons d'utiliser les prononciations initiales produites, sur un lexique en langue iban, par un système G2P malais puis de faire postéditer (corriger) ces prononciations par des locuteurs natifs iban. Nous pensons cette approche raisonnable car malais et iban sont étroitement liés et tous deux appartiennent à un même groupe de langues d'après (Lewis *et al.*, 2013) (Adelaar, 2005) . Dans un premier temps, nous analysons la "distance" entre les prononciations générées par notre système G2P en malais et les corrections de ces prononciations par un locuteur natif iban. Dans (Heeringa et de Wet, 2008), une étude similaire est menée pour mesurer les distances entre l'afrikaans et le néerlandais, l'afrikaans et le frison, ainsi que l'afrikaans et l'allemand. Les auteurs observent que le néerlandais présente des prononciations plus proches de l'afrikaans que les deux autres langues. A partir des post-éditions de prononciations obtenues sur un lexique initial en iban, nous construisons un système G2P empirique pour l'iban afin de phonétiser d'autres entrées. Les deux systèmes G2P malais et G2P iban, ainsi qu'une méthode hybride (fondée sur l'observation que de nombreux mots du malais sont utilisés en iban), sont tout d'abord évalués. Les lexiques phonétisés obtenus sont ensuite intégrés dans le premier système de RAP jamais développé pour l'iban.

Cet article est organisé comme suit : la section 2 décrit brièvement la phonologie de la langue iban. La section 3 présente les ressources collectées pour cette langue (pour la RAP) ainsi que notre stratégie de construction d'un système G2P. La section 4 présente les différents modèles acoustiques construits pour l'iban à partir des 3 différents dictionnaires de prononciation obtenus et les performances de reconnaissance automatique de la parole associées. La partie 5 analyse plus en détail les sorties et les erreurs des différents systèmes de RAP évalués et, enfin, la section 6 conclut et donne quelques perspectives à ce travail.

Une partie de cet article, consacrée à l'obtention d'un système G2P pour l'iban a déjà été publiée en anglais, en Octobre 2013, pour le *4th Workshop on South and Southeast Asian Natural Language Processing* qui s'est tenu en même temps que la conférence IJCNLP 2013. Cependant, contrairement à la présente soumission, cet article ne contenait encore aucune expérience de RAP.

2 La langue iban

L'iban [1] est une langue austronésienne (branche Malayo-Sumbawan) parlée par plus de 600 000 locuteurs sur l'île de Bornéo, principalement en malaisie (état de Sarawak), ainsi qu'au Brunei et en Indonésie. La langue appartient au même groupe que le malais (groupe *Malayic* - voir (Lewis *et al.*, 2013) et (Dryer et Haspelmath, 2013) . En plus de l'apprentissage du malais, langue officielle, l'iban est enseigné dans les écoles primaire et secondaire (mais non obligatoire) depuis le début des années 90. Dans plusieurs universités, notamment à l'université de Sarawak [2], des cours de langue iban débutants sont proposés aux étudiants de premier cycle. L'iban est influencé par le malais en termes de phonologie, de morphologie et de syntaxe (Education-Department, 2007). Selon (Education-Department, 2007), beaucoup de mots sont communs à ces deux langues (mêmes formes de surface) et on compte de nombreux emprunts au malais. Quelques exemples sont répertoriés dans le tableau 1 avec les prononciations associées.

TABLE 1 – *Exemples de mots communs malais/iban avec les prononciations associées*

No.	Mots	iban	malais
1	ke	/kə/	/kə/
2	nya	/ɲaʔ/	/ɲə/ or / ɲa/
3	iya	/ija/	/ija/
4	bilik	/biliəʔ/	/bileʔ/
5	dua	/duwa/	/duwə/ or /duwa/
6	sida	/sidaʔ/	/sida/
7	puluh	/puluəh/	/puloh/
8	raban	/raban/	/raban/
9	lalu	/lalu/	/lalu/
10	orang	/uraŋ/	/oraŋ/

En 1981, (Omar, 1981) publie une description complète de cette langue. Selon l'auteur, la langue comprend 19 consonnes (/p/, /b/ /t/, /d/, /k/, /g/, /ʔ/, /ʧ/, /dʒ/ , /s/, /h/, /m/, /n/, /ɲ/, /ŋ/, /l/, /r/, /j/, /w/), 6 voyelles (/i/, /e/, /a/, /o/, /u/, / ə/) et 11 séquences de voyelles [3] (/ui/, /ia/, /ea/, /ua/, /oa/, /iu/, /iə/, /uə/, /oə/, /ai/, /au/). Cette liste n'inclut pas les consonnes empruntées au malais telles que /f/, /v/, /θ/, /ð/, /z/, /ʃ/, /x/, /ɣ/. Comme le montre la Table 1, iban et malais utilisent le système d'écriture latin ; par ailleurs, ce ne sont pas des langues tonales. En dehors des séquences de voyelles, bien moins nombreuses en malais, une autre différence entre les deux langues est la présence plus fréquente du son /ʔ/ à la fin de mots iban se terminant par une voyelle.

3 Ressources pour l'iban

3.1 Corpus textuel

Des données textuelles de type buletin d'information ont été collectées sur un site web d'information [4] qui publie plusieurs articles en iban chaque jour. Nous avons récupéré ces articles pour la période 2009-2012 et rassemblé un total de 7000 articles dans divers domaines (général,

1. iban language code : [iba] (ISO 639-3)
2. http ://www.unimas.my
3. nommées *vowel clusters* par (Omar, 1981)
4. www.theborneopost.com/news/utusan-borneo/berita-Iban/

sport et divertissement). Après extraction, le texte a été nettoyé et normalisé en plusieurs étapes, comme suit : (1) supprimer les balises HTML, (2) convertir les dates et les nombres en mots (e.g : 1982 à *sembilan belas lapan puluh dua*), (3) convertir les abbréviations (ex. Dr. à *Doktor*, Prof. à *Profesor*, Kpt. à *Kapten*), (4) segmenter en phrases, (5) mettre le texte en minuscules (6) enlever les marques de ponctuation hormis dans les titres. En définitive, un corpus de 2,08 millions de mots est disponible. C'est peu pour la modélisation statistique du langage, mais c'est à ce jour le seul corpus textuel de plus d'un million de mots disponible pour cette langue.

Grâce à ces données de texte, nous avons construit un modèle de langage à base de trigrammes avec la boîte à outils SRILM (Stolcke, 2002). La perplexité du modèle, mesurée sur les transcriptions du corpus oral collecté (voir paragraphe suivant) est de 162 et le taux de mots hors vocabulaire (%MHV) est de 2,3 %.

3.2 Corpus oral transcrit

Nous disposons de 8 heures transcrites de parole (échantillonnée à 16 kHz). Les données ont été fournies par la RTM (*Radio Televisyen Malaysia*), une station de radio et de télévision locale, à travers l'une de ses chaînes de radio appelée WaiFM. La chaîne diffuse cinq à dix minutes de nouvelles quotidiennes en langue iban. Les tables 2 et 3 donnent plus de détails sur le corpus oral et sa répartition entre apprentissage et test.

TABLE 2 – iban : corpus oral

Genre	Locuteurs	Phrases	Tokens	Durée (minutes)
Femme	14	1,382	36,194	222
Homme	9	1,750	44,408	257

TABLE 3 – Répartition apprentissage / test pour nos expériences

Set	Locuteurs	Genre (H :F)	Phrases	(minutes)
Test	6	2 :4	473	71
Apprentissage	17	7 :10	2659	408

Les données de parole ont été transcrites par 8 locuteurs natifs de l'iban. Au cours d'un atelier préliminaire tenu à l'Université de Sarawak, les transcripteurs ont reçu une session de formation surune formation à Transcriber ((Barras *et al.*, 2000), (Barras *et al.*, 1998)). Au total, on dispose de 3132 phrases transcrites, prononcées par 23 locuteurs différents. Un sous ensemble de 473 phrases, représentant environ une heure de corpus, est choisi pour l'évaluation.

3.3 Dictionnaire de prononciation

3.3.1 Construction du système G2P pour l'iban

Notre corpus textuel iban contient 37k mots différents. Parmi ces mots iban, on trouve aussi du malais (23%) et de l'anglais (19%). Ces statistiques sont obtenues en comparant notre liste de mots iban avec un dictionnaire malais disponible et issu de (Tan *et al.*, 2009) et avec le dictionnaire de CMU en anglais.

Le lexique malais initial de (Tan *et al.*, 2009) nous permet d'entrainer un système G2P pour le malais en utilisant Phonetisaurus, un outil open source fondé sur des transducteurs à états finis

((Novak, 2012), (Novak *et al.*, 2011)) . Un sous-ensemble de 68k est prélevé du dictionnaire initial de 76k tandis que 8k mots sont conservés pour l'évaluation du système G2P construit. Le taux d'erreur sur les phonèmes (PER) est de 6,20 % tandis que le taux d'erreur sur les mots (WER) est de 24,98 % (voir (Juan et Besacier, 2013) pour plus de détails).

Par la suite, avec une méthodologie similaire à celle de (Heeringa et de Wet, 2008), nous sélectionnons les 100 mots les plus fréquents de notre corpus textuel et appartenant à la fois à notre lexique iban (37k) et au lexique malais de (Tan *et al.*, 2009) (76k prononciations disponibles). Nous indiquerons à l'avenir *iban-malais* pour faire référence à ces mots communs aux deux lexiques iban et malais. Le système G2P malais est utilisé pour phonétiser nos 100 mots sélectionnés et les sorties du système G2P malais sont ensuite corrigées par un locuteur natif iban [5]. La comparaison entre les sorties corrigées indique un taux d'erreur sur les mots de 47 % (WER), ce qui signifie que 53 prononciations sur les 100 ont été considérées comme parfaites par l'annotatrice. Ce système G2P malais peut donc être considéré comme un point de départ acceptable pour produire des prononciations pour la langue iban.

En conséquence, le système G2P malais est alors utilisé pour phonétiser un sous-ensemble de 1k du lexique iban initial. Parmi les 1k mots, 500 sont des mots iban inexistants en malais, tandis que 500 autres sont de la catégorie *iban-malais* définie précédemment. Les sorties du système G2P malais sont à nouveau corrigées et les 1k mots correctement phonétisés en iban sont ensuite utilisés pour construire un premier système G2P en iban, toujours avec Phonetisaurus. Notre évaluation, sur un autre sous-ensemble de 1k mots, est présentée dans la table 4 (issue de (Juan et Besacier, 2013)). Les résultats montrent que le nouveau système G2P obtenu pour l'iban est plus performant pour les mots spécifiques à cette langue (notés avec l'indice *I* dans la table) tandis que le système G2P malais phonétise mieux les mots communs aux deux langues (notés avec l'indice *IM* dans la table).

TABLE 4 – Phonétisation de l'iban : performances des systèmes G2P malais et G2P iban

Phonétiseur	Corpus	PER (%)	WER (%)	Post-édition (minutes)
G2P malais	500_{IM}	6.52	27.2	30
	500_{I}	15.8	56.0	42
G2P iban	500_{IM}	13.6	44.2	45
	500_{I}	8.2	31.8	32

Note : IM pour mots malais-iban et I pour mots "pur iban". (Juan et Besacier, 2013)

3.3.2 Obtention d'un lexique iban complet

Le lexique complet iban (37k) est phonétisé avec plusieurs systèmes G2P. En plus d'un premier lexique iban phonétisé avec le système G2P malais (S1) et d'un second lexique iban phonétisé avec le système G2P iban (S2), nous considérons aussi un système G2P hybride (S3) qui repose sur la stratégie suivante : le système G2P malais phonétise les mots de type *iban-malais* tandis que le système G2P iban phonétise les autres mots ("pur" iban). Par ailleurs, les mots d'origine anglaise sont phonétisés par le système G2P malais (comme montré dans (Tan *et al.*, 2009) car le lexique ayant servi à l'apprentissage de ce système comprend beaucoup de mots anglais). Au final, nous disposons d'un vocabulaire iban de 37k mots phonétisés (1K manuellement et les 36k autres automatiquement). Une dernière évaluation réalisée sur 2000 nouveaux mots indique un taux d'erreur PER de 8.1% et un WER de 29.4% obtenu avec le système hybride S3. Dans la

5. la première auteure de cet article

section suivante, ces trois systèmes G2P (S1, S2 et S3) sont évalués dans le cadre d'un système de reconnaissance automatique de la parole.

4 Systèmes de reconnaissance automatique de la parole (RAP)

4.1 Systèmes Kaldi réalisés

Les systèmes sont développés pour l'iban à partir de l'outil open-source Kaldi (Povey *et al.*, 2011). Les modèles acoustiques sont entraînés sur la partie apprentissage du corpus décrit au paragraphe 3.2 avec les 3 dictionnaires de prononciations issus des systèmes G2P S1 (malais), S2 (iban) et S3 (Hybride). Nous reportons les performances obtenues avec les différentes étapes de l'apprentissage. Les paramètres extraits sont 13 MFCCs pour les premiers modèles de type monophone et triphone (4,200 états et 40,000 Gaussiennes). L'enrichissement des paramètres par les coefficients delta et delta-delta, ainsi que plusieurs transformations telles que LDA et MLLT (Gopinath, 1998) sont aussi appliquées et évaluées. Enfin, un apprentissage adapté au locuteur (*Speaker Adaptive Training* - SAT) est réalisé et combiné avec une adaptation dans l'espace des paramètres de type fMLLR (Gales, 1998). Le décodage est réalisé avec plusieurs facteurs de poids pour le modèle de langage (16 au total) et nous reportons ici les meilleurs résultats obtenus pour chaque configuration.

4.2 Performances

Les résultats sont résumés dans la Table 5. Les performances augmentent progressivement au fur et à mesure des étapes. Il est suprenant de constater que la différence entre les trois systèmes, qui utilisent des stratégies de phonétisation relativement différentes, est en fait très faible. Les meilleures performances sont bien obtenues avec la méthode de phonétisation la plus pointue (S3 - Hybride - 20.6% WER) mais la différence est très faible avec le système fondé sur le phonétiseur malais (S2 - malais - 20.82% WER). Ceci montre que l'utilisation d'un phonétiseur déjà disponible dans une langue proche est une option tout à fait viable pour amorcer le développement d'un système de RAP dans une nouvelle langue très peu dotée.

TABLE 5 – RAP de la langue iban - Performances (%WER) pour différents systèmes G2P

Etapes	Lexique		
	malais G2P	iban G2P	Hybride G2P
Monophone	42.17	41.79	41.97
Triphone	36	36.44	36.11
Triphone + Δ + Δ	36.47	36.98	36.77
+ MLLT + LDA	27.24	27.71	26.80
+ SAT	25.86	27.02	25.82
+ fMLLR	20.82	21.90	20.60

4.3 Combinaison de systèmes

Afin de voir si les systèmes construits à partir de plusieurs lexiques sont complémentaires, nous combinons les treillis issus des 3 meilleurs systèmes (dernière ligne de la table 5) et utilisons la technique de décodage par minimisation du risque de Bayes (MBR (Goel *et al.*, 2003)). Les résultats sont présentés dans la table 6 pour la combinaison entre 2 ou 3 systèmes. On observe une amélioration des résultats de 1,38 % en absolu dans le meilleur des cas (combinaison de

trois systèmes). En définitive, compte tenu de la taille limitée de notre corpus d'apprentissage, nous considérons que ce premier système de RAP de la langue iban est assez performant. Ses erreurs sont analysées plus en détail dans le paragraphe suivant.

TABLE 6 – Combinaisons de systèmes de RAP en utilisant différentes stratégies de phonétisation

Combinaison	%WER
Hybride G2P + iban G2P	19.83
malais G2P + iban G2P	19.76
Hybride G2P + malais G2P	19.55
Hybride G2P + malais G2P + iban G2P	19.22

5 Analyse des résultats obtenus

5.1 Faible corrélation entre performances de RAP et de G2P

La figure 1 présente les performances de reconnaissance automatique de la parole (WER en ordonnées) en fonction des performances du système de phonétisation G2P (à nouveau, WER en abcisse). Il est intéressant de constater la relative faible corrélation entre les performances des systèmes. En d'autres termes, les systèmes de phonétisation (G2P) ne semblent avoir que très peu d'influence sur les performances du système de transcription automatique. La bonne nouvelle est que l'utilisation d'une langue similaire dotée de ressources (comme le malais) semble être un bon point de départ pour générer les prononciations et construire un système de reconnaissance pour une langue très peu dotée (comme l'iban).

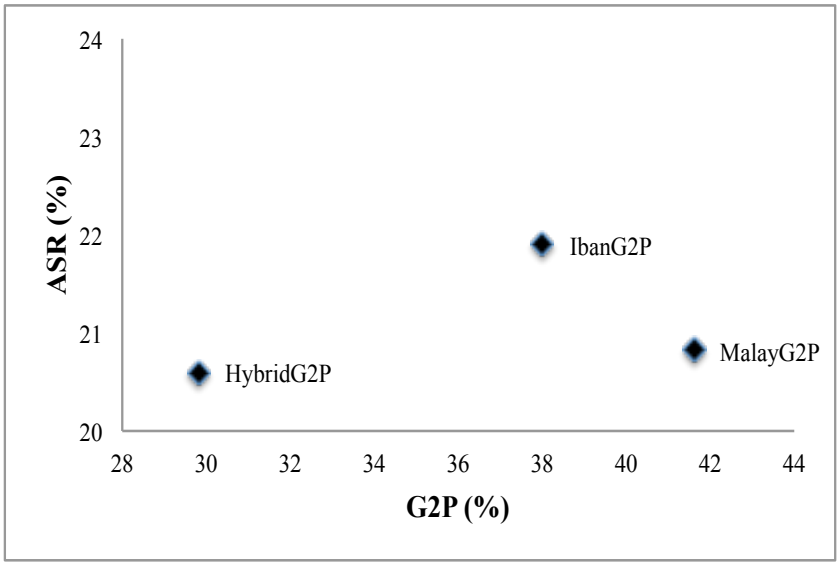


FIGURE 1 – Performance de RAP (ASR) en fonction de la performance de phonétisation (G2P) pour l'iban

5.2 Analyse des erreurs

Le Tableau 7 présente les principales confusions pour les systèmes fondés sur un G2P Hybride, malais, iban ; ainsi que sur un système combiné. Il apparaît que les principales erreurs sont dues à des problèmes de normalisation othographique et de découpage morphologique. Par exemple, le mot *rakyat* (« peuple ») est un mot malais reconnu par le système comme *rayat* qui signifie la même chose en iban. C'est donc une erreur dans la référence qui pénalise l'évaluation du système

TABLE 7 – Principales confusions pour les systèmes fondés sur un G2P Hybride, malais, iban ; ainsi que sur un système combiné

Hybride	malais	iban	Combiné (H+M+I)
rakyat => rayat	rakyat => rayat	rakyat => rayat	rakyat => rayat
ka => ke	ka => ke	ari => hari	ka => ke
ti => ke	ari => hari	ka => ke	ari => hari
ari => hari	serta => sereta	serta => sereta	ti => ke
urang => orang	ti => ke	ti => ke	serta => sereta
serta => sereta	urang => orang	urang => orang	urang => orang
mohamad => mohd	datuk => dato	ke => ka	ke => ka
ka => madahka	ka =>madahka	mohamad => mohd	mohamad => mohd
ke => bejalaika	ke => bejalaika	agensi => ijinsi	agensi => ijinsi
antara => entara	mohamad => mohd	ka => madahka	ka => madahka

de RAP. Une raison possible est que les transcripteurs ont été influencés par l'orthographe du malais. Les autres exemples sont *urang* et *orang* (« personne »), *serta* et *sereta*, *mohamad* et *mohd*, *penerbai* et *penerebai* (« compagnie aérienne »), *agensi* et *ijinsi* (« agence »). Des confusions liées à un problème de découpage morphologique sont par exemple *ka* et *madahka*, *ka* et *bejalaika*, *ka* et *ngambika*, *waifm* et *fm*. *ka* est un suffixe utilisé dans les verbes transitifs et a la même fonction que *kan* en malais. Ce suffixe est souvent séparé de sa racine dans nos textes iban et dans nos transcriptions de parole. Enfin, les différentes colonnes (correspondant à différentes comparaisons entre systèmes) ne font pas apparaître des erreurs fondamentalement différentes d'un système à l'autre. Encore une fois, ceci semble indiquer que le dictionnaire de prononciation a finalement peu d'influence sur les principales erreurs produites par le système de RAP.

6 Conclusions et perspectives

Cet article a décrit notre collecte de ressources pour la langue iban, dans l'objectif de construire un système de reconnaissance automatique de la parole pour cette langue. Nous nous sommes plus particulièrement focalisés sur une méthodologie d'amorçage du lexique phonétisé à partir d'une langue proche (le malais). Les performances des premiers systèmes de reconnaissance automatique de la parole construits pour l'iban sont encourageantes et montrent aussi que la qualité du dictionnaire de prononciation a une influence limitée sur la qualité de transcription. Ce résultat peut apparaître comme positif car il suggère une méthode générique pour amorcer un système G2P en utilisant une langue proche. Une première analyse des erreurs fait ressortir des problèmes bien connus pour les langues peu dotées : problèmes de normalisation de l'orthographe, erreurs liées à la morphologie (séparation ou non des affixes de la racine), etc. C'est notamment sur ces points que se porteront nos efforts dans le futur. Nous venons, par exemple, de lancer une campagne d'annotation collaborative de données iban où il est demandé aux annotateurs d'indiquer d'éventuels problèmes de normalisation et de choisir, à partir d'un couple de mots similaires, la bonne orthographe. Cette campagne vient d'être lancée sur le groupe Facebook *Iban People*.

Références

ADELAAR, A. (2005). The austronesian languages of asia and madagascar. *In The Austronesian Languages of Asia and Madagascar : A Historical Perspective*, pages 1–42, London, UK. Routledge Language Family Series.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. *In Proc. First International Conference on Language Resources and Evaluation (LREC)*, pages 1371–1376.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (2000). Transcriber :development and use of a tool for assisting speech corpora production. *In Proc. Speech Communication special issue on Speech Annotation and Corpus Tools*, volume 33. disponible sur : trans.sourceforge.net/en/publi.php.

BESACIER, L., BARNARD, E., KARPOV, A. et SCHULTZ, T. (2014). Automatic speech recognition for under-resourced languages : A survey. *Speech Communication*, 56:85–100.

DAVEL, M. et MARTIROSIAN, O. (2009). Pronunciation dictionary development in resource-scarce envionments. *In In Proc. INTERSPEECH*, pages 2851–2854.

DRYER, M. et HASPELMATH, M. (2013). The world atlas of language structures online. leipzig : Max planck institute for evolutionary anthropology. available at : http ://wals.info.

EDUCATION-DEPARTMENT, S. (2007). *Sistem Jaku Iban di Sekula*. Sarawak, Malaysia, 1ère édition.

GALES, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12:75–98.

GOEL, V., KUMAR, S. et BYRNE, W. (2003). Segmental minimum bayes-risk decoding for automatic speech recognition. *In IEEE Transactions on Speech and Audio Processing*.

GOPINATH, R. A. (1998). Maximum likelihood modeling with gaussian distributions for classification. *In Proceedings of ICASSP*, pages 661–664.

HEERINGA, W. et de WET, F. (2008). The origin of afrikaans pronunciation : a comparison to west germanic languages and dutch dialects. *In Proc. Conference of the Pattern Recognition Association of South Africa*, pages 159–164.

JUAN, S. S. et BESACIER, L. (2013). Fast bootstrapping of grapheme to phoneme system for under-resourced languages - application to the iban language. *In Proc. 4th Workshop on South and Southeast Asian Natural Language Processing 2013*, Nagoya, Japan.

LEWIS, M. P., SIMONS, G. F. et FENNIG, C. D. (2013). Ethnologue : Languages of the world, sil international. available at : http ://www.ethnologue.com.

MASKEY, S. R., BLACK, A. W. et TOMOKIYO, L. M. (2004). Bootstrapping phonetic lexicons for language. *In Proc. INTERSPEECH*, pages 69–72.

NOVAK, J. R. (2012). Phonetisaurus : A wfst-driven phoneticizer. disponible sur : https ://code.google.com/p/phonetisaurus.

NOVAK, J. R., MINEMATSU, N. et HIROSE, K. (2011). Evaluations of an open source wfst-based phoneticezer. PDF, General Talk No. 452, The Institute of Electronics, Information and Communication Engineers.

OMAR, A. H. (1981). *Phonology*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Schwarz, O. G. N. G. M. H. P., Silovsky, J., Stemmer, G. et Vesely, K. (2011). The kaldi speech recognition toolkit. *In* Society, I. S. P., éditeur : *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, volume IEEE Catalog No. : CFP11SRW-USB.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. *In Proc. of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

Tan, T.-P., Li, H., Tang, E. K., Xiao, X. et Chng, E. S. (2009). Mass : a malay language lvsr corpus resource. *In Proc. Oriental COCOSDA International Conference 2009*, pages 26–30.

Une étude prosodique comparative des questions en français en contact avec l'occitan

Philippe Boula de Mareüil,[1] Albert Rilliard,[1] Fanny Ivent [1,2]

(1) LIMSI-CNRS, BP 133, 91403 Orsay CEDEX
(2) LPP, 19 rue des Bernardins, 75005 Paris
philippe.boula.de.mareuil@limsi.fr, albert.rilliard@limsi.fr
fanny.ivent@limsi.fr

RÉSUMÉ

Cet article rapporte une première analyse de données collectées en Provence et en Languedoc, lors d'une enquête de terrain menée auprès de locuteurs parlant, en plus du français, l'occitan (provençal ou languedocien). Nous avons notamment comparé la prosodie de questions totales terminées par un mot accentué sur l'avant-dernière syllabe (ex. *caserna* en occitan, « caserne » avec un schwa final prononcé en français méridional). Sur les deux dernières syllabes des questions, il apparaît que le patron mélodique montant-montant est le plus fréquent et, d'après une expérience perceptive utilisant la modification/resynthèse de prosodie, qu'il est préféré aux patrons montant-descendant et descendant-montant par des auditeurs méridionaux (sans différences significatives entre Provençaux et Languedociens). Les tests perceptifs n'ont pas permis d'attribuer des fonctionnalités spécifiques (demande d'information, demande de confirmation ou question incrédule) à ces différents patrons intonatifs.

ABSTRACT

A comparative prosodic study of questions in French in contact with Occitan

This study reports on a first analysis of data collected in Provence and Languedoc, during a field survey conducted among speakers of Occitan (Provençal and Languedocian), in addition to French. In particular, we compared the prosody of yes/no questions ending in a word stressed on the penultimate syllable (e.g *caserna* 'barracks' in Occitan, *caserne* with a pronounced final schwa in southern French). On the last two syllables of questions, it turns out that the rising-rising pitch pattern is the most common and, according to a perception experiment using prosody modification/resynthesis, that it is preferred to rising-falling and falling-rising patterns by southern French listeners (without significant differences between Provence and Languedoc). Perceptual tests failed to assign specific functions (information request, confirmation query or incredulous question) to the different intonation patterns.

MOTS-CLÉS : prosodie en contact, accents méridionaux en français, langues en danger.
KEYWORDS: prosody in contact, southern accents in French, endangered languages.

1 Introduction

L'interrogation constitue un aspect privilégié de l'intonation car, dans nombre de langues, elle confère à la prosodie la fonction de distinguer les questions des assertions correspondantes. « En français, par exemple, il est fréquent que le caractère interrogatif de l'énoncé ne soit marqué que par une montée mélodique de la voix sur le dernier mot »

(Martinet, 1970). Dans les phrases du français parlé « il vient » et « il vient ? », la montée mélodique de la voix a dans le deuxième cas la même valeur qu'une marque syntaxique (l'inversion du verbe et du pronom sujet) ou lexicale (l'introducteur *est-ce que*). Dans le français des conversations spontanées, où les questions indiquées par la seule intonation sont les plus fréquentes (Grundstrom et Léon, 1973), il en va comme dans d'autres langues romanes telles l'italien et l'occitan. Il existe ainsi une longue tradition de recherche sur les caractéristiques spécifiques de l'intonation interrogative (Delattre, 1966 ; Fónagy, 2003 ; Savino, 2012), même si l'occitan est évidemment moins étudié.

En français, dans les questions totales (appelant des réponses de type *oui* ou *non*), la montée mélodique est généralement associée à la dernière syllabe accentuée de l'énoncé (Di Cristo, 1998). Le ton haut peut être placé plus tôt dans la question, si celle-ci est introduite par *est-ce que*, si l'on demande une confirmation plus qu'une information et même, indépendamment de ce biais, dans certaines variétés de français comme le français parlé en Corse (Boula de Mareüil *et al.*, 2012). Mais en français « standard » (parisien), la dernière syllabe accentuée de l'énoncé coïncide presque toujours avec la syllabe finale. L'accent tombe en effet sur la dernière syllabe de groupes de mots ou sur la syllabe précédant un schwa final ; or ce schwa final est presque toujours élidé en français non-méridional (Di Cristo, 1998). Quel est le comportement intonatif de ce schwa final, dans des variétés méridionales de français où le schwa est davantage prononcé ? La présente étude vise d'abord à répondre à cette question.

Dans une étude pilote portant sur le français en contact avec l'occitan, Sichel-Bazin *et al.* (2012) ont observé, dans des questions totales, des mouvements intonatifs montants ou (parfois) descendants sur ce schwa final, avec éventuellement des fonctions épistémiques différentes : demande d'information dans le premier cas, demande de confirmation dans le second. Ceci vaut-il en Languedoc comme en Provence ? Il n'en va déjà pas nécessairement ainsi du français en contact avec le catalan parlé dans le Roussillon. Les phrases intonatives du catalan montrent d'importantes variations en matière d'intonation, selon leur structure, leur fonction et le dialecte (Prieto, 2001 ; Vanrell *et al.*, 2013) ; mais le patron descendant-montant semble prévaloir dans les questions terminées par un paroxyton — avec une descente mélodique sur l'avant-dernière syllabe, accentuée, et une remontée sur la finale (Martínez Celdrán *et al.*, 2005).

Cet article rapporte les premières analyses d'enregistrements recueillis lors d'enquêtes de terrain menées en Provence et en Languedoc, auprès de 14 bilingues français-occitan.[1] Les sections suivantes présentent le contexte et les hypothèses (section 2), l'enquête et le corpus (section 3). Différents patrons intonatifs ont été relevés pour les questions totales, aussi bien en français méridional qu'en occitan. Ces observations (fournies en section 4) nous ont conduits à mettre au point une expérience perceptive, utilisant la modification de prosodie. L'expérience (section 5) visait à déterminer si, à partir de questions en français prononcées par des Provençaux et des Languedociens, des auditeurs provençaux et languedociens se montraient plus habitués à tel ou tel patron mélodique. Les résultats sont résumés dans la section 6, qui propose également quelques perspectives.

2 Contexte et hypothèses

L'occitan est une langue gallo-romane parlée principalement dans le sud de la France, mais également dans le Val d'Aran catalan, en Espagne, et quelques vallées cisalpines, en

[1] Depuis, 14 bilingues français-catalan ont été enregistrés en Roussillon (Boula de Mareüil *et al.*, 2013), mais il n'en sera pas question ici.

Italie. En France, où seule la langue de la République bénéficie d'un statut officiel reconnu par la Constitution, l'occitan est en danger (Meisenburg, 2011). Même si l'enseignement en occitan y est autorisé depuis 1951, la population parle désormais majoritairement le français, langue d'usage quotidien dans l'espace public : en France, l'occitan compterait à présent 526 000 locuteurs adultes selon Héran *et al*. (2002) — ou, moins, 110 000 locuteurs selon Bernissan (2012) —, sur un territoire traditionnellement occitanophone d'environ 15 millions d'habitants. Nous considérons ici le provençal comme un dialecte de l'occitan, sans occulter que certains y voient une langue distincte (Blanchet, 2004). De même, si Coquillon (2005) note des différences prosodiques entre français parlés dans les domaines provençal et languedocien, d'autres études mettent plutôt en avant les similitudes (Woehrling et Boula de Mareüil, 2006 ; Sichel-Bazin, 2014).

Nous nous inspirerons dans cet article du modèle métrique-autosegmental (Pierrehumbert, 198n1), récemment adapté à l'occitan, en contact avec le français méridional (Hualde, 2003 ; Sichel-Bazin *et al*., 2012, 2014). Si une montée sur la dernière syllabe accentuée de l'énoncé se poursuivant jusqu'à la fin est la configuration la plus commune (Sichel-Bazin *et al.,* 2012, 2014), d'autres schémas existent (Hualde, 2003).

D'après Sichel-Bazin *et al*. (2012, 2014), qui ont examiné les interférences entre substrat dialectal et français régional, ainsi que leur spécificité fonctionnelle, il en est de l'occitan comme du français en contact avec l'occitan. Le contour montant-montant (que nous noterons H*H%), est majoritairement utilisé quand le locuteur demande une information à son interlocuteur, sans présupposé quant à la réponse. Dans les demandes de confirmation, où la réponse attendue est affirmative, un mouvement montant-descendant (que nous noterons H*L%, montant sur la dernière syllabe accentuée de l'énoncé, descendant ensuite) est l'option préférée. Dans les questions incrédules enfin, où la réponse attendue est de type *non*, la nuance de doute est indiquée par un ton bas sur la dernière syllabe accentuée de l'énoncé, suivi d'une remontée de la mélodie (que nous noterons L*H%). Ces différents patrons sont résumés dans la figure 1.

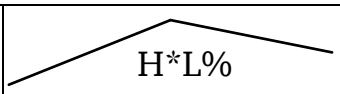
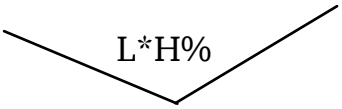
Forme du patron	H*H%	 H*L%	 L*H%
Fonction	demande d'information	demande de confirmation	question incrédule

FIGURE 1 – Représentation schématique de contours intonatifs de fin de question en occitan et en français en contact avec l'occitan, avec leur interprétation fonctionnelle d'après Sichel-Bazin *et al*. (2012, 2014).

3 Enquête et corpus

Des enregistrements ont été effectués, en occitan et en français autour d'Avignon (Provence) et de Montpellier (Languedoc). Issus de milieux socioprofessionnels variés, 5 Provençaux et 9 Languedociens (8 hommes et 6 femmes, âgés de 49 ans en moyenne) ont été enregistrés — avec un micro de haute qualité, à 44,1 kHz :

- prononçant des phrases aux structures très contrôlées, répétées selon les modalités interrogative et assertive (des phrases conçues pour être relativement transparentes dans diverses langues romanes) ;
- lisant un texte en français, avant de le traduire en occitan ;
- au cours d'entretiens semi-directifs à la fois en français et en occitan.

Les données ont été collectées en alternant entre occitan et français, à chaque changement de tâche. Les phrases contrôlées étaient présentées dans un ordre aléatoire. Les locuteurs devaient lire chaque série (en occitan ou en français, avec au moins une répétition), en commençant pour chaque phrase par la forme interrogative, suivie immédiatement de la même phrase à la forme assertive. On élicitait ainsi, pour les phrases contrôlées, des suites de questions-réponses.

Ces phrases contrôlées se pliaient aux exigences du projet AMPER (Atlas Multimédia Prosodique de l'Espace Roman) (Contini *et al.*, 2002), un des buts de l'enquête étant d'enrichir cet atlas et de permettre des comparaisons avec d'autres langues romanes, comme le catalan et le corse (Boula de Mareüil *et al.*, 2013). En accord avec le protocole AMPER, les phrases que nous avons élaborées, d'une douzaine de syllabes en moyenne, devaient avoir un verbe dissyllabique, des noms et expansions trisyllabiques avec différents patrons accentuels. Des exemples de telles phrases sont donnés dans la table 1, en graphie dite « classique » pour l'occitan. Dans la pratique, ces phrases étaient présentées aux sujets de façon adaptée aux particularités dialectales, suivant la graphie mistralienne préférée par nos locuteurs provençaux, d'où par exemple *trovo* pour « trouve ».

Français	La touriste malade trouve la caserne	Le podestat trouve la cavité profonde
Occitan	La torista malauta tròba la caserna	Lo podestat tròba la cavitat prigonda
Provençal	*La touristo malalto trovo la caserno*	*Lou poudesta trovo la cavita prefoundo*

TABLE 1 – Exemples de phrases transparentes en français et en occitan, suivant la graphie mistralienne (en italique) pour le provençal.

En français, l'accent tombe toujours sur la dernière syllabe du groupe — ou sur la syllabe précédant un schwa final prononcé (Di Cristo, 1998, *inter alia*). En occitan, sauf dans quelques variétés périphériques où les mots peuvent être accentués sur l'antépénultième, l'accent porte soit sur la dernière soit sur l'avant-dernière syllabe des mots pleins, dits selon les cas oxytons ou paroxytons, respectivement. Nous avons ainsi inclus des oxytons comme *garnison* (« garnison ») et des paroxytons comme *francesa* (« française ») dans nos phrases contrôlées, lesquelles étaient au nombre de 34 en occitan (et autant en français). Chaque série de 34 questions et 34 assertions a été répétée trois fois par les locuteurs en occitan, conformément aux directives du projet AMPER, au moins une fois en français.

4 Observations générales et phrases sélectionnées

Dans la présente étude, nous avons seulement analysé les phrases contrôlées, aux modalités assertive et interrogative. Comme dans des études antérieures (Sichel-Bazin *et al.*, 2012, 2014 ; Boula de Mareüil *et al.*, 2012, 2013), l'intonation des assertives (avec un contour descendant de finalité) présente de fortes ressemblances entre français régional et langues régionales. Dans les questions terminées par un oxyton, on trouve en Provence et en Languedoc une montée sur la dernière syllabe (accentuée) de l'énoncé, quelle que soit la variété de langue. Dans les autres questions, en revanche, certaines des différences d'intonation mentionnées plus haut ont pu être observées. Dans les questions terminées par une voyelle atone, les Provençaux et les Languedociens ont montré une grande cohérence intra-locuteur entre leurs différentes répétitions : 9 d'entre eux (7 Languedociens, 2 Provençaux) optaient régulièrement pour le patron H*H%, 5 d'entre eux (3 Provençaux, 2 Languedociens) pour le patron H*L%, en occitan comme en français.

En vue d'expériences perceptives, nous avons sélectionné deux phrases (celles de la table 1), terminées par des paroxytons — caserne et *profonde*, où le schwa final est (presque) toujours prononcé en français méridional, ou leurs contreparties en occitan. Plus de 200 réalisations en résultaient, dans les deux modalités interrogative et assertive, en français ou en occitan — avec les différentes répétitions. Ces réalisations ont été segmentées manuellement en noyaux vocaliques à l'aide du logiciel PRAAT (Boersma, 2001). Des mesures de fréquence fondamentale (F_0) ont alors été prises (en utilisant également PRAAT) au début, au milieu et à la fin de chaque voyelle. Après corrections manuelles, les résultats ont été représentés graphiquement comme exemplifié dans la figure 2 pour la schématisation de la F0 d'une phrase en français et en occitan.

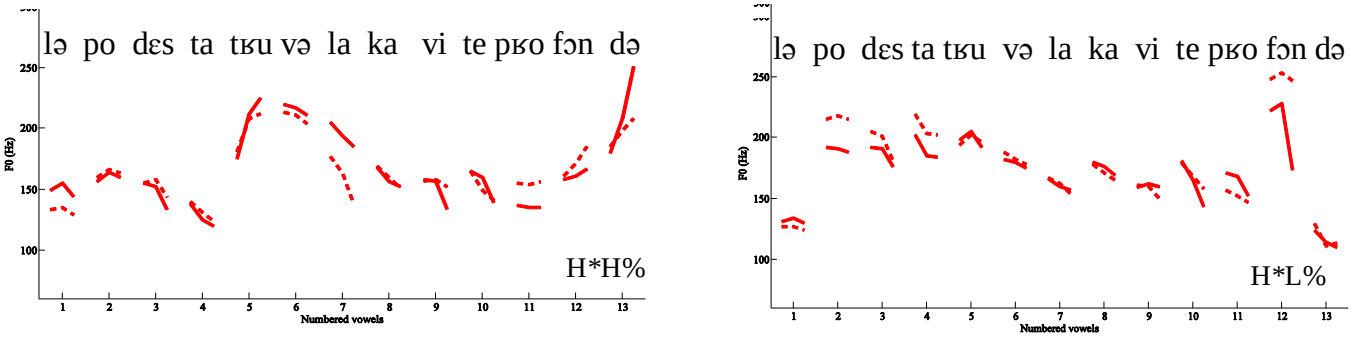


FIGURE 2 – Représentation de la F_0 fondée sur les voyelles (indépendamment de leurs durées) pour la phrase « le podestat trouve la cavité profonde ? », prononcée par un Languedocien (à gauche, avec un patron H*H%) et par un Provençal (à droite, avec un patron H*L%), en français (ligne pleine) et en occitan (en pointillé).

Ces phrases, dans leur modalité interrogative en français, ont été utilisées dans une expérience perceptive, notamment pour vérifier si différents contours intonatifs peuvent être associés à telle ou telle région. Dans le test, nous avons manipulé les contours terminaux par modification/resynthèse de la prosodie, comme explicité ci-dessous.

5 Expérience perceptive

5.1 Méthode et protocole

Le test d'écoute avait pour but de vérifier si *oui* ou *non* des différences de perception pouvaient être mises en évidence entre Provençaux et Languedociens. Il s'agissait d'un test par paires dans lequel il était demandé aux auditeurs, entre deux versions d'une même question, laquelle leur semblait plus proche, dans son intonation, de ce qu'ils étaient habitués à entendre dans leur région. Pour cette expérience, nous avons sélectionné les deux questions mentionnées ci-dessus (*cf.* table 1), en français méridional, de façon à obtenir 10 patrons H*H% et 10 patrons H*L%. Autant de patrons inversés ont été générés par resynthèse, au moyen de PRAAT, de façon à rendre montant-descendant chaque patron H*H% et montant-montant chaque patron H*L%. De plus, des patrons L*H%, absents des données de ces locuteurs occitans,[2] ont été créés. Les valeurs moyennes en demi-tons de ces différents patrons sont illustrées dans la figure 3.

Les locuteurs ont été appariés selon leur hauteur de voix moyenne, de façon à présenter des contours intonatifs différents (H*H% ou H*L%). Dans les paires de questions ainsi constituées, il était demandé de comparer des contours intonatifs différents, issus de locuteurs différents. Cette précaution visait à ne pas trop attirer l'attention des sujets sur

[2] Le patron descendant-montant L*H%, cependant, a été relevé dans les données recueillies en Roussillon.

l'objet même de notre étude. Les auditeurs écoutaient, par exemple, les deux questions illustrées dans la figure 2, séparées par un intervalle d'une seconde. Au total, 40 paires de questions (égalisés en intensité) étaient présentées aux auditeurs : 10 de type H*H%-H*L% originaux, 10 de type H*H%–H*L% synthétisés, 10 de type L*H%–H*L% synthétisés et 10 de type L*H%–H*H% synthétisés.

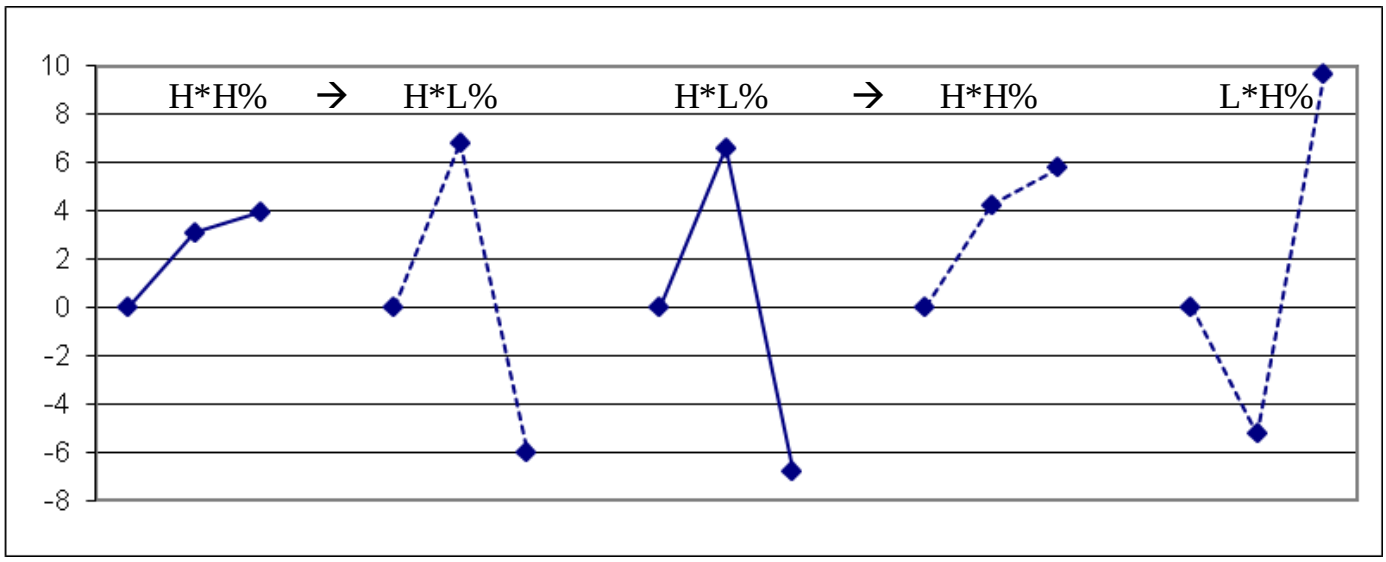


FIGURE 3 – Pentes moyennes (en demi-tons) des différences de F_0 prélevée au milieu des 3 voyelles finales de question, pour les patrons intonatifs originaux (lignes pleines) et modifiés par synthèse (en pointillés).

Le test, mené à travers une interface web, durait environ 20 minutes. Les participants étaient informés que l'expérience concernait le français parlé dans le Midi (plus précisément des questions enregistrées en Provence et en Languedoc) et qu'ils seraient amenés à écouter de la parole éventuellement modifiée acoustiquement. Ils étaient invités à utiliser un casque ou des écouteurs. Les sujets devaient d'abord fournir quelques informations à caractère biographique (âge, lieu de résidence, etc.). Suivait une phase de familiarisation avec le type de stimuli : des originaux et des modifications de prosodie non-utilisés dans le test proprement dit, pour illustrer différents patrons intonatifs des différentes régions étudiées. Dans le test proprement dit, les stimuli étaient présentés par paires, dans un ordre aléatoire différent pour chaque sujet. Les auditeurs pouvaient écouter chaque paire autant de fois qu'ils le souhaitaient, mais il était impossible de corriger des réponses, une fois une nouvelle paire présentée. Pour chaque paire de questions, les sujets devaient indiquer quelle version leur semblait la plus proche, dans l'intonation, de ce à quoi ils étaient habitués, en cliquant sur l'un des deux boutons proposés.

5.2 Auditeurs

Au total, 40 sujets ont pris part au test (17 hommes et 23 femmes, âgés de 37 ans en moyenne). Il s'agissait de 20 Provençaux et de 20 Languedociens, sans problèmes d'audition connus, qui étaient assez bien ancrés dans leur région, ayant vécu 33 ans en moyenne dans le Midi. Tous étaient de langue maternelle française, seul un sujet se déclarant bilingue français-occitan.

Avant le test, les sujets devaient préciser leur familiarité avec l'accent du Midi et avec l'occitan. Il leur était demandé s'ils pensaient pouvoir distinguer un Provençal d'un Languedocien, parlant français — ce à quoi 16 auditeurs sur 40 ont répondu par la positive. Il leur était d'autre part demandé s'il leur arrivait d'avoir (respectivement,

d'entendre) des conversations en occitan : une fois par semaine ou plus, quelques fois par mois, rarement ou jamais — à quoi 36 (respectivement, 27) auditeurs ont répondu que cela ne leur arrivait jamais ou que rarement.

5.3 Résultats

Les résultats de l'expérience perceptive sont présentés dans la figure 4, dans les différentes configurations où deux patrons intonatifs étaient appariés. Ils ont fait l'objet de tests de χ^2 qui ont montré une préférence significative, uniquement dans les stimuli modifiés par synthèse, pour le patron H*H% (face à H*L% [$\chi^2(1) = 26$; $p < 0{,}001$] et à L*H% [$\chi^2(1) = 34$; $p < 0{,}001$]).

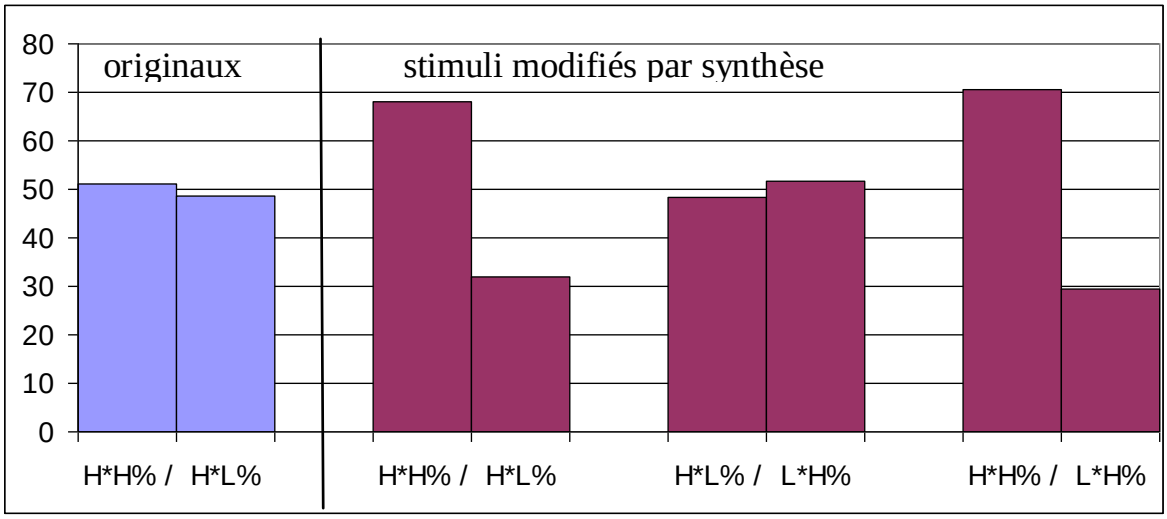


FIGURE 4 – Résultats de l'expérience perceptive : pourcentages de cas où les auditeurs se sont dits plus familiers de patrons H*H%, H*L% ou L*H%, dans différentes paires de questions où ces patrons étaient en compétition (stimuli originaux à gauche, resynthétisés à droite).

Entre les originaux H*H% et H*L%, les réponses sont équilibrées : les différences ne sont pas significatives, de même qu'entre H*L% et L*H%. Les différences entre les réponses des Provençaux et celles des Languedociens ne sont pas non plus significatives. Ces résultats suggèrent donc que, dans les deux régions étudiées, une certaine variation est tolérée en matière de contours intonatifs pour les questions. Si le patron H*H% est le mieux acceptés, les autres ne sont pas totalement rejetés, y compris le patron L*H% créé par synthèse (puisqu'il était absent des originaux).

L'analyse des résultats par locuteur n'a pas montré de tendances claires: tantôt un Languedocien est globalement perçu comme plus familier par les Provençaux, tantôt un Provençal est globalement perçu comme plus familier par les Languedociens. Dans tous les cas, aucun biais lié à l'origine commune des locuteurs et des auditeurs n'a été observé.

5.4 Discussion

D'après cette étude, le patron H*H% est le plus caractéristique des questions en français en contact avec l'occitan. On a ensuite affaire au patron H*L%. Quant au patron L*H%, il a été jugé moins familier aux auditeurs que la montée H*H%, mais rivalise avec H*L%, confronté à ce dernier patron. Ces résultats soulèvent le problème de l'interprétation et de la fonction de ces différents patrons intonatifs.

Une nouvelle expérience perceptive est en cours, interrogeant les auditeurs sur la fonctionnalité des questions. Comme précédemment, les deux phrases en français présentées dans la table 1 ont été retenues, ainsi que leurs modifications/resynthèses de prosodie. Elles ont été présentés isolément (et dans un ordre aléatoire) aux sujets de ce nouveau test, dont la tâche consistait notamment à identifier l'acte de langage réalisé (3 possibilités : demande de confirmation, question incrédule, demande d'information). Plus précisément, les auditeurs devaient indiquer si, d'après l'intonation, le locuteur (ou la locutrice) qui posait la question s'attendait plus probablement à une réponse en *oui* (demande de confirmation), *non* (incrédulité) ou bien *oui/non* (ne sait pas).

Au moment de la rédaction, 20 auditeurs provençaux et languedociens ont pris part au test. Les résultats préliminaires ne nous permettent pas d'étayer les hypothèses de Sichel-Bazin *et al*. (2012, 2014), illustrées dans la figure 1. Les auditeurs ont par exemple jugé que le patron H*H% appelait une réponse en *oui* dans 43 % des cas et en *non* dans 21 % des cas, alors que le patron H*L% appelait une réponse en *oui* dans 30 % des cas et en *non* dans 47 % des cas. D'après ces chiffres, le patron H*H% serait plutôt une demande de confirmation et le patron H*L% plutôt une question incrédule. Une analyse plus fouillée, avec davantage d'auditeurs, est cependant nécessaire avant de conclure.

6 Conclusion

Cet article a présenté une enquête de terrain menée en Provence et en Languedoc, comprenant des phrases assez transparentes en occitan et en français, dont les structures prosodiques ont été comparées. Les questions totales, en particulier, ont été analysées : des patrons mélodiques montant-montant (H*H%) et montant-descendant (H*L%) à la fin ont été observés, à la fois en occitan et en français méridional — à la différence de ce qu'on note de façon prototypique en français standard, où en raison de la chute du schwa final, la dernière syllabe de l'énoncé est celle qui porte l'accent nucléaire.

Une expérience perceptive centrée sur l'intonation des questions a été menée, à partir des enregistrements de français, utilisant la modification/resynthèse de la prosodie. Bien qu'une certaine variation soit autorisée, le test d'écoute suggère que des auditeurs provençaux et languedociens (sans différences significatives entre les deux groupes) sont plus familiers du patron H*H%. Ce résultat, en perception, rejoint l'observation que ce patron est majoritaire dans nos données en français en contact avec l'occitan, comme en occitan. Plutôt atypique dans les langues romanes, pour les paroxytons, ce patron mélodique de l'interrogation demande à être étudié de façon plus approfondie, par rapport à sa fonctionnalité.

Un test perceptif en cours, qui n'a pas (encore) permis de dégager le rôle de demande d'information associé à la montée H*H%, doit être poursuivi, étendu au Roussillon et à la parole spontanée. La méthode utilisée pour enregistrer le corpus ne cherchait pas à distinguer entre questions de significations pragmatiques différentes — chose d'ailleurs difficile à réaliser hors contexte. Il est donc nécessaire de recueillir de nouvelles données sur le terrain. Sans nul doute, en parole spontanée, l'interprétation fonctionnelle des énoncés représentera un tout autre défi.

Remerciements

Ce travail a été financé par le projet ANR PADE. Nous sommes redevables à P. Mairano, J.-P. Lai, I. Lehka-Lemarchand, à tous les locuteurs que nous avons enregistrés et à tous les auditeurs qui ont bien voulu prendre part à l'expérience perceptive.

Références

BERNISSAN, F. (2012). Combien de locuteurs compte l'occitan en 2012 ? *Revue de Linguistique Romane,* 76 : 467–512.

BLANCHET, P. (2004). L'identification sociolinguistique des langues et des variétés linguistiques : pour une analyse complexe du processus de catégorisation fonctionnelle. *MIDL*, Paris, 31–36.

BOERSMA, P. (2001). Praat, a system for doing phonetics by computer. *Glot International,* 5(9/10) : 341–345.

BOULA DE MAREÜIL, P., RILLIARD, A., MAIRANO, P., LAI, J.-P. (2012). Questions corses : peut-on mettre en évidence un transfert prosodique du corse vers le français ? *29es Journées d'Études sur la Parole*, Grenoble, 609–616.

BOULA DE MAREÜIL, P., RILLIARD, A., LEHKA-LEMARCHAND, I., IVENT, F. (2013). Regional accents and languages in France: a contrastive prosodic analysis of Romance varieties. *ISCA Workshop on Phonetics, Phonology and Languages in Contact (PPLC)*, Paris, 72–75.

CONTINI, M., LAI, J.-P., ROMANO, A., ROULLET, S., DE CASTRO MOUTINHO, L., COIMBRA, R. L., PEREIRA BENDIHA, U., SECCA RUIVO, S. M. (2002). Un projet d'Atlas Multimédia Prosodique de l'Espace Roman. *1st International Conference on Speech Prosody*, Aix-en-Provence, 227–230.

COQUILLON, A. (2005). *Caractérisation prosodique du parler de la région marseillaise*, Thèse de doctorat, Université de Provence.

DELATTRE, P. (1966). Les dix intonations de base du français. *The French Review,* 401 : 1–14.

DI CRISTO, A. (1998). Intonation in French. *In* Hirst, D. J. & Di Cristo, A., éditeurs, *Intonation systems: A survey of twenty languages*, Cambridge University Press, Cambridge, 195–218.

FÓNAGY, I. (2003). Des fonctions de l'intonation : Essai de synthèse. *Flambeau,* 29 : 1–20.

GRUNDSTROM, A. et LÉON, P. (1973). *Interrogation et intonation*, Didier, Montréal.

HÉRAN, F., FILLON, A., DEPREZ, C. (2002). La dynamique des langues en France au fil du xxe siècle. *Population et Sociétés,* 376 : 1–4.

HUALDE, J. I. (2003). Remarks on the diachronic reconstruction of intonational patterns in Romance with special attention to Occitan as a bridge language. *Catalan Journal ofLinguistics*, 2 : 181–205.

MARTINET, A. (1970). *Éléments de linguistique générale*, Armand Colin, Paris.

MARTÍNEZ CELDRÁN, E., FERNÁNDEZ PLANAS, A. M., CARRERA SABATÉ, J., ESPUNY MONSERRAT, J. (2005). Approche de la carte prosodique dialectale de la langue catalane en Catalogne. *Géolinguistique — Hors série* 3 : 103–151.

MEISENBURG, T. (2011). Prosodic phrasing in the spontaneous speech of an Occitan French bilingual. *In* Gabriel, C. & Lleó, C., éditeur, *Intonational phrasing in Romance and Germanic, Cross-linguistic and bilingual studies*, John Benjamins, Amsterdam, 127–151.

PIERREHUMBERT, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America*, 70 : 985–995.

PRIETO, P. (2001). L'entonació dialectal del català: el cas de les frases interrogatives absolutes. *In* Bover, A., Lloret, M. R., Vidal-Tibbits, M., éditeurs, *Actes del Novè Col·loqui de la North American Catalan Society*, Publicacions de l'Abadia de Montserrat, Barcelone, 347–377.

SAVINO, M (2012). The intonation of polar questions in Italian: Where is the rise? *Journal of the International Phonetic Association,* 42(1) : 23–48.

SICHEL-BAZIN, R., BUTHKE, C., MEISENBURG, T. (2012). La prosodie du français parlé à Lacaune : influences du substrat occitan. *In* Simon, C., éditeur, *La variation prosodique régionale en français*, De Boeck, Bruxelles, 137–157.

SICHEL-BAZIN, R., MEISENBURG, T., PRIETO. P. (2014). Intonational phonology of Occitan: Towards a prosodic transcription system. *In* Frota, S., & Prieto, P., éditeurs, *Intonational variation in Romance*. Oxford University Press, Oxford, sous presse.

VANRELL, M. M., MASCARÓ, I, TORRES-TAMARIT, F., PRIETO, P. (2013). Intonation as an encoder of speaker certainty: information and confirmation yes-no questions in Catalan. *Language and Speech,* 56(3) : 163–190.

WOEHRLING, C. et BOULA DE MAREÜIL, P. (2006). Identification d'accents régionaux en français : perception et analyse. *Revue Parole,* 37 : 25–65.

Caractérisation acoustico-phonétique de parole provenant de patients atteints de paralysies faciales

Julie Mauclair[1, 2] Thomas Pellegrini[1] Maxime Le Coz[1]
Marina Robert[3] Peggy Gatignol[3]
(1) IRIT-Université Paul Sabatier, Toulouse, France
(2) Université Paris Descartes, Paris , France
(3) Service ORL et de chirurgie de la face et du cou, Hôpital de la Pitié Salpétrière, Paris, France
{mauclair,pellegrini,lecoz}@irit.fr
{marina.robert,peggy.gatignol}@psl.aphp.fr

RÉSUMÉ

La caractérisation de la parole pathologique est un problème important pour la détection et le diagnostic précoces. La parole pathologique étudiée dans cet article, provient de patients paralysés faciaux, souffrant notamment de problèmes de prononciation des plosives. Plusieurs paramètres sont extraits d'un détecteur automatique de bursts en vue de caractériser cette altération. L'une des autres hypothèses étudiées ici est que les cibles vocaliques souffrent de la mauvaise prononciation des consonnes, et ont une réalisation approximative. Une analyse des voyelles et des bursts est donc proposée en fonction des différents grades de sévérité de la paralysie des patients. Une diminution significative du nombre de bursts, des écart-type des durées de ceux-ci et du F2 des voyelles /i, e, ɛ, a/ ont été observés pour les patients du corpus. Le F3 des voyelles /i, y/ est affecté, montrant que les patients ont du mal à produire ces voyelles, faisant intervenir les lèvres.

ABSTRACT

Phonetico-acoustic characterization of speech pronounced by patients impaired with facial paralysis

Characterization of pathological speech is important for early detection and diagnosis. The pathological speech that is studied in this paper is from patients who have a peripheral facial paralysis which, among other pronunciation impairments, affects their ability to pronounce bilabial sounds. The idea is then to analyse the bursts that should characterize the biliabial sounds and the vowels that suffer from the bad pronounciation of the consonants. Indeed, this article analyses the vowels and the bursts that are uttered by patients, in correlation to the various levels of the paralysis impairement. No significant differences were observed concerning the vowel durations. A decrease in the number and in the standard deviation of the duration of the bursts, as well as a decrease in the second formant values were observed for the following vowels : /i, e, ɛ, a/. Differences in the third formant values appeared for the posterior vowels /i, y/, showing that the patients have difficulties with moving their lips.

MOTS-CLÉS : Paramètres acoustico-phonétiques, traitement automatique de la parole, parole pathologique.

KEYWORDS: Acoustical and phonetical features, automatic speech processing, pathological speech.

1 Introduction

La paralysie faciale périphérique est une pathologie fréquente qui a de nombreuses conséquences dans le quotidien des patients. Parmi les conséquences notables de cette paralysie, l'articulation des patients est affectée (Gatignol et Lamas, 2004). De nos jours, il y existe toute une batterie d'outils disponible, permettant de mesurer le degré de l'atteinte mais la plupart ne tienne compte que de l'aspect physique du visage et ne mesure pas la capacité du patient à prononcer correctement des mots ou des phrases.

L'échelle d'évaluation clinique la plus utilisée au niveau internationale est celle de House et Brackmann (House et Brackmann, 1985) qui permet une cotation du degré de dysfonction faciale rapide et intégrant l'évaluation des séquelles, allant du grade I (fonctionnement normal) au grade VI (paralysie flasque totale). C'est une échelle à 6 niveaux qui permet uniquement de noter les aspects physiques de la paralysie (principalement le front, les yeux et la bouche) grâce à une observation directe effectuée par un praticien. Le tableau 1 met l'accent sur les difficultés au niveau de la bouche au travers des différents grades de l'échelle de House et Brackmann. Pour les patients atteints de paralysie faciale périphérique à partir du grade III, la littérature fait état d'un retentissement psychologique et de troubles fonctionnels tels que les troubles oculaires, les syncinésies, l'hémi-spasme, les troubles de la mastication, de la déglutition et de l'articulation (Gatignol et Lamas, 2004).

Grade	**Caractéristisques**
I Normal	Fonction faciale normale
II Atteinte discrète	Légère parésie. Au repos, symétrie et tonus normaux. **Bouche : légère asymétrie**
III Atteinte modérée	Asymétrie évidente mais non choquante. Au repos, symétrie et tonus normaux **Bouche : faiblesse des muscles quand l'effort est maximum**
IV Atteinte modérément sévère	Faibllesse ou asymétrie défigurante Au repos, tonus et symétrie normaux **Bouche : asymétrique quand l'effort est maximum**
V Atteinte sévère	Mouvement à peine perceptible Asymétrie au repos **Bouche : mouvement très faible**
VI Paralysie totale	Aucun mouvement

TABLE 1 – Les mouvements de la bouche référencés dans l'échelle de House et Brackman

Dans le domaine du traitement automatique de la parole, l'analyse de la voix et notamment de la parole pathologique est utilisée pour aider au diagnostic médical. L'analyse automatique des voix est une méthode non-invasive et est largement expérimentée en ce qui concerne les pathologies laryngées et la voix dysphonique (Drugman *et al.*, 2011; Henriquez *et al.*, 2009) afin de fournir une évaluation objective aux praticiens. L' analyse acoustique des voix pathologiques peut être pratiquée sur des voyelles tenues ou des phrases et utilise des méthodes de traitement du signal linéaires ou non linéaires (Henriquez *et al.*, 2009; Arias-Londoño *et al.*, 2011). Souvent le but de ces expériences est de discriminer la voix pathologique de la voix non-pathologique (Drugman *et al.*, 2011; Henriquez *et al.*, 2009; Arias-Londoño *et al.*, 2011). Les troubles de la parole, notamment la caractérisation de la parole dyshartrique ont par exemple été étudiés dans le projet DesPho APady (Fougeron *et al.*, 2010). Plusieurs articles font de la classification de parole dysarthrique pour caractériser plusieurs types de dysarthrie. La classification porte notamment sur des paramètres caractérisant les spécificités acoustico-articulatoires de ce type de parole (Auzou *et al.*, 2007) et plusieurs études utilisent des systèmes de traitement automatique de la parole (Gu *et al.*, 2005) pour objectiver la sévérité d'un trouble de parole. Pour les atteintes faciales, une analyse automatique de la parole pourrait aider le patient à mesurer objectivement ses progrès lors d'une thérapie. Celle-ci est un très long processus et il est souvent difficile pour les patients d'être objectifs sur l'amélioration de leur condition, même si celle-ci est faite par leurs proches ou même par les médecins.

Ici, le but de cet article est de vérifier s'il y a une corrélation entre l'atteinte physique faciale prise en compte dans l'échelle de House et Brackmann avec les difficultés articulatoires dont souffrent les patients. Afin d'appréhender les difficultés articulatoires, plusieurs paramètres caractérisant les réalisations acoustiques sont calculés pour représenter le degré d'atteinte des patients. A l'écoute de la parole des patients, l'impact le plus clair de la paralysie porte sur la réalisation approximative des consonnes et en particulier des plosives (Gatignol et Lamas, 2004). Les voyelles proches sont alors aussi altérées et leur réalisation est réduite. Pour vérifier ces hypothèses dégagées grâce à une analyse et un alignement manuel dans (Robert *et al.*, 2011), un système de détection automatique de bursts est utilisé (paragraphe 3.1). Ce système, déjà décrit dans une précédente étude pour une classification préliminaire (Mauclair *et al.*, 2013), permet ici une description plus précise du phénomème d'altération au niveau des plosives. De plus, une étude des altérations des voyelles au niveau des différents grade de paralysie faciale est également proposée (paragraphe 3.2).

Le plan de l'article est construit comme suit. La base de données dont un corpus de phrases isolées est extrait pour cet article, est décrite dans le paragraphe 2. Ensuite, nous décrirons les paramètres étudiés pour les expériences dans le paragraphe 3. Enfin, une discussion des résultats et une conclusion à ce travail seront abordées.

2 Corpus utilisé

Les expériences ont été réalisées sur une base de données enregistrée à l'hôpital de la Pitié Salpétrière de Paris. Celle-ci contient les enregistrements de 70 personnes, des patients et des témoins sains. Les patients souffrent de paralysie faciale périphérique d'origine virale, dûe à des neurinomes (du nerf acoustique ou du nerf facial) ou encore d'origine traumatique. De plus amples détails sur les patients de ce corpus sont disponibles dans (Robert, 2011). Quelques

enregistrements proviennent du même patient à plusieurs grades de paralysie. Les patients ont été enregistrés dans une cabine insonorisée avec un microphone hypercardoïde, grâce à un enregistreur numérique au format 16 bits/44,1kHz PCM WAV.
Plusieurs types d'enregistrements ont été réalisés. Les patients ont eu plusieurs textes à lire, comme des mots isolés, des phrases isolées, des passages de journaux. Ils ont aussi répondu de façon spontanée à la question : "Comment faîtes vous une omelette ?"

Dans cet article, seule la partie où les patients et témoins ont lu les phrases isolées a été exploitée. 32 enregistrements de 32 patients ont été utilisés, équitablement répartis dans quatre classes que nous détaillerons par la suite. Chaque personne prononce 17 phrases (telles que : "Pierre et Paul préparent la purée de petits pois"), ce qui correspond approximativement à 1 minute par patient. L'épreuve de lecture de phrases comporte 14 phrases testant le lieu d'articulation des phonèmes : bilabiaux, alvéolaires, vélaires, labiodentaux et palataux. Trois autres phrases opposent les modes d'articulation : occlusives et constrictives constrictives sonores et sourdes et nasales.

Les quatre classes étudiées dans ce papier correspondent à 5 grades de paralysie faciale (voir Tableau 1), les grades I, III, IV et les V&VI. Ce choix d'une classification un peu différente des grades de la paralysie a été fait pour deux raisons. La première est que, les patients de grade II ne ressentent plus le besoin de venir consulter, ils n'ont donc pas pu être enregistrés . Ensuite, les différences entre les grades V et VI ne correspondent pas aux mouvements de la bouche et nous avons choisi de les regrouper dans la même classe. Le grade I est constitué de patients témoins n'ayant jamais présenté de paralysie faciale, car il est quasi-impossible pour les paralysés de revenir à une fonction de la face normale.

3 Expériences

Pour étudier ces différentes classes, plusieurs paramètres acoustico-phonétiques sont calculés. Dans un premier temps, nous avons choisi des paramètres caractérisants les plosives. En effet, ces phonèmes sont, à l'écoute des enregistrements des patients, assez mal prononcés(Robert *et al.*, 2011). Dans une précédente étude(Mauclair *et al.*, 2013), cette paramétrisation était faite à l'aide d'un détecteur de bursts. Nous reprenons ici cette paramétrisation pour en extraire une étude plus fine dans le paragraphe 3.1. Une autre hypothèse étudiée ici est que nous pensons que si plusieurs consonnes sont altérées par la paralysie faciale, les voyelles alentour peuvent l'être aussi, d'une part par le phénomène de coarticulation provenant de la consonne altérée précédente ou suivante et d'autre part par la nécessité pour la production de quelques-unes de faire intervenir les lèvres. Nous avons donc étudié l'espace vocalique de celles-ci (voir paragraphe 3.2). Pour les analyses statistiques, nous nous sommes basés sur un T-test avec un seuil de significativité de 10%.

3.1 Analyse des bursts

La détection des bursts se fait automatiquement grâce à la méthode développée dans (Hosom et Cole, 2000). Cette méthode permet d'obtenir les valeurs en amplitude des pics d'énergie en fonction du temps. Ici, le but n'est pas de connaître le nombre de bursts correctement détectés mais plutôt de voir s'il y a une différence dans le nombre de bursts détectés en fonction des

différents grades et si la représentation des bursts diffère. La figure 1 montre qu'il existe une diminution du nombre de bursts détectés en fonction du grade de la paralysie.

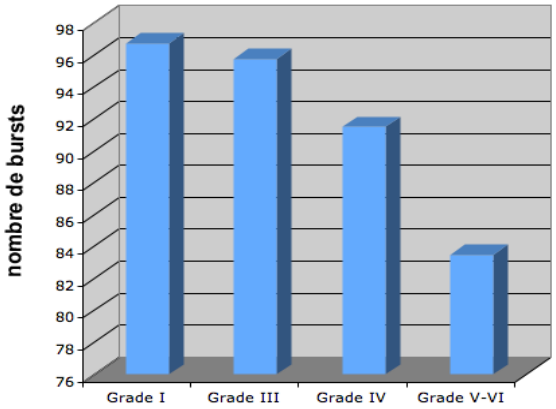


FIGURE 1 – *Représentation du nombre de bursts par grade de paralysie faciale*

Ces bursts forment des gaussiennes (voir figure 2)dont l'amplitude et la dispersion sont intéressantes à étudier pour caractériser le degré d'altération dans la prononciation des plosives. En effet, l'écart type des durées des bursts (et donc la dispersion de ces bursts) semble être plus élevé avec des grades de paralysies élevés, car la paralysie faciale entraîne un ralentissement dans la production de cette explosion.

Plusieurs paramètres sont extraits du résultat du détecteur de bursts pour caractériser ces plosives. Nous avons choisi d'étudier le nombre de bursts détectés, le nombre de bursts détectés dont l'amplitude est supérieure à la moyenne de nombre de bursts, la moyenne des écart type des durées des bursts et le nombre de bursts dont l'écart type des durées est inférieur à la moyenne des écart type des durées.

La différence en terme de nombre de bursts détectés n'est pas significative entre les locuteurs sains (grade I) et ceux des grades III et IV. Par contre elle est significative entre les patients des grades III et V&VI et entre les locuteurs sains et ceux du grade V&VI (p-value respectivement de 0.08 et 0.06). Si on tient compte maintenant uniquement du nombre de bursts dont l'amplitude est supérieure à l'amplitude moyenne, l'écart devient significatif entre les locuteurs sains et les grades IV ainsi qu'entre les locuteurs sains et ceux du grade V (p-value de 0.09 et p-value de 0.08). Ce qui est très intéressant est la mesure des écart-types de chacun des bursts pour les différents grades. La différence est significative entre la moyenne des écart-types des durées des bursts des locuteurs du grade III et du grade I (p-value de 0.004), ainsi qu'entre le grade IV et le grade I (p-value de 0.009) et entre le grade V&VI et le grade I (p-value de 0.02). Ceci est également vérifié entre les grades pathologiques, la différence est significative. Quand on comptabilise seulement les écart-types qui sont inférieurs à la moyenne des écart-types sur la durée des bursts, seul le grade V&VI a une différence significative avec les locuteurs sains (p-value de 0.07).

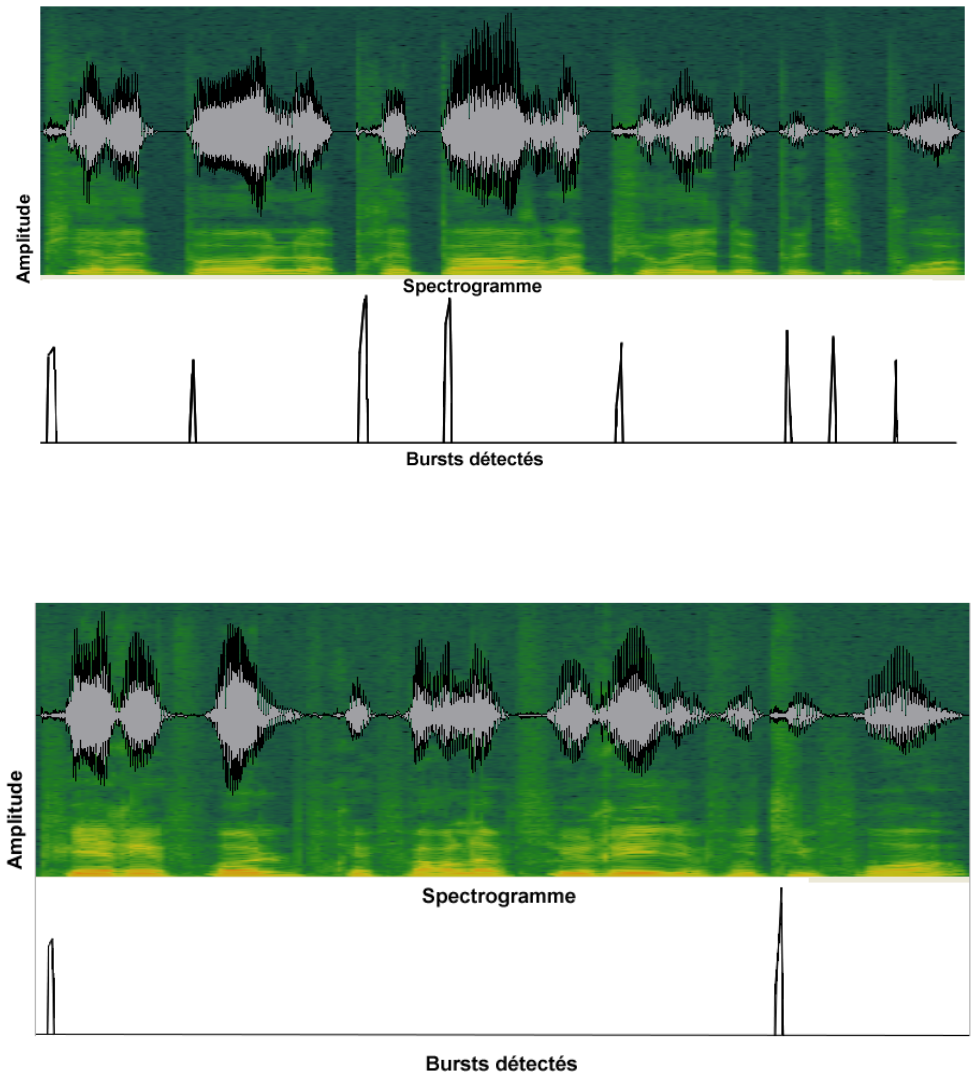


FIGURE 2 – *Représentation des bursts détectés pour un sujet grade I (en haut) et un patient grade VI (en bas) prononçant :" Pierre et Paul préparent la purée de petits pois*"

3.2 Espace vocalique

Pour calculer les valeurs des formants analysés, un alignement forcé est réalisé à l'aide du plugin Easyalign de Praat (Goldman, 2011; Boersma, 2001). Nous avons considéré les onze voyelles orales suivantes : a, e, i, o, u, y, ɛ, ə, ɔ, ø, œ. Le nombre total de voyelles segmentées est d'environ 5000. Le nombre d'occurrences de chaque voyelle correspond approximativement au pourcentage observé en français : les voyelles /a/ et /œ/ sont la plus fréquente (27%) et la moins fréquente (1%) respectivement. À l'aide du logiciel Praat, nous avons extrait F0, les trois premiers formants, l'énergie et les durées de ces voyelles. Un filtrage sur le deuxième formant a été réalisé pour écarter les voyelles donnant des valeurs aberrantes, en prenant une tolérance de plus ou moins 300Hz autour de valeurs canoniques tirées de (Gendrot et Adda-Decker, 2005).

La figure 3 représente les triangles vocaliques des locuteurs masculins (à gauche) et féminins (à droite) du corpus. Les triangles vocaliques des groupes III et IV sont très similaires au triangle du

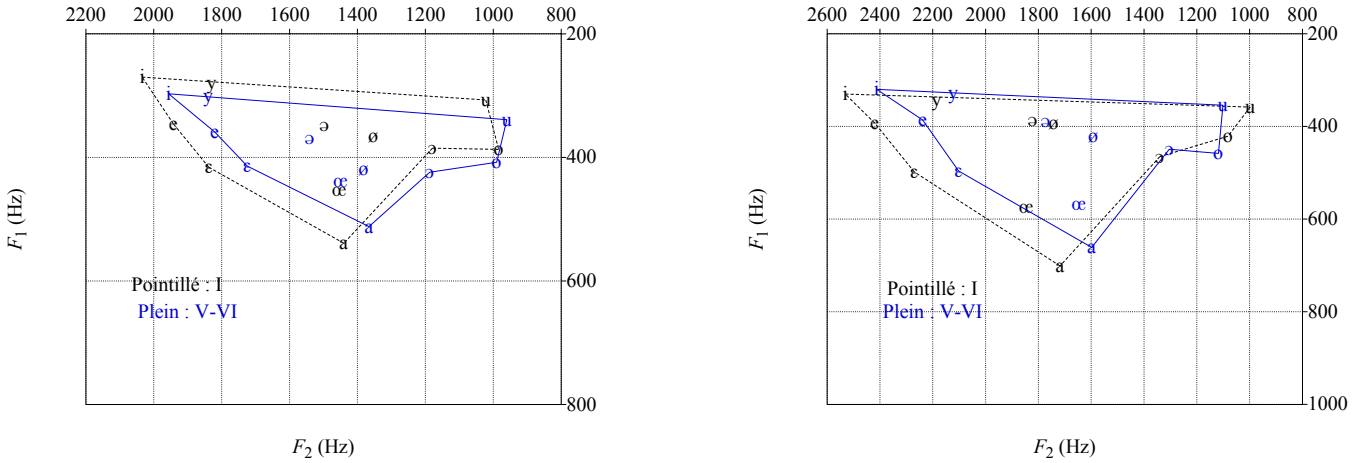


FIGURE 3 – *Représentation F1/F2. Gauche : locuteurs masculins, droite : locuteurs féminins*

groupe V&VI pour les femmes. Pour les hommes, le triangle du groupe III est similaire à celui du groupe V&VI, alors que le triangle IV est plus proche du I. Il nous semble à l'écoute que l'un des locuteurs classé dans le groupe IV a plus d'aisance dans sa lecture que les autres locuteurs de son groupe, ce qui pourrait expliquer ce résultat. Nous avons choisi de ne représenter que les triangles I et V&VI pour faciliter la lecture des figures. Tant chez les hommes que chez les femmes, une différence apparaît nettement entre le groupe témoin I et le groupe de sujets V&VI. Les espaces vocaliques des locuteurs V&VI sont déplacés vers des valeurs de F2 réduites pour les voyelles /i, e, ɛ, a/. Cette diminution des valeurs du F2 est significative pour toutes ces voyelles chez les hommes comme chez les femmes (p-value<0.001). Elle reste également significative si l'on compare le groupe témoin avec les deux groupes III et IV. Ces voyelles sont donc plus postérieures, ce qui peut s'expliquer par le fait que ces personnes souffrent d'hypokinésie due à leur paralysie faciale. Ce résultat est similaire à celui obtenu dans (Audibert et Fougeron, 2012) avec un groupe de locuteurs cérébelleux atteints de dysarthrie ataxique caractérisée par une altération de la coordination temporo/spatiale lors de l'exécution des mouvements.

Il nous a semblé également intéressant de considérer le troisième formant des voyelles /i,y/. En effet, ces voyelles qui peuvent être arrondies ou étirées font particulièrement intervenir la mobilité labiale. L'étirement des lèvres augmente la valeur de F3 (voyelle /i/) et l'arrondissement la diminue (/y/). On s'attendrait donc à une diminution du F3 du /i/ et une augmentation du F3 du /y/ qui s'accentuent avec l'importance de la paralysie. Dans (Robert *et al.*, 2011), nous avions trouvé des valeurs de F3 significativement différentes entre les locuteurs sains et les patients III, IV et V&VI mais sur un corpus de parole lue différent. Dans la présente étude, le /i/ a un F3 effectivement plus bas chez les patients V&VI par rapport aux groupe I (femmes : 3175Hz contre 3475Hz avec p-value=0.04, hommes : 2705Hz contre 2982Hz avec p-value=0.006). Cette diminution est aussi valable si l'on compare les locuteurs sains et les groupes III et IV.

En ce qui concerne la voyelle /y/, on s'attend à une augmentation du F3 due à la difficulté qu'ont les patients à arrondir les lèvres. La figure 4 illustre la différence du F3 pour ces deux voyelles en fonction du grade de sévérité. Une augmentation est en effet observée chez les hommes de manière progressive avec la pathologie (sauf pour le groupe III qui a une valeur plus faible) :

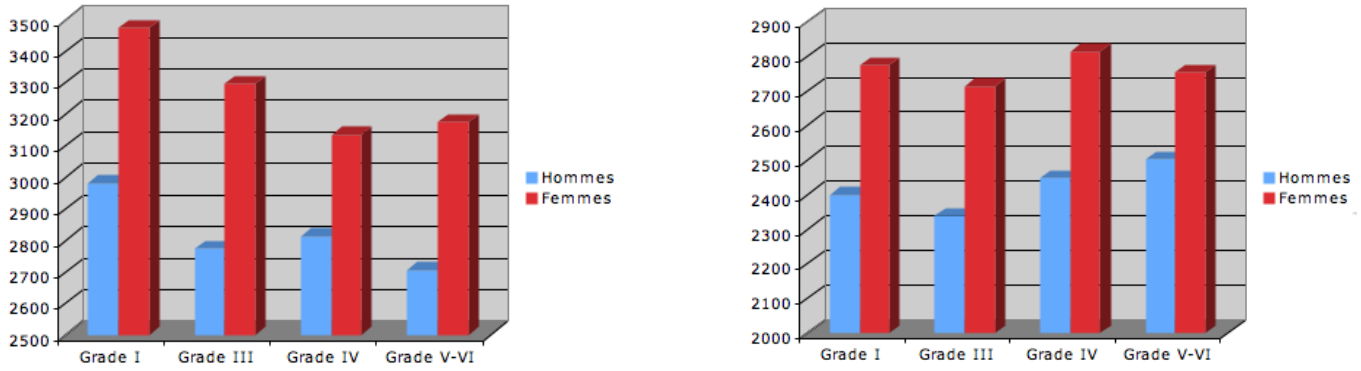


FIGURE 4 – *Valeurs des troisièmes formants pour les voyelles /y/ (à gauche) et /i/ (à droite) pour les hommes et les femmes*

2502Hz pour le groupe V&VI contre 2400Hz pour le groupe I (p-value=0.02). Chez les femmes, une augmentation est observée si l'on compare le groupe I (2775Hz) et le groupe IV (2814Hz). En revanche, une faible diminution de 20Hz est observée pour les groupes III et V&VI.

En ce qui concerne les durées des voyelles, aucune différence significative n'a été mise en évidence entre les différents groupes de locuteurs (durée moyenne autour de 100ms).

4 Discussion et conclusion

Nous avons étudiés ici deux groupes de phonèmes qui nous semblent altérés par la paralysie faciale. Les différences sont en générales très significatives entres les paramètres des locuteurs sains et ceux des locuteurs de grade V&VI. Cependant, elles s'atténuent pour les grades qui se trouvent entre ces extrêmes. Les seuls paramètres qui permettent de distinguer les classes médianes sont les moyennes des écart-types des bursts, les valeurs du F2 pour les voyelles /e, ɛ/ et les valeurs du F3 pour la voyelle /i/. Il est également vrai qu'à l'écoute des enregistrements, il est parfois difficile d'entendre la difficulté d'articulation, notamment de certains patients de grade III. Ces patients ont en effet parfois obtenu cette gradation plutôt à cause de la mauvais mobilité de leur oeil que de celle de leur bouche et ont, à l'écoute des enregistrement, plus d'aisance en lecture que d'autres patients de leur groupe.

Dans cet article, nous avons proposé une caractérisation de plusieurs phénomènes acoustico-phonétiques en vue d'étudier des enregistrements de patients atteints de paralyse faciale. La classification en terme de sévérité de l'atteinte proposée par les orthophonistes est une classification sur une échelle tenant compte non seulement des mouvements de la bouche mais également des mouvements de la joue et de l'oeil. Cette classification a donc des limites en terme de ce que l'on peut espérer corréler avec la prononciation des patients. Cependant, entre les grades extrêmes, la caractérisation au niveau des voyelles et en terme de détection des bursts donne des résultats significativement différente pour espérer une classification plus efficace que celle auparavant réalisée avec la détection de bursts seule dans (Mauclair *et al.*, 2013). Une paramétrisation agrégeant cette caractérisation en terme de bursts à une caractérisation de certaines voyelles pourraient améliorer les résultats.

Références

Arias-Londoño, J., Godino-Llorente, J., Sáenz-Lechón, N., Osma-Ruiz, V. et Castellanos-Domínguez, G. (2011). Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Transactions on Biomedical Engineering*, 58(2):370 –379.

Audibert, N. et Fougeron, C. (2012). Distorsions de l'espace vocalique : quelles mesures ? application à la dysarthrie. *Journées d'Étude sur la Parole (JEP 2012)*.

Auzou, P., Ozsancak, C., Pinto, S. et Rolland, V. (2007). *Les dysarthries*. Solal.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International 5*, 9:341–345.

Drugman, T., Dubuisson, T. et Dutoit, T. (2011). Phase-based information for voice pathology detection. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4612 –4615, Prague, Czech Republic.

Fougeron, C., Crevier-Buchman, L., Fredouille, C., Ghio, A., Meunier, C., Chevrie-Muller, C., Audibert, N., Bonastre, J.-F., Simon, A., Delooze, C., Duez, D., Gendrot, C., Legou, T., Lévêque, N., Pillot, C., Pinto, S., Pouchoulin, G., Robert, D., Vaissiere, J., Viallet, F. et Vincent, C. (2010). Developing an acoustiphonetic characterization of dysarthric speech in french. *In 7th international conference on Language Resources and Evaluation*, pages 2831–2838.

Gatignol, P. et Lamas, G. (2004). *Paralysies faciales*. Solal Editions.

Gendrot, C. et Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in french and german. *Interspeech 2005*, pages 2453–2456.

Goldman, J. (2011). Easyalign : an automatic phonetic alignment tool under praat. *Interspeech 2011*, pages 3233–3236.

Gu, L., Harris, J.-G., Rahul, S. et Sapienza, C. (2005). Disordered speech evaluation using objective quality measures. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 321–324.

Henriquez, P., Alonso, J., Ferrer, M., Travieso, C., Godino-Llorente, J. et Diaz-de Maria, F. (2009). Characterization of healthy and pathological voice through measures based on nonlinear dynamics. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1186–1195.

Hosom, J.-P. et Cole, R. A. (2000). Burst detection based on measurements of intensity discrimination. *In International Conference on Spoken Language Processing (ICSLP)*, pages 564–567, Beijing, China.

House, J. et Brackmann, D. (1985). Facial nerve grading system. *Otolaryngology-Head and Neck Surgery*, 93:146 – 147.

Mauclair, J., Koenig, L., Robert, M. et Gatignol, P. (2013). Burst-based features for the classification of pathological voices. *Interspeech 2013*, pages 2167–2171.

Robert, M. (2011). Analyse acoustique des troubles articulatoires chez les patients atteints de paralysie faciale périphérique. Mémoire de Master, Ecole d'Orthophonie de Paris VI.

Robert, M., Mauclair, J., Lannadere, E., Tankéré, F., Lamas, G. et Gatignol, P. (2011). Analyse des troubles articulatoires au cours des paralysies faciales périphériques. *LXVII Congrès de la Société Française de Phoniatrie*.

ROOTS : un outil pour manipuler facilement, efficacement et avec cohérence des corpus annotés de séquences

Jonathan Chevelu Gwénolé Lecorvé Damien Lolive
IRISA / Université de Rennes 1 – 6, rue de Kerampont, Lannion, France
{jonathan.chevelu, gwenole.lecorve, damien.lolive}@irisa.fr

RÉSUMÉ

De nombreux outils de traitement automatique de la parole et du langage naturel permettent aujourd'hui d'annoter des documents écrits et oraux. Cependant, cette richesse logicielle conduit à une grande diversité de formats de fichiers et de types d'information. Du fait de cette hétérogénéité, le développement de processus de traitements complexes nécessite souvent de convertir et d'aligner, parfois de manière répétée, de nombreuses informations. Pour pallier ce problème, cet article présente ROOTS, un outil libre dédié à la manipulation homogène de données séquentielles annotées. Écrit en C++ et disposant d'une interface de programmation (API) dans plusieurs langages, ROOTS est rapide et facile à prendre en main. Ces propriétés sont étayées par deux exemples applicatifs.

ABSTRACT

Roots: a tool for easy, efficient and consistent manipulation of annotated sequence corpora

Numbers of speech and natural language processing tools are nowadays available to annotate spoken and textual documents. However, this diversity of software leads to numerous different file formats and information types which are inherently not interoperable. Hence, developing complex data processing processes often implies to convert and align, sometimes repeatedly, these pieces of information. To overcome this problem, this paper presents ROOTS, a tool for homogeneous management and processing of annotated sequential data. As ROOTS is written in C++ and provides an efficient application programming interface (API) in several programming languages, it is fast and user-friendly. These porperties are demonstrated on two sample tasks.

MOTS-CLÉS : Outil logiciel, gestion de corpus, représentation de données annotées.

KEYWORDS: Software tool, corpus management, annotated data representation.

1 Introduction

Mettre au point de nouvelles méthodes de traitement automatique de la parole ou du langage naturel passe fréquemment, au-delà de l'aspect scientifique, par la résolution de divers problèmes techniques liés à la gestion des données. Tout d'abord, parce que le type des données, leur distribution et la nature des descripteurs nécessaires à deux tâches différentes sont rarement les mêmes. Par exemple, dans le cadre du traitement automatique des langues (TAL), le développement d'un extracteur d'entités nommées requiert généralement un corpus de textes annotés en informations morphosyntaxiques et syntaxiques et dont la proportion d'entités nommées est élevée. Alors que l'apprentissage d'un phonétiseur requiert avant tout un corpus aligné de graphèmes et de phonèmes, éventuellement agrémentés d'informations étymologiques sur les mots représentés,

les distributions *a priori* de ces différents éléments restant, elles, plus secondaires.

Ensuite, il existe beaucoup de systèmes spécialisés ou de ressources expertes capables de fournir des annotations (des étiqueteurs en partie du discours (POS), des dictionnaires de prononciation, des listes de noms propres...) mais la plupart de ces outils utilisent des formats spécifiques, travaillent sur des échelles temporelles différentes ou s'appuient sur des jeux d'étiquettes qui leur sont propres. Ce constat est également valable pour les outils d'annotation manuelle puisque de multiples logiciels co-existent sans véritablement partager de format en commun.

Cette hétérogénéité induit souvent un travail long et rébarbatif de conversion et d'alignement des informations issues de différents outils. Éventuellement, certaines informations sont simplifiées ou perdues durant ces étapes, d'autres sont redondantes car stockées de différentes manières dans divers fichiers. Au delà de la perte de temps, cette complexité peut conduire à des erreurs dans les programmes et dans les résultats, ceci d'autant plus fréquemment que les tâches à résoudre sont aujourd'hui de plus en plus élaborées.

L'objectif de cet article est de présenter un outil, appelé ROOTS, qui homogénéise la gestion de ces données annotées et permet de simplifier le développement d'applications dans le cadre du traitement de la parole et du langage naturel, notamment à des fins de recherche. ROOTS est le fruit de plusieurs années de développement en interne et, comme nous pensons qu'il devrait profiter à la communauté, il est depuis peu librement téléchargeable sur Internet (`http://roots-toolkit.gforge.inria.fr`). Cet article présente cette première version publique de l'outil.

Précisément, la section 2 détaille tout d'abord l'architecture de ROOTS. Notre outil est ensuite comparé à d'autres outils existants à la section 3, puis son fonctionnement pratique et ses performances sont illustrés à la section 4 à travers deux tâches : l'annotation non supervisée d'un large corpus de livres numériques et l'extraction de statistiques à partir d'annotations.

2 Présentation de l'outil Roots

Afin d'homogénéiser et de synchroniser les différents niveaux d'annotation d'un corpus, (Barbot *et al.*, 2011) ont initialement proposé ROOTS (pour *Rich Object Oriented Transcription System*) comme une solution fondée sur des relations algébriques matricielles. Un prototype en Perl implémentant cette solution a été validé dans (Boëffard *et al.*, 2012) par la mise au point d'un processus de construction automatique d'un corpus de livres audios à des fins de synthèse vocale. Ce processus tirait profit de ROOTS pour représenter de manière homogène de nombreux niveaux d'annotations.

Depuis, une nouvelle implémentation en C++ a été développée, celle-ci faisant de ROOTS un outil complet, facile à prendre en main et capable de traiter de larges collections de données séquentielles annotées, en particulier issues de documents oraux ou écrits. Cette section en donne une vue d'ensemble en présentant son architecture logique des données, les fonctionnalités mises à disposition des utilisateurs et ses aptitudes en terme de stockage.

2.1 Architecture logique des données

Comme le résume la figure 1, les données sont structurées hiérarchiquement dans ROOTS. Fondamentalement, les données sont modélisées comme des séquences d'items. Ces items sont

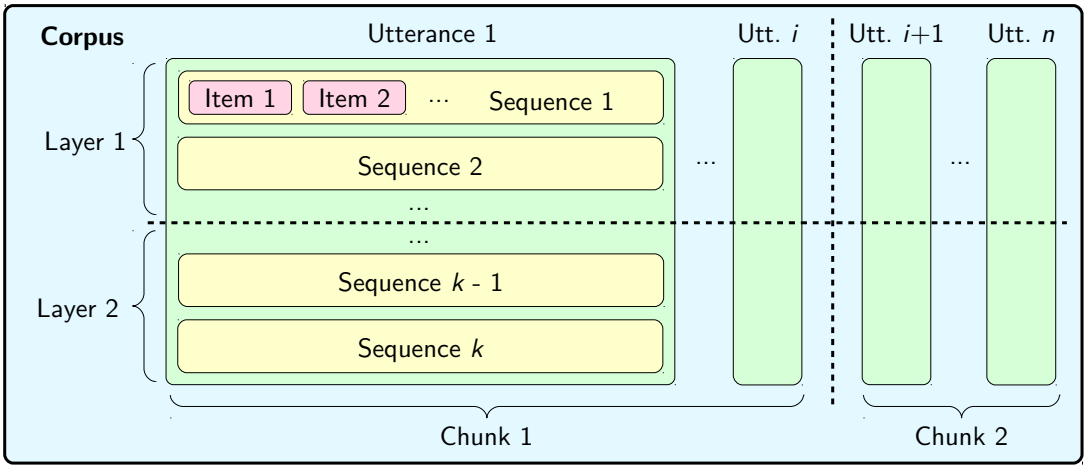


FIGURE 1 – Organisation logique des données dans ROOTS.

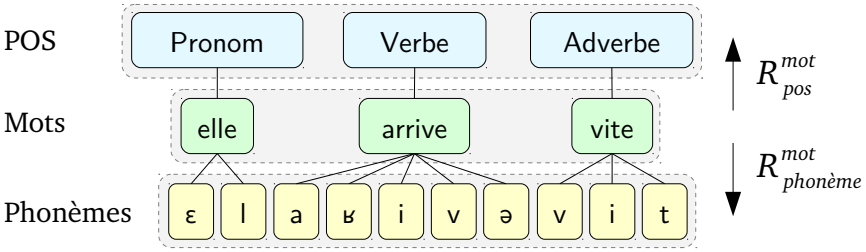


FIGURE 2 – Exemple de 3 séquences liées par 2 relations, R^{mot}_{pos} et $R^{mot}_{phonème}$.

typés, par exemple, en mots, en graphèmes, en classes d'entités nommées…, et peuvent donc représenter différents niveaux d'annotations des mêmes données comme autant de séquences de types différents. Les correspondances entre items des différentes séquences sont définies comme des relations algébriques représentées sous forme de matrices. L'ensemble de ces relations permet de produire un graphe non orienté dont les nœuds sont des items et les arêtes sont dérivées des relations. La composition des relations algébriques qui forment le plus court chemin entre deux séquences permet de trouver des correspondances entre items même en l'absence de relation directe. L'exemple d'une séquence de mots annotée en POS et en phonèmes est donnée en figure 2. Les lecteurs intéressés trouveront tous les détails théoriques dans (Barbot *et al.*, 2011).

Les séquences d'un même contenu sont regroupées en énoncés, ou *utterances* dans la terminologie de ROOTS. Ils peuvent correspondre à n'importe quelle unité pertinente pour un domaine donné (un mot isolé, une phrase, un groupe de souffle…). Rassemblés en une liste, les énoncés forment un corpus. Beaucoup de structures peuvent être modélisées ainsi : des paragraphes, des chapitres, des livres, des reportages télévisés, des brèves d'actualités… Pour hiérarchiser ces structures, un corpus peut être divisé en sous-corpus, par exemple pour représenter un chapitre comme une liste de paragraphes. Ces divisions « verticales » (*cf.* figure 1) d'un corpus sont appelées *chunks* dans ROOTS. Les corpus peuvent aussi être découpés « horizontalement » en couches d'informations, ou *layers*. Ces couches permettent, au besoin, d'isoler certaines annotations relevant d'un même niveau d'abstraction. Par exemple, dans le cadre d'un corpus de journaux télévisés, il peut être intéressant de découper le corpus en 3 couches : l'une pour les informations acoustiques et phonétiques (fréquences fondamentales, spectres, allophones…) ; une autre regroupant les informations d'ordre linguistique (transcription orthographique, étiquetages en POS, arbres syntaxiques…) ; une dernière pour des métadonnées (heure de diffusion, lieu d'un reportage…).

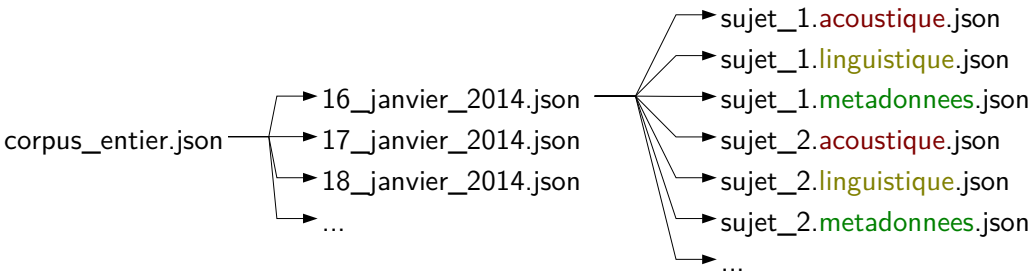


FIGURE 3 – Exemple de structuration des fichiers pour un corpus de journaux télévisés.

2.2 Interface de programmation et outils

ROOTS s'appuie sur une bibliothèque écrite en C++ et rassemble un ensemble d'utilitaires. La bibliothèque représente environ 33 000 lignes de code. Elle s'accompagne d'une API riche et largement documentée qui propose en particulier les fonctionnalités suivantes pour les différents éléments manipulés :

- **Items :** obtenir et définir les propriétés d'un item, comme le genre et le nombre pour un POS ; obtenir les items liés dans d'autres séquences.
- **Relation :** tester, obtenir les items liés ; lier ou délier des items ;
- **Séquence :** ajout, suppression, extraction et mise à jour d'un ou plusieurs items.
- ***Utterance* :** ajout, suppression, extraction et mise à jour des séquences et relations internes ; obtenir des relations directs ou composées entre séquences ; combiner, scinder et sauvegarder.
- **Corpus :** ajout, suppression, extraction et mise à jour d'un énoncé ; gestion de la structure logique et du stokage (voir la section 2.3) ; sauvegarde et chargement.

Pour permettre le prototypage simple et rapide de programmes, cette API est aussi disponible pour le langage Perl grâce au générateur automatique d'interface SWIG. Ainsi, ROOTS peut être porté à moindres frais vers beaucoup d'autres langages de programmation. En particulier, un portage vers le langage Python est prévu.

Des scripts utilitaires fondés sur l'API Perl sont disponibles pour les opérations les plus utiles (fusions et découpages de corpus, la recherche dans un corpus, la visualisation textuelle comme graphique des données...). D'autres scripts permettent également les imports et exports depuis et vers les outils de traitement de la parole HTK (Young *et al.*, 2009), Praat (Boersma, 2002), WaveSurfer (Sjölander et Beskow, 2000) et Transcriber (Barras *et al.*, 2001). ROOTS permet d'ajouter simplement d'autres formats, par exemple ceux de type CSV.

Toutes les sources, la documentation et des tutoriaux sont disponibles[1] en ligne sur `http://roots-toolkit.gforge.inria.fr`.

2.3 Stockage des données

Les informations sont sauvegardées sous forme de fichiers dans ROOTS, et non dans des bases de données. Ainsi les opérations rudimentaires sur des fichiers (déplacer, supprimer, copier...) sont connues de tous les utilisateurs et ne requièrent pas de compétence particulière. Par défaut, un corpus est simplement stocké comme un unique fichier contenant tous les énoncés, séquences, items et relations. Si le corpus est structuré en sous-corpus ou en couches, un fichier par subdivision est créé. Cette stratégie permet d'alléger les opérations de lecture et d'écriture en ne ciblant que certaines portions utiles du corpus entier. De plus, chaque sous fichier est lui-même

1. Sous licence *GNU Lesser General Public Licence* (LGPL) v3.0.

vu comme un corpus ROOTS et peut donc être utilisé indépendamment. Cela permet, par exemple, de tester et régler un programme sur une fraction seulement d'un vaste corpus avant de le lancer sur son intégralité. En reprenant l'exemple d'un corpus de journaux télévisés, nous pouvons imaginer une répartition des fichiers telle que présentée par la figure 3. Le fichier en point d'entrée (à gauche) représente le corpus entier. Celui-ci énumère simplement la liste des fichiers contenant les annotations du journal de chaque jour. De même, chacun de ces fichiers pointe vers les fichiers contenant ses différentes couches d'informations, sujet par sujet. En pratique, cette structuration des fichiers est transparente pour les utilisateurs qui souhaitent consulter le corpus. Enfin, l'utilisation des ressources est optimisée par un cache sur les *chunks* retardant au maximum le chargement des fichiers, et donc les allocations de mémoire correspondantes.

Deux formats de fichiers sont actuellement disponibles : XML et JSON. Ils sont lisibles sans outil particulier et sont appropriés pour la mise en place de services web. D'autres formats vont être ajoutés dans les versions futures, notamment un format binaire souhaitable pour développer des applications très rapides.

3 Comparaison avec les outils existants

D'autres outils permettent le traitement de données annotées. Parmi eux, le système GATE (Cunningham, 2002) propose d'interconnecter des flux d'annotations pour le développement de briques de TAL. Malgré cela, GATE s'appuie sur un environnement de développement intégré et ne permet donc pas facilement d'intégrer des outils tiers.

Plus récemment, la boîte à outils NXT a été proposée pour la gestion de corpus multimodaux (Carletta *et al.*, 2005; Calhoun *et al.*, 2010). Celle-ci s'intéresse à de grands corpus multiplement annotés en proposant un modèle générique d'organisation des données. Ce modèle s'apparente à celui d'une base de données, avec un accès aux données par un langage de requête. À l'inverse, ROOTS laisse à l'utilisateur le soin de parcourir les données comme bon lui semble. Par ailleurs, les mécanismes permettant de mettre en relation les annotations de différents ordres ne sont pas optimisés dans NXT alors que cet aspect est au cœur de notre outil.

Dans une approche plus générale, UIMA (Ferrucci et Lally, 2004; Ferrucci *et al.*, 2006) propose des normes de génie logiciel pour la gestion de données non structurées, y compris leur annotation et leur traitement. Cependant, UIMA est, à notre sens, trop élaboré techniquement pour permettre de prototyper rapidement et facilement des idées de solutions, comme l'exige souvent un travail de recherche. Il est donc plutôt dédié à des développements industriels. En outre, UIMA prône un format commun pour les outils de traitement de données alors que la philosophie de ROOTS est de permettre la construction à moindre coût de ponts entre les formats spécifiques de ces outils.

Dans cet esprit, ROOTS se rapproche des travaux réalisés dans le cadre du système de synthèse de la parole Festival (Black *et al.*, 2002). Ce système s'appuie sur le formalisme HRG, pour *Heterogenous Relation Graphs*, qui vise à offrir un mode de représentation unique des différents niveaux d'informations intervenant dans un système de synthèse (Taylor *et al.*, 2001). HRG propose ainsi une bibliothèque C++ regroupant des classes pour de multiples structures linguistiques, phonologiques, acoustiques. . ., et leur mise en relation. Notre outil s'en distingue néanmoins car HRG est partie intégrante du système Festival alors que ROOTS est complètement autonome. Par ailleurs, ROOTS dispose d'une véritable interface de programmation (API), en C++ et en Perl.

Enfin, notons à toute fin utile que les problèmes engendrés par la manipulation de très grands

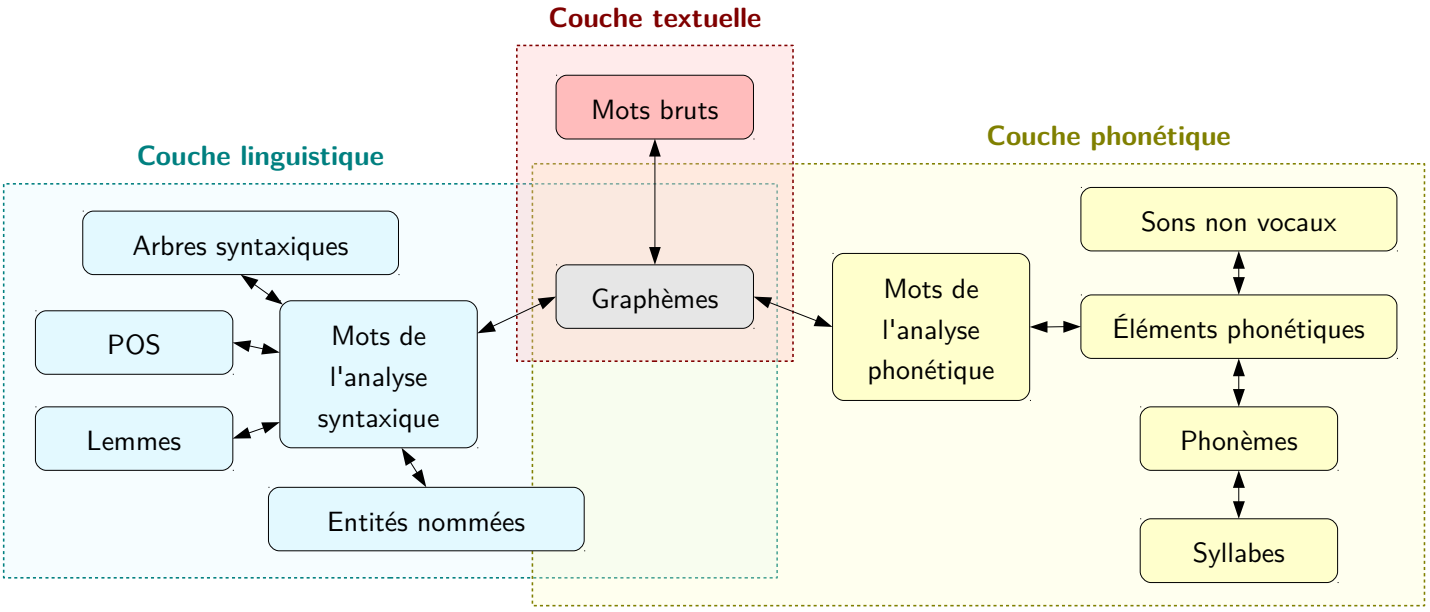


FIGURE 4 – Vue schématique des relations entre séquences.

volumes de données (Jacobs, 2009), de l'ordre du pétaoctet, ne sont volontairememnt pas pris en compte, ni dans cet article ni dans ROOTS plus généralement.

4 Exemples d'applications

ROOTS a été développé pour créer et traiter facilement et efficacement de grands corpus recensant des informations nombreuses, hétérogènes et issues d'outils divers. Cette section démontre ces capacités sur un exemple de construction d'un tel corpus, puis sur une tâche de calcul de statistiques à partir d'un corpus créé *via* ROOTS. Les expériences ont été réalisées sur une machine équipée de 64 cœurs Intel Xeon cadencés à 2 GHz et de 120 Go de RAM.

4.1 Construction d'un corpus annoté de livres numériques

Nous avons conçu un processus non supervisé qui, à partir d'un texte brut en français, produit et relie les informations suivantes : graphèmes, POS, lemmes, entités nommées, arbres syntaxiques, phonèmes, sons non vocaux et syllabes. En utilisant les concepts de ROOTS et comme illustré par la figure 4, ce processus consiste à compléter une couche d'informations textuelles rudimentaires, composées des seules séquences de mots bruts et des graphèmes correspondants, avec une couche linguistique et une couche phonologique. Ces deux dernières couches sont produites respectivement à partir des résultats de deux outils spécialisés : une suite logicielle pour le TAL de la société Synapse Développement et un service web de phonétisation, de la société Voxygen. Comme souvent, chaque outil s'appuie sur un découpage en mots qui lui est propre [2]. Aussi, trois séquences différentes de mots sont stockées (mots bruts, issus de l'analyse linguistique ou issus du phonétiseur) et mises en relation en utilisant les graphèmes comme pivot.

Les sorties des outils d'annotations sont converties en objets ROOTS à l'aide de scripts Perl écrits pour l'occasion (environ 2 000 lignes de code).

Ce processus d'annotation a été appliqué sur 1 300 livres numériques, libres et en français,

2. Par exemple, l'un peut conserver les acronymes en un mot alors que l'autre les explose en séquences d'initiales.

Type	Nombre (millions)
Mots (bruts)	94
Graphèmes	534
Mots (phonologiques)	112
Syllabes	134
Sons non vocaux	21
Phonèmes	309

Type	Nombre (millions)
Mots (linguistiques)	111
POS	111
Lemmes	111
Entités nommées	4
Arbres syntaxiques	6

TABLE 1 – Nombre d'items automatiquement générés dans le corpus de livres numériques.

Syllabe	Nombre	Syllabe	Nombre	Syllabe	Nombre
/mɑ̃/	1 098 968	/sjɔ̃/	649 986	/rɛ/	417 144
/te/	1 041 858	/se/	495 601	/ve/	338 582
/tɛ/	885 349	/ne/	430 312	/re/	303 055
/vɛ/	822 038	/si/	418 720	/tɑ̃/	288 008

TABLE 2 – Liste des 12 syllabes les plus fréquentes en fin de mot dans le corpus de livres.

collectés depuis `http://www.ebooksgratuits.com`. Le texte brut de ces livres a ainsi été transformé en un corpus ROOTS dont les statistiques sont présentées dans la table 1. Ce corpus est réparti en 35 388 fichiers JSON qui occupent un total de 220 Go (environ 27 fichiers et 173 Mo par livre). Cette taille pourrait être largement réduite par compression puisque JSON est un format très verbeux. Le temps d'exécution du processus d'annotation a été de quinze heures et est principalement dû aux appels aux outils externes comme le service en ligne de phonétisation.

4.2 Extraction de statistiques

Cette section démontre maintenant la simplicité d'utilisation de l'API de ROOTS. Pour cela, et afin d'illuster au mieux les fonctionnalités élémentaires, nous supposons vouloir répertorier et compter les syllabes en fin de mot dans un corpus, excepté pour les mots mono-syllabiques.

Cette tâche est réalisée par un court script dont le code complet est donné par l'algorithme 1. Il est composé de deux parties : la fonction `compter_dernieres_syllabes` (lignes 3-16) et le code principal (l. 18-27). Le code principal déclare une table associative pour stocker le nombre d'occurrences des syllabes (l. 18), puis charge un corpus ROOTS à partir d'un fichier JSON (l. 19). Le corpus est ensuite parcouru, phrase par phrase, de la manière suivante : une copie de chaque phrase est récupérée (l. 21), la table est mise à jour par la fonction `compter_dernieres_syllabes` (l. 22), puis la copie est détruite pour libérer la mémoire (l. 23). Enfin, toutes les syllabes rencontrées et leur nombre d'occurrences respectif sont affichés (l. 25-27). La fonction `compter_dernieres_syllabes` prend en entrée la référence vers la table des comptes et une référence vers un objet *utterance*, contenant une phrase (l. 4-5). La séquence de mots de cette phrase est récupérée (l. 6) et parcourue (l. 7). Pour chaque mot, la liste des syllabes correspondantes est calculée (l. 8). Après avoir vérifié que le mot ne contient pas qu'une seule syllabe (l. 9), la dernière syllabe est lue et convertie en une chaîne de caractères (l. 10), puis son nombre d'occurrences est soit initialisé soit incrémenté (l. 11-13).

Ce script a été appliqué en parallèle sur chaque livre numérique du corpus décrit à la section 4.1.

```perl
use roots;

sub compter_dernieres_syllabes {
    my $p_nb_occurrences = shift;
    my $phrase = shift;
    my $seq_mots = $utterance -> get_sequence("Mots");
    foreach my $mot (@{$seq_mots -> get_all_items()}) {
        my @syllabes = @{$mot -> get_related_items("Syllabes")};
        if($#syllabes > 0) {
            my $derniere_syllabe = pop(@syllabes) -> to_string();
            if(!exists($$p_nb_occurrences{$derniere_syllabe}))
                { $$p_nb_occurrences{$derniere_syllabe} = 1; }
            else { $$p_nb_occurrences{$derniere_syllabe} += 1; }
        }
    }
}

my %nb_occurrences;
my $corpus = new roots::Corpus("MonFichierRoots.json");
for (my $i = 0; $i < $corpus -> count_utterances(); $i++) {
    my $phrase = $corpus -> get_utterance($i);
    compter_dernieres_syllabes(\%nb_occurrences, $phrase);
    $phrase -> destroy();
}
while (my ($syllabe, $nombre) = each(%nb_occurrences)) {
    print $syllabe."\t".$nombre."\n";
}
```

Algorithme 1 – Script Perl utilisant l'API de ROOTS pour compter les syllabes.

L'ensemble a duré 20 min, soit environ 1 min par livre. En complément, l'utilisation moyenne de la mémoire centrale a été d'environ 300 Mo par livre. Le résultat des 12 syllabes les plus fréquentes est donné par la table 2. Il ressort de son analyse rapide que la plupart des mots terminent par /mɑ̃/, suffixe typique des adverbes, ou par /e/, /ɛ/ ou /ɑ̃/, marques des flexions verbales du participe passé, de l'imparfait et du gérondif. Ainsi, ces résultats sont cohérents avec le style narratif du corpus.

5 Conclusion et développements futurs

Dans cet article, nous avons présenté l'outil ROOTS, qui permet de générer, gérer et traiter facilement, efficacement et avec cohérence de grands corpus de séquences annotées, principalement à partir de données orales ou écrites. Après avoir présenté l'outil et l'avoir comparé à l'existant, ses performances ont été démontrées sur deux exemples applicatifs. ROOTS est entièrement opérationnel, libre et facile à étendre. Pour ces multiples raisons, nous pensons qu'il devrait être profitable à la communauté.

Dans les versions futures, une interface graphique permettant une édition manuelle d'annotations et de nouvelles fonctionnalités d'import/export seront ajoutées. Il est également envisagé d'étendre la notion de séquence à celle de treillis, utile pour modéliser l'incertitude de certaines informations. Enfin, les corpus générés avec ROOTS à partir de données libres seront prochainement mis à disposition, notamment le corpus de livres présenté dans nos expériences.

Références

Barbot, N., Barreaud, V., Boëffard, O., Charonnat, L., Delhay, A., Le Maguer, S. et Lolive, D. (2011). Towards a versatile multi-layered description of speech corpora using algebraic relations. *In Proceedings of Interspeech*, pages 1501–1504.

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

Black, A. W., Taylor, P., Caley, R. et Clark, R. (2002). The Festival speech synthesis system. Rapport technique, University of Edinburgh.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Boëffard, O., Charonnat, L., Le Maguer, S., Lolive, D. et Vidal, G. (2012). Towards fully automatic annotation of audiobooks for TTS. *In Proceedings of LREC*, pages 975–980.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M. et Beaver, D. (2010). The NXT-format Switchboard corpus : a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Carletta, J., Evert, S., Heid, U. et Kilgour, J. (2005). The NITE XML toolkit : Data model and query language. *Language Resources and Evaluation*, 39(4):313–334.

Cunningham, H.and Maynard, D. B. K. T. V. (2002). GATE : an architecture for development of robust HLT applications. Proceedings of the Annual Meeting of the ACL, pages 168–175.

Ferrucci, D. et Lally, A. (2004). UIMA : an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E. W., Hampp, T., Doganata, Y. *et al.* (2006). Towards an interoperability standard for text and multi-modal analytics. *IBM Research Report*.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52(8):36–44.

Sjölander, K. et Beskow, J. (2000). Wavesurfer - an open source speech tool. *In Proceedings of Interspeech*, pages 464–467.

Taylor, P., Black, A. W. et Caley, R. (2001). Heterogeneous relation graphs as a formalism for representing linguistic information. *Speech Communication*, 33(1-2):153–174.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. et Woodland, P. C. (2009). *The HTK Book, version 3.4*. Cambridge University Engineering Department.

Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique

Irina Illina[1], Dominique Fohr[1], Georges Linarès[2]

(1) Equipe parole, LORIA-INRIA, 54602 Villers-les-Nancy, France
(2) LIA, Université d'Avignon, 84911 Avignon, France

RESUME

La reconnaissance de noms propres est une tâche difficile dans le domaine de la recherche d'information dans de grandes bases de données audio/vidéo. Les noms propres sont souvent indispensables pour comprendre l'information contenue dans un document. Notre travail se concentre sur l'augmentation du vocabulaire d'un système de transcription automatique de la parole. L'idée est de récupérer automatiquement des noms propres à partir de documents diachroniques. Nous avons proposé des méthodes qui augmentent de façon dynamique le vocabulaire du système de reconnaissance en utilisant des informations lexicales et temporelles. Nous faisons l'hypothèse que les mêmes noms propres apparaissent fréquemment dans des documents relatifs à la même période temporelle. Nous avons étudié une méthode fondée sur l'information mutuelle et nous avons proposé une nouvelle méthode utilisant la similarité cosinus. Dans cette nouvelle méthode, le contexte d'un nom propre est représenté par un modèle vectoriel (sac de mots). Nous avons également étudié différents paramètres de sélection de noms propres afin de limiter l'augmentation du vocabulaire et donc l'impact sur les performances d'un système de reconnaissance. Les résultats de reconnaissance pour la transcription de bulletins d'informations radiophoniques montrent une réduction significative du taux d'erreur de mots en utilisant un vocabulaire augmenté.

ABSTRACT

Proper names are usually keys to understand the information contained in a document. Our work focuses on increasing the vocabulary size of a speech transcription system by automatically retrieving proper names from contemporary diachronic text documents. We proposed methods that dynamically augment the automatic speech recognition system vocabulary, using lexical and temporal features. We assume that the same proper names frequently appear in documents relating to the same time period. We studied a method based on Mutual Information and we proposed a new method based on cosine similarity to retrieve new proper names. In this new method, proper name context is represented by vector space model (Bag of Words). We also studied different metrics for proper name selection in order to limit the vocabulary augmentation and therefore the impact on the recognition performance. Recognition results show a significant reduction of the word error rate using augmented vocabulary with retrieved proper names.

MOTS-CLES : reconnaissance de la parole, mots hors vocabulaire, noms propres, augmentation du vocabulaire.

KEYWORDS: speech recognition, out-of-vocabulary words, proper names, vocabulary augmentation.

1 Introduction

Dans notre travail, nous nous intéressons à la reconnaissance automatique de la parole (RAP) et plus particulièrement à la transcription de documents audio. Même en utilisant un très grand vocabulaire, les systèmes RAP sont confrontés au problème des *mots hors vocabulaire* (*Out Of Vocabulary*, OOV). Ces mots OOV sont des mots qui se trouvent dans le signal de parole, mais pas dans le vocabulaire du système RAP. Dans ce cas, le système RAP ne pourra pas les transcrire correctement et les remplacera par un ou plusieurs mots du vocabulaire, impactant négativement l'intelligibilité de la transcription.

La reconnaissance de *noms propres* (NP) est une tâche complexe, car les noms propres sont en constante évolution et aucun vocabulaire statique ne pourra contenir tous les noms propres existants : par exemple, les NP représentent environ 10 % des mots des articles de journaux en anglais ou en français et ils sont vitaux pour caractériser le contenu d'un texte (Friburger, 2002). Bechet et Yvon (Bechet, 2000) ont montré que 72% des mots OOV sont des NP dans le cas d'un vocabulaire de 265K mots.

Notre travail utilise la modélisation du contexte temporel afin de capturer l'information lexicale de manière à récupérer des noms propres OOV et augmenter la taille du vocabulaire du système de reconnaissance. Nous nous concentrons sur l'exploitation du contexte lexical grâce à des informations temporelles de documents *diachroniques* (documents qui évoluent dans le temps) (Allauzen, 2005). Notre hypothèse est que l'information temporelle est un élément important pour capturer des dépendances NP-contexte (Kobayashi, 1998). Nous supposons que les NP sont souvent liés à un événement qui se produit dans une période de temps spécifique dans des documents diachroniques. Nous émettons l'hypothèse que les NP évoluent dans le temps, et que pour une date donnée, les mêmes NP vont apparaître dans des documents qui appartiennent à la même période. Les contextes temporels ont été proposés auparavant par Federico et Bertoldi (Bertoldi, 2001) pour augmenter le vocabulaire, et par (Prada, 2010) pour la prédiction des OOV dans les sorties du système de reconnaissance. Contrairement à ces travaux, notre travail étend le vocabulaire du RAP en utilisant des périodes temporelles plus courtes et plus précises pour éviter l'augmentation excessive du vocabulaire. Nous recherchons un bon compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire. Dans le cas contraire, cela pourrait conduire à augmenter considérablement les ressources requises pour un système de reconnaissance.

Nous utilisons également la notion du contexte pour extraire de nouveaux noms propres (Bigot, 2013) et (Oger, 2008). Cependant, nos approches se concentrent sur l'exploitation de la temporalité des documents en utilisant des documents diachroniques.

Les expériences sont réalisées dans le cadre de la transcription de bulletins d'informations radiophoniques et le corpus diachronique utilisé est un corpus d'articles d'agence de presse.

Cet article est organisé de la façon suivante : la section suivante donne la méthodologie proposée pour la sélection des nouveaux NP à partir des documents diachroniques. Les sections 3 et 4 décrivent les expériences et les résultats.

2 Méthodologie

Notre idée consiste à extraire des noms propres OOV automatiquement à partir des documents diachroniques, en utilisant le contexte lexical et temporel. Nos méthodes d'extraction des noms OOV sont fondées sur l'idée que les noms propres manquants se retrouvent probablement dans des documents contemporains, c'est-à-dire correspondant à

la même période de temps que le document que nous voulons transcrire. Nous émettons l'hypothèse que les NP évoluent dans le temps, et que pour une date donnée, les mêmes noms propres apparaitront dans les documents qui appartiennent à la même période. Par exemple, dans un document de mars 2011 contenant les NP *Japon* et *Fukushima*, il y a de fortes chances que les NP *TEPCO*, *Daiichi* et *Naoto Kan* apparaissent.

Nous proposons d'utiliser les documents du corpus diachronique qui sont contemporains de chaque document de test, et de construire un vocabulaire augmenté pour chaque document audio de test. En résumé, nous avons un document audio de test (à transcrire) qui contient des mots OOV, et nous avons un corpus de textes diachroniques, utilisé pour rechercher de nouveaux NP. Un vocabulaire augmenté est construit pour chaque document de test.

Nous supposons que, pour une date donnée, un nom propre du corpus de test co-occurrera avec d'autres NP des documents diachroniques correspondant à la même période de temps. Parmi ces NP, nous faisons l'hypothèse qu'un certain nombre seront présents dans le document de test, c'est-à-dire seront des NP OOV. L'idée est d'exploiter la relation entre les NP pour un enrichissement performant du vocabulaire.

Dans cet article, différentes stratégies de sélection de NP seront proposées pour construire ce vocabulaire augmenté :

- Méthode de référence : Sélection des documents diachronique en utilisant uniquement une période de temps correspondant au document de test.

- Méthode fondée sur l'information mutuelle : même stratégie que la méthode de référence, mais l'information mutuelle est utilisée pour mieux choisir les noms propres OOV.

- Méthode fondée sur la similarité cosinus : même stratégie que la méthode de référence mais les documents sont représentés par le modèle vectoriel (Shingal, 2011).

Dans une étude précédente (Nkairi, 2013), nous avons présenté les résultats des méthodes de référence et celle fondée sur l'information mutuelle. Ici, une nouvelle méthode, la méthode de similarité cosinus, est proposée et comparée à la méthode fondée sur l'information mutuelle.

2.1 Méthode de référence

Cette méthode consiste à extraire une liste de tous les noms propres OOV qui apparaissent dans les documents du corpus diachronique, qui correspondent à la même période que le document de test. Ensuite, notre vocabulaire est augmenté avec tous les NP extraits de ces documents. Nous considérons cette méthode comme méthode de référence. Le problème de cette approche est que si le corpus diachronique est de grande taille, nous risquons d'augmenter la taille du vocabulaire de façon démesurée.

2.2 Méthode fondée sur l'information mutuelle

Pour obtenir un meilleur compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire, nous allons filtrer les NP à l'aide de l'information mutuelle (MI) :

A) *Extraction des NP de chaque document de test* : Pour chaque document de test, nous extrayons la liste des NP. L'objectif est d'utiliser ceux-ci comme points d'ancrage pour chercher de nouveaux noms propres dans le corpus diachronique.

B) *Extraction du contexte à partir des documents diachroniques* : Après extraction de la liste des noms propres du document de test, nous pouvons commencer à extraire leurs «contextes» dans le corpus diachronique. Seuls les documents qui correspondent à la même

période de temps que le document de test sont pris en compte. A partir de l'étiquetage morphosyntaxique de ces documents diachroniques (*TreeTagger*), nous extrayons les mots qui ont été étiquetés comme « nom propre ».

Afin de limiter l'augmentation excessive du vocabulaire, nous proposons d'utiliser l'information mutuelle. Nous calculons cette information entre les NP du document de test (qui appartiennent au vocabulaire du système de reconnaissance) et les NP des documents contemporains du corpus diachronique. Si deux NP ont une information mutuelle élevée, cela augmente la probabilité qu'ils apparaissent tous les deux dans le document de test.

Dans la théorie des probabilités et la théorie de l'information, l'information mutuelle de deux variables aléatoires est une grandeur qui mesure la dépendance mutuelle des deux variables aléatoires. Formellement, l'information mutuelle de deux variables aléatoires discrètes X et Y est défini comme :

$$I(X;Y) = \sum_{x,y} p(X=x, Y=y)\, log\left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)}\right)$$

Plus la probabilité de la cooccurrence de deux noms propres dans le corpus diachronique est grande, plus la probabilité de leur cooccurrence dans un document de test est élevée.

Enfin, on calcule l'information mutuelle entre toutes les combinaisons de la variable X (X est un NP du vocabulaire apparaissant dans le document de test) et la variable Y (Y est un NP OOV extrait des documents contemporains diachroniques).

C) *Augmentation du vocabulaire* : Les nouveaux NP dont l'information mutuelle est supérieure à un seuil sont retenus. Afin de sélectionner les NP les plus pertinents, nous éliminons ceux qui ont une faible fréquence d'apparition. Les prononciations de ces NP sont générées en utilisant un dictionnaire phonétique ou un outil graphème-phonème (Illina, 2011). En utilisant cette méthodologie, nous nous attendons à extraire une liste réduite (par rapport à la méthode de référence) de tous les NP potentiellement manquants.

2.3 Méthode fondée sur la similarité cosinus

Par rapport à la méthode MI, seule l'étape B est modifiée. Nous représentons chaque document du corpus diachronique comme un *vecteur de mots* (sac de mots, BOW). Seuls les mots significatifs sont conservés : les verbes, les adjectifs, les noms et les NP. Pour chaque NP OOV présent dans les documents diachroniques contemporains du document de test, un *vecteur NP* est calculé comme étant la somme de *vecteurs de mots* des documents dans lesquels ce NP apparait. Le document de test est représenté par son *vecteur de mots*. Puis, la similarité cosinus entre ce vecteur de mots et chaque *vecteur NP* est calculée. Les NP dont la similarité cosinus est supérieure à un seuil sont sélectionnés et ajoutés au vocabulaire.

3 Expériences

3.1 Corpus de test

Doc1	Doc2	Doc3	Doc4	Doc5
2007/12/20	2007/12/21	2008/01/17	2008/01/18	2008/01/24

TABLE 1 – Date des documents de test

Pour valider la méthodologie proposée, nous avons utilisé comme corpus de test 5 documents audio du corpus ESTER2 (voir la Table 1). L'objectif de cette campagne était d'évaluer la transcription automatique d'émissions de radio en français (Galliano, 2009). La

Table 2 présente les occurrences de tous les NP (appartenant au vocabulaire et OOV) dans chaque document de test par rapport au vocabulaire de 97k mots de notre système de RAP. Pour augmenter le taux d'OOV, nous avons choisi au hasard 75 NP, qui apparaissent dans les corpus de test et nous les avons retirés de notre vocabulaire 97k mots. Ce vocabulaire réduit sera appelé par la suite *vocabulaire standard*. Cela conduit à un taux de NP OOV d'environ 1% (404/38525). Au total, le corpus de test contient 148 NP OOV différents.

	Nombre de mots différents	**Nombre d'occurrences**	**NP appartenant au vocabulaire**	**NP OOV**	**Occ. des NP OOV**
Doc1	1350	4099	86	44	93
Doc2	1446	4604	89	39	70
Doc3	1958	11803	43	25	63
Doc4	2107	10152	90	39	71
Doc5	1432	7867	48	27	107
All	-	38525	-	-	404

TABLE 2 – Couverture des noms propres du corpus de test.

Nous extrayons les NP de la transcription automatique générée par notre système ANTS (*Automatic News Transcription System*). Pour cela, le vocabulaire utilisé pour la reconnaissance a été étiqueté en termes de NP et non-NP. Comme nous utilisons des documents diachroniques, nous construisons un vocabulaire spécifique pour chaque document de test en fonction de la période choisie.

Les résultats sont présentés en termes de rappel (%): nombre de NP OOV retrouvés divisé par le nombre total de NP OOV du document de test. Pour les expériences de reconnaissance, le taux d'erreur de mots (*Word Error Rate,* WER) est calculé.

3.2 Corpus diachronique

Comme corpus diachronique, nous avons utilisé le corpus *GigaWord* : Agence France Presse (AFP) et *Associated Press Worldstream* (APW). Le corpus français *GigaWord* est une archive d'articles d'agences de presse : pour l'AFP de mai 1994 à décembre 2008, pour l'APW de novembre 1994 à décembre 2008. Le choix de *GigaWord* a été motivé par le fait qu'il est *contemporain* du corpus de test Ester, qu'il est rédigé dans le *même style* (journalistique) et qu'il traite de domaines similaire (politique, sports, etc.). De plus, sa granularité temporelle est fine, journalière. En revanche, *GigaWord* est centré sur l'information instantanée.

3.3 Système de transcription

Le système ANTS (Illina, 2004) utilisé pour ces expériences est fondé sur des modèles HMM dépendants du contexte appris sur un corpus audio de 200 heures d'émissions de radio. Le moteur de reconnaissance est Julius (Lee, 2009). Le modèle de langage est estimé en utilisant la boîte à outils SRILM (Stolcke, 2002) sur des corpus de textes d'environ 1.8 milliard de mots (*GigaWord* inclus). Le modèle de langage est ré-estimé pour chaque vocabulaire augmenté. Le lexique phonétique de base contient 218K prononciations pour les 97K mots.

4 Résultats expérimentaux

4.1 Méthode de référence

En utilisant l'outil d'étiquetage morphosyntaxique *TreeTagger*, nous avons extrait 40982 NP différents à partir de 2 mois du corpus diachronique. Parmi ces 40982 NP, 23210 ne sont pas dans notre vocabulaire. Parmi ces 23210, seulement 102 NP sont présents dans le corpus de

test (OOV). Cela montre qu'il est nécessaire de filtrer la liste des NP pour avoir un meilleur compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire. Afin d'étudier si la période temporelle joue un rôle important, nous avons étudié trois intervalles de temps dans les documents diachroniques : le même jour que le document de test (1 jour); 3 jours avant et 3 jours après la date du document de test (1 semaine); le mois courant du document de test (1 mois).

Période temporelle	Nombre moyen de NP sélectionnés par fichier de test	Nombre moyen de NP OOV retrouvés par fichier de test	Rappel (%)
1 jour	925	16	44.0
1 semaine	4305	21	58.6
1 mois	13069	**24**	**67.6**

TABLE 3 –Résultats pour la méthode de référence.

Nous appelons ***NP sélectionnés*** les noms propres que nous avons récupérés à partir des documents diachroniques en utilisant notre méthode et qui ne sont pas dans notre vocabulaire.

Nous appelons ***NP OOV retrouvés*** les noms propres OOV de la liste *NP sélectionnés* qui sont présents dans les documents de test.

La Table 3 montre que l'utilisation des documents diachroniques d'une journée nous permet de sélectionner en moyenne 925 nouveaux NP par fichier de test. Parmi ces NP, en moyenne, 16 NP sont présents dans un fichier de test (OOV). Cela correspond à un rappel de 44%.

Nous pouvons remarquer que limiter l'intervalle de temps des documents diachroniques utilisés à une semaine réduit l'ensemble de candidats NP à 4305 (Table 3), tout en récupérant 58.6% des OOV manquants. Ce résultat confirme l'idée que l'utilisation de l'information temporelle peut aider à réduire la liste des nouveaux candidats NP pour l'enrichissement du vocabulaire tout en conservant un bon taux de rappel.

4.2 Méthode fondée sur l'information mutuelle

La Table 4 montre les résultats de la méthode fondée sur l'information mutuelle en utilisant différentes périodes de temps et différentes valeurs de seuil.

Période temporelle	Seuil	Nb moyen de NP sélectionnés par fichier de test	Nb moyen de NP OOV retrouvés par fichier de test	Rappel (%)
1 jour	0.05	10.4	4.8	13.8
	0.01	294.6	12.8	36.8
	0.005	421.0	14.2	40.8
	0.001	531.6	15.4	44.2
1 semaine	0.05	3.6	2.8	8.1
	0.01	50.6	8.8	25.3
	0.005	229.0	12.0	34.5
	0.001	1749.4	**19.2**	**55.2**
1 mois (occ>2)	0.05	2.6	1.6	4.6
	0.01	18.8	6.8	19.5
	0.005	41.4	9.0	25.8
	0.001	221.8	14.0	40.2

TABLE 4 – Résultats pour la méthode fondée sur l'information mutuelle.

Comme nous construisons un vocabulaire augmenté pour chaque fichier de test, les résultats présentés dans la Table 4 sont donnés en termes de moyenne des valeurs calculées sur les 5

fichiers de test. Pour les résultats correspondant à une période d'un mois, nous avons introduit un seuil de fréquence (*occ* dans la Table 4) afin de limiter les efforts de calcul. Utiliser uniquement les documents diachroniques d'un seul jour s'avère être insuffisant pour extraire un grand nombre de NP OOV. Cela peut être dû au fait que nous sommes dans le cadre des agences de presse et qu'un nouvel événement fait l'objet d'articles pendant plusieurs jours. Avec le seuil de 0,05 très peu de NP sont sélectionnés, mais une grande partie de ces NP sont des NP OOV. Par exemple, pour une période d'une semaine, parmi les 4 NP sélectionnés, 3 sont de NP OOV. Le meilleur rappel est obtenu en utilisant une période de temps d'une semaine (55,2%).

4.3 Méthode fondée sur la similarité cosinus

Les résultats pour la méthode fondée sur la similarité cosinus sont présentés dans la Table 5. Afin de réduire encore le nombre de NP sélectionnés, nous ne gardons que les NP ayant un nombre d'occurrences supérieur à un seuil dépendant de la période temporelle.

Comme dans le cas de la méthode MI, ne considérer qu'une seule journée pour récupérer des NP semble insuffisant. Le meilleur compromis entre le rappel et le nombre de NP sélectionnés est obtenue pour la période d'un mois et un seuil de 0,05 (59,8% de rappel).

Période temporelle	Seuil	Nb moyen de NP sélectionnés par fichier de test	Nb moyen de NP OOV retrouvés par fichier de test	Rappel (%)
1 jour (occ>0)	0.025	813.4	15.4	44.3
	0.05	437.6	14.4	41.4
	0.075	131.4	11.2	32.2
	0.1	51.8	8.4	24.1
1 semaine (occ>1)	0.025	1880.0	19.4	55.8
	0.05	1127.6	18.8	54.0
	0.075	431.6	17.0	48.9
	0.1	152.0	13.4	38.5
1 mois (occ>2)	0.025	3795.6	**21.4**	**61.5**
	0.05	2473.8	20.8	59.8
	0.075	1010.2	19.4	55.8
	0.1	334.4	17.0	48.9

TABLE 5 – Résultats pour la méthode fondée sur la similarité cosinus.

4.4 Résultats de reconnaissance

Pour valider les approches proposées, nous avons réalisé la transcription automatique des 5 documents de test. Pour chaque document de test et pour chaque période temporelle, un vocabulaire augmenté est généré.

	Vocabulaire standard	Vocabulaire augmenté		
		1 jour	1 semaine	1 mois
Doc 1	19.7	17.9	18.0	18.4
Doc 2	20.9	20.3	19.7	20.0
Doc 3	28.3	28.2	28.1	28.2
Doc 4	24.5	24.0	24.2	24.2
Doc 5	36.5	36.1	36.1	36.0
All	27.1	**26.6**	**26.6**	26.7

TABLE 6 – WER (%) pour la méthode MI, seuil de 0.001.

La Table 6 donne les résultats obtenus pour la méthode fondée sur l'information mutuelle. Il montre que, en moyenne, le vocabulaire augmenté en utilisant des données diachroniques réduit de manière significative le WER. Nous n'observons pas de différence significative entre les WER correspondant aux trois périodes temporelles évaluées.

Les résultats pour la méthode fondée sur la similarité cosinus sont donnés dans la Table 7. Comme pour la méthode MI, une amélioration significative est obtenue en utilisant les vocabulaires augmentés. Les résultats obtenus avec les périodes temporelles d'une semaine et d'un mois sont légèrement meilleurs que ceux d'une journée.

	Vocabulaire standard	**Vocabulaire augmenté**		
		1 jour seuil: 0.025 occ>0	**1 semaine seuil: 0.025 occ>1**	**1 mois seuil: 0.05 occ>2**
Doc 1	19.7	18.1	18.0	18.2
Doc 2	20.9	20.1	19.5	19.4
Doc 3	28.3	28.2	27.9	28.0
Doc 4	24.5	24.0	23.8	23.7
Doc 5	36.5	36.1	35.8	35.8
All	27.1	26.7	**26.4**	**26.4**

TABLE 7 – WER (%) pour la méthode cosinus.

Pour les deux méthodes, les performances dépendent du document de test. Pour les documents 1 et 2, quelle que soit la période de temps utilisée pour créer le vocabulaire augmenté, l'amélioration est significative. En revanche, l'amélioration pour le document 3 est plus faible. Enfin, les deux méthodes proposées donnent une amélioration significative par rapport au vocabulaire standard, mais la différence de WER entre les deux méthodes n'est pas significative.

5 Conclusion

Dans le cadre de la reconnaissance automatique de la parole, notre étude a porté sur le problème de l'augmentation du vocabulaire à l'aide de documents diachroniques. Nous avons étudié des méthodes qui augmentent le vocabulaire avec des noms propres en utilisant la notion de contexte et des informations temporelles. L'idée est d'utiliser les noms propres reconnus du document de test comme ancres pour rechercher de nouveaux NP dans le corpus diachronique. Notre modèle de contexte est fondé sur l'information mutuelle. Nous avons proposé une nouvelle méthode utilisant la similarité cosinus (modèle vectoriel). Les nouveaux NP sont ajoutés au vocabulaire du système.

Des expériences ont été menées sur des émissions de radio en utilisant des données textuelles d'agences de presse comme corpus diachronique. Les résultats valident l'hypothèse que la période temporelle et le contexte lexical permettent de récupérer des noms propres manquants et d'obtenir une réduction significative du taux d'erreur de mots en ajoutant ces NP dans le vocabulaire. Une perspective intéressante pourrait être d'exploiter des informations "sémantiques" contenues dans le document de test : quand une date précise est identifiée, les documents diachroniques temporellement proches de cette date pourraient être utilisés pour extraire de nouveaux noms propres.

Remerciements

Les auteurs tiennent à remercier l'ANR *ContNomina* SIMI-2 de l'Agence Nationale de la Recherche française (ANR) pour son soutien.

Références

Allauzen, A. and Gauvain, J.-L. (2005). Diachronic vocabulary adaptation for broadcast news transcription, *Proc. of Interspeech*.

Bechet, F. and Yvon, F. (2000). Les Noms Propres en Traitement Automatique de la Parole, *Revue Traitement Automatique des Langues*, vol. 41, num. 3 pp. 672-708.

Bertoldi, N. and Federico, M. (2001). Lexicon adaptation for broadcast news transcription*, In Adaptation-2001*, pp. 187- 190.

Bigot, B., Senay, G., Linares, G., Fredouille, C., and Dufour, R. (2013). Person name recognition in ASR outputs using continuous context models, *Proc. of ICASSP*.

Friburger, N. and Maurel, D. (2002). Textual Similarity Based on Proper Names, *Proc. of the workshop Mathematical/Formal Methods in Information Retrieval*, pp. 155-167.

Galliano, S., Gravier, G., Chaubard, L. (2009). The ESTER2 Evaluation Campaign for the Rich Transcription of French Radio Broadcast, *Proc. of Interspeech*.

Illina I., Fohr D., Mella O., Cerisara C. (2004). The Automatic News Transcription System: ANTS, some Real Time experiments, *Proc. ICSLP*.

Illina I., Fohr D., Jouvet D. (2011). Grapheme-to-Phoneme Conversion using Conditional Random Fields, *Proc. Interspeech*

Kobayashi, A., Onoe, K., Imai, T., Ando, A. (1998). Time dependent language model for broadcast news transcription and its post-correction, *Proc ICSLP*

Lee, A. and Kawahara, T. (2009). Recent Development of Open-Source Speech Recognition Engine Julius, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).*

Nkairi, I., Illina, I., Linarès, G., Fohr, D. (2013). Exploring temporal context in diachronic text documents for automatic OOV proper name retrieval, *Proc. of LTC*.

Oger, S., Linarès, G. and Béchet, F. (2008). Local methods for on-demand out-of-vocabulary word retrieval, *Proc. of the Language Resources and Evaluation Conference (LREC)*.

Parada, C., Dredze, M., Filimonov, F. and Jelinek, F. (2010). Contextual Information Improves OOV Detection in Speech, *Proc. of NAACL*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees*, Proc. of ICNMLP*.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit, *Proc. of ICSLP.*

Le couple YASP/YADE
Yet Another Semantic Parser
Yet Another Dialogue Engine

Jérôme Lehuen
Équipe LST – LIUM, Université du Mans
Avenue Laënnec, 72085 Le Mans Cedex 9
Jerome.Lehuen@univ-lemans.fr

RÉSUMÉ

Cet article présente un modèle original de représentation des connaissances sémantique, un module de compréhension du langage parlé reposant sur ce modèle, et un gestionnaire de dialogue reposant également sur ce même formalisme. Ces composants, associés à un module de reconnaissance de la parole et à un module de synthèse vocale, constituent un système de dialogue oral complet. Le système de compréhension a été développé et évalué avec les données du corpus MEDIA et s'est bien positionné par rapport aux systèmes évalués lors de la campagne MEDIA. Les énoncés du dialogueur combinent restitution et génération de texte dans une optique de reformulation de parties d'énoncés de l'usager, afin d'en « négocier le sens » et potentiellement de récupérer certaines erreurs. L'approche unifiée qui consiste à faire reposer compréhension et interaction sur les mêmes connaissances rentabilise le coût de production de ces connaissances tout en assurant une cohérence sémantique entre ces deux aspects.

ABSTRACT

The YASP/YADE couple : Yet Another Semantic Parser/Yet Another Dialogue Engine

This paper presents an original model of semantic knowledge, a spoken language understanding component based on this model, and a dialogue manager also based on the same formalism. These components are combined to a speech recognition module, and to a speech synthesis module, in order to build a comprehensive oral dialogue system. The understanding component was developed and tested with the MEDIA corpus, and obtained good results. The dialogue manager combines restitution and text generation, in order to reformulate parts of the user's statements, to « negotiate the meaning » and potentially to recover errors. The unified approach, which consists in basing the understanding and the dialogue modules on the same knowledge, pays for the cost of producing this knowledge, while ensuring semantic consistency between these two aspects of a spoken dialogue system.

MOTS-CLÉS : Système de Dialogue Oral, Compréhension de la Langue Naturelle.

KEYWORDS: Spoken Dialogue System, Natural Language Understanding.

1 Introduction

La compréhension automatique de la parole – et par extension le DHM (dialogue humain-machine) – devient un enjeu d'importance, dès lors que les systèmes de RAP (reconnaissance automatique de la parole) deviennent performants. Un système de DHM orienté tâche est composé de modules qui, activés à tour de rôle, permettent la réalisation d'une tâche déterminée. Un tel système comprend typiquement un module de RAP, un module de compréhension, un module de dialogue et un module de synthèse de la parole (Jokinen et McTear, 2009).

Si les approches probabilistes de type HMM *(hidden Markov model)* sont désormais devenues incontournables en ce qui concerne la RAP, il n'en est pas de même pour la compréhension et pour le dialogue. En effet, le choix de la méthode pour passer d'une phrase à une représentation sémantique est possible : approches essentiellement symboliques à base de connaissances (lexiques, grammaires, *etc.*), approches essentiellement stochastiques à base de corpus annotés et d'apprentissage automatique, comme par exemple les CRF *(conditional random fields)* (Lafferty et al., 2001), ou encore approches mixtes, relativement peu explorées. En ce qui concerne les modèles de dialogue, les approches stochastiques, comme par exemple POMDP *(partially observable Markov decision processes)* (Young et al., 2006), sont devenues des alternatives sérieuses aux approches logiques, les modèles fondés sur la planification intentionnelle issus de (Allen et Perrault, 1980) étant complexes à mettre en œuvre et à maintenir. En revanche, les modèles fondés sur des schémas *(frame-based/slot-filling models)* continuent à se développer (Cole, 1999), (Pellom et al., 2001), (Hanna et al., 2007), notamment au travers des plateformes implémentant le standard VoiceXML (www.vxml.org), tout comme les modèles fondés sur des agents intelligents *(agent-based models)* (Seneff et al., 1998), (Pellom et al., 2001), (Jokinen, 2009), ces deux approches n'étant pas exclusives.

Cet article propose une approche essentiellement symbolique pour le DHM. Nous avons cherché en particulier à mutualiser les connaissances entre les modules de compréhension, de dialogue et de génération. La représentation autour de laquelle s'articulent ces modules est la *structure de granules*, décrite dans la deuxième partie de l'article. La troisième partie expose la mise en œuvre du modèle des granules au niveau des composants YASP *(yet another semantic parser)* et YADE *(yet another dialogue engine)*. Les règles de ce dernier recherchent des motifs dans les structures de granules, les considérant ainsi comme des schémas hiérarchiques. La tâche choisie pour notre démonstrateur est la tâche MEDIA qui porte sur la réservation de séjours dans des établissements hôteliers, tâche de référence de la campagne d'évaluation EVALDA-MEDIA (Bonneau-Maynard et al., 2006). Une première version de YASP avait été évaluée hors campagne d'évaluation, mais en respectant son protocole (Lehuen et Lemeunier, 2010). Depuis, le modèle des granules et le module YASP ont été améliorés et intégrés au sein d'une chaîne complète de dialogue oral.

2 Le modèle des granules

Le modèle des granules est une généralisation et une opérationnalisation du *schéma actanciel* proposé par (Tesnière, 1988) dans le cadre de sa syntaxe structurale. Le modèle repose sur une grammaire sémantique qui s'inspire à la fois des grammaires génératives (Chomsky, 1965), des grammaires de cas (Fillmore, 1968), et des grammaires de dépendances (Tesnière, 1988). Un granule représente toute unité porteuse de sens (acte de langage, entité nommée, adjectif, *etc.*) ayant des *verbalisations*, des *offres* et éventuellement des *attentes*, au travers de *dépendances*.

Les **offres** et les **attentes** peuvent être des classes, des catégories, des traits sémantiques, voire même des catégories grammaticales. Le modèle n'impose aucune contrainte sur le choix des offres et des attentes : c'est d'une part une expertise linguistique, d'autre part une analyse de la tâche et/ou d'un corpus de dialogues, qui permettront de les déterminer.

Une **dépendance** est caractérisée par un ensemble d'attentes (obligatoires ou optionnelles) et éventuellement par un rôle. Le principe qui dirige la construction des structures de granules étant basé sur l'offre et la demande, c'est sur un critère de correspondance entre les attentes d'une dépendance d'un granule A et les offres d'un granule B, que ce dernier peut être rattaché au granule A, et endosser éventuellement un rôle au sein de la structure.

Un **rôle** peut être un actant (agent, patient, objet, bénéficiaire, *etc.*), un circonstant (lieu, moment, façon, *etc.*) ou plus simplement un concept de la tâche. Par exemple, dans *« Un aller simple pour Marseille »*, Marseille est une *ville* mais endosse le rôle *arrivée* dans le contexte de l'énoncé.

Une **verbalisation** est une expression régulière composée de mots (symboles terminaux) et de références à des dépendances (symboles non-terminaux). En imposant des contraintes sur la place des dépendances, l'ensemble des verbalisations définit une grammaire non contextuelle, utilisable en analyse et potentiellement en génération. On peut également associer à chaque verbalisation des **méta-données** qui permettent de préciser des informations complémentaires sur la forme utilisée (morphologie, modalité, style, degré de politesse, *etc.*).

Ces connaissances sont décrites dans un formalisme XML simple et extensible (cf. Figure 1). La grammaire MEDIA comporte actuellement 239 granules, 133 dépendances et 826 verbalisations, organisés en fonction de leur domaine d'application ou de leur champ sémantique (localisation spatiale, temporelle, marqueurs de référence, objets de la tâche, *etc.*). Ces verbalisations sont actuellement renseignées à la main, mais nous envisageons d'automatiser cette tâche à l'aide d'un lemmatiseur et d'un dictionnaire de synonymes.

```
<granule concept="demander" offers="action command-tache">
        <dependency code="A1" demands="demandable"/>
        <dependency code="A2" demands="demandable" required="morpho:inf"/>
        <syntax pattern="je veux/voudrais/cherche/recherche/souhaite A1"/>
        <syntax pattern="avez/auriez vous A1" metadata="mode:inter"/>
        <syntax pattern="serait il possible de A2" metadata="mode:inter"/>
</granule>

<granule concept="objetBD:reservation" offers="objet demandable">
        <dependency code="A1" demands="nombre" role="nombre-reservation"/>
        <dependency code="A2" demands="reservable" role="objet"/>
        <dependency code="A3" demands="contrainte-sejour" role="contrainte" mult="TRUE"/>
        <syntax pattern="réservation(s)/>
        <syntax pattern="A1 réservation(s)"/>
        <syntax pattern="A1 réservation(s) pour A2"/>
        <syntax pattern="réserver" metadata="morpho:inf"/>
        <syntax pattern="réserver (pour) A2" metadata="morpho:inf"/>
</granule>

<granule concept="objetBD:chambre" offers="objet demandable reservable">
        <dependency code="A1" demands="nombre" role="nombre-chambre"/>
        <dependency code="A2" demands="contrainte-chambre" role="contrainte" mult="TRUE"/>
        <syntax pattern="chambre(s)"/>
        <syntax pattern="A1 chambre(s)"/>
</granule>
```

FIGURE 1 – Extrait de la grammaire MEDIA

La Figure 2 représente une structure de granules construite par YASP. Les granules sont de type `[concept]` ou `[concept:valeur]`. Les méta-données, indiquées par le signe + fournissent des informations morphologiques ou sur le mode utilisé. Les relations sont étiquetées par les rôles et les relations en pointillés correspondent à des rattachements hypothétiques, non guidés par la grammaire, mais uniquement par le mécanisme d'offres et d'attentes (cf. § 3.1). Cette représentation correspond à une sortie brute, c'est-à-dire sans inférence *a posteriori*. Or, les cinq personnes correspondent aux deux couples et à l'enfant, ce qui n'apparaît pas dans la structure. Le mois de la première date peut être inféré à partir de la seconde. Les dates des « deux nuits suivantes » peuvent également être inférées. A l'instar des représentations de type *Case Frame*, un des avantages d'une telle structure est de donner aux concepts des contextes expressifs, indépendants de l'ordre des mots, qui favorisent inférences et réécritures formelles.

« Bonjour, serait-il possible de réserver une chambre pour cinq personnes, deux couples et un enfant, du vingt au vingt et un février sur Marseille, puis les deux nuits suivantes sur Avignon. Merci. »

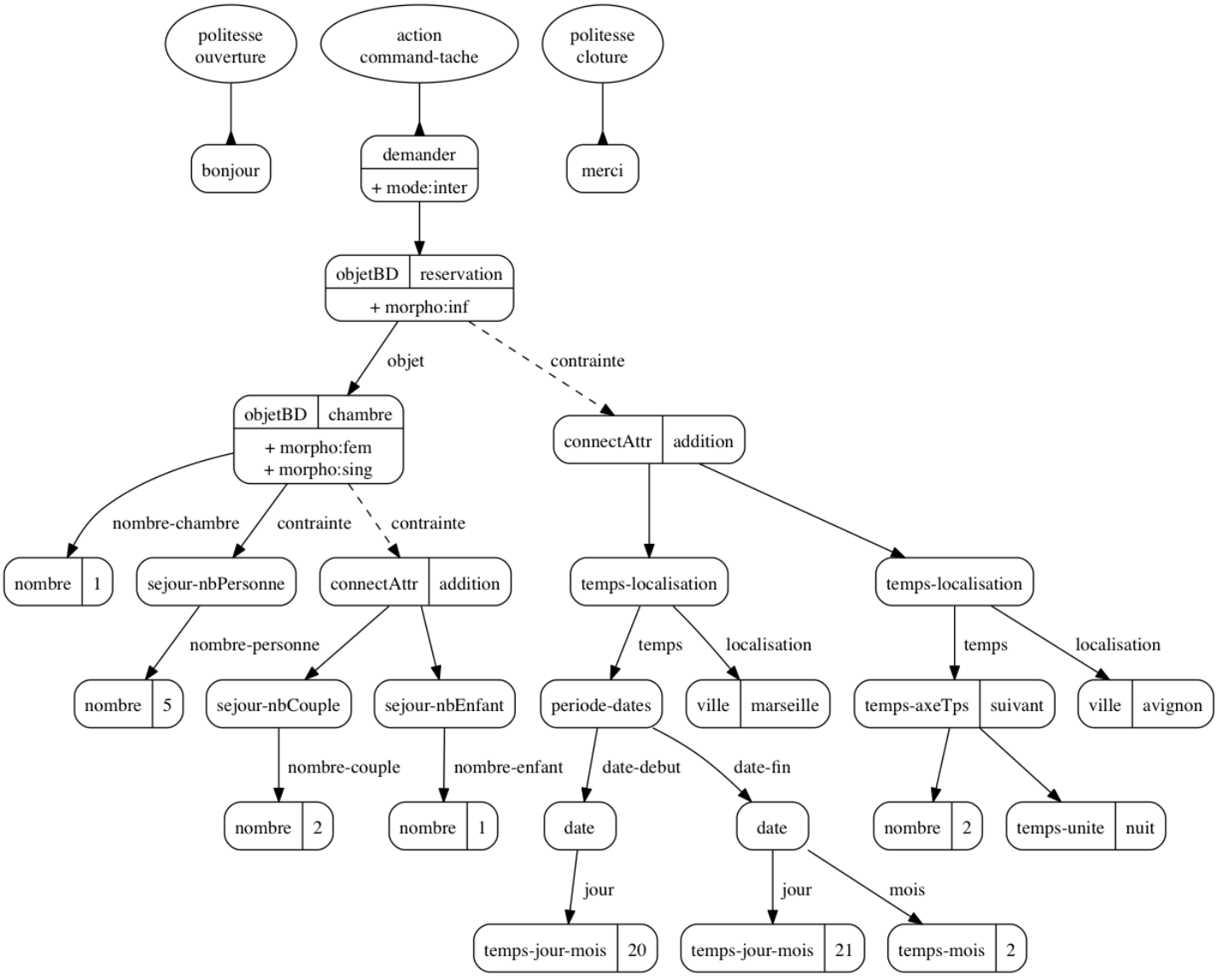


FIGURE 2 – Exemple de structure de granules produite par YASP

3 Mise en œuvre du modèle des granules

3.1 Le module de compréhension YASP

Le principe d'analyse implémenté dans YASP s'apparente à une analyse tabulaire ou *chart parsing* (Kay, 1986). L'idée est de développer et de garder en même temps plusieurs hypothèses ainsi que les conclusions partielles de l'analyse. Dans notre cas, cela consiste à développer des structures de granules de façon concurrente, en parallèle avec l'application de la grammaire, selon une stratégie *bottom-up* (à partir des mots), jusqu'à ce que plus aucune règle ne puisse être appliquée. Ce choix a été motivé, d'une part par les caractéristiques des données à traiter : énoncés oraux, souvent agrammaticaux, tout du moins dans leur globalité, comportant hésitations, insertions, redondances, et d'autre part par le modèle des granules qui intègre une grammaire hors-contexte. Les quatre principales étapes de l'analyse sont les suivantes :

1. **Instanciation** : consiste à instancier des granules à partir des mots (éléments terminaux) et d'autres granules (éléments non-terminaux) présents dans la mémoire de travail. Cette étape peut donner lieu à des conflits (recouvrements). Pour chaque granule, un score est calculé en fonction du nombre de mots pris en compte et des mots en trop (insertions).
2. **Résolution des conflits** : consiste à départager les granules conflictuels sur la base de leurs scores respectifs (certains conflits peuvent être résolus dès la première étape). Des conflits peuvent subsister, les granules en conflits sont alors indiqués dans la sortie XML pour une prise en charge ultérieure, par exemple par un gestionnaire de dialogue ;
3. **Récupération** : consiste à rattacher certains granules isolés (granules libres) sur la base des offres et des attentes. Cette étape peut également donner lieu à des conflits, qui sont résolus en optimisant une fonction d'évaluation qui prend en compte le score du granule rattaché, ainsi que son éloignement (dans l'énoncé) par rapport au granule-père ;
4. **Finalisation** : consiste à propager les rôles, à ajouter quelques autres étiquettes, à sérialiser la structure objet résultante selon un format XML (cf. Figure 3) puis à effectuer des inférences et transformations sur cette sortie (post-traitements).

La sortie XML du module YASP (cf. Figure 3) décrit une structure hiérarchique de granules instanciés et contextualisés, les attributs d'un granule instancié étant les suivants :

— `id` : identificateur unique du granule (utilisé pour les transformations) ;
— `code` : identificateur de la dépendance à laquelle le granule est rattaché (ex : `A1`) ;
— `concept` : concept du granule, éventuellement associé à une valeur (ex : `nombre:1`) ;
— `role` : rôle attribué au granule au sein de la structure (ex : `nombre-chambre`) ;
— `offers` : toutes les offres du granule, y compris celles qui ne sont pas utilisées ;
— `retained` : sous-ensemble des offres qui ont permis au granule d'être rattaché ;
— `metadata` : ensemble des méta-données associées à la verbalisation utilisée ;
— `text` : segment d'énoncé qui a été reconnu par la verbalisation ;
— `pos` : position du segment dans l'énoncé complet.

D'un point de vue technique, le module YASP est un serveur écrit en langage C qui embarque un moteur d'inférence CLIPS (Giarratano et Riley, 1989). Les étapes successives de l'analyse sont toutes décrites sous la forme de règles CLIPS. La structure de *chart* est implémentée dans la base de faits (chaque granule est un fait) et chaque verbalisation est préalablement compilée sous la

forme d'une règle de production. Les post-traitements (inférences et réécritures) sont décrits en XML sous la forme de règles dont les préconditions correspondent à des motifs à rechercher dans les structures de granules (cf. Figure 4). Ces règles sont également traduites sous la forme de règles CLIPS, à l'aide de transformations XSLT.

```
<structure text="bonjour je voudrais une chambre double">
   <granule id="3" concept="bonjour" offers="politesse ouverture" text="bonjour" pos="1"/>
   <granule id="14" concept="demander" offers="action command-tache"
            text="je voudrais une chambre double" pos="2">
      <granule id="13" code="A1" concept="objetBD:chambre" offers="objet demandable reservable"
               retained="demandable" text="une chambre double" pos="4">
         <granule id="1" code="A1" concept="nombre:1" offers="nombre quantite"
                  retained="nombre" role="nombre-chambre" text="une" pos="4"/>
         <granule id="6" code="A2" concept="chambre-type:double" offers="contrainte-chambre"
                  retained="contrainte-chambre" role="contrainte" text="double" pos="6"/>
      </granule>
   </granule>
</structure>
```

FIGURE 3 – Exemple de sortie XML issue du module YASP

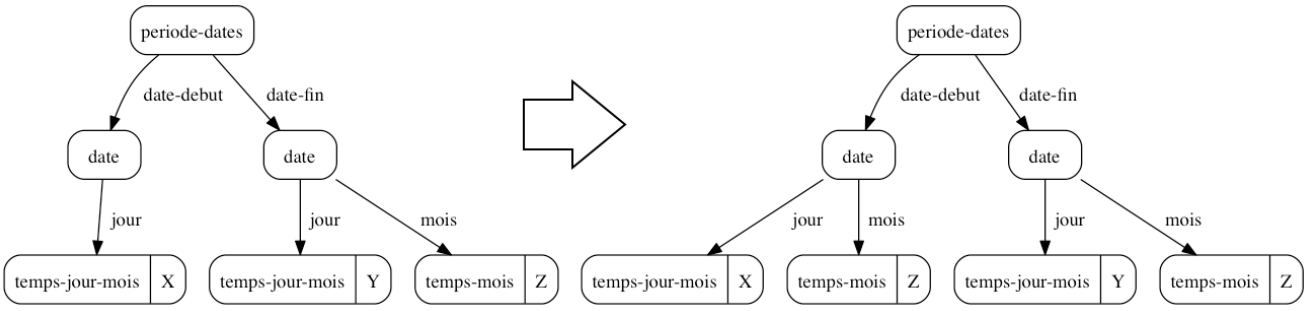


FIGURE 4 – Exemple de post-traitement sur une structure de granules

Les performances de YASP ont été évaluées avec une première version de la grammaire MEDIA (Lehuen et Lemeunier, 2010) afin de le positionner en tant que sixième système de la campagne d'évaluation EVALDA-MEDIA (Bonneau-Maynard et al., 2006). Les systèmes devaient produire des étiquettes [mode, concept, valeur] (Bonneau-Maynard et Rosset, 2003) et quatre CER *(concept error rate)* devaient être calculés, selon le nombre de modes (2 ou 4) et la précision des concepts (Relax/Full). Le système YASP, qui ne s'était inséré qu'en cinquième place en « Relax 2 modes » était remonté en troisième place en « Full 4 modes » (cf. Figure 5), sachant que les deux premiers systèmes (STO-1 et HYB-5) produisaient naturellement leurs sorties dans le formalisme cible de la campagne d'évaluation, évitant ainsi le bruit lié à la conversion de formalisme, problème également évoqué dans (Villaneau et Lamprier, 2005).

3.2 Le module de dialogue YADE

Une des fonctions d'un module de dialogue est d'intégrer les informations issues de l'interprétation sémantique dans le contexte de la tâche, mais aussi dans le contexte du dialogue, afin de maintenir une certaine logique interactionnelle (*i.e.* enchaînement logique des actes de langage), et rendre le dialogue « satisfaisant » pour un interlocuteur humain. Un module de dialogue a également intérêt à pouvoir assurer une fonction métalinguistique (ou *métadialogique*) qui lui permet de « parler de la langue et du dialogue dans le dialogue », de reformuler des parties d'énoncés de

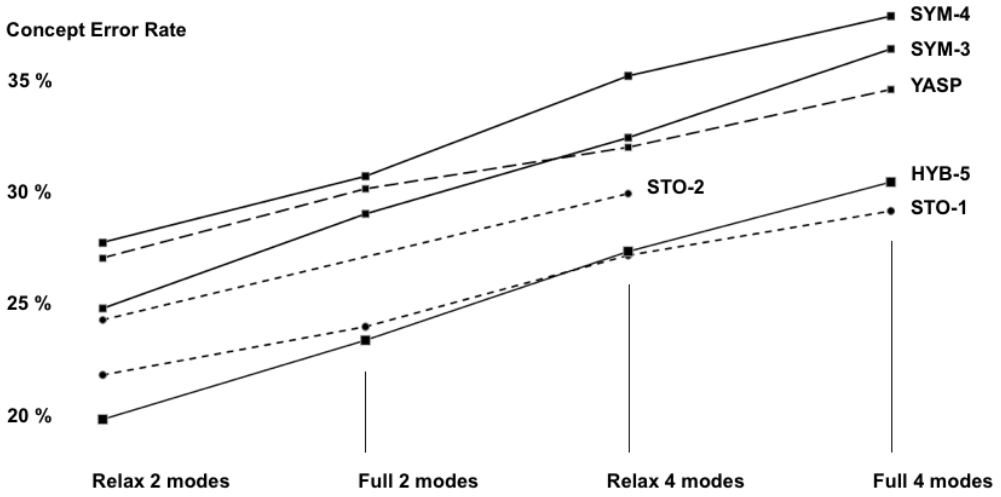


FIGURE 5 – Positionnement de YASP par rapport aux cinq systèmes de la campagne EVALDA-MEDIA soit deux systèmes STO-chastiques, deux systèmes SYM-boliques et un système HYB-ride

l'usager, d'en négocier ou d'en préciser le sens. Cette fonction est d'autant plus sensible que nous nous situons dans un contexte de dialogue oral où la grammaticalité est relative, et que notre entrée est susceptible d'être bruitée. L'exemple suivant, produit avec le couple YASP/YADE, illustre les capacités métalinguistiques de notre module de dialogue :

M1 : *Bonjour, que puis-je pour votre service ?*
H2 : *Je voudrais ...*
M3 : *J'ai compris que vous me demandez quelque chose, mais je n'ai pas tout compris. Que désirez-vous exactement ?*
H4 : *Une ...*
M5 : *Pardon, vous demandiez quelque chose. Pouvez-vous me dire ce que vous désirez ?*
H6 : *Une chambre double*
M7 : *Ah, d'accord. Vous m'avez demandé une chambre double, très bien ...*

L'énoncé H2 ayant été partiellement compris, YADE signifie à l'usager en M3 qu'il attend un complément d'information ou une reformulation. Ce faisant, il impose un contexte d'interprétation à la réponse H4 de l'usager, ce qui lui permet en M5 de reformuler ses attentes. En M7, il formule un « accusé de compréhension » afin de pouvoir continuer. Ce dernier énoncé a ceci de particulier qu'il associe restitution de phrase et génération automatique à partir de la portion de la structure de granule qui était manquante, et qui a été ajoutée après H6.

Ainsi, les structures de granules participent également une autre fonction : la génération, à condition que la grammaire ait été pensée en conséquence. Un attribut booléen `gen` a été ajouté aux verbalisations pour indiquer celles qui sont utilisables par l'algorithme de génération. Ce dernier parcourt une sous-structure à partir d'un granule, et reconstruit une séquence de mots en sélectionnant les meilleures verbalisations selon un critère de contrainte locale et/ou globale sur les méta-données. Cette fonctionnalité impose nécessairement des contraintes sur l'écriture des grammaires, puisque les expressions du type *« chambre(s) »* ou *« réserver (pour) »* ou encore *« avez/auriez »* ne peuvent pas être utilisées en génération. C'est la raison pour laquelle il s'agit d'identifier les granules susceptibles d'être regénérés par YADE, et de leur associer des verbalisations *réversibles*. Par exemple, le granule `objetBD:chambre` doit être redéfini comme dans la Figure 6, les méta-données morphologiques des différentes verbalisations autorisant des accords en genre et en nombre entre un granule-père et ses éventuels granules-fils.

```
<granule concept="objetBD:chambre" offers="objet demandable reservable">
        <dependency code="A1" demands="nombre" role="nombre-chambre"/>
        <dependency code="A2" demands="contrainte-chambre" role="contrainte" mult="TRUE"/>
        <syntax pattern="chambre" metadata="morpho:fem morpho:sing" gen="TRUE"/>
        <syntax pattern="chambres" metadata="morpho:fem morpho:plu" gen="TRUE"/>
        <syntax pattern="A1 chambre" metadata="morpho:fem morpho:sing" gen="TRUE"/>
        <syntax pattern="A1 chambres" metadata="morpho:fem morpho:plu" gen="TRUE"/>
        <syntax pattern="A1 chambre A2" metadata="morpho:fem morpho:sing" gen="TRUE"/>
        <syntax pattern="A1 chambres A2" metadata="morpho:fem morpho:plu" gen="TRUE"/>
</granule>
```

FIGURE 6 – Réécriture du granule `objetBD:chambre` pour la génération

En utilisant les structures de granules comme mémoire de travail, YADE peut être considéré comme un dialogueur à base de schémas hiérarchiques. Son comportement est décrit à l'aide de règles de type $\langle motif \rangle + \langle contexte \rangle \rightarrow \langle actions \rangle + \langle contexte \rangle$ qui recherchent des motifs particuliers dans les structures de granules, identifient des thèmes, déclenchent des actions et imposent des contraintes sur l'interprétation des réponses de l'usager et sur la suite du dialogue. Tout comme les règles de post-traitement de YASP, les règles de dialogue sont écrites dans un formalisme XML, puis traduites en CLIPS à l'aide de transformations XSLT.

4 Conclusion

La compréhension du langage naturel parlé, ainsi que son utilisation dans une application de dialogue oral, nécessitent une modélisation de la langue à la fois faiblement contraignante, afin de prendre en charge des formulations grammaticalement imprévisibles, et suffisamment formalisée, afin de fournir des représentations structurées et expressives. Pour l'heure, nous avons souhaité développer une approche originale, essentiellement symbolique, en proposant un modèle sémantique robuste, qui s'inspire à la fois des grammaires génératives au travers des verbalisations, des grammaires de cas au travers des offres et des attentes, ainsi que des grammaires de dépendances. Le *modèle des granules* intègre ainsi des connaissances sémantiques et des connaissances syntaxiques au sein d'une même structure de données. Le module de compréhension YASP, qui implémente le modèle des granules, a été évalué avec le corpus de test MEDIA, hors campagne d'évaluation, mais en respectant son protocole. Les performances de YASP ont dépassé celles des systèmes symboliques dans les modes d'évaluation « 4 modes ». Le module de dialogue YADE repose à double titre sur le modèle des granules, en exploitant les représentations issues de YASP en tant que mémoire de travail, et en utilisant des connaissances présentes dans les définitions des granules, notamment pour générer des parties des interventions du système. Ainsi, le coût lié à la production des connaissances nécessaires à la phase de compréhension est rentabilisé en ce sens qu'une partie de ces connaissances est réutilisée par le module de dialogue. Par cela, nous avons essayé de démontrer l'intérêt d'une représentation sémantique commune aux modules de compréhension et de dialogue. Le couple YASP/YADE est intégré au sein d'un système de dialogue complet, avec entrée et sortie vocales. Ces outils sont diffusés en open-source sous licence GNU-GPL, et téléchargeables sur le site du LIUM [1].

1. www-lium.univ-lemans.fr

Références

ALLEN, J. et PERRAULT, R. (1980). Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178.

BONNEAU-MAYNARD, H. et AL. (2006). Results of the french evalda-media evaluation campaign for literal understanding. *In Proceedings of LREC 2006*, pages 2054–2059, Genova, Italy.

BONNEAU-MAYNARD, H. et ROSSET, S. (2003). A semantic representation for spoken dialogs. *In Eurospeech'03*, Genève.

CHOMSKY, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

COLE, R. (1999). Tools for research and education in speech science. *In International Conference of Phonetic Sciences*, San Francisco, CA.

FILLMORE, C. (1968). The case for case. *Universals in Linguistic Theory*, 1:1–88.

GIARRATANO, J. et RILEY, G. (1989). *Expert systems, Principles and programming*. PWS, Boston.

HANNA, P. et AL. (2007). Promotion extension and reuse in a spoken dialog manager : an evaluation of the queen's communicator. *ACM Trans. Speech Lang. Process*, 4(3).

JOKINEN, K. (2009). *Constructive Dialog Management - Speech Interaction and Rational Agents*. John Wiley and Sons, Chichester, UK.

JOKINEN, K. et MCTEAR, M. (2009). *Spoken Dialogue Systems*, volume 5 de *Synthesis Lectures on Language Technologies*. Morgan and Claypool.

KAY, M. (1986). Algorithm schemata and data structures in syntactic processing. pages 35–70. Morgan Kaufmann, San Francisco.

LAFFERTY, J. et AL. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In International Conference on Machine Learning*, pages 282–289.

LEHUEN, J. et LEMEUNIER, T. (2010). A language model for human-machine dialog : The reversible semantic grammar. *In 7th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, pages 17–26, Funchal, Madeira, Portugal.

PELLOM, B. et AL. (2001). University of colorado dialog systems for travel and navigation. *In Human Language Technology Conference (HLT2001)*, San Diego, CA.

SENEFF, S. et AL. (1998). Galaxy-ii : A reference architecture for conversational system devlopment. *In 5th International Conference on Spoken Language Pocessing*, Sydney, Australia.

TESNIÈRE, L. (1988). *Éléments de syntaxe structurale*. Klincksieck, Paris.

VILLANEAU, J. et LAMPRIER, S. (2005). Corpus de dialogue homme-machine : annotation sémantique et compréhension. *In Journées Internationales de Linguistique de Corpus*, pages 221–229, Lorient, France.

YOUNG, S. et AL. (2006). Partially observable markov decision processes with continuous observations for dialogue management.

Etude préliminaire de l'influence des caractéristiques phonétiques sur le bégaiement : le cas du français et du slovaque

Ivana Laborde Didirkova[1] Fabrice Hirsch[1]
(1) Laboratoire PRAXILING UMR 5267 CNRS – Université Montpellier 3

ivana.didirkova@univ-montp3.fr, fabrice.hirsch@univ-montp3.fr

RESUME

Le bégaiement est une altération de la fluence de la parole qui touche environ 1% de la population mondiale. Si ce trouble de la communication a donné lieu à un grand nombre de travaux visant à repérer les éléments (phonologiques, lexicologiques ,…) pouvant être à l'origine d'une disfluence, il s'avère que la plupart d'entre eux ont été menés sur des locuteurs parlant l'anglais des Etats-Unis. Peu de recherches ont cherché à savoir si le bégaiement pouvait se manifester différemment d'une langue à une autre. L'objectif de notre travail consiste à étudier la fréquence d'apparition des disfluences sur des locuteurs francophones et slovacophones afin de déterminer s'il existe une influence de la structure phonologique sur l'apparition des bégayages. Sur cinq critères analysés, des variations inter-langues ont été constatées pour trois d'entre eux : nature du phonème, lieu d'articulation des voyelles et celui des consonnes.

ABSTRACT

Influence of phonetic characteristics on stuttering in French and Slovak languages: a preliminary study

Stuttering is defined as a disturbance in the normal speech fluency that affects about 1% of the world's population. Many studies are available on this communication disorder, trying to identify the elements (phonological, lexicological…) that could be at the origin of a disfluency. However, most of these studies have been conducted on the American English and only a few of researches have tried to know if manifestations of stuttering could change from one language to another. The purpose of our research is to study the frequency of disfluencies in French- and Slovak- speaking people who stutter in order to determine if the phonological structure has an influence on stuttering. Five criteria have been analyzed in this study; the cross-linguistic variations were observed for three of them: phoneme category, place of articulation for vowels and place of articulation for consonants.

MOTS-CLES : bégaiement ; disfluences ; étude inter-langues.
KEYWORDS: stuttering; disfluencies; cross-language research.

1 Introduction

Le bégaiement est un trouble du rythme de la parole qui se caractérise par la survenue fréquente de disfluences. Ce sont ces disfluences qui constituent l'objet de notre recherche. Si

la littérature (Anderson & Byrd, 2008 ; Blomgren, 2012) a montré qu'un certain nombre de phonèmes était bégayé plus que d'autres et ce, du fait notamment de leurs caractéristiques articulatoires, la majorité de ces études ne se fonde que sur une langue, à savoir l'anglais des Etats-Unis. L'objectif de notre recherche consiste à vérifier si les résultats obtenus pour l'anglais sont confirmés pour d'autres systèmes linguistiques.

1.1 Aspects linguistiques du bégaiement

Le bégaiement est défini comme une perturbation du rythme de la parole dans lequel l'individu sait exactement ce qu'il veut dire mais il est incapable de le dire en raison d'une répétition involontaire, d'une prolongation ou de l'arrêt du son (Organisation Mondiale de la Santé, 1977). En tant que tel, le bégaiement est un trouble de la communication puisqu'il apparaît essentiellement en situation d'interaction : les sujets bègues ne présentent généralement pas de disfluences lorsqu'ils parlent seuls ou à un animal.

Dans le domaine des Sciences du langage, un certain nombre de travaux a cherché à établir un lien entre ces accidents de parole et les segments propres au discours. L'une des premières études menées sur ce sujet a été réalisée par Brown en 1938. Cette recherche a montré que le bégaiement varie en fonction de plusieurs critères. En effet, Brown (1938) a par exemple relevé que la longueur d'un mot ou encore sa position dans l'énoncé sont des facteurs pouvant expliquer la survenue ou non de disfluences dans le discours. Par la suite, d'autres études, plus récentes, ont également esquissé des différences dans les bégayages en fonction de facteurs linguistiques. Ainsi, Howell *et al.* (1999) ont utilisé l'index de complexité phonologique (Index of Phonological Complexity) développé par Jakielski (1998) pour montrer que les mots les plus bégayés présentent une complexité phonologique accrue. Ces observations ont également été faites par Wolk (2000) qui précise que ce sont les groupes consonantiques situés en initiale de mot qui sont davantage bégayés par rapport à ceux situés à l'intérieur ou à la fin des mots. Wolk (2000) explique ce phénomène par des exigences élevées en termes de rapidité de mouvements nécessaire à la prononciation d'un groupe consonantique. Les travaux de Jayaram (1983), portant sur le bégaiement chez des personnes mono- et bilingues qui bégaient, ont précisé que les disfluences survenaient davantage sur les consonnes sourdes, et plus précisément sur les fricatives sourdes et ce, chez les deux groupes. Wingate (1984), quant à lui, a mis en relation les pics de hauteur et la fréquence d'apparition des bégayages : en d'autres termes, il y aurait davantage d'accidents de parole lorsque la fréquence fondamentale augmente. Plus récemment, une étude menée par Blomgren (2012) s'est intéressée à l'influence des facteurs phonologiques sur la fréquence des bégayages. Cette étude précise que les consonnes labiodentales, les consonnes non-voisées et les groupes consonantiques seraient bégayés plus fréquemment.

1.2 Problématique et hypothèses

Dans le cadre de cette recherche, nous nous sommes fixés pour *objectif* de nous interroger sur une éventuelle influence de certaines structures phonologiques sur l'apparition des bégayages. Notre *problématique* globale peut être formulée à travers la question suivante : les caractéristiques articulatoires des sons ont-elles une influence sur la fréquence d'apparition des disfluences ?

Pour répondre à cette question et pour rendre nos résultats plus robustes, nous avons choisi de travailler sur deux langues, à savoir le français et le slovaque. Disposer d'un corpus

multilingue permettra ainsi de mettre en évidence les phénomènes d'apparition du bégaiement qui sont universels, c'est-à-dire indépendants de la langue parlée, et de vérifier si certaines caractéristiques propres à une langue spécifique peuvent être à l'origine de disfluences supplémentaires.

Plusieurs éléments laissent penser que la distribution des disfluences pourrait être différente dans la parole bègue, selon la langue parlée. Nous savons premièrement que la complexité phonologique joue un rôle important dans le bégaiement (Anderson et Byrd, 2008). Or, la fréquence d'utilisation des sons est différente en français (Wioland, 1985 ; Gromer & Weiss, 1998) et en slovaque (Stefanik, 1999 ; Didirkova, 2013). Il est donc permis de penser que, selon les phonèmes présents dans une langue ou non et selon leur fréquence d'apparition, la distribution du bégaiement va se trouver modifiée d'une langue à l'autre. Par ailleurs, l'étude de Browman et Goldstein (1989) a révélé l'existence de différences inter-langues au niveau des patterns articulatoires lors de la production d'une même syllabe. Par conséquent, cela suppose que certains patrons articulatoires seront plus complexes dans une langue par rapport à une autre, entrainant ainsi davantage de disfluences. Notre *hypothèse* est donc que nous ne retrouverons pas qualitativement les mêmes disfluences dans deux langues différentes. Nous pensons en effet que le fait d'avoir un système phonologique et des patterns articulatoires différents dans chaque langue va se traduire par un bouleversement des fréquences de disfluences pour chaque phonème.

2 Méthodologie

2.1 Corpus, sujets, méthode

2.1.1 Participants et corpus

Notre corpus est constitué d'enregistrements de locuteurs francophones et slovacophones. Le choix des langues étudiées a été guidé par la volonté d'avoir deux systèmes phonologiques différents, ce qui permettra de définir avec plus de précision l'importance ou non de la présence de certaines configurations phonologiques dans l'apparition d'une disfluence.

Nos données ont été obtenues à l'aide de l'Association Parole – Bégaiement, qui a permis de contacter des personnes bègues francophones, et du Centre pour les Disfluences de Bratislava, qui a rendu possible l'acquisition de données en langue slovaque. Douze sujets ont pris part à cette étude : 7 locuteurs français (moyenne d'âge : 35 ans) et 5 locuteurs slovaques (moyenne d'âge : 33 ans) ont ainsi accepté de participer à cette investigation. Tous les participants sont adultes et seules deux sont de sexe féminin, un pour chaque langue. Soulignons que, devant la difficulté d'obtenir des enregistrements en parole pathologique, et plus précisément lorsqu'il s'agit de locuteurs bègues, nous avons fait le choix d'harmoniser nos données en travaillant sur une durée de corpus comparable pour chaque langue. Nos analyses reposent ainsi sur 22 minutes et 17 secondes de temps de parole en français et sur 20 minutes 53 secondes en slovaque, soit une durée totale d'enregistrements de 43 minutes et 10 secondes. Cet article étant une étude préliminaire, il va de soi que nous enregistrerons de nouveaux sujets si nos hypothèses étaient vérifiées, et ce afin d'avoir des données similaires d'un point de vue temporel dans les deux langues.

Concernant le contenu du corpus, il a été demandé à nos sujets de se présenter, de décrire

leur journée-type, de raconter l'histoire du dernier film vu ou encore du dernier livre lu et ce, dans le but d'obtenir de la parole spontanée. Les enregistrements ont été effectués à l'aide d'un ordinateur portable relié à un microphone. Les entretiens se déroulaient dans un environnement calme.

Les locuteurs ont prononcé 5325 mots, dont 2969 en langue française et 2356 en langue slovaque. Au total, 421 mots bégayés ont été analysés. 189 ont été prononcés par les locuteurs francophones et 232 par les sujets slovacophones.

2.1.2 Méthode d'analyse

Les enregistrements sonores ont donné lieu à deux transcriptions, l'une orthographique et l'autre en alphabet phonétique international. Dans les transcriptions, il a été tenu compte de ce que le locuteur prononçait ainsi que des disfluences produites à l'oral.

La transcription orthographique a servi d'appui pour le décompte de mots prononcés par le locuteur, ainsi que des disfluences produites. Il est à noter que, dans le cadre de cette étude, seules les répétitions et les prolongations ont été inventoriées. Les sidérations, blocages, mots d'appui et autres disfluences ont été exclus de l'étude (Blomgren, 2012).

Par la suite, les phonèmes bégayés ont été analysés selon les critères suivants :

- nature du phonème : nous avons quantifié le nombre de consonnes, de voyelles, de semi-consonnes et de diphtongues (pour la langue slovaque uniquement) bégayées ;
- lieu d'articulation des voyelles : nous avons classifié les disfluences produites selon qu'elles apparaissent sur des voyelles antérieures, centrales ou postérieures ;
- lieu d'articulation des consonnes : le nombre de disfluences produites par point de resserrement maximal a été décompté ;
- mode articulatoire des consonnes : nous avons dénombré les occlusives et les constrictives bégayées ;
- voisement : la quantité d'éléments voisés et non-voisés bégayés a été définie.

2.2 Résultats

Le taux moyen de mots bégayés chez les locuteurs francophones a été de 6,36% *vs.* 9,85% pour les sujets slovacophones.

2.2.1 Répartition des bégayages sur les consonnes, voyelles, diphtongues et semi-consonnes

Une différence a été constatée pour ce premier critère entre locuteurs slovacophones et francophones. Pour les premiers cités en effet, ce sont les diphtongues qui subissent le plus grand nombre de disfluences (20%), alors que ce type de son n'existe pas en français. Concernant les autres catégories de phonèmes, on retrouve des résultats similaires dans les deux langues, les consonnes étant davantage bégayées (8,12% pour les locuteurs francophones *vs.* 9,45% pour les locuteurs slovacophones) que les voyelles (2,25% en français *vs.* 2,53% en slovaque) (cf. Table 1).

	locuteurs francophones	locuteurs slovacophones
voyelles	2,25%	2,53%
consonnes	8,12%	9,45%
semi-consonnes	0%	NA
diphtongues	NA	20%

TABLE 1 – Répartition des bégayages en fonction de la nature du phonème. Les résultats sont donnés en termes de pourcentage d'occurrences bégayées par rapport au nombre d'occurrences prononcées.

De manière générale, le Tableau 1 montre que, par rapport aux consonnes, la production des voyelles est moins problématique pour les locuteurs bègues.

2.2.2 Répartition des bégayages en fonction du lieu d'articulation des consonnes

L'impact du lieu d'articulation des consonnes varie en fonction de la langue. Tandis que chez les locuteurs francophones, les consonnes les plus altérées étaient les consonnes post-alvéolaires (28,31%), les locuteurs slovacophones ont produit davantage de disfluences sur les consonnes labiodentales (15,84%) avec également un grand taux de bégayages sur les consonnes bilabiales (12,81%) ainsi que sur la laryngale (13,7%). (cf. Table 2).

	locuteurs francophones	locuteurs slovacophones
consonnes bilabiales	7,48%	12,81%
consonnes labio-dentales	6,45%	15,84%
consonnes alvéo-dentales	8,93%	NA
consonnes alvéolaires	8,33%	11,43%
consonnes post-alvéolaires	28,31%	NA
consonnes alvéo-palatales	NA	5,77%
consonnes palatales	0%	2,02%
consonnes vélaires	4,59%	6,58%
consonne laryngale	NA	13,7%

TABLE 2 – Répartition des bégayages en fonction du lieu d'articulation des consonnes. Les résultats sont donnés en termes de pourcentage d'occurrences bégayées par rapport au nombre d'occurrences prononcées.

2.2.3 Répartition des bégayages en fonction du mode articulatoire des consonnes

Le Tableau 3 nous montre des résultats similaires pour le français et le slovaque. Dans les deux langues, les consonnes occlusives sont légèrement plus altérées (8,68% pour le français et 10,12% pour le slovaque) que les consonnes constrictives (respectivement à 7,96% et 8,64%). (cf. Table 3).

	locuteurs francophones	locuteurs slovacophones
consonnes occlusives	8,68%	10,12%
consonnes affriquées	0%	0%
consonnes constrictives	7,96%	8,64%
consonne vibrante	NA	1,68%

TABLE 3 – Répartition des bégayages en fonction du mode articulatoire des consonnes. Les résultats sont donnés en termes de pourcentage d'occurrences bégayées par rapport au nombre d'occurrences prononcées.

2.2.4 Répartition des bégayages en fonction du voisement

En ce qui concerne l'apparition des disfluences sur les éléments voisés ou non-voisés, aucune différence n'a été observée entre les deux langues étudiées : en français comme en slovaque, ce sont les consonnes non-voisées (7,78% pour le français et 11,5% pour le slovaque) qui présentent un taux d'altération plus élevé que leurs homologues voisés (respectivement 4,51% et 4,91%) (cf. Table 4).

	locuteurs francophones	locuteurs slovacophones
phonèmes voisés	4,51%	4,91%
phonèmes non-voisés	7,78%	11,5%

TABLE 4 – Répartition des bégayages en fonction du voisement. Les résultats sont donnés en termes de pourcentage d'occurrences bégayées par rapport au nombre d'occurrences prononcées.

3 Synthèse, discussion et perspectives

Des bégayages produits par des locuteurs bègues francophones et slovacophones ont été analysés dans cette étude et ce, notamment du point de vue des caractéristiques phonologiques de ces disfluences. Conformément aux résultats obtenus préalablement dans d'autres études menées sur le sujet pour l'anglais des Etats-Unis (Brown, 1935 ; Jayaram, 1983 ; Blomgren, 2012), notre hypothèse de départ supposait une certaine influence de la

structure phonologique de la langue sur l'apparition des bégayages. Dans ce cadre, les résultats obtenus ont permis de relever plusieurs points intéressants.

Notre travail a permis de relever des phénomènes observables sur les deux langues étudiées et d'autres qui semblent être plus dépendants de la langue parlée. Ainsi, le mode articulatoire des consonnes, de même que le critère du voisement présentent les mêmes résultats chez les locuteurs bègues francophones et slovacophones. En effet, en ce qui concerne le mode articulatoire des consonnes, les disfluences produites par les locuteurs participant à l'étude se trouvent essentiellement sur des consonnes occlusives. En ce qui concerne le trait de sonorité, on observe que les éléments non-voisés ont présenté une fréquence de bégayages plus élevée que leurs homologues voisés et ce, à nouveau dans les deux langues étudiées. Il est également à noter que, dans son étude, Blomgren (2012) avait obtenu le même résultat pour l'anglais parlé aux Etats-Unis avec 16% de bégayages sur les phonèmes non-voisés contre 13% de bégayages sur les phonèmes voisés. Ce résultat pourrait s'expliquer par la présence d'une certaine rigidité musculaire chez les personnes bègues (Max *et al.*, 2005 ; Hirsch, 2007), qui empêcherait une transition rapide des sons voisés aux sons non-voisés.

En revanche, des différences inter-langues ont également été observées pour d'autres critères. Les variations les plus importantes ont été relevées pour le lieu d'articulation des consonnes. Les participants francophones ont en effet présenté le plus grand nombre de disfluences sur les consonnes post-alvéolaires. Ce résultat diffère de ceux obtenus pour les locuteurs slovacophones, pour qui les consonnes labiodentales semblent plus problématiques. Signalons que Blomgren (2012) a pu relever des difficultés semblables pour cette catégorie de consonnes chez des locuteurs parlant l'anglais des Etats-Unis. Par ailleurs, il nous semble intéressant de rappeler que les voyelles les plus touchées par un bégayage chez les locuteurs slovacophones sont les diphtongues, alors que ces dernières sont absentes du système phonologique du français. Ces différences inter-langues pourraient sans doute s'expliquer par la présence de patterns articulatoires différents d'une langue à une autre pour un même son ou un même groupe de sons (Browman & Goldstein, 1989) : par conséquent, on pourrait penser que la dynamique utilisée en slovaque pour produire des labio-dentales et celle en français pour les post-alvéolaires sont plus problématiques pour les locuteurs bègues. Enfin, le fait que les diphtongues soient les phonèmes les plus bégayés en slovaque confirme les propos de Howell *et al.* (1999) indiquant un lien entre la complexité phonologique et le nombre de disfluences.

En perspective, nous pensons qu'il est avant tout nécessaire d'enregistrer davantage de locuteurs dans les deux langues et ce, afin d'aboutir à un corpus d'une durée plus longue. Cela permettra de vérifier si les premiers résultats délivrés dans le cadre de ce travail sont robustes ou non. En outre, et si des différences inter-langues sont effectivement constatées, il serait intéressant de refaire ce type d'études sur d'autres langues. Par ailleurs, nous pensons qu'il serait également souhaitable d'affiner notre analyse en tenant compte de la localisation de la disfluence, tout en normalisant les occurrences de bégaiement avec les fréquences d'apparition des phonèmes dans chaque langue. En d'autres termes, il serait intéressant de relever les caractéristiques phonétiques des sons disfluents en fonction de l'endroit où l'accident de parole est localisé (début de tour de parole, à l'intérieur d'un énoncé,…). De même, nous savons qu'il existe différents types de disfluences dans la parole bègue (répétitions, prolongements,…). Il serait donc judicieux de reprendre la méthodologie

employée dans le cadre de ce travail tout en tenant compte de la nature de l'accident de parole. Enfin, des études mettant en lien la complexité syntaxique et le bégaiement, de même que des recherches portant sur la relation entre la nature des mots et ce trouble de la communication permettraient de compléter cet article en se plaçant sur d'autres niveaux linguistiques.

Remerciements

Nous tenons à remercier Madame Ľubomíra Štenclová pour nous avoir ouvert les portes du Centre de Disfluences de Bratislava. Nous exprimons également notre reconnaissance à Madame Jacqueline Bru qui nous a permis d'enregistrer des patients bègues francophones.

Références

ANDERSON, J. D., BYRD, C. T. (2008). Phonotactic probability effects in children who stutter. *Journal of speech, language and hearing research*, 51:851 – 866.

AU-YEUNG, J., HOWELL, P., PILIGRIM, L. (1998). Phonological words and stuttering on function words. *Journal of speech, language and hearing research,* 41:1019–1030.

BLOMGREN, M. (2012). Do speech sound characteristics really influence stuttering frequency ? *Proceedings of the 7th World Congress of Fluency Disorders,* CD-ROM.

BLOODSTEIN, G., GROSSMANN, M. (1981). Early stuttering : some aspects of their form and distribution. *Journal of speech and hearing research,* 10:786-789.

BROWMAN, C. P., GOLDSTEIN (1989). Articulatory gestures as phonological units. Phonology, 6, 201-251.

BROWN, S.F. (1938). Stuttering with relation to word accent and word position. *The Journal of Abnormal and Social Psychology,* 33:112-120.

BRUNDAGE, S. B., RATNER, N.B. (1989). Measurement of stuttering frequency in children's speech. *Journal of fluency disorders,* 14:351-358.

DIDIRKOVA, I. (2013). Le bégaiement : étude inter-langues des disfluences réalisées par des locuteurs bègues francophones et slovacophones. *Mémoire de Master II, Université Paul Valéry, Montpellier.*

GROMER B., WEISS M. (1990). *Lire*, tome 1 : apprendre à lire - Armand Colin.

HIRSCH, F. (2007). Le bégaiement : perturbation de l'organisation temporelle et conséquences spectrales. *Thèse soutenue à l'Université Marc Bloch, Strasbourg, 343 p.*

HOWELL, P., AU-YEUNG, J., PILIGRIM, L. (1999). Utterance rate and linguistic properties as determinants of lexical dysfluencies in children who stutter. *The journal of the acoustical society of America,* 105:481-490.

HOWELL, P., AU-YEUNG, J., YARUSS, J. S., ELDRIDGE, K. (2006). Phonetic difficulty and stuttering. *Clinical Linguistics and Phonetics,* 20:703-716.

JAKIELSKI, K. J. (1998). Motor organization in the acquisition of consonant clusters. *University of Texas Austin. Dissertation / PhD thesis.*

JAYARAM, M. (1983). Phonetic influences on stuttering in monolingual and bilingual stutterers. *Journal of communication disorders,* 16(4):287-297.

MAX, L., GRACCO, VL. (2005). Coordination of oral and laryngeal movements in the perceptually fluent speech of adults who stutter. *Journal of speech and hearing research,* 48(3):524-542.

NATKE, U. et al. (2004). Linguistic stress, within-word position, and grammatical class in relation to early childhood stuttering. *Journal of Fluency Disorders* 29(2):109-22.

PERKINS, W. (1992). *Stuttering prevented.* San Diego : Singular Publishing Group, Inc.

PIERART, B. (2011). *Les Bégaiements de l'adulte.* Wavre : Mardaga.

ŠTEFÁNIK, J., RUSKO, M., POVAŽANEC, D. (1999). Frekvencia slov, grafém, hlások a ďalších elementov slovenského jazyka, Jazykovedný časopis, 50ème Année d'édition, n°2, 81-93.

WINGATE, M. (1979). The first three words. *Journal of speech and hearing research,* 22:604-612.

WIOLAND, F. (1985). Les structures syllabiques du français. Genève : Slatkine.

WOLK, L., BLOMGREN, M., SMITH, A. (2000). The frequency of simultaneous disfluency and phonological errors in children : A preliminary investigation. *Journal of fluency disorders,* 25:269-281.

Application du critère BIC pour la segmentation en tours de chant

Marwa Thlithi, Thomas Pellegrini, Julien Pinquier, Régine André-Obrecht, Patrice Guyot
IRIT – Université de Toulouse – 118, route de Narbonne – 31062 Toulouse Cedex 9 – France
{thlithi, pellegrini, pinquier, obrecht, guyot}@irit.fr

RESUME

Dans le cadre d'un projet sur l'indexation de documents ethnomusicologiques sonores (ANR CONTINT DIADEMS), le repérage des chanteurs et des chœurs est apparu comme essentiel et nous a amené à s'interroger sur la notion de « tours de chant ». Dans cet article, nous présentons nos premiers pas dans le domaine en proposant une méthode de segmentation fondée sur le Critère d'Information Bayésien (BIC) qui vise à détecter des changements de chanteurs dans des enregistrements musicaux. Le corpus de cette étude est composé d'enregistrements musicaux fournis par des ethnomusicologues et il nous permet d'illustrer l'importance du coefficient de pénalité du critère BIC : sa valeur optimale varie en fonction du contenu des enregistrements. Pour s'affranchir de l'apprentissage d'une unique valeur de ce paramètre, nous proposons de recueillir plusieurs segmentations pour plusieurs valeurs du paramètre et de consolider la détection *a posteriori*. Un gain relatif en termes de F-mesure, de 15% (7% absolu) est obtenu entre cette décision *a posteriori* et une décision prise après apprentissage du coefficient de pénalité.

ABSTRACT

Segmentation in singer turns with the Bayesian Information Criterion

As part of a project on indexing ethno-musicological audio recordings (ANR CONTINT DIADEMS), determining singers and choirs automatically appeared to be essential and led us to reflect about the notion of "singer turns". In this article, we report our first experiments in this direction by exploring a method based on the Bayesian Information Criterion (BIC) to detect singer turns. The BIC penalty coefficient was shown to vary when determining its value to achieve the best performance for each recording. In order to avoid the decision about which single value is best for all the documents, we propose a combination of several segmentations obtained with different values of this parameter. This method uses majority voting. A gain of 15% relative (7% absolute) in terms of F-measure was obtained compared to a single coefficient determined on a development sub-corpus.

MOTS-CLES : segmentation audio, tours de chant, critère BIC.
KEYWORDS: audio segmentation, singer turns, BIC criterion.

1 Introduction

Un document sonore peut être structuré automatiquement de bien des manières en fonction de l'objectif final. S'il s'agit d'indexation d'un fond de documents par exemple, on se posera vraisemblablement les questions : est-on en présence de parole, de musique, à quels moments, qui parle, qui chante, etc. L'une des premières étapes d'inférence de cette structure est de découper

puis d'étiqueter des zones ou segments dits « homogènes acoustiquement ». En traitement automatique de la parole, il s'agira d'identifier les changements de locuteurs ou tours de parole pour savoir qui parle et à quels moments dans le document, pour ensuite faciliter éventuellement d'autres traitements qui pourraient suivre comme la transcription automatique de ce qui est dit (Anguera, 2012). Dans le contexte du projet ANR CONTINT DIADEMS [1] (Description, Indexation, Accès aux Documents EthnoMusicologiques et Sonores) sur l'indexation de documents ethnomusicologiques sonores, nous nous sommes posés la question de savoir si nous pouvions réaliser le même type de structuration en identifiant les changements de chanteurs (solistes et/ou chœurs) au sein des enregistrements musicaux qui sont l'objet d'étude du projet. Nous utiliserons l'expression « segmentation en tours de chant » par analogie à la segmentation en tours de parole. Par chant, nous désignons la voix chantée au sens large, accompagnée ou non par des instruments, en groupe ou en soliste.

La Figure **1** illustre notre problématique de segmentation en tours de chant. La vérité « terrain » consiste en une annotation manuelle des tours de chant et entrées/sorties d'instruments éventuelles. Dans cet article, nous présentons une méthode de segmentation en tours de chant qui précèderait une étape ultérieure de regroupement de segments figurant un même chanteur ou groupe de chanteurs.

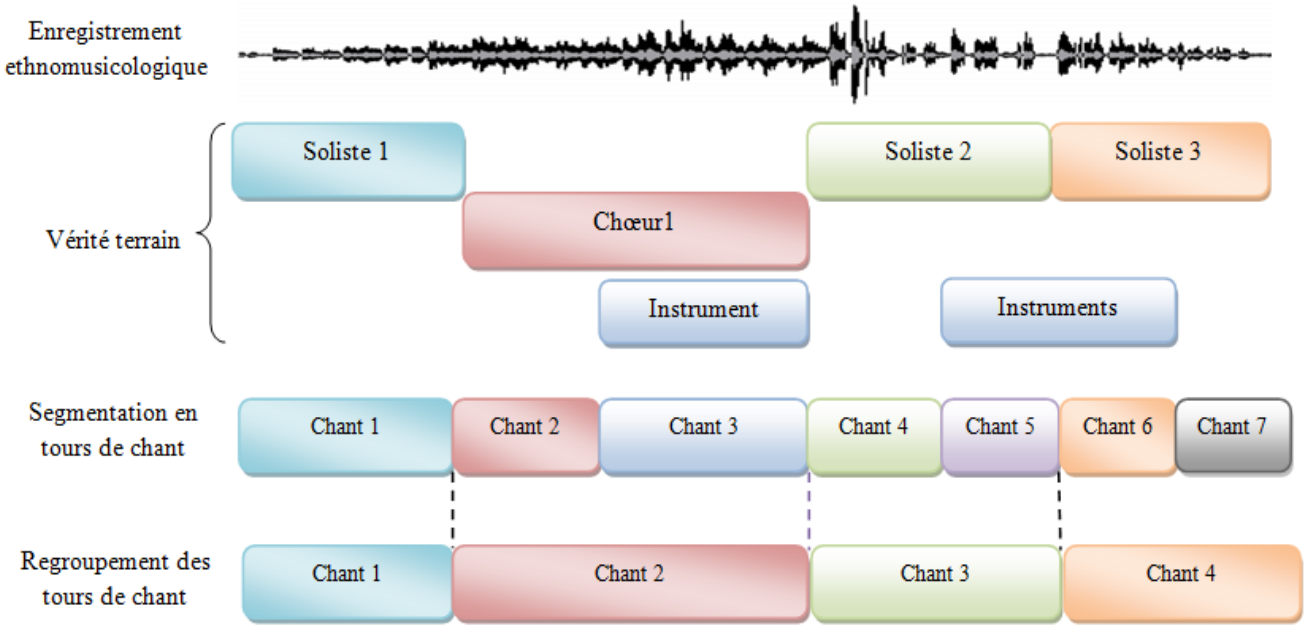


Figure 1 : Illustration de la segmentation en tours de chant et de l'étape de regroupement.

Depuis quelques années notre équipe travaille sur des questions liées à la voix chantée, en particulier sur la détection de voix chantée (Lachambre, 2009) et la segmentation chant solo/chœur (Le Coz, 2012) dans un contexte musical. Ce dernier traitement est réalisé à l'aide de critères similaires à la détection de parole superposée. Dans cette étude, nous continuons à utiliser une analogie avec la parole en testant la méthode de segmentation fondée sur le Critère d'Information Bayésien (BIC), très utilisée en segmentation en tours de parole (Zhu, 2005). Son application sur des enregistrements de chant nécessite une revisite de ce critère et une adaptation appropriée de ses paramètres. En particulier, nous avons observé qu'il était difficile de déterminer une valeur optimale du coefficient de pénalité du critère selon les enregistrements. Ce qui nous a amené à

[1] http://www.irit.fr/recherches/SAMOVA/DIADEMS/

proposer une méthode de type vote majoritaire impliquant plusieurs segmentations réalisées avec plusieurs valeurs de ce paramètre.

Dans la section 2, nous commençons par rappeler brièvement le critère BIC théorique et l'algorithme de référence que nous avons utilisé. Ensuite, le contexte applicatif est présenté en détail puis nous détaillons la mise en œuvre spécifique à notre tâche et les premiers résultats. Dans la section 4, le problème de la variabilité du facteur de pénalité est discuté et la méthode de combinaison est décrite. Enfin, nous comparons les performances globales obtenues.

2 Critère BIC pour la segmentation audio

2.1 Présentation générale du critère

Le Critère d'Information Bayésien, comme son nom l'indique, se place dans un contexte bayésien de sélection de modèles. Variante du critère d'Akaïke, il est utilisé dans de nombreux contextes applicatifs et ce depuis très longtemps (Akaïke, 1974), (Schwarz, 1978). Ces dernières années, il est au cœur de nombreux travaux de segmentation audio (Cettolo, 2005), (Siu, 1991), (Chen, 1998), (Delacourt, 2000) et bien des systèmes de segmentation et de regroupement en locuteurs, actuels et performants, implique le BIC.

La segmentation audio consiste à diviser le flux audio en des segments homogènes en effectuant un test d'hypothèses. Pour chaque point de changement potentiel, il y a deux hypothèses possibles : la première suppose que, de part et d'autre de ce point, le signal obéit au même modèle probabiliste, noté M_0 (H_0), la deuxième suppose qu'il y a un changement de modèle et que deux modèles différents M_1 et M_2 sont nécessaires (H_1). Dans la pratique, les modèles sont estimés sur trois fenêtres d'analyse et des critères tels que le critère BIC sont utilisés pour déterminer si le signal est « mieux » représenté par deux modèles distincts ou par un modèle unique ; un seuil est déterminé de manière empirique ou adapté dynamiquement.

Il s'en suit que si le signal analysé correspond à une séquence de N vecteurs d'observation (vecteurs acoustiques ou trames) de dimension d, noté $X_0(x_1, x_2, \dots, x_N)$; un point de changement potentiel positionné après la trame t induit deux sous suites consécutives $X_1(x_1, x_2, \dots, x_t)$ et $X_2(x_{t+1}, x_2, \dots, x_N)$. En supposant que X_0, X_1 et X_2 suivent des lois gaussiennes données respectivement par $M_0(\mu_0, \Sigma_0)$, $M_1(\mu_1, \Sigma_1)$ et $M_2(\mu_2, \Sigma_2)$, le critère ΔBIC à l'instant t est donné par :

$$\Delta BIC(t) = R(t) - \lambda P$$

$R(t)$ est le rapport de log vraisemblance entre les deux hypothèses $(LL(H_1)/LL(H_0)$ et s'écrit :

$$R(t) = \frac{1}{2}\left(N \log\left(\left|\widehat{\Sigma_0}\right|\right) - t \log\left(\left|\widehat{\Sigma_1}\right|\right) - (N-t)\log\left(\left|\widehat{\Sigma_2}\right|\right)\right.$$

avec $\left|\widehat{\Sigma_i}\right|$, déterminant de la matrice de covariance Σ_i, estimée sur la suite X_i.

P est proportionnel à la différence des nombres de paramètres estimés pour chaque hypothèse et vaut, dans le cas de matrices de covariance pleines :

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N$$

Le facteur de pénalité λ est appris de telle sorte que le critère ΔBIC soit positif dès lors que l'hypothèse H_1 est vérifiée, indiquant l'existence de deux modèles différents. Sinon, l'hypothèse

H_0 est validée ainsi que l'existence d'un seul modèle pour la fenêtre X_0.

2.2 Algorithme de référence

La mise en œuvre du BIC implique de déterminer deux paramètres importants : la taille N de la fenêtre de signal dans laquelle une frontière de segment est recherchée, et le facteur de pénalité λ du critère. Dans un premier temps, nous avons cherché à fixer la valeur de ces deux paramètres sur un sous-ensemble de notre corpus décrit dans la prochaine section. Une première version de l'algorithme faisait suite aux travaux d'El-Khoury dans laquelle la taille de la fenêtre était constante (El-Khoury, 2009). Cependant, la recherche d'une taille optimale de cette fenêtre, indépendante de l'enregistrement considéré, s'est révélée difficile. Nous avons alors implémenté une autre version de l'algorithme dans laquelle la taille de fenêtre d'analyse augmente tant qu'aucune frontière potentielle n'est trouvée. Ce procédé s'inspire d'études en segmentation de la parole (Andre-Obrecht, 1988), et il est illustré dans la Figure **2**.

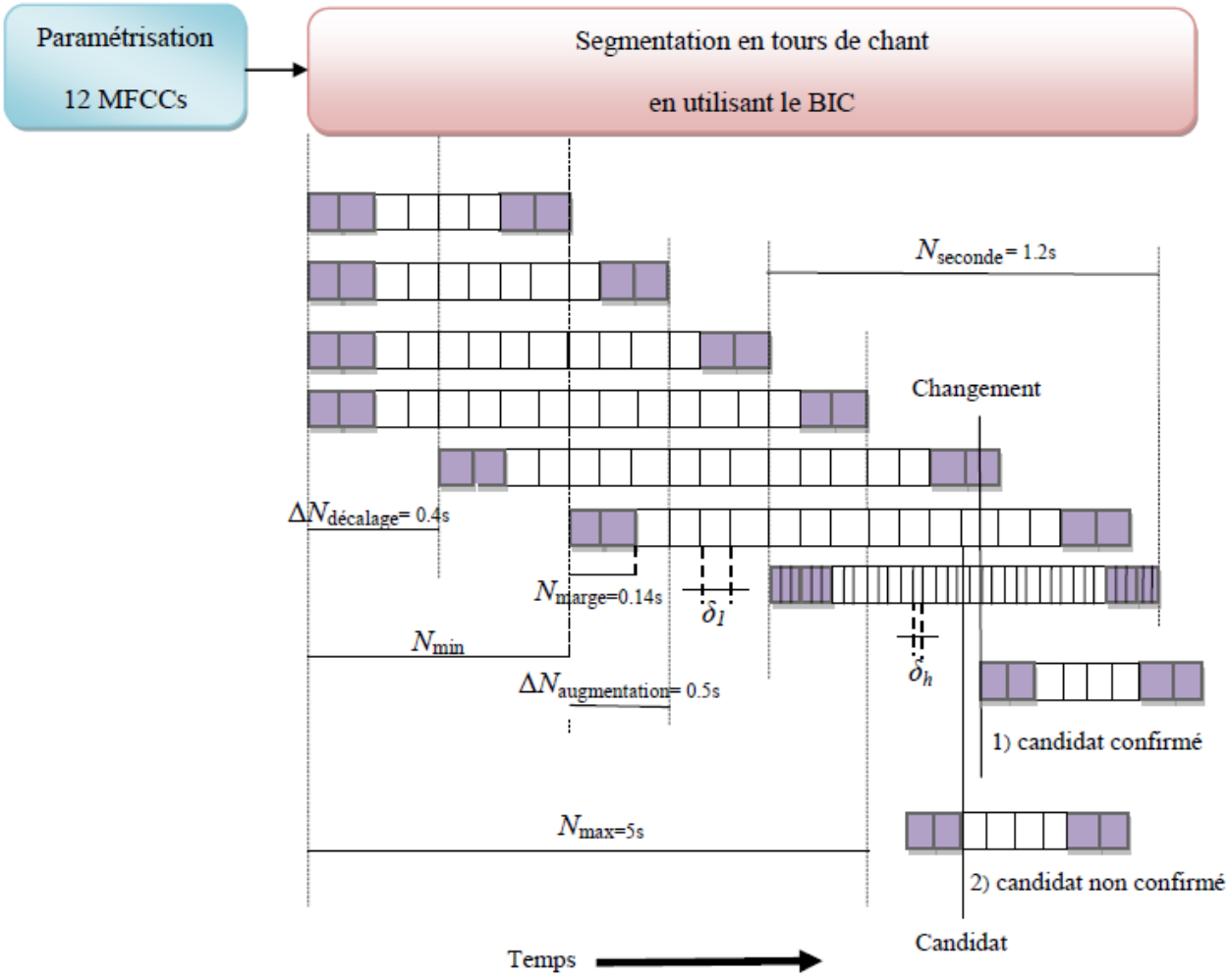


Figure 2 : Illustration de l'algorithme de notre système de segmentation en tours de chant. Figure adaptée de (Cettolo, 2005).

La recherche se déroule en deux temps, impliquant deux résolutions temporelles différentes :

1. La longueur initiale de la fenêtre d'analyse est fixée à N_{min} . Elle augmente de $\Delta N_{augmentation}$ tant que le test d'hypothèse ΔBIC ne valide aucune frontière interne, jusqu'à une valeur maximale N_{max} . Les valeurs de ΔBIC sont calculées à intervalles

réguliers pour des valeurs échantillonnées de t, à savoir une fois toutes les $\delta 1$ observations. Si aucune frontière n'est détectée lorsque N_{max} est atteint, la fenêtre d'analyse est décalée de $\Delta N_{décalage}$ et l'analyse est réinitialisée.

2. Si une frontière potentielle est détectée, une fenêtre de longueur $N_{seconde}$ est centrée sur cette frontière et les valeurs de ΔBIC sont calculées au sein de cette fenêtre à une résolution haute, toutes les δh observations, afin d'affiner la position de cette frontière.

Nous imposons que toute frontière ne produise aucun segment d'une durée inférieure à N_{marge}, ce qui implique qu'aucune frontière n'est recherchée entre les zones $[1, N_{marge}]$et $[N - N_{marge}, N]$.

3 Performances du système de référence

3.1 Contexte applicatif et corpus d'étude

Nous avons effectué nos expériences sur un corpus d'enregistrements sélectionnés spécifiquement pour la détection des tours de chant par les ethnomusicologues partenaires du projet DIADEMS. Des exemples sont accessibles en ligne[2]. Ces enregistrements ont été réalisés dans plusieurs pays sub-sahariens (Congo, Gabon, Cameroun), avec une qualité acoustique variable (enregistrements en extérieur en général, présence de bruits de fonds et d'évènements sonores autres que la musique). Ils contiennent principalement des tours de chant soliste/chœur, des zones de voix chantée en alternance ou superposée avec des instruments ou de la parole. Les fichiers ont été annotés manuellement en tours de chant. Une frontière est posée dans les situations suivantes :

- Passage d'un seul chanteur (soliste) à plusieurs chanteurs (chœur) et vice versa,
- Passage d'un chanteur A à un chanteur B,
- Passage du chant à la parole et vice versa.

Ce corpus est constitué de 9 enregistrements dont la durée totale est de 20 minutes que nous avons divisées en un corpus de développement (DEV) et un corpus d'évaluation (EVAL) dans les proportions 20% et 80% respectivement.

3.2 Mise en œuvre de l'algorithme

3.2.1 Choix des paramètres acoustiques

Toujours par analogie entre tours de parole et tours de chant, nous avons travaillé avec des paramètres acoustiques utilisés en segmentation de la parole. Nous avons étudié comme vecteur d'observation, les 12 MFCC, accompagnés éventuellement de l'énergie de la trame, des dérivées premières et secondes de ces coefficients, les MELSPEC, les PLP et les RASTA-PLP. L'estimation est réalisée sur 20 ms, toutes les 10 ms.

3.2.2 Ajustement des paramètres

Les paramètres de l'algorithme ont été déterminés sur le corpus de développement :

- N_{min} : la taille minimale de la fenêtre de recherche d'une frontière est fixée à 0,8s tandis que sa longueur maximale N_{max} est de 5s,
- $\Delta N_{augmentation}$: le nombre d'observations ajoutées à la fenêtre de détection tant qu'il

[2] http://diadems.telemeta.org/archives/fonds/CNRSMH_DIADEMS/

n'y a pas de frontière de segments détectée et tant que la taille maximale n'est pas atteinte, correspond à 0,5s,

- $\Delta N_{décalage}$: le décalage de la fenêtre sur le signal lorsque la taille maximale est atteinte et qu'aucune frontière potentielle n'est découverte correspond à 0,4s,
- N_{marge} : la taille minimale d'un segment de chant est de 0,14s,
- $N_{seconde}$: la taille de la fenêtre d'analyse fine correspond à 1,2s,
- la faible résolution temporelle du calcul du ΔBIC est de $\delta 1 = 5$ (1 trame sur 5, soit 50ms) tandis que la haute résolution considère toutes les trames de paramètres : $\delta h = 1$ (soit 10ms de précision).

3.2.3 Premiers résultats

Les performances ont été évaluées en termes de Rappel, Précision et F-mesure pondérés sur les durées des fichiers. Nous acceptons une tolérance de plus ou moins 30 ms sur les positions des frontières. Les meilleurs résultats ont été obtenus en limitant le vecteur d'observation à 12 coefficients MFCC. Seuls les paramètres RASTA-PLP ont donné des résultats équivalents à ceux obtenus avec les MFCC mais sans gain significatif.

Afin de définir une valeur « optimale » du coefficient de pénalité λ, les performances du système ont été calculées en le faisant varier sur les enregistrements de l'ensemble de développement. La valeur choisie de 1,02 pour présenter les résultats correspond au point de fonctionnement où la précision et le rappel sont égaux sur les enregistrements de DEV. Avec une telle configuration, des F-mesures de 66,6% et de 43,7% ont été obtenues sur DEV et EVAL respectivement. Ces valeurs sont rapportées dans la première partie du Tableau **1** de la section 4. La performance sur EVAL est beaucoup plus faible que sur DEV : une analyse des résultats montre que deux enregistrements qui présentent des alternances de chant solo/chœur très rapides, ont des performances très faibles, avec des F-mesures entre 30% et 40%.

4 Pertinence du coefficient de pénalité λ

L'ajustement du coefficient de pénalité s'est avéré délicat ; nous illustrons et analysons dans la première partie de cette section sa pertinence ainsi que sa grande sensibilité aux variations des conditions et de contenus des enregistrements. Afin de remédier à ce problème de variabilité, il nous est apparu nécessaire de relâcher cette contrainte, en raisonnant systématiquement sur plusieurs valeurs de λ et en confrontant plusieurs résultats de segmentation ; ce nouvel algorithme est présenté dans cette même section.

4.1 Variabilité des performances en fonction de λ

Le rôle du coefficient λ dans le critère BIC est de pénaliser une modélisation trop complexe : dans le cadre gaussien et multi modèles qui est le nôtre, plus la valeur de λ augmente, plus l'hypothèse H_1 est pénalisée et plus l'insertion d'une frontière est difficile : l'algorithme a tendance à moins segmenter. Globalement, choisir la bonne valeur de λ revient à trouver le bon compromis entre Rappel et Précision. La Figure **3** présente la variation de la performance globale de notre système sur le corpus DEV en fonction du coefficient de pénalité dans l'intervalle $[0{,}8 \quad 1{,}2]$. Nous remarquons que la performance varie de manière importante en fonction de ce facteur. Une des raisons majeures est sans nul doute liée à la petite taille de l'ensemble DEV.

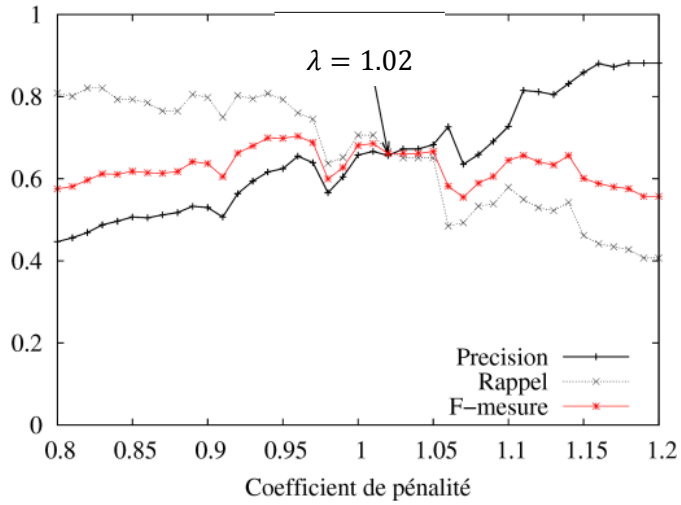


Figure 3 : Variation de la performance du système de segmentation en fonction du coefficient de pénalité sur le corpus DEV.

Néanmoins nous avons remarqué que la valeur de λ peut varier énormément d'un enregistrement à l'autre dès lors que l'on cherche à l'optimiser sur l'enregistrement seul. Pour certains enregistrements, nous trouvons de bonnes performances avec des valeurs proches de 1,2 ; or les segments attendus sont longs ; les valeurs moins élevées de l'ordre de 0,8 se révèlent meilleures là où les segments attendus sont plus courts. Cela engendre une variabilité importante de la performance globale de notre système en fonction de λ.

À des fins de comparaison, nous avons, pour chaque enregistrement du corpus DEV et EVAL, déterminé la meilleure valeur de F-mesure obtenue en faisant varier le coefficient de pénalité ; nous appellerons ce système artificiel, le système « *oracle* ». Il en résulte les performances globales données dans le Tableau **1**. Sur DEV et EVAL respectivement, la F-mesure atteint 79,5% et 53,8%. En comparant ces résultats avec ceux obtenus au point de fonctionnement λ constant à 1,02, nous trouvons un écart de performance d'environ 17% relatifs (13,4% absolus) pour DEV et 19% relatifs (10,1% absolus) pour EVAL. Cette différence absolue de 10% en termes de F-mesure confirme la nécessité de ne pas fixer *a priori* le coefficient de pénalité, quitte à inclure un post-traitement pour obtenir une segmentation *a posteriori*.

4.2 Décision *a posteriori* par consolidation

Afin d'éviter le problème de variabilité et le choix *a priori* du facteur de pénalité, nous avons effectué plusieurs segmentations en faisant varier le coefficient λ sur l'intervalle [0,8 1,2] avec un pas de 0,01. Sont ainsi obtenues 41 segmentations d'un même enregistrement. Pour décider quelles sont les frontières retenues, un vote est effectué sur les candidats obtenus : une frontière sera validée si elle est trouvée par au moins S_0 segmentations parmi les 41, en tolérant un écart de 0,3s. Nous parlerons de Décision Consolidée *A Posteriori* (DCAP). La valeur de S_0 comprise entre 1 et 41 a été déterminée en utilisant les enregistrements du DEV et fixée à 17.

4.3 Performances globales

Les résultats expérimentaux sont rassemblés dans le Tableau **1**. La performance globale obtenue par DCAP est de 50,4% en F-Mesure sur le corpus EVAL. L'augmentation est d'environ 6% relatifs (4,0% absolus) pour le corpus DEV et 15% relatifs (6,7% absolus) pour le corpus EVAL,

par rapport à celles trouvées avec une valeur de λ fixée à 1,02.

Naturellement cette performance reste plus faible que la performance du système « *oracle* » : il existe encore une marge de gain possible de 7% relatifs (3,4% absolus), si l'on considère la F-mesure obtenue avec le système « *oracle* » comme la limite supérieure à atteindre.

Système	DEV			EVAL		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
$\lambda = 1{,}02$	65,6	66,6	66,1	39,7	48,8	43,7
oracle	77,3	81,8	79,5	51,1	56,8	53,8
DCAP	65,7	75,2	70,1	59,3	43,8	50,4

Tableau 1 : Performance globale de notre système de segmentation en tours de chant.

Certains enregistrements du corpus EVAL montrent une F-mesure de presque 80% et d'autres restent à 30%. Les erreurs sur ces enregistrements sont principalement des fausses alarmes : leur écoute révèle beaucoup de superpositions de chanteurs, d'alternances très rapides entre les solistes et le chœur, la présence d'instruments percussifs comme des cloches, des frappements des mains, et du bruit de fond (parole, cris). En outre, ces enregistrements se sont avérés être les plus difficiles à annoter manuellement : dans certains cas d'alternances rapides de chanteurs, il n'est pas évident de décider s'il faut réellement insérer une frontière de segment : cette observation nécessite une analyse en termes d'usage pour comprendre les limites de l'attendu en termes de segmentation et nécessitera une discussion avec les ethnomusicologues.

5 Conclusions et perspectives

Dans cet article, nous avons présenté le problème de la segmentation en tours de chant par analogie aux tours de parole, en vue de traitements ultérieurs, avec comme objectif final l'indexation par le contenu d'enregistrements musicaux ethnomusicologiques. Nous avons appliqué une méthode de segmentation fondée sur le critère BIC. Le choix d'une unique valeur du paramètre de pénalité de ce critère, obtenu par ajustement global sur un corpus de développement, s'est avéré non satisfaisant, puisque les performances variaient sensiblement d'un enregistrement à un autre. Afin d'éviter de choisir une unique valeur, nous avons combiné des segmentations obtenues avec différentes valeurs, et obtenu la segmentation finale en ne gardant que les frontières présentes dans plusieurs d'entre elles. Avec cette dernière méthode, un gain en F-mesure de 15% relatifs a pu être obtenu par rapport au système de référence.

De nouvelles versions de cette segmentation sont testées actuellement, comme la combinaison avec des segmentations réalisées avec d'autres paramètres (RASTA-PLP) ainsi que la calibration/fusion discriminante des scores de ΔBIC des différentes segmentations. Nous souhaitons confirmer nos résultats sur une plus grande quantité de données et également sur des données différentes de celles du projet, comme des enregistrements en studio de musique dans des conditions acoustiques plus contrôlées, afin de préciser les limites de ce type d'approche. Dans un avenir plus lointain, il s'agira d'étudier le problème de regroupement de chanteurs.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0022).

Références

André-Obrecht, R. (1998). A new statistical approach for the Automatic Segmentation of Continuous Speech Signals, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 36-1, pp. 29-40.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic and Control*, AC-19, pp. 716-723.

Anguera, X et Bozonnet, S. et Evants, N. et Fredouille, C. et Friedland, G. et Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20:2.

Cettolo, M. et Vescovi, M. et Rizzi, R. (2005). Evaluation of BIC-based algorithms for audio segmentation. *In Computer Speech And Language*, pp. 147-170.

Chen, S. Scott et Gopalakrishnan, P. S. (1998). Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. *In The DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, pp. 127-132.

Delacourt, P. et Wellekens, C. (2000). DISTBIC: a speaker-based segmentation for audio data indexing. *In Speech Communication*, vol. 32, pp. 111-126.

El-Khoury, E. et Senac, C. et Pinquier, J. (2009). Improved speaker diarization system for meetings. *In Proc. ICASSP*, Taipei, pp. 4097-4100.

Lachambre, H. et André-Obrecht, R. et Pinquier, J. (2009). Singing voice detection in monophonic and polyphonic contexts. *In Proc. European Signal Processing Conference*, Glasgow, pp. 1344-1348.

Le Coz, M. et André-Obrecht, R. et Pinquier, J. (2012). Feasibility of the Detection of Choirs for Ethnomusicologic Music Indexing. *In Proc. International Workshop on Content-Based Multimedia Indexing*, Annecy, pp. 145-148.

Schwarz, Gideon. (1978). Estimating the dimension of a model. *In The annals of Statistic*, vol. 6, pp. 461-464.

Siu, M-H. Yu, G. et Gish, H. (1991). Segregation of speakers for speech recognition and speaker identification. *In Proc. ICASSP*, Toronto, Canado, pp. 873-876,

Zhu, X. et Barras, C. et Meignier, S. et Gauvain, J.-L. (2005). Combining speaker identification and BIC for speaker diarization. *In Proc. INTERSPEECH*, Lisbon, pp. 2441-2444.

Estimation de la source glottique par décomposition modale empirique

Kemiha Mina[1] Abdellah Kacha[1], Mounir Boudjerda[1,2]
(1) Laboratoire de Physique de Rayonnement et Applications, Université de Jijel, Algérie
(2) Welding and NDT Research Centre (CSC), Alger, Algérie
kemihamina@yahoo.fr, kacha_a@yahoo.com, boudjerda@yahoo.com

RESUME

Dans cet article, la décomposition modale empirique est proposée comme alternative pour estimer la source glottique à partir du signal de parole. En utilisant l'algorithme de décomposition modale empirique, le logarithme de l'amplitude du spectre du signal de parole est décomposé en composantes oscillatoires appelées fonctions de modes intrinsèques. Une procédure adaptative est ensuite utilisée pour sélectionner les fonctions de modes intrinsèques appropriés qui constituent l'amplitude dans le domaine log spectral de la source glottique. L'exploitation de l'information de phase obtenue à partir du signal acoustique conjointement avec la somme des fonctions de modes sélectionnées permet d'obtenir une estimation de la source glottique. La méthode proposée est testée sur des signaux de parole synthétiques et comparée avec la méthode d'estimation de la source glottique basée sur le cepstre.

ABSTRACT

In this paper, the empirical mode decomposition algorithm is proposed as an alternative to estimate the glottal source from the speech signal. Using the empirical mode decomposition algorithm, the logarithm of the magnitude spectrum of the speech signal is decomposed into oscillatory modes names intrinsic mode functions. An adaptive procedure is then used to select the appropriate intrinsic mode functions that constitute the magnitude of the glottal source in the log spectral domain. The exploitation of the phase information jointly with the sum of the selected intrinsic mode functions provides an estimate of the glottal source. The proposed method is tested on a synthetic speech signals and compared to the cepstrum-based glottal source estimation method.

MOTS-CLES : décomposition modale empirique, fonctions de modes intrinsèques, estimation de la source glottique.

1 Introduction

La séparation de la parole en ses deux contributions réponse fréquentielle du conduit vocal et excitation glottique est un sujet important en traitement de la parole. Isoler explicitement les deux composantes permet de les modéliser de façon indépendante. La caractérisation de l'excitation glottique présente plusieurs avantages en reconnaissance du locuteur (Plumpe et al. 1999), dans la caractérisation et l'analyse des troubles de la voix (Moore et al. 2003), en reconnaissance de la parole (Yamada, et al. 2002) et en synthèse de la parole (Drugman et al. 2009). Ces raisons justifient la nécessité de développer des algorithmes capables

d'estimer et de paramétrer le signal glottique de manière robuste et fiable.

Bien que les techniques de modélisation du conduit vocal soient assez bien établies, ce n'est pas le cas de la représentation de la source glottique. Certains travaux ont abordé le problème d'estimation de la contribution de la glotte directement à partir de la forme d'onde du signal de parole. La plupart des approches s'appuie sur une première modélisation paramétrique du conduit vocal, puis utilise le filtrage inverse afin d'éliminer l'effet du conduit vocal et obtenir une estimation du signal glottique. Dans (Alku et al. 1994), le modèle discret tout pôle a été utilisé pour modéliser le conduit vocal. La technique itérative adaptive par filtrage inverse décrit dans (Alku et al. 1992) isole le signal source en estimant de manière itérative à la fois les composantes dues au conduit vocal et à la source. Dans (Brookes et Chan, 1994), l'estimation du signal glottique est affinée sur plusieurs cycles glottiques. Une technique non paramétrique basée sur les zéros de la transformée en Z (ZZT) et le cepstre complexe a été proposée dans (Bozkurt et al. 2005 ; Doval et al. 2003). Cette approche repose sur l'observation que la parole est un signal à phase mixte comprenant une composante causale et une composante anti-causale où la composante anti-causale correspond à la phase d'ouverture de la glotte et la composante causale comprend à la fois la fermeture de la glotte et les contributions du conduit vocal.

Récemment, une méthode de décomposition du signal, appelée décomposition modale empirique (EMD), a été introduite pour analyser les données issues de processus non stationnaires et/ou non linéaires (Huang N.E. et al. 1998). La décomposition modale empirique a reçu plus d'intérêt en termes d'applications, d'interprétation et d'amélioration. L'avantage majeur de la décomposition modale empirique est que les fonctions de base sont obtenues à partir du signal lui-même et non fixées a priori comme dans les méthodes d'analyses conventionnelles (transformée de Fourier, transformée en ondelettes, etc.). Dans (Kacha et al. 2012 ; Kacha et al. 2013), la décomposition modale empirique a été proposée pour décomposer le logarithme du spectre du signal de parole en trois composantes qui sont la composante harmonique, la réponse fréquentielle du conduit vocal, et le bruit. Ces composantes ont été utilisées par la suite pour définir un indice acoustique appelé rapport harmonique sur bruit.

Dans cet article, la décomposition modale empirique est proposée comme alternative pour estimer la source glottique à partir du signal de parole. La méthode d'estimation proposée opère dans le domaine log-spectral. L'efficacité de l'approche proposée est évaluée sur la parole synthétique et sa performance est comparée à celle du cepstre complexe (Drugman et al. 2009 ; Drugman et al. 2012) dont l'utilisation impose un ensemble de contraintes sur la fenêtre d'analyse pour que le signal de parole puisse être modélisé par un modèle à phase mixte.

Le reste du papier est organisé comme suit. L'algorithme de décomposition modale empirique est présenté dans la section 2. L'approche d'estimation de la source glottique basée sur la décomposition modale empirique est présentée dans la section 3. Les résultats basés sur les signaux synthétiques sont présentés dans la section 4. Enfin, les conclusions sont données dans la section 5.

2 Décomposition modale empirique

La décomposition modale empirique est une méthode qui repose sur une décomposition adaptée en décrivant localement le signal comme une succession de contributions d'oscillations rapides (hautes fréquences) sur des oscillations plus lentes (basses fréquences) proposée par (Huang N.E. et al. 1998). En raison de sa nature empirique et algorithmique, il n'existe actuellement aucun fondement théorique complet de cette méthode. La méthode de décomposition modale empirique dispose d'un processus de tamisage fini. Les signaux peuvent être décomposés en une série de fonctions de modes intrinsèques (IMF) qui doivent satisfaire à deux conditions : (i) le nombre de passages par zéro et le nombre d'extrema sont égaux ou ne diffèrent pas plus d'un, (ii) la valeur moyenne de l'enveloppe produite par les extrema locaux est égale à zéro. Dans ce procédé, les fonctions de modes intrinsèques sont essentiellement obtenues à partir de points extrema locaux qui représentent des échelles de temps caractéristiques du signal. Au moyen de l'EMD, le signal à analyser est décomposé en N modes empiriques et un résidu. Les IMFs représentent les oscillations locales intrinsèques incrustées dans le signal, le mode résiduel peut être une fonction monotone ou d'énergie très faible peut être ignoré. Lorsque le mode résiduel satisfait à cette condition, les différentes étapes du procédé de tamisage peuvent s'écrire sous la forme du pseudo code suivant :

Etape 1 : initialisation $j \leftarrow 1\ (j^{ième}\ IMF), r_{j-1}(t) \leftarrow x(t)$(résidu), Fixer le seuil Δ.

Etape 2 : extraire la $j^{ième}\ IMF$:

(a) $h_{j,i-1}(t) \leftarrow r_{j-1}(t),\ i \leftarrow 1$ (i, itération de la boucle de tamisage)

(b) extraire les maxima et minima locaux de $h_{j,i-1}(t)$

(c) calculer les enveloppes supérieures et inférieures : $U(t)$ et $L(t)$ par interpolation entre les maxima et les minima locaux $h_{j,i-1}(t)$ respectivement.

(d) calculer l'enveloppe moyenne : $\mu(t) \leftarrow (\ U(t) + L(t))/2$

(e) mettre à jour : $h_{j,i}(t) \leftarrow h_{j,i-1}(t) - \mu(t), i \leftarrow i + 1$.

(f) calculer le critère d'arrêt (par exemple) $SD = \sum_{t=0}^{T} \frac{|h_{j,i-1}(t) - h_{j,i}(t)|^2}{(h_{j,i-1}(t))^2}$

où T représente la durée totale du signal.

(g) décider : répéter l'étape (b)-(f) jusqu'à ce que $SD < \Delta$ et alors mettre $IMF_j(t) \leftarrow h_{j,i}(t)\ \ (j^{ième} IMF)$.

Etape 3 : mettre à jour le résidu : $r_j(t) \leftarrow r_{j-1}(t) - IMF_j(t)$.

Etape 4 : répéter l'étape 2 avec $j \leftarrow j + 1$ jusqu'à ce que le nombre d'extrema dans $r_j(t)$ soit inférieur à 2.

Le processus de reconstruction du signal est donné par l'équation suivante,

$$x(t) = \sum_{j=1}^{N} IMF_j(t) + r_N(t), N \in N^*$$

3 Estimation de la source glottique

3.1 Principe

Selon le modèle source-filtre de production de la parole, le signal de parole peut être considéré comme le résultat de la convolution de l'excitation de l'appareil vocal (excitation glottique) et de sa réponse impulsionnelle (Deller et al. 1993) :

$$x(t) = e(t) * v(t)$$

où $x(t)$ est le signal de parole, $v(t)$ est la réponse impulsionnelle du système modélisant le conduit vocal, et $e(t)$ est le signal d'excitation ayant pour origine les cordes vocales, et $*$ désigne le produit de convolution. La transformée de Fourier du signal transforme la convolution en produit décrit par l'équation suivante

$$X_w(f) = E_w(f) \times V(f)$$

où f dénote la fréquence, $X_w(f)$ *et* $E_w(f)$ sont, respectivement, les spectres d'amplitude des trames d'analyse du signal et de l'excitation glottique pondérées par la fenêtre $w(t)$ et $V(f)$ est la réponse fréquentielle du conduit vocal.

En prenant le logarithme complexe des deux membres de l'équation, il vient :

$$\log(X_w) = log|X_w(f)| + j\angle X_w$$

où $\angle X_w$ dénote la phase de $X_w(f)$. La partie réelle du logarithme complexe représente le spectre d'amplitude du signal et la partie imaginaire représente la phase du signal. La phase doit être corrigée en ajoutant des multiples de$\pm 2\pi$ (Stark et al. 2011).

3.2 Estimation du spectre d'amplitude de la source glottique

Le spectre d'amplitude d'une trame pondérée du signal de parole peut être écrit comme

$$|X_w(f)| = |E_w(f) \times v(f)|$$

En prenant le logarithme des deux membres de l'équation, il vient :

$$log|X_w(f)| = log|E_w(f)| + \log|v(f)|$$

On observe que le logarithme du spectre d'amplitude d'une trame pondérée du signal de parole est la somme de deux composantes spectrales : le logarithme du spectre d'amplitude de l'excitation pondérée $log|E_w(f)|$, et l'enveloppe spectrale $log|v(f)|$ (Kacha et al. 2012). Le logarithmique du spectre d'amplitude du signal de parole voisée peut être considéré comme constitué d'une variation lente (par rapport à la fréquence) représentant le contour due à la contribution du conduit vocal et d'une série d'harmoniques, caractérisée par une structure périodique. L'algorithme de décomposition modale empirique donne un outil efficace qui permet de séparer les deux composantes du spectre d'amplitude. En effet, l'algorithme de décomposition modale empirique agit comme un banc de filtres (Flandrin et

al. 2004), de sorte que la décomposition du logarithme du spectre d'amplitude résulte en plusieurs composantes oscillantes (IMF) qui peuvent être regroupés en deux catégories (classes) où chaque classe de composants est associée à une partie du spectre d'amplitude.

Il a été montré que la variance des IMFs pour les signaux de parole diminue de manière significative après la quatrième IMF au fur et à mesure que l'ordre des IMFs augmente (Chatlani et al. 2012). Il a été constaté expérimentalement que pour le signal de parole, les statistiques des IMFs sont caractérisées par un pic d'énergie à une IMF d'ordre élevé. Cette propriété est utilisée pour sélectionner l'indice optimal qui permet de séparer la composante harmonique et l'enveloppe spectrale. Les différentes étapes de la méthode de séparation de la composante harmonique et de l'enveloppe spectrale illustrées par la figure 1 peuvent être résumées comme suit (Chatlani et al. 2012) :

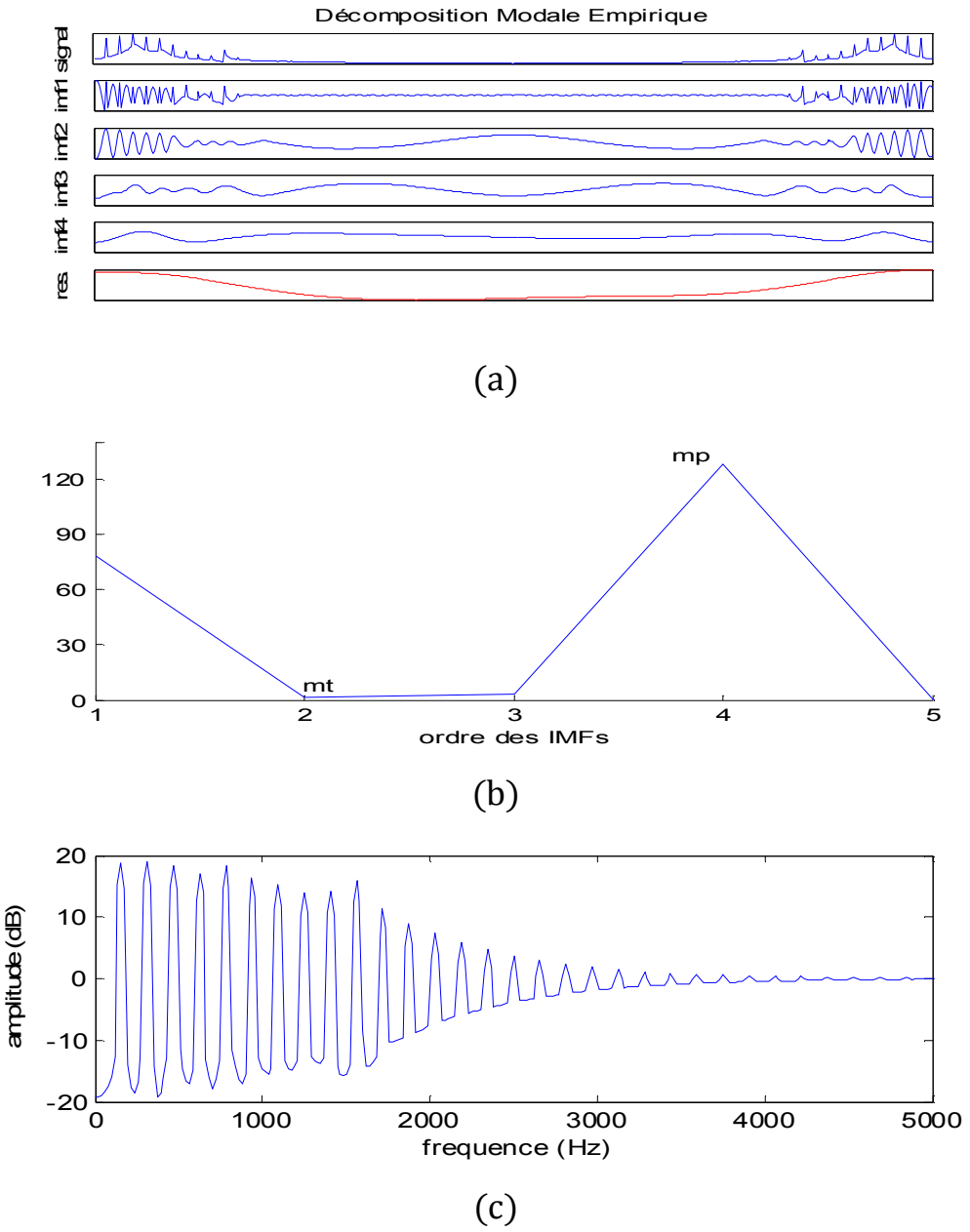


FIGURE 1 – Illustration de l'algorithme d'estimation du logarithme du spectre d'amplitude de la source glottique: (a) Décomposition modale empirique du logarithme du spectre d'amplitude d'une trame extraite d'une voyelle /a/ synthétique. (b) Variance des IMFs. (c) Logarithme du spectre d'amplitude de la source glottique.

1. Calculer la variance V de chaque IMF (Figure 1 (a)).

2. Identifier l'indice du maximum, mp dans V pour ordre de l'IMF supérieur a 4.

3. Identifier l'indice du minima mt.

4. Calculer mb tel que $mb = mp - mt$.

5. Déterminer d'indice M avec $M = mp - mb$.

Le logarithme du spectre d'amplitude de la source glottique (Figure 1 (c)) est estimé par la relation suivante:

$$log|E_w(f)| = \sum_{j=1}^{M-1} IMF_j(f)$$

3.3 Estimation de la source glottique

L'approche d'estimation de la source glottique combine le logarithme du spectre d'amplitude de la source glottique et la phase estimée comme la partie imaginaire du logarithme complexe du spectre du signal de parole. L'approche d'estimation est illustrée par la figure 2. Le filtre passe bas a pour rôle d'éliminer l'effet de la fenêtre de pondération $w(t)$.

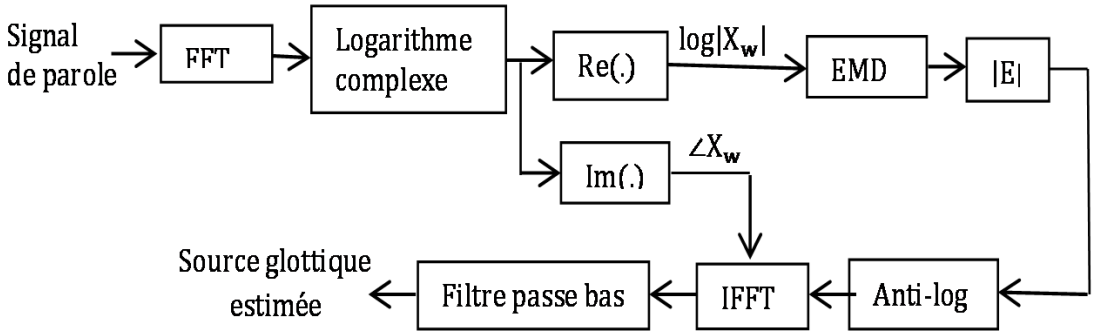
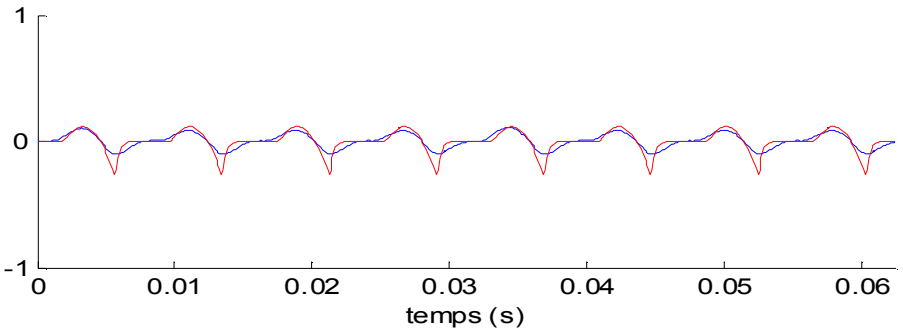


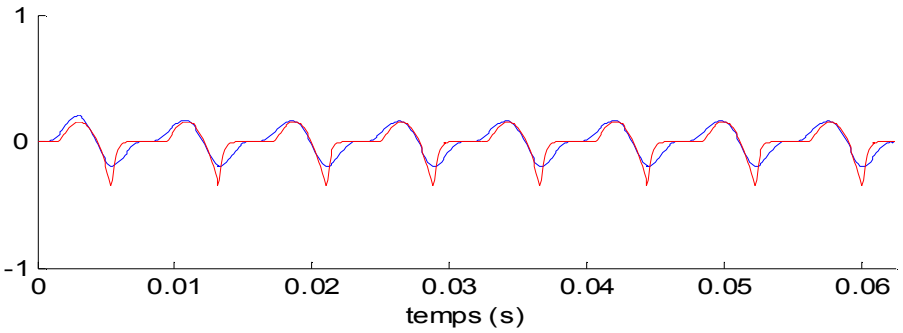
FIGURE 2 – Estimation de la source glottique.

4 Résultats et discussions

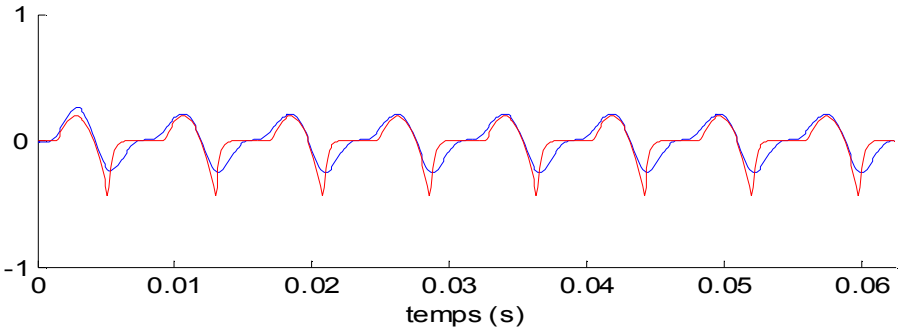
Évaluer objectivement et quantitativement une méthode d'estimation du signal glottique exige de travailler avec des signaux synthétiques, puisque la véritable source n'est pas disponible pour les signaux de parole réelle. Dans ce travail, le signal artificiel utilisé dans le test est la voyelle synthétique /a/ de durée 1 seconde et de fréquence fondamental f0=128 Hz générée selon le modèle source-filtre de production de la parole. La fréquence d'échantillonnage de l'ensemble des signaux de parole utilisés dans l'expérience est de 20 kHz. Le modèle source-filtre est constitué d'une source qui génère un train d'impulsions périodique modélisant le flux d'air glottique et un conduit vocal modélisé comme un filtre



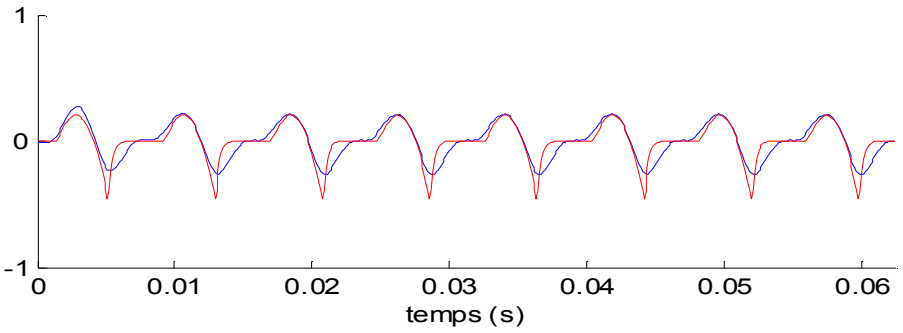
L=256



L=512



L=1024



L=2048

FIGURE 3 – Forme d'onde de la source glottique estimée par la méthode proposée superposée au modèle de l'excitation glottique pour différentes longueurs L de la fenêtre de pondération.

tout-pôle, caractérisé par trois pôles (L. R. Rabiner. 1968; Deller et al. 1993) correspondant aux fréquences de formants 981,6 Hz, 1631,3 Hz et 3165,9 Hz présentant des bandes fréquentielles de 140 Hz, 180 Hz et 55 Hz, respectivement. Le rayonnement aux lèvres est modélisé par un dérivateur du premier ordre $R(z) = 1 - z^{-1}$.

Le signal de parole a été divisé en k trames sans chevauchement en utilisant une fenêtre de Hamming et la source glottique est estimée pour chaque trame. Afin d'étudier les performances de la méthode proposée, l'estimation de la source glottique a été effectuée pour différentes tailles de la fenêtre. Les résultats obtenus sont présentés par la figure 3. La décomposition modale empirique permet d'avoir des estimations précises de la source glottique quelle que soit la longueur de la fenêtre utilisée. La figure 4 montre le modèle de la source glottique utilisée ainsi que les estimations basées sur la décomposition modale empirique et le cepstre complexe pour une trame de longueur 1024 échantillons. Il est observé que l'estimation basée sur la décomposition modale empirique est plus précise que celle basée sur le cepstre complexe.

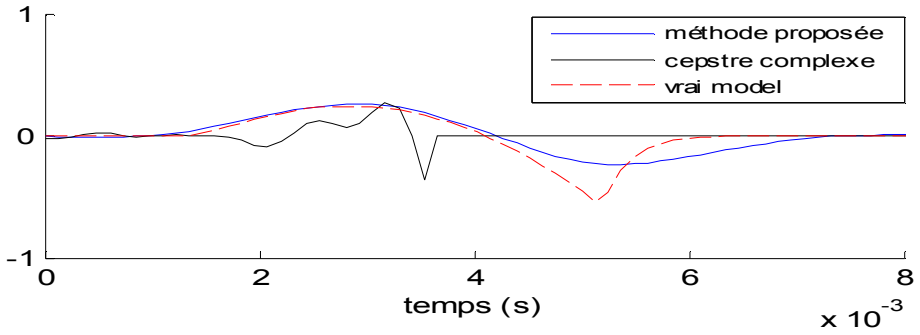


FIGURE 4 –Forme d'onde de la source glottique estimée par la méthode proposée comparée à celle estimée par le cepstre complexe et au vrai modèle pour une trame de longueur L=1024.

5 Conclusion

Dans cette présentation, l'algorithme de décomposition modale empirique a été proposé comme alternative pour estimer l'excitation glottique à partir du signal de parole. La performance de la méthode proposée a été comparée à celle de l'estimation basée sur le cepstre complexe. La méthode proposée est simple et systématique. Les résultats obtenus montrent que la méthode proposée fournit une estimation précise de l'excitation glottique aussi bien pour les trames de courte durée que pour les trames de longue durée.

Références

L. R. Rabiner. (1968). Digital-Formant Synthesizer for Speech-Synthesis Studies. *J. Acoust. Soc. Amer*, 43(4): 822-828.

Alku, J. Svec, E. Vilkman, F. Sram. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3): 109-118.

J. H. Deller, J. G. Proakis, J. H. L. Hansen. (1993). *Discrete-time processing of speech signals*. Prentice-Hall.

D. Brookes, D. Chan. (1994). Speaker characteristics from a glottal airow model using glottal

inverse filtering. *Proc. Institue of Acoust*, 15:501-508.

P. Alku, E. Vilkman. (1994). Estimation of the glottal pulseform based on discrete all-pole modeling. *Third international Conference on Spoken Language Processing*: 1619-1622.

Huang N.E. et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proc. R. Soc. London Ser. A* ,454: 903-995.

M. Plumpe, T. Quatieri, D. Reynolds. (1999). Modeling of the glottal flow derivative wave form with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7: 569-586.

D. Yamada, N. Kitaoka, S. Nakagawa. (2002). Speech Recognition Using Features Based on Glottal Sound Source. *Trans. of the Institute of Electrical Engineers of Japan*, 122(12): 2028-2034.

B. Doval, C. dAlessandro, N. Henrich. (2003). The voice source as a causal/anticausal linear filter. *Proceedings ISCA ITRW VOQUAL03*:*15-19*.

E. Moore, M. Clements, J. Peifer, L.Weisser. (2003). Investigating the role of glottal features in classifying clinical depression. *Proc. of the 25th International Conference of the IEEE Engineering in Medicine and Biology Society*, 3:2849-2852.

P. Flandrin, G. Rilling, and P. Goncalves. (2004). Empirical mode decomposition as a filter bank. *IEEE Sig. Proc. Lett*, 11(2): 112–114.

B. Bozkurt, B. Doval, C. DAlessandro, T. Dutoit. (2005). Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech. *IEEE Signal Processing Letters*, 12(4): 2005.

T. Drugman, G. Wilfart, A. Moinet, T. Dutoit. (2009). Using a pitch-synchronous residual for hybrid HMM/frame selection speech synthesis. *IEEE International Conference on Acoustic Speech and Signal* Processing,ICASSP 2009: 3793-3796.

T. Drugman, B. Bozkurt, T. Dutoit, (2009), Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation*. Interspeech 2009*: 116-119.

M. Stark, M. Wohlmayr, F. Pernkopf. (2011). Source–Filter-Based Single-Channel Speech Separation Using Pitch Information. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 19(2):242-255.

N. Chatlani, J. Soraghan. (2012). EMD-Based Filtering (EMDF) of Low-Frequency Noise for Speech Enhancement. *IEEE Trans. audio, speech, and lang. proc*,20(4): 1158-1166.

A. Kacha, F. Grenez, J. Schoentgen. (2012). Assessment of Disordered Voices Using Empirical Mode Decomposition in the Log-Spectral Domain. *Interspeech 2012.*

T. Drugman, B. Bozkurt, T. Dutoit. (2012). A comparative study of glottal source estimation techniques. *Computer Speech & Language* ,26(1): 20-34.

A. Kacha, F. Grenez, J. Schoentgen. (2013). Empirical Mode Decomposition-Based Spectral Acoustic Cues for Disordered Voices Analysis, *Interspeech 2013.*

Décodage hybride dans les SRAP pour l'indexation automatique des documents multimédia

Mohamed BOUAZIZ[1, 2] Antoine LAURENT[1, 3] Yannick ESTEVE[1]
(1) LIUM, Le Mans, France
(2) LIA, Avignon, France
(3) LIMSI-CNRS, Orsay, France
mohamed.bouaziz@alumni.univ-avignon.fr, antoine.laurent@limsi.fr,
yannick.esteve@lium.univ-lemans.fr

RÉSUMÉ

Certains Systèmes de Reconnaissance Automatique de la Parole (SRAP) atteignent des taux d'erreur de l'ordre de 10%. Toutefois, notamment dans le cadre de l'indexation automatique des documents multimédia sur le web, les SRAP se trouvent face à la problématique des mots hors-vocabulaire. En effet, les entités nommées en constituent une grande partie et sont remarquablement importantes pour les tâches d'indexation. Nous mettons en œuvre, dans ce travail, la solution du décodage hybride en utilisant les syllabes comme unités sous-lexicales. Cette méthode est intégrée au sein du SRAP LIUM'08 développé par le Laboratoire d'Informatique de l'Université du Maine. Avec une légère dégradation de la performance générale du système, environ 31% des noms de personne hors vocabulaire sont correctement reconnus.

ABSTRACT

Hybrid decoding in ASR systems for multimedia documents auto-indexing

Some state of the art Automatic Speech Recognition (ASR) systems reach even about 10% in word error rate. Yet, these systems suffer from out-of-vocabulary (OOV) words particularly in the task of auto-indexing of multimedia documents on the web. Indeed, not only do named entities take a big part of OOVs, but they are also of significance in multimedia documents indexing. In this work, we implement a syllable-based hybrid decoding approach within the LIUM'08 ASR system. Despite a small decrease in the system's general performance, about 31% of proper nouns are correctly recognized.

MOTS-CLÉS : mot hors-vocabulaire, Indexation automatique, vocabulaire ouvert, décodage hybride, unité sous-lexicale.

KEYWORDS: out-of-vocabulary, auto-indexing, open vocabulary, hybrid decoding, sub-lexical unit.

1 Introduction

Durant les dernières décennies, l'information audio a pris une place importante parmi les interfaces de communication. Parallèlement à ce progrès, le traitement de ces grandes quantités de

données est devenu indispensable. Selon (Srinivasan et Brown, 2002), le domaine d'application de la reconnaissance de la parole se base principalement sur deux axes. Le premier axe concerne l'utilisation de la parole comme entrée, notamment, pour les systèmes de dictée vocale, les systèmes de navigation, les applications à but commercial, etc. Le deuxième champ d'application présente la parole comme une source de données ou de connaissances. Les Systèmes de Reconnaissance Automatique de la Parole (SRAP) sont par exemple utilisés pour transcrire les documents multimédia mis à disposition sur le web afin d'améliorer les systèmes d'indexation de ce type de documents. Dans ce dernier contexte, les SRAP se trouvent face au problème des mots hors-vocabulaire à cause du changement continu du vocabulaire à traiter. En effet, les SRAP à large vocabulaire modélisent une langue par le moyen d'un vocabulaire de taille fixe. Ces systèmes ne peuvent pas, en conséquence, couvrir la totalité des mots prononcés.

Les mots hors-vocabulaire sont en partie responsables de la dégradation de la performance d'un SRAP. Non seulement ces mots ne sont pas reconnus, mais aussi ils influent négativement sur la reconnaissance des mots voisins. En effet, parmi les dix mots voisins d'un mot hors-vocabulaire, moins de cinq mots, en moyenne, sont correctement reconnus (Dufour, 2008). Par conséquent, cette anomalie aura une influence sur tout traitement potentiel (traduction automatique de la parole, indexation automatique de documents, etc.) (Bisani et Ney, 2005).

Dans les deux sections suivantes, nous présentons les différents travaux qui s'articulent autour des finalités ci-dessus introduites. Ensuite, nous exposons, dans les sections 4 et 5, notre contribution qui consiste à intégrer la solution du « décodage hybride » dans le SRAP développé au sein du Laboratoire d'Informatique de l'Université du Maine (LIUM). Enfin, la section 6 est réservée pour la présentation des résultats de la mise en œuvre de notre contribution visant à remédier à la problématique des mots hors-vocabulaire.

2 Décodage hybride

Les SRAP à vocabulaire ouvert permettent de pallier la couverture partielle des mots d'une langue. Une des techniques du vocabulaire ouvert consiste à utiliser un modèle de langage basé sur des unités sous-lexicales (syllabes, morphèmes, phonèmes,...) et non pas sur des mots entiers. La combinaison de cette solution avec l'approche classique correspond au « décodage hybride ».

(Bisani et Ney, 2005) décomposent les mots du vocabulaire en des unités sous-lexicales dénommées « graphonèmes » qui combinent un graphème avec sa prononciation. Ces unités sont ensuite insérées au sein du même vocabulaire pour former une modélisation linguistique hybride. En adoptant ces nouveaux concepts sur des données de la langue anglaise, les auteurs réussissent à atteindre une réduction relative de 30% du taux d'erreur de mots pour des corpus d'évaluation comportant un taux de mots hors-vocabulaire supérieur à 10%. En outre, (Bisani et Ney, 2005) ont montré que le nouveau système reconnaît correctement en moyenne un mot de plus, parmi les mots voisins du mot hors-vocabulaire, par rapport au système de reconnaissance de base. Les travaux présentés dans (Shaik *et al.*, 2011) représentent une extension du système construit par (Bisani et Ney, 2005). En effet, la décomposition des mots n'est plus limitée aux graphonèmes qui doivent avoir un nombre maximal de lettres. Les mots du vocabulaire sont ainsi décomposés en syllabes ou en morphèmes en utilisant des outils spécialisés. Ce système opère sur des données en langue allemande et atteint une réduction relative de 5% en taux d'erreur de mots par rapport au système de base. Ce système reconnaît 40% des mots désignés comme hors-vocabulaire qui

représentent 2,3% du corpus de test. En appliquant des techniques d'apprentissage par unités sous-lexicales sur des données d'entraînement de petite taille de l'amharique, une langue peu-dotée, (Gelas *et al.*, 2012) réussissent à reconnaître jusqu'à 75% des mots hors-vocabulaire. Ce système est testé sur des données de test contenant 9% de mots hors-vocabulaire et arrive à réduire de 50,3% le taux d'erreur de mots.

La majorité des travaux s'intéressant à la reconnaissance des mots hors-vocabulaire adoptent les syllabes et/ou les morphèmes comme des unités sous-lexicales (par exemple (Zablotskiy *et al.*, 2012) pour le russe, (Rotovnik *et al.*, 2007) pour le slovène, (Gelas *et al.*, 2012) pour le swahili, etc.). En revanche, certains travaux utilisent des unités sous-lexicales de taille plus petite. Par exemple, (Bazzi et Glass, 2000) présentent une stratégie de construction des mots hors-vocabulaire en partant d'une suite de phonèmes. Pour ce faire, les auteurs utilisent un modèle de langage en bi-gramme afin de modéliser les contraintes phonotactiques. Avec un léger taux de fausse alarme, la moitié des mots hors-vocabulaire sont correctement reconnus.

Divers travaux s'intéressent ainsi au traitement des mots hors-vocabulaire par le moyen du décodage hybride. Malgré le grand nombre de langues concernées, aucune tentative, au moins à notre connaissance, n'a encore mis en œuvre cette solution dans le cadre de la langue française.

3 Indexation automatique des documents multimédia

Nous avons évoqué, dans l'introduction, la possibilité d'avoir recours aux SRAP dans le but de raffiner les systèmes d'indexation des documents multimédia. Un des premiers essais dans ce contexte est SpeechBot, un moteur de recherche développé par (Logan *et al.*, 1996) dont le système d'indexation se base sur la transcription automatique des documents audio. Malgré la faible qualité des transcriptions, le système arrive selon (Van Thong *et al.*, 2002) à satisfaire 77.5% des requêtes effectuées sur des pages web contenant des émissions radio. En ce qui concerne le traitement des mots hors-vocabulaire, (Logan *et al.*, 2002) étendent leur système en utilisant la technique du décodage hybride. Dans un premier temps, les documents audio sont transcrits par un SRAP dont le modèle de langage est appris sur une combinaison de mots entiers et d'unités sous-lexicales. Ensuite, les mots des requêtes sont découpés avant leur exécution. Les phonèmes et des unités semblables aux syllabes, dénommés particules, sont utilisés comme unités sous-lexicales. L'utilisation de ces techniques, selon les mêmes auteurs, contribue à une légère amélioration dans la précision et le rappel des requêtes comportant des mots hors-vocabulaire. En revanche, elle cause une augmentation relativement importante dans le taux de fausses alarmes.

Dans un contexte similaire, une autre tentative de traitement des mots hors-vocabulaire, décrite dans (Allauzen et Gauvain, 2005), consiste à introduire dynamiquement de nouveaux mots au vocabulaire du système sans adapter le modèle de langage. Ces mots sont extraits à partir d'un ensemble de métadonnées relatives à des documents multimédia. Le nouveau système réussit non seulement à réduire le taux de mots hors-vocabulaire de 30% mais aussi à reconnaître 84% des entités nommées nouvellement introduites dans le vocabulaire.

4 Méthodologie

Les travaux exposés ci-dessus ne reprennent pas la solution du décodage hybride de la même façon. Nous présentons, dans cette section, les principales décisions que nous étions amenés à prendre dans la conception de notre système.

4.1 Choix des mots hors-vocabulaire

L'apparition des mots hors-vocabulaire peut provenir de plusieurs facteurs. D'une part, (Shaik *et al.*, 2011) affirment que ces mots apparaissent plus souvent dans le traitement des langues morphologiquement riches. Par exemple, dans des langues comme le français, l'arabe et l'allemand, la majorité des mots possèdent une grande variabilité. D'autre part, et particulièrement dans une tâche traitant de la reconnaissance de documents multimédia du web, les données à traiter peuvent provenir de diverses sources (journaux parlés ou télévisés, débats d'actualité, vidéos d'amateurs,...). Dans ce cas, le caractère dynamique du vocabulaire utilisé se manifeste par une émergence continue de nouveaux mots appartenant à la catégorie des entités nommées (noms propres, noms de villes, etc.). Cette catégorie constitue, selon (Réveil *et al.*, 2013), une grande partie des mots hors-vocabulaire. Or, ce type de mots est d'une grande importance parmi les index sur lesquels se base un moteur de recherche. Ainsi, nous nous concentrons au sein de ce travail aux entités nommées, et spécifiquement, aux noms de personne. Enfin, nous envisageons que cette approche permettra au nouveau système de composer les noms de personne hors-vocabulaire en combinant la séquence d'unités sous-lexicales correspondante.

4.2 Choix de l'unité sous-lexicale

À travers l'étude de l'état de l'art que nous avons abordée sur l'utilisation du décodage hybride, nous remarquons que le choix de l'unité sous-lexicale représente un élément décisif dans les résultats potentiels du décodage. Cela étant, les morphèmes et les syllabes sont les unités les plus utilisées dans ce contexte.

En ce qui concerne les morphèmes, ce choix participe bien à l'amélioration de la reconnaissance des mots hors-vocabulaire, notamment dans (Gelas *et al.*, 2012) et (Shaik *et al.*, 2011). En revanche, cette solution est mise en œuvre afin de traiter plutôt de la richesse morphologique des langues flexionnelles. En effet, elle permet au SRAP de synthétiser plus efficacement les mots qui disposent d'une large variabilité grammaticale en partant d'un lemme et d'une combinaison de terminaisons grammaticales. Ainsi, nous avons plutôt besoin d'une méthode qui sera, autant que possible, adaptée à la reconnaissance des entités nommées.

Pour ce qui est des phonèmes, ce choix s'annonce prometteur dans le sens où chaque langue possède un nombre fini de phonèmes. Cependant, deux problèmes se manifestent dans ce cas. Premièrement, la transcription de la parole produit une séquence de phonèmes. Pour avoir une transcription d'un signal sonore en mots, une deuxième phase est alors nécessaire. En effet, après le décodage, il faut reconstruire les graphèmes (suite de lettres produisant un phonème) à partir des séquences de phonèmes en sortie. En outre, un deuxième souci se manifeste à cause du manque de contraintes au sein d'un modèle de langage appris sur des phonèmes vu la petite taille des unités sous-lexicales utilisées.

Partant de cette analyse, nous choisissons d'adopter les syllabes en tant qu'unités sous-lexicales. Nous envisageons ainsi à travers cette stratégie d'avoir des unités dont la combinaison est susceptible de former les mots à retrouver tout en assurant un nombre suffisant de contraintes.

5 Protocole expérimental

Afin de mettre en œuvre nos idées, nous partons du SRAP du LIUM (Deléglise *et al.*, 2009), le meilleur SRAP open-source dans la campagne d'évaluation ESTER 2, basé sur le système CMU Sphinx. Les différents modèles du LIUM'08 sont appris à partir de données provenant de plusieurs origines (ESTER1, ESTER2, EPAC, articles de journaux, sites internet). Ce système utilise 39 paramètres acoustiques issus d'une paramétrisation de type PLP. Les modèles acoustiques dépendent du type du canal et sont adaptés au genre du locuteur par la méthode Maximum A Posteriori (MAP). En se basant sur un vocabulaire de 122k mots, le modèle de langage est appris à l'aide de la boîte à outils SRILM (Stolcke *et al.*, 2002) et en recourant à la méthode de discounting Kneser-Ney (Kneser et Ney, 1995). Avec 35 types de phonèmes et 5 types de fillers, le dictionnaire phonétique est construit à l'aide du dictionnaire de phonétisation BDLEX (Perennou et Calmes, 1987) et l'outil de phonétisation LIA_PHON (Béchet, 2001). Le système de segmentation et de regroupement en locuteurs (Meignier et Merlin, 2010), basé sur le Critère d'Information Bayésien, a obtenu le meilleur taux d'erreur lors de la compagne d'évaluation ESTER 2. Enfin, un processus de décodage en 5 passes est utilisé pour transcrire le signal de la parole.

Dans la section précédente, nous avons choisi de prendre les syllabes en tant qu'unités sous-lexicales. Nous passons maintenant à la mise en œuvre de notre approche. La campagne REPERE [1] étant conçue pour évaluer les systèmes de reconnaissance des noms de personne, les transcriptions du corpus fourni dans le cadre de cette compagne sont enrichies par une annotation de ces noms. Ainsi, nous tirons parti de ces annotations pour apprendre, optimiser et évaluer les nouveaux modèles de langage. Les données correspondent à 7 programmes télévisés diffusés sur les chaînes BFM TV et LCP. 42 heures du corpus REPERE servent à l'apprentissage (R_train), 9 heures à l'optimisation (R_dev) et 9 heures au test (R_test).

Concernant le système de base (S0), nous enrichissons le corpus d'apprentissage avec le corpus R_train et nous introduisons les nouveaux mots présents dans ce dernier dans le vocabulaire de base. En revanche, pour ce qui est de l'apprentissage du modèle de langage hybride (S1), nous découpons les noms de personne en syllabes au niveau de R_train, par le moyen de la fonction de syllabation fournie par l'outil LIA_PHON. Ce corpus compte 7947 occurrences relatives à 1225 noms de personne et le découpage en syllabes de ces mots produit 1074 syllabes différentes. Nous désignons par "nom de personne", dans notre protocole expérimental, un nom ou un prénom d'une personne. Ensuite, nous ajoutons le nouveau R_Train au corpus d'apprentissage et nous introduisons les syllabes générées dans le vocabulaire de base. Afin d'optimiser le modèle, nous utilisons le corpus de développement R_dev après y avoir découpé, de la même manière, les noms de personne. Une deuxième expérience (S2), à laquelle nous recourons, consiste à marquer les syllabes résultant du découpage des noms de personne, par l'étiquette « SyllEtiqu », dans R_train et R_dev. En effet, nous envisageons par cette technique que notre système soit plus apte à faire une distinction entre les mots et les syllabes qui ont une même orthographe. La taille du vocabulaire dans toutes les expériences que nous mettons en œuvre ne dépasse pas 124k mots.

1. www.defi-repere.fr

La phonétisation des nouvelles entrées du vocabulaire est effectuée en utilisant l'outil de phonétisation LIA_PHON. Toutes les variantes de prononciation produites sont prises en considération. En ce qui concerne les syllabes employées dans le système appris au cours de la deuxième expérience, les étiquettes sont temporairement enlevées pour effectuer la phonétisation automatique. Elles sont ensuite reprises pendant l'intégration des syllabes correspondantes dans le nouveau dictionnaire phonétique.

La dernière étape consiste à concevoir une stratégie afin d'évaluer les sorties du nouveau système. Ce travail a pour finalité d'améliorer le rendement des systèmes d'indexation automatique des documents multimédia par rapport aux noms de personne. Nous nous intéressons ainsi aux requêtes, saisies par l'utilisateur, qui comportent un ou plusieurs noms de personne considérés comme hors-vocabulaire pour le SRAP utilisé. En revanche, nous considérons le cas où un moteur de recherche, disposant d'un tel système d'indexation, n'est pas apte à repérer les noms de personne parmi les mots de la requête. Ainsi, nous supposons que ledit moteur découpe tous les mots d'une requête en syllabes avant d'effectuer la recherche dans les index basés sur les transcriptions automatique. Dès lors, nous procédons à une stratégie d'évaluation particulière. D'une part, nous découpons en syllabes les mots des hypothèses de transcription en tenant compte de la norme de présentation des fichiers d'hypothèse produits en format ctm[2] (Time Marked Conversation). D'autre part, nous préparons la transcription de référence des données de test (R_test), fournie en format stm [2] (Segment Time Mark), en découpant chaque mot en syllabes. Ces formats, ainsi que les évaluations effectuées, se reposent sur le moteur d'alignement « sclite » [3] (Score-Lite) fourni par NIST (National Institute of Standards and Technology). En adoptant ces considérations, la problématique de reconstruction des mots à partir des séquences de syllabes ne fait donc pas partie de nos préoccupations. Les données de test comptent 279 noms de personne différents. Parmi ces mots, 27 sont originellement hors-vocabulaire. Par ailleurs, 13 parmi les noms de personne de R_test n'existent pas dans le vocabulaire du nouveau système mais sont présents en syllabes dans le corpus R_train. Le reste des noms de personne de R_test sont artificiellement mis hors vocabulaire.

6 Résultats

Le tableau 1 récapitule les résultats obtenus en ce qui concerne les taux de reconnaissance ou d'erreur de syllabes pour le système de base ainsi que pour les deux systèmes reposant sur la stratégie du décodage hybride.

Système	TR de NP HV	TR de syllabes dans les NP HV	SER
Système de base (S0)	0%	21,05%	18,70%
Système avec syllabes (S1)	**31,39%**	42,82%	**19,90%**
Système avec syllabes étiquetées (S2)	31,10%	**42,95%**	24,90%

TABLE 1: Résultats de reconnaissance (TR : taux de reconnaissance, NP : nom de personne, HV : hors-vocabulaire, SER : Taux d'erreur de syllabes)

2. www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/infmts.htm
3. www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

Partant du fait que les noms de personne évalués n'appartiennent pas au vocabulaire, le taux de reconnaissance des noms de personne hors-vocabulaire (HV) (nombre de noms de personne HV dont toutes les syllabes sont correctement reconnues / nombre de noms de personne HV) dans le système de base est nul. Comme décrit dans le tableau 1, ce système reconnaît tout de même 21,05% des syllabes au sein de ces mêmes noms de personne. Ce taux correspond au nombre de syllabes correctement reconnues dans l'ensemble des noms de personne HV divisé par le nombre total de syllabes dans l'ensemble des noms de personne HV.
Le système S1, appris sur des noms de personne en syllabes, réussit à reconnaître 31,39% des noms de personne et 42,82% des syllabes dans l'ensemble de ces noms de personne. En revanche, il cause une augmentation, de 1,2% en absolue, du taux d'erreur général de syllabes (SER). Cette légère dégradation de la performance générale du système est négligeable par rapport au gain que nous obtenons par rapport à la reconnaissance des noms de personne, qui sont d'une grande importance pour les systèmes d'indexation. Quant au système S2, avec une augmentation remarquable dans le SER et une légère diminution du taux de reconnaissance de noms de personne, ce système réussit tout de même à obtenir un meilleur taux de reconnaissance de syllabes dans les noms de personne (42,95%).

Enfin, nous n'avons pas eu la possibilité de comparer les résultats de cette mise en œuvre du décodage hybride avec d'autres travaux s'intéressant à la reconnaissance des mots hors-vocabulaire. En effet, à notre connaissance, aucun des travaux de l'état de l'art n'a adopté une stratégie d'évaluation se focalisant sur l'étude de la performance des SRAP par rapport à la reconnaissance des syllabes au sein des noms de personne hors-vocabulaire.

7 Conclusion

Prenant en compte les particularités des systèmes d'indexation automatique des documents multimédia, ce travail constitue, à notre connaissance, un des premiers travaux qui essayent de pallier le problème des mots hors-vocabulaire par le moyen du décodage hybride au sein des SRAP traitant de la langue française. Malgré la diversité et l'importance des choix à faire, la mise en oeuvre de l'approche proposée réussit à reconnaître environ un tiers des noms de personne hors-vocabulaire.
L'intérêt de ces résultats se manifeste, entre autres, dans deux aspects. D'une part, l'énonciation de mots appartenant à la catégorie des entités nommées, dans les données multimédia disponibles sur le web, est une des causes fondamentales de l'apparition des mots hors-vocabulaire. D'autre part, cette catégorie de mots a bien une grande importance pour les systèmes d'indexation traitant de ces données.
Enfin, il est envisagé d'étendre ces travaux en prenant en considération les autres sous-catégories des entités nommées. Par ailleurs, le traitement des mots hors-vocabulaire peut être au profit d'autres axes. Notamment dans le cadre de la traduction automatique de la parole, le système de traduction s'appuie sur la sortie des SRAP. Ainsi, en procédant à une reconstitution des noms de personne découpés, un système de traduction de la parole profiterait de la réduction de l'influence des mots hors-vocabulaire dans la performance du SRAP.

Références

ALLAUZEN, A. et GAUVAIN, J.-L. (2005). Open vocabulary asr for audiovisual document indexation. *In Proceedings of the ICASSP*, volume 1, pages 1013–1016.

BAZZI, I. et GLASS, J. (2000). Modeling out-of-vocabulary words for robust speech recognition. *In The Proceedings of the sixth International Conference on Spoken Language Processing*, volume 1.

BÉCHET, F. (2001). Lia phon : un systeme complet de phonétisation de textes. *Traitement automatique des langues*, 42(1):47–67.

BISANI, M. et NEY, H. (2005). Open vocabulary speech recognition with flat hybrid models. *In INTERSPEECH*, pages 725–728.

DELÉGLISE, P., ESTEVE, Y., MEIGNIER, S. et MERLIN, T. (2009). Improvements to the lium french asr system based on cmu sphinx : what helps to significantly reduce the word error rate ? *In INTERSPEECH*, pages 2123–2126.

DUFOUR, R. (2008). From prepared speech to spontaneous speech recognition system : a comparative study applied to french language. *In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, pages 595–599. ACM.

GELAS, H., ABATE, S. T., BESACIER, L. et PELLEGRINO, F. (2012). Analyse des performances de modèles de langage sub-lexicale pour des langues peu-dotées à morphologie riche. *JEP-TALN-RECITAL*, page 53.

KNESER, R. et NEY, H. (1995). Improved backing-off for n-gram language modeling. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, page 181–184.

LOGAN, B., MORENO, P. et DESHMUKH, O. (2002). Word and sub-word indexing approaches for reducing the effects of oov queries on spoken audio. *In Proceedings of the second international conference on Human Language Technology Research*, pages 31–35.

LOGAN, B., MORENO, P., VAN THONG, J.-M. *et al.* (1996). An experimental study of an audio indexing system for the web. *In in Proc. ICSLP*. Citeseer.

MEIGNIER, S. et MERLIN, T. (2010). Lium spkdiarization : an open source toolkit for diarization. *In CMU SPUD Workshop*, volume 2010.

PERENNOU, G. et CALMES, M. d. (1987). Bdlex lexical data and knowledge base of spoken and written french. *In European conference on Speech Technology*.

RÉVEIL, B., DEMUYNCK, K. et MARTENS, J.-P. (2013). An improved two-stage mixed language model approach for handling out-of-vocabulary words in large vocabulary continuous speech recognition. *Computer Speech & Language*.

ROTOVNIK, T., MAUČEC, M. S. et KAČIČ, Z. (2007). Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech communication*, 49(6):437–452.

SHAIK, M. A. B., MOUSA, A. E.-D., SCHLÜTER, R. et NEY, H. (2011). Hybrid language models using mixed types of sub-lexical units for open vocabulary german lvcsr. *In INTERSPEECH*, pages 1441–1444.

SRINIVASAN, S. et BROWN, E. (2002). Is speech recognition becoming mainstream ? *Computer*, 35(4):38–41.

STOLCKE, A. *et al.* (2002). Srilm-an extensible language modeling toolkit. *In INTERSPEECH*.

VAN THONG, J.-M., MORENO, P. J., LOGAN, B., FIDLER, B., MAFFEY, K. et MOORES, M. (2002). Speechbot : an experimental speech-based search engine for multimedia content on the web. *Multimedia, IEEE Transactions on*, 4(1):88–96.

ZABLOTSKIY, S., SHVETS, A., SIDOROV, M., SEMENKIN, E. et MINKER, W. (2012). Speech and language resources for lvcsr of russian. *In LREC*, pages 3374–3377.

Perception et Réalisation du /R/ standard français en finale de mot.

Cedric Gendrot [1]
(1) LPP, ILPGA, 19 rue des bernardins, 75005 Paris.
cgendrot@univ-paris3.fr

RESUME

La variabilité du /R/ en français (comme dans d'autres langues) est reconnue dans la littérature, et ses différentes réalisations sont souvent considérées comme des variantes libres.
Dans cette étude, afin de mieux appréhender la variabilité du /R/, nous nous concentrons sur le /R/ final de mot, et nous avons choisi de l'étudier à l'aide d'une approche en 2 étapes : (1) au moyen d'un test de perception. Dans un premier temps des séquences de 2 mots séparées par la présence du /R/ (par exemple "par les" / "pas les") ont été sélectionnées afin de constituer un continuum entre un /R/ pleinement réalisé et un /R/ élidé. Ces items ont été soumis à 23 auditeurs pour un test d'identification afin de relever les paramètres pertinents à la perception du /R/ et ainsi proposer une mesure de rhoticité.
(2) Cette mesure de rhoticité, une fois utilisée sur de grands corpus de parole continue (spontanée et journalistique) nous permet d'appréhender la variabilité du /R/ d'après des prédicteurs tels que la fréquence de digrammes, le débit, le contexte phonémique, et la position prosodique.

ABSTRACT

Variability of (French) /R/ is a frequently studied phenomenon showing that /R/ can have multiple realizations. In French, all these studies were undertaken using small read corpora and we have reason to believe that these corpora don't allow to look at the full picture. Indeed factors such as local diphone and digram frequency, as well as speech rate can have almost as much influence as phonemic context in the realization of /R/
According to Ohala's AFP's principle, /R/ would tend to be either an unvoiced fricative or a voiced approximant. We chose to analyze word final /R/s as they tend to embrace the largest spectrum of variation. The study realized here is two-fold: a perception study in a specific phonemic context, between /a/ and /l/, where /R/ is realized as an approximant, so as to better understand the parameters and their thresholds necessary for /R/ identification, and provide a measure of rhoticity.
In a second step, keeping the rhoticity measurement in mind, we analyzed the realizations of word final /R/s in two continuous speech corpora (journalistic and spontaneous) and modelized the realization of /R/ using predictors such as diphone and digram frequency, phonemic context and speech rate.

MOTS-CLES : /R/, phonétique, production, perception, prédicteurs, grands corpus.
KEYWORDS: /R/, phonetics, production, perception, predictors, large corpora.

1 Introduction

1.1 Acoustique du /R/

Cette étude est un travail préliminaire au sein de l'ANR *REPER* visant à comprendre la variabilité du /R/ français standard dans une approche didactique pour les apprenants de Français Langue Etrangère.

Walter (1988), Chafcouloff (1980,1983), Fougeron (2007), Meunier (1994) pour le français ont évoqué la variabilité du /R/ qui peut être réalisé de voisé ([ʁ̬]) à dévoisé ([ʁ̥]), fricatif ([ʁ]), ou approximant ([ɹ̠]). De nombreuses études ont également mentionné la variabilité du /R/ pour d'autres langues (Maddieson, 1984 et références incluses, Mielke et al. 2010, Lawson et al. 2008), avec une tendance notable à se dérhoticiser. D'un point de vue phonologique, le /R/ pourrait être à la fois considéré comme une fricative ou comme une approximante et les évolutions mentionnées dans les travaux ci-dessus tendent à le faire basculer dans l'une ou l'autre de ces catégories. De même, la spécification de son voisement pourrait être reconsidérée.

Le /R/ final de mot, qu'il soit ou non suivi d'une pause, semble particulièrement sujet à variation, d'autant qu'il peut également être élidé. Nous considérons ici que la variabilité du /R/ français évoquée dans la littérature n'est pas parfaitement appréhendée, du fait qu'elle est analysée principalement dans des corpus lus et parfois basée sur des jugements impressionnistes. Nous prônons dans cette étude l'utilisation d'un test de perception ainsi qu'un recours à une analyse sur grands corpus afin de mieux comprendre la variabilité du /R/ en français. L'utilisation de corpus de parole continue peut avoir deux conséquences importantes : (1) observer une variabilité plus importante que pour de la parole lue (2) mais également de meilleures possibilités de comprendre les modulations de ce phonème, par des critères plus variés et de plus haut niveau que ceux fréquemment mentionnés dans les études précédentes. Au delà de critères tels que le contexte phonémique et la position dans le mot, nous pensons que des prédicteurs tels que le débit local, la fréquence (lexicale et/ou de digrammes), ou la position dans la phrase prosodique sont importants.

1.2 Perception du /R/

Les critères d'identification du /R/ restent flous, notamment si l'on prend en compte sa variabilité. Existe-t-il des critères qui restent stables quelque soit la réalisation du /R/ ? Nous pensons qu'il est nécessaire de réaliser un test perceptif qui permettra de mettre en parallèle l'identification du /R/ et des mesures acoustiques. On trouve peu d'études sur la perception du /R/, particulièrement en français. Une des sources de difficultés de mettre en place un test de perception sur le /R/ est qu'il est particulièrement difficile de le réaliser par la synthèse de Klatt. Il apparait hasardeux de contrôler tous les paramètres afin de réaliser un continuum tel que pour la recherche de prototype ou la perception

catégorielle. L'utilisation de grands corpus permet de contourner en partie, bien qu'imparfaitement, ce problème en permettant de relever un nombre important d'occurrences diversifiées et de constituer soi-même ce continuum par une sélection appropriée des items. Les détails de cette sélection seront évoqués dans la partie consacrée (2.1.)

1.3 Physiologie du /R/

La variabilité du /R/ pourrait être due à sa réalisation physiologique. En s'inspirant du principe de contrainte aérodynamique de voisement (Ohala, 2013), le statut approximant ou fricatif du /R/ impliquera *in fine* une dichotomie entre fricative sourde et approximante voisée. Notamment parce que le /R/ est un phonème postérieur (uvulaire) et que la taille de la cavité postérieure est alors très réduite, la pression orale peut arriver rapidement à hauteur de la pression sous-glottique et ainsi "éteindre" le voisement. Le /R/ pourrait ainsi être réalisé comme une fricative sourde ou comme une approximante voisée selon le degré de constriction. C'est ce qui selon nous explique la forte variabilité du /R/, le paramètre de stricture va influencer non seulement le continuum approximante/fricative, mais également le trait de voisement, ces 2 dimensions étant corrélées.

2 Méthode d'investigation

2.1 Test d'identification du /R/

Le but de cette expérience est de fournir des digrammes naturels distingués par la présence du /R/ (de type "par les" versus "pas les") à des auditeurs pour qu'ils les identifient. Trois types de séquences de digrammes ont été choisis pour leur fréquence d'usage dans les corpus : "par les" versus "pas les", "par la" vs. "pas la", "par le" vs. "pas le". Il était difficile d'envisager de présenter des monogrammes de par leur courte durée. Quatre locuteurs ont été choisis - au sein des 2 corpus présentés ci-dessous - sur la base du plus grand nombre d'occurrences obtenues et de leur variété. Les digrammes avec une pause entre les deux mots n'ont pas été retenus.

Trois hommes et une femme ont été sélectionnés : trois locuteurs (2 hommes, 1 femme) sur le corpus de parole journalistique et le locuteur masculin restant sur le corpus de parole spontanée. Un minimum de trois occurrences (et jusqu'à cinq quand cela était possible) par locuteur et par type de digramme a été retenu avec une occurrence où le /R/ est absent de façon sous-jacente ("pas"), une occurrence où le /R/ est réalisé hyper-articulé ([ʁ]) et une occurrence où le /R/ est jugé réduit ([ɻ]) par 3 experts phonéticiens (dont l'auteur). Quand cela était possible, une occurrence où le /R/ a été jugé élidé par les mêmes experts et une autre occurrence où le /R/ est réduit ont été ajoutées. La sélection des items s'est faite en choisissant des items au débit et à la f0 comparables. Au total, 43 items ont été sélectionnés.

Le test s'est déroulé sur ordinateur auprès de 23 étudiants de sciences du Langage de 1ère à 3ème année. Un script PRAAT a été écrit pour présenter les différents stimuli aux sujets selon la procédure suivante. Cinq répétitions par item ont été proposées dans ce test, dans un ordre semi-aléatoire afin d'éviter les doublons, et varié pour chaque sujet.

L'expérience s'est déroulée en trois parties distinctes ("par les", "par la" et "par le"). Après avoir choisi entre les deux propositions en cliquant sur la case appropriée ("pas les" ou "par les", "pas la" ou "par la" et "pas le" ou "par le"), les sujets devaient indiquer sur une échelle de 1 à 5 le degré de certitude avec lequel ils avaient répondu, 5 étant la certitude maximale. Chaque stimulus était précédé d'un bip et aucune autre écoute n'était possible pour le sujet. Une phase d'entrainement avec des items nouveaux permettait aux sujets de se familiariser à la tâche demandée.

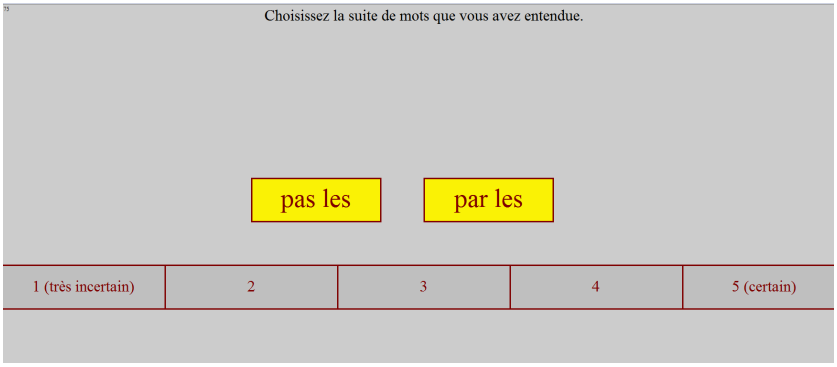


FIGURE 1 – Capture d'écran du test d'identification

2.2 Corpus et mesures

Deux types de corpus sont utilisés pour cette étude : le corpus ESTER (Galliano et al. 2005), un corpus de parole journalistique considéré comme de la parole préparée plutôt que lue, avec quelques séquences de parole libre ; et le corpus NCCF (*Nijmegen Corpus of Casual French*), détaillé dans Torreira et al. (2010) et qui s'approche plus de la parole spontanée. Dans les 2 cas, la segmentation et transcription orthographique a été dans un premier temps effectuée par des auditeurs humains et l'alignement en phonèmes et en mots a été réalisée automatiquement par le système d'alignement automatique du LIMSI (Gauvain et al. 2002).

La segmentation du /R/ est actuellement en cours de correction manuelle, elle n'a été effectuée que sur 11 heures du corpus de parole journalistique et 7 heures de parole spontanée. Cette correction de segmentation montre que si les frontières de début et de fin de /R/ sont souvent trop longues (environ 15ms au total), la position centrale du /R/ reste bonne, et la position des frontières précédant et suivant le phonème sont également bonnes. L'élision du /R/ est rarement prévue par les systèmes de reconnaissance automatique et la présence du /R/ est ainsi parfois corrigée par les correcteurs. La prise de décision reste cependant difficile puisque des "traces" acoustiques permettent d'identifier auditivement la présence du /R/ sans pour autant être capable de le localiser. Lorsque /R/ est réalisé comme une fricative sourde, l'information spectrale contenue dans le phonème ne pose pas de problème pour sa détection et sa segmentation. Par contre, lorsqu'il est réalisé comme une approximante voisée, il tend à se dérhoticiser et à être difficile à segmenter, voire à identifier, ce qui est fréquemment le cas dans le contexte que nous avons choisi ici (i.e. précédé par une voyelle et suivi par une sonante). Ces difficultés nous ont amené à réaliser le test de perception mentionné en 2.1 et à proposer des mesures dites syntagmatiques, i.e. relatives aux phonèmes environnants, décrites en 3.

Des mesures acoustiques ont été effectuées sur les items utilisés lors du test de

perception. Celles-ci ont été effectuées automatiquement avec Praat et vérifiées manuellement. Des mesures prosodiques (durée des différents segments, f0 et intensité) et spectrales (fréquence, largeur de bande et amplitude des formants, moments spectraux) ainsi que des mesures de HNR (harmonic-to-noise ratio) ont été choisies afin d'avoir un panel le plus descriptif possible dans notre tentative de corréler les résultats du test perceptif aux mesures acoustiques. Des mesures syntagmatiques (i.e. à partir des segments adjacents) ont été effectuées. Précisément, les mesures ont été prises entre le début du /a/ (de "pas" ou "par") et la fin de /l/ (de "les", "le" ou "la") à 20%, 40%, 60% et 80% de la durée de la séquence /aRl/. Des ratios entre chaque point de mesure ont permis d'obtenir des mesures relatives entre /R/ et son phonème précédent /a/. Le but de ces mesures syntagmatiques est d'obtenir des mesures relatives pour le cas où la variation des valeurs (plutôt que des valeurs absolues) serait plus pertinente. Les mesures relatives effectuées en comparant le /R/ et le /l/ ne se sont pas avérées significatives et ne sont pas développées ci-après.

3 Résultats

3.1 Perception de la différence "pas" / "par"

Les résultats du test de perception sont détaillés dans la table 1. Dans l'ensemble, les séquences "pas les" sont identifiées comme telles, i.e. sans le /R/ (à 96.25%). Les séquences "par les" où le /R/ avait été jugé pleinement réalisé par les experts sont identifiées comme telles à 97.1%, 87.25% pour les cas où /R/ avait été jugé approximant, et plus surprenant jusqu'à 74.4% pour les /R/ que les experts avaient jugé élidé. Ces premiers résultats, et particulièrement pour le dernier cas, indiquent que des traces acoustiques subsistent dans le signal, quand bien même la présence du /R/ serait considérée comme discutable par les experts phonéticiens. Il est à noter que seul un item "par" (un /R/ jugé comme approximant par les experts) a été identifié comme tel par moins de 50% des sujets. Le second item ayant recueilli le moins d'identifications "par" est un item pour lequel /R/ avait été jugé élidé, et son taux d'identification monte tout de même à 62.5%. Les scores de certitude indiquent une moyenne stable et proche du maximum (autour de 4.5) pour les réponses correspondant à la majorité des votes (en gras dans le tableau 1).

	pas les /pas le / pas la	par (ø) les / par (ø) le / par (ø) la	par (ʁ̥) les / par (ʁ̥) le / par (ʁ̥) la	par (ʁ) les / par (ʁ) / le par (ʁ) la
identification "par"	3.75% (par) **96.25% (pas)**	**74.4% (par)** 25.6% (pas)	**87.25% (par)** 12.75% (pas)	**97.1% (par)** 2.9% (pas)
certitude (de 1 à 5)	2.95 (par) **4.48 (pas)**	**4.14 (par)** 3.19 (pas)	**4.3 (par)** 2.87 (pas)	**4.5 (par)** 2.8 (pas)

TABLE 1 – Résumé du nombre d'identifications de "par" (et ses différentes réalisations) et "pas"

Dans un premier temps, nous avons effectué une régression logistique entre la forme "pas" et la forme "par" (jugée fricative) pour identifier les critères acoustiques entre les 2 formes sous-jacentes. L'identification du /R/ se caractérise pour tous les locuteurs par

une baisse de F2 ($p = 0.012$), et une diminution de la durée de la voyelle précédente ($p = 0.01$) et de la séquence /aRl/ ($p = 0.04$) pour l'item "par" en comparaison de "pas". Une montée de F1 et du HNR ('harmonic-to-noise-ratio'), ainsi qu'une diminution de f0 est observée pour trois locuteurs sur quatre seulement.

3.2 Perception de la différence entre les différents types de "par"

Dans un deuxième temps, les différentes formes de /R/ jugées par les experts phonéticiens (élidé, approximant, fricatif) ont été comparées afin d'identifier les critères de variation du /R/. Une régression linéaire a permis de mettre en évidence les facteurs acoustiques impliquant un nombre plus grand de détections du /R/ (i.e. du mot "par"). Seule une baisse de F2 ($p = 0.005$) et de la durée de la séquence /aRl/ ($p = 0.013$) permet de prédire significativement un taux plus important de détections de /R/. La durée de la voyelle précédente s'approche du seuil de significativité sans toutefois l'atteindre. Les 2 items "par" mentionnés en 3.1 ayant recueilli moins d'identification du /R/ lors du test de perception sont d'ailleurs caractérisés par une valeur plus haute de F2 véhiculant un timbre moins postérieur de la voyelle.

3.3 Regroupement de tous les items et calcul d'une mesure de rhoticité

Puisque les paramètres récurrents impliqués dans la réalisation et la variation du /R/ sont la durée de la voyelle précédente et de la séquence, et l'abaissement de F2 (ou plus précisément le rapprochement de F1 et F2 depuis la voyelle jusqu'au /R/), nous avons regroupé ces mesures en une mesure de rhoticité calculée de la façon suivante : durée de la voyelle précédente/(variation F2 - variation F1). Dans une travail en cours, nous espérons que cette mesure calculée et affichée sur un diagramme en temps réel permettra à des apprenants du Français Langue Etrangère de s'entrainer à la production du /R/ français.

Un arbre de décision (cf. figure 2) calculé dans le logiciel R (version 2.15.1) à l'aide de la fonction *ctree* a permis de déterminer les seuils d'identification du /R/ à 0.5 pour la mesure de rhoticité (/R/ est détecté au dessus de ce seuil), avec une variation nécessaire de F2 (depuis la voyelle /a/) estimée à 73 Hz et une diminution de durée vocalique d'environ 30ms.

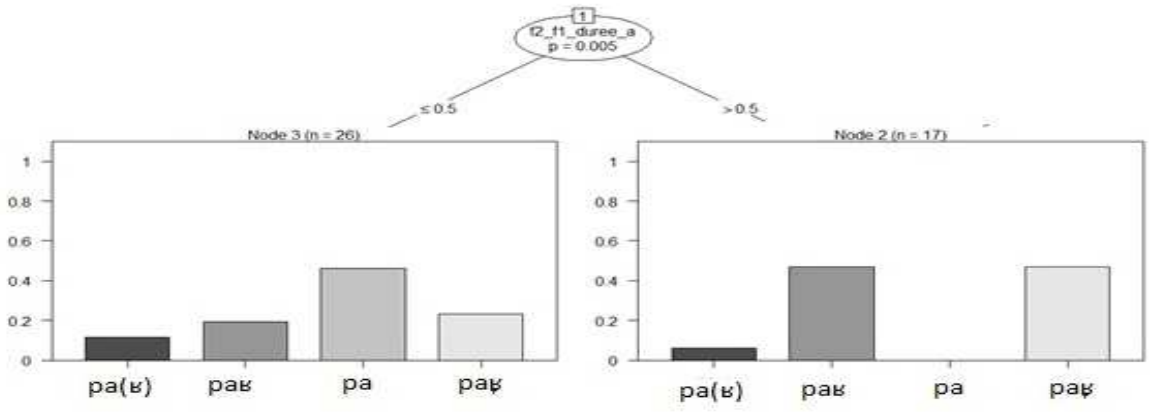


FIGURE 2 – Arbre de décision indiquant le seuil de détection du /R/ dans le test de perception

4 Analyse sur corpus

4.1 Variations du /R/ en finale de mot

Les mesures acoustiques sur corpus ont été effectuées de façon identique à celles effectuées sur les stimuli du test de perception, i.e. en mesurant les variations sur des séquences /aR/ en fin de mot pour les quatre locuteurs du test de perception, soit environ 6000 occurrences au total. La mesure de rhoticité calculée dans la section précédente a été utilisée en variable dépendante afin de tester sa variation en fonction de prédicteurs que l'on peut appréhender sur des corpus de parole continue, à savoir : le débit mesuré comme le nombre de phonèmes en 1 seconde (en incluant les pauses), le contexte phonémique suivant, la position dans le mot et dans la phrase prosodique, le nombre de syllabes dans mot, sa catégorie grammaticale, la fréquence *dans le corpus* du mot (contenant le /R/ ainsi que le mot suivant). La position du /R/ dans le groupe prosodique a été mesurée d'après une annotation spécifique détaillée dans Gendrot et Gerdes (2011). Deux positions ont été retenues : position finale ou médiane, mais une catégorisation plus fine pourra être envisagée prochainement.

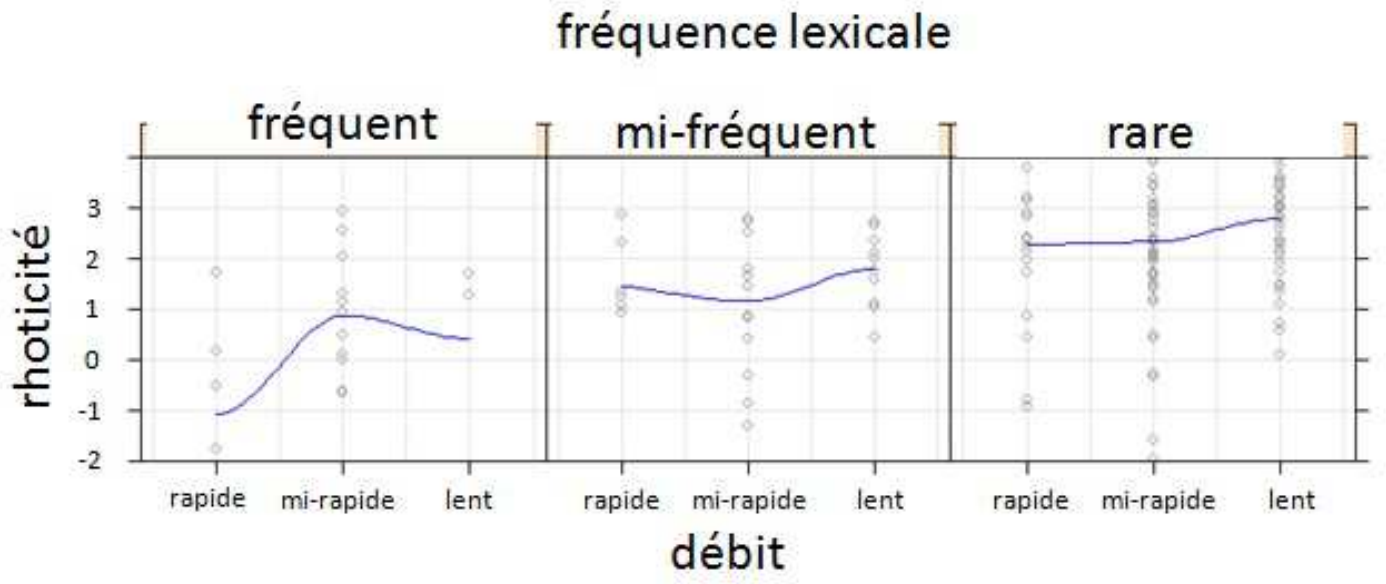


FIGURE 3 – Variation du /R/ selon notre mesure de rhoticité en fonction du débit et de la fréquence du mot porteur du /R/.

Des régressions linéaires (cf. figure 3) ont été effectuées pour montrer que la fréquence du mot porteur du /R/ ($p = 0.001$), la position dans le groupe prosodique ($p = 0.012$) et le débit ($p = 0.04$) prédisent une variation du /R/ (avec un R^2 à 40.2 %). La fréquence lexicale recouvre ici la catégorie grammaticale du mot qui serait significative sinon. Les occurrences identifiées comme ayant une valeur de rhoticité basse ont été vérifiées manuellement sur le corpus et apparaissent comme élidées par l'auteur mais devront dans une prochaine étape faire l'objet d'un nouveau test d'identification.

5 Discussion et Conclusion

Les conclusions de cette étude restent limitées puisqu'évaluées principalement sur un

seul contexte phonémique (entre /a/ et /l/ pour le test de perception ou après /a/ dans l'analyse sur corpus) et ne permettent pas de prendre en considération les réalisations de /R/ en tant que fricatives sourdes par exemple. Le but de ce projet est de poursuivre ces analyses sur d'autres contextes : s'il apparait difficile de demander à des auditeurs naïfs de distinguer des variations de friction et/ou de voisement du /R/, nous envisageons la mise en place d'un test de perception de type AXB qui permettra ainsi de mettre en évidence des regroupements de catégories perceptives sur les différentes réalisations de /R/ (fricative sourde, fricative voisée, approximante voisée, etc).

Des mesures et des analyses statistiques - non détaillées ici - ont été effectuées sur d'autres contextes phonémiques que ceux du test de perception, notamment la séquence /ɛʁ/ en position finale de mot (la voyelle /ɛ/ remplaçant /a/) puisque le /a/ est caractérisé par un F1 maximalement élevé pour une voyelle, ce qui a limité l'observation d'une montée de F1 souvent notée dans la littérature. Si les trois prédicteurs mentionnés dans cette étude sont retrouvés dans les régressions linéaires, d'autres prédicteurs telle que la f0 franchissent le seuil de significativité, suggérant une variabilité plus complexe encore que celle mesurée ici. Il sera intéressant de vérifier sur des /R/s réalisés comme fricatives sourdes (où la segmentation entre voyelle et /R/ est plus nette) si les mesures acoustiques observées sont reportées plus franchement sur la voyelle adjacente.

Au banc des limitations de cette étude, la prédiction de la variation du /R/ dans nos données n'apparait finalement que peu élevée avec un R^2 à 40%. Il semble que les mesures acoustiques proposées n'arrivent pas à la hauteur de la précision de l'oreille de nos auditeurs. Nous avions porté certains espoirs sur la mesure de HNR qui selon nous permettrait de mesurer progressivement le degré de friction dans le signal, mais la mesure de HNR est inadaptée en plusieurs points : elle détecte théoriquement le bruit au niveau glottique avant tout et se mesure en principe sur une fenêtre de périodicité proche de 100ms (Severin, 2005), ce qui dans les deux cas n'est pas valable pour le /R/. Nous travaillons donc sur l'élaboration d'une mesure de ratio entre friction et voisement plus appropriée pour notre cas.

Toujours est-il que des résultats intéressants sont visibles dès cette étude préliminaire. Pour le /R/ approximant voisé ici visé dans le test des perception, nous avons pu proposer une mesure de rhoticité - valable pour les rhotiques uvulaires seulement - caractérisée par une baisse du 2ème formant d'environ 70Hz et une baisse de la durée de la voyelle précédente (d'environ 30ms) permettant de favoriser l'identification du /R/. Cette mesure est à son tour influencée de façon systématique par la fréquence lexicale, le débit et la position dans le groupe prosodique ce qui laisse à penser que la variabilité observée pour le /R/ se rapproche pour beaucoup de la variabilité observée sur d'autres phonèmes. Elle semble seulement plus remarquable perceptivement à cause de contraintes aérodynamiques comme mentionné en 1.3.

Remerciements

Cette étude a été financée grâce au projet ANR jeunes chercheurs REPER (édition 2013) ANR-13-JSH2-0005-01 et également grâce au Labex EFL CGI.

Références

CHAFCOULOFF, M. (1980). Les caractéristiques acoustiques de [j, ɥ, w, l, r] en français. Dans *Travaux de l'Institut de Phonétique d'Aix*.

CHAFCOULOFF, M. (1983). A propos des indices de distinction /1-R/ en français. *Speech Communication.* Volume 2. pp. 137–139

FOUGERON, C. (2007) Word boundaries and contrast neutralization in the case of enchaînement dans French. In Cole, J. & Hualde J.I. (eds) *Papers in Laboratory Phonology IX: Change in Phonology*, Berlin: Mouton de Gruyter, pp. 609---642.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-P., GRAVIER G.(2005), The ESTER Phase II evaluation campaign for the rich transcription of French broadcast news, *Proceedings of Eurospeech-Interspeech*, Lisbonne, pp. 1149-1152.

GAUVAIN, J.L., LAMEL, L. ET ADDA, G. (2002) The Limsi Broadcast News Transcription System, *Speech Communication*, 37(1-2): pages 89-108.

LAWSON, E., STUART-SMITH, J., SCOBBIE, J. (2008) Articulatory Insights into Language Variation and Change: Preliminary Findings from an Ultrasound Study of Derhoticization in Scottish English. *University of Pennsylvania Working Papers in Linguistics.* Vol. 14(2)

MEUNIER, C. (1994) Les groupes de consonnes : problématique de la segmentation et variabilité acoustique. Thèse de Doctorat. Université d'Aix-en-Provence.

MADDIESON, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press

MIELKE, J, ADAM BAKER, AND DIANA ARCHANGELI. (2010) Variability and homogeneity in American English allophony and retraction. Variation, Detail, and Representation. *LabPhon 10*. Berlin: Mouton de Gruyter. 699-719.

OHALA, J. J. (1983), The origin of sound patterns in vocal tract constraints. In: MacNeilage, Peter (ed.): *The production of speech*. New York: 189-216.

RUSSELL WEBB, E. (2004a). Explanatory adequacy in phonology: a deductive approach to /R/. *Linguistica Atlantica*, 25: 77–94.

SEVERIN, F., BOZKURT, B., DUTOIT, T. (2005) HNR extraction in voiced speech, oriented towards voice quality analysis. Proceedings of EUSIPCO'05, Antalya,Turkey.

TORREIRA, F., ADDA-DECKER, M., & ERNESTUS, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52, pages 201-212.

TRANEL, B. (1987). *The Sounds of French: An Introduction*. Cambridge: Cambridge University Press.

WALTER, H. (1988). *Le français dans tous les sens*. Paris : Robert Laffont.

Opposition /y/-/u/ chez 7 apprenantes japonophones avec et sans l'aide de l'ultrason lingual: 1. analyse acoustique comparative voyelles isolées / logatomes

Claire Pillot-Loiseau, Tanja Kocjančič Antolík[1], Takeki Kamiyama[1,2]
(1) Laboratoire de Phonétique et Phonologie (LPP) UMR 7018 CNRS / Université Paris 3 Sorbonne Nouvelle : 19 rue des Bernardins, 75005 PARIS, FRANCE
(2) Linguistique Anglaise Psycholinguistique (LAPS), EA 1569, Université Paris 8 : 2 rue de la Liberté, 93526 Saint-Denis Cedex
claire.pillot@univ-paris3.fr, takeki.kamiyama@univ-paris8.fr, tkocjancic@googlemail.com

RESUME

Pour étudier la production de /y/ et /u/ français par une native et 7 apprenantes japonophones, quatre ont bénéficié de leçons concernant /u/ avec l'ultrason lingual (sujets EXP) et trois n'en n'ont pas bénéficié (CTR). Toutes ont été enregistrées plusieurs fois pour comparer l'effet de la présence ou de l'absence des leçons sur la production de l'opposition /y/-/u/. Les mesures acoustiques de dix répétitions de /y/ et /u/ isolés et dans des logatomes CV_1CV_2 (C = /p,t,k/, V = /y,u/) ont utilisé la Distance Euclidienne (DE) /y/-/u/, les différences entre F3 et F2 de /y/ (F3-F2) et F2 et F1 de /u/ (F2-F1). /y/ n'est pas mieux produit par EXP contrairement à F2-F1 de /u/ (moins grande différence) et DE (plus élevée) après les leçons. L'opposition /y-u/ est mieux réalisée à l'isolée pour toutes, et le moins bien produite pour les logatomes avec /t/. Les données des apprenantes sont plus variables que la native.

ABSTRACT

/y/-/u/ contrast by 7 Japanese learners with & without the help of Ultrasound : 1. Acoustical comparative analysis between isolated vowels and non-words.

In this study we investigated the production of French /y/ and /u/ by one native speaker and 7 Japanese-speaking learners; 4 of them received lessons on /u/ with ultrasound and 3 didn't receive these lessons. Acoustic measurements of 10 repetitions of isolated /y/ and /u/ are compared to those of non-words CV_1CV_2 (C = /p,t,k/, V = /y,u/), both before and after the ultrasound lessons, in terms of Euclidean distance (ED) between /y/ and /u/, the difference between F2 and F3 of /y/ (F3-F2), and F1 and F2 of /u/ (F2-F1). Our result show that /y/ isn't better produced by EXP unlike F2-F1 of /u/ (more little difference) and DE (higher) after the lessons. The opposition /y-u/ is best achieved in isolation for all, and it's bad produced for non-words with /t/. Data of the learners are more variable than those of the native subject.

MOTS-CLES : Apprenantes japonaises, ultrason, acoustique, distance euclidienne, formants, voyelles isolées, logatomes
KEYWORDS: Japanese learners, Ultrasound, acoustics, Euclidean Distance, formants, isolated vowels, non-words.

1 Introduction et but de l'étude

Le système vocalique d'une langue maternelle a un impact significatif sur l'acquisition des voyelles de langues étrangères. /y/ et /u/ sont des voyelles difficiles pour des locuteurs de langues qui ne possèdent pas ces voyelles (Kamiyama et Vaissière 2009). La voyelle antérieure arrondie fermée /y/ et la voyelle postérieure arrondie fermée /u/ sont phonologiquement en opposition en français ; parvenir à produire cette opposition est d'une grande importance pour les apprenants du Français Langue Etrangère (FLE). Des japonophones natifs de Tokyo (Kamiyama 2012) ont tendance à produire le /u/ français avec un deuxième formant (F2) plus élevé que les francophones natifs (Kamiyama et Vaissière 2009), ce qui est directement lié à la nature de la voyelle fermée non antérieure /u/ en japonais de Tokyo, transcrite [ɯ]. Acoustiquement, [ɯ] montre un F2 plus élevé (supérieur à 1000Hz) que celui du /u/ français ; d'un point de vue articulatoire, la langue est située moins loin en arrière et les lèvres sont moins arrondies (Bothorel et al. 1986 ; Uemura & Takada 1990).

Il est donc nécessaire que les japonophones apprenant le français positionnent leur langue aussi loin en arrière que le font les francophones natifs quand ils produisent /u/. Toutefois, parce que la position linguale postérieure ne peut pas être observée durant la parole, la correction de cette position se pratique couramment uniquement sur la base d'une évaluation perceptive. D'un autre côté, l'imagerie ultrason, technique d'observation durant la parole la moins invasive, la plus fiable et la plus économique, permet le contrôle en temps réel de la position linguale. Il n'existe pas à notre connaissance de recherches prouvant l'efficacité de cet équipement pour améliorer la prononciation de /y/ et /u/ du français chez des apprenants du FLE.

Notre but est d'évaluer acoustiquement la production de l'opposition entre /y/ et /u/ français par quatre japonophones ayant suivi trois leçons à l'aide de l'ultrason lingual pour la production de ces voyelles isolées et dans des logatomes (groupe EXP), et par trois autres apprenantes de même origine n'ayant pas bénéficié de cette aide spécifique (groupe CTR). Nous supposons 1° que les facteurs groupe (meilleure production de EXP que CTR), enregistrement et contexte influencent la production de /y/ et /u/; 2° qu'il existe des variations intra-groupes en fonction de la locutrice.

2 Méthode

Sept japonophones (niveau de français intermédiaire) ont suivi 12 semaines de formation traditionnelle de phonétique du FLE. EXP1 (33 ans, niveau B2), EXP2 (31 ans, B1), EXP3 (28 ans, B1) et EXP4 (29 ans, B2), ont suivi en plus trois leçons spécifiques avec l'ultrason, contrairement à CTR1 (29 ans, niveau B1), CTR2 (33 ans) et CTR3 (33 ans, niveau B1). Toutes sont natives de Tokyo ou des environs, sauf CTR2 de Fukuoka, et EXP4 de Kobe. Ayant commencé à apprendre le français adulte, elles vivent en France depuis au moins deux ans, sauf CTR1 (5 mois). Une locutrice native française (NAT, 42 ans) a aussi été enregistrée pour obtenir des données de référence.

Pour mettre en évidence les différences dans l'articulation de [y] et [u], les données ultrason et acoustiques ont été simultanément enregistrées avec une synchronisation de ces deux signaux, avec le logiciel *Articulate Assistant Advanced* (AAA). Les huit

participantes portaient un casque pour stabiliser la sonde sous le menton pour pouvoir analyser les images ultrason de manière fiable. Le tableau 1 résume l'ensemble des enregistrements réalisés par locutrice et leur chronologie.

Chronologie	**1[er] enregistrement**	**2[ème] enregistrement**	**3[ème] enregistrement**	**Groupe**
Enregistre-ments et sujets	**ENR 1** (1 semaine avant 1[ère] leçon et formation 12 semaines)	**ENR 2** (1 semaine après 3[ème] leçon et formation 12 semaines)	**ENR 3** (2 mois après le ENR 2)	
NAT	X			Référence
EXP 1 à EXP 4	X	X	X	Expéri-mental
CTR 1 à CTR 3	X (sans leçon ultrason)	X (sans leçon ultrason)		Contrôle

TABLE 1 – Chronologie des 19 enregistrements réalisés pour chaque locutrice.

Les quatre locutrices du groupe expérimental (EXP) ont bénéficié de trois leçons de 45mn dans lesquelles l'ultrason lingual a été utilisé pour réaliser et contrôler la position linguale pour /y/ et /u/. L'entraînement a débuté avec la répétition avec puis sans image ultrason, puis la lecture de : 1) voyelles /y/ et /u/ isolées puis alternées, 2) syllabes avec puis sans allongement de la voyelle cible, où /y/ et /u/ apparaissaient en position initiale, finale puis médiane. Les voyelles cibles de ces syllabes apparaissaient dans des contextes phonétiques facilitants, neutres puis difficiles tels que définis en français par Callamand (1981). 3) Ces leçons se sont poursuivies avec la répétition puis la lecture de mots respectant ces contextes, puis de phrases et textes avec ces mots. Le protocole exact de cet apprentissage était ajusté aux capacités et aux préférences de chaque participant. Le temps nous a finalement permis de ne véritablement travailler que /u/.

Le corpus enregistré était composé de dix répétitions (deux séries de cinq) de 1) [y], [u], [a], [i] et, pour les sept japonophones, [ɯ] isolées ; 2) alternance entre [y] et [u] ; 3) logatomes bisyllabiques CV_1CV_2 (V = /y,u/ et C = /p,t,k/) ; 4) 28 mots avec /y/ et /u/, et 5) quatre phrases. Seuls les résultats concernant les voyelles isolées et les logatomes sont ici présentés.

Les données acoustiques, enregistrées à 22050Hz, 16bits, ont permis l'analyse des quatre premiers formants mesurés à 25%, 50% et 75% de la durée de chaque voyelle. Nous avons ensuite mesuré la distance euclidienne (DE) entre [y] et [u] pour calculer le degré de réalisation de l'opposition entre ces deux voyelles (Marushima *et al.* 2010, en prenant en compte les trois premiers formants), et calculé la différence F3-F2 pour [y] (Gendrot *et al.* 2008), et F2-F1 pour [u]. En effet, « les faibles écarts entre F1 et F2 (voyelles postérieures), et F2 et F3 (/y/) pour les voyelles françaises focales semblent une caractéristique définitoire de ces voyelles. » (Georgeton *et al.* 2012).

3 Résultats

3.1 Influence du groupe, du contexte et de l'enregistrement

En premier lieu, nous proposons une étude par groupe (inter-sujets) afin de déterminer si la session d'enregistrement, le groupe (NAT, CTR, EXP) et le contexte (voyelle isolée vs

logatome, V1 vs V2 et consonne dans les logatomes) influencent les valeurs de F3-F2 de /y/, F2-F1 de /u/ et de DE. Le tableau 2 montre les résultats d'une ANOVA calculée pour les mesures répétées des facteurs groupe, enregistrements 1 et 2, et contexte sur les valeurs acoustiques rappelées ci-dessus. Les figures 1 à 3 donnent la médiane et l'écart-type des valeurs pour lesquelles les tendances les plus significatives sont observées.

Mesures répétées	Degré de liberté	Σ des carrés	F	p
F3-F2 de /y/				
Groupe (NAT, CTR, EXP)	2	43203486	1.24	0.36
Enregistrement (1 vs 2)	1	2128135	0.55	0.51
Contexte	**6**	**52610915**	**4.55**	**0.002**
Groupe x contexte	12	5531268	0.24	0.99
Groupe x enregistrement	1	2009416	0.47	0.52
Contexte x enregistrement	6	1503796	0.41	0.86
Groupe x contexte x enregistrement	6	5300181	1.45	0.23
F2-F1 de /u/				
Groupe (NAT, CTR, EXP)	2	76906418	2.22	0.20
Enregistrement (1 vs 2)	**1**	**43429978**	**17.8**	**0.008**
Contexte	**6**	**117617795**	**36.91**	**1.5e-12**
Groupe x contexte	12	2540304	0.39	0.95
Groupe x enregistrement	1	15065301	6.18	**0.05**
Contexte x enregistrement	6	525481	0.25	0.95
Groupe x contexte x enregistrement	6	763793	0.37	0.89
DE				
Groupe (NAT, CTR, EXP)	**2**	**126466905**	**9.14**	**0.021**
Enregistrement (1 vs 2)	**1**	**48357368**	**20.36**	**0.006**
Contexte	**6**	**59638739**	**9.61**	**6.5e-06**
Groupe x contexte	12	3129782	0.25	0.99
Groupe x enregistrement	1	1730813	0.73	0.43
Contexte x enregistrement	6	145188	0.04	0.99
Groupe x contexte x enregistrement	6	1157438	0.33	0.91

TABLE 2 – ANOVA calculée pour les mesures répétées des facteurs groupe, enregistrements 1 et 2, et contexte (isolé/logatome, V1/V2 et consonne) sur les variables dépendantes F3-F2 de /y/ ; F2-F1 de /u/ et DE

D'après ce tableau, le contexte est l'unique facteur affectant significativement F3-F2 de /y/. Des tests Post-hoc confirment que cette mesure est significativement différente entre V1, V2, et la voyelle isolée, excepté pour V1 après /p/ vs. /y/ isolé, V2 après /t/ vs. /y/ isolé, V1 après /p/ vs V2 après /t/ et les voyelles dans les logatomes kV1kV2. Les facteurs groupe, enregistrement et toutes les interactions de facteurs (groupe x contexte, groupe x enregistrement, contexte x enregistrement, groupe x contexte x enregistrement) n'affectent pas significativement cette mesure.

Les trois groupes ne montrent pas de différence entre eux sur la différence F2-F1. Toutefois, cette différence concernant la voyelle /u/ est significativement influencée par la session d'enregistrement pour EXP ($p < 0{,}01$). La figure 1 illustre les médiane et écart-type de cette différence formantique pour le /u/ des sujets CTR et EXP en fonction

de l'enregistrement : F2-F1 ne diminue que pour EXP de ENR1 à ENR2 : sa valeur moyenne se rapproche de F2-F1 du /u/ français (488Hz, Georgeton et al. 2012).

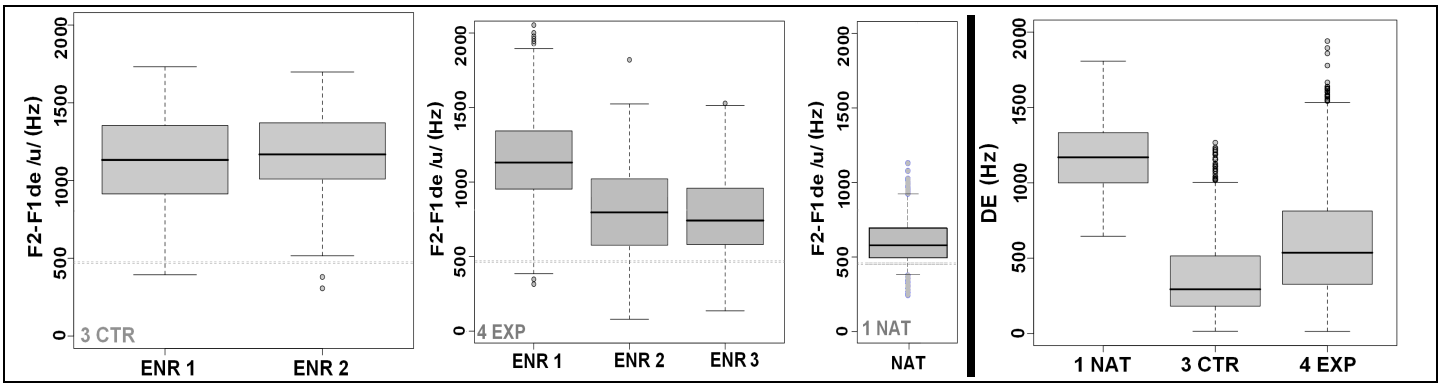


FIGURE 1 – Médiane et écart-type de F2-F1 de /u/ (tous contextes condondus) en fonction de l'enregistrement pour les sujets CTR (gauche), EXP et NAT (milieu). N = 1260 pour les 3 CTR ; N = 2520 pour les 4 EXP et N = 210 pour NAT. Ligne horizontale en pointillés : F2-F1 de /u/ français (800Hz : Georgeton *et al.* 2012). Droite : Médiane et écart-type de DE en fonction du groupe (NAT, CTR, et EXP). N = 1260 pour les 3 CTR ; N = 2520 pour les 4 EXP et N = 210 pour NAT.

F2-F1 est aussi surtout influencée par le contexte d'émission de /u/ ($p < 0{,}0001$) : la figure 2 montre les valeurs les plus basses de F2-F1 de /u/ dans le contexte isolé (sauf pour CTR), et des valeurs élevées pour V1 de /tutu/. Les tests Post-hoc pour le facteur contexte dans les logatomes révèlent que F2-F1 est significativement différente (p est au moins inférieur à 0,0001) entre V1 et V2 uniquement avec /t/, et entre C1, C2 et /u/ isolée. Aucune des interactions entre les facteurs n'est significative.

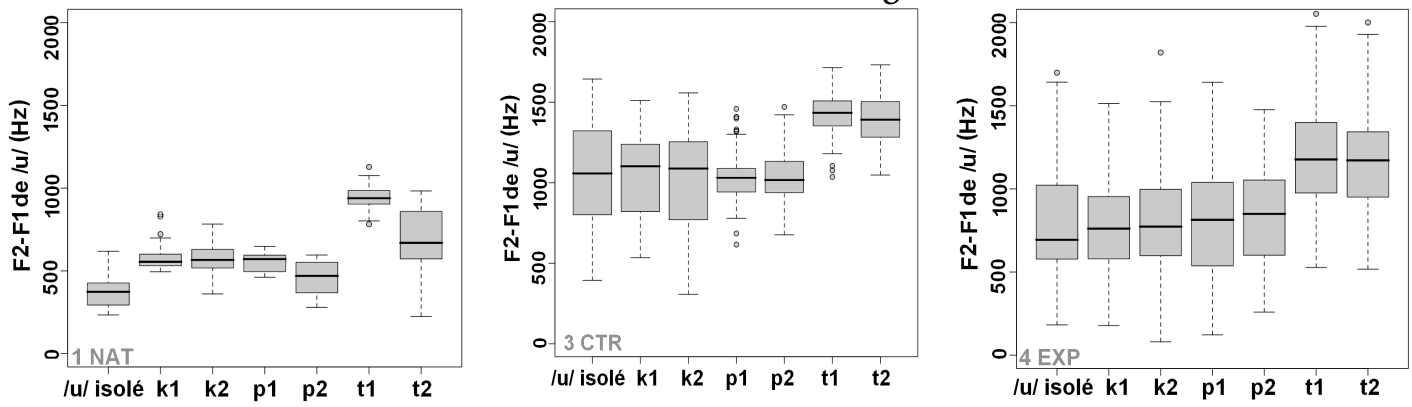


FIGURE 2 – Médiane et écart-type de F2-F1 de /u/ en fonction du contexte pour les sujets NAT (à gauche), CTR (au milieu), et EXP (à droite). N = 2520 pour les 3 CTR ; N = 5040 pour les 4 EXP et N = 420 pour NAT.

Les facteurs groupe (figure 1), ENR et surtout contexte ($p < 0{,}0001$) ont un effet significatif sur la valeur de la distance euclidienne DE. Des tests Post hoc de Tukey montrent que DE de chacun des trois groupes (NAT, EXP et CTR) est significativement différente de celle d'un autre groupe ($p < 0{,}0001$). La figure 1 montre que DE est plus importante chez NAT, puis chez EXP. En outre, dans les logatomes, DE est significativement différente (p au moins inférieur à 0,005) entre V1 et V2 sauf quand la consonne est /t/ ou /k/. Aucune interaction entre les facteurs n'est significative.

3.2 Evolution au cours des différents enregistrements

Cette section s'intéresse à l'évolution des paramètres acoustiques d'un enregistrement à l'autre, sujet par sujet.

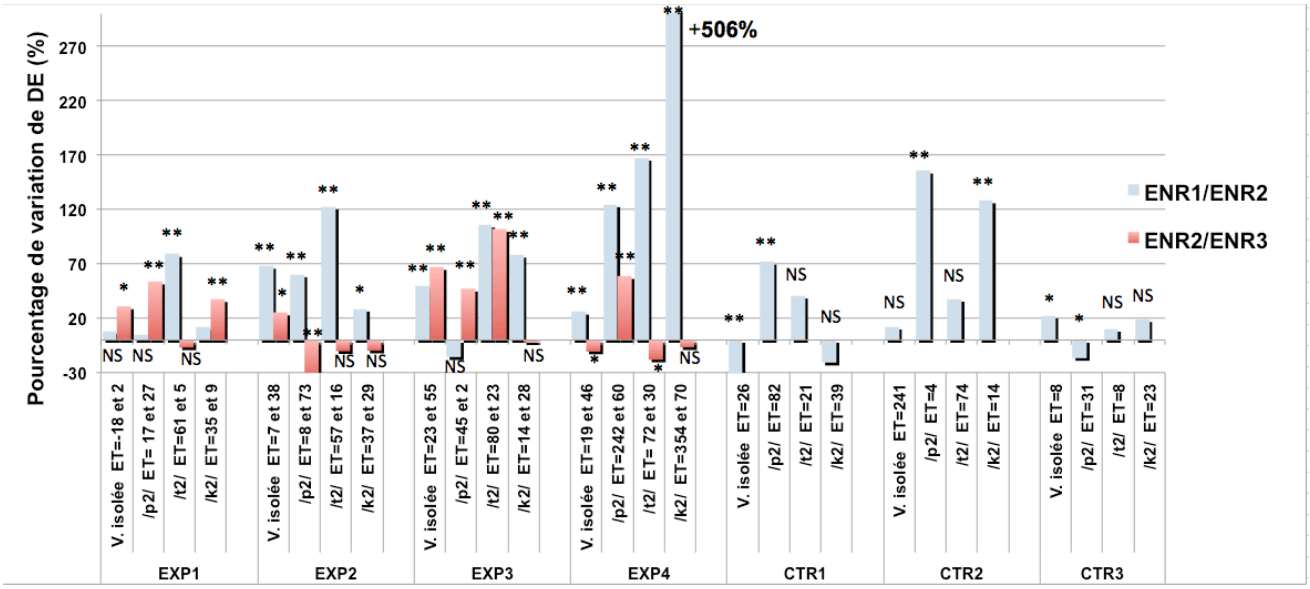


FIGURE 3 – Pourcentage de variation de DE (%) entre ENR 1 et ENR 2 (bleu) et entre ENR 2 et ENR 3 (rouge) pour les sept japonophones. ANOVA à un facteur et test PLSD de Fisher: NS : $p > 0{,}05$; * : $0{,}0001 < p < 0{,}05$; ** : $p < 0{,}001$; $z < 0{,}05$ pour toutes les valeurs de F ; DDL = 3,116 (EXP), DDL = 2,87 (CTR) ; N = 120 (EXP), N = 90 (CTR).

Dans la figure 3 (pourcentage de variation de DE d'un enregistrement au suivant), par souci de visibilité, les écarts-types ont été inscrits sur l'axe des abscisses. Si DE augmente, l'opposition entre /y/ et /u/ est mieux produite d'un enregistrement au suivant. D'après une ANOVA à un facteur, DE augmente significativement de ENR1 à ENR2, sauf pour les voyelles isolées, les logatomes avec /p/ et /k/ de EXP1, ceux avec /p/ de EXP 3, ceux avec /t/ de CTR1, CTR2 et CTR3, les voyelles isolées de CTR2, et les logatomes avec /k/ chez CTR1 et CTR3. L'augmentation de DE est particulièrement remarquable pour les logatomes avec /k/ chez EXP 4 (500%). Les sujets EXP augmentent en moyenne leur DE de 89%, tandis que les sujets CTR l'augmentent de 35%. Toutefois, il existe une importante variabilité intra-sujet d'une part (notamment pour les logatomes avec /p/ et /k/ de EXP4 n'ayant pas prononcé la même voyelle entre les cinq premières répétitions et les cinq dernières), et entre les locutrices d'autre part : EXP4 montre les plus grands progrès dans la production de l'opposition /y/-/u/, ainsi que CTR2. DE augmente significativement de ENR2 à ENR3 pour les sujets EXP, sauf pour les logatomes avec /t/ de EXP1 et EXP2, et ceux avec /k/ de EXP2, EXP3 et EXP4. DE augmente en moyenne de 20% de ENR2 et ENR3, et est la plus importante pour EXP3.

Si F3-F2 diminue, /y/ a été produite plus focale et plus antérieure d'un enregistrement à l'autre, se rapprochant du /y/ français. La figure 4 montre une production inconstante de cette voyelle en fonction de l'item (voyelle isolée ou logatome) ou du sujet : F3-F2 de /y/ diminue en moyenne de 7,8% tant pour les sujets CTR que les sujets EXP de ENR1 à ENR2. Seuls EXP1 et EXP4 réalisent de significatifs progrès entre ENR2 et ENR3 pour les logatomes avec /p/ et /k/. EXP2 augmente F3-F2 d'un enregistrement à l'autre pour tous les items, excepté pour les logatomes avec /t/.

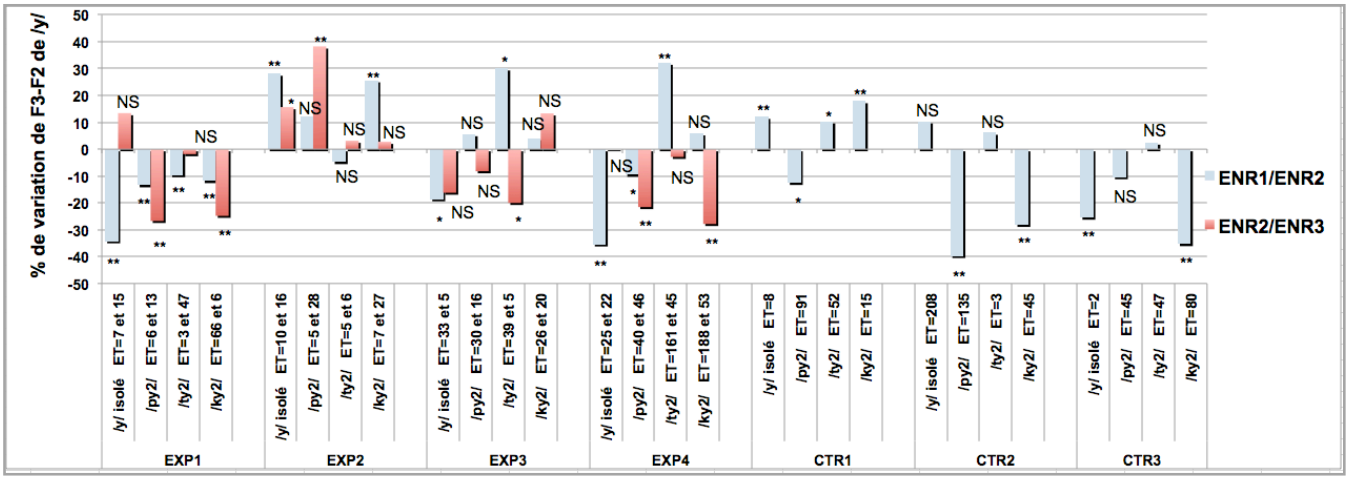


FIGURE 4 – Pourcentage de variation de F3-F2 de /y/ entre ENR 1 et ENR 2 (bleu) et entre ENR 2 et ENR 3 (rouge) pour chaque sujet. Mêmes statistiques que figure 3.

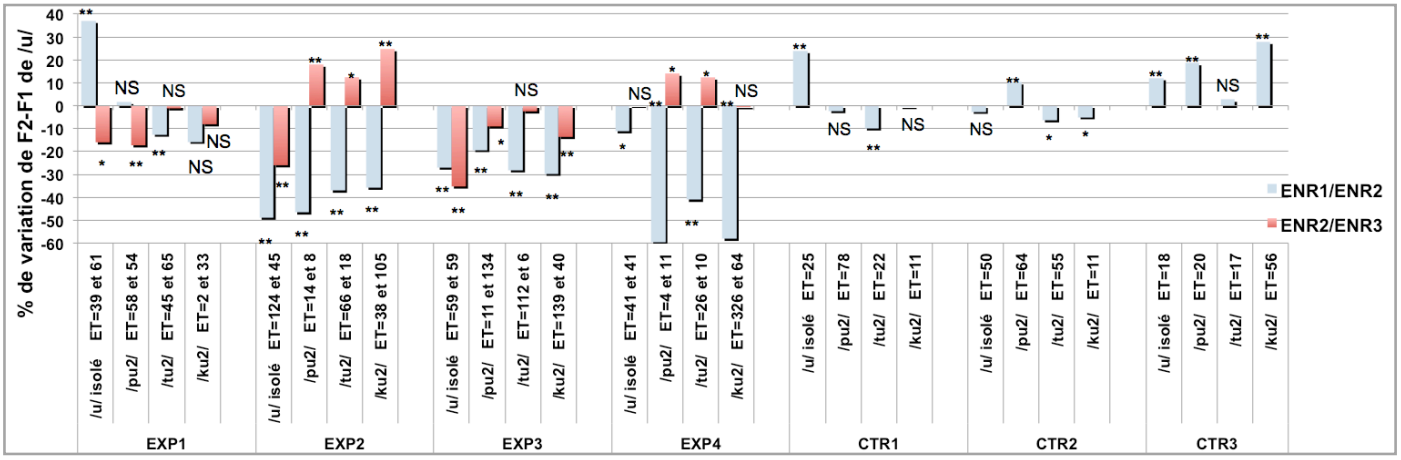


FIGURE 5 – Pourcentage de variation de F2-F1 de /u/ (Hz) entre ENR 1 et ENR 2 (bleu) pour les sept japonophones. Même test statistique que pour les figures 3 et 4.

Si F2-F1 diminue, /u/ a été produite plus focale et plus postérieure d'un enregistrement à l'autre, se rapprochant du /u/ français. D'après la figure 5, contrairement à la production de /y/, les progrès sont significatifs pour les sujets EXP 2, EXP 3 et EXP 4, de ENR 1 à ENR 2. F2-F1 de /u/ diminue en moyenne de 27% pour les sujets EXP, et augmente de 5,8% pour les sujets CTR de ENR 1 à ENR 2.

4 Discussion

Pour les 2/3 de nos sujets, /y/ et /u/ isolées sont davantage opposées (DE plus grande) et focales (différences formantiques plus petites) que quand ces voyelles se trouvent en position finale de logatomes CV_1CV_2, quelle que soit la consonne associée. L'articulation est plus précise à l'isolée, comme l'ont déjà remarqué Georgeton *et al.* (2012, voyelles isolées « articulées hors contexte ») et Gendrot et Adda-Decker (2005). Ces derniers auteurs expliquent ces variations par le fait que les valeurs formantiques des voyelles sont inversement proportionnelles à la durée. Bien que nous comparions les voyelles isolées et les deuxièmes voyelles des logatomes, et non les voyelles isolées aux corpus journalistiques comme l'ont fait ces auteurs, nous avons également trouvé des variations de durée en faveur de celle des voyelles (voyelles isolées NAT : 340ms (ET 30), pour EXP + CTR : 251ms (ET : 43) ; moyenne des V_2 (logatomes) : NAT : 190ms (ET : 17), pour EXP + CTR : 203ms, ET : 30). En outre, l'entourage

consonantique des voyelles cibles (logatomes) influence l'articulation de la plupart des huit sujets : en raison d'une coarticulation favorable, /y/ est plus facilement articulée avec /t/, et /u/ plus facilement produite avec /k/. Ces contextes favorisants (Callamand, 1981), sont à utiliser lors des premières leçons de prononciation de ces voyelles auprès d'apprenants du FLE pour en favoriser la production.

F3-F2 de /y/ et F2-F1 de /u/ de la native à l'isolée sont proches de celles des 40 femmes françaises enregistrées par Georgeton *et al.* (2012 : valeur de 488Hz pour ces deux différences). Elles augmentent pour les logatomes, mais restent inférieures à 800Hz, valeurs typiques de voyelles focales françaises (Georgeton *et al.* 2012).

Selon Marushima *et al.* (2010), /u/ et /y/ sont les plus difficiles à produire pour les japonophones, même chez des apprenants avec une grande durée d'apprentissage du français. Notre étude confirme ces assertions : à ENR1, DE des apprenantes est en moyenne de 379Hz, soit un peu plus de trois fois inférieure à DE moyenne de NAT (1234Hz, tous items confondus). F3-F2 de /y/ est en moyenne deux fois plus importante (moyenne chez les 7 apprenantes, tous items confondus : 939Hz, ET : 155) que celle de NAT, se révélant proche d'une telle différence pour la voyelle [ø] (1097Hz en moyenne, Georgeton *et al.* 2012). Gao *et al.* (2013) ont trouvé chez 4 japonophones un écart entre F2 et F3 de /y/ isolée de 1455 Hz, et de 1168Hz pour l'écart entre F1 et F2 de /u/ chez ces mêmes apprenantes, contre 1122Hz en moyenne tous items et japonophones confondus (ET : 2012) dans notre étude. F2-F1 de /u/ est proche d'une telle valeur pour [ø] (1183 Hz en moyenne, Georgeton *et al.* 2012).

Il est cependant remarquable de constater, de ENR1 à ENR3, une amélioration de ces valeurs se rapprochant de celles des voyelles françaises cibles pour les sujets EXP ayant bénéficié des trois leçons spécifiques à l'aide de l'ultrason lingual : l'amélioration concernant la voyelle /y/ n'est pas remarquable, dans la mesure où nous n'avons eu le temps de travailler spécifiquement que la voyelle /u/ d'une part, et que l'attention à l'arrondissement des lèvres n'a pas été enseignée en priorité, la première place du travail pédagogique étant consacrée à la visualisation de la position postérieure de /u/ sur les images de l'ultrason. L'amélioration de /u/ est cependant notable pour les sujets EXP2, EXP3 et EXP4 contrairement aux trois sujets CTR. Les sujets CTR2 et EXP4 ne venant pas de Tokyo, il est probable que leur accès à /u/ français leur ait été facilité car F3-F2 de /y/ chez CTR2 est de 625Hz (ET : 71) dès ENR1, et DE de EXP4 pour les voyelles isolées est de 1091Hz (ET : 175Hz) dès ENR1.

5 Conclusions et perspectives

Notre étude a confirmé l'impact acoustique d'une meilleure réalisation de l'opposition /y/-/u/ isolées et dans des logatomes par les apprenantes japonophones quand celles-ci ont pu visualiser la position de la langue de ces voyelles à l'aide de l'ultrason. La quantification de l'analyse des images linguales correspondantes est en cours, et la validation perceptive de ces productions est ici présentée dans un autre article.

Remerciements

Ce travail a bénéficié d'une aide de l'Etat gérée par l'ANR au titre du programme Investissements d'Avenir portant la référence ANR-10-LABX-0083.

Références

BOTHOREL, A., SIMON, P., WIOLAND, F., ZERLING, J.P. (1986). *Cinéradiographie des voyelles et consonnes du français*. Strasbourg, Publications de l'Institut de Phonétique de Strasbourg.

CALLAMAND, M. (1981). *Méthodologie de l'enseignement de la prononciation : organisation de la matière phonique du français et correction phonétique*, Paris : Création Loisirs Enseignement International, 191 pages.

GAO, J., GEORGETON, L., KAMIYAMA, T., PAILLEREAU, N. (2013). Étude inter-langues sur la production des voyelles focales et des voyelles moyennes du français en français langue étrangère (FLE). *PPLC 13 (Phonetics, phonology and languages in contact; Contact varieties, multilingualism, and second language learning)*, 21-23/08/2013, Paris, 128-131.

GENDROT, C. et ADDA-DECKER, M. (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 Interspeech 2005*, Lisbonne, Portugal, 2353-2456.

GENDROT, C., ADDA-DECKER, M., VAISSIÈRE, J. (2008). Les voyelles /i/ et /y/ du français : focalisation et variations formantiques. *In XXVIIèmes Journées d'étude sur la Parole (JEP 2008)*, Avignon, France, 205-208.

GEORGETON, L., PAILLEREAU, N., LANDRON, S., GAO, J., KAMIYAMA, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE, *In XXX[e] Journées d'Études sur la Parole*, Grenoble, 145-152.

KAMIYAMA, T., VAISSIERE, J. (2009). Perception and production of French close and close-mid rounded vowels by Japanese-speaking learners, *Acquisition et interaction en langue étrangère Aile... Lia* 2:9-41.

KAMIYAMA, T. (2012). Production des voyelles du français par des apprenants japonophones : effet du dialecte d'origine. *In XXX[e] Journées d'Études sur la Parole (JEP 2012)*, Grenoble, 771-778.

MARUSHIMA, N., DETEY, S., KAWAGUCHI, Y. (2010). Caractéristiques phonétiques des voyelles orales arrondies du français chez des apprenants japonophones, *Flambeau,* 36:53-72.

UEMURA, Y., TAKADA, M. [National Institute for the Japanese Language] (1990). [Vowels, consonants and syllables in Japanese: experimental phonetics study in articulatory movements]. Tokyo, Shûei shuppan.

Perturbations du Voice Onset Time (V.O.T) et du Voice Termination Time (V.T.T.) dans des séquences VCV produites par des patients francophones souffrant de paralysies récurrentielles

Camille Fauth[1], Béatrice Vaxelaire[2], Jean-François Rodier[3], Pierre-Philippe Volkmar[3], Rudolph Sock[2 et 4]

[1] LORIA – Equipe Parole CNRS UMR 7503 Nancy - France

[2]Université de Strasbourg, Institut de Phonétique de Strasbourg – IPS & U.R. 1339 Linguistique, Langues et Parole – LilPa, E.R. Parole et Cognition Strasbourg - France

[3]Centre Paul Strauss, Département de Chirurgie Oncologique, Centre Régional de Lutte Contre le Cancer – Strasbourg - France

[4]Université Pavla Jozefa Safarika, Faculté des Lettres Košice - Slovaquie

camille.fauth@gmail.com

RESUME

Ce travail entend étudier les caractéristiques temporelles de la voix de patients présentant une paralysie récurrentielle post-thyroïdectomie, à l'aide de différents indices spatio-temporels tels que le V.OT. et le V.T.T. Il s'agit en réalité d'analyser les conséquences d'une paralysie récurrentielle sur le timing oro-laryngé des patients. 7 patients ont été enregistrés lors de trois phases post-opératoires. Notre corpus est composé de douze logatomes de type VCV.
Notre démarche se veut articulatori-acoustique, puisqu'il s'agit d'observer à partir du signal acoustique continu, les différents indices articulatori-acoustiques qui peuvent nous permettre de remonter aux timing et aux configurations articulatoires. Une modification statistiquement significative de la durée du V.O.T. apparait pour les consonnes voisées uniquement dans les phases d'enregistrement post-opératoires proches de l'opération. La durée du V.T.T. n'est en revanche pas modifiée. Le temps et la rééducation vocale permettent un retour à des valeurs proches de celles obtenues pour les locuteurs contrôles.

ABSTRACT

A spatiotemporal study of speech in patients with recurrent nerve paralysis after thyroid surgery focussing on V.O.T. and V.T.T.

This study analyses the temporal characteristics of the voice of patients with laryngeal post-thyroidectomy surgery immobility, using different spatiotemporal cues such as V.O.T. and V.T.T. Actually, we examine the consequences of laryngeal paralysis on oro-laryngeal timing of 7 patients, recorded during three postoperative phases. Our corpus consists of twelve VCV sequences. The approach, an articulatory-to-acoustic one, allowed us to observe, in the acoustic speech signal, different cues that may allow tracing the articulatory timing patterns and configurations. A statistically significant change in the duration of VOT appears only for voiced consonants, in the early postoperative phase. V.T.T. durations are however not changed. Time and voice therapy have a positive impact as patients' values in late recording phases are close to those

obtained for control speakers.

MOTS-CLES : production de la parole, parole pathologique, phonétique clinique, glande thyroïde, paralysie récurrentielle, perturbations, réajustements, V.O.T., V.T.T.
KEYWORDS: speech production, speech disorder, clinical phonetics, thyroid gland, laryngeal immobility, perturbations, readjustments, V.O.T., V.T.T.

1 Introduction

La présente étude propose d'analyser les caractéristiques spatio-temporelles de la voix de patients présentant une paralysie récurrentielle, suite à une opération de la glande thyroïde. Il s'agit ici d'évaluer les conséquences de la paralysie récurrentielle sur le *timing* des articulateurs, à l'aide d'un corpus de logatomes (ou séquences VCV). Rappelons qu'une paralysie récurrentielle a des conséquences acoustiques et aérodynamiques sur la voix des patients (voir Hartl, 2004 ou Fauth, 2012 par ex.). Dans la plupart des cas, la paralysie récurrentielle post-thyroïdectomie est transitoire, une amélioration de la qualité vocale est donc mesurable au cours du temps (Laccourreye, Le Clech, & Santini, 2008). Il s'agit ici de savoir si cette amélioration de la qualité vocale est manifeste dans les aspects temporels du signal de parole.

Notre investigation, bien qu'acoustique, entend analyser les différents indices articulatori-acoustiques qui peuvent être observés à partir d'un signal acoustique continu, et ainsi remonter aux stratégies et aux configurations articulatoires (Abry et al., 1985). Une attention particulière sera donc accordée aux différentes perturbations temporelles qu'une paralysie récurrentielle peut provoquer sur la voix des patients, grâce à l'étude de la réorganisation temporelles des paramètres retenus. Notre étude se veut longitudinale puisque la voix des patients est enregistrée lors de différentes phases post-opératoires (post-opératoire 1, 2 et 3), ce qui devrait nous permettre de mettre au jour, non seulement les *perturbations* provoquées par la paralysie récurrentielle, mais également les *réajustements* que les patients pourraient mettre en place au cours du temps et de leur rééducation orthophonique.

2 Procédure expérimentale

2.1 Locuteurs

Cette étude est conduite à partir de sept patients dont la voix a été enregistrée, soit cinq femmes (UPPBAS, UPPHEI, UPPHUB, UPPWAL et UPPWAN) et deux hommes (UPPLAT et UPPPAI). Tous ces patients présentent une paralysie récurrentielle unilatérale consécutive à une thyroïdectomie totale. Une précédente étude (Fauth, 2012) effectuée à partir de voyelles soutenues a montré que la qualité vocale de ces patients étaient détériorée après l'opération avant de s'améliorer grâce au temps et à la rééducation orthophonique. Pour la présente étude, les patients ont été enregistrés une fois par mois tout au long de leur rééducation orthophonique. Afin d'obtenir une cohorte comparable et ainsi de pouvoir conduire des analyses de variance (ANOVA à mesures répétées), seules trois phases post-opératoires ont pu être retenues : en post-opératoire 1 (soit quinze jours après l'intervention), en post-opératoire 2 (soit un mois après l'intervention) et en post-opératoire 2 (c'est-à-dire deux mois après l'intervention). Pour certains locuteurs, la rééducation orthophonique n'était alors pas terminée.

L'incidence de la paralysie récurrentielle unilatérale post-thyroïdectomie reste relativement faible.

Le taux de paralysie récurrentielle en postopératoire immédiat (soit un mois après l'opération), après une opération de la glande thyroïde est compris entre 0,5% et 8,3%. (voir par exemple (Benninger, Gillen, & Altaian, 1998) ou (Lo, Kwok, & Yuen, 2000). De fait, il n'a pas été possible d'acquérir des données en phase préopératoire. La voix des patients a été comparée avec celle de locuteurs contrôles (dorénavant LC), appariés en sexe et en âge avec les patients correspondants. Les locuteurs contrôles constituent alors la voix de référence.

2.2 Corpus

Notre corpus est constitué de douze logatomes (ou séquence VCV), insérés dans une phrase porteuse : par ex : « Cet VCV ça. ». Les logatomes sont créés comme suit : si V1 est /i/ alors V2 est /a/, et vice versa. La consonne est l'une des 6 occlusives du français /p, t, k, b, d, g/. Ces consonnes ont été choisies car elles permettent l'alternance de séquences entièrement voisées V1 [b d g] V2 et de séquences mixtes V1 [p t k] V2, où les plis vocaux ne sont pas sollicités de façon continue. Les douze logatomes sont donc les suivants : [abi], [adi], [agi], [api], [ati], [aki], [iba], [ida], [iga], [ipa], [ita] et [ika]. Il s'agit pour le locuteur de lire les logatomes qui lui sont présentés dix fois chacun, la présentation étant faite de façon aléatoire.

2.3 Mesures

Le Voice Onset Time (V.O.T.), ou délai d'établissement du voisement, a été mesuré selon Klatt (1975), soit comme l'intervalle allant du relâchement consonantique jusqu'à l'apparition de la structure formantique stable de la voyelle subséquente. Rappelons que Liberman et al., (1958) notent le V.O.T. comme un indice majeur de la classification perceptive des occlusives voisées/non voisées.

Les durées du V.O.T. obtenues ont également été normalisées en calculant le pourcentage pris par cet intervalle dans les différentes réalisations de la tenue consonantique entière. La tenue consonantique est considérée comme l'intervalle allant de la fin de la structure formantique stable de V1 au début de la structure formantique stable de V2. Le résultat ainsi obtenu est une valeur exprimée en pourcentage. La durée relative du V.O.T. a été calculée comme suit : Durée absolue du V.O.T. X 100/Tenue consonantique. Il nous semble important de prendre en considération les valeurs normalisées en production de la parole afin de pouvoir évaluer correctement les effets de l'élasticité du signal temporel de la parole (Gaitenby, 1965 ; Sock, 1998).

Le Voice Termination Time (V.T.T.), ou délai d'arrêt du voisement, a été décrit par Agnello (1975) comme l'intervalle allant de la closion articulatoire, ou la pression intra-orale maximale, jusqu'à la dernière vibration glottique. Pour certains auteurs, le V.T.T. coïnciderait avec le temps nécessaire à l'arrêt de vibration des plis vocaux, après la closion dans le conduit vocal.

3 Hypothèses

Etant donné qu'une paralysie récurrentielle entraine des modifications de la périodicité de la voix et des difficultés dans le contrôle du mouvement des plis vocaux et donc du voisement, nous formulons les hypothèses suivantes :

Les difficultés de voisement rencontrées par les patients pourraient avoir un impact sur la durée du V.O.T. En effet, celle-ci considérée comme un indice de vibration des plis vocaux et pouvant permettre ainsi la distinction des consonnes voisées des consonnes non voisées, serait modifiée.

Traditionnellement, le V.O.T. de Klatt (1975), allant du relâchement supraglottique de la consonne à l'apparition d'une structure vocalique formantiquement stable, est plus court pour les occlusives voisées que pour les occlusives non voisées. Pour les occlusives voisées, la transition CV implique uniquement un changement d'un état obstrué du conduit vocal à un état suffisamment dégagé de celui-ci pour permettre l'apparition d'une structure formantique claire. En revanche, pour les occlusives non voisées, cette transition CV exige, outre le changement d'état du conduit vocal, une modification de la configuration de la glotte d'une position ouverte pour la consonne non voisée à une position fermée pour la voyelle. En conséquence, en raison de difficultés de maintien de vibrations laryngées, le V.O.T. devrait alors être plus long pour les occlusives voisées (plus ou moins désonorisées, et produites avec difficulté par les patients) par rapport à celles produites par les locuteurs contrôles. Ainsi, la durée du V.O.T. de Klatt (1975) des occlusives voisées des patients pourrait alors être comparable à celle mesurée pour leurs occlusives non voisées, rendant ainsi la distinction entre ces deux catégories potentiellement difficiles dans une perspective quantitative. Il s'agira également d'observer la robustesse des potentielles différences de durées observées en valeurs absolues du V.O.T. face à l'élasticité du signal. Une attention particulière sera alors accordée aux durées en valeurs relatives, afin de comprendre si l'observation précédente est significativement pertinente ou simplement liée à des stratégies différentes en matière de vitesse d'élocution. Il est en effet possible que la paralysie récurrentielle contraigne les locuteurs à réajuster leurs vitesses d'élocution, ce qui se traduirait par des compressions/expansions du signal de parole.

Le V.T.T. (Agnello, 1975), reflet du timing oro-laryngé, pourrait être modifié par la paralysie récurrentielle. Rappelons qu'il s'agit en effet d'un paramètre qui correspond, dans une transition VC, à l'intervalle allant de la disparition de la structure formantique clairement définie de la voyelle (disparition provoquée par la closion supraglottique de la consonne), à l'arrêt du voisement dans la phase consonantique. Dans la mesure où ce paramètre est lié à l'arrêt des vibrations périodiques, donc par extrapolation à l'amortissement des vibrations des plis vocaux (Sock, 1998), il devrait être modifié par la paralysie récurrentielle, avec un délai d'arrêt du voisement plus court pour les patients par rapport aux locuteurs contrôles.

En raison de stratégies compensatoires différentes selon les individus, la variabilité inter locuteurs devrait être plus marquée dans les phases d'enregistrement post-opératoires précoces (post-opératoire 1 notamment). Il est également possible que lors de certaines productions, le locuteur se rapproche des cibles attendues, alors que pour d'autres il en serait particulièrement éloigné. La variabilité intra locuteurs pourrait alors également être importante.

4 Résultats

4.1 Durées absolues du V.O.T.

En contexte [a-i], l'interaction entre les trois facteurs, voisement, lieu d'articulation et phases d'enregistrement, est statistiquement significative [$F(6,36) = 51{,}41$ $p < 0{,}032$]. L'interaction repose notamment sur le fait que, quelle que soit la consonne voisée considérée, les productions des locuteurs contrôles sont significativement différentes de celles des patients, lors de la phase d'enregistrement post-opératoire 1. A partir de la phase d'enregistrement post-opératoire 3, les tests post-hoc indiquent que le facteur de voisement ne permet plus de distinguer les productions des locuteurs contrôles de celles des patients, pour les consonnes voisées. En revanche, les consonnes non voisées des locuteurs contrôles et des patients en phases post-opératoires 1,2 et 3 ne

sont pas significativement différentes. En ce qui concerne le facteur lieu d'articulation, le test post-hoc révèle qu'en post-opératoire 1, la variable VOT ne permet ne permet pas toujours de distinguer les consonnes les unes des autres. Il n'existe alors plus de différence significative entre [d] et [g]. Enfin, signalons que le facteur lieu d'articulation pour la variable VOT indique que leurs productions sont significativement différentes en fonction de la consonne étudiée.

Nous avons observé un effet principal du voisement [$F(1,6) = 105,08$ $p < 0,0002$]. Suite aux tests LSD, nous pouvons affirmer que cet effet principal de voisement repose, comme attendu, sur une différence significative entre les consonnes voisées et les consonnes non voisées. Une différence significative a également pu être mise au jour pour le facteur lieu d'articulation [$F(2,12) = 77,03$ $p < 10^{-5}$]. Les tests post-hoc révèlent que la durée du VOT est significativement différente en fonction du lieu d'articulation des diverses occlusives. Ainsi, la durée du VOT de [p b] est différente de celle de [t d], qui est elle-même différente de celle de [k g]. Un effet principal de la phase a pu être mis au jour pour la variable VOT en contexte [a-i] [$F(3,18) = 188,02$ $p < 0,013$]. La phase d'enregistrement post-opératoire 1 est significativement différente de toutes les autres phases d'enregistrement.

En contexte [i-a], on n'observe pas d'interaction significative entre les trois facteurs, voisement, lieu d'articulation et phases d'enregistrement (p=ns).

En post-opératoire 1, la durée du V.O.T. des patients est plus longue pour les trois occlusives voisées [b d g] par rapport à celle des locuteurs contrôles, et cela en contextes [a-i] et [i-a]. Cette phase d'enregistrement est également caractérisée par des écarts-types plus importants, ce qui témoigne d'une certaine variabilité dans les productions des patients. Notons que la différence entre la durée du V.O.T. des patients et des locuteurs contrôles est plus grande dans le contexte [a-i]. Il est donc probable que la transition [C=> i] exige un contrôle plus précis et plus problématique que la transition [C=> a]. En effet, la voyelle [i] requérant une constriction réduite dans la zone alvéolaire, sans qu'il y ait de contact entre la langue et les alvéoles, les patients semblent avoir plus de difficulté à réaliser cette transition de la consonne vers la voyelle cible [i]. Cela ne semble pas être le cas pour le [a], dont l'abaissement mandibulaire, accompagné de celui de la masse de la langue avec qui la mandibule est fortement couplée, crée sans difficulté notoire la constriction radico-pharyngale nécessaire pour l'émergence acoustique de cette voyelle.

Les difficultés de voisement rencontrées par les patients, suite à la paralysie récurrentielle ont donc un impact sur les valeurs de V.O.T. uniquement pour les consonnes voisées. Les difficultés de maintien des vibrations laryngées conduisent les patients à produire des V.O.T. plus longs. En effet, les patients sont contraints d'arrêter l'activité laryngienne périodique pour la consonne voisée plus tôt, permettant ainsi un relâchement consonantique relativement précoce. Cela a pour conséquence d'augmenter la durée du V.O.T. Avec le temps et la rééducation vocale, une réduction des valeurs de V.O.T. est observée, et cela dès la phase d'enregistrement post-opératoire 2, ce qui laisse supposer une amélioration du contrôle des mouvements laryngés. Ce phénomène se poursuit jusqu'en post-opératoire 3, où les valeurs des V.O.T. sont alors comparables à celles des locuteurs contrôles, notamment pour les occlusives [b] et [d].

L'étude des productions des patients pris individuellement a permis de montrer que l'importante variabilité relevée en post-opératoire 1 était le reflet de différences inter et intra-individuelles. De façon générale, cette variabilité diminue lors des phases d'enregistrement plus tardives, ce qui va dans le sens d'un meilleur contrôle des vibrations laryngées : les productions sont alors plus régulières.

La durée du V.O.T. des occlusives non voisées n'a pas permis de mettre au jour des différences significatives entre les productions des patients et des locuteurs contrôles, et ce quelle que soit la phase d'enregistrement étudiée.

Rappelons que le paramètre du V.O.T. peut varier en fonction du lieu d'articulation. En effet, il a été observé, entre autres, que plus la closion a lieu à l'arrière de la cavité buccale, plus le V.O.T. est long (Peterson & Lehiste, 1962). Cela voudrait dire que plus l'aire du contact est étendue, plus le V.O.T. est élevé (Stevens, 1989). Ainsi le V.O.T. est généralement plus court pour les consonnes bilabiales et plus long pour le [k], le [t] ayant une durée intermédiaire dans la plupart des cas. Par ailleurs, on sait que le V.O.T. de Klatt (1975) est plus long pour les occlusives non voisées que pour les occlusives voisées (Sock & Benoit, 1986). Traditionnellement en français, la durée du V.O.T. d'un [b] est plus petite que celle du V.O.T. d'un [d], qui est elle-même plus petite que celle d'un [g]. Le V.O.T. est également plus court pour un [p] que pour un [t] ou un [k]. Si les données qui viennent d'être mentionnées valent pour des sujets sans trouble de la parole, la figure 1 indique que ces constats se maintiennent chez des locuteurs présentant une paralysie récurrentielle après thyroïdectomie, et cela pour les deux contextes étudiés, ainsi que pour toutes les phases d'enregistrement. Si la durée du V.O.T. est effectivement allongée en post-opératoire 2 pour la production des occlusives voisées des patients, celles-ci n'en restent pas moins quantitativement différenciées.

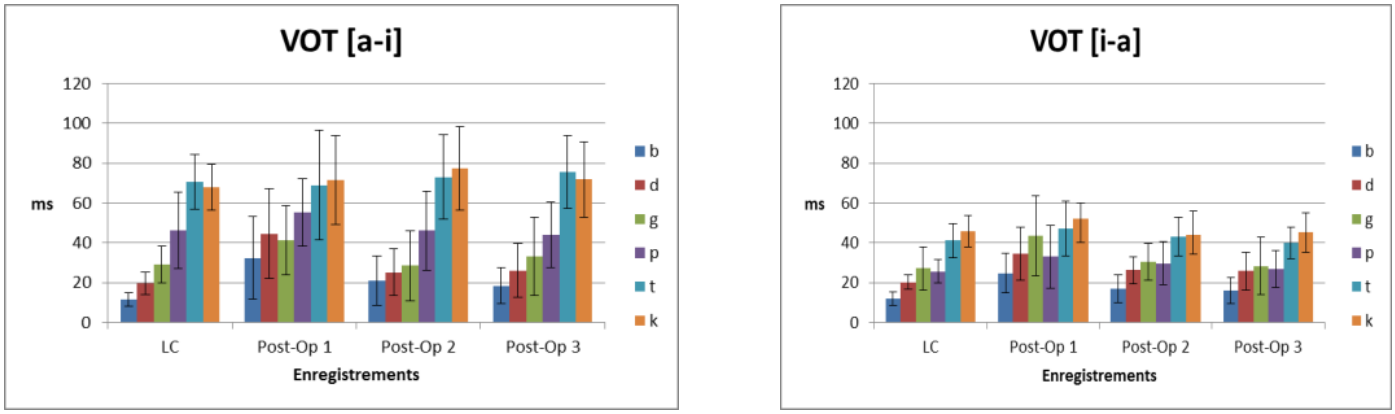


Figure 1 : Durées moyennes du V.O.T. (Klatt, 1975) pour six occlusives du français en contexte [a-i] (à droite) et en contexte [i-a] (à gauche)

4.2 Durées relatives du V.O.T.

Les observations faites pour les valeurs absolues sont confirmées lors de l'étude des valeurs relatives. Les durées relatives du V.O.T. des consonnes voisées restent plus longues pour les patients présentant une paralysie récurrentielle que celles des locuteurs contrôles. Cela est particulièrement remarquable en contexte [a-i]. Notons que la dernière phase est également caractérisée par une variabilité réduite.

Les mêmes observations, dans une moindre mesure toutefois (p=*ns*), sont observables en contexte [i-a]. Ces constats étaient notés en valeurs absolues également. Il est donc probable que la transition [$C_{voisée}$=> i] exige un contrôle plus précis et plus problématique que la transition [$C_{voisée}$=> a]. La durée relative du V.O.T. de [b d g] est plus importante lors de la phase post-opératoire 2. Ces valeurs se maintiennent en post-opératoire

Les durées relatives du V.O.T. pour les consonnes non voisées ne permettent pas de distinguer les différentes phases d'enregistrement (p=*ns*), et ce dans les deux contextes [a-i] et [i-a]. Les écarts-

types les plus importants sont relevés lors de la phase d'enregistrement post-opératoire 1, la variabilité se réduisant dans les phases d'enregistrement suivantes.

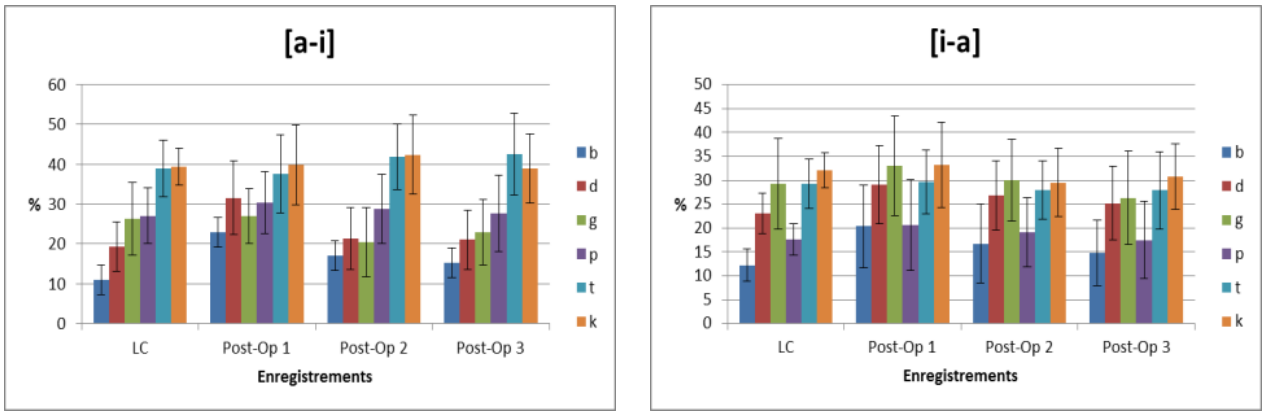


Figure 2: Durées relatives moyennes du V.O.T. (Klatt, 1975) pour six occlusives du français en contexte [a-i] (à droite) et en contexte [i-a] (à gauche)

4.3 Durées absolues du V.T.T.

L'étude du V.T.T. n'a pas permis de montrer de modification significative de ce paramètre suite à la paralysie récurrentielle post-thyroïdectomie. Dans cette section, nous proposons d'étudier cet autre paramètre intrasegmental, le V.T.T. (Agnello, 1975), considéré comme un indice du timing oro-laryngé. Ce paramètre est lié à l'arrêt des vibrations périodiques, donc par extrapolation, à l'amortissement des vibrations laryngées (Sock, 1998), après obstruction du conduit vocal. L'étude de cet intervalle dans les séquences prononcées par les locuteurs atteints d'une paralysie récurrentielle, suite à une thyroïdectomie, indique qu'il n'est pas modifié par la paralysie récurrentielle. Les productions des patients, lors des phases post-opératoires, restent en effet proches des valeurs relevées pour les productions des locuteurs contrôles, soit environ 20 ms dans tous les contextes étudiés. Les productions des patients se distinguent toutefois par des écarts-types légèrement plus importants, reflets d'une variabilité plus grande dans l'ensemble des phases d'enregistrement post-opératoires. Néanmoins, les écarts-types ne permettent pas de dégager des tendances claires pour ce paramètre.

L'étude des intervalles vocaliques, qui s'étendent de la fin de la structure formantique stable de V1 au début de la structure formantique stable de V2, dans les séquences prononcées par les locuteurs atteints d'une paralysie récurrentielle, suite à une thyroïdectomie, indique qu'ils ne sont pas modifiés (p=*ns*) par l'immobilité laryngée. Le contexte vocalique, [a-i] ou [i-a], n'a pas permis de montrer des différences, entre locuteurs contrôles et patients, en ce qui concerne la durée de l'intervalle vocalique. Enfin, les écarts-types les plus importants sont apparu dans les productions des patients, notamment lors de la phase d'enregistrement post-opératoire 1. La variabilité tend à se réduire au cours du temps.

5 Discussion

Il convient à présent de vérifier si nos hypothèses initiales ont été confirmées ou infirmées, en ce qui concerne le timing des paramètres retenus.

Les difficultés de voisement rencontrées par les patients ont un impact sur la durée du V.O.T. (*Hypothèse 1*). En raison de difficultés de maintien de vibrations laryngées, le V.O.T. est plus long pour les occlusives voisées, produites par les patients que celles produites par les locuteurs

contrôles. En effet, les vibrations périodiques s'estompent plus tôt, « laissant ainsi la place », de manière précoce, au relâchement consonantique. Cela est particulièrement vrai lors de la phase d'enregistrement post-opératoire 1. La durée du V.O.T. de (Klatt, 1975) des occlusives voisées des patients reste quantitativement différente de celle mesurée pour leurs occlusives non voisées ; la distinction entre ces deux catégories devrait rester possible dans une perspective perceptive. Signalons que dans certains cas, les occlusives voisées, produites par les patients dans les phases d'enregistrement post-opératoires précoces, sont plus ou moins désonorisées puisque produites avec une certaine difficulté de maintien de l'activité laryngienne. Les différences de durées observées en valeurs absolues dans l'hypothèse 1 sont maintenues en valeurs relatives. Rappelons que le V.O.T. est plus long en post-opératoire 1 pour les occlusives voisées des patients. Il semble donc que l'allongement du V.O.T. pour les occlusives voisées dans les productions des patients en post-opératoire 1 soit une manœuvre robuste.

Le V.T.T. (Agnello, 1975), reflet également du timing oro-laryngé, n'est pas modifié par la paralysie récurrentielle. Les productions des patients restent proches de celles des locuteurs contrôles, et ce quels que soit le contexte et la phase d'enregistrement considérés. L'amortissement des vibrations laryngées, quantifiable à partir de l'intervalle nécessaire à l'arrêt des vibrations périodiques, soit la mesure du V.T.T. (Sock, 1998) ne varie pas avec la paralysie récurrentielle. Ces résultats infirment *l'hypothèse 2.* Nous supposions qu'en raison des difficultés de maintien du voisement, le V.T.T. des patients serait plus court que celui des locuteurs contrôles.

Nous avons vu que, quel que soit le paramètre étudié, la variabilité inter et/ou intra-locuteurs est toujours plus marquée pour tous les paramètres chez les patients. même si les productions des patients sont apparemment proches de celles des contrôles, leurs productions restent tout de même marquées par une variabilité plus importante. La phase post-opératoire 2, notamment, indique des écarts-types souvent plus importants que ceux mesurés chez les locuteurs contrôles, et que ceux observés dans les phases d'enregistrement post-opératoires suivantes. La variabilité inter et intra individuelle se réduit dès la phase post-opératoire 2. Cette observation, qui se confirme en post-opératoire 3, témoigne d'une régularité retrouvée dans les productions des patients, et pourrait être un signe de récupération d'un timing oro-laryngé « standard ».

6 Conclusions

Un V.T.T. apparait parfois dans certaines occlusives voisées produites par les patients, témoin d'une certaine désonorisation de celles-ci. Ce phénomène reste toutefois à quantifier. Puisque cet intervalle n'a pas été modifié dans les séquences non voisées produites par les patients, il est possible que les difficultés des patients reposent plutôt sur le maintien du voisement que sur son arrêt.

D'autres intervalles spatio-temporels acoustiques, comme la durée de l'occlusion ou la tenue consonantique et vocalique, devraient également nous permettre de mieux appréhender de façon plus globale les stratégies de réajustements que les patients mettent en place suite à l'apparition d'une paralysie récurrentielle. Il serait utile aussi de procéder à une vérification de la validité perceptive de ces ré-organisations spatiotemporelles en conduisant des tests de perception.

Remerciements

Programme de la MISHA, 2013-2017, ANR "DOCVACIM" et le projet du CS de UdS Gutenberg-Strasbourg, 2009-2011.

Références

ABRY, C., AUTESSERRE, D., BARRERA, C., BENOIT, C., BOE, L. J., CAELEN, J., VIGOUROUX, N. (1985). Propositions pour la segmentation et l'étiquetage des sons du français. In 14e JEP du GCP du GALF (p. 156‑163).

AGNELLO, J. (1975). Voice onset and voice termination features of stutterers. In Vocal tract dynamics and dysfluency: the proceedings of the first annual Hayes Martin Conference on Vocal Tract Dynamics. Speech and Hearing Institute.

BENNINGER, GILLEN, & ALTAIAN. (1998). Changing etiology of vocal fold immobility. The Laryngoscope, 108(9), 1346‑1350.

FAUTH, C. (2012). Perturbation de la production de la parole suite à une opération de la glande thyroïde. Université de Strasbourg, Strasbourg.

GAITENBY, J. (1965). The elastic word. Speech Research, 2, 1‑12.

HARTL, D. (2004). Caractérisation de la voix et de la parole dans les paralysies récurrentielles unilatérales. Université de Paris III-Sorbonne Nouvelle, Paris.

KLATT, D. (1975). Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. J Speech Hear Res, 18(4), 686‑706.

LACCOURREYE, O., LE CLECH, G., & SANTINI, J. (2008). Complications de la chirurgie des affections et tumeurs bénignes de la thyroïde. In Les risques chirurgicaux en oto-rhino-laryngologie : information, prise en charge et prévention (Société francaise d'oto-rhino-laryngologie., p. 369‑386). Paris: Laccourreye O, Chabolle F.

LAPRIE, Y., & BUSSET, J. (2011). Construction and evaluation of an articulatory model of the vocal tract. Présenté à 19th European Signal Processing Conference - EUSIPCO-2011.

LIBERMAN, A. M., DELATTRE, P. C., & Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. Language and Speech, 1, 153_167.

LO, C.-Y., KWOK, K.-F., & YUEN, P.-W. (2000). A Prospective Evaluation of Recurrent Laryngeal Nerve Paralysis During Thyroidectomy. Arch Surg, 135(2), 204‑207.

PERRIER, P., PAYAN, Y., BUCHAILLARD, STÉPHANIE, I., NAZARI, M. A., & CHABANAS, M. (2011). Biomechanical models to study speech. Faits de Langues, 37, 155‑171.

Peterson, G., & Lehiste, I. (1962). Duration of syllable Nuclei in American English. Journal of the Acoustical Society of America, (32), 693‑703.

SOCK, R. (1998). Organisation temporelle en production de la parole: émergence de catégories sensori-motrices phonétiques. Presses univ. du Septentrion, Université Stendhal, Grenoble III,.

SOCK, R., & BENOIT, C. (1986). VOT et VTT en français. In 15 èmes Journées d'Etude sur la Parole du Groupe Communication Parlée (p. 307‑310). Aix-en-Provence.

STEVENS, K. (1989). On the quantal nature of speech. Journal of Phonetics, (17), 3‑45.

Evolution des activations cérébrales lors de la production de parole après exérèse au niveau de la cavité orale

Audrey Acher[1] Marc Sato[1] Laurent Lamalle[2,3,4,5] Coriandre Vilain[1] Christophe Savariaux[1] Silvain Gerber[1] Arnaud Attyé[3,6] Alexandre Krainik[3,4,6,7] Georges Bettega[8] Christian Adrien Righini[9] Brice Carlot[10] Muriel Brix[11] et Pascal Perrier[1]

(1) GIPSA-Lab, UMR 5216 CNRS/Université Grenoble-Alpes, France
(2) Inserm, US 017 – Grenoble, France
(3) Université Joseph Fourier – Grenoble, France
(4) CHU de Grenoble, UMS IRMaGe – Grenoble, France
(5) CNRS, UMS 3552 – Grenoble, France
(6) Clinique Universitaire de Neuroradiologie et d'IRM – CHU Grenoble, France
(7) Grenoble Institute of Neurosciences, INSERM U836 – Grenoble, France
(8) Service de chirurgie plastique et maxillo-faciale – CHU Grenoble, France
(9) Clinique Universitaire d'ORL, Pôle Tête et cou et Chirurgie Réparatrice – CHU Grenoble, France
(10) Hôpital Privé de la Casamance – Aubagne, France
(11) Service de chirurgie maxillo-faciale et plastique – CHU Nancy, France

Audrey.Acher@gipsa-lab.grenoble-inp.fr, Pascal.Perrier@gipsa-lab.grenoble-inp.fr

RESUME

A partir de données d'IRM fonctionnelle, acoustiques et praxiques, nous avons étudié les corrélats cérébraux de l'adaptation de la parole après une modification structurelle importante du conduit vocal lors de trois tâches : mouvements oro-faciaux silencieux, voyelles et syllabes. Onze patients ont été enregistrés lors de trois sessions, pré-opératoire et post-opératoire à 1 mois et 3 mois (pour sept d'entre eux aussi à 9 mois). Onze sujets contrôles ont été enregistrés en parallèle. Une analyse de groupe (patients/contrôles) « cerveau entier » révèle des patrons d'activation spécifiques aux patients au cours des différentes sessions en particulier pour la tâche de production de voyelles. De plus, une moindre activation cérébrale par rapport aux sujets sains a été observée pour toutes les tâches dans des régions motrices, sensorielles et d'intégration sensori-motrice. Nous interprétons ces résultats en relation avec la redéfinition des buts de parole et l'adaptation de modèles internes du système moteur périphérique.

ABSTRACT

Evolution of brain activations during speech production after intra-oral surgery

Using functional MRI, acoustic data and motor oral assessment, brain correlates of speech recovery after dramatic structural changes in the vocal tract have been investigated during three tasks: orofacial movements and speech production (vowels and syllables). Eleven patients were recorded during three sessions, preoperatively and postoperatively, 1 month and 3 months after surgery (for seven of them also 9 months after surgery). Eleven healthy subjects were recorded in parallel. A "whole brain" group analysis (patients/healthy subjects) reveals a specific activation pattern for patients, in particular during vowel task, which is session-dependent. Moreover, global lower brain activation compared to healthy subjects was found for all tasks in primary, secondary and integrative sensorimotor regions. Differences are interpreted in relation with the emergence of new speech motor goals and changes in the internal models.

MOTS-CLES : chirurgie de la cavité orale, glossectomie, IRMf, analyse en région d'intérêt, analyse acoustique, adaptation de la parole, neurophonétique.
KEYWORDS: oral surgery, glossectomy, fMRI, Region of Interest analysis, acoustic analysis, speech recovery, neurophonetics.

1 Introduction

La chirurgie de la cavité orale occasionne notamment des difficultés durables de l'articulation de la parole. De plus, l'effet de la radiothérapie, qui a généralement lieu entre le 1er et le 3ème mois après l'opération, se rajoute à celui de la chirurgie. Le processus de récupération dépend du type de lésion, de l'âge des patients et de l'impact de la chirurgie sur les muscles de la cavité orale (Buchaillard et al., 2007). Les patients doivent donc adapter leur parole à la nouvelle configuration de leur cavité orale. Haupage et al., 2010, ont réalisé une étude en IRM fonctionnelle (IRMf) sur six patients afin de connaître les activations cérébrales associées à la déglutition avant et six mois après glossectomie partielle. Les résultats montrent qu'après chirurgie les patients présentent une forte activation cérébrale dans le lobule pariétal supérieur (LPS), l'aire motrice supplémentaire (AMS) et le cortex cingulaire antérieur (CCA) par rapport à la condition pré-chirurgicale et en comparaison avec un groupe de sujets sains. De plus, une analyse en région d'intérêt (ROI) au niveau de l'aire linguale du gyrus précentral (M1) a montré un volume d'activation cérébrale moyen plus faible pour les patients par rapport aux sujets sains quelle que soit la session d'enregistrement. Le but de notre étude est d'étudier l'adaptation de la parole pour un groupe de patients opérés de la cavité orale.

2 Fondements théoriques et hypothèses

Des changements sont attendus dans les zones sensori-motrices classiquement associées aux mouvements oro-faciaux, le cortex moteur primaire (M1) dans le gyrus précentral, le cortex somatosensoriel primaire (S1) dans le gyrus postcentral, l'AMS et le cortex prémoteur dans le cortex frontal postérieur et moyen. Mais nous attendons aussi des patrons d'activation plus spécifiques à la parole, avec une hypothèse forte au niveau des aires temporo-pariétales, en particulier au niveau du lobule pariétal inférieur (LPI) où sont traitées conjointement les informations sensorielles et motrices (Hickok et Poeppel, 2007). La définition de nouveaux buts articulatori-acoustiques, adaptés à la nouvelle morphologie du conduit vocal, pourrait ainsi être mise en évidence par une modulation spécifique des activations dans cette zone. De même, la réappropriation du conduit vocal perturbé pourrait impliquer une modulation de l'activité au niveau du cervelet, où s'élaborent les modèles internes du système moteur périphérique. Ces modèles internes transmettent à M1 les informations permettant de générer les commandes motrices appropriées à la réalisation d'un but moteur (Wolpert et al., 1998). Ils sont également impliqués dans la correction de ces commandes par des mécanismes de prédiction (Kawato, 1999).

3 Matériel et méthode

3.1 Evaluation IRMf

Les acquisitions des images IRM ont eu lieu sur l'imageur corps entier 3T (Philips Achieva TX) de la plateforme IRMaGe de Grenoble. Pour l'IRMf nous avons utilisé en réception une antenne « tête » à 32 canaux. Lors de l'expérience, les sujets étaient en position allongée, leur tête était

maintenue dans l'antenne avec des mousses latérales. Ils portaient des bouchons d'oreille et un système casque-microphone compatible IRM utilisé pour communiquer avec eux et enregistrer leurs productions. Les consignes visuelles de la tâche ont été projetées à l'aide du logiciel Presentation (Neurobehavioral Systems, Albany, EU) sur un écran situé derrière le sujet allongé qui pouvait le voir par réflexion sur un miroir placé à l'aplomb de ses yeux. Un monitoring des réalisations des tâches de parole était effectué par un expérimentateur en temps réel. Les enregistrements en IRMf étaient basés sur le paradigme de *sparse sampling* (Gracco et al., 2005), afin de minimiser de possibles artefacts de mouvement sur les images fonctionnelles et d'éviter la production de parole dans le bruit. Cette technique exploite le délai existant entre l'activité neuronale liée à une tâche motrice et la réponse hémodynamique associée (signal BOLD). Les paramètres des images fonctionnelles étaient les suivants : séquence d'acquisition EPI (imagerie échoplanaire) en écho de gradient FE-EPI (pondérée en T2*), temps de répétition 10 s[1], temps d'acquisition 2.7 s, temps d'écho 30 ms, angle de rotation 90°. Les paramètres d'encodage spatial de base étaient les suivants : 53 coupes d'épaisseur 3 mm, adjacentes, parallèles au plan bi-commissural CA-CP, acquises en mode non entrelacé avec une résolution dans le plan de 3 mm isotrope (champ de vue 216 mm de côté, encodé par une matrice de 72 x 72). Pour l'IRM anatomique des images tridimensionnelles à haute résolution spatiale (1 mm isotrope), pondérées en T1 ont été acquises. 3 tâches motrices, une tâche oro-faciale silencieuse et 2 tâches de parole, devaient être réalisées de façon aléatoire dans 2 runs fonctionnels de 14 minutes chacun. La tâche oro-faciale silencieuse était explicitement présentée comme « non parole » et impliquait soit la langue (rétraction et protraction), dans une tâche considérée comme complexe pour les patients soit les lèvres (étirement et protrusion, tâche plus simple). Certaines tâches de production de parole (voyelles et syllabes) n'exigeant pas une forte précision articulatoire étaient considérés comme simples (/a/, /ə/ ; /fa/, /pa/) ; d'autres plus complexes car nécessitant un positionnement lingual précis (/i/, /u/ ; /sa/, /ʃa/). Une tâche de repos (sans mouvement ni production sonore) servait de tâche de référence. Une consigne visuelle d'une durée d'une seconde indiquait pour chaque essai le stimulus à produire. Chaque tâche était produite à partir d'une position initiale bouche fermée, mandibule et langue relâchées, vers laquelle le sujet retournait après la tâche. Un item était produit toutes les 10 secondes selon un ordre pseudo-aléatoire et maintenu pendant une seconde. Pour les syllabes, il était demandé aux locuteurs de maintenir la consonne. Les participants savaient qu'ils ne devaient pas bouger afin d'éviter les artéfacts de mouvement. Ils ont été entraînés à réaliser les différentes tâches quelques jours avant la date de l'expérience et le jour de l'expérience. Aucun participant n'a fait part de difficulté à réaliser les tâches. Chaque tâche a été répétée 24 fois. 168 scans fonctionnels ont ainsi été acquis (6 tâches * 24 répétitions + 1 tâche de repos * 24 répétitions) pour une durée totale d'environ 50 minutes. 3 scans ont été ajoutés au début de chaque session pour équilibrer le signal IRM. Ils ne sont pas inclus dans les analyses.

3.2 Evaluation acoustique et praxique

Les données acoustiques du protocole ont été acquises dans une pièce calme avec un micro électroacoustique placé à environ 20 cm de la bouche du sujet et relié à un enregistreur PMD MARANTZ 671. La fréquence d'échantillonnage choisie était de 44.1kHz. Le corpus comprenait les 10 voyelles du français [i, e, ɛ, a, y, ø, œ, u, o, ɔ] en isolé et 12 consonnes du français [p, b, t, d, s, z, ʃ, ʒ, k, g, l, ʁ] en contexte symétrique aCa, iCa et aCi. 10 répétitions de chaque item ont été enregistrées. L'évaluation praxique a pour but d'examiner les mouvements des articulateurs de la

[1] Le temps de répétition de 10 secondes est dû à l'utilisation du *sparse sampling*.

parole lors de la réalisation de tâches motrices simples à complexes[2]. Elle évalue le mouvement de 3 articulateurs (mandibule, lèvres et langue) selon une cotation à 5 degrés (0 point pour un mouvement impossible, 1 point pour un mouvement très altéré, 2 points pour un mouvement modérément altéré, 3 points pour un mouvement légèrement altéré et 4 points pour un mouvement correct). La sous-partie mandibule est notée sur 40, celle sur les lèvres sur 32, et celle sur la langue sur 96. Les sous-totaux ont été additionnés afin d'obtenir un score global sur 168.

3.3 Participants

11 patients (dont quatre femmes ; moyenne d'âge : 56.18 ans) et 11 volontaires sains appariés en âge et en genre ont donné leur consentement écrit pour leur participation à l'étude. Une visite médicale de pré-inclusion a été réalisée par un médecin afin de vérifier que les participants ne présentaient aucune contre-indication à l'étude. Tous les participants avaient une vision normale ou corrigée, étaient de langue maternelle française et aucun antécédent de troubles du langage, d'audition, de déficit neurologique ou de pathologie psychiatrique n'a été rapporté. La latéralité droite a été vérifiée grâce à la version française du questionnaire de latéralité manuelle d'Olfield (Olfield, 1971). Différents types de chirurgie ont été réalisés en fonction de la localisation de la tumeur cancéreuse. 4 patients ont subi une hémiglossectomie (exérèse de l'hémi-langue dans le sens longitudinal) ; 3 patients ont subi une pelvi-glosso-mandibulectomie interruptrice (exérèse d'une partie du plancher de la bouche, d'une partie de la langue et d'une partie de la mandibule interrompant sa continuité) ; enfin, 4 pelvi-glossectomies ont été réalisées (exérèse d'une partie du plancher et d'une partie de la langue). Les patients inclus ne présentaient pas d'antécédent de cancer des voies aéro-digestives supérieures ni de maladie neurologique. 9 des patients de la cohorte ont suivi 6 semaines de radiothérapie post-opératoire. Cette étude a reçu un avis favorable du Comité de Protection des Personnes Sud Est V et de l'Agence Française de Sécurité Sanitaire des Produits de Santé. 7 patients ont été suivis selon le protocole longitudinal complet au cours de 4 sessions d'enregistrement. La cohorte complète de patients (11) a passé les trois premières sessions. Les 11 sujets sains ont passé seulement la partie IRMf du protocole.

4 Analyse des données

Les données ont été analysées à l'aide du logiciel SPM8 (Statistical Parametric Mapping; Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK) sous Matlab 7.13 (Mathworks, Natick, MA, USA) et du logiciel R (R Core Team, 2012).

4.1 Prétraitements individuels

Pour chacun des participants, les images fonctionnelles ont d'abord été réalignées après estimation des 6 paramètres de mouvements afin de contrôler les mouvements de tête entre les scans. L'image anatomique a été recalée sur l'image fonctionnelle moyenne après segmentation pour correspondre à un cerveau normalisé dans l'espace commun du Montreal Neurological Institute (MNI). Les images fonctionnelles ont ensuite été normalisées (repère MNI) et lissées via un filtre gaussien

[2] La grille d'évaluation a été réalisée en collaboration avec le Pr Vacher (Hôpital Beaujon APHP, Paris) et le Dr Deneuze (Centre Léon Bérard, Lyon) et est inspirée de la partie « Evaluation de la motricité bucco-linguo-faciale » du Test de récupération Fonctionnelle après chirurgie bucco-pharyngée (T.R.F.) élaboré par le département de phoniatrie-orthophonie du Centre Lacassagne de Nice.

passe-bas de 6 mm^3 afin d'augmenter le rapport signal sur bruit et de compenser la variabilité anatomique des cerveaux des participants. Pour chaque participant, les corrélats neuronaux reliés aux 6 tâches ont été analysés selon un modèle linéaire général (GLM ; Friston et al., 1995). Le modèle linéaire général inclut des régresseurs d'intérêt reliés aux tâches et des régresseurs de non-intérêt liés aux paramètres de réalignement ; les tâches de repos forment une ligne de base. Chaque tâche (régresseurs d'intérêt) était représentée par 24 images fonctionnelles. La réponse de type hémodynamique associée à chaque évènement a été modélisée par une réponse impulsionnelle finie de type impulsion unique (FIR) pour chaque scan fonctionnel. Avant l'estimation du modèle, un filtre des basses fréquences a priori non-reliées aux conditions expérimentales (variations lentes d'origine physiologique) a été appliqué (passe-haut de fréquence de coupure de 1/128 Hz). Des cartes d'analyse statistique individuelles ont été calculées pour chaque participant, pour les 6 tâches.

4.2 Analyse de groupe cerveau entier 11 patients / 11 sujets sains

Nous avons réalisé cette analyse sur les 3 premières sessions incluant le maximum de sujets. Suite à l'estimation pour chaque participant des activations observées lors de la production des différentes tâches par rapport à la condition de repos, une analyse de groupe considérant le sujet comme effet aléatoire a été réalisée via une ANOVA à mesures répétées. 7 contrastes 'F' ont été calculés pour mettre en évidence l'effet principal du groupe, de la session et de la complexité et des interactions entre ces trois facteurs. Ces analyses ont été effectuées en respectant un seuil statistique défini à $p<0.005$ non corrigé et une taille minimale des clusters de 10 voxels. Pour tous les contrastes, tous les pics d'activation ont été déterminés dans chaque cluster, leur localisation a ensuite été labellisée avec la boite à outils Anatomy de SPM (Eickhoff et al., 2005). Si une région n'avait pas pu être assignée avec Anatomy, elle a été déterminée avec le logiciel Talairach Daemon (Lancaster et al., 2000) grâce aux coordonnées du pic d'activation converties de l'espace MNI à l'espace stéréotaxique standard de Talairach & Tournoux (1988). Un masque explicite permettant de ne retenir que les activations situées dans la substance grise a été utilisé.

4.3 Analyse en régions d'intérêt avec intégration d'une covariable

Sur la base de données de la littérature sur les réseaux cérébraux de la production de la parole, des données confirmées par l'analyse cerveau entier de nos sujets, et de nos hypothèses sur les mécanismes sous-jacents à l'adaptation de la parole après chirurgie du conduit vocal nous avons sélectionné neuf régions d'intérêt bilatérales pour une analyse statistique spécifique : le lobule VI du cervelet contenant les représentations sensori-motrices de la région oro-faciale (Manni & Petrosini, 2004), M1 à l'origine des commandes motrices efférentes, S1 qui reçoit les informations oro-sensorielles, le gyrus temporal supérieur (GTS) qui reçoit les informations acoustiques, le gyrus frontal moyen (GFM) impliqué dans la planification motrice, l'AMS qui coordonne et séquence les actes moteurs complexes, le LPS impliqué notamment dans le séquencement des actions sensorimotrices, le LPI, contenant des zones d'interaction sensori-motrices responsables d'une redéfinition des buts sensori-moteurs de la parole, enfin la partie postéro-inférieure du gyrus frontal (pGFI, aire de Broca). Nous avons choisi de réaliser une analyse avec un modèle linéaire à effets mixtes. La définition du modèle a été réalisée avec la fonction lmer du package lme4 (Bates et al., 2013). Trois effets fixes, Session, Groupe et Complexité de la tâche ont été considérés. Un terme aléatoire, Sujet, a permis de prendre en compte le cas d'une analyse à mesures répétées. Nous avons réalisé une analyse séparée pour chaque région cérébrale. Nous avons étudié l'effet de

ces facteurs sur l'intensité du signal BOLD. L'inspection visuelle des résidus n'a pas révélé de déviations d'homoscédasticité ni de normalité. Les p-valeurs ont été obtenues grâce à la fonction pamer.fnc du package LMERConvenienceFunctions (Tremblay et al., 2013). Les tests *a posteriori* ont été réalisés avec la fonction glht du package multcomp (Hothorn et al., 2008). Le seuil statistique a été fixé à $p \leq 0.05$, seuil classiquement utilisé en sciences humaines. L'analyse en régions d'intérêt présentée ici a été réalisée en intégrant une covariable par tâche prenant en compte l'impact de la chirurgie sur la parole et son amélioration. L'analyse a été réalisée pour les 7 patients et les 7 sujets sains ayant passé les 4 sessions d'enregistrement. La covariable pour la tâche de voyelles a été calculée à partir de l'aire du triangle vocalique ; pour la tâche syllabique, à partir de la différence des centres de gravité fréquentiels de /s/ et /ʃ/ et ; pour la motricité oro-faciale avec le score à l'examen des praxies bucco-linguo-faciales. Nous avons défini une covariable par patient, à partir de deux sous-indices (indice de dégradation calculé en faisant une différence de score entre les deux premières sessions et indice d'amélioration en faisant une différence de score entre la session à 1 mois et la session à 9 mois). Un indice élevé témoigne d'un fort impact de la chirurgie et d'une récupération significative. Un indice moyen témoigne d'un impact important de la chirurgie mais d'une récupération insatisfaisante. Un indice faible traduit un faible impact de la chirurgie. Pour les sujets sains, cet indice est égal à 0. Ainsi, tous les sujets sont caractérisés avec une variable continue, dont la valeur est d'autant plus haute que les mécanismes cérébraux engagés dans l'adaptation sont complexes (le facteur Groupe n'est ainsi plus utile).

5 Résultats

Dans cette section seuls les résultats statistiquement significatifs seront présentés.

5.1 Analyse de groupe cerveau entier 11 patients / 11 sujets sains

Avant de débuter l'analyse des effets principaux, nous avons vérifié les effets simples sur la population de patients et de sujets sains. Ils sont congruents avec le réseau du contrôle moteur de la parole classiquement observé dans des études précédentes (Grabski et al., 2013, Acher et al., 2012). Pour les trois tâches, le réseau implique des activations bilatérales dans les zones motrices, sensorielles et d'intégration sensori-motrice. Pour les tâches de parole, en plus de ces activations, on trouve une activation du cortex auditif (GTS et gyrus de Heschl).

5.1.1 Effet principal du groupe

L'effet majeur est, dans certaines régions, une activation cérébrale inférieure chez les patients par rapport aux sujets sains pour toutes les sessions. Ce phénomène touche, pour toutes les tâches, le cortex sensori-moteur bilatéral et l'opercule pariétal droit, l'insula (à gauche pour la parole, à droite pour les mouvements silencieux) et le gyrus supramarginal droit. Pour la parole spécifiquement, on retrouve en plus, parmi ces régions, le GTS et le LPS (droit pour les voyelles, bilatéral pour les syllabes), le GFI, le GFM et le gyrus angulaire gauches et les lobules IV, V, VI, VIIa et VIIIa du cervelet. Enfin, pour la tâche de mouvements oro-faciaux, le GFI (aire 9) droit présente aussi moins d'activation.

5.1.2 Interaction Groupe, Session et Complexité

Au cours de la session pré-opératoire, lors de la production des mouvements silencieux, nous

observons chez les patients une activité plus forte dans le LPS et le cortex cingulaire postérieur droits pour les mouvements linguaux comparés aux mouvements labiaux. Ce phénomène n'existe pas pour les sujets sains. A 1 mois, une activation moindre dans le GFM et le lobule VIII droits est observée chez les patients pour les voyelles complexes par rapport aux voyelles simples. A 3 mois, le LPS droit est de nouveau plus activé pour les patients lors de la production des mouvements linguaux. Au cours de cette session à 3 mois, pour la tâche de production des voyelles complexes, on observe une nouvelle fois plus d'activation pour les patients, dans les lobules VIII et IX du cervelet, le CCA et le GFM droits. Aucune interaction entre Groupe, Session et Complexité n'est observée pour les syllabes.

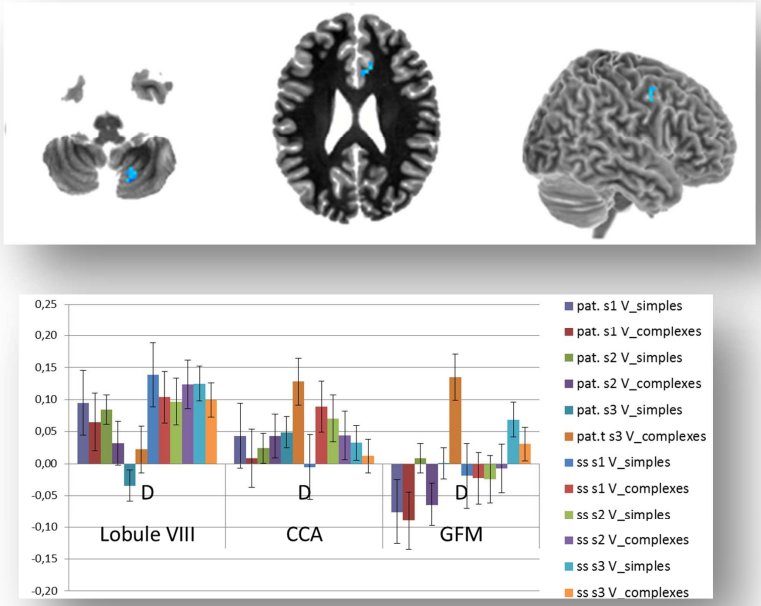


FIGURE 1: Interaction groupe, session et complexité lors de la production des voyelles ($p<0.005$ non corrigé, taille minimale des clusters : 10 voxels), abréviations : D : droit, CCA : cortex cingulaire antérieur, GFM : gyrus frontal moyen, pat. : patients, ss : sujets sains, s : session, V : voyelles. Réseaux cérébraux des activations observées, vues : z=-45, z=-25, hémisphère droit, en bleu foncé les pics les plus activés avec représentation graphique des niveaux d'activation, les barres d'erreur représentent une erreur type.

5.2 Analyse en régions d'intérêt prenant en compte les résultats des évaluations acoustiques et praxiques

L'analyse avec intégration de covariable pour les voyelles montre plusieurs interactions significatives entre la session et l'indice de parole. L'activation du LPI droit croit avec cet indice au cours de la session à 3 mois uniquement. Pour M1 à gauche, une décroissance de l'activation est observée avec cet indice au cours de la session pré-opératoire alors que cette activité réaugmente avec cet indice 9 mois après l'opération. Le même phénomène est observé pour l'AMS bilatérale. Enfin, pour l'activité dans le GTS gauche, l'influence de l'indice de parole se caractérise ainsi : diminution en session pré-opératoire, augmentation à 1 mois puis de nouveau diminution à 3 et 9 mois. Pour les syllabes, l'analyse avec intégration de covariable montre que l'activité dans le LPS bilatéral croit avec l'indice de parole dans la session post-opératoire à 1 mois et baisse pour la session à 3 mois.

6 Discussion et perspectives

La diminution d'activation globale pour toutes les sessions et toutes les tâches observée chez les patients au niveau du cortex sensori-moteur primaire et de l'opercule pariétal (considéré comme le cortex somatosensoriel secondaire) pourrait être révélateur d'une baisse du retour sensoriel consécutif à la présence de la tumeur, puis à l'exérèse chirurgicale et aux effets à court terme de la radiothérapie. Notons que Mosier et al. (2005) et Haupage et al. (2010) avait déjà relevé une activité réduite dans M1 pour des patients similaires par rapport à des sujets contrôles lors de tâches de déglutition. La réduction d'activation au niveau des autres zones suggère que le traitement de la maladie pourrait affecter globalement le réseau de perception-production de la parole.

Les deux types d'analyses, « cerveau entier » et « régions d'intérêt avec intégration d'une covariable », ont permis de mettre en évidence l'effet de la session et du type de tâche spécifiquement pour le groupe des patients. Pour les voyelles complexes, les résultats montrent une activation moindre du cervelet (lobule VIII droit) à 1 mois de l'opération. Ceci pourrait s'interpréter comme le résultat d'un arrêt temporaire de l'usage des modèles internes devenus inadaptés, à cause de l'exérèse. L'activation en hausse à 3 mois dans la même zone cérébelleuse mais aussi dans le CCA et le GFM droits pourrait, à l'inverse, être le signe d'une construction de nouveaux modèles internes avec une hausse de l'attention pour le contrôle de l'action (*scaffolding-storage* évoqué par Petersen et al., 1998 en phase de récupération). Golfinopoulos et al., 2011, ont montré des activations bilatérales dans le cervelet postérieur (lobule VIII) et le CCA droit quand le feedback somatosensoriel est perturbé de façon non prédictible après blocage artificiel de la mandibule par un ballon gonflé d'air. Une forte activation du LPI droit est aussi retrouvée à 3 mois, ce qui pourrait être la marque d'une redéfinition des buts sensori-moteurs lors de l'adaptation des modèles internes. Toujours pour les voyelles, la forte activation de l'AMS bilatérale, de M1 à gauche et la baisse de l'activité du GTS gauche sont autant d'éléments cohérents avec notre hypothèse d'une réutilisation, 9 mois après la chirurgie, de modèles internes adaptés à la nouvelle morphologie du conduit vocal. En particulier, la moindre activation du cortex auditif correspond au phénomène de décharge corollaire, dans lequel les modèles internes pourraient être impliqués (Blakemore et al., 2000).

Remerciements

A la Fondation des Gueules Cassées et le Pôle Grenoble Cognition (projet REPARLE). A R. Laboissière et C. Pichat (LPNC, Grenoble) pour leur aide dans l'analyse statistique.

Références

ACHER, A., SATO, M., LAMALLE, L., KRAINIK, A., & PERRIER, P (2012). Méthodologie en IRM fonctionnelle pour l'étude des activations corticales associées au réapprentissage de la parole. *In XXIXe Journées d'Études sur la Parole (JEP 2012)*, 1:723-730, Grenoble, France.

BATES, D., MAECHLER, M., BOLKER, B. & WALKER, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. *R package version 1.0-5.*

BLAKEMORE, S. J., WOLPERT, D. & FRITH, C. (2000). Why can't you tickle yourself? *Neuroreport*, 11, 11-16.

BUCHAILLARD, S., BRIX, M., PERRIER, P. & PAYAN, Y. (2007). Simulations of the consequences of tongue surgery on tongue mobility: Implications for speech production in post-surgery conditions. *International Journal of Medical Robotics and Computer Assisted Surgery*, 3(3):252-261.

EICKHOFF, S.B., STEPHAN, K.E., MOHLBERG, H., GREFKES, C., FINK, G.R., AMUNTS, K. & ZILLES, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, 25:1325-1335.

FRISTON, K.J., HOLMES, A.P., POLINE, J.B., GRASBY, P.J., WILLIAMS, S.C., FRACKOWIAK, R.S. & TURNER, R. (1995). Analysis of fMRI time-series revisited. *NeuroImage*, 2:45-53.

GOLFINOPOULOS, E., TOURVILLE, J. A., BOHLAND, J. W., GHOSH, S. S., NIETO-CASTANON, A. & GUENTHER, F. H. (2011). fMRI investigation of unexpected somatosensory feedback perturbation during speech. *Neuroimage*, 55, 1324-1338.

GRABSKI, K., SCHWARTZ, J.-L, LAMALLE, L., VILAIN, C., VALLEE, N., BACIU, M. LE BAS, J.-F & SATO, M. (2013). Shared and distinct neural correlates of vowel perception and production. *Journal of Neurolinguistics*, 26(3):384-408.

GRACCO, V.L., TREMBLAY, P. & PIKE, G.B. (2005). Imaging speech production using fMRI. *Neuroimage*, 26:294-301.

HAUPAGE, S., PECK, K.K., BRANSKI, R.C., HSU, M., HOLODNY, A. & KRAUS, D. (2010). Functional MRI of tongue motor tasks in patients with tongue cancer: observation before and after partial glossectomy. *Neuroradiology*, 52:1185–1191.

HOTHORN, T., BRETZ, F. & WESTFALL, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, 50:3, 346--363.

HICKOK, G. & POEPPEL, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8:393-402.

KAWATO, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9:718-772.

LANCASTER, J.L., WOLDORFF, M.G., PARSONS, L.M., LIOTTI, M., FREITAS, C.S., RAINEY, L., KOCHUNOV, P.V., NICKERSON, D., MIKITEN, S.A. & FOX, P.T. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping*, 10:120-131.

MANNI, E., & PETROSINI, L. (2004). A century of cerebellar somatotopy: a debated representation. *Nature Reviews Neuroscience*, 5:241-249.

MOSIER, K., LIU, W.-C., BEHIN, B., LEE, C. & BAREDES, S. (2005). Cortical adaptation following partial glossectomy with primary closure: implications for reconstruction of the oral tongue. *Ann Otol Rhinol Laryngol*, 114, 681-687.

OLDFIELD, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9:97-114.

PETERSEN, S. E., VAN MIER, H., FIEZ, J. A. & RAICHLE, M. E. (1998). The effects of practice on the functional anatomy of task performance. *Proc Natl Acad Sci USA*, 95, 853-860.

R CORE TEAM (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

TALAIRACH, J., TOURNOUX, P. (1988). *Co-planar stereotaxic atlas of the human brain*. Thieme, New York.

TREMBLAY, A., DALHOUSIE UNIVERSITY, RANSIJN, J. & UNIVERSITY OF COPENHAGEN (2013). LMERConvenienceFunctions: A suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions. *R package version 2.5*.

WOLPERT, D. M., MIALL, R. C. & KAWATO, M. (1998). Internal models in the cerebellum.*Trends in Cognitive Sciences*, 2 :338-347.

Structures syllabiques et caractéristiques du cycle mandibulaire : une étude articulatoire des asymétries

Nathalie Vallée[1] Thi Thuy Hien Tran[1] Solange Rossato[2] Paolo Mairano[1,3]
(1) GIPSA-lab DPC, UMR 5216, CNRS & Université Grenoble Alpes, BP25 38040 Grenoble cedex 9, France
(2) LIG GETALP, UMR 5217, CNRS & Université Grenoble Alpes, BP 53 - 38041 Grenoble cedex 9, France
(3) LFSAG, Università di Torino. Via Sant'Ottavio 20, 10124, Turin, Italy
nathalie.vallee@gipsa-lab.grenoble-inp.fr, thi-thuy-hien-tran@gipsa-lab.grenoble-inp.fr, solange.rossato@imag.fr, paolomairano@gmail.com

RESUME

Notre étude tente de définir des éléments de la nature physique de la syllabe, et plus particulièrement, d'établir un lien entre asymétries phonétique et phonologique capable d'expliquer le *MOP*, principe phonologique empirique qui affecte les segments consonantiques à la position initiale de syllabe plutôt que finale. Les deux cadres théoriques qui s'intéressent à ce lien, *Frame, then Content theory* et *Articulatory Phonology*, laissent quelques interrogations qui nous amènent à le revoir sous l'aspect articulatoire, plus précisément, au niveau des caractéristiques du geste mandibulaire. Des études antérieures sur l'anglais-américain ont montré l'existence d'asymétries au niveau des phases du cycle mandibulaire qui pourraient expliquer les grandes tendances des structures syllabiques. Nous exposons ici des résultats préliminaires sur le français et le vietnamien.

ABSTRACT

Syllabic structures and properties of jaw cycle: An articulatory study of Asymmetries

The present study tries to highlight a link between phonetic and phonological asymmetries which could account for the *MOP* which allocates consonantal segments to syllable-initial position rather than syllable-final. The theoretical frameworks concerned are the *Frame, then Content theory* and *Articulatory Phonology*, but they do not account for this link thoroughly. For this reason, we shall reconsider it in the light of new articulatory data focusing on the mandibular gesture. Previous studies on American-English have shown evidence of asymmetries within the jaw cycle that might explain general trends of syllabic structures. We present first results on French and Vietnamese.

MOTS-CLES: Syllabe, cycle mandibulaire, asymétries phonétique et phonologique, français, vietnamien
KEYWORDS: Syllable, jaw cycle, phonetic and phonological asymmetries, French, Vietnamese

1 Introduction

Établir le lien entre asymétries phonétique et phonologique au niveau de l'organisation structurelle de la syllabe devrait permettre d'expliquer un standard sur lequel se base la plupart des théories sur la syllabation des langues : le *Maximum Onset Principle* ou *MOP* (Kahn, 1976 ; Selkirk, 1982 ; Clements & Keyser, 1983 ; Goldsmith, 1990 ; Blevins, 1995). Principe universel de syllabation, le MOP postule l'attraction des segments consonantiques dans la position antéposée au noyau vocalique (attaque) au dépend de la position postposée (coda). À ce jour, le MOP reste une observation phonétique et phonologique sur laquelle se base bon nombre de travaux même expérimentaux en linguistique, psycholinguistique, traitement automatique des langues. Deux cadres théoriques s'intéressent à la nature du MOP : *Frame, then Content theory* (MacNeilage, 1998 ; 2008) et *Articulatory Phonology* (Browman & Golstein, 1988 ; 1995 ; 2000).

Dans le premier, l'organisation universelle de la parole en successions de consonne et voyelle est postulée comme étant le produit de l'oscillation de la mandibule, reliant directement la structure

CV universelle au cycle du geste mandibulaire (qui constitue le cadre) : la consonne est produite lorsque la mandibule est en position haute alors que la voyelle est réalisée lorsque la mandibule est en position basse (le contenu). Cependant, cette théorie n'apporte pas d'explication à la forte prédominance de la structure syllabique CV par rapport à la structure inverse VC laquelle représente moins de 5 % des syllabes des langues du monde (Rousset, 2004) et qui s'inscrit tout aussi bien dans le cycle mandibulaire. La théorie prédit également des combinaisons sonores CV favorisées dans les langues qui sont semblables aux productions des enfants au stade du babillage, et ce quelle que soit la langue de leur environnement (Davis & MacNeilage, 1995 ; MacNeilage & Davis, 2001). Ces combinaisons favorisées appelées *Pure Frame* sont produites avec un cycle d'oscillation mandibulaire, sans mouvement antéro- postérieur de la langue entre la réalisation de la consonne et celle de la voyelle, et sont considérées comme des productions articulatoirement simples leur conférant un avantage qui expliquerait leur surreprésentation. Rousset (2004), confirmé par Vallée, Rossato & Rousset (2009), montre que les proportions de *Pure Frame* parmi les syllabes VC relevées dans 15 langues sont significativement plus importantes que pour les syllabes CV, indiquant une forte cohésion articulatoire plus fréquente entre voyelle et consonne lorsque cette dernière est en coda par rapport à l'attaque. Il faut aussi ajouter comme point faible à cette théorie que le type de syllabe CVC est au deuxième rang des fréquences des structures syllabiques dans les langues du monde (après CV), et au premier rang de certaines (ex. les langues du sud-est asiatique : nyakhur, thai, vietnamien, wa...).

Dans le modèle de la phonologie articulatoire, au contraire, la mandibule n'a pas de rôle direct dans la production de la syllabe, car considérée comme support des actions des articulateurs lèvre inférieure et langue. La prédominance de la structure CV y est expliquée par les caractéristiques naturelles des segments, consonantiques et vocaliques, qui s'associent en phase (*in phase-coupling*) : les gestes articulatoires pour C et V sont très précisément coordonnés, donc compatibles, sans interférence négative mutuelle sur leur production respective car réalisés grâce à des mécanismes articulatoires indépendants qui permettent, en même temps, un déclenchement synchronisé des gestes pour C et V. Le chevauchement des gestes, estimé par la mesure du *C-Center Effect* (Browman & Goldstein, 1988) confère à la combinaison CV "*the most stable coupling mode*" et les différentes positions intra-syllabe possibles pour C correspondent alors à différents modes de configuration des gestes (Browman & Golstein, 1995 ; Byrd, 1995). La syllabe est ici, à l'inverse de la démonstration sur laquelle s'appuie la théorie *Frame, then Content*, le produit de plusieurs gestes individuels correspondant aux segments coarticulés ("*a gestural constellation*"). Les combinaisons CV favorisées sont celles pour lesquelles les gestes sont les plus naturellement en phase donc stables (Goldstein, Byrd & Saltzman, 2006 ; Whalen, Giulivi, Goldstein, Nam & Levitt, 2011). La structure renversée VC demande une coordination motrice plus complexe dans la mesure où la production de la consonne, si elle est lancée trop tôt, masque la réalisation acoustique de la voyelle. Pour cette raison, les structures VC montrent beaucoup moins de stabilité dans la coordination des gestes (pas de *C-Center Effect*) expliquant la vulnérabilité de la coda sujette à la lénition. Cette explication peut-elle être prolongée à l'observation de la proportion des *Pure-Frame* plus importante dans les structures VC par rapport à CV ?

Une autre piste d'explication de cette co-articulation noyau-coda, qui serait aussi à même de rendre compte du MOP, pourrait venir des études de Kelso, Vatikiotis-Bateson, Saltsman & Kay (1985), Gracco (1994), Redford (1999) ou encore Redford & Donkelaar (2008) qui, en cherchant à évaluer l'impact des propriétés biomécaniques de la mandibule sur l'articulation des segments, mettent en évidence l'existence d'une asymétrie entre les deux phases du cycle mandibulaire : la

phase de remontée (fermeture) est observée plus rapide, plus courte, moins ample, plus raide (*stiffness*) que la phase d'abaissement (ouverture). Si ces asymétries inhérentes aux propriétés du cycle mandibulaire étaient vérifiées indépendamment des cibles linguistiques, elles pourraient expliquer plusieurs tendances observées dans les langues du monde dont le MOP (cf. entre autres Rousset, 2004). Notre étude se situe dans cette problématique, à savoir si le cycle mandibulaire joue un rôle ou non dans la compréhension de la nature physique de la syllabe. Les travaux précédents ayant porté essentiellement sur des locuteurs de l'anglais-américain, notre étude s'inscrit dans un projet plus vaste multilingue (français, polonais, portugais brésilien, ruwund, tachelhit, vietnamien). Nous présentons ici les premiers résultats obtenus pour le français et le vietnamien.

2 Procédure et méthode

2.1 Matériel

Les mouvements des articulateurs mâchoire, langue, lèvres ont été mesurés avec le système d'articulographie électromagnétique (EMA) AG200 de la société Carstens grâce auquel il a été procédé à l'acquisition en 2D, à une fréquence de 200 Hz, de 5 bobines collées sur les articulateurs (mâchoire, lèvre inférieure, lèvre supérieure, apex et dos de la langue) et 2 bobines collées sur le plan de référence médio-sagittal du sujet. Le signal acoustique de parole a été enregistré avec un enregistreur numérique stéréo PMD670 de Marantz, micro directionnel C1000S d'AKG et numérisé à 44.1 KHz.

2.2 Corpus et participants

Les résultats préliminaires présentés ci-après ont été obtenus pour le français à partir de séquences extraites du corpus de logatomes {CV.CVL.CVC, CV.LVC, CLV.CVLC, CLV.CVCL, CLV.CLVC} avec C=/p b t d s z/, L=/l/, V=/i a/ et le point représentant une frontière syllabique. Un corpus de 43 phrases contenant chacune des syllabes dans différentes distributions (ex. Son analyse digitale est utile et suffisamment probable) complète ces données, ainsi que la répétition d'un enchainement des 2 voyelles. Cinq locutrices natives du français (CC, EB, LV, ML, SB) âgées de 20 à 43 ans ont participé à l'expérience. Les logatomes et phrases ont été enregistrées dans 2 sessions à part. Cinq répétitions de chaque logatomes, ainsi que les phrases, ont été présentées dans un ordre aléatoire. Les résultats préliminaires exposés ici portent sur cinq répétitions de 3 logatomes prononcés à un débit normal d'élocution : /aiaiaiaiai/, /tataltat/, /sasalsas/. Le même protocole a été utilisé pour le vietnamien. Les logatomes sont monosyllabiques CV et CVC avec $C_{initiale}$=/b t d s z/, C_{finale}=/p t/, V=/i a / et ton modal B1-D1. Cinq locutrices natives du vietnamien âgées de 28 à 35 ans, variété du nord (HOA, LINH, THANH, HOAI, HUONG) ont participé à l'expérience. Nous présentons ici les premiers résultats obtenus pour cette langue avec la séquence /aiaiaiaiai/.

2.3 Mesures

Pour l'heure, seul le mouvement de la mâchoire a fait l'objet d'analyses. À partir de la segmentation semi-automatique avec EasyAlign et Praat, les séquences ont été extraites et étiquetées avec Trap v.6, outil développé sous environnement Matlab au GIPSA-lab par C. Savariaux. Les mesures effectuées avec Trap sont les suivantes : (1) durée des phases d'abaissement (ouverture) et de remontée (fermeture) de la mandibule mesurées à partir des maxima d'ouverture et de fermeture qui correspondent aux points de passage par zéro de la courbe de vitesse ; (2) pic de vitesse et vitesse moyenne de chacune des phases ; (3) amplitude de chaque phase, qui correspond au déplacement vertical de la mandibule, estimée entre les maxima d'ouverture et de fermeture.

2.4 Analyse statistique

Notre objectif est d'étudier les variations des variables réponses (durée, vitesse et amplitude de chacune des phases du cycle mandibulaire) et l'influence de deux facteurs sur celles-ci : d'une part le type syllabique CV, CVL, CVC ainsi que la séquence VV ; d'autre part la langue (français, vietnamien) pour la séquence VV. Notre protocole ayant permis de recueillir plusieurs valeurs de variable réponse pour un même sujet, il ne nous garantit pas l'indépendance des observations qui pourraient alors être liées à un ou quelques sujets, comme par exemple la durée des phases. Notre choix s'est porté sur le modèle linéaire à effets mixtes et pour permettre de respecter l'hypothèse selon laquelle les résidus suivent une loi normale (condition d'application des modèles mixtes), nous avons choisi de transformer la variable réponse en son logarithme :

$$log\, Y_{ijkl} = \mu + \alpha_k + \beta_l + (\alpha\beta)_{kl} + \tau_i + \varepsilon_{ijkl} \qquad \text{avec :}$$

Y_{ijkl} : j ième valeur du sujet i prise par la variable réponse pour la modalité k du facteur A et l du facteur B

α_k : effet de la modalité k du facteur A sur la variable réponse,

β_l : effet de la modalité l du facteur B sur la variable réponse,

$(\alpha\beta)_{kl}$: effet de l'interaction de la modalité k du facteur A et de la modalité l du facteur B sur la variable réponse,

τ_i : effet aléatoire du sujet i sur la variable réponse. τ_i suit une loi normale $\tau_i \sim \mathcal{N}(0, \sigma_\tau^2)$ avec σ_τ^2 la variable réponse entre sujets,

ε_{ijkl} : représente le terme d'erreur et suit une loi normale, $\varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma_e^2)$.

Pour analyser la différence entre les deux modalités (phases d'ouverture et fermeture), à l'intérieur de chaque modalité type syllabique et langue, sera appliquée la méthode de Hothorn, Bertz et Westfall (2008) qui permet de réaliser des comparaisons multiples de moyennes avec le modèle mixte en garantissant également que le risque de première espèce lié à la prise simultanée de toutes les décisions ne dépasse pas le seuil fixé à l'avance à 5 % en ajustant les p-values. La méthode sera appliquée aux données avec la fonction *glht* du package *multcomp* du logiciel R ainsi que la fonction *lsmeans* du package *lsmeans*.

3 Résultats

3.1 Durée

Les comparaisons multiples des durées moyennes entre phase d'ouverture et phase de fermeture pour chacun des trois cycles mandibulaires relevés dans les logatomes CV.CVL.CVC (C= {/t/, /s/}, V=/a/) et les cinq cycles de la séquence /aiaiaiaiai/ montrent des différences significatives entre les deux phases, avec une fermeture plus longue que l'ouverture (Table 1, Figures 1 et 2), sauf pendant la réalisation de la syllabe initiale /ta/.(t) ($F\text{-}O_{ta.(t)}$=-0.06164, SD=0.066, z=-0.932, p=0.95).

	Estimate	Std. Error	z	value Pr(>|z|)
F - O | iai	0.21906	0.04275	5.125	2.09e-06 ***
F - O | sa.(s)	0.22734	0.06759	3.364	0.00537 **
F - O | sal	0.47024	0.06759	6.957	2.43e-11 ***
F - O | sas	0.46185	0.06759	6.833	5.81e-11 ***
F - O | ta.(t)	-0.06164	0.06610	-0.932	0.95156
F - O | tal	0.31084	0.06610	4.702	1.80e-05 ***
F - O | tat	0.50349	0.06610	7.617	1.83e-13 ***

TABLE 1 – Estimations ponctuelles des différences de moyennes de durée (log) entre phase de fermeture et phase d'ouverture pour tous les cycles mandibulaires des logatomes /aiaiaiaiai/, /tataltat/ et /sasalsas/ avec écarts-type des différences, valeur de la statistique et p-value (hypothèse du test : F-O=0).

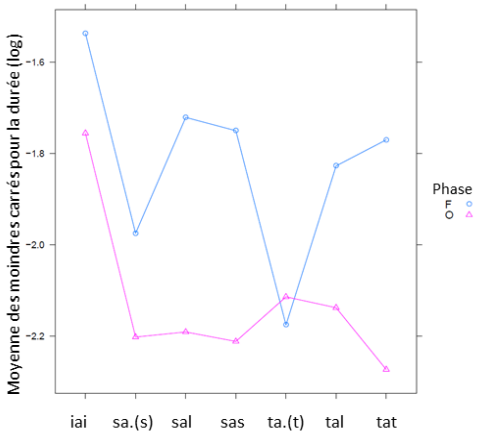


FIGURE 1 – Estimations des durées moyennes (log) des phases d'ouverture et de fermeture pour chaque type de cycle mandibulaire. Les fermetures sont en général plus longues que les ouvertures.

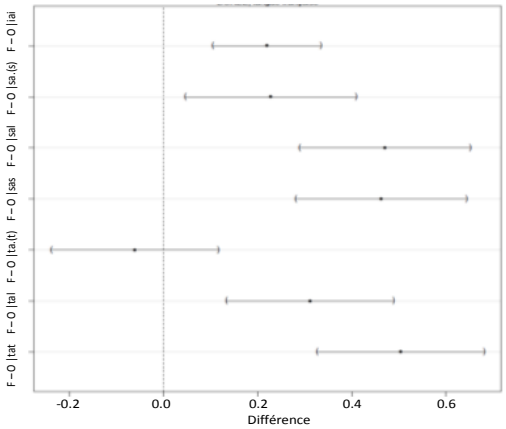
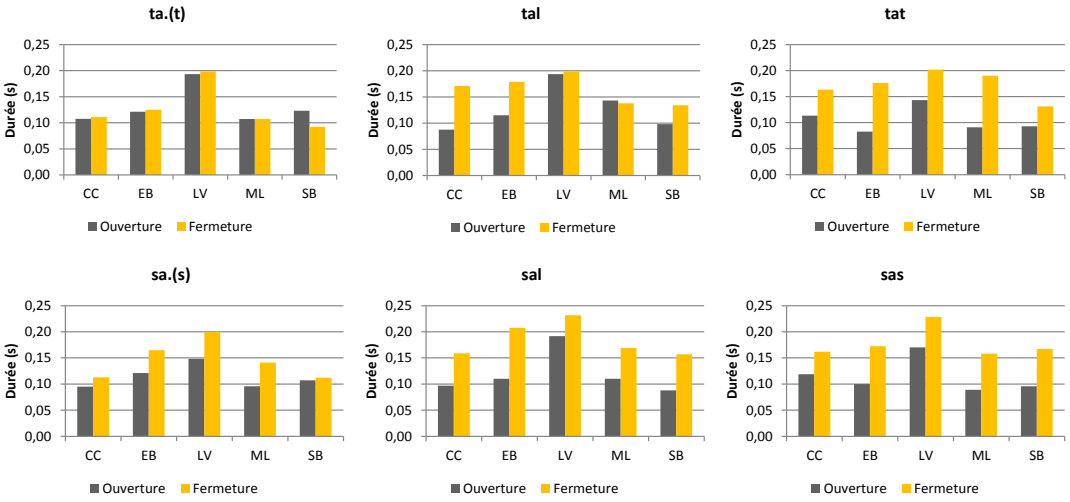


FIGURE 2 – Estimations ponctuelles des différences de moyennes entre durée de fermeture et durée d'ouverture (exprimées en log) et intervalles de confiance à 95 % pour chaque type de cycle.

La figure 2 comporte les valeurs des comparaisons multiples entre les différents types de cycles mandibulaires (lignes horizontales), le point représentant l'estimation ponctuelle de la différence et les parenthèses, les bornes de l'intervalle de confiance à 95 %. Seul l'intervalle de confiance pour /ta/.(t) contient la valeur 0 indiquant une différence non significative. La figure 1 montre que les remontées de la mandibule sont en moyenne plus longues en présence d'une consonne en coda (/tal/ et /tat/ *vs.* /ta/.(s), et /sal/ et /sas/ *vs.* /sa/.(s)).



FIGURES 3 – Durées moyennes des phases d'ouverture et de fermeture des cycles mandibulaires en fonction des types syllabiques et des locutrices du français pour les deux consonnes /t/ et /s/.

Comparativement aux deux séquences avec consonnes, les durées moyennes des deux phases du cycle sont plus longues en l'absence de geste consonantique (séquence /aiaiaiaiai/). C'est aussi la séquence dont l'intervalle de confiance est le plus réduit (Figure 2), indiquant une variabilité moindre dans les réalisations. Un effet de la position finale du logatome, position dans laquelle on pourrait s'attendre à un allongement de la durée en français (patron accentuel iambique), est seulement observé pour la phase de fermeture avec la plosive en coda (Figure 1). Hormis le cas de /ta/.(t), l'observation des durées moyennes des phases d'ouverture *vs.* phases de fermeture montre que le patron avec fermeture plus longue que l'ouverture est assez stable chez les sujets français, quelles que soient la structure syllabique, la consonne plosive ou fricative, la position dans le logatome (Figures 3) : on observe une légère tendance au patron inverse (ouverture plus longue que la fermeture) chez la locutrice ML pour le cycle qui produit la syllabe /tal/ (143 *vs.* 138 ms). SB et LV montrent des différences moyennes entre fermeture et ouverture de 5 ms respectivement pour /sa/.(s) et /tal/, alors que chez ces locutrices, pour les autres syllabes, les

différences moyennes sont d'au moins 20 ms chez SB et 40 ms chez LV. Aucune des cinq locutrices françaises ne montre une corrélation forte entre durées d'ouverture et durées de fermeture (CC : r=.08 ; EB : r=.25 ; LV : r=0.31 ; ML : r=.02 ; SB : r=.33), dont une consistance encore moindre avec la plosive. Il en est de même pour la séquence /aiaiaiaiai/ (CC : r=.19 ; EB : r=.34 ; LV : r=0.22 ; ML : r=.22 ; SB : r=.05).

3.2 Vitesse

Comme pour la durée, les mesures de la vitesse moyenne des phases d'ouverture et phases de fermeture sont significativement différentes (Table 2, Figure 5) avec une ouverture plus rapide que la fermeture (Figure 4), sauf pour le cycle mandibulaire correspondant à la réalisation de la syllabe initiale /ta/.(t) ($F\text{-}O_{ta.(t)}$=-0.0215, SD=0.293, z=0.073, p=1).

	Estimate	Std. Error	z	value Pr(>|z|)
F - O | iai	-0.14975	0.05811	-2.577	0.067726
F - O | sa.(s)	-0.26517	0.08620	-3.076	0.014586 *
F - O | sal	-0.59029	0.08620	-6.848	< 1e-10 ***
F - O | sas	-0.44852	0.08620	-5.203	1.37e-06 ***
F - O | ta.(t)	0.02152	0.29353	0.073	1.000000
F - O | tal	-0.38905	0.08431	-4.615	2.76e-05 ***
F - O | tat	-0.33608	0.08431	-3.986	0.000469 ***

TABLE 2 – Estimations ponctuelles des différences de moyennes des vitesses (log) entre phases de fermeture et d'ouverture pour tous les cycles mandibulaires des logatomes /aiaiaiaiai/, /tataltat/ et /sasalsas/, avec écarts-type des différences, valeur de la statistique et p-value (l'hypothèse est une différence F-O nulle).

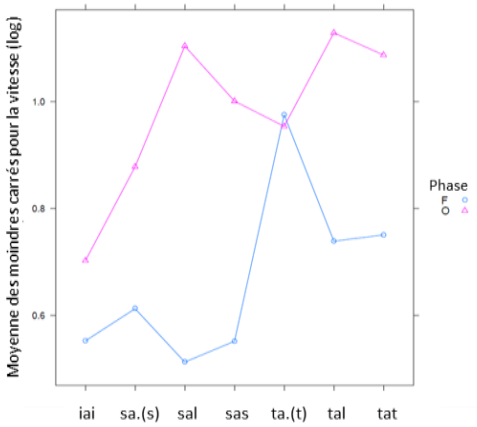


FIGURE 4 – Estimations des vitesses moyennes (log) des phases d'ouverture et de fermeture pour chaque type de cycle mandibulaire. Sauf pour /ta/.(t), les ouvertures sont plus rapides.

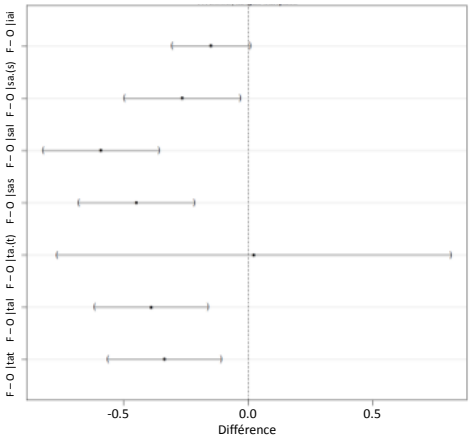
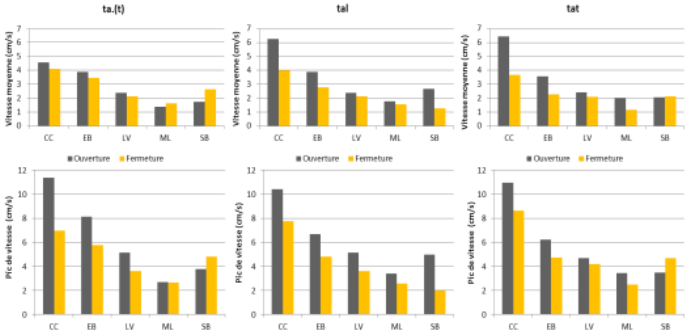


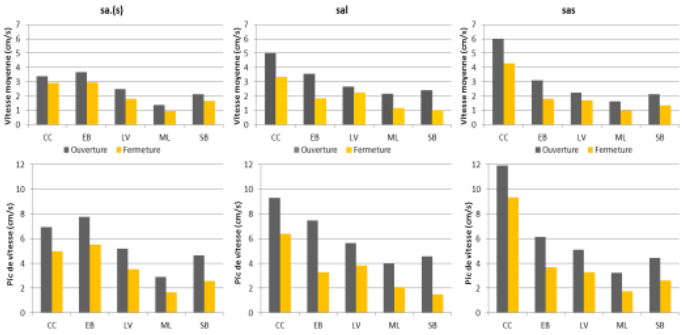
FIGURE 5 – Estimations ponctuelles des différences de moyennes entre vitesse moyenne de fermeture et vitesse moyenne d'ouverture (log) et intervalles de confiance à 95 % pour chaque type de cycle mandibulaire.

Les figures 6 montrent que deux locutrices ML et SB présentent, pour cette syllabe, un patron de vitesse mandibulaire inversé par rapport aux autres locutrices (fermeture plus rapide), patron qui n'est pas observé par ailleurs sauf pour SB dans le cycle correspondant à /tat/ ($F\text{-}O_{tat}$=0.073 cm/s). Ces différences entre locutrices ne sont plus observées pour le logatome avec la fricative coronale /s/ : l'ouverture est toujours plus rapide que la fermeture (Figure 7). Les figures 6 et 7 montrent également que les pics de vitesse, mesurés et moyennés pour chaque phase, chaque syllabe et chaque locutrice, suivent les patterns décrits ci-dessus pour les vitesses moyennes. Seules trois locutrices montrent une corrélation entre vitesses moyennes d'ouverture et de fermeture (CC : r=.79 ; EB : r=.52 ; LV : r=0.65). La même observation est faite pour les pics de vitesse (CC : r=.64 ; EB : r=.51 ; LV : r=0.72). Cette tendance n'est pas observée chez les locutrices ML (r=.23) et SB (r=.30) lesquelles présentent un patron de vitesse inversé pour la syllabe initiale /ta/.(t). La même observation est valable pour les pics de vitesse : ML r=.18 ; SB

r=.25. L'exclusion de cette syllabe dans le calcul de la corrélation augmente sensiblement la valeur du coefficient pour ML (r=.45) mais pas pour SB (r=.27), même remarque pour les pics de vitesse avec ML : r=.50 et SB : r=.23. Cependant, pour la séquence/aiaiaiaiai/, une corrélation des vitesses moyennes (confirmée par les pics de vitesse) est observée chez les cinq locutrices françaises, corrélation particulièrement forte chez ML (CC : r=.79 ; EB : r=.74 ; LV : r=0.60 ; ML : r=.96 ; SB : r=.68). Les valeurs des coefficients de corrélation pour les pics de vitesse sont CC : r=.81 ; EB : r=.53 ; LV : r=0.58 ; ML : r=.95 ; SB : r=.65.



FIGURES 6 – Vitesses moyennes (en haut) et pics de vitesse moyennés (en bas) des phases d'ouverture et de fermeture des cycles mandibulaires en fonction des types syllabiques et des locutrices du français pour la plosive /t/.



FIGURES 7 – Vitesses moyennes (en haut) et pics de vitesse moyennés (en bas) des phases d'ouverture et de fermeture des cycles mandibulaires en fonction des types syllabiques et des locutrices du français pour la fricative /s/.

3.3 Amplitude

L'amplitude (déplacement vertical de la mandibule), ne montre pas de patron régulier entre phases d'ouverture et phases de fermeture du cycle mandibulaire (Figure 8).

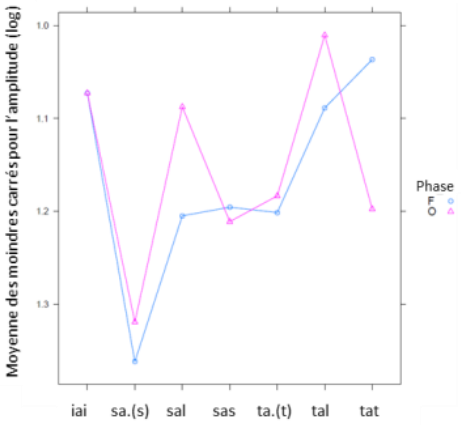


FIGURE 8 – Estimations du déplacement vertical de la mandibule (log) dans les phases d'ouverture et de fermeture pour chaque type de cycle mandibulaire. Les syllabes finales de logatome montrent des ouvertures plus amples que les fermetures.

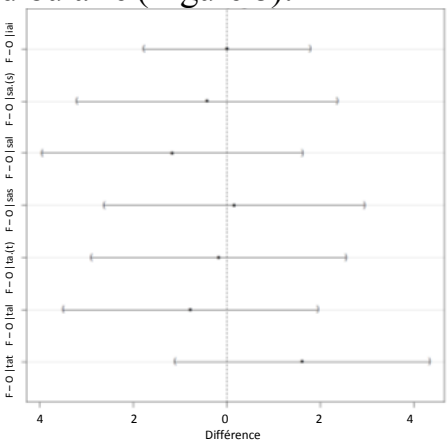
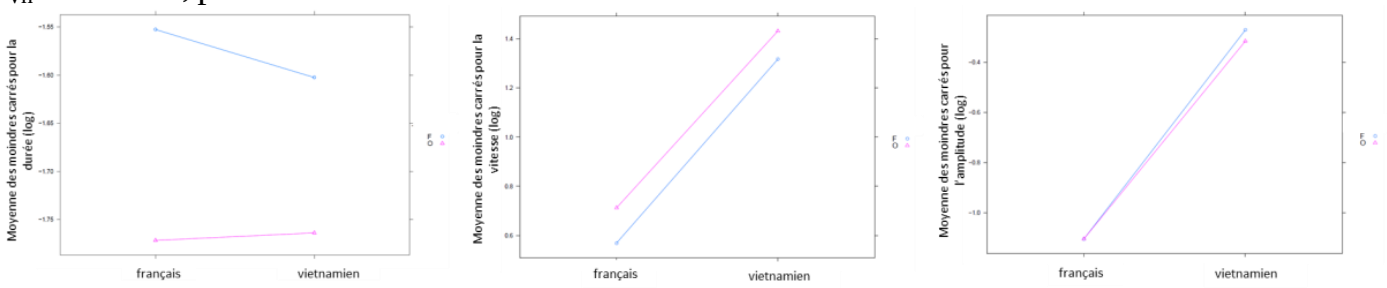


FIGURE 9 – Estimations ponctuelles des différences de moyennes entre amplitude de fermeture et d'ouverture (en log) et intervalles de confiance à 95 % pour chaque type de cycle mandibulaire. Aucune différence n'est significative.

On note une amplitude moyenne plus petite à l'ouverture qu'à la fermeture pour les cycles des syllabes /sa/.(s), /sal/, /ta/.(t) et /tal/ et le contraire, une amplitude plus grande à l'ouverture pour les syllabes /tat/ et dans une moindre mesure /sas/, toutes deux en finale de logatome. La figure 9 indique qu'aucune différence d'amplitude moyenne entre phases de fermeture et d'ouverture n'est significative. Aucune régularité n'est observée dans les différences d'amplitudes de phase en fonction du locuteur ou du type de syllabe. L'amplitude semble dépendante des segments et des structures syllabiques. On relève cependant chez toutes les locutrices une corrélation forte entre amplitude d'ouverture et amplitude de fermeture dans les logatomes avec consonnes (CC : r=.98 ; EB : r=.96 ; LV : r=0.88 ; ML : r=.88 ; SB : r=.73) ainsi que dans la séquence /aiaiaiaiai/ (CC : r=.97 ; EB : r=.70 ; LV : r=0.83 ; ML : r=.87 ; SB : r=.65).

3.4 Comparaison français-vietnamien

Seules les répétitions de la séquence /aiaiaiaiai/ ont pu être prises en compte dans cet article. Ce résultat préliminaire nous permet une première comparaison avec les données obtenues auprès des cinq locutrices du français pour cette même séquence. Les profils globaux des locutrices du vietnamien sont similaires à ceux des locutrices du français : les phases d'ouverture sont plus courtes, plus rapides et moins amples que les phases de fermeture (Figures 10). Comme pour le français, les différences de durées (D) et de vitesses moyennes (V) entre fermeture et ouverture sont significatives (D_{fr} : z=5.88, p=7.90e-09 ; D_{vn} : z=5.26, p =2.81e-07 ; V_{fr} : z=-2.5, p=0.024 ; V_{Vn} : z=-2.59, p=0.018.



FIGURES 10 – Estimation des moyennes de durées, vitesses moyennes et amplitudes (exprimées en log) des phases d'ouverture et de fermeture en français et vietnamien pour la séquence /aiaiaiaiai/.

4 Conclusion

Si ces résultats préliminaires vont bien dans le sens d'asymétries dans le cycle mandibulaire, avec une phase d'ouverture (abaissement de la mandibule) généralement plus courte, plus rapide et moins ample que la phase de fermeture, ceux-ci sont en contradiction avec les résultats des travaux sur l'anglais-américain de Kelso et al. (1985), Gracco (1994), Redford (1999), Redford & Donkelaar (2008) qui mettent en évidence un patron d'asymétrie inverse (c'est-à-dire une ouverture plus longue, moins rapide, plus ample) qu'ils supposent être la conséquence des propriétés biomécaniques de la mandibule (ouverture et fermeture ne sont pas soumises aux mêmes contraintes). Dans notre étude, seule la syllabe initiale /ta/.(t) montre un patron inverse chez deux locutrices ML et SB pour les facteurs durée et vitesse, mais sans significativité. Les différences d'amplitudes, bien que non significatives, montrent un patron inverse pour les syllabes finales (fermeture moins ample) qui peut s'expliquer par le relâchement articulatoire (Krakow, 1999). Peu de modifications liées au contexte (plosive *vs*. fricative, latérale, syllabe ouverte *vs*. fermée, position initiale *vs*. finale) sont donc observées pour les facteurs durée et vitesse. La séquence vocalique /aiaiaiaiai/ montre une plus grande stabilité du cycle mandibulaire et de ses asymétries au niveau durée, vitesse et amplitude, liée très certainement à la réitération (Kelso et al., 1985). Les résultats des corrélations sont en adéquation avec les études précédentes citées. Si les travaux antérieurs sur le cycle mandibulaire offraient des pistes d'explications du MOP, il n'en va pas de même de nos premiers résultats obtenus. La similitude entre français et vietnamien exclut pour l'instant toute explication liée à la structure syllabique de base : le français est une langue majoritairement CV et le vietnamien CVC (Rousset, 2004). L'influence des caractéristiques du cycle mandibulaire sur la phonologie (comme mis en avant par la théorie *Frame, then Content*) n'est pas éclairci et nécessite de démêler les aspects language-dependent, context-dependent et subject-dependent auxquels la Phonologie Articulatoire peut sans doute apporter des éléments de réponse. Ce constat ne peut que nous inciter à poursuivre nos investigations multilocuteurs et inter-langues.

Remerciements

Cette étude fait partie du Projet ANR-10-BLAN-1916 *APPSy*. Nous remercions vivement Silvain Gerber, Christophe Savariaux et Quentin Tura pour leur assistance, précieuse dans ce projet.

Références

Blevins J. (1995). The syllable in Phonological Theory. In J.A. Goldsmith (ed.): Handbook of Phonological Theory, 206-235. Blackwell Publishers: Oxford.

Browman, C. P. & L. M., Goldstein (1988). Some notes on syllable structure in articulatory phonology. Phonetica 45, 140-155.

Browman, C. P. & L.M., Goldstein (1995). Gestural syllable position effects in American English. In Bell-Berti, F. & L.J., Raphael: Producing Speech: Contemporary Issues. For Katherine Safford Harris, 19-33. AIP Press: New York.

Browman, C. P. & L. M., Goldstein (2000). Competing constrainsts on intergestural coordination and self-organization of phonological structures. Bulletin de la Communication Parlée 5, 25-34.

Byrd, D. (1995). C-centers revisited. Phonetica 52, 285-306.

Clements, N. & S. J., Keyser (1983). CV Phonology: A generative theory of the syllable. MIT Press.

Davis, B. L. & P. F., MacNeilage (1995). The articulatory bases of babbling. Journal of Speech and Hearing Research 38, 1199-1211.

Goldsmith, A. (1990). Auto segmental and Metrical Phonology. Basil Blackwell LTD: Oxford.

Goldstein, L.M., Byrd, D. & E. L. Saltzman (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In Arbib M., (ed): From action to language: The mirror neuron system, 215–249. Cambridge University Press: Cambridge.

Gracco, V. L. (1994). Some organizational characteristics of speech movement control. Journal of Speech and Hearing Research 37, 4-27.

Hothorn, T., Bertz, F. & P., Westfall (2008). Simultaneous inference in general parametric models. Biometrical Journal 50(3), 346-363.

Kahn, D. (1976). Syllable-based Generalizations in English Phonology. Indiana University Linguistics Club: Bloomington.

Kelso, J. A. S., Vatiokiotis-Bateson, E., Saltzman, E. L. & B., Kay (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics and dynamic modeling. JASA 77, 266-280.

Krakow, R. A. (1999). Physiological organization of syllables : a review. Journal of Phonetics 27, 23-54.

MacNeilage, P. F. (1998). The Frame/Content Theory of Evolution of Speech Production. Behavioral and Brain Sciences 21, 499-511.

MacNeilage, P. F. (2008). The origin of speech. Oxford University Press: Oxford, England.

MacNeilage, P. F. & B. L., Davis (2001). Motor mechanisms in speech ontogeny: phylogenetic, neurobiological and linguistic implications. Current Opinion in Neurobiology 11, 696-700.

Redford, M. A. (1999). An Articulatory Basis for the Syllable. PhD thesis, The University of Texas: Austin.

Redford, M. A. & P., van Donkelaar (2008). Jaw cycles and linguistic syllables in adult English. In Davis, B.L. & K. Zajdo (eds.): The Syllable in Speech Production: Perspectives on the Frame/Content Theory, 355-76. Taylor & Francis: London.

Rousset, I. (2004). Structures syllabiques et lexicales des langues du monde : données, typologies, tendances universelles et contraintes substantielles. Thèse de doctorat, Université Stendhal Grenoble.

Selkirk, E. (1982). The syllable. In H., van der Hulst & N., Smith (eds): The structure of phonological representations, 337-383. Foris Publications: Dordrecht.

Vallée, N., Rossato, S. & I., Rousset (2009). Favoured syllabic patterns in the world's languages and sensorimotor constraints. In Pellegrino, F., Marsico, E., Chitoran I., & C., Coupé (Eds): Approaches to Phonological Complexity, 111-139. Mouton de Gruyter: Berlin.

Whalen, D. H., Giulivi, S., Goldstein, L. M., Nam, H. & A.G., Levitt (2011). Response to MacNeilage and Davis and to Oller. Language Learning and Development 7(3), 243–249.

Mesures de protrusion par capture optique de mouvements : quelle métrique est la plus représentative de l'opposition d'arrondissement en français ?

Georgeton Laurianne[1] Audibert Nicolas [1]
(1) Laboratoire de Phonétique et de Phonologie (UMR 7018)
19 rue des Bernardins - 75005 Paris
{laurianne.georgeton, nicolas.audibert}@univ-paris3.fr

RESUME

Le but de cette étude est d'évaluer laquelle des trois mesures de protrusion relevées dans la littérature (l'avancement de la lèvre inférieure, celui de la lèvre supérieure ou celui des commissures) rend le mieux compte de l'implémentation du trait d'arrondissement sur les voyelles orales antérieures fermées, mi-fermées et mi-ouvertes du français. 491 voyelles orales antérieures arrondies et non-arrondies ont été produites en contexte par 3 locutrices du français standard équipées du dispositif sans fil de capture optique de mouvements Qualisys pour l'obtention de mesures d'articulation labiale. La comparaison des voyelles antérieures arrondies vs. non-arrondies sur ces 3 mesures et leur mise en correspondance avec les variations des formants F3 et F2 montrent que l'avancement des commissures est la mesure la plus représentative de l'opposition d'arrondissement. Des différences interindividuelles de stratégies d'articulation labiale pour les paires étudiées sont également observées.

ABSTRACT

Protrusion measurements using optical motion capture: which metric is the most representative of the rounding feature in French?

The aim of this study is to evaluate which of the three protrusion measurements used in the literature (forwarding of the lower lip, the upper lip or the lips corners) better accounts for the implementation of the rounding feature in French on closed, mid-closed, mid-open front vowels. 491 front rounded and unrounded vowels were produced in context by 3 female speakers of standard French, equipped with the wireless optical motion capture system Qualisys to get labial articulation measurements. The comparison of rounded vs. unrounded front vowels using the 3 possible measurements and their link with variations of formants F3 and F2 show that the forwarding of lip corners best accounts for the rounding opposition. Inter-individual differences in labial articulation strategies for pairs studied are also observed

MOTS-CLES : labialité, protrusion, capture de mouvements, voyelles, arrondissement
KEYWORDS: labiality, protrusion, motion capture, vowels, rounding

1 Introduction

Les études qui se sont intéressées à la distinction des voyelles selon les mesures labiales ont mis en évidence trois paramètres essentiels, correspondant aux trois degrés de liberté physiologiques des lèvres proposés par Ladefoged (1979) : l'aperture, l'étirement et la protrusion. Parmi ces paramètres, plusieurs permettent de rendre compte de la distinction entre les paires /i-y/, /e-ø/ : l'étirement, l'aire aux lèvres (combinant les paramètres d'étirement et d'aperture), le rapport entre la protrusion et l'écartement, et le rapport entre l'aire et la protrusion (Fant, 1960, Abry & Boe, 1986).

Cette étude s'intéresse plus particulièrement à la protrusion et à sa capacité à rendre compte plus ou moins largement du contraste d'arrondissement des voyelles antérieures, en comparant différentes mesures possibles de ce paramètre d'articulation labiale. La protrusion peut être définie comme « *une poussée des lèvres vers l'avant, se traduisant par le rapprochement à la fois du milieu des lèvres (plan vertical) et des commissures (plan horizontal)* » (Abry et al. 1980, p 44). Dans la littérature, trois mesures ont le plus souvent été considérées: (i) le déplacement en avant de la lèvre supérieure (Upper Lip : UL), (ii) le déplacement en avant de la lèvre inférieure (Lower Lip : LL) et (iii) le déplacement en avant de la commissure des lèvres (mesure similaire à la distance C utilisée dans Abry & Boe, 1980).

Les données analysées dans cette étude s'inscrivent dans le cadre de recherches doctorales sur l'influence de la position initiale de constituants prosodiques sur la réalisation labiale et acoustique des voyelles du français (Georgeton & Fougeron, 2014). Pour observer l'effet de la position prosodique sur la protrusion des voyelles du français, nous avons cherché à établir quelle mesure permettait de refléter le plus fidèlement les mouvements de protrusion. Constatant une certaine hétérogénéité dans la littérature, nous avons alors cherché à déterminer laquelle des trois mesures possibles de protrusion labiale rend le mieux compte de l'opposition d'arrondissement dans nos données. A notre connaissance, seule l'étude d'Abry et Boë (1980) a directement comparé la capacité de discrimination des voyelles antérieures arrondies vs. non-arrondies de ces trois mesures, sans toutefois inclure la paire /ɛ-œ/.

Après les travaux pionniers sur l'anglais américain, de nombreux travaux ont porté sur des langues dans lesquelles certaines voyelles ne s'opposent que par l'arrondissement, telles que le français et le suédois. Nous esquissons ci-dessous un tableau des principaux usages de ces mesures dans la littérature.

(LL) Fromkin (1964) conclut que la protrusion est principalement réalisée par l'avancement de la lèvre inférieure, à partir de mesures des distances entre dents et lèvres extraites de données radiographiques et de moulages labiaux. En effet, la variabilité de ces mesures est plus importante pour la lèvre inférieure que pour la lèvre supérieure. L'auteur souligne également que, bien que l'avancement de la lèvre inférieure résulte fréquemment du rapprochement des lèvres ou du mouvement mandibulaire, il est également possible de projeter les lèvres vers l'avant indépendamment de l'aperture.

(C) L'étude d'Abry et Boë (1980) montre que seule la mesure liée au mouvement de la commissure permet de séparer les voyelles non-arrondies (/i-e/) des voyelles arrondies (/y-ø/). Cependant, cette distinction n'est observée pour un locuteur sur cinq testés. Les auteurs soulignent également que pour opérer une séparation complète entre l'espace de réalisation des voyelles arrondies vs. non-arrondies, il est nécessaire de prendre en compte l'écartement labial en complément de la protrusion. À partir de résultats d'une analyse en composantes principales, Robert et al. (2007) concluent que le mouvement de protrusion est dans une aussi large mesure associé aux mouvements des commissures qu'à ceux des lèvres inférieure et supérieure, et choisissent d'intégrer la position des commissures ainsi que celle de LL et UL dans la mesure de protrusion qu'ils retiennent pour étudier la coarticulation labiale.

(UL) L'avancement de la lèvre supérieure a été retenu comme mesure de protrusion dans diverses études cherchant à caractériser des mouvements articulatoires fins comme ceux induits par la prosodie articulatoire ou l'expression émotionnelle. Les résultats ne sont pas uniformes. En position finale de frontière prosodique, la voyelle /u/ montre une protrusion plus élevée (Tabain et Perrier, 2007). Pour les études portant sur l'articulation des syllabes focalisées, les résultats de Dohen (2005) montrent un effet fort de la protrusion alors qu'aucun effet n'est observé dans les travaux de Ménard et al. (2013). Nordstrand et al. (2004), montrent dans leur étude des modification articulatoires induites par des expressions émotionnelles en suédois que la mesure d'avancement de la lèvre supérieure permet de regrouper les voyelles arrondies d'une part et les voyelles non-arrondies d'autre part.

Nous comparons tout d'abord le potentiel de discrimination entre voyelles antérieures des différentes métriques proposées. Dans le cas du français, les études portant sur le contraste entre les voyelles [±arrondi] se sont limités à l'étude des paires les plus fermées /i-y/ et /e-ø/. Dans cette étude, nous prenons également en compte la paire de voyelles antérieures mi-ouvertes /ɛ-œ/. Nous évaluons ensuite dans quelle proportion ces mesures de protrusion rendent compte des corrélats acoustiques des différences articulatoires entre voyelles antérieures non-arrondies et arrondies.

2 Méthodologie

2.1 Corpus

Afin de pouvoir étudier les variations de protrusion par paire de voyelles, les séries de voyelles suivantes ont été considérées /i, e, ɛ/ et /y, ø, œ/. Les voyelles cibles Vc apparaissent dans des séquences $VC_1\#V_CC_2$ où V_C= /i, e, ɛ/ /y, ø, œ/, V= /i/, C_1= /p/ et C_2= /p/ excepté pour les voyelles mi-ouvertes /ɛ/ et /œ/ pour lesquelles la consonne suivante C_2 est respectivement /v/ et /ʃ/ afin de respecter les contraintes phonotactiques du français (Georgeton & Fougeron, 2014).

Quatre locutrices de 25 à 40 ans sans accent régional identifiable ont lu un total de 16 exemplaires de chaque énoncé dans un ordre aléatoire. Suite à un dysfonctionnement du système d'acquisition, les données de l'une des locutrices, en quantité trop faible, ont été éliminées de l'analyse. Toutes locutrices confondues,

nous obtenons un ensemble de 491 énoncés répartis comme indiqué dans la Table 1.

	/i/	/e/	/ɛ/	/y/	/ø/	/œ/	Total
S1	32	29	29	30	36	32	188
S2	28	28	26	30	22	28	162
S3	23	25	21	23	23	26	141
Total	83	82	76	83	81	86	491

Table 1 : Nombre de répétitions analysées par voyelle, en fonction de la locutrice.

2.2 Acquisition des données

Le système Qualisys a été utilisé pour suivre les trajectoires d'un total de 13 marqueurs, au moyen de 4 émetteurs/récepteurs infra-rouge disposés autour de la locutrice. Comme illustré sur la figure 1, les quatre marqueurs retenus dans cette étude sont les suivants : les commissures des lèvres (droite et gauche), sur la lèvre supérieure au niveau de l'arc de cupidon sur le bord vermillon et à l'opposé sur la lèvre inférieure. Les marqueurs de référence positionnés en des points fixes (casque et arête du nez) permettent de compenser les mouvements de la tête du locuteur. Afin d'extraire le déplacement des marqueurs situées sur les lèvres, l'ensemble des données a été normalisé par rapport à ces six points de référence.

Un enregistrement audio effectué avec un micro-casque Shure SM 10A a été couplé à l'acquisition de la position des marqueurs. Afin de permettre la mise en correspondance de ces données, le système Qualisys enregistre de façon synchrone au flux audio un signal de synchronisation (trigger) présentant un front descendant à chaque instant correspondant à l'acquisition d'une mesure Qualisys.

2.3 Segmentation et analyse acoustique des voyelles analysées

Les voyelles cibles V_C (/i, e, ɛ, y, ø, œ/) ont été segmentées avec Praat à partir du signal audio. Les limites acoustiques des voyelles ont été fixées d'après le début et la fin du second formant sur les spectrogrammes. Les valeurs des formants représentent la valeur moyenne entre le premier et le deuxième tiers de la voyelle. La détection a été effectuée grâce à l'algorithme de Burg implémenté dans le logiciel Praat (Boersma & Weenink, 2013). Les valeurs aberrantes des formants ont ensuite été corrigées en déterminant les valeurs correctes par inspection visuelle des spectrogrammes.

2.4 Extraction des mesures de protrusion P_{UL}, P_{LL} et P_C

Comme illustré sur la figure 1, les trois mesures considérées sont calculées comme l'avancement respectif de trois marqueurs par rapport à un plan de référence perpendiculaire au plan sagittal et passant par l'arête du nez: UL situé sur la lèvre supérieure, LL situé sur la lèvre inférieure, et le point C situé à équidistance des commissures gauche et droite. Les mesures P_{UL}, P_{LL} et P_C correspondent donc à la

distance sur l'axe z entre la référence et le point UL, LL ou C.

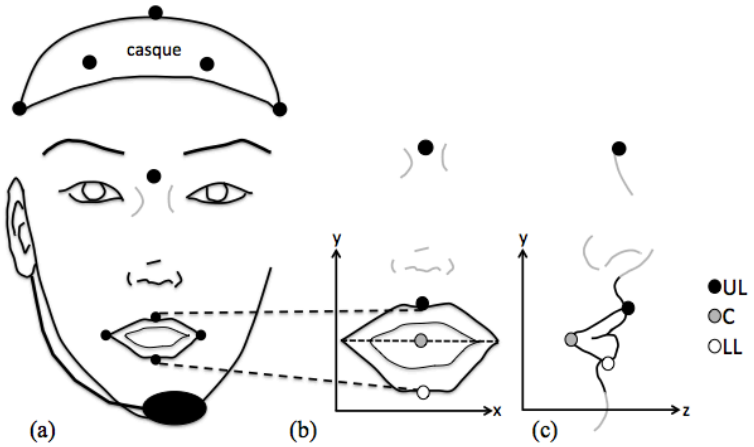


Figure 1 : En (a), schéma de la disposition des marqueurs Qualisys sur le contour externe des lèvres. En (b), représentation de face et en (c), représentation de profil des marqueurs situés sur la lèvre supérieure (UL, en noir), sur la lèvre inférieure (LL, en blanc) et sur la commissure (C, en gris).

3 Résultats

Compte tenu de l'importante variabilité inter-locuteur relevée dans les analyses articulatoires précédemment réalisées sur nos données (Georgeton & Fougeron, 2014), ainsi que dans les mesures de protrusion relevées dans la littérature (voir par exemple Abry et Boë, 1980, Dohen, 2005), nous présentons des analyses séparées par locuteur. En accord avec les résultats de Robert et al. (2007), les mesures de protrusion sont inter-corrélées pour tous les locuteurs, la corrélation la moins importante étant observée pour S3 entre les mesures P_{LL} et P_C (r=.90).

3.1 Distinction des voyelles antérieures selon les mesures de protrusion

Afin de prendre en compte les différences d'aperture entre les voyelles antérieures en complément des différences d'arrondissement, nous avons comparé pour chaque mesure de protrusion (P_{UL}, P_{LL} et P_C) les valeurs obtenues pour la voyelle arrondie et pour la voyelle non-arrondie dans les paires /i-y/, /e-ø/, /ɛ-œ/. Ces comparaisons multiples ont été effectuées au moyen de tests-t avec correction de type Bonferroni (Benjamini & Hochberg, 1995). Les résultats sont récapitulés dans la table 2.

Les comparaisons par paires de voyelles de même aperture pour chaque locutrice indiquent que seule la mesure P_C distingue la paire /i-y/ pour les trois locutrices. Les voyelle /e/ et /ɛ/ se distinguent de leurs contre-parties arrondies /ø/ et /œ/ pour une locutrice S1 uniquement: sur les trois mesures P_{UL}, P_{LL} et P_C pour /e-ø/, sur P_C pour /ɛ-œ/. Pour la locutrice S1, la mesure P_C est donc la seule à distinguer les trois paires /i-y/, /e-ø/ et /ɛ-œ/. Pour S2 et S3, seule la paire /i-y/ est discriminée, par les trois mesures pour S2 et uniquement par P_C pour S3. Ces comparaisons sont illustrées par la figure 2 pour la mesure de protrusion Pc.

		/i-y/	/e-ø/	/ɛ-œ/
S1	P_{LL}	t(60)=-3 ; p=.022 /i/ < /y/	t(62)=-3 ; p=.030 /e/ < /ø/	t(58)=-1 ; ns
	P_{UL}	t(60)=-4 ; p<.001 /i/ < /y/	t(62)=-3 ; p=.002 /e/ < /ø/	t(59)=-.5 ; ns
	P_C	t(60)=-6 ; p<.001 /i/ < /y/	t(62)=-5 ; p<.001 /e/ < /ø/	t(62)=-5 ; p=.006 /ɛ/ < /œ/
S2	P_{LL}	t(55)=-3 ; p=.039 /i/ < /y/	t(47)=-1 ; ns	t(52)=-.4 ; ns
	P_{UL}	t(54)=-3 ; p=.018 /i/< /y/	t(46)=-1 ; ns	t(52)=-.6 ; ns
	P_C	t(54)=-5 ; p<.001 /i/ < /y/	t(44)=-2 ; ns	t(52)=.3 ; ns
S3	P_{LL}	t(44)=-.3 ; ns	t(46)= .4 ; ns	t(52)=-.4 ; ns
	P_{UL}	t(44)=-2 ; ns	t(46)=-1 ; ns	t(52)=-.8 ; ns
	P_C	t(44)=-3 ; p=.007 /i/ < /y/	t(46)=-1 ; ns	t(52)=.7 ; ns

Table 2 : Résultats du test-t avec correction de type Benjamini & Hochberg (1995) pour chaque locuteur et chaque mesure de protrusion : avancement de la lèvre supérieure (P_{UL}), de la lèvre inférieure (P_{LL}) et la commissure (P_C). ns : non significatif.

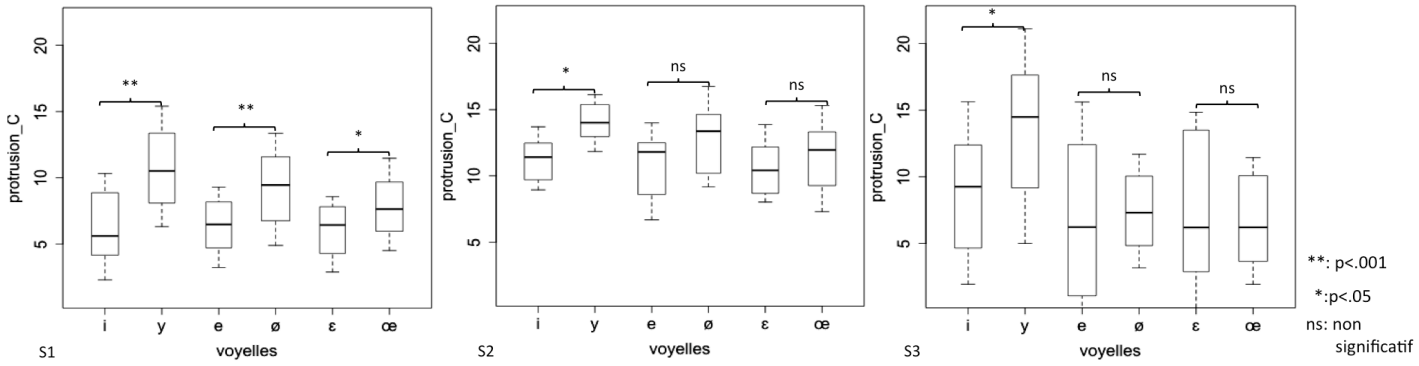


Figure 2 : Distribution des valeurs de protrusion P_C, pour les paires /i-y/, /e-ø/ et /ɛ-œ/, et différences significatives au sein des paires de voyelles pour les locuteurs S1, S2 et S3.

3.2 Relation entre mesures de protrusion et formants pour le contraste [±**arrondi**]

Pour les voyelles antérieures, le mouvement de protrusion implique une baisse des formants associés à la cavité antérieure, c'est-à-dire F2 mais également F3 (Vaissière, 2011). Les voyelles /i/ et /y/ sont également considérées comme des voyelles focales

caractérisées par un rapprochement de formants (Vallée, 1994, Schwartz et al., 1997). Entre ces deux voyelles, la distance entre F3 et F2 sera maximalement basse pour /y/ et maximalement élevée pour /i/.

Les corrélations entre chaque mesure de protrusion et les mesures acoustiques F3, F2 et F3-F2 sont présentées dans la table 3, séparément pour chaque paire de voyelles antérieures de même aperture.

		P_{LL}			P_{UL}			P_C		
		S1	S2	S3	S1	S2	S3	S1	S2	S3
/i-y/	F2	-.3*	ns	ns	-.4**	ns	ns	-.6**	-.3*	-.3*
	F3	-.4*	-.3*	ns	-.5**	-.3*	ns	-.6**	-.5**	-.4*
	F3-F2	ns	-.3*	ns	ns	-.3*	ns	-.2**	-.5**	ns
/e-ø/	F2	-.3*	ns	ns	-.4*	ns	ns	-.5**	ns	ns
	F3	ns	ns	ns	-.3*	ns	ns	-.4*	ns	ns
/ɛ-œ/	F2	ns	ns	ns	ns	ns	ns	-.4*	ns	-.4*
	F3	ns	ns	ns	ns	ns	ns	-.3*	ns	.2*

Table 3 : Résultats des corrélations par paire de voyelles de même aperture entre les paramètres acoustiques F2, F3, F3-F2 et chacune des mesures articulatoires : protrusion de la lèvre supérieure (P_{UL}) de la lèvre inférieure (P_{LL}) et de la commissure (P_C) pour les paires /i-y/, /e-ø/ et /ɛ-œ/. * : $p<.05$; ** : $p<.001$; ns (case grisée) : corrélation non-significative.

Les mesures acoustiques et les mesures de protrusion sont corrélées négativement montrant que, comme attendu, l'augmentation de la protrusion est associée à une diminution des valeurs de F2 et F3 (et de l'écart entre F2 et F3 pour la paire /i-y/) et vice-versa. Pour la paire /i-y/, les formants F2 et F3 sont significativement corrélés à la mesure P_C pour les 3 locutrices (de -.3 à -.6 suivant la locutrice, comme indiqué dans la table 3). La mesure de protrusion la plus corrélée à F3-F2 pour la paire /i-y/ est P_C, pour deux des locutrices. Pour la paire /e-ø/, les formants F2 et F3 ne sont significativement corrélés aux mesures de protrusion que pour la locutrice S1. Pour la paire /ɛ-œ/, seule P_C est significativement corrélée à F2 et F3 pour deux des locutrices. De plus, quelle que soit la locutrice et la mesure acoustique considérée, les corrélations les plus importantes sont observées pour la mesure P_C.

4 Conclusions

Cette étude montre que parmi les trois mesures de protrusion correspondant au déplacement en avant (i) de la lèvre supérieure, (ii) de la lèvre inférieure et (iii) des commissures des lèvres, la mesure la plus à même de rendre compte de la différence entre voyelles antérieures arrondies et non-arrondies du français dans nos données est l'avancement de la commissures des lèvres P_C. Cependant, cette distinction n'est

observée pour les trois paires /i-y/, /e-ø/ et /ɛ-œ/ que pour l'une des trois locutrices (S1). Ces différences interindividuelles sont en accord avec les résultats de la plupart des études multi-locuteurs de l'articulation labiale (voir par exemple Abry et Boë, 1980, ou sur les mêmes données Georgeton & Fougeron, 2014). Les autres mesures de protrusion P_{UL} et P_{LL} distinguent uniquement la paire /i-y/ pour les locutrices S1 et S2 et /e-ø/ pour la locutrice S2. Ce résultat peut probablement s'expliquer par des différences anatomiques entre nos locutrices. Le positionnement des marqueurs sur le vermillon des lèvres est un positionnement largement choisi dans les études voulant rendre compte de variations articulatoires au niveau des lèvres (par exemple, Dohen (2005) avec des mesures effectuées par capture optique de mouvements ou encore Tabain & Perrier (2007) avec des mesures par articulographie électromagnétique). Cependant, ce positionnement pourrait également ne refléter que partiellement les mouvements des lèvres et commissures elles-mêmes.

L'extension de l'analyse à un plus grand nombre de locuteurs serait nécessaire avant de généraliser l'emploi de cette mesure. L'ampleur de la variabilité anatomique interindividuelle observée pourrait néanmoins remettre en question la possibilité de définir une mesure unique, qui serait la plus représentative tous locuteurs confondus.

Pour expliquer cette meilleure séparation entre voyelles arrondies et non-arrondies également opérée par la mesure P_C sur leurs données, Abry et Boë (1980) considèrent que le déplacement en avant de la commissure des lèvres reflète plus fidèlement la composante de rétraction. Notre étude étend ces résultats à la paire /ɛ-œ/, et montre que cette mesure reflète également mieux les conséquences acoustiques attendues, c'est-à-dire l'abaissement des formants F3 et F2 pour les voyelles arrondies vs. non-arrondies (ainsi que la réduction de la distance focale F3-F2 pour /i-y/), suggérant ainsi une meilleure prise en compte de l'allongement de la cavité antérieure.

Le contraste d'arrondissement apparaît comme variable selon le degré d'aperture, la protrusion labiale ne permettant pas de discriminer la paire /ɛ-œ/. Zerling (1990) classifie les voyelles selon les paramètres d'aperture aux lèvres et d'étirement des lèvres à partir de données de 105 locuteurs et identifie trois groupes de voyelles du français, sans toutefois prendre en compte la protrusion : les voyelles /i, e, ɛ/ non-labialisées, la voyelle /œ/ labialisée et /y, ø/ fortement labialisées. Nos résultats suggèrent que seules les voyelles fortement labialisées dans la classification de Zerling (1990) se distinguent des autres voyelles antérieures par le geste de protrusion, avec un degré de labialisation intermédiaire pour la voyelle /ø/.

Remerciements

Merci à Cécile Fougeron pour ses conseils avisés. Merci à nos gentilles locutrices.

Références

Abry, C. & Boë, L.-J (1980). *A la recherche de corrélats géométriques discriminants pour l'opposition d'arrondissement vocalique en français*. In Abry, C., Boë, L.-J., Corsi, P., Descout, R., Gentil, M., & Graillot, P. *Labialité et phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales*, 217-237. Publication de l'Université des Langues et Lettres de Grenoble.

Abry C. & Boë L.-J. (1986). Laws for lips. Speech Communication, 5(1) :97-104.

Benjamini Y. & Hochberg Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), p. 289-300.

Boersma, P. & Weenink D. (2013). Praat: doing phonetics by computer (Version 5.3.56)

Dohen M. (2005). *Deixis prosodique multisensorielle: production et perception audiovisuelle de focalisation contrastive en français*. Thèse de l'INP Grenoble.

Fant G. (1960). Acoustic theory of speech production, The Hague, Paris.

Fromkin V. (1964). Lip positions in American English vowels. *Language and Speech*, 7(3), 215-225.

Georgeton L. & Fougeron C. (2014). Domain-initial strengthening on French vowels and phonological contrasts: evidence from lip articulation and spectral variation. *Journal of phonetics*.

Ladefoged P. (1979). Articulatory parameters. Status report. In *Proceedings International Congress of Phonetic Science,* Copenhague, Danmark, 41-47.

Ménard L., Leclerc A., Tiede M., Prémont A., Turgeon C., Trudeau-Fisette P. & Côté D. (2013). Correlates of contrastive focus in congenitally blind adults and sighted adults. *14th Annual Conference of the International Speech Communication Association*, Lyon, France.

Nordstrand M., Svanfeldt G., Granström B. & House D. (2004). Measurements of articulatory variation in expressive speech for a set of Swedish vowels. *Speech Communication* 44, 187-196.

Robert V., Bonneau A., Wrobel-Dautcourt B. & Laprie Y. (2007). Prédiction phonétique de la coarticulation labiale. *Perturbations et réajustements : langue et langage*. Publications de L'université Marc Bloch, Strasbourg, 155-167.

Schwartz J-L., Boë L-J, Vallée N. & Abry C. (1997). The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics*, 25, 255-286.

Tabain, M. & Perrier, P. (2005). Articulation and acoustic study of /u/ pre-boundary position in French. The interaction of compensatory articulation, neutralization avoidance and featural enhancement. *Journal of Phonetics, 35, 135-161.*

Vallée N. (1994). *Systèmes vocaliques : de la typologie aux prédictions*, Thèse de l'Université Stendhal Grenoble III.

Vaissière J. (2011). On the acoustic and perceptual characterization of reference vowels in a crosslanguage perspective. In *Proceedings of the 17th international congress of phonetic sciences,* Hong-Kong, China. 52 59

Zerling J.-P. (1990). *Aspects articulatoires de la labialité vocalique en français. Contribution à la modélisation à partir de labio-photographies, labiofilms et films radiologiques. Etude statique, dynamique et contrastive.* Thèse de l'Université de Strasbourg II.

Perception de la parole rapide chez les enfants présentant une dysphasie expressive

Hélène Guiraud[1], Emmanuel Ferragne[2], Nathalie Bedoin[1], Sonia Krifi-Papoz[3], Vania Herbillon[4], Aurélia Bascoul[5], Sibylle Gonzalez-Monge[6], Véronique Boulenger[1]

(1) Laboratoire Dynamique Du Langage, CNRS/Université Lyon 2 UMR 5596, Lyon, France
(2) CLILLAC-ARP, Université Diderot Paris 7, Paris, France
(3) Service de Neurologie Pédiatrique, HFME, Bron, France
(4) Service Épilepsie, Sommeil et Explorations fonctionnelles neurologiques, HFME, Bron, France
(5) Service d'Audiologie et d'Explorations orofaciales, Hôpital Edouard Herriot, Lyon, France
(6) Centre de Référence des Troubles des Apprentissages, L'Escale, Bron, France

helene.guiraud@univ-lyon2.fr, veronique.boulenger@cnrs.fr

RESUME

Notre système cognitif est capable de s'adapter rapidement aux variations de débit de parole afin de comprendre le message. Des études suggèrent que cette adaptation active les régions prémotrices du cerveau, impliquées dans la production de parole. Cette étude examine comment les enfants présentant des troubles expressifs du langage traitent la parole produite à un débit rapide. Dix-sept enfants (8-13 ans) présentant une dysphasie expressive, appariés à des enfants contrôles, ont réalisé une tâche de jugement sémantique de phrases produites à trois débits de parole (normal, naturel rapide et compressé artificiellement). Les performances des enfants dysphasiques sont plus faibles, en termes de temps de réponse et d'indice de sensibilité (d'), que les contrôles, plus particulièrement lorsque le débit est rapide. Un déficit de perception de parole rapide semble donc exister chez ces enfants, suggérant un rôle des processus impliqués dans la production de la parole dans les processus de perception.

ABSTRACT

Processing of fast speech speech in children with speech output disorders

Our cognitive system is able to adapt rapidly to changes in speech rate to eventually understand the message. This adaptation has been shown to activate brain premotor regions, involved in speech production. This study examines how children with specific language impairment process speech produced at a rapid rate. Seventeen children (8-13 years old) with specific speech output disorders, matched to healthy controls, performed a semantic judgment task on sentences produced at three speech rates (normal, natural fast and time-compressed). Performance of children with specific speech output disorders were generally lower, in terms of response times and sensitivity index (d'), than their controls, particularly when they had to discriminate incoherent sentences produced at a rapid rate. A deficit in the perception of fast speech therefore seems to exist in these children, suggesting a role of speech production processes in speech perception.

MOTS-CLES : perception de la parole ; dysphasie expressive ; parole rapide.
KEYWORDS: speech perception ; speech output disorders ; fast speech.

1 Introduction

Dans la vie quotidienne, tout auditeur, adulte comme enfant, est confronté à des interlocuteurs variés s'exprimant à des vitesses différentes. La compréhension du message peut être difficile en début de conversation, mais en quelques minutes, l'accès à l'information linguistique s'améliore, facilitant l'échange entre les interlocuteurs. Cette adaptation rapide aux variations du signal de parole est garantie par la remarquable flexibilité de notre système cognitif à normaliser un large éventail de paramètres du signal acoustique afin de rendre le discours compréhensible. Lorsque le débit de parole augmente, les gestes articulatoires effectués plus rapidement sont moins précis (Kohler, 1992), ce qui entraîne des phénomènes de réduction ainsi qu'une coarticulation et une assimilation plus importantes : les modifications sont non seulement temporelles mais aussi spectrales. Ces phénomènes sont en outre non-linéaires : en anglais et en français, les consonnes et les syllabes accentuées (anglais) sont moins réduites que les voyelles et les syllabes non accentuées (Duez, 1998 ; Max, 1997). Ces modifications spectro-temporelles sont spécifiques à la parole naturellement accélérée et créent une difficulté supplémentaire pour l'auditeur, exigeant un temps de traitement plus important et des ressources cognitives supplémentaires.

Des études ont montré que les adultes, jeunes ou âgés, sont parfaitement capables de s'adapter à la parole accélérée artificiellement (par compression temporelle du signal de parole naturelle produite à un débit normal) et ce jusqu'à un taux de compression de 65% (Dupoux & Green, 1997; Golomb et al, 2007; Janse et al, 2003). Dupoux et Green (1997) ont demandé à des adultes d'écouter puis de rappeler par écrit des phrases compressées à 38% et 45% de la durée originale du signal. Une amélioration des performances était observée après l'écoute de 5 à 15 phrases, reflétant la mise en place rapide d'un processus d'adaptation à ces deux taux de compression, avec toutefois une adaptation plus longue pour la parole plus fortement compressée. Le processus d'adaptation a également été étudié de façon plus écologique, avec l'écoute de parole produite naturellement à débit rapide (Adank & Janse, 2009; Janse, 2004). Dans ce cas, la compression temporelle du signal dépend du type de segments, et des modifications spectrales s'y ajoutent. Adank et Janse (2009), dans une tâche de jugement sémantique de phrases présentées à un débit normal, rapide ou compressé, ont ainsi montré, chez des adultes néerlandais, une adaptation plus longue et plus difficile à la parole naturelle rapide qu'à la parole compressée. Récemment, nous avons montré que les enfants de langue maternelle française de 8 à 11 ans s'adaptaient aussi rapidement que les adultes à la parole rapide (naturelle et compressée) (Guiraud et al., 2013). Si l'adaptation était comparable pour les deux types de parole accélérée, les scores de bonnes réponses plus faibles pour la parole naturelle rapide que pour la parole compressée reflètent néanmoins des différences qualitatives dans le traitement de ces deux types de parole accélérée.

Pour faire face aux particularités temporelles et spectrales de la parole naturelle rapide, l'auditeur doit sans doute passer d'un ensemble de règles acoustico-phonétiques à un autre, plus approprié à l'extraction d'indices pertinents pour l'identification phonologique (Francis et al., 2000 ; Golomb et al., 2007 ; Guiraud et al., 2013). Ce processus exigeant au niveau cognitif nécessiterait un temps de traitement plus long et recruterait des aires cérébrales supplémentaires. C'est en revanche sans doute moins nécessaire pour traiter la parole compressée artificiellement, même si des études en imagerie par résonance magnétique fonctionnelle indiquent que cette dernière sollicite parfois des substrats cérébraux spécifiques. Adank et Devlin (2010) ont ainsi montré une activation du cortex prémoteur gauche lors de la phase d'adaptation à la parole compressée, absente pour la parole à débit normal (voir aussi Osnes et al., 2011). Ces résultats concordent avec le modèle à double voie de la perception de la parole (Hickok & Poeppel, 2007) qui postule l'existence de deux voies

corticales de traitement de la parole : la voie ventrale temporale (voie du « quoi ») permettrait la reconnaissance du signal et participerait au couplage des représentations phonologiques et sémantiques, et la voie dorsale pariéto-frontale (voie du « comment ») servirait d'interface sensorimotrice appariant les représentations phonologiques et articulatoires et suppléant la voie ventrale pour comprendre le message en conditions d'écoute difficiles.

Si les régions articulatoires du cerveau jouent un rôle dans la perception de la parole, en particulier dégradée, les enfants présentant une dysphasie expressive (dysphasie phonologico-syntaxique et dyspraxie verbale) devraient être particulièrement perturbés par la parole rapide. La dysphasie est en effet un trouble spécifique, sévère et durable du développement du langage oral, caractérisé par des déficits de production et/ou de compréhension du langage parlé (Maillart et al., 2012). La forme de dysphasie la plus fréquente, la dysphasie phonologico-syntaxique, touche principalement le versant expressif du langage et serait liée à un trouble de la programmation phonologique (Parisse & Maillart, 2010). La dyspraxie verbale est quant à elle définie par une réduction, voire une absence de langage expressif en raison d'un déficit d'organisation, de programmation et de production des gestes articulatoires associés aux sons de parole et aux séquences verbales. Des liens sont mis en évidence entre les déficits en perception et en production de la parole dans la dyspraxie verbale, ce qui est cohérent avec le modèle de la double voie et l'idée selon laquelle des régions corticales spécialisées dans les représentations articulatoires interviendraient en perception. Des corrélations sont en effet décrites entre les performances de tels enfants dans des tâches de production et de perception de parole. Ils discriminent par exemple mal la frontière entre /d/ et /b/ et leurs performances sont corrélées au nombre de phonèmes qu'ils substituent en production (Groenen et al., 1996; voir aussi Nijland, 2009). Un trouble du traitement du rythme est aussi rapporté chez des enfants dysphasiques, dans des situations impliquant la motricité aussi bien que la perception auditive. Les dysphasiques présentent ainsi des difficultés à s'ajuster au rythme d'un métronome lors d'une tâche de « tapping » (battre la mesure avec le doigt) sur un tempo lent (2 Hz) (Corriveau & Goswami, 2009 ; Hill, 2001). Ils ont aussi des difficultés dans des tâches de discrimination auditive non verbale mesurant la sensibilité à des indices rythmiques par ailleurs importants pour la segmentation syllabique (temps de montée de l'enveloppe d'amplitude et durée du signal), alors que leurs performances sont bonnes dans des tâches non verbales n'impliquant pas d'aspect rythmique (jugement d'intensité ; Corriveau et al., 2007).

Dans notre étude, au vu des liens étroits unissant les systèmes de perception et de production de la parole (Pulvermüller & Fadiga, 2010 pour une revue) et de l'implication des régions prémotrices dans la perception de la parole dégradée (Hickok et Poeppel, 2007), nous avons examiné comment les enfants présentant une dysphasie phonologico-syntaxique ou une dyspraxie verbale traitent la parole accélérée (naturellement ou artificiellement). Ces deux types de dysphasie expressive résultant d'un déficit de la programmation phonologique ou articulatoire, notre hypothèse est qu'elles devraient se traduire par des difficultés spécifiques à percevoir la parole rapide, d'autant plus marquées en cas d'accélération naturelle : les phénomènes de coarticulation plus importants devraient en effet rendre le traitement particulièrement difficile compte tenu des appariements auditivo-moteurs déficitaires chez ces enfants.

2 Matériel et méthode

2.1 Participants

Dix-sept enfants présentant une dysphasie expressive (phonologico-syntaxique, N = 13 ;

dyspraxie verbale, N = 4) de 8 à 13 ans (moyenne = 11.1 ans, ET 1.45) ont participé à l'étude. Chacun était apparié à un contrôle de même âge et de même sexe (moyenne = 11 ans, ET 1.3). Le protocole a reçu un avis favorable du Comité de Protection des Personnes Sud-Est II et tous les enfants et leurs parents ont signé un formulaire de consentement avant l'expérience. Le diagnostic a été posé par les neuropsychologues et/ou orthophonistes de différents services hospitaliers. Tous les enfants étaient de langue maternelle française, droitiers, sans trouble de l'audition, avec un QI supérieur à 70, et ne présentaient ni de déficit de l'attention ni de trouble autistique. Une batterie d'évaluation neuropsychologique complémentaire a permis de s'assurer du maintien de certains aspects du diagnostic lors de l'inclusion dans l'étude.

2.2 Stimuli

Trois cents phrases (7-9 mots) ont été créées suivant la même structure syntaxique : Déterminant – Nom 1 – Verbe – Déterminant – Nom 2 – Préposition – Déterminant – Nom 3. La cohérence sémantique du Nom 3 (mot cible CVCV) au sein du contexte phrastique a été manipulée : ce mot était cohérent avec le contexte dans la moitié des phrases (n = 150) et incohérent dans l'autre moitié (n = 150). Chaque mot cible apparaissait dans les deux conditions de cohérence sémantique (e.g., « Sa fille déteste la nourriture de la cantine » et « Le public applaudit le joueur pour sa cantine »).

Les phrases ont été enregistrées (44.1 kHz, mono, 16 bits) par un locuteur français de sexe masculin (32 ans), dans une salle sourde à l'aide du logiciel ROCme ! (Ferragne et al., 2012). Chaque phrase a été enregistrée à un débit normal puis à un débit rapide. La durée des 2×300 phrases ainsi que le nombre de syllabes de chaque phrase ont été calculés. Le débit normal moyen était de 6.76 syllabes/s (ET 0.57) et le débit rapide moyen de 9.15 syllabes/s (ET 0.60), soit un ratio moyen de 0.74 (i.e., facteur d'accélération de 1.35). La compression temporelle a ensuite été appliquée à partir des enregistrements à débit normal, à l'aide de l'algorithme PSOLA implémenté dans Praat. Cette compression correspond à une resynthèse de chaque stimulus de parole à débit normal, ne modifiant pas la hauteur du son, mais seulement son aspect temporel. Chaque phrase compressée était appariée en termes de débit à sa phrase correspondante en débit naturel rapide. Un filtre passe-haut de 80 Hz, ainsi qu'un fondu d'entrée et de sortie sur l'enveloppe d'amplitude, ont été appliqués sur la totalité des fichiers sons ; le pic d'intensité a également été normalisé.

Les 900 phrases ont été réparties dans 12 listes expérimentales de 75 stimuli chacune selon la méthode du carré latin. La fréquence d'occurrence des mots cibles dans la langue, leurs nombres de phonèmes et de voisins phonologiques ne différaient pas significativement entre les listes. Chaque liste était composée de 3 blocs expérimentaux de 25 stimuli chacun (13 phrases cohérentes/12 phrases incohérentes) correspondant aux 3 conditions de débit de parole. Les blocs étaient toujours présentés dans le même ordre : débit normal, débit naturel rapide et parole compressée. Chaque liste contenait chaque phrase une seule fois afin d'éviter les effets de répétition et aucune paire de phrases (cohérente/incohérente) contenant le même mot cible n'était présentée dans la même liste. Au travers des 12 listes, chaque mot cible était présenté dans les 6 conditions (3 conditions de débit de parole × 2 conditions de cohérence sémantique). Enfin, dans chaque bloc, l'ordre des phrases variait aléatoirement entre les participants.

2.3 Procédure

L'enfant était assis devant un ordinateur portable, un casque audio sur les oreilles, dans une salle

calme. Il se voyait présenter l'une des 12 listes expérimentales. Tous les essais se déroulaient de la même manière : une croix de fixation blanche sur fond noir apparaissait au centre de l'écran et, une seconde après, le stimulus verbal était présenté via le casque à un niveau d'écoute confortable, la croix de fixation restant toujours affichée. Après l'écoute de la phrase, l'enfant devait juger le plus rapidement et précisément possible de la cohérence sémantique de la phrase en appuyant sur l'une de deux touches prédéfinies du clavier (réponse avec l'index gauche pour une phrase cohérente et avec l'index droit pour une phrase incohérente). Si aucune réponse n'était donnée dans les 7 secondes, l'essai suivant était automatiquement présenté. Après 5 phrases d'entraînement (différentes des stimuli expérimentaux), la liste expérimentale était présentée. La présentation des stimuli et l'enregistrement des réponses manuelles étaient effectuées avec E-Prime 2 (Psychology Software Tools, Inc., www.pstnet.com). L'expérience durait 15 minutes.

Les temps de réponse (TR, délai entre le début de présentation du mot cible et la réponse manuelle) et les scores ont été enregistrés. Les performances ont été analysées dans le cadre de la théorie de la détection du signal, en calculant l'indice de sensibilité d' et le biais de réponse β de chaque participant pour chaque condition expérimentale. Ces analyses sont basées sur la proportion de *hits* (i.e., réponses correctes pour les phrases incohérentes) et de fausses alarmes (i.e., erreurs pour les phrases cohérentes ; Macmillan & Creelman, 1991).

2.4 Analyses statistiques

Les essais où le participant ne donnait pas de réponse ou une mauvaise réponse ont été exclus de l'analyse des TR, de même que les TR non compris dans l'intervalle [moyenne +/- 2.5 × ET].

D'abord, nous avons analysé les TR moyens des participants à l'aide d'une analyse de variance (ANOVA) à trois facteurs : le facteur à mesures répétées « Débit» à trois modalités (Normal, Naturel Rapide et Compressé), le facteur à mesures répétées « Plausibilité sémantique » à deux modalités (Cohérence et Incohérence) et le facteur à mesures indépendantes « Groupe » à deux modalités (Dysphasiques et Contrôles). Ensuite, les indices de sensibilité (d') moyens ont été définis comme variables dépendantes dans une ANOVA avec les facteurs Débit et Groupe. En cas d'interaction significative, des tests post-hoc de Tukey ont été réalisés afin de déterminer les différences entre les conditions.

3 Résultats

3.1 Analyse des temps de réponse

L'ANOVA Groupe × Plausibilité × Débit montre un effet significatif de la Plausibilité ($F = 17.02$, $p < .001$) : l'ensemble des enfants éprouve plus de difficultés à catégoriser les phrases incohérentes (1461 ms, ET 507) que les phrases cohérentes (1283 ms, ET 383). Aucun effet du Débit ni du Groupe ne sont observés. Une interaction tendancielle est notée entre les facteurs Plausibilité et Groupe ($F = 3.12$, $p = .08$) ; les dysphasiques sont plus longs à traiter les phrases incohérentes que cohérentes ($p = .048$) alors que ce n'est pas le cas chez les contrôles. Une interaction significative est révélée entre les facteurs Débit et Plausibilité ($F = 0.77$, $p < .05$). Le test de Tukey comparant les moyennes à l'issue de cette interaction ne montre pas que l'effet négatif de l'incohérence soit seulement dû à l'une des conditions de débit. Toutefois, une différence de TR plus importante est observée entre les phrases cohérentes et incohérentes pour la parole compressée (279 ms) que pour la parole normale et naturelle rapide (respectivement 111 ms

et 121 ms). De plus, l'observation de la cette différence est plus marquée chez les dysphasiques (411 ms en parole compressée, 168 ms en parole normale et 161 ms en parole naturelle rapide) que chez les contrôles (147 ms, 54 ms et 81 ms). Au vu de ces résultats, nous avons effectué une ANOVA Débit × Plausibilité pour chaque groupe. Aucun effet significatif n'est observé chez les contrôles, alors qu'un effet de la Plausibilité (F = 12.53, p < .01) ainsi qu'une interaction tendancielle Débit × Plausibilité (F = 3.1, p = .058) sont notés chez les dysphasiques. Ceci suggère bien que les dysphasiques tendent à se distinguer en étant particulièrement longs à identifier les phrases incohérentes sémantiquement lorsque la parole est compressée temporellement (Figure 1).

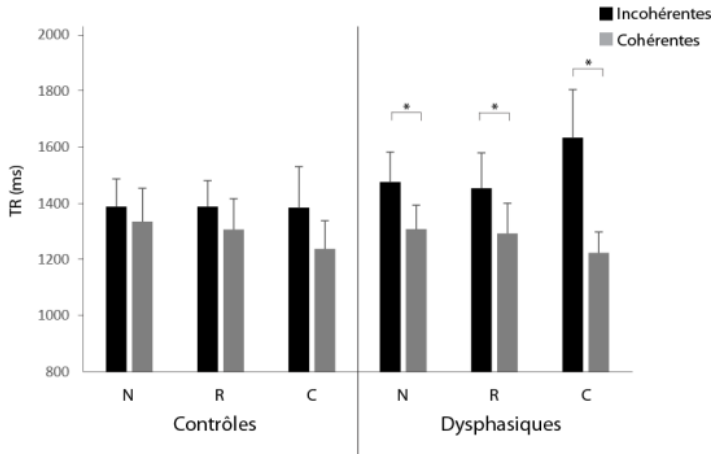


FIGURE 1 – Temps de réponse moyens (TR) dans chaque condition de débit de parole (normal (N), naturel rapide (R), compressé (C)) pour les phrases cohérentes et incohérentes, pour chaque groupe de participants. (*) indique une différence significative entre les conditions. Les barres d'erreur indiquent la différence standard entre les participants.

3.2 Analyse de la sensibilité

L'ANOVA conduite sur les d' révèle un effet significatif du facteur Groupe (F = 13.13, p < .01), qui s'explique par une moins bonne sensibilité à l'incohérence sémantique chez les dysphasiques (d' = 1.13, ET 0.9) que chez les contrôles (d' = 1.96, ET 0.9). L'effet principal du Débit est également significatif (F = 8.94, p < .001) : le test de Tukey montre qu'il s'explique principalement par une forte tendance de l'ensemble des enfants à être moins sensibles à l'incohérence sémantique en cas de débit naturel rapide qu'en cas de débit normal (p = .06 ; d' = 1.27 en débit naturel rapide vs. d' = 1.88 en débit normal ; Figure 2). En revanche, sur l'ensemble des enfants, la condition de parole compressée ne diffère pas significativement des deux autres conditions de débit. Les facteurs Débit et Groupe n'interagissent pas significativement. Suite à ces observations, et selon nos hypothèses, nous avons effectué une ANOVA sur le facteur Débit pour chaque groupe. Aucune différence significative entre les différentes conditions de débit n'est observée chez les contrôles. En revanche, chez les dysphasiques, l'effet du facteur Débit est significatif (F = 6013, p < .01). Des tests t de Student appariés révèlent une différence significative entre les conditions de débit normal et rapide (p < .01) et, normal et compressé (p < .01). Il semble donc que les enfants dysphasiques éprouvent plus de difficultés à discriminer les phrases incohérentes sémantiquement lorsque le débit est accéléré (naturellement et artificiellement).

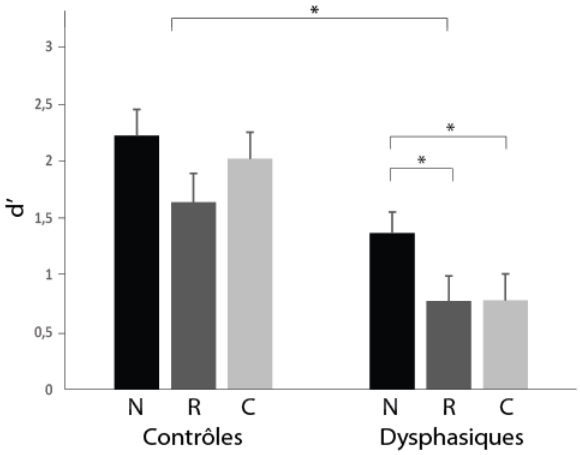


FIGURE 2 – Indice de sensibilité *d'* dans chaque condition de débit de parole pour chaque groupe de participants. (*) indique une différence significative entre les conditions. Les barres d'erreur indiquent la différence standard entre les participants.

4 Discussion

Notre étude avait pour but d'examiner les processus de perception de la parole produite à un débit rapide, dans une tâche de jugement sémantique de phrases, chez des enfants présentant une dysphasie expressive. Nos résultats révèlent un effet de la plausibilité sémantique des phrases chez les deux groupes d'enfants (dysphasiques et contrôles), avec des TR plus longs lors de l'identification des phrases incohérentes. Ce résultat est en accord avec la littérature montrant un traitement plus difficile des stimuli verbaux non plausibles sémantiquement par rapport aux stimuli plausibles (Adank & Janse, 2009 ; Kutas & Hillyard, 1980). L'analyse du *d'* montre que, globalement, les enfants dysphasiques sont moins sensibles à l'incohérence sémantique des phrases entendues que leurs contrôles, reflétant l'existence d'une difficulté particulière de traitement sémantique lors de la perception de la parole dans la dysphasie expressive. Notre principal objectif était de déterminer comment les enfants présentant une dysphasie expressive traitent la parole produite à un débit rapide. Nos résultats tendent à montrer que lorsque le débit de parole est accéléré, ces enfants présentent une sensibilité plus faible à l'incohérence sémantique comparée à la perception de parole produite à un débit normal (Figure 2). Autrement dit, leurs capacités à discriminer les phrases incohérentes des phrases cohérentes diminuent, et ils commettent un plus grand nombre d'erreurs de jugement sémantique. Les analyses réalisées indépendamment sur chaque groupe d'enfants montrent un effet significatif du débit sur les *d'* des enfants dysphasiques mais pas sur ceux des enfants contrôles. Nos données suggèrent donc que les enfants présentant une dysphasie expressive pourraient être plus fortement affectés par un débit de parole rapide, que l'accélération soit naturelle ou artificielle, que les enfants sains lorsqu'il s'agit de détecter des anomalies sémantiques dans des phrases. Cette interprétation semble corroborée par les résultats obtenus sur les TR : l'interaction tendancielle Groupe × Plausibilité suggère un allongement des TR, pour décider qu'une phrase est incohérente, plus important chez les dysphasiques que chez les contrôles. De plus, l'interaction tendancielle Débit × Plausibilité chez les dysphasiques montre bien une difficulté particulière pour ces enfants à traiter l'incohérence lorsque la parole est accélérée (particulièrement pour la parole compressée), alors qu'aucun effet n'est observé chez les enfants contrôles.

Nos données soulignent ainsi que les enfants présentant une dysphasie expressive éprouvent de plus grandes difficultés à détecter des anomalies sémantiques dans des phrases que les enfants

sains lorsque le débit de parole est accéléré (naturellement ou artificiellement). Ceci semble cohérent avec l'idée d'une implication de la voie dorsale du modèle d'Hickok et Poeppel (2007) lors de la perception de la parole, particulièrement lorsque celle-ci est dégradée, afin de garantir un traitement du signal plus rapide et plus efficace. Chez les enfants présentant des troubles expressifs du langage, la voie dorsale, qui apparie les représentations phonologiques et articulatoires, pourrait ne pas être aussi fonctionnelle que chez les enfants contrôles, conduisant à des difficultés pour traiter la parole dégradée (rapide). Le modèle de l'AST (Poeppel, 2003) suggère que les traitements phonémique, syllabique et phrastique seraient associés aux oscillations cérébrales, dans les régions dédiées au traitement de la parole, dans les bandes de fréquence γ (25-35Hz), θ (4-8Hz) et δ (1-3Hz) respectivement. Il existerait en outre une asymétrie fonctionnelle selon laquelle les cortex auditifs gauche et droit échantillonneraient le signal de parole à des fréquences différentes (γ et θ respectivement), offrant parallèlement des modes de découpage phonémique et syllabique. En accord avec ces modèles, des études en électro- et magnétoencéphalographie révèlent une synchronisation des oscillations θ des cortex auditifs sur les modulations lentes d'enveloppe d'amplitude du signal de parole, qui caractérisent le rythme de la parole et lui confèrent son intelligibilité (Peelle & Davis, 2012 pour une revue). Cet alignement a été observé pour de la parole intacte et dégradée (i.e. parole vocodée et compressée). Or, des études récentes ont montré un dysfonctionnement des mécanismes oscillatoires dans les bandes de fréquence γ et δ dans la dyslexie, un trouble spécifique de l'apprentissage de la lecture et de l'écriture principalement causé par un déficit phonologique (Lehongre et al., 2011 ; Soltész et al., 2013). Ceci était observé au niveau du cortex auditif gauche mais aussi dans les régions articulatoires et somatosensorielles gauches (pour la bande γ) appartenant à la voie dorsale. Il a alors été suggéré que ces déficits oscillatoires sous-tendent, au moins en partie, les déficits de segmentation syllabique, de perception de la prosodie et de traitement phonologique classiquement observés dans la dyslexie (Goswami, 2011). Les enfants dysphasiques et les dyslexiques présentent par ailleurs un trouble du rythme au niveau auditif et moteur (Corriveau & Goswami, 2009). Nous pouvons alors supposer l'existence, chez les enfants présentant une dysphasie expressive, d'une anomalie des oscillations θ (et potentiellement δ) dans les cortex auditifs ainsi que des oscillations γ dans les régions articulatoires. Ce dysfonctionnement pourrait conduire à des difficultés de synchronisation des oscillations θ sur le rythme de la parole ainsi qu'à l'élaboration de schémas articulatoires incorrects ou imprécis, ce qui affecterait particulièrement la perception de la parole produite à un débit rapide. Ces hypothèses restent évidemment à vérifier en examinant la dynamique des oscillations cérébrales lors de la perception de la parole rapide chez les enfants présentant des troubles expressifs du langage. Notre hypothèse d'une difficulté particulière de traitement de la parole naturelle rapide, par rapport à la parole compressée, n'est pas vérifiée. Ceci pourrait s'expliquer par le fait que dans cette étude, sur les 17 enfants dysphasiques testés, seuls 4 présentaient une dyspraxie verbale. Or ce trouble affectant la programmation et la production des gestes articulatoires, ces enfants pourraient trouver la parole (naturelle) rapide plus difficile à percevoir que les enfants ayant une dysphasie phonologico-syntaxique dans laquelle les troubles de la production sont moins prédominants. Il nous paraît donc nécessaire de poursuivre cette étude en recrutant un nombre plus important d'enfants présentant une dyspraxie verbale afin de comparer leurs performances à celles des enfants dysphasiques phonologico-syntaxique et ainsi, pouvoir préciser la difficulté particulière éventuelle du traitement de la parole naturelle rapide.

Remerciements

Cette étude a été menée dans le cadre de l'ANR JCJC ODYSSEE (PI : V. Boulenger). H. Guiraud est financée par une bourse doctorale du Labex ASLAN.

Références

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *J Acoust Soc Am*, *126*(5), 2649–2659.

Adank, P., Devlin, J.T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage, 49*, 1124-1132.

Corriveau, K., Pasquini, E., Goswami, U. (2007). Basic Auditory Processing Skills and Specific Language Impairment : A New Look at an Old Hypothesis. *J Speech Lang Hear Res*, *50*, 647–666.

Corriveau, K. H., & Goswami, U. (2009). Rhythmic motor entrainment in children with speech and language impairments: tapping to the beat. *Cortex*, *45*(1), 119–130.

Duez, D. (1998). Consonant sequences in spontaneous french speech. Sound patterns of Spontaneous Speech, ESCA Workshop, LA beaume-les-Aix, 63-68.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J Expl Psychol Hum Percept Perform*, *23*(3), 914–927.

Ferragne, E., Flavier, S. & Fressard, C. (2012). Recording of Oral Corpora Made Easy : logiciel d'enregistrement de corpus. http://ddl.ish-lyon.cnrs.fr.

Francis, A.L., Daldwin, K., Nusbaum, H.C. (2000). Effects of training on attention to acoustic cues. *Percept Psychophys, 62(8),* 1668-80.

Golomb, J. D., Peelle, J. E., & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *J Acoust Soc Am*, *121*(3), 1701–1708.

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends Cogn Sci, 15(1),* 3-10.

Groenen, Maassen, Crul, & Thoonen. (1996). The specific relation between perception and production errors for place of articulation in developmental apraxia of speech. *J Speech Lang Hear Res*, *39*, 468–482.

Guiraud, H., Ferragne, E., Bedoin, N., Boulenger, V. (2013). Adaptation to natural fast speech and time-compressed speech in children. In *Interspeech* (pp. 1–5). Lyon.

Hickock, G., Poeppel, D. (2007). The cortical organization of speech processing. *Nature Rev Neurosc, 8*, 393-402.

Hill, E. L. (2001). Non-specific nature of specific language impairment : a review of the literature with regard to concomitant motor impairments, *Int J Lang Comm Dis*, *36*(2), 149–171.

Janse, E. (2004). Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Comm*, *42*(2), 155–173.

Janse, E., Nooteboom, S., & Quené, H. (2003). Word-level intelligibility of time-compressed speech: prosodic and segmental factors. *Speech Comm*, *41*(2-3), 287–301.

Kohler, K. (1992). Gestual reorganization in connected speech: a functional viewpoint on articulatory phonology. *Phonetica, 49*, 205-211.

Kutas, M. & Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207(4427),* 203-205.

Lehongre, K., Ramus, F., Villiermet, N., Schwartz, D., Giraud, A.L. (2011). Altered Low-Gamma Sampling in Auditory Cortex Accounts for the Three Main Facets of Dyslexia. *Neuron, 72*, 1080–1090.

Macmillan, N.A., & Creelman, C.D. (1991). Detection theory: A user'guide. Cambridge, England: Cambridge University Press.

Maillart, C., Schelstraete, M.A. (2012). Les dysphasies : De l'évaluation à la rééducation. Issy-les Moulineaux : Elsevier Masson.

Max, L. and Caruso, A.J. (1997). Acoustic measures of temporal intervals across speaking rates: variability of syllable- and phrase-level relative timing. *J Speech Lang Hear Res, 40(5)*, 1097-1110.

Nijland, L. (2009). Speech perception in children with speech output disorders, *Clin Linguist Phon, 23*, 222-239.

Osnes, B., Hugdahl, K., Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *NeuroImage, 54*, 2437-2445.

Parisse C . et Maillart C . (2010). Nouvelles propositions pour la recherche et l ' évaluation du langage chez les enfants dysphasiques. In C. Gruaz & C. Jacquet-pfau (Eds.), *Autour du mot : pratiques et compétences.* (Lambert-Lu., pp. 201–222). Limoges : France.

Peelle, J.E., Davis, M.H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Front Psychol, 3*, 320.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as asymmetric sampling in time. *Speech Comm, 41,* 245-255.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature reviews. Neuroscience*, *11*(5), 351–60.

Soltesz, F., Szucs, D., Leong, N., White, S., Goswami, U. (2013). Differential Entrainment of Neuroelectric Delta Oscillations in developmental Dyslexia. *Plos One, 8(10),* 1-11.

Analyse prosodique des affects sociaux dans l'interaction face à face en japonais

Dominique Fourer[1], Marine Guerry[2],Takaaki Shochi[1,4]
Jean-Luc Rouas[1], Jean-Julien Aucouturier[3], Albert Rilliard[5]
(1)LaBRI - CNRS UMR5800, Univ. Bordeaux 1, France
(2) Univ. Bordeaux Montaigne, France ; (3) IRCAM - CNRS UMR9912 - UPMC, Paris, France
(4) CLLE-ERSSàB UMR5263 CNRS, Bordeaux, France ; (5) LIMSI - CNRS UPR3251, Orsay, France

RÉSUMÉ
Le but de cet article est de caractériser la prosodie attitudinale en langue japonaise. Ce travail s'appuie sur des travaux décrivant 16 attitudes correspondant à des situations de communication différentes. Ces situations peuvent, ou non être conventionnalisées dans la langue japonaise. Les paramètres estimés de fréquence fondamentale, d'amplitude et de durée ont été extraits d'énoncés exprimant ces 16 attitudes en japonais. Dans cette étude, nous présentons les effets sur ces paramètres des facteurs sexe du locuteur et expression attitudinale. Nous analysons également lesquels de ces paramètres prosodiques sont les plus discriminants pour caractériser acoustiquement chaque attitude.

ABSTRACT
Prosody analysis of Japanese social affects

The aim of this paper is to examine cues for prosodic characterization of attitudes in Japanese. This work is based on previous studies where 16 communicative social affects were defined. The audio signal parameters (fundamental frequency, amplitude and duration) of previously recorded Japanese attitudes, are statistically analyzed. Interesting interactions among the parameters, the gender and the expression of specific attitude (e.g. politeness) were found, and we report on which parameters most significantly characterize each attitude.

MOTS-CLÉS : prosodie, attitude, parole, japonais.

KEYWORDS: prosody, attitude, speech, Japanese.

1 Introduction

L'expression prosodique d'attitudes, telles qu'elles sont définies par (Wichmannn, 2000), est un moyen utilisé par les locuteurs pour effectuer certains actes de langage intentionnels (Fónagy *et al.*, 1984). Ces performances prosodiques sont partiellement liées à la maîtrise qu'a le locuteur de la langue, à sa personnalité, à son sexe, ainsi qu'au contexte communicationnel et linguistique - pour une discussion voir (Moraes et Rilliard, 2014). Chaque langue possède des formes et des variations prosodiques conventionnelles, dépendantes du contexte d'interaction. Les études

observant ces variations prosodiques reposent généralement sur des inventaires d'expressions conventionnalisées de la langue étudiée (Fujisaki et Hirose, 1993; Morlec *et al.*, 2001; de Moraes, 2008; Gu *et al.*, 2011). La plupart des études citées utilisent des corpus de laboratoire pour lesquels des locuteurs élicitent des variations prosodiques correspondant à un label donné. Ces labels sont interprétés par le locuteur dans son cadre culturel propre. De ce fait, la comparaison des performances entre deux langues (Shochi *et al.*, 2009; Rilliard *et al.*, 2009) se heurte aux critiques formulées par Wierzbicka sur certaines comparaisons interculturelles d'expressions émotionnelles (Wierzbicka, 1985, 1992, 2005). Les objets étudiés ne sont pas nécessairement comparables car actualisant des concepts différents. Si ce problème peut être contourné expérimentalement pour des tests de perception (Rilliard *et al.*, 2014b), les corpus enregistrés de cette manière ne permettent pas de contourner ce problème pour tout ce qui concerne les aspects de production. Dans ce but, le corpus PADE (Rilliard *et al.*, 2013, 2014a) définit des contextes d'interaction amenant les locuteurs à exprimer des variations prosodiques spécifiques à ces contextes. Les contextes restent identiques quelles que soient la culture et la langue étudiées, dans le but d'observer les variations prosodiques qui caractérisent les variations expressives correspondantes. Cet article s'attache à l'analyse acoustique des variations prosodiques produites par 19 locuteurs de langue maternelle japonaises ayant été enregistrés dans le cadre de cette étude. Une des finalités de cette analyse consiste à pouvoir synthétiser des variations prosodiques pertinentes pour exprimer ces variations attitudinales. Nous avons décidé d'utiliser uniquement les descripteurs prosodiques suivants : la fréquence fondamentale (F_0), l'amplitude *Root Mean Square* (RMS) et la durée des mores. En utilisant ces caractéristiques pour une analyse statistique basée sur une analyse de variance à mesures répétées (RM-ANOVA), nous sommes capables de déterminer comment chaque paramètre contribue à la différenciation des attitudes. Cet article est organisé comme suit : le cadre de travail utilisé pour l'expression des attitudes par les locuteurs est décrit dans la section 2. Le protocole d'enregistrement et les méthodes d'extraction des caractéristiques sont respectivement décrits dans les sections 3 et 4. Enfin, les résultats des analyses statistiques sont présentés dans la section 5 et discutés dans la section 6.

2 Contextes sociaux des expressions attitudinales

De manière similaire au protocole expérimental proposé par (Gu *et al.*, 2011), chaque sujet interagit avec un interlocuteur durant un dialogue aboutissant sur une phrase cible prononcée dans un contexte induisant une variation attitudinale. Afin d'immerger les sujets dans un contexte donné, un scénario a été préparé pour chaque attitude permettant d'aboutir sur la production des phrases cibles. 16 contextes correspondant à des expressions attitudinales conventionnalisées dans différentes langues (mais pas nécessairement dans chacune) ont été préparés : Admiration (ADMI), Arrogance (ARRO), Autorité (AUTH), Mépris (CONT), Doute (DOUB), Ironie (IRON), Irritation (IRRI), Déclaration (DECL), Question (QUES), Évidence (OBVI), Politesse (POLI), Séduction (SEDU), Sincérité (SINC), Surprise (SURP), Incertitude (UNCE), Marcher sur les œufs (WOEG). La définition de ces 16 attitudes, la description des concepts correspondants ainsi que la procédure d'enregistrement sont détaillées dans (Rilliard *et al.*, 2013). A titre d'exemple, voici un extrait de la situation créée pour l'attitude « Marcher sur des œufs » :

– Votre chef (locuteur *A*) vous (locuteur *B*) a chargé de préparer une salle pour une importante conférence. Votre chef est un homme nerveux qui ne supporte pas que les choses ne soient pas faites minutieusement et il se met facilement en colère si les choses ne sont pas à son goût.

Votre chef inspecte la salle où va commencer la conférence, lorsqu'il trouve dans une poubelle une vieille banane à moitié mangée. Il est furieux.

Il existe des contextes qui ne correspondent à aucune expression conventionnelle dans certaines langues étudiées (pour l'instant l'anglais des États-Unis, le japonais et le franÃğais hexagonal). C'est typiquement le cas de l'expression désignée par « marcher sur des œufs »qui correspond approximativement au concept japonais de *kyoshuku*, décrit par (Sadanobu, 2004) comme « correspondant à un mélange entre une honte douloureuse et de l'embarras venant du fait que le locuteur est conscient du fardeau qu'il impose à son interlocuteur de par sa requête »(p.34, traduction des auteurs). Dans ce travail, nous nous concentrons sur les résultats du japonais réalisés par des locuteurs natifs et sur l'une des phrases cibles utilisées pour l'enregistrement des attitudes : « Banana ».

3 Procédure d'enregistrement

Les performances audio-visuelles de 19 locuteurs natifs du japonais (11 femmes, 8 hommes) parlant le dialecte de Tokyo ont été enregistrées. La plupart des locuteurs sont des étudiants, qui ont été payés pour leur participation. Les enregistrements ont eu lieu dans une chambre sourde à l'Université Waseda au Japon. Le son a été enregistré à l'aide d'un microphone omnidirectionnel *Earthworks QTC1*, placé à un mètre de la bouche du locuteur (cette distance a été choisie pour limiter les perturbations d'intensité liées aux mouvements du locuteur). Le niveau du microphone a été calibré avant chaque session d'enregistrement à l'aide d'un calibreur acoustique *Bruel & Kjaer*. Ainsi, le niveau de pression acoustique peut être corrigé après enregistrement à un niveau comparable entre tous les locuteurs. L'énoncé de la phrase cible « Banana »produite par chaque locuteur pour chaque attitude (total de 304 stimuli) a été manuellement isolée en fichiers individuels. Ceux-ci ont été enregistrés au format WAV en utilisant une quantification sur 16 bits avec une fréquence d'échantillonnage de 16 kHz. Chaque stimulus a été découpé afin de retirer les silences de début et de fin phrase, puis manuellement segmenté au niveau des phonèmes à l'aide du logiciel PRAAT (Boersma et Weenink, 2001).

4 Estimation des paramètres prosodiques

Chaque stimulus est caractérisé par sa fréquence fondamentale (F_0), son amplitude instantanée ainsi que la durée de chaque phonème. La F_0 mesurée en Hertz est estimée pour chaque stimulus toutes les 10 ms par l'algorithme SWIPE (Camacho et Harris, 2008). Pour l'analyse statistique (*cf.* Section 5), seule la valeur de F_0 correspondant à la partie voisée de chaque phonème est considérée. Par la suite, nous calculons la moyenne des F_0 pour chaque voyelle (3 valeurs par stimulus). Enfin, nous appliquons l'algorithme de *MultiDimensional Scaling algorithm* (MDS) (Bronstein *et al.*, 2006) sur chaque vecteur de paramètre F_0 normalisé en divisant chaque élément par la F_0 moyenne du sujet considéré (*cf.* Section 5.4). Cette normalisation permet de comparer uniquement les modulations de F_0 autour de la fréquence moyenne, entre tous les sujets du corpus. Ainsi, l'algorithme MDS permet une interprétation géométrique de la distance entre les attitudes basée sur la corrélation calculée entre les vecteurs de F_0. Le paramètre d'amplitude est décrit par l'amplitude RMS calculée pour une fenêtre de signal de 20 ms utilisant un recouvrement de 50%. Pour ce calcul, la valeur de chaque échantillon de signal est un nombre réel normalisé sur

[−1;+1]. Pour l'analyse statistique, nous considérons uniquement les valeurs correspondant aux parties voisées. Nous calculons ensuite le RMS moyen associé à chaque voyelle (3 par stimulus). La durée de chaque syllabe est mesurée en millisecondes sur la base de la segmentation manuelle.

5 Résultats de l'analyse statistique

Les différences prosodiques sont mesurées en fonction des attitudes et du sexe des locuteurs, en considérant séparément la F_0, le RMS et la durée syllabique. Chaque stimulus est représenté par une série de 3 mesures de chaque paramètre. Pour chaque paramètre, nous réalisons un test RM-ANOVA, avec les facteurs attitude (16) × position de la more (3) et le facteur inter-sujet du sexe (2). Pour la suite de cette section, $F(a,b)$ correspond à la statistique calculée lors du test. Nous utilisons la notation M pour désigner la valeur moyenne du paramètre considéré.

5.1 La fréquence fondamentale

Le test RM-ANOVA appliqué sur la F_0 révèle une forte interaction attitude × position moraïque : $F(30,210)=8.12$, $p<0.05$. Cela signifie que le profil de F_0 sur la phrase change de manière significative en fonction de l'attitude. D'après la Figure 2(a), l'interaction porte principalement sur les deux premières mores. Ainsi, la plupart des attitudes sont associées à une F_0 plus basse sur la seconde voyelle (forme V) excepté pour les attitudes AUTH, DECL, QUES et (dans une moindre mesure) SINC et SEDU qui montrent des contours en forme de V inversé. Afin de mieux caractériser ce constat, nous avons calculé (en %) le rapport de la F_0 normalisée entre la première et la seconde voyelle pour chaque stimulus (cf. Figure 2(d)). Un test RM-ANOVA entre l'attitude (16) et le sexe (2) révèle une interaction significative ($F(15,225)=13.6$, $p<0.05$). Ainsi, AUTH(+26%), DECL (+17%), QUES (+34%) ont une pente de F_0 plus importante que pour les autres attitudes ($p<0.05$, après correction Bonferroni). SINC (+7%) montre également une augmentation de la F_0 plus importante que pour CONT (-14%) ($p<0.05$,après correction Bonferroni). De plus, la différence de F_0 entre les attitudes est en interaction avec le sexe du sujet ($F(30,210)=1.95$, $p<0.05$), ce qui indique qu'un profil de F_0 différent est utilisé par les hommes et par les femmes pour la même attitude. On constate également que cette interaction de F_0 n'est pas caractérisée par la première voyelle (attitude × sexe : $F(15,240)=1.36$, $p=0.16$), la seconde voyelle (attitude × sexe : $F(15,240)=1.83$, $p=0.03$) ou la troisième voyelle individuellement (attitude × sexe : $F(15,240)=0.38$, $p=0.98$), mais par les différences de F_0 entre les voyelles. Comme le montre la Figure 2(d), la différence entre les sexes est particulièrement importante en ce qui concerne le rapport entre la seconde et la première voyelle pour l'ensemble des attitudes (attitude × sexe : $F(15,225)=2.58$, $p<0.05$) mais aussi individuellement pour les attitudes ADMI (hommes : +5%, femmes : -11%, p=0.03), AUTH (hommes=+10%, femmes : +37%, $p<0.05$) et QUES (hommes=+49%, femmes=+23%, $p<0.05$; toutes : Fisher LSD posthocs).

5.2 L'amplitude

Le test RM-ANOVA sur les valeurs du RMS révèle un rôle important du sexe ($F(1,17)=24.5$, $p<0.05$) qui indique que les différences d'amplitude observées chez les locutrices sont plus

importantes que chez les locuteurs. On note aussi un effet important de la position moraïque ($F(2,34)=88.51$, $p<0.05$) : les amplitudes des premières ($M=.00071$) et des secondes ($M=.00066$) mores sont plus importantes que pour la troisième ($M=.00034$). L'attitude joue également un rôle non négligeable ($F(15,255)=1.99$, $p<0.05$) : certaines attitudes sont réalisées avec une amplitude plus importante que d'autres pour toute syllabe et pour les deux sexes. Les attitudes ADMI ($M=.00063$), POLI ($M=.00060$), SINC ($M=.00060$) et WOEG ($M=.00059$) ont des amplitudes plus fortes que AUTH ($M=.00054$), CONT ($M=.00052$), IRRI ($M=.00052$), QUES ($M=.00054$), SEDU ($M=.00053$), SURP ($M=.00056$) et UNCE ($M=.0052$) (test de Fisher LSD sans correction Bonferroni, $p<.05$). On note également une interaction de la position maraïque × sexe ($F(2,34)=3.51$, $p=0.04$) telle que la seconde voyelle est en général d'amplitude plus faible pour les hommes que pour les femmes. L'interaction attitude × sexe n'est pas significative ($F(15,255)=0.88$, $p=0.58$). Une interaction importante attitude × position de la more ($F(30,510)=6.57$, $p<.05$) montre que les variations intonatives changent de position dans la phrase, selon l'attitude (*cf.* Figure 2(b)). De plus, comme pour la F_0, les différences sont plus importantes lorsque nous observons le rapport RMS entre la seconde et la première more bien que la seconde more soit presque toujours d'amplitude plus faible que la première pour toutes les attitudes excepté AUTH, DECL and QUES. Contrairement à la F_0, cet effet n'entre pas en interaction avec le sexe ($F(15,255)=0.33$, $p=0.99$ - attitude × sexe en utilisant le test RM-ANOVA sur le rapport du RMS de la seconde sur première voyelle). Des formes similaires sont utilisées par les hommes et les femmes pour produire la même attitude (*cf.* Figure 2(e)).

5.3 La durée

Le test RM-ANOVA appliqué sur la durée des mores révèle un effet significatif du sexe ($F(1,17)=4.71$, $p<0.05$), pour lequel les femmes ($M=137$ ms) réalisent des mores plus longues que celle des hommes ($M=117$ ms). On note une influence de la position de la more sur sa durée ($F(2,34)=95.6$, $p<0.001$) : la dernière more ($M=188$ ms) est presque deux fois plus longue que la première ($M=99$ ms) ou la seconde more ($M=98$ ms). La relation de la durée avec l'attitude ($F(15,255)=13.0$, $p<.05$), montre que certaines attitudes sont produites avec un allongement plus important que pour d'autres, quels que soient la more ou le sexe. Les attitudes AUTH ($M=162$ ms), CONT ($M=166$ ms), DECL ($M=171$ ms), DOUB ($M=155$ ms), QUES ($M=147$ ms) et SURP ($M=165$ ms) montrent des durées plus longues que les autres ($p<.05$, avec correction Bonferroni). On note un effet significatif de l'attitude sur la durée de la première more ($F(15,255)=11.09$, $p<.05$), de la seconde more ($F(15,255)=9.48$, $p<.05$) et de la troisième more ($F(15,255)=13.08$, $p<.05$). On ne note aucune interaction position des mores × sexe ($F(2,34)=1.57$, $p=0.22$), ni attitude × sexe ($F(15,255)=1.17$, $p=0.29$). Cependant, comme mentionné précédemment, il y a une interaction significative attitude × position moraïque sur la durée ($F(30,510)=12.07$, $p<.05$) qui montre que pour différentes attitudes, les variations de durées ne sont pas identiques sur les mores successives (*cf.* Figure 2(c)). Comme pour la F_0 (mais pas pour le RMS), on observe une interaction entre position moraïque × sexe ($F(30,510)=2.43$, $p=<.05$). Cette différence entre les sexes s'observe principalement sur la dernière more ($F(15,255)=1.89$, $p<0.05$) (*cf.* Figure 2(f)). Pour les autres mores, cette interaction est moins importante (première more ($F(15,255)=1.68$, $p=0.05$; deuxième more : $F(15,255)=1.04$, $p=0.4$).

5.4 Algorithme MDS

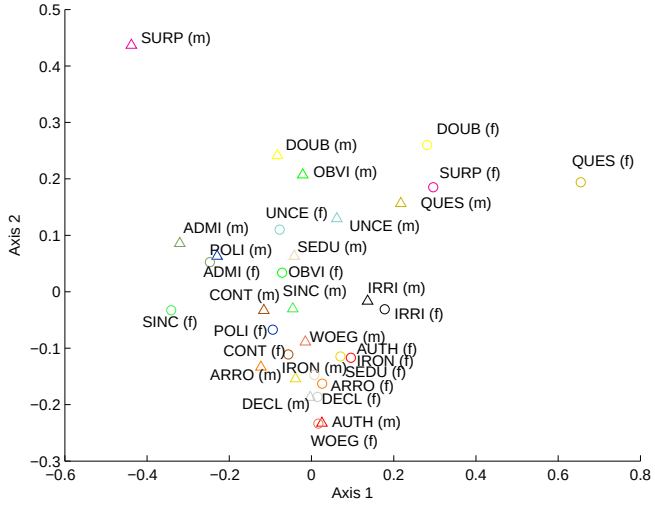
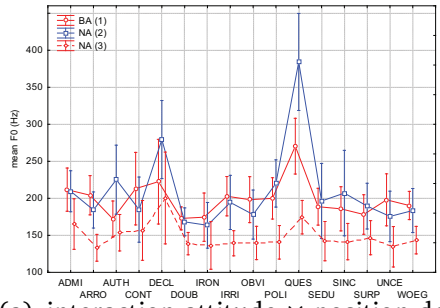


FIGURE 1 – Résultat de l'algorithme de MDS appliqué sur une fonction de corrélation entre les profils de F_0 (cercles : femmes, triangles : hommes).

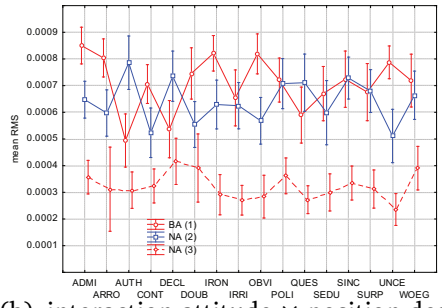
Le résultat de l'algorithme MDS (Bronstein *et al.*, 2006) appliqué sur les profils F_0 moyens de chaque attitude en dissociant hommes et femmes est présenté sur la Figure 1. La distance entre chaque point est associée à la fonction de corrélation (Hermes, 1998) calculée entre les profils F_0 normalisés de chaque sujet. Pour les calculs, les profils F_0 des parties voisées ont été normalisés en durée (par rééchantillonnage) afin d'obtenir 100 mesures de F_0 associées à la durée moyenne de la phrase sur tout le corpus (M=531.25 ms).

6 Discussion

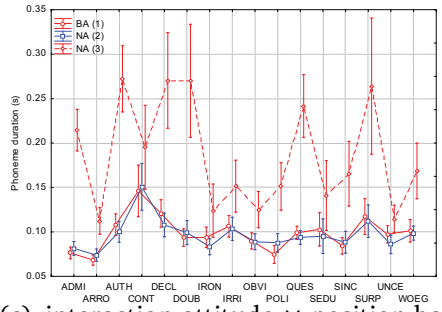
Cette étude sur la langue japonaise nous a permis d'explorer les caractéristiques acoustiques d'un ensemble d'attitudes non nécessairement spécifique de cette langue, dans le but d'identifier les similarités et les différences entre leurs profils prosodiques, basés sur les paramètres de F_0, de RMS et de durée syllabique. Les résultats de l'analyse statistique montrent des interactions entre les attitudes et le sexe pour la F_0. Cependant, nous avons vu que l'expressivité des femmes n'est pas seulement caractérisée par une F_0 (normalisée) plus élevé mais aussi par une amplitude plus élevée et des durées moraïques plus longues que celles des hommes. Cette observation est valable même si ces différences constituent plus un style lié au sexe que des variations de stratégie liées aux différentes attitudes. Nous avons également observé que les expressions de politesse (ADMI, POLI, SINC, WOEG) ont une amplitude plus élevée que les autres, y compris les expressions d'impolitesse (AUTH, CONT, IRRI). Cette observation diffère des résultats observés pour les mêmes situations chez des locuteurs d'anglais des États-Unis (Rilliard *et al.*, 2013), pour lesquels les attitudes véhiculant un trait sémantique d'imposition (AUTH, ARRO, CONT), sont réalisées avec une intensité plus élevée que les attitudes politesse (POLI, SINC, WOEG). Ce constat ne vaut cependant pas pour l'attitude d'ADMI pour laquelle les deux groupes linguistiques montrent des stratégies comparables. Les résultats montrent d'importantes variations de durée moraïque parmi les attitudes, principalement entre les deux premières voyelles et la troisième



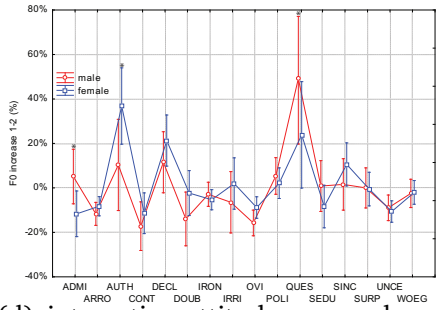
(a) interaction attitude × position de syllabe basé sur les valeurs de F_0



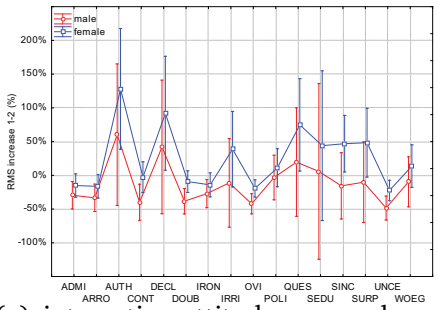
(b) interaction attitude × position des syllabes basée sur les vaeurs de RMS



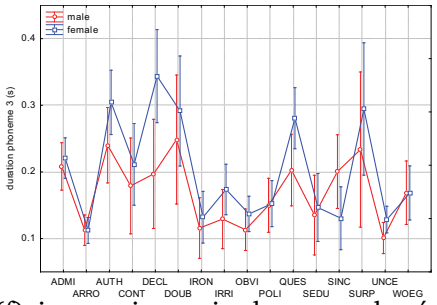
(c) interaction attitude × position basée sur la durée de chaque syllabe



(d) interaction attitude × sexe du sujet basé sur les valeurs de F_0



(e) interaction attitude × sexe du sujet basée sur les valeurs de RMS



(f) interaction attitude × sexe basée sur la durée de la troisième syllable

FIGURE 2 – 2(a) et 2(d) : Profils de F_0 en fonction de l'attitude et du sexe du sujet. 2(a) : F_0 moyenne (en Hz) pour chaque voyelle de « banana », pour les 16 attitudes. 2(d) : différences de la F_0 moyenne entre la première et la seconde voyelle (en % de croissance). 2(b) et 2(e) : Profils RMS entre les attitudes et le sexe. 2(b) : RMS moyen pour les 3 voyelles de « banana »pour les 16 attitudes. 2(e) : différence du RMS moyen entre la seconde et la première voyelle (en % de croissance). 2(c) : durée de chacune des 3 syllabes de « banana »pour les 16 attitudes. 2(f) : durée de la troisième more pour les 16 attitudes.

voyelle (qui sont approximativement deux fois plus longue que les deux premières). Cela permet de déduire que les affects sociaux peuvent être à l'origine d'ajustement de durée basé sur la structure rythmique du japonais. Cela est en accord avec une étude antérieure de (Maekawa et Kagomiya, 2000) qui montre que la durée de la troisième voyelle est un facteur important permettant de caractériser l'expression de certaines attitudes. Le japonais standard est décrit comme une langue à accents toniques caractérisée par la montée de la F_0 au début de chaque phrase et marquée par une chute importante de la F_0 après une voyelle accentuée (Fujisaki et Sudo, 1971). La phrase « Banana »possède en japonais un accent lexical localisé sur la première more. Ainsi, nous nous attendons à observer une montée de la F_0 sur la première voyelle suivie d'une chute. Cependant, les résultats montrent que les contours de F_0 varient en fonction de l'attitude. Notre hypothèse est que l'instant où le pic de F_0 est atteint pour la première ou la seconde voyelle est lié au débit de parole ou à l'intensité. D'après nos données, il semble que certaines attitudes (AUTH, QUES et DECL) pour lesquelles le pic de F_0 arrive tardivement (sur la deuxième voyelle) ont des débits de parole plus lents. Cette position temporelle concorde avec un pic d'amplitude sur la deuxième voyelle, contrairement à la plupart des attitudes qui réalisent le pic ainsi que la chute de F_0 à l'intérieur de la première more (forme de V inversé).

L'analyse MDS basée sur la corrélation des profils de F_0 des attitudes nous permet d'identifier 3 catégories : les expressions de déférence, d'imposition et les expressions dubitatives (Brandt, 2008; Rilliard *et al.*, 2009). La première catégorie des expressions de déférence contient les

expressions ADMI, SINC, POLI (et SEDU uniquement pour les hommes) ainsi que deux attitudes (OBVI, UNCE) qui sont plus complexes à interpréter. La seconde catégorie d'imposition contient les attitudes AUTH, ARRO, IRON, IRRI ainsi que 2 autres (DECL, WOEG) qui ne correspondent pas sémantiquement aux autres expressions du groupe. Il est intéressant de noter que WOEG, qui correspond au concept de politesse de *Kyoshuku* se trouve dans la région des expressions d'imposition. On notera, d'après (Sadanobu, 2004; Rilliard *et al.*, 2014b,a), que cette expression de *kyoshuku* véhicule un trait sémantique d'imposition, certainement véhiculé par l'expression prosodique. Cela explique aussi les résultats d'études antérieures sur cette expression (Shochi *et al.*, 2006; Rilliard *et al.*, 2009), pour laquelle des incompréhensions sont observées chez des auditeurs ne parlant pas japonais. Il faut aussi noter que l'expression de WOEG se distingue de l'impolitesse par ses caractéristiques de qualité vocale spécifiques(Rilliard *et al.*, 2009). Malgré des profils de F_0 proches entre hommes et femmes pour l'attitude SEDU, d'après l'analyse RM-ANOVA, (*cf.* Figure 2(d)), l'analyse de MDS dissocie les performances de ces deux groupes (Figure 1) : les expressions des femmes se trouvent prosodiquement plus proches de la catégorie d'imposition, alors que les performances masculines sont groupées avec les expressions de déférences. On pourra apprécier les potentielles interprétations de ce résultat, que nous n'aurons pas la place de développer. Nous noterons tout de même que la qualité de voix devrait ici aussi jouer un rôle majeur (présence d'un souffle important chez les femmes). Dans nos prochains travaux, nous effectuerons une étude contrastive similaire entre différentes langues.

Remerciements

Merci à Donna Erickson pour ses conseils avisés ainsi qu'à Mariko Kondo et Sylvain Detey pour l'enregistrement du corpus à l'université de Waseda. Cette recherche est financée par le projet ANR PADE et le projet bordelais PEPS IDEX/CNRS Mavoix.

Références

Boersma, P. et Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Brandt, P. A. (2008). Thinking and language. a view from cognitive semio-linguistics. *In Proc. Speech Prosody*, pages 649–654.

Bronstein, A. M., Bronstein, M. M. et Kimmel, R. (2006). Generalized multidimensional scaling : a framework for isometry-invariant partial surface matching. *Proc. Natl. Acad. Sci. U.S.A.*, 103(5):1168–1172.

Camacho, A. et Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *Journal of the Acoustical Society of America*, 124:1638–1652.

de Moraes, J. A. (2008). The pitch accents in Brazilian Portuguese : analysis by synthesis. *In Proc. Speech Prosody*, pages 389–397.

Fónagy, I., Bérard, E. et Fonagy, J. (1984). Clichés mélodiques. *Folia Linguistica*, 17:153–185.

Fujisaki, H. et Hirose, K. (1993). Analysis and perception of intonation expressing paralinguistic information in spoken Japanese. *In Proc. ESCA Workshop on Prosody*.

Fujisaki, H. et Sudo, H. (1971). A model for the generation of fundamental frequency contours of japanese word accent. *Journal of acoustical society of Japan*, 27(9):445–453.

Gu, W., Zhang, T. et Fujisaki, H. (2011). Prosodic analysis and perception of Mandarin utterances conveying attitudes. *In Proc. Interspeech*, pages 1069–1072.

Hermes, D. J. (1998). Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language and Hearing Research*, 41(1):73.

Maekawa, K. et Kagomiya, T. (2000). Influence of paralinguistic information on segmental articulation. *In Proc. 6th International Conference on Spoken Language Processing*, pages 349–352.

Moraes, J. A. et Rilliard, A. (2014). *Spoken Corpora and Linguistic Studies*, chapitre Illocution, Attitudes and Prosody. John Benjamins, Amsterdam. (à paraître).

Morlec, Y., Bailly, G. et Aubergé, V. (2001). Generating prosodic attitudes in French : Data, model and evaluation. *Speech Communication*, 33(4):357–371.

Rilliard, A., de Moraes, J., Erickson, D. et Shochi, T. (2014a). Social affect production and perception across languages and cultures âĂŞ the role of prosody. *Leitura*, 52. (à paraître).

Rilliard, A., Erickson, D., de Moraes, J. et Shochi, T. (2014b). *Linguistic approaches to emotions in context*, chapitre Cross-Cultural Perception of some Japanese Expressions of Politeness and Impoliteness, pages 251–276. John Benjamins, Amsterdam.

Rilliard, A., Erickson, D., Shochi, T. et Moraes, J. A. D. (2013). Social face to face communication - American English attitudinal prosody. *In Proc. Interspeech*.

Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D. et Aubergé, V. (2009). Multimodal indices to Japanese and French prosodically expressed social affects. *Language and speech*, 52(2–3):223–243.

Sadanobu, T. (2004). A natural history of Japanese pressed voice. *Journal of the Phonetic Society of Japan*, 8(1):29–44.

Shochi, T., Aubergé, V. et Rilliard, A. (2006). How prosodic attitudes can be false friends : Japanese vs. French social affects. *In Proc. Speech Prosody*, pages 692–696.

Shochi, T., Rilliard, A., Aubergé, V. et Erickson, D. (2009). *The role of prosody in affective speech*, volume Linguistic Insights 97, chapitre Intercultural perception of English, French and Japanese social affective prosody, pages 31–59. Peter Lang.

Wichmannn, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. *In Proc. ISCA Workshop on Speech and Emotion*, pages 143–148.

Wierzbicka, A. (1985). A semantic metalanguage for a cross-cultural comparison of speech acts and speech genres. *Language in Society*, 14(4):491–513.

Wierzbicka, A. (1992). Defining emotion concepts. *Cognitive Science*, 16:539–581.

Wierzbicka, A. (2005). Empirical universals of language as a basis for the study of other human universals and as a tool for exploring cross-cultural differences. *Ethos*, 33(2):256–291.

Labial-Coronal *vs.* Labial-Vélaire : Étude du phasage des gestes en Français

Manon Carrissimo-Bertola[1], Nathalie Vallée[1], Ioana Chitoran[2]
(1) GIPSA-Lab DPC, UMR 5216, CNRS-Université Grenoble-Alpes, BP25 38040 Grenoble Cedex 9
(2) Université Paris Diderot - Paris 7, CLILLAC-ARP case 7003, 5 rue T. Mann 75250 Paris Cedex 13
manon.carrissimo-bertola@gipsa-lab.grenoble-inp.fr,
nathalie.vallee@gipsa-lab.grenoble-inp.fr,
Ioana.chitoran@linguist.univ-paris-diderot.fr

RESUME

Différentes études ont mis en avant la surreprésentation de séquences Labial-Coronal par rapport à Coronal-Labial dans les productions d'enfants au stade des premiers mots ainsi que dans les unités lexicales de plusieurs langues du monde, alors que les structures avec Labiale et Vélaire ne montrent pas de tendance à favoriser les séquences avec une première consonne Labiale. L'effet LC a été l'objet de recherches récentes en phonétique expérimentale, à l'aide notamment de protocoles EMA, pour permettre de mieux comprendre les origines de ce phénomène. Notre travail consiste à observer la stabilité articulatoire de séquences CVCV de patrons Labial-Coronal et Coronal-Labial ainsi que Labial-Vélaire et Vélaire-Labial à partir d'une tâche de répétition, sous la contrainte d'un paradigme de vitesse. Les déplacements des articulateurs ont été enregistrés à l'aide d'un EMA 2D. Nous présentons les premiers résultats de ce travail.

ABSTRACT

Labial-Coronal Effect and Labial-Velar Sequences: An articulatory study in French
Studies have shown a Labial-Coronal order preference over Coronal-Labial both in children's production at the first-words stage and in the lexicons of various world languages while a similar Labial-Velar order preference has not been observed. The Labial-Coronal effect (LC effect) has recently been studied in EMA experiments, in order to better understand the origin of this effect. In our study, we compare in a production task the stability of: (i) CVCV sequences in Labial-Coronal order *vs.* Coronal-Labial order and, (ii) CVCV sequences with Labial-Velar *vs.* Velar-Labial order. Participants performed a repetition task at an increasing speech rate. We recorded trajectories and timing of articulatory movements with the EMA 2D articulograph. In this paper are presented our preliminary results.

MOTS-CLES : Effet Labial-Coronal, Séquence Labiale Vélaire, français, phasage articulatoire
KEYWORDS: Labial Coronal Effect, Labial-Velar Sequence, French, articulatory phasing

1 Introduction

Plusieurs études portant sur les tendances générales dans l'organisation interne des syllabes de langues d'origines géographique et génétique différentes ont montré la prédominance de certaines combinaisons de segments consonantiques et parmi elles, une tendance forte à favoriser un ordre pour lequel une consonne labiale précède une consonne coronale par rapport à l'ordre inverse (MacNeilage, Davis, Kinney & Matyear, 1999 ; Rousset, 2004). MacNeilage et Davis

(2000) nomment cette tendance *LC Effect* (Effet Labial-Coronal).

Parallèlement à l'effet LC, la recherche d'autres cooccurrences de consonnes et donc l'existence d'autres effets de consonnes dans les unités lexicales a suscité notre intérêt, notamment celle de l'effet Labial-Vélaire. Or, aucune tendance impliquant une consonne labiale produite en amont de la production d'une consonne vélaire n'a pu être mise en avant ou généralisée dans une vingtaine de langues (Carrissimo-Bertola, 2010). Etudier dans une démarche expérimentale les mécanismes articulatoires de ces successions de consonnes peut permettre de mieux comprendre pourquoi un ordre de consonnes est plus fréquemment rencontré qu'un autre dans les langues.

1.1 L'effet Labial-Coronal

Dans les productions de jeunes enfants, au stade des premiers mots, MacNeilage et al. (1999) observent une tendance à organiser les sons consonantiques entre deux syllabes consécutives en fonction de leur lieu d'articulation. Les séquences avec consonne labiale devant consonne coronale sont nettement favorisées par rapport à celles où une labiale succède à une coronale. La même tendance a été observée dans le lexique d'une dizaine de langues pour des unités lexicales dissyllabiques. L'Effet LC a été expliqué dans un premier temps dans le cadre de la théorie *Frame, then Content* (MacNeilage, 1998), avec l'argument du plus simple d'abord : le degré de complexité articulatoire d'une labiale est moindre que celui d'une coronale car il ne nécessite pas de geste lingual superposé à la remontée de la mandibule, et initier une séquence avec une labiale serait moins coûteux d'un point de vue articulatoire qu'avec une coronale.

Les résultats de Rousset (2004), confirmés par Vallée, Rossato & Rousset (2009), complètent ces travaux puisqu'ils démontrent un effet LC fort dans les lexiques syllabés de la base G-ULSID (*Grenoble and UCLA Lexical and Syllabic Inventory Database*) également étendu aux séquences inter-syllabes d'unités lexicales plus longues mais aussi aux séquences intrasyllabes de structure CVC. Carrissimo-Bertola (2010) constate le même effet LC dans le même échantillon encore enrichi (cf. TABLE 1) ainsi qu'en italien (Carrissimo-Bertola, 2011). Dans cet échantillon, seuls le wa et le vietnamien ne présentent pas d'effet LC.

Études	Nb langues	CVC	CV.CV	Dissyllabes
MacNeilage & al. (1999)	10	*	*	2,23
Rousset (2004)	10 à 13	1,44	1,73	2,39
Vallée & al. (2009)	17	1,89	1,68	2,79
Carrissimo-Bertola (2010)	19	6,59	1,70	2,56

TABLE 1 - Valeurs moyennes des ratios LC/CL calculés dans le cadre des différentes études citées ci-dessus (* signifie l'absence de données).

Au-delà du cadre de la théorie *Frame, then Content*, de récentes études ont permis de mettre en avant l'implication de contraintes liées aux systèmes de production et de perception de la parole dans l'effet LC (Lancia & Fuchs, 2011; Nazzi, Bertoncini, & Bijeljac-Babic, 2009 ; Rochet-Capellan & Schwartz, 2007 ; Sato, Vallée, Schwartz, & Rousset, 2007 ; Tsuji, Gomez, Medina, Nazzi, & Mazuka, 2012).

Le travail de Rochet-Capellan & Schwartz (2007) a consisté à tester la stabilité articulatoire des patrons LC *vs.* CL afin de déterminer si des contraintes motrices pourraient être un facteur explicatif de la surreprésentation du patron LC. Leur protocole expérimental se base sur le principe selon lequel, pendant une tâche de coordination de gestes, l'accélération du rythme contraint le sujet à adapter ses gestes et à adopter un mode de coordination plus simple et donc

plus stable que ce soit pour des séries de gestes manuels (Haken, Kelso & Bunz, 1985) ou de parole (Kelso, Saltzman, & Tuller, 1986). Les travaux de Tuller & Kelso (1991) ont montré que dans une tâche de répétition accélérée, des séquences VC tendent à se réorganiser en CV, patron syllabique supposé articulatoirement plus stable. Rochet-Capellan & Schwartz (2007) ont étudié les réorganisations articulatoires mises en place lors d'une tâche de répétition de pseudo-mots de type /pata/ et /tapa/, produite par des locuteurs de langue maternelle française. Leur étude acoustique et articulatoire montre une stabilité plus forte pour les dissyllabes /pata/, donc pour le patron LC. L'analyse du phasage des gestes mandibule, lèvres et apex leur a permis de proposer une explication basée sur la possibilité d'anticiper la production de la consonne coronale durant l'articulation labiale. L'analyse du phasage entre mandibule et autres articulateurs met en avant l'évolution d'un cycle oscillatoire mandibulaire par syllabe vers une unique oscillation pour un dissyllabe lorsque le rythme de production devient particulièrement soutenu. Une étude similaire auprès de locuteurs allemands sur le phasage/antiphasage des gestes dans les patrons LC *vs.* CL a démontré que les deux gestes, labial et lingual, demeuraient en antiphase sans réelle anticipation de geste (Lancia & Fuchs, 2011) contrairement aux résultats obtenus par Rochet-Capellan & Schwartz (2007). Par ailleurs, Tsuji, Gomez, Medina, Nazzi, & Mazuka (2012) ont constaté également que les japonais tendent à adopter et stabiliser le patron LC plutôt que CL, lors d'une tâche de production similaire. Leur explication est basée sur la plus grande facilité articulatoire des séquences LC par rapport à CL.

1.2 Les séquences Labial-Vélaire

MacNeilage et al. (2000) incluent l'effet LC dans un effet de consonnes plus large décrit dans les productions des enfants au stade des premiers mots par Ingram (1974) et Macken (1977) : *the Fronting effect*, qui réfère à l'observation de séquences favorisées pour lesquelles la première consonne est plus antérieure que celle qui suit. De même que pour l'effet LC, d'autres tendances sur une organisation des consonnes en séquences dépendante de leur lieu articulatoire ont été recherchées dans les langues de G-ULSID, afin de vérifier si cet effet de *fronting* dépasse le cadre de l'acquisition au stade des premiers mots. Les résultats de l'étude de Carrissimo-Bertola (2010) ne montrent pas de tendance LV ou VL généralisable à l'ensemble des 19 langues examinées (cf. TABLE 2). Le ratio moyen du nombre de séquences LV sur le nombre de séquences VL est de 1,55 (SD=2,82) pour les séquences CVCV, mais seules 20 % des langues de l'échantillon présentent un ratio supérieur à 1. Ce ratio est de 0,89 (SD=0,35) si le navaho, présentant une tendance LV très forte, est écarté des données. Les observations de Carrissimo-Bertola (2010) ne confortent donc pas le *Fronting effect* dans les unités lexicales des langues pour les traits labial et vélaire.

	CVC	**CVCV**
Ratio moyen LV/VL	1,48	1,55
Tendance LV	9	4
Tendance VL	6	7
Sans tendance	*	7
*	4	1

TABLE 2 : Ratio moyen LV/VL pour les 19 langues de la base G-ULSID et nombre de langues concernées par la tendance ; * signifie absence de données (d'après Carrissimo-Bertola, 2010).

1.3 Hypothèses pour Labial-Vélaire

Notre protocole expérimental est adapté de celui de Rochet-Capellan & Schwartz (2007), repris

par Lancia & Fuchs (2011). La conclusion formulée sur le lien entre stabilité en production et anticipation des gestes nous incitent à observer non seulement des cibles de type /pata/ et /tapa/ mais aussi des cibles /paka/ et /kapa/ pour tester l'hypothèse suivante : si l'anticipation du geste lingual lors de la production de l'articulation labiale est la clef de la prédominance des séquences LC sur CL, alors cette réponse est aussi valable pour LV par rapport à VL même si un effet LV n'est pas tendanciel dans les unités lexicales des langues. En effet, une consonne extra-buccale libérant l'activation de la langue, l'anticipation d'un geste coronal dans un geste labial est possible et devrait permettre une anticipation similaire entre une consonne vélaire et une consonne labiale.

2 Méthodologie

2.1 Sujets et stimuli

Cinq sujets de langue maternelle française et nés en France ont participé à cette expérience. Les locuteurs étaient tous de sexe féminin, âgées de 23 à 27 ans (âge moyen : 24,8). Dans cet article sont présentés les résultats préliminaires obtenus avec deux des locutrices.

Le matériel présenté aux locutrices se compose de six entrées dissyllabiques C_1VC_2V : /pata/ *vs.* /tapa/, /paka/ *vs.* /kapa/ et /kata/ *vs.* /taka/. Le choix des phonèmes s'est porté sur trois consonnes plosives sourdes qui facilitent le découpage des segments en contexte vocalique. La voyelle [a] a été retenue pour sa configuration ouverte du conduit vocal et sans geste labial. De plus, ses propriétés articulatoires, et notamment son degré d'aperture, marquent un fort contraste avec la fermeture nécessaire à la réalisation des consonnes plosives.

Les cibles étaient présentées dans un ordre aléatoire en série de 6 éléments générée automatiquement trois fois pour chaque locutrice. La consigne donnée était de répéter un pseudo-mot inscrit sur un écran face à elle, tout en essayant de respecter le plus possible un rythme imposé par un métronome visuel. Une série test permettait de vérifier la maitrise des consignes et d'entrainer les locutrices à la tâche avant d'être équipées de l'articulographe.

2.2 Protocole

Les enregistrements se sont déroulés dans la chambre anéchoïde du Gipsa-Lab, Grenoble. Les données ont été recueillies grâce à un système EMA 2D, AG-200 de Carstens, à une fréquence d'échantillonnage de 200 Hz. Les locuteurs étaient équipés de 6 bobines collées sur les articulateurs lèvre inférieure, lèvre supérieure, mâchoire, apex, corps et dos de la langue ainsi que 2 bobines de référence pour déterminer le plan médio-sagittal de la locutrice. Les données acoustiques ont été recueillies avec un micro AKG C10005 et un enregistreur numérique PMD 670 (fréquence d'échantillonnage 22 khz).

Les locutrices équipées de l'articulographe étaient assises face à un écran 17 pouces AG NEOVO X17-A où s'affichait les cibles à répéter avec en dessous, sous la forme d'un flash clignotant, le métronome visuel. Ce métronome imposait un rythme de production en trois phases. Une première phase d'accélération enregistrait une évolution d'un flash toutes les 600 ms à un flash pour 100 ms. Une phase de stabilité d'une durée de 500 ms maintenait un rythme d'un flash par 100 ms. Enfin une phase de décélération dictait un rythme décroissant selon une pente symétrique à la phase d'accélération. La durée totale de la tâche pour une cible était de 18,5 s. Le

rôle du métronome était d'inciter les locutrices à répéter le mot le plus rapidement possible.

2.3 Mesures et analyses

Les trajectoires oscillantes des articulateurs apex, mâchoire et lèvres ont été segmentées et étiquetées à l'aide des outils d'annotation EMA développés par M. Tiede (Haskins-lab) dans l'environnement Matlab®. Pour chaque cycle d'oscillation, les événements minimum et maximum des courbes de déplacement ont été repérés automatiquement à partir des passages par zéro de la courbe de vitesse de chacun des articulateurs. Les cibles /t/ et /k/ sont considérées comme atteintes dès lors que respectivement apex et dos de la langue atteignent leur position maximale et la cible /p/ est atteinte lorsque la lèvre inférieure est au point le plus haut du cycle oscillatoire. Pour analyser la coordination articulatoire entre gestes consonantiques d'un dissyllabe répété nous avons choisi d'examiner le timing des gestes lèvres/apex pour /pt/ et /tp/, lèvres et dos de la langue pour /pk/ et /kp/. Les articulateurs lèvre inférieure et langue sont aussi observés par rapport à leur coordination avec le cycle oscillatoire de la mandibule. Les transitions de /p/ à /t/ ou /p/ à /k/ sont estimées à partir de la mesure de l'espace temporel qui sépare les maxima des articulateurs considérés :

- Δt_1 correspond à l'intervalle temporel depuis la position la plus élevée de la lèvre inférieure (Max_{LL}) à la position maximale de l'apex (Max_{TT}). Pendant Δt_1, la lèvre inférieure s'abaisse et l'apex s'élève.
- Δt_2 est l'intervalle temporel depuis la position maximale de l'apex (Max_{TT}) jusqu'à l'atteinte de la position maximale de la lèvre inférieure (Max_{LL}). Δt_2 correspond à la phase de relâchement de l'occlusion coronale et d'élévation de la lèvre inférieure.
- Δt_3 correspond au laps de temps écoulé depuis la position maximale de la lèvre inférieure (Max_{LL}) à la position maximale du dos de la langue (Max_{TD}). Dans le délai Δt_3, la lèvre inférieure redescend et le dos de la langue remonte.
- Δt_4 est le délai temporel depuis la position la plus élevée du dos de la langue (Max_{TD}) à celle de la lèvre inférieure (Max_{LL}). Δt_4 correspond à la phase de relâchement de l'occlusion vélaire et d'élévation de la lèvre inférieure.

$$\Delta t_1 = tmax_{TT} - tmax_{LL} \text{ , } \Delta t_2 = tmax_{LL} - tmax_{TT}$$

$$\Delta t_3 = tmax_{TD} - tmax_{LL} \text{ , } \Delta t_4 = tmax_{LL} - tmax_{TD} \qquad \text{avec } \Delta t_i \geq 0$$

Nous formulons l'hypothèse que les effets de consonnes sont dus à une coordination articulatoire plus stable que nous proposons d'estimer à partir du timing des gestes, en nous inspirant des travaux expérimentaux en phonologie articulatoire (Browman & Goldstein, 1992). Les durées de /p/ à /t/ (Δt_1) et /p/ à /k/ (Δt_3) devraient être plus courtes respectivement que de /t/ à /p/ (Δt_2) et de /k/ à /p/ (Δt_4), ce qui marquerait une coarticulation plus forte dans un sens Labial-Coronal ou Vélaire (LC ou LV) plutôt que Coronal ouVélaire -labial (CL ou VL).

3 Résultats

3.1 Evolution du cycle oscillatoire de la mandibule

Le nombre total de cycles enregistrés pour les articulateurs a été calculé pour chaque locutrice en fonction de chaque dissyllabe (FIGURE 1). Bien que la locutrice CG produise moins de cycles que la locutrice ASB quel que soit le dissyllabe, les deux locutrices réalisent un nombre total de

cycles plus petit pour les cibles qui impliquent le dos de la langue que pour celles qui sollicitent l'apex. Au pic de vitesse, les locutrices adoptent les deux types de comportements suivants: soit la réduction de 2 cycles oscillatoires à un seul par dissyllabe, soit le blocage de la mâchoire et l'arrêt de son oscillation. Ces deux comportements expliquent le nombre réduit de cycles de la mâchoire par rapport à celui des autres articulateurs.

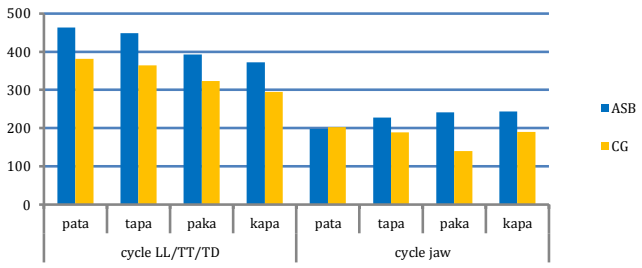


FIGURE 1 : Nombre total de cycles réalisés par les articulateurs lèvre (LL), apex (TT), dos de la langue (TD) et mâchoire (Jaw) en fonction de la condition mot initiale.

Pour chaque locutrice, les résultats portés sur les figures suivantes 2 à 5 prennent en compte l'ensemble des répétitions pour chaque pattern.

3.2 Cibles /pata/ vs. /tapa/

3.2.1 /pata/

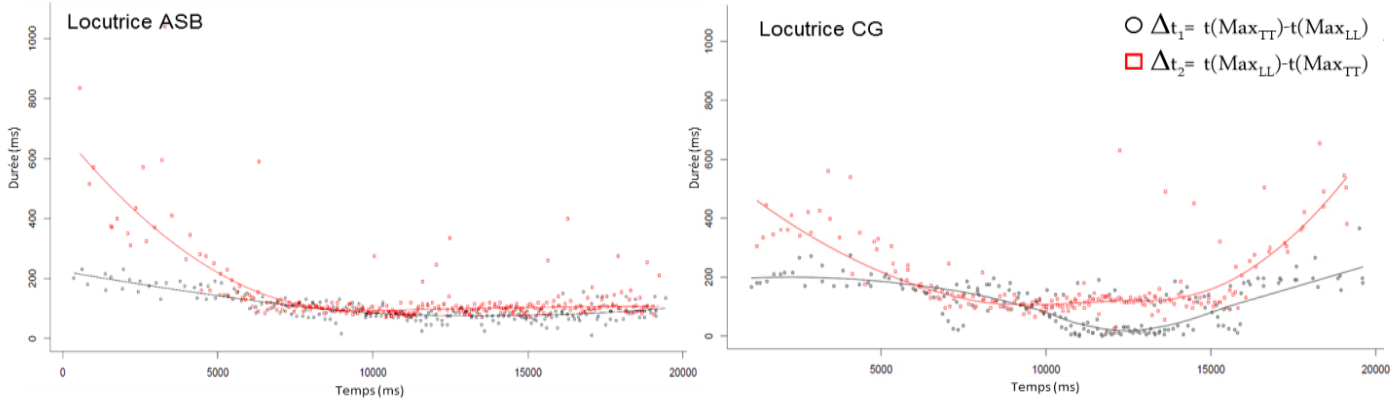


FIGURE 2 : Evolutions temporelles des Δt_1 et Δt_2 pour la cible /pata/ chez les deux locutrices.

A vitesse lente, Δt_2 est plus long que Δt_1, phénomène relié au patron accentuel français iambique (Δt_2 comprend la frontière entre disyllabes). Les résultats pour les deux locutrices (FIGURE2) montrent que pendant l'accélération, Δt_2 diminue et tend vers Δt_1. Cependant, même au pic de l'accélération (à environ 11 s), la courbe de tendance des valeurs de Δt_1 reste sous celle de Δt_2. La cible de l'apex est plus rapidement atteinte après Max_{LL} que la remontée labiale derrière Max_{TT}. En d'autres termes, la succession des gestes labial puis coronal est toujours plus rapide que l'enchainement du geste labial derrière le geste coronal, révélant la stabilité du patron LC.

3.2.2 /tapa/

Dans le cas des cibles /tapa/, à vitesse lente et en raison du patron accentuel iambique, Δt_1 est plus long que Δt_2 (cf. FIGURE 3). Sous la contrainte de vitesse, ce rapport s'inverse. Les délais Δt_1 mesurés depuis Max_{LL} à Max_{TT} se réduisent, et au pic de l'accélération, la courbe de tendance Δt_1 passe sous celle des valeurs de Δt_2. Cette évolution des durées entre les positions maximales de la lèvre et de l'apex reflète une cohésion articulatoire LC plus forte et un basculement de CL vers

LC chez les deux locutrices.

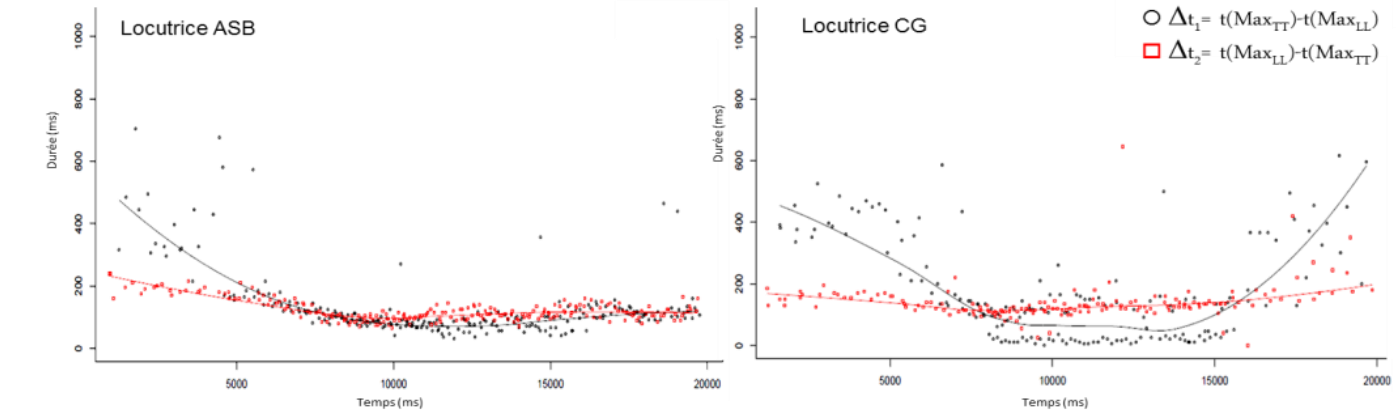


FIGURE 3: Evolutions temporelles des Δt_1 et Δt_2 pour la cible /tapa/ chez les deux locutrices.

3.3 Cibles /paka/ vs. /kapa/

3.3.1 /paka/

A vitesse normale, Δt_3 est plus court que Δt_4. L'accélération provoque une réduction de la durée inter-dissyllabe Δt_4, mais la courbe de tendance des valeurs de Δt_4 se maintient au dessus de celle de Δt_3 chez les deux locutrices, indice d'un intervalle temporel moins long entre le relâchement labial et l'élévation du dos de la langue que celui nécessaire pour la transition inverse. Cette tendance est plus marquée pour la locutrice CG (cf. FIGURE 4). Pour cette cible, la cohésion LV est suffisamment forte pour résister à la vitesse.

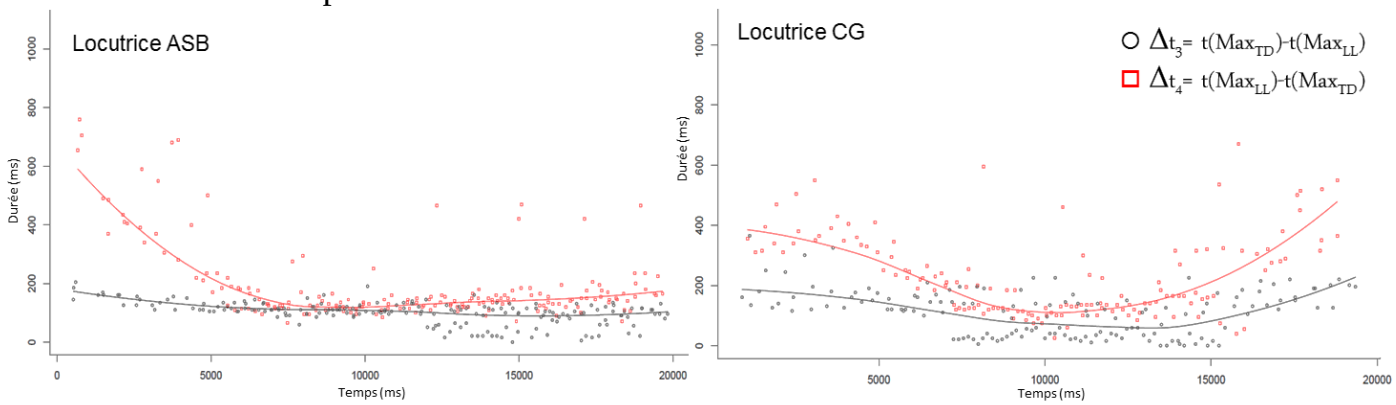


FIGURE 4 : Evolutions temporelles des Δt_3 et Δt_4 pour la cible /paka/ chez les deux locutrices.

3.3.2 /kapa/

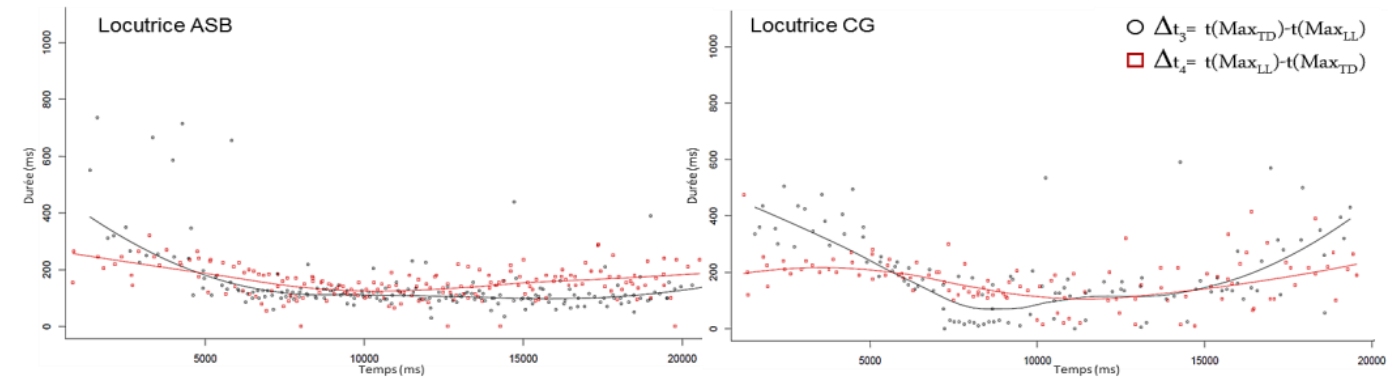


FIGURE 5 : Evolutions temporelles des Δt_3 et Δt_4 pour la cible /kapa/ chez les deux locutrices.

Pour cette cible (FIGURE 5), à débit de parole normal, l'écart temporel Δt_3 est plus long que Δt_4, en raison du patron iambique. Sous l'effet de la vitesse, les valeurs de Δt_3 convergent vers les

valeurs de Δt_4. Pour la locutrice ASB, la courbe de tendance Δt_3 passe sous celle de Δt_4, marqueur d'une plus grande anticipation de la remontée de la langue que pour le geste labial. Chez cette locutrice, la cohésion articulatoire LV est plus forte que la cohésion VL : le patron VL bascule vers LV à débit rapide. Pour les résultats obtenus pour la locutrice CG, les intervalles Δt_3 et Δt_4 tendent plutôt à s'égaler. Les tendances chez cette participante sont moins nettes, conséquence possible d'un échantillon de données plus restreint (moins de cycles produits).

4 Discussion et conclusion

Les résultats obtenus donnent un premier aperçu des comportements articulatoires mis en place par les locuteurs pour réaliser la tâche demandée, ces comportements pouvant être interprétés comme des stratégies de production sous la contrainte de vitesse. Ils mettent en évidence des stratégies articulatoires individuelles et une réalisation de la tâche plus ou moins aisée selon les locuteurs et les cibles. Cette variabilité entre sujets avait été pointée par Rochet-Capellan & Schwartz (2007).

Dans le cas de /pata/ *vs.* /tapa/, les productions des locutrices évoluent de telle sorte que la transition à partir du geste labial vers le geste lingual est plus rapide que dans le sens contraire. Ces observations s'orientent dans la direction des résultats obtenus par Rochet-Capellan & Schwartz (2007) et de Lancia & Fuchs (2011). Pourtant, alors que ces travaux insistent sur la coordination lèvre-mandibule et lèvre-langue, nos premières analyses semblent minimiser le rôle de la mâchoire (arrêt de l'oscillation à débit rapide). D'ailleurs Lancia & Fuchs ont remarqué que le couplage mandibule et lèvre est plus faible que celui de la pointe de la langue et de la mandibule. Nos premières analyses montrent une coordination de la mandibule avec le dos de la langue moins forte qu'avec l'apex.

Rochet-Capellan & Schwartz (2007) mettent en avant l'économie produite par l'alternance des lieux d'articulations qui rend possible l'anticipation des gestes. Or pour des locuteurs francophones, si nos résultats pour les séquences LC/CL confirment la stabilité LC et l'attraction des patrons CL vers LC sous l'effet de l'accélération, ce comportement est moins observé pour les séquences avec consonne vélaire. Si nos deux locutrices adoptent une stratégie similaire pour les séquences /pata/ et /tapa/, les mécanismes des locuteurs sont différents pour les séquences LV et VL. Nos résultats soulèvent la question de savoir pourquoi cette stratégie ne s'applique pas pour une séquence impliquant les lieux d'articulation labial et vélaire. Ce point fait écho à la recherche des séquences LV et VL dans les unités lexicales qui montre que contrairement à l'effet LC fort dans l'ensemble des langues, aucun effet LV ou VL ne peut être généralisé entre les langues. Les différents comportements de réorganisation articulatoire entre les sujets peuvent refléter cette absence d'effet généralisé.

Ce premier résultat expérimental sur les séquences labial-vélaire interroge aussi sur la question des phonèmes à double plosions majoritairement labio-vélaire, /k͡p/ ou /g͡b/, dans les langues africaines (Ladefoged & Maddieson, 1996). Bien que plus rares, d'autres doubles plosives existent mais, Maddieson & Ladefoged soulignent que leur durée est bien plus longue que celle de /k͡p/ ou /g͡b/, se rapprochant plus de la durée de deux segments consonantiques que d'un seul, ce qui questionne le degré de cohésion articulatoire entre les deux gestes de fermeture.

Ces premiers résultats en accord avec les travaux antérieurs cités montrent bien une réorganisation articulatoire entre les différents gestes pour pallier la contrainte de vitesse. L'asymétrie d'attractivité entre les patrons /pt/ et /tp/ et entre les patrons /pk/ et /kp/ est bien mise

en avant chez les deux locutrices.

Remerciements

Nous remercions tout particulièrement C. Savariaux pour son investissement dans le protocole expérimental et dans le recueil des données. Nous tenons à remercier l'équipe de recherche A93 pour leur accueil à Haskins Labs. Toute notre gratitude à M. Tiede pour nous avoir permis d'utiliser ses outils d'analyse de données EMA, ainsi que pour son aide et ses conseils. Enfin, la présentation des résultats de ce travail a été possible grâce à la collaboration de S. Gerber.

Références

BROWMAN, C.P.,, & GOLDSTEIN, L. (1992). Articulatory phonology: an overview. Phonetica, 49(3-4), 155-180.

CARRISSIMO-BERTOLA, M. (2010, juin 24). *Structures syllabiques des unités lexicales : « the fronting effect »* (Mémoire Master1). Université Stendhal, Grenoble.

CARRISSIMO-BERTOLA, M. (2011). *Tendances et universaux des structures syllabiques : l'effet Labial-Coronal en italien*. Université Stendhal, Grenoble.

HAKEN, H., KELSO, J. A. S., & BUNZ, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*(5), 347‑356.

KELSO, J. A. S., SALTZMAN, E. L., & TULLER, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, *14*(1), 29‑59.

LADEFOGED, P., & MADDIESON, I. (1996). *The sounds of the world's languages* (Vol. 1-1). Oxford: Blackwell.

LANCIA, L., & FUCHS, S. (2011). The labial coronal effect revised. In *Proceedings of the 9th International Seminar on Speech Production (ISSP)* (p. 187‑194). Montréal.

MACNEILAGE, P. F. (1998). The Frame/Content Theory of Evolution of Speech Production. *Behavioral and Brain Sciences*, *21*(04), 499‑511.

MACNEILAGE, P. F., & DAVIS, B. L. (2000). Deriving Speech from Nonspeech: A View from Ontogeny. *Phonetica*, *57*(2-4), 284‑296.

MACNEILAGE, P. F., DAVIS, B. L., KINNEY, A., & MATYEAR, C. L. (1999). Origin of Serial-Output Complexity in Speech. *Psychological Science*, *10*(5), 459 ‑460.

ROCHET-CAPELLAN, A., & SCHWARTZ, J.-L. (2007). An articulatory basis for the labial-to-coronal effect: /pata/seems a more stable articulatory pattern than /tapa/. *Journal of the Acoustical Society of America*, *121*(6), 3740‑3754.

ROUSSET, I. (2004). *Structures syllabiques et lexicales des langues du monde : données, typologies, tendances universelles et contraintes substancielles* (Thèse doctorat). Université Stendhal, Grenoble.

TSUJI, S., GOMEZ, N. G., MEDINA, V., NAZZI, T., & MAZUKA, R. (2012). The labial–coronal effect revisited: Japanese adults say pata, but hear tapa. *Cognition*, *125*(3), 413‑428.

TULLER, B., & KELSO, J. A. S. (1991). The production and perception of syllable structure. *Journal of Speech and Hearing Research*, *34*, 501‑508.

VALLÉE, N., ROSSATO, S., & ROUSSET, I. (2009). Favored syllabic patterns in the world's languages and sensori-motor constraints. In F. Pellegrino, E. Marsicoa, I. Chitoran, & C. Coupé (Éd.), *Approaches to Phonological Complexity* (p. 111‑140). Berlin: Mouton de Gruyter.

L'aspiration du /s/ coda en andalou occidental est-elle fonction du type de /s/ ?

Fanny Ivent[1] Cécile Fougeron[1] Mark Gibson[2]
(1) LPP, UMR 7018, 19 rue des Bernardins, 75005 Paris
(2) Universidad de Navarra, Departamento de Lingüística, 31009 Pamplona (España)
fanny.ivent@hotmail.fr, cecile.fougeron@univ-paris3.fr, mgibson@unav.es

RÉSUME

L'aspiration du /s/ en position de coda est une caractéristique qui définit de nombreux dialectes de l'espagnol. Nous étudions ce processus phonologique en andalou occidental sur des /s/ dans trois contextes morpho-phonologiques différents : des /s/ lexicaux internes et des /s/ morphémiques finaux qui seront soit des /s/ verbaux (2ème pers. sing.), soit des /s/ pluriels. Par une étude perceptive et acoustique menée sur les productions d'une locutrice et les jugements de 40 auditeurs, nous évaluons si le contraste phonétique entre des paires minimales qui se distinguent par la présence ou l'absence de /s/ est maintenu, lorsque le /s/ est réduit. Nos résultats montrent que le processus d'aspiration et ses conséquences perceptives dépendent du type de /s/. Le contraste phonétique est maintenu pour les /s/ lexicaux internes alors qu'il est menacé pour les /s/ morphémiques finaux, surtout pour les /s/ pluriels.

ABSTRACT

Is coda /s/ aspiration in Western Andalusian Spanish dependent on the type of /s/?

Syllable-final /s/ aspiration is a defining characteristic of several Spanish dialects. We study this phonological process in Western Andalusian Spanish on /s/ in three distinct morpho-phonological contexts: word-internal /s/ (lexical), final verbal morpheme /s/ (2nd pers. sing.) and final plural morpheme /s/. In a perceptive and acoustic study based on the productions of one female speaker and the judgments of 40 listeners, we assess whether the phonetic contrast between minimal pairs distinguished by the presence or absence of /s/, is maintained when /s/ is reduced. The results show that both syllable-final aspiration and perceptual discernibility are influenced by the different categories of /s/ examined in this study. The phonetic contrast is maintained for word-internal /s/, whereas for final /s/ the phonetic contrast is negligible especially when /s/ appears as the plural morpheme.

MOTS-CLÉS : processus phonologique, réduction, aspiration du /s/, andalou occidental, analyse acoustique, perception, types de /s/, ambiguïté sémantique.
KEYWORDS: phonological process, reduction, /s/ aspiration, Western Andalusian Spanish, acoustic analysis, perception, types of /s/, semantic ambiguity.

1 Introduction

Dans l'espagnol d'Andalousie occidentale, le /s/ en position de coda subit un processus de réduction appelé l'aspiration du /s/. Alors qu'il est généralement décrit comme donnant lieu à deux variantes allophoniques : (1) une forme avec débuccalisation de la fricative notée [h] et (2) une forme avec élision complète de la fricative notée [Ø], le processus semble plutôt graduel (Narbona, Cano, Morillo, 1998 ; Gerfen, 2002 ; O'Neill, 2008). Ainsi la réalisation [h] regroupe un ensemble de réalisations présentant des indices acoustiques variables du /s/ et le [Ø] est utilisé lorsque le segment semble avoir totalement disparu.

Différents types de /s/ subissent ce processus d'aspiration en espagnol. Nous traiterons ici trois types de /s/ différant selon leur contexte morpho-phonologique : (1) des /s/ lexicaux en position interne de mot (que l'on appellera /s/ lexicaux internes) dont la présence peut faire contraste entre des paires minimales du type /pasta/ « pâte » et /pata/ « patte » ; et (2) deux types de /s/ morphémiques en position finale de mot : des /s/ marqueurs du pluriel qui permettent un contraste entre les formes au singulier (ex. /faldas/ « jupes » vs. /falda/ « jupe »), et des /s/ marqueurs de la seconde personne du singulier des verbes au présent de l'indicatif qui permettent un contraste entre la troisième personne du singulier (ex. /kantas/ « tu chantes » vs. /kanta/ « il chante »). Afin de se référer à ces deux derniers types de /s/ nous parlerons de « /s/ pluriels » pour les noms et de « /s/ personnes » pour les verbes.

Si le processus d'aspiration du /s/ est très répandu en espagnol, les manifestations de ce phénomène sont très différentes d'un dialecte à l'autre (Torreira, 2007b ; Hammond, 1978 ; Gerfen, 2002 ; Colantoni et Kochetov, 2011). Il est donc important de ne pas généraliser les résultats issus d'une étude dans un dialecte en particulier à l'ensemble des dialectes affectés par ce processus de lénition. En andalou occidental, à notre connaissance, aucune étude portant sur les différents types de /s/ et alliant production et perception n'a été effectuée. De plus, les études perceptives sont très rares en règle générale dans ce dialecte. Nous citerons l'étude d'O'Neill (2005), portant sur les /s/ morphémiques finaux (pluriels et personnes mélangés), dont les résultats montrent un fort taux d'élision du segment et une incapacité des auditeurs à récupérer l'information dans le signal (taux d'identification correcte <50%). Ces résultats s'opposent fortement à ceux reportés par Torreira (2007a, 2007b) à l'issue d'études de production concernant les /s/ lexicaux internes en andalou occidental. En effet, l'auteur trouve de nombreux indices acoustiques du /s/ élidé (ex. /pasta/ produit [pahta] ou [paØta]) sur les segments adjacents et notamment sur la consonne suivant le /s/ dont la tenue et le VOT sont allongés. Nous avons donc d'une part, des /s/ morphémiques pour lesquels le contraste phonétique semble perdu et de l'autre, des /s/ lexicaux présentant de nombreux indices acoustiques du segment réduit. Ceci dit, nous n'avons pas de données en perception permettant de faire un éventuel lien entre ces indices acoustiques et la récupération de l'information du segment réduit.

Cette divergence de traitement entre les /s/ lexicaux et les /s/ morphémiques, révélée dans ces deux études, mérite une investigation plus poussée. Afin de s'assurer tout d'abord que la variation entre les participants de ces études ne soit pas la cause de ces résultats, il nous semble important de comparer le traitement de ces différents types de

/s/ au sein d'une même étude dans laquelle ils seraient produits par un même locuteur et testés perceptivement par les mêmes auditeurs. Si nous rencontrons des résultats semblables à ceux observés dans la littérature, il conviendra de trouver pourquoi les /s/ lexicaux et les /s/ morphémiques sont traités différemment. Cette différence peut être liée (1) à la position du segment dans le mot et/ou (2) au statut linguistique du /s/. Dans notre revue de la littérature, les /s/ lexicaux sont toujours étudiés en position interne de mot suivis d'une consonne occlusive sourde, alors que les /s/ morphémiques (personnes et pluriels) sont obligatoirement positionnés en fin de mot. Si des indices secondaires sur la consonne suivant le /s/ s'avèrent importants pour le maintien du contraste phonétique entre des paires minimales du type /pasta/, produit [pahta] ou [paØta], vs. /pata/, alors la position expliquerait pourquoi les /s/ lexicaux sont mieux traités que les /s/ morphémiques. Le statut linguistique pourrait aussi être une explication. L'information véhiculée par ces trois types de /s/ n'étant pas la même, il se peut que selon l'importance de celle-ci, les locuteurs adaptent leur production pour que la communication ne soit pas perturbée. En effet, nous pensons qu'en cas de forte ambiguïté sémantique, le locuteur chercherait à préserver au maximum les contrastes entre formes avec /s/ réduit et formes sans /s/ sous-jacent, et donc maintiendrait un plus grand nombre d'indices acoustiques du /s/. Nous émettons l'hypothèse qu'une réalisation sans [s] d'une forme verbale à la seconde personne du singulier est plus ambiguë qu'une réalisation sans [s] d'une forme nominale au pluriel étant donné que dans ce dernier cas la marque du pluriel peut aussi se retrouver dans certains articles (ex. articles définis « los » vs. « el »). L'information du pluriel est donc distribuée dans la phrase à plusieurs endroits, alors que l'information de la seconde personne du singulier du présent est uniquement contenue dans le verbe. En espagnol, le sujet n'apparaît pas dans la phrase sauf dans de rares cas où il a une fonction pragmatique (focus, insistance). Concernant les /s/ lexicaux, le contexte sémantique dans lequel apparaît le mot, permettrait aussi de lever l'ambiguïté sémantique. Nous regarderons donc si les /s/ lexicaux, les /s/ personnes et les /s/ pluriels sont ou non traités différemment. La comparaison des /s/ personnes avec les /s/ pluriels, tous deux en position finale, sera donc très importante pour déterminer l'impact de ce second facteur.

En résumé, nous chercherons dans cette étude à déterminer si le contraste phonétique entre une forme sans /s/ sous-jacent (exemple /pata/) et une forme avec un /s/ sous-jacent (exemple /pasta/), dont la réalisation est réduite ([pahta] ou [paØta]), est ou non maintenu en perception et quels indices acoustiques sont disponibles dans le signal produit, ceci pour les différents types de /s/ (lexicaux, pluriels et personnes), afin de voir si l'on observe une différence de traitement entre chacun d'eux.

2 Matériel et Méthode

Le matériel utilisé consiste en 42 paires minimales se distinguant par la présence ou l'absence de /s/ dans la forme sous-jacente du mot. Parmi celles-ci, 10 paires mettent en jeu des /s/ lexicaux internes (/roska/ « galette » vs. /roka/ « roche ») et 32 paires comprennent des /s/ morphémiques finaux, dont 16 /s/ personnes (« komes/ « tu manges » vs. /kome/ « il mange ») et 16 pluriels (/fotos/ « photos » vs. /foto/ « photo »). La position de l'accent dans les items a été contrôlée au mieux. Les /s/ lexicaux internes se trouvent en général en syllabe accentuée (huit mots bisyllabiques et un mot

trisyllabique accentués sur la première syllabe contenant le /s/ et un mot trisyllabique accentué sur la pénultième syllabe avec un /s/ en syllabe pré-tonique). Les /s/ morphémiques (pluriels et personnes) par contre se trouvent tous en syllabe non-accentuée puisque toutes les formes choisies portent l'accent lexical sur la pénultième syllabe. Nous parlerons d'items « tests » pour se référer aux formes avec un /s/ sous-jacent et d'items « contrôles » pour se référer aux formes sans /s/ sous-jacent. Les items tests et contrôles ont été produits dans des phrases cadres lues au sein d'un corpus plus vaste comprenant un total de 402 phrases. Les enregistrements ont eu lieu à l'université de Pampelune (Navarre, Espagne). Une locutrice andalouse (de Cordoue) a été sélectionnée pour notre étude. Nous avons extrait un total de 151 réalisations de /s/ (20 /s/ lexicaux, 67 /s/ personnes, 64 /s/ pluriels) de ses productions. Les réalisations de /s/ contenues dans l'échantillon ont été catégorisées, à partir de la visualisation des spectrogrammes et l'écoute des productions, par un juge expert maîtrisant bien l'espagnol. Quatre catégories phonétiques ont été définies : [s] pour les formes non réduites, [h] pour les formes débuccalisées de la fricative et les celles qui contenaient en plus de la voix soufflée sur la voyelle précédant le segment, [V̤] lorsque l'expérimentateur notait uniquement de la voix soufflée sur cette voyelle, et [Ø] lorsque le segment semblait totalement élidé. Ces différents types de réalisations réduites de /s/ sont illustrés en figure 1.

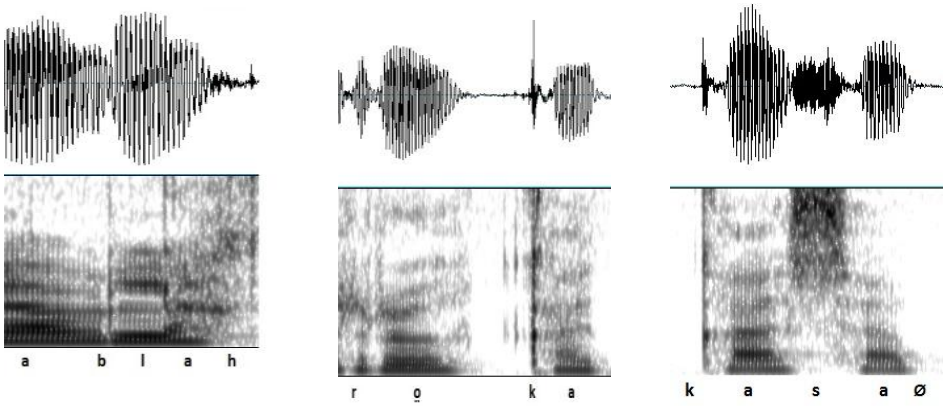


FIGURE 1 – Illustration des trois catégories possibles. À gauche la forme /ablas/ « tu parles » réalisée [abl̤ah] avec une fricative débuccalisée et de la voix soufflée sur V_1 ; au centre la forme /roska/ « galette » réalisée [ro̤ka] ; et à droite, la forme /kasas/ « maisons » réalisée [kasaØ].

La première catégorie ([s] non réduit) a été éliminée par la suite car la locutrice a réduit l'ensemble des items avec un /s/ sous-jacent, traduisant la productivité du phénomène d'aspiration en andalou occidental. Les 151 formes avec /s/ étant toutes réduites, nous avons sélectionné pour le test de perception deux répétitions de 34 des paires minimales produites (10 formes lexicales, 12 verbes et 12 noms), soit 68 items tests (tous présentant des formes réduites du segment) et 68 items contrôles (sans /s/ sous-jacent). Le critère de sélection des formes tests était de privilégier la variation dans les réalisations de /s/, selon les 3 variantes phonétiques définies plus haut, de façon à avoir des formes avec plus ou moins d'informations acoustiques du /s/ (voir la distribution en 4). 300 distracteurs créés à partir de mots et de non mots du corpus de production ont

été inclus dans le test. Tous mettaient en jeu des variations segmentales semblables à notre objet d'étude (ex. élision du /r/ en position de coda). Au total, 136 items (tests et contrôles) et 300 distracteurs ont été présentés auditivement aux participants pour une tâche d'identification. 40 participants andalous (20) et non-andalous (20) ont pris part à l'expérience. Ces deux populations sont regroupées dans cet article car elles n'ont pas montré de différence de traitement (Ivent, 2013). Tous les participants vivaient à Pampelune et étaient pour la plupart des étudiants et des professeurs d'université. Les participants devaient choisir parmi deux réponses orthographiques présentées à l'écran, celle qu'ils pensaient avoir entendu. Par exemple les participants entendaient la forme [paØta] et ils devaient dire s'il s'agissait du mot *pasta* ou *pata*. Les réponses (correctes / incorrectes) et les temps de réaction (non présentés ici, voir Ivent 2013) ont été récoltés sous E-Prime. L'analyse statistique des données de perception consiste en des analyses de variance et des tests-t indépendants effectués après une normalisation arc-sinus des taux d'identification correcte. Pour les ANOVAs, les tests post-hoc sont des tests de Fisher avec un seuil de significativité fixé à .05.

L'analyse acoustique des items tests et contrôles, utilisés pour le test de perception, comprend une analyse qualitative de la distribution des formes réduites de /s/ pour les trois types de /s/ et une analyse quantitative des différences acoustiques entre formes tests et contrôles pour les items lexicaux. Différentes dimensions acoustiques relevées comme pertinentes au contraste dans la littérature ont été examinées. La durée totale du mot ainsi que la durée des différents intervalles indiqués sur la figure 2 sont comparées entre items tests et contrôles. Le timbre de V_1, indice pertinent en andalou oriental (Navarro Tomás, 1939 ; Alvar, 1955 ; Salvador, 1977), est examiné par des mesures spectrales de F1 et F2. La qualité de voix sur V_1 a également été présentée comme un indice de l'aspiration du /s/ par O'Neill (2008). Nous avons donc mesuré la pente du fondamental entre 50 et 95% de la durée de V_1, afin de voir si la réduction du /s/ en espagnol andalou s'accompagne d'une chute du fondamental qui est un indice de voix soufflée dans certaines langues (Ohala, 1973). Une série de test-t indépendants a été effectuée sur ces mesures afin de tester la significativité des différences observées entre les formes tests et les formes contrôles.

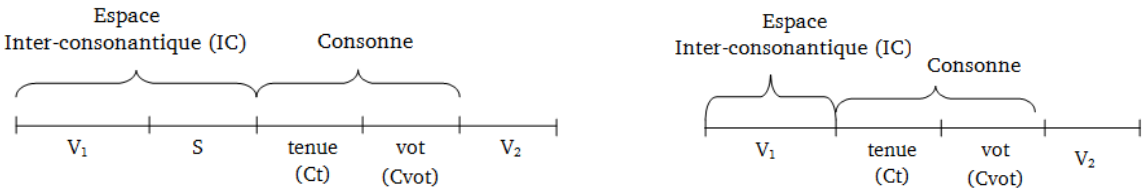


FIGURE 2 – Différents intervalles temporels mesurés sur les items tests (à gauche) et leurs contrôles (à droite) pour les mots à /s/ lexicaux internes.

3 Identification des formes avec /s/ réduit

Comme attendu, les items contrôles (mots sans /s/ dans leur forme sous-jacente, comme /pata/, /kanta/, /falda/) ont été bien identifiés, sans pour autant atteindre 100%. Par contre, le taux d'identification des items tests comprenant un /s/ réduit est significativement inférieur ($F(1, 130)=183$; $p<.0001$), avec une interaction significative en fonction du type de /s/ ($F(2, 130)=26$; $p<.0001$). Les items avec un /s/ lexical sont

identifiés correctement à 90% mais restent moins bien identifiés que les items contrôles (t_{38}=2.68 ; p=.01). L'identification des items tests morphémiques (personnes et pluriels) est également inférieure à celle de leurs contrôles (t_{46}=8.12 ; p<.0001 et t_{46}=12.68 ; p<.0001, respectivement) avec un écart entre tests et contrôles bien plus important. En effet, les items tests avec un /s/ morphémique final ont été très mal identifiés avec seulement 30% d'identification correcte pour les items à /s/ pluriels et 39% pour les items à /s/ personnes. Une comparaison entre les items tests des trois catégories (lexicaux, personnes, pluriels) montre que l'identification des formes avec /s/ réduit est fonction du type de /s/ (F(2, 65)=62.81 ; p<.0001) et que les /s/ lexicaux internes réduits sont bien mieux identifiés que les deux /s/ morphémiques finaux (p<.0001). En revanche, le taux d'identification des /s/ morphémiques (pluriels et personnes) ne diffère pas (p>.05).

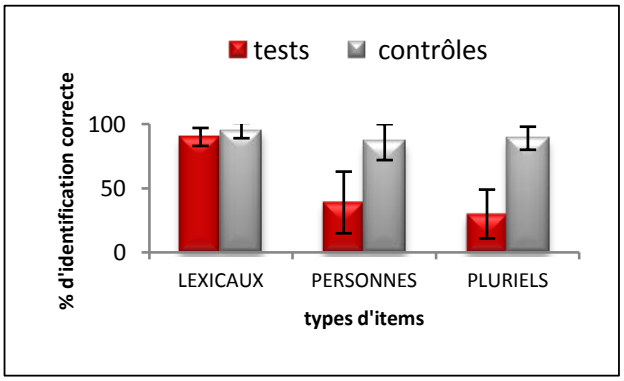


FIGURE 3 – Taux d'identifications correctes (en %) des différents types d'items (lexicaux, personnes et pluriels) tests (en rouge) et contrôles (en gris).

4 Analyse acoustique des items de perception

Pour rappel, toutes les formes avec un /s/ sous-jacent produites par la locutrice, et donc toutes les formes sélectionnées pour le test de perception, comportent des /s/ réduits. La distribution des réalisations des 68 items tests au sein des 3 catégories de variantes acoustiques définies [h], [V̤], [Ø], (voir section 2) est présentée dans le tableau 1.

	/s/ lexicaux (N=20)	/s/ personnes (N=24)	/s/ pluriels (N=24)
[h]	70%	54%	29%
[V̤]	10%	8%	4%
[Ø]	20%	38%	67%

TABLE 1 – Distribution des réalisations phonétiques des trois types de /s/ de la locutrice.

La distribution des formes réduites de /s/ varie en fonction du type de /s/ avec pour les /s/ lexicaux une majorité de réductions partielles, soit un maintien de traces acoustiques pendant l'intervalle correspondant au /s/ et/ou la voyelle précédente (70% de [h] et 10% de [V̤]). En revanche, la distribution des /s/ pluriels est renversée avec une majorité d'élisions totales (67% de [Ø]). Enfin, pour les /s/ personnes, la répartition est plus

équilibrée même si les élisions totales (38% de [Ø]) restent moins fréquentes que les réductions graduelles présentant des traces acoustiques de la fricative (54% de [h] et 8% de [V̤]). Nous pouvons expliquer le fort taux d'identification correcte des /s/ lexicaux par la préservation de traces acoustiques de la fricative traduisant un processus graduel de réduction du segment dans ce type de mot. Ces indices, susceptibles d'être utilisés par les auditeurs, comprennent la présence d'une fricative débuccalisée ([h]) mais aussi des différences acoustiques distribuées sur les segments environnants. La figure 4 présente les résultats des mesures acoustiques effectuées sur ces items lexicaux.

Durée (ms.)					
Mot	V_1	IC	Ct	Cvot	V_2
ns	ns	ns	p<.03	p<.0001	p=.0002

V_1		
F1	F2	F0
ns	ns	p<.0001

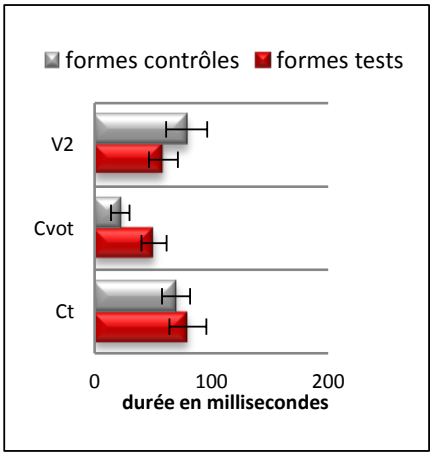


FIGURE 4 – Comparaison acoustique entre formes tests et formes contrôles pour les items lexicaux. Résultats des tests-t pour toutes les mesures à gauche et représentation graphique des durées moyennes de V_2, Cvot et Ct avec écart-type à droite.

Les items tests se distinguent notamment de leurs contrôles par des différences de durée sur les segments suivants le /s/. En effet, si la durée totale du mot, de V_1 et de l'intervalle IC (incluant V_1 et /s/ réduit) sont similaires au sein des formes tests et contrôles, le contraste ressort au niveau de l'organisation temporelle des segments C et V_2. La consonne occlusive sourde des formes tests est plus longue tant au niveau de sa tenue que de son VOT. Cet allongement compensatoire de la consonne suivant le /s/ aspiré a aussi été montré par Torreira (2007a, 2007b) dans ce dialecte. La voyelle post-consonantique, V_2, est par contre significativement plus courte dans les items tests. À notre connaissance, cette mesure n'a pas été soumise à l'analyse dans de précédentes études. Ce résultat peut s'expliquer par un ajustement des durées segmentales au sein des formes avec /s/ sous-jacent qui contiennent plus de segments que les formes contrôles, et donc, des segments plus courts. De plus, cet ajustement de durée sur V_2 peut être d'autant plus important au sein de la syllabe CV_2 lorsque C est allongée. Outre ces indices temporels sur CV_2, les items tests se distinguent des contrôles par une pente de F0 sur V_1 plus abrupte (p<.0001) que nous interprétons comme un indice de souffle sur la voyelle précédant le /s/ aspiré. Par contre, contrairement à ce qui a été observé en andalou oriental, le timbre de V_1 ne semble pas être un indice d'aspiration en andalou occidental comme cela a déjà été reporté (Alvar, 1955).

En ce qui concerne les /s/ morphémiques finaux, nous avons vu que le contraste perceptif entre items tests et contrôles est en péril avec moins de 40% d'identification correcte des formes avec /s/. Pour autant, lorsque l'on examine les formes présentées aux auditeurs, 62% des items tests avec /s/ de personne présentent une réduction graduelle

du /s/ (54% de [h] et 8% de [V̤]). Des cas de réduction graduelle apparaissent aussi sur 33% des formes avec /s/ de pluriel. Ces formes contiennent donc des indices acoustiques du segment aspiré (la présence d'une fricative débuccalisée et/ou de la voix soufflée sur V_1). Une analyse des taux d'identifications entre formes à réduction graduelle ([h] et [V̤]) vs. totale ([Ø]), pour les deux /s/ morphémiques finaux, montre que les auditeurs ont pu avoir recourt à ces indices. En effet, les formes à réduction graduelle sont significativement mieux identifiées que les formes [Ø] ($p<.0001$) même si leur taux d'identification moyen reste faible à 51% (contre 20% pour les formes à réduction totale).

5 Conclusion

Les résultats de cette comparaison entre /s/ lexicaux internes et /s/ morphémiques finaux au sein d'une même étude confirment ceux de la littérature, à savoir une différence de traitement en fonction du type de /s/ en andalou occidental.

Au niveau de la production, nous observons chez notre locutrice que le processus d'aspiration du /s/ produit des variantes phonétiques qui se répartissent sur un continuum de formes plus ou moins réduites. Nous avons émis l'hypothèse que la réduction phonétique du /s/ pouvait être fonction de l'ambiguïté provoquée par l'aspiration. La comparaison de deux types de /s/ morphémiques finaux (personnes et pluriels) offre des pistes intéressantes en ce sens, qu'il nous faudra poursuivre sur un plus grand nombre de locuteurs. En effet, il apparaît que les cas de réductions les plus extrêmes ([Ø]) sont plus fréquents pour les /s/ pluriels que pour les /s/ personnes. Dans les phrases de notre corpus, les mots avec /s/ pluriels étaient précédés d'un article défini qui porte également la marque du pluriel. L'information portée par ce /s/ final est donc redondante ce qui pourrait favoriser son élision. Par comparaison, la réduction du /s/ final des verbes paraît plus compromettante puisque l'information de personne est uniquement portée par ce segment, ce qui peut expliquer son plus faible taux d'élision complète. Cependant, si un effet du poids informatif du /s/ semble affecter la réalisation acoustique du segment réduit, il ne se traduit guère au niveau perceptif. En effet, si les auditeurs identifient un peu mieux les items avec /s/ personnes (39%) que les items avec /s/ pluriels (30%), cette différence n'est pas significative et ces deux types d'items sont globalement très mal identifiés.

Cette mauvaise identification des /s/ morphémiques finaux contraste fortement avec celle des /s/ lexicaux internes. Nous observons que l'aspiration des /s/ lexicaux internes, en andalou occidental, est graduelle avec une préservation d'indices acoustiques sur le segment réduit (produit comme une fricative débuccalisée) et/ou sur les segments adjacents comme l'a montré Torreira (2007a, 2007b). De plus, nous confirmons que dans ce dialecte, le contraste entre forme à /s/ lexical aspiré et formes sans /s/ sous-jacent est relativement bien préservé en perception. Dans cette étude nous n'avons pas testé la contribution relative des différents indices à l'identification des formes. Pour autant, la comparaison des différents types de /s/ et l'analyse acoustique des items suggèrent que la présence d'indices acoustiques sur la consonne suivant le /s/ (et sur V_2) sont des indices importants pour la préservation du contraste. Présents dans les formes avec un /s/ interne lexical, ces indices sont indisponibles quand le /s/ est en position finale.

Références

AlVAR, M. (1955). Las encuestas del Atlas Lingüístico y Etnográfico de Andalucía. *Revista de Dialectología y Tradiciones Populares,* 11:231–274.

GERFEN, C. (2002). Andalusian codas1. *Probus,* 14:247–277.

HAMMOND, R. (1978). An experimental verification of the phonemic status of open and closed vowels in Caribbean Spanish. *Corrientes actuales en la dialectología del Caribe hispánico*, 33–125.

IVENT, F. (2013). L'aspiration du /s/ en position coda en espagnol d'Andalousie occidentale : analyse acoustico-perceptive autour de l'ambiguïté du message. *Mémoire de M2,* Université Sorbonne Nouvelle – Paris 3.

KOCHETOV, A., COLANTONI, L. (2011). Gestural coordination in Spanish /s/ weakening: An electropalatographic study. *In Proceedings of the international Seminar on Speech Production (ISPP 2011),* Montreal, 273–280.

NARBONA, A., CANO, R., MORILLO, R. (1998). *El español hablado en Andalucía*. Ariel Lingüística.

NAVARRO TOMÁS, T. (1939). Desdoblamiento de fonemas vocálicos. *Revista de Filología Hispánica*, 1:165–167.

OHALA, J. (1973). The physiology of tone. *Southern California Occasional Papers in Linguistics,* 1:1-14.

O'NEILL, P. (2005). Utterance final /s/ in Andalusian Spanish. The phonetic neutralization of a phonological contrast. *Language Design: Journal of Theoretical and Experimental Linguistics*, 151–166.

O'NEILL, P. (2008). La fonación murmurada y la preaspiración en andaluz. *Language Design*, 9.

SALVADOR, G. (1977). Unidades fonológicas en el andaluz oriental. *Revista Española de Lingüística*, 7:1–23.

TORREIRA, F. (2007a). Pre- and post-aspirated stops in Andalusian Spanish. *Prosodic and segmental issues in Romance*, 67–82.

TORREIRA, F. (2007b). Coarticulation between aspirated-s and voiceless stops in Spanish: An interdialectal comparison. *The 9th Hispanic Linguistics Symposium*, 113–120.

Modélisation acoustico-phonétique de langues peu dotées : Études phonétiques et travaux de reconnaissance automatique en luxembourgois

Martine Adda-Decker[1, 2] Lori Lamel[2] Gilles Adda[2]
(1) LPP, CNRS-Paris 3/Sorbonne Nouvelle
(2) Groupe TLP, LIMSI-CNRS
martine.adda-decker@univ-paris3.fr, lori.lamel@limsi.fr ,gilles.adda@limsi.fr

RÉSUMÉ

Le luxembourgeois est une langue germano-franconique et l'une des langues européennes sous-décrites. Cet article étudie la similitude entre les segments phonétiques en luxembourgeois avec leurs équivalents en allemand, français et anglais *via* des techniques d'alignement forcés. En utilisant les modèles acoustiques monolingues d'amorçage de ces trois langues, ainsi que des modèles "multilingues" entraînés sur un corpus de parole obtenu par concaténation, nous avons examiné si le luxembourgeois était mieux représenté par l'une des langues prises individuellement ou par le modèle multilingue. Au niveau global, les modèles allemands fournissent la meilleure correspondance, mais une analyse par segments montre des préférences spécifiques. Les premiers résultats en transcriptions illustrent les performances des différents jeux de modèles acoustiques monolingues et multilingues, ainsi que les modèles luxembourgeois construits à partir de 1200 heures de parole non transcrites en luxembourgeois, et des méthodes non supervisées.

ABSTRACT

Acoustic-phonetic modeling for under-resourced languages : an overview of recent phonetic studies and automatic speech recognition experiments in Luxembourgish

Luxembourgish, a Germanic-Franconian language, is embedded in a multilingual context on the divide between Romance and Germanic cultures and remains one of Europe's under-described languages. This paper investigates the similarity between Luxembourgish phone segments with German, French and English via forced speech alignment techniques. Making use of monolingual acoustic seed models from these three languages, as well as "multilingual" models trained on pooled speech data we investigated whether Luxembourgish was globally better represented by one of the individual languages or by the multilingual model. While globally, the German models provide the best match, a phone-based analysis, shows language-specific preferences. First ASR results illustrate the accuracy of the various sets of monolingual and multilingual acoustic models and Luxemburgish acoustic models built from 1200 hours of untranscribed Luxemburgish audio data using unsupervised methods.

MOTS-CLÉS : Langues peu dotées ; modélisation acoustique ; modèles multilingues ; système de transcription de la parole ; luxembourgeois ; alignements forcés..

KEYWORDS: under-resourced languages ; acoustic modeling ; multilingual models ; large vocabulary speech recognition ; Luxembourgish ; Forced alignment..

1 Introduction

Le Luxembourg est un pays au centre de l'Europe de l'Ouest, composé de 65% d'habitants autochtones et de 35% d'immigrés. La langue nationale, le luxembourgeois ("Lëtzebuergesch"), n'est la langue officielle que depuis 1984 et n'est parlé que par les autochtones ; les immigrés selon leur pays d'origine parlent l'une des deux autres langues officielles, le français et l'allemand. A ces langues officielles s'est ajouté il y a peu l'anglais en tant que langue de communication fréquente, en particulier dans les milieux professionnels (banque, commission européenne). Le luxembourgeois peut être considéré comme une langue partiellement peu dotée (Adda-Decker *et al.*, 2008b). En effet, il y a peu ressources linguistiques telles que des lexiques ou des corpus de langue écrite en luxembourgeois au profit des écrits en français et en allemand.

Dans ce travail, nous allons examiner les propriétés acoustiques qui définissent la langue luxembourgeoise dans sa relation avec ses influences germaniques et romanes. Nous regardons en particulier l'influence de l'allemand, le français et l'anglais sur la réalisation acoustique des phonèmes du luxembourgeois. Pour cela nous utiliserons des données audio alignées au niveau phonémique. Nous examinerons ensuite comment on peut utiliser cette connaissance pour construire des modèles d'amorçage pour produire des modèles du luxembourgeois en utilisant un apprentissage non-supervisé.

La section suivante présente l'inventaire phonémique du luxembourgeois et sa correspondance avec les trois langues exogènes présentes aux Luxembourg. Nous présentons des résultats d'alignement pour des modèles d'amorçage acoustiques mono- et multilingues. Ensuite nous présentons les résultats en reconnaissance utilisant des modèles acoustiques issus de ces modèles d'amorçage, et appris à l'aide de méthodes non-supervisées sur un grande quantité de données en luxembourgeois. Enfin, nous concluons et présentons les perspectives tant en transcription que pour les études linguistiques sur le luxembourgeois.

2 Similarité entre segments phonémiques

2.1 Inventaire phonémique du luxembourgeois

L'inventaire phonémique du luxembourgeois que nous avons adopté (Schanen, 2004) contient 60 symboles dont 3 symboles extra-phonémiques pour le silence, la respiration et l'hésitation. La Table 1 présente un échantillon de l'inventaire phonémique avec des exemples. Le luxembourgeois est caractérisé par un grand nombre de diphtongues. Nous avons choisi de coder les diphtongues et les affriqués à l'aide d'un seul symbole. Étant donnée l'importance de l'apport du français, nous avons inclus les nasales dans l'inventaire, bien qu'elles ne soient théoriquement pas présentes dans les mots luxembourgeois.

2.2 Modèles acoustiques d'amorçage

Nous avons construit 3 jeux de 60 modèles acoustiques suivant le travail initié dans (Adda-Decker *et al.*, 2010, 2011b), en utilisant les langues exogènes présentes au Luxembourg et pour lesquelles il existe une quantité suffisante de données d'apprentissage. Ainsi, nous avons utilisé 150 heures

exemple (Français)	Lux	Fra	All	Ang
VOYELLES ORALES				
li**i**cht (lumière)	i	i	i	i
schü**t**zen (bus)	ʏ	y	ʏ	ɪ
f**ä**eg (capable)	ɛ :	ɛ	ɛ :	ɛ
DIPHTONGUES				
l**éi**en (mentir)	e͡ɪ	e	e	e
l**ou**nen (louer)	ɔ͡ʊ	o	o	o

TABLE 1 – Echantillon de correspondances cross-lingues entre phonèmes : Les cibles en luxembourgeois sont mises en correspondance avec un phonème identique ou similaire dans les 3 langues exogènes (Fra, All, Ang).

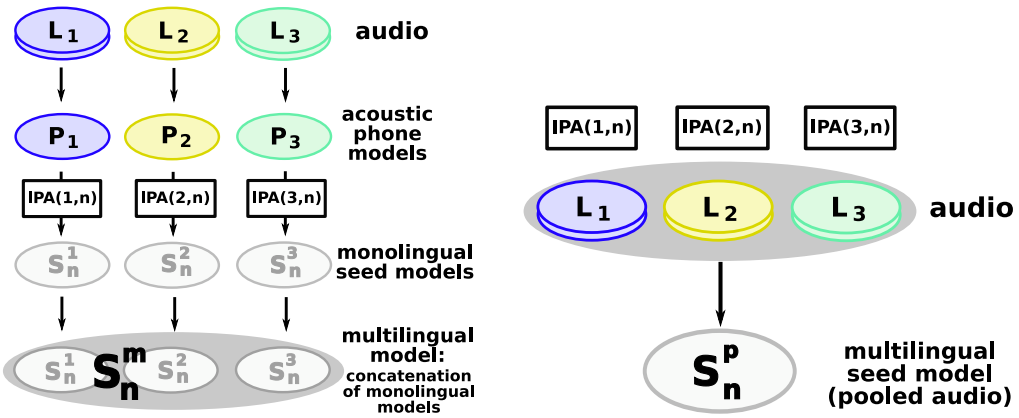


FIGURE 1 – Modèles acoustique d'amorçage pour la langue cible n (luxembourgeois) étant donnés des modèles de phones P_i d'une langue L_i ($i = 1, 2, 3$ anglais, français, allemand), et les mises en correspondance des symboles IPA entre les langues i et n IPA(i,n). gauche : monolingue S_n^i et multilingue par concaténation des modèles S_n^m ; droite : modèles multilingues par concaténation des données S_n^p .

pour l'anglais et le français, et 40 heures pour l'allemand. Les jeux de phonèmes monolingues sont de taille différente : 48 pour l'anglais, 37 pour le français et 49 pour l'allemand. Les modèles de phones sont des modèles de Markov cachés représentant des allophones contextuels avec une structure gauche-droite à états liés utilisant des mélanges de gaussiennes (typiquement 64).

La figure 1 (gauche) illustre le développement de 3 modèles acoustiques de pseudo-luxembourgeois, contenant chacun 60 phones, pour l'anglais, le français et l'allemand, en mettant en correspondance les phonèmes du luxembourgeois avec leur plus proche équivalent dans la langue source (IPA(i, n) dans la Fig. 1). Un quatrième modèle a été construit en concaténant les 3 modèles précédents, en laissant au décodeur le choix parmi ceux-ci. Enfin un jeu de modèles multilingues a été appris en mettant en commun les trois corpus d'apprentissage (figure 1 droite, puis dénommé *pooled* dans le reste de l'article), en utilisant les labels fournis par les correspondances IPA(i, n). Ces jeux multilingues reprennent des concepts introduits par exemple dans (Schultz et Waibel, 2001).

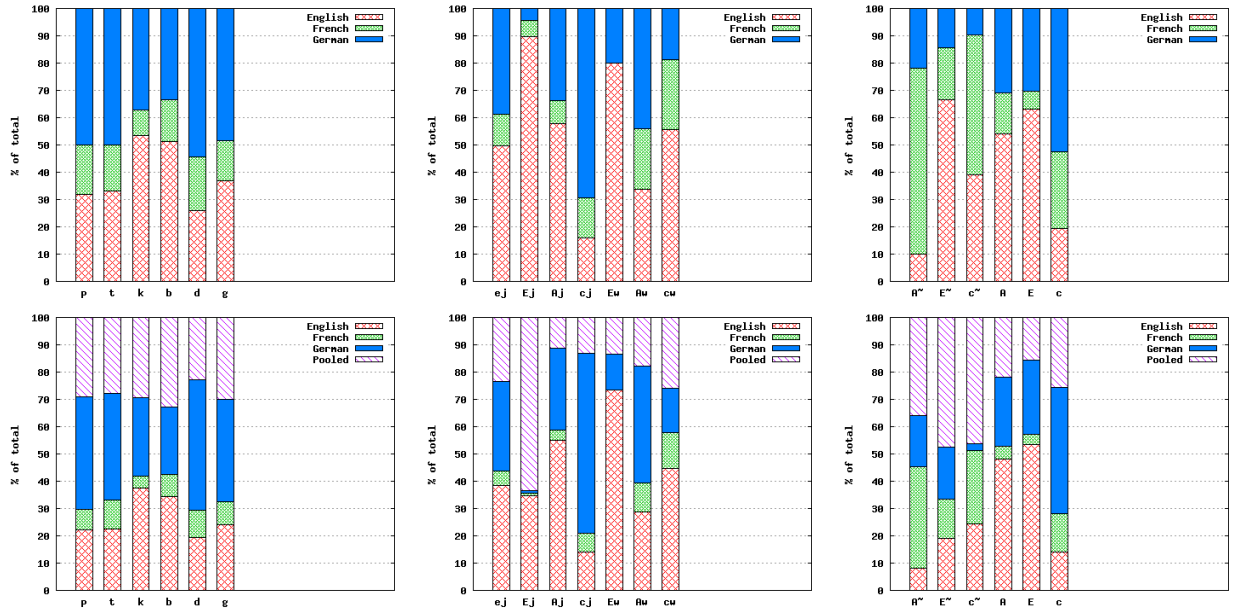


FIGURE 2 – Alignement en utilisant les modèles multilingues par concaténation des modèles monolingues (haut) et en ajoutant le modèle multilingue obtenu par concaténation des données d'apprentissage (pooled, bas). Les taux de sélection sont montrés pour chaque phonème ; gauche : plosives (/p,t,k,b,d,g/) ; milieu : diphtongues /e^{j},ɛ^{j},a^{j},ɔ^{j},ɛ^{w},a ^{w},ɔ^{w}/ ; droite : nasales /ɛ̃ ,ã,õ/ et les voyelles orales correspondantes /ɛ,a,ɔ/.

2.3 Données de parole en luxembourgeois pour l'alignement forcé

Les alignements forcés ont été réalisés sur 80 minutes de parole transcrites manuellement, en provenance du parlement luxembourgeois (débats de la Chambre (70')) et des actualités en provenance de la chaîne RTL (10') (Adda-Decker *et al.*, 2008b).

La durée moyenne d'un phone est assez stable (70-80ms) quels que soient les modèles d'amorçage utilisés. Le corpus contient un total de 56 000 segments de phones.

2.4 Résultats

Dans cette étude où nous désirons examiner les propriétés phonémiques des différents modèles, l'identité de chaque langue peut changer à la frontière de chaque phone. De ce fait, on observe de nombreux changements de langue et les modèles allemands sont moins utilisés que dans une étude précédente (Adda-Decker *et al.*, 2010), où le changement avait lieu à chaque frontière de mot. Dans l'alignement avec les modèles obtenus par concaténation des données, les modèles allemands sont alignés avec 36% des segments, les modèles anglais avec 27%, et les modèles français uniquement 10,5%. Dans toutes les conditions, les langues germaniques sont préférées, ce qui est compatible avec un influence corrélée à une distance typologique.

Dans la mise en correspondance entre les symboles IPA du luxembourgeois et des langues exogènes, il y a des symboles qui sont partagés (les plosives) et d'autres pour lesquels la correspondance est approximative (les diphtongues). Nous donnons ici (voir figure 2) quelques détails sur les résultats d'alignement en fonction de la proximité des symboles lors de la mise en

correspondance. Pour chaque symbole, les segments correspondants sont alignés avec l'un des modèles acoustiques (anglais, français, allemand, *pooled*). Les taux sont calculés en fonction de la langue d'origine choisie. Le taux correspondant au modèle *pooled* donne une indication de l'écart entre la langue cible et les modèles monolingues sources.

(i) symboles partagés entre les langues : les plosives Les plosives /p/, /t/ and /k/ et leur correspondant voisé existent dans les 4 langues (sources et cible). Le luxembourgeois étant considéré comme une langue germanique, on peut faire l'hypothèse que les plosives doivent être réalisées d'une façon similaire à celle de l'allemand ou de l'anglais. Les résultats détaillés dans la figure 2 (gauche) montrent que les segments sont partagés entre l'allemand et l'anglais et que seuls 10 à 20% des segments utilisent le modèle français. En ajoutant les modèles *pooled* (bas), environ 30% des données utilisent ce modèle.

(ii) correspondance approximative : les diphtongues Le répertoire phonémique du luxembourgeois contient un grand nombre de diphtongues. La figure 2 (milieu) montre un tableau moins homogène que pour les plosives. Il y a une tendance à utiliser les modèles anglais, les segments du phonème /$\underset{\smile}{\varepsilon\text{ɪ}}$ / étant alignés à 90% avec des modèles anglais, alors que /$\underset{\smile}{\text{ɔ}\text{ɪ}}$ / sont clairement plus allemand. Les modèles *pooled* semble absorber /$\underset{\smile}{\varepsilon\text{ɪ}}$ /, alors que la situation reste inchangée pour /$\underset{\smile}{\text{ɔ}\text{ɪ}}$ /.

(iii) correspondance approximative : les voyelles nasales Les voyelles nasales sont utilisées en luxembourgeois pour certains mots importés du français. La figure 2 (droite) montrent les résultats pour les nasales et les voyelles orales correspondantes. Le taux d'utilisation du modèle français est très élevé pour les segments ã et õ. En introduisant les modèles *pooled* ce taux tend à chuter ce qui montre le faible lien entre les nasales du français et du luxembourgeois.

On peut voir dans la figure 2 (bas) que le modèle *pooled* est utilisé pour aligner un grand nombre de segments (par exemple /$\underset{\smile}{\varepsilon\text{ɪ}}$ /) pour certains phonèmes, alors que pour d'autres le modèle *pooled* n'est utilisé que pour une faible partie des données.

3 Application à la transcription automatique du luxembourgeois

Des expériences préliminaires (Adda-Decker *et al.*, 2011a) nous ont permis de voir les limites de modèles multilingues tels que développés dans la section précédente. Afin d'atteindre des taux d'erreurs acceptables et en l'absence de transcription de corpus audio en luxembourgeois, les modèles développés dans la section précédentes doivent servir de modèles d'amorçage pour la construction de modèles acoustiques en luxembourgeois, en utilisant la méthodologie développée dans (Lamel *et al.*, 2002) : une grande quantité de données audio en luxembourgeois est décodée de manière itérative à l'aide de modèles de langage en luxembourgeois et des modèles acoustiques obtenus lors de la précédente itération. A chaque itération, et au fur et à mesure que les modèles acoustiques sont plus précis et plus efficaces, une plus grande quantité de données audio et un plus grand nombre de contextes phonétiques sont utilisés. Le nombre l'itération nécessaires et la

quantité de données audio incorporée à chaque itération est largement empirique, et la solution retenue ici correspond à la méthodologie utilisée dans (Oparin *et al.*, 2013).

3.1 Donnée audio en luxembourgeois

Nous avons recueilli dans le cadre du projet Quaero une grande quantité de données audio en luxembourgeois. Elles correspondent à des données audio recueillies sur le Web en 2012 et 2013, principalement en provenance de RTL (flash 1000h, journal 880h) mais également d'autres origine (Radio100.7 15h, talk show 5h,...). Le nombre total d'heures récoltées est de 1678. Ces données sont utilisées en partie pour l'apprentissage non-supervisé des modèles acoustiques en luxembourgeois. Les jeux de développement (196mn) et d'évaluation (205mn) sont les jeux officiels fournis dans le cadre du projet Quaero, provenant des mêmes sources que celles présentes dans l'apprentissage.

3.2 Modèle de langage

Nous avons utilisé un système d'identification de la langue (Lavergne *et al.*, 2010) afin de filtrer de manière efficace le luxembourgeois des langues exogènes mentionnées précédemment, afin de pouvoir utiliser les données hétérogènes recueillies sur le Web.

Utilisation de textes multilingues hétérogènes Nous avons utilisé divers textes en luxembourgeois, certains décrits dans (Adda-Decker *et al.*, 2008b) et d'autres recueillis plus récemment sur le Web (Adda-Decker *et al.*, 2011a). Les textes appartiennent à 3 domaines :

1. **Actualités, sources écrites** :
   - RTL2008 : données RTL anciennes (avant 2008) et filtrées manuellement (0,6Mmots)
   - RTL2012 : Sites Web de RTL (aspirés en 2012) (10,3Mmots).
   - WIKIPEDIA : le Wikipedia luxembourgeois (3,6Mmots).
   - MISC : des rapports, livres, revues. . .recueillis sur le Web (1,7Mmots).
2. **Transcription de l'oral** :
   - CHAMBRE : transcriptions *bona fide* (Adda-Decker *et al.*, 2008a) des débats au parlement luxembourgeois (22,1Mmots).
3. **média sociaux** :
   - BLOGS : 90 blogs (provenant d'une liste de 400 blogs en répertorié comme luxembourgeois) (10,2Mmots)
   - BLOGS_COMMENT : les commentaires des internautes sur les blogs préselectionnés (3,1Mmots).

En utilisant le système de détection, on filtre en moyenne 33% des données. La quantité rejetée dépend fortement de la source : pour WIKIPEDIA uniquement 3% des données sont rejetées, principalement en provenance d'articles traitant de langues exotiques, ou de citations en grec ancien ; 68% du texte des BLOGS luxembourgeois est rejeté comme n'étant pas du luxembourgeois, alors même que nous n'avions retenu que 90 des 400 blogs répertoriés, afin de ne conserver que les blogs contenant effectivement une quantité significative de données en luxembourgeois. Enfin 27% des textes en provenance de la CHAMBRE ont été rejetés : au-delà des passages en français,

Modèles	nb. contextes	nb. heures	WER(%)
amorçage (Ang)	63	200h	79,9
amorçage (Fra)	63	300h	76,1
amorçage (All)	63	52h	68,9
amorçage (pooled)	63	552h	70,7
Non-Sup1	7k	80h	33,6
Non-Sup1	22k	80h	33,2
Non-Sup2	31k	193h	29,3
Non-Sup3	39k	500h	28,4
Non-Sup3, mlp	39k	500h	27,43
Non-Sup4, mlp	50k	1200h	25,6

TABLE 2 – Modèles acoustiques du luxembourgeois

présents dans les débats, ce taux élevé est dû à la présence de rapports écrits en français parmi les transcriptions. Après filtrage, 34 millions de mots ont été conservés à partir des presque 52 millions des textes bruts.

Lexique et modèle de langage Les textes filtrés ont été utilisés pour construire un lexique et des modèles de langage. Pour la liste de mots, les 200k mots les plus probables en provenance ces 7 sources décrites précédemment ont été sélectionnés, de manière à minimiser la perplexité des unigrammes sur un texte de développement. Un taux de mots hors vocabulaire (MHV) 2,35% a ainsi pu être obtenu, significativement plus bas que dans des expériences précédentes (Adda-Decker *et al.*, 2008b). Le dictionnaire de prononciation a été obtenu à partir de cette liste de mots, où les prononciation ont été obtenues à l'aide d'un programme de transformation graphème-phonème, puis corrigées à la main. été En ce qui concerne le modèle de langage, le meilleur 3-gramme obtenu par interpolation des 3-gramme calculés sur les 7 sources obtient un perplexité de 369 sur le jeu de développement.

3.3 Modèles acoustiques et taux d'erreurs

La table 2 résume les données audio utilisées lors des itérations successives pour l'apprentissage non supervisé des modèles acoustiques, avec leur taille respective. La dernière colonne donne le taux d'erreur de mots (*Word Error Rate, WER*) sur le jeu de développement Quaero. La partie supérieure de la table donne les WER obtenus avec les 4 modèles d'amorçage appris sur des données de taille significativement plus importante que celle utilisée utilisée dans la section 2. Bien que le modèle allemand obtienne un taux légèrement plus bas, et par souci de cohérence, nous avons utilisé le modèle *pooled* comme modèle d'amorçage.

Dans les 3 premières itérations, des modèles utilisant des paramètres PLP0 ont été construit, alors que le dernier décodage a utilisé des paramètres MLPPLP0 (les paramètres MLP du système allemand ont été utilisés pour le système luxembourgeois) (Fousek *et al.*, 2011). A chaque itération, la quantité de données audio non-transcrites utilisée pour l'apprentissage a été doublée.

Les modèles sont structurellement similaires à ceux décrits à la section 2. Les modèles de phone

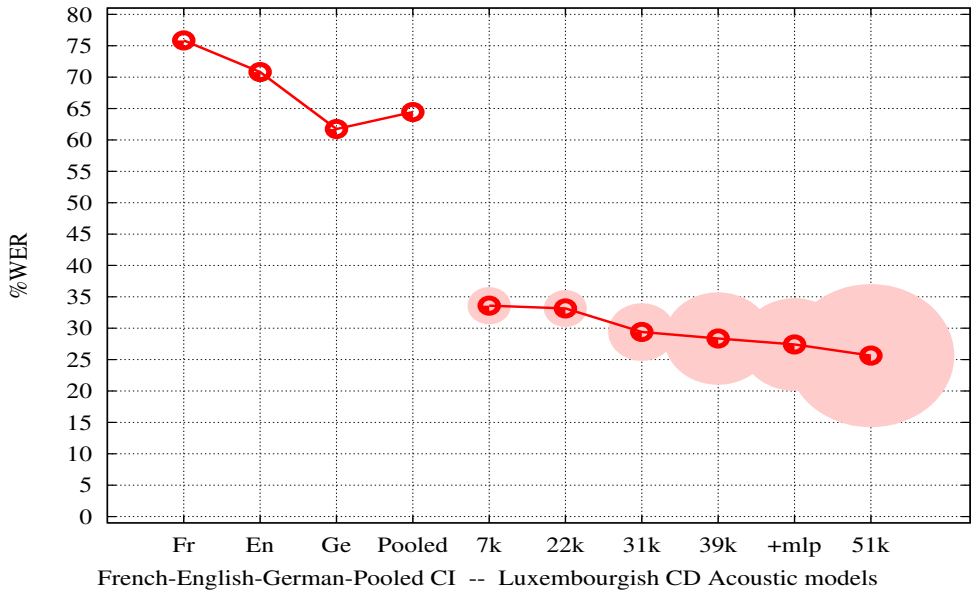


FIGURE 3 – Taux d'erreurs de mots pour les modèles acoustiques non-natifs (gauche) et natifs (droite) la taille du corpus d'apprentissage est représenté par un disque autour du point de résultat.

en contexte triphone sont indépendants du mot, mais dépendants de la position dans le mot et indépendants du genre, étant donné que des expériences précédentes sur l'apprentissage non-supervisé de modèles acoustiques, ont montré que les modèles ainsi obtenus avaient des résultats identiques qu'ils dépendent du genre ou non.

La figure 3 synthétise les taux d'erreurs obtenus. La partie gauche montre les taux pour les modèles non-natifs (anglais, français, allemand, multilingue mutualisé) qui sont compris entre 60% et 75% d'erreurs-mots. La partie droite montre le même type de résultats, mais avec des modèles natifs luxembourgeois obtenus par apprentissage non-supervisé (sélection d'une partie du corpus d'apprentissage luxembourgeois transcrit automatiquement avec les modèles multilingues). Différents jeux de modèles contexte-dépendants luxembourgeois (de 7k à 51k contextes) ont été estimés, le nombre de contextes augmentant progressivement avec la quantité de données sélectionnés pour l'apprentissage. Ce volume d'apprentissage croissant est représenté dans la figure par le diamètre des disques autour du résultat correspondant.

4 Conclusion et perspectives

Le travail présenté ici décrit la construction de modèles acoustiques pour le luxembourgeois, une langue peu dotée ayant subi de fortes influences germaniques et romanes. Nous avons tout d'abord exploré les modèles d'amorçage possibles, et défini un inventaire phonémique que nous avons mis en correspondance avec les inventaires des 3 principales langues exogènes présentes au Luxembourg (allemand, français et anglais). Ensuite , différents modèles d'amorçage ont été construits, monolingues ou multilingues (par concaténation des données d'apprentissage). Ces modèles ont été utilisés lors d'alignements forcés de données audio en luxembourgeois. L'identité du modèle utilisé lors de l'alignement donne des indications sur les correspondances phonémiques

inter-langues, et les modèles acoustiques. Les taux d'utilisation du modèle *pooled* donnent une mesure indicative de l'écart entre les réalisations acoustiques d'une paire donnée de phonèmes dans la langue source et la langue cible. Par exemple, on observe une correspondance importante entre le luxembourgeois et l'anglais pour le phonème /ɛ͡ɪ /. Comme perspective, nous comptons utiliser cette mesure pour des applications d'apprentissage de langues, en mettant en avant pour une paire L1/L2 donnée, la liste des phonèmes potentiellement difficile. Enfin, nous avons utilisé ces modèles d'amorçage pour construire des modèles acoustiques en luxembourgeois à partir d'une très grande quantité de données audio non-transcrites par des méthodes d'apprentissage non-supervisé et un modèle de langage du luxembourgeois appris sur des textes issus du Web et filtrés de manière à éliminer les textes d'une autre langue. Un taux d'erreurs-mots de 25.6% a été obtenu sur le jeu officiel de développement de Quaero.

Références

Adda-Decker, M., Barras, C., Adda, G., Paroubek, P., de Mareïl, P. B. et Habert, B. (2008a). Annotation and analysis of overlapping speech in political interviews. *In LREC*. European Language Resources Association.

Adda-Decker, M., Lamel, L. et Adda, G. (2011a). A first lvcsr system for luxembourgish, an under-resourced european language. *In Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (L&TC 2011)*, pages 47–50, Poznan, Poland.

Adda-Decker, M., Lamel, L. et Snoeren, N. (2010). Comparing mono- & multilingual acoustic seed models for a low e-resourced language : a case-study of Luxembourgish. *In InterSpeech'10, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Japan.

Adda-Decker, M., Pellegrini, T., Bilinski, E. et Adda, G. (2008b). Developments of letzebuergesch resources for automatic speech processing and linguistic studies. *In LREC*.

Adda-Decker, M., Snoeren, N. et Lamel, L. (2011b). Studying Luxembourgish phonetics via multilingual forced alignments. *In ICPhS'11, 17th International Congress of Phonetic Sciences*, Hong Kong, China.

Fousek, P., Lamel, L. et Gauvain, J.-L. (2011). Combining mlp and plp features for speech transcription. *In Handbook of natural language processing and machine translation : DARPA global autonomous language exploitation*, pages 408–416. Springer.

Lamel, L., Gauvain, J. et Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16(1):115–229.

Lavergne, T., Cappé, O. et Yvon, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

Oparin, I., Lamel, L. et Gauvain, J.-L. (2013). Rapid development of a latvian speech-to-text system. *In ICASSP*, pages 7309–7313. IEEE.

Schanen, F. (2004). *Parlons Luxembourgeois*. L'Harmattan.

Schultz, T. et Waibel, A. (2001). Experiments on cross-language acoustic modeling. *In Proceedings of Eurospeech*, Aalborg.

Caractérisation et classification automatique des modes phonatoires en voix chantée

Léonidas Ioannidis Jean-Luc Rouas Myriam Desainte-Catherine
LaBRI - CNRS UMR 5800, Univ. Bordeaux, France
leonidas.ioannidis@labri.fr, jean-luc.rouas@labri.fr,
myriam@labri.fr

RÉSUMÉ

Dans cette étude, nous présentons une analyse de paramètres de la voix pouvant permettre de caractériser et détecter automatiquement les modes phonatoires de la voix chantée. L'objectif est d'étudier les différences dans le contenu spectral entre les types de phonation et de valider leur capacité à les classer automatiquement avec des méthodes d'apprentissage. Cela revêt un intérêt lorsque l'on cherche à caractériser la qualité vocale du chant car le mode phonatoire est un des principaux éléments. De nombreuses applications sont envisageables au niveau de l'indexation et de la recherche d'information dans les signaux audio. Nous présentons ici une méthode basée sur des méthodes d'apprentissage automatique en utilisant un ensemble de paramètres acoustiques et une base de données adéquate. Cette méthode nous permet d'obtenir des résultats de classification des quatre modes phonatoires principaux avec une F-mesure supérieure à 0.8.

ABSTRACT

Caractérisation and automatique classification of phonation modes in singing voice

In this study we present an analysis of voice related features that can help in a classification task to automatically detect and characterize phonation modes found in singing voice. The goal here is to examine the differences in spectral content between different phonation modes, and validate their ability to classify the phonation using machine learning techniques. This can be of interest when characterising the singing voice quality, since phonation is a essential part of the overall quality. It can have numerous applications for indexing and retrieving audio using content based search. We present here a method based on machine learning training, using an acoustic feature set on an audio library. This method is able to classify four widely used phonation modes with a F-measure above 0.8.

MOTS-CLÉS : Modes phonatoires, Voix chantée, Qualité vocale.

KEYWORDS: Phonation Modes, Singing Voice, Voice quality.

1 Introduction

La voix chantée est un des instruments de musique les plus utilisés. C'est pour cela que ses attributs sont étudiés d'une manière différente de ceux de la voix parlée (Sundberg, 1987). Bien que partageant le même appareil phonatoire, la voix parlée et la voix chantée l'utilisent de manière différente et elles exigent différentes démarches que ce soit au niveau de l'analyse ou

de la synthèse. La voix humaine est un instrument capable de produire une gamme très grande de sons. Dans le cadre de la voix chantée, cette variété est souvent exploitée comme un moyen d'expressivité, comme un violoniste peut pincer ou tirer avec l'archet afin de produire une couleur sonore différente. Cette variabilité est rendue possible par un ensemble de muscles qui pouvant se contracter ou se dilater (Sundberg, 1995).

La phonation est le sujet d'étude de cet article. Elle est définie comme étant le son produit par les vibrations des cordes vocales. Ce son peut varier selon la dynamique de vibration de la glotte et le contrôle de la musculature qui l'entoure.

La phonation est aussi perceptivement distinguée par des auditeurs entraînés (Sofranko et Prosek, 2012). Dans des études précédentes, des techniques permettant d'identifier les modes phonatoires ont été proposées, mais elles utilisent toutes des méthodes intrusives, telles que les mesures aérodynamiques, la résistance laryngée et l'électroglottographie (Grillo et Verdolini, 2008). Chacune de ces méthodes offre des performances variables. Cependant le problème est loin d'être considéré résolu. En dehors des considérations sur l'efficacité de ces méthodes, la classification de données déjà enregistrées est une nécessité qui ne peut pas être traitée par les méthodes intrusives. Des travaux ont été menés dans cette direction (Proutskova *et al.*, 2012b). Les auteurs de cette étude proposent une méthode basée sur la décomposition source-filtre, afin d'estimer la contribution de la source, et une classification dépendant de la voyelle prononcée, basée sur des descripteurs extraits du signal source. Cependant, les inconvénients d'une déconvolution inverse source-filtre appliquée à la voix chantée sont connus. Cette opération est sujette à des estimations erronées du signal source à cause de sa structure harmonique parcimonieuse, spécialement dans le cas des notes de haute fréquence (Alku, 2011). Pour pallier ces erreurs, il est possible d'obtenir une meilleure estimation de la source en utilisant des connaissances a priori sur le signal, mais la méthode devient alors inadaptée pour des sons inconnus. Les travaux précédents ont montré ces inconvénients et ont tenté de les contourner en proposant un nouveau descripteur basé sur la transformation en ondelettes pour mesurer l'amplitude sur différentes bandes de fréquences, afin de capturer la pente spectrale (Kane et Gobl, 2011).

Dans cet article, nous proposons un cadre de classification automatique des modes phonatoires avec l'extraction des descripteurs audio sans estimation préalable de la source glottique ni aucun a priori sur le signal. De par la nécessité croissante d'indexer des sons dans de grandes bases de données, ce travail est utile pour l'analyse sémantique de la voix chantée.

Dans la section 2, une explication détaillée des modes phonatoires est présentée. Dans la section 3, la base de données utilisée et les descripteurs acoustiques que nous utilisons sont expliqués et présentés. Dans la section 4, les expériences de classification automatique sont présentées et leurs résultats sont commentés. Finalement, des conclusions et des perspectives sont données dans la section 5.

2 Les modes de phonation de la voix chantée

Dans cette section, nous décrivons les modes phonatoires tels que définis par Sundberg. Selon lui, 4 modes phonatoires existent. D'un point de vue physiologique, ils résultent d'une différence dans les pressions sous-glottique et de la force appliquée pendant la phonation en contraction et dilations de la glotte. Pendant la phonation soufflée, les plis vocaux restent partiellement ouverts

pendant la vibration.Les muscles qui contrôlent l'adduction des plis vocaux sont détendus, ce qui permet à l'air de s'écouler à travers de glotte. Ceci donne une voix semblable au chuchotement (Sundberg, 1987). La voix modale, qui est la plus fréquente, correspond à une vibration régulière des plis vocaux tout au long de leur surface avec une feremeture complète à chaque période. Le debit d'air est notament moindre que dans le cas de la voix soufflée. Une voix serrée est produite quand une tension forte est appliquée à la musculature qui entoure le larynx. L'ensemble du larynx est alors soumis à une tension et le haut du larynx va se contracter fortement tandis que les plis ventriculaires pressent aussi sur les surfaces hautes des plis vocaux, rendant leur vibration inefficace et le debit d'air devient plus petit. (Sundberg, 1995). La phonation "Flow", aussi connue comme voix résonante, est une phonation assez fréquente dans la voix lyrique occidentale. Elle est caractérisée par une pression sous-glottique plus faible que la voix serrée et une faible force d'adduction aux plis vocaux (Titze, 2001). Elle permet d'obtenir des niveaux sonores comparables à la voix serrée mais autorise le chanteur a mieux contrôler son débit d'air. Cela contribue à en faire une voix "économique" comme la voix modale (Proutskova *et al.*, 2012a). Elle est très utilisée par les chanteurs lyriques car elle leur permet d'être entendus malgré la présence d'un ensemble orchestral important, en augmentant la puissance spectrale autour de 3KHz. Ce phénomène est aussi appelé "le formant du chanteur" (Sundberg, 1987). Des termes synonymes ou équivalents peuvent être trouvés dans la littérature. Par exemple "whisper" (chuchotée), "neutre", "harsh" au lieu de serrée. Mêmes s'ils ne sont pas toujours réellement équivalents, nous les mentionnons car ils sont relativement usités. Pour éviter les confusions, nous allons cependant pour cet article uniquement utiliser les termes que nous avons définis.

Les mécanismes de la voix sont très flexibles et nous permettent de produire une gamme des sons très variée. Lorsque l'on parle de mode phonatoires, nous devons aussi penser que ces modes sont pas toujours bien distincts et que l'on pourrait les positionner dans un espace continu, où à un extrême se trouverait la voix soufflée et à l'autre la voix serrée. Ainsi, vers le milieu de cet espace, nous trouverions la voix modale et "flow" mais également tous les cas intermédiaires. Les frontières ne sont pas strictement definies (Proutskova *et al.*, 2012a).

Pour analyser le contenu spectral de la voix chantée, il est nécessaire d'utiliser des outils adaptés aux modes de production humaine. Le modèle source-filtre (Fant, 1981) est souvent employé. Dans ce cadre, la production est modélisée en deux parties. La source, représentant la production du signal sortant de la glotte, et le filtre, qui est l'implémentation du filtrage effectué par la cavité vocale et les résonateurs. La démarche inverse est appelée filtrage-inverse. Après une estimation de l'enveloppe spectrale, en général en utilisant des coefficients de prédiction linéaire (Linear Predictive Coding -LPC), permettant d'obtenir une approximation de la cavité vocale, le signal est filtré pour obtenir une estimation de la source (Alku, 2011). Un autre modèle pouvant être utilisé pour la modélisation de la voix est le modèle sinusoïdal présenté par (McAulay et Quatieri, 1986). Dans ce modèle, les caractéristiques spectrales du signal sont approximées par des composantes sinusoïdales.

3 Base de données et Descripteurs

3.1 La base de données

Nous utilisons une base de données – distribuée librement – construite spécifiquement pour étudier les modes phonatoires et l'estimation de la source glottique. Le manque de données de référence et des données d'apprentissage en nombre suffisant ont conduit les auteurs à l'élaboration de cette base, constituée d'enregistrements de voix chantée étiquetés selon le mode phonatoire. Le fait d'utiliser cette base nous permet d'évaluer les résultats de notre méthode et permettra d'effectuer des comparaisons avec d'autres approches. Cette base est constituée de 790 enregistrements courts de 9 voyelles différentes que l'on peut trouver dans la langue russe. Ces voyelles sont toutes chantées avec les quatre modes phonatoires. Pour chaque mode, la totalité de l'étendue vocale de la chanteuse a été enregistrée lorsque cela était possible. Les auteurs notent cependant que pour les modes "Flow" et serrée, il n'était pas possible de produire des sons au delà d'une certaine fréquence car la phonation devenait alors instable. Tous les enregistrements ont été effectués par une seule chanteuse de tessiture soprano. La durée moyenne de chaque enregistrement est de 1.34 secondes, variant de 0.9 à 1.6 secondes. La base de données est fournie avec des méta-données indiquant la fréquence, la voyelle et le mode phonatoire exprimé dans chaque fichier. Les détails de la procédure d'enregistrement et des informations techniques sur la base de données peuvent être trouvés dans (Proutskova *et al.*, 2012a). Les enregistrements sont effectués avec un taux d'échantillonage de 96kHz. Avant d'effectuer nos traitements, nous avons sous-échantillonné l'ensemble des sons à 16kHz. C'est un taux d'échantillonage adéquat pour le traitement de la voix puisque 8kHz suffisent pour capturer le contenu spectral de la voix chantée. Enfin, nous n'utilisons qu'un extrait de 500ms situé au milieu de chaque fichier afin d'éliminer automatiquement les silences et les parties transitoires au début et à la fin de la phonation.

3.2 Descripteurs

Une analyse spectrale est conduite avant de choisir un ensemble final de descripteurs acoustiques permettant de modéliser les modes phonatoires. La pente spectrale est un des paramètres les plus pertinents pour différencier les quatre modes phonatoires (Kane et Gobl, 2011). Il est également connu que la différence entre les deux premières harmoniques du signal source est un moyen fiable de distinguer les modes soufflés et serrés (Alku, 2011), (Sundberg, 1987). C'est une autre manière de déterminer la rapidité de la chute de l'énergie sur les hautes fréquences et la richesse harmonique du signal. En analysant les spectrogrammes de la Fig. 2, nous pouvons observer des différences entre les modes soufflé et serré. La prédiction linéaire LPC est utilisée pour visualiser les différences de fréquence des formants et leurs amplitudes selon les modes phonatoires. Dans ce cas, la comparaison est effectuée pour la même voyelle, puisque les valeurs des fréquences formatiques dépendent fortement de l'articulation. Les fréquences, les amplitudes et les bandes passantes des quatres premiers formants sont calculées. Dans la Fig. 1, nous observons que ces amplitudes varient pour chaque mode (même voyelle et même fréquence).

Nous pouvons observer que les formants sont sujets à des déplacements pour les modes serrée et soufflé. Cela peut être dû à l'étirement du larynx. Une pression sous-glottique plus importante peut également être responsable de changements dans les fréquences des formants (Sundberg,

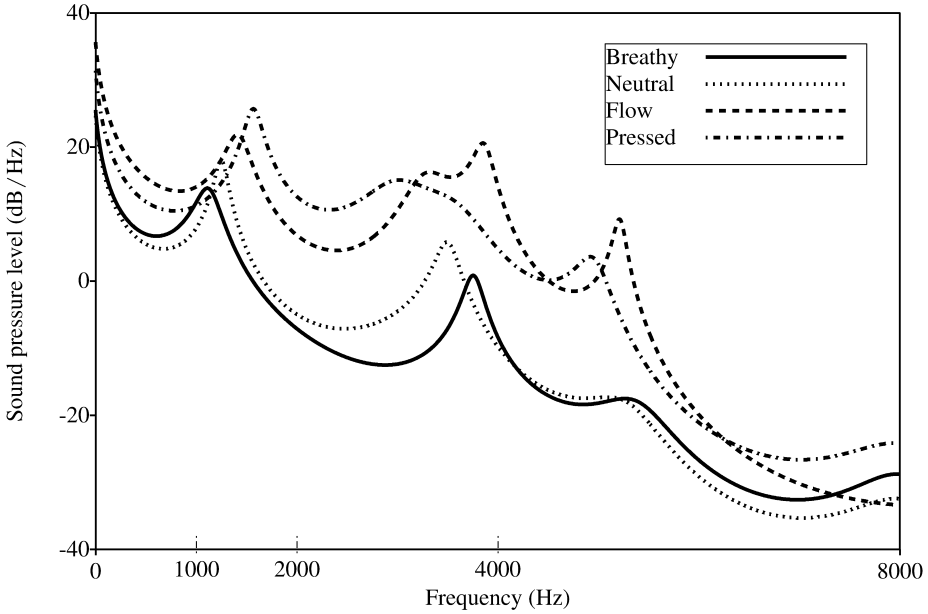


FIGURE 1 – Analyse LPC de la voyelle /a/ pour différents types de phonation.

1987). Toutefois, en considérant que l'activité musculaire diffère pour chaque mode phonatoire, la forme du conduit vocal est alors également modifiée, ce qui se reflète par des réponses fréquentielles différentes. Dans la Fig. 1, nous observons l'espace typique F1-F2 où les voyelles */a/, /u/* et */i/* sont produites. Nous pouvons observer que la phonation "normale" occupe plus d'espace que les autres modes. Les modes serré et "flow" partagent des caractéristiques similaires, notamment un déplacement particulièrement important de la fréquence de F1 pour les voyelles */a/* et */i/*.

La richesse harmonique observée dans les modes serré et "flow" peut également être capturée en terme de Rapport Harmonique sur Bruit (HNR). Ce paramètre est communément utilisé dans des scénarios d'identification automatique d'instruments de musique, où certaines familles d'instruments peuvent avoir des excitations plus importantes dans les harmoniques élevées. Le HNR est mesuré par une technique basée sur le cepstre habituellement utilisé pour les signaux de voix parlée (de Krom, 1993). Dans cette méthode, la taille de la fenêtre d'analyse est fixée comme étant égale à 5 périodes du fondamental. La mesure de HNR est alors effectuée en prenant le composant fréquentiel du cepstre et en comparant le niveau d'énergie au seuil du bruit. Les mesures sont prises sur différentes bandes de fréquences : 0-0.5KHz, 0-1.5KHz, 0-2.5KHz et 0-3.5KHz. La fréquence fondamentale est calculée sur l'ensemble des données grâce à l'algorithme STRAIGHT (Kawahara *et al.*, 1998). En application une analyse FFT et étant donné le F0 calculé, les trois premières harmoniques H1 H2 et H3 sont extraites et leurs amplitudes normalisées sont calculées en dB ainsi que les différences H1-H2 et H2-H4. De plus, les différences entre l'amplitude du fondamental et l'amplitude des formants sont également calculées : H1-Amp1, H1-Amp2 et H1-Amp3. Enfin, nous ajoutons à l'ensemble de ces paramètres le pic de proéminence cepstral (Cepstral Peak Prominence - CPP), un paramètre qui a été élaboré pour la détection de voix soufflée et qui est basé sur une mesure de périodicité du ceptre (Hillenbrand *et al.*, 1994). L'ensemble de ces paramètres est calculé sur des fenêtres Hanning de 25ms avec recouvrement toutes les 5 ms. Au final, la moyenne de chaque paramètre est calculée sur chaque fichier, résultant en un vecteur de 24 paramètres par fichier. Dans le tableau 1 l'ensemble des

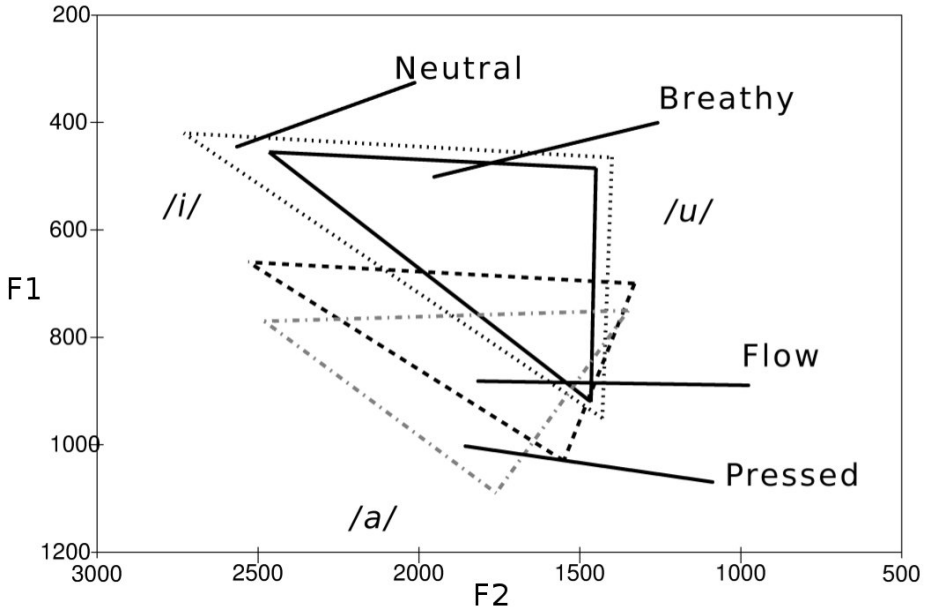


FIGURE 2 – Espace occupé dans le plan F1-F2 par chaque mode phonatoire pour les voyelles /a/, /i/, /u/.

descripteurs sont donnés.

ind.	nom	Desc.	ind.	nom	Desc.
1	H1	ampl. 1 harm.	13	HNR05	HNR (0-0,5KHz)
2	H2	ampl. 2 harm.	14	HNR15	HNR (0-1,5KHz)
3	H4	ampl. 3 harm.	15	HNR25	HNR (0-2,5KHz)
4	A1	ampl. 1 formant.	16	HNR35	HNR (0-3,5KHz)
5	A2	ampl. 2 formant.	17	F1	1 Formant
6	A3	ampl. 3 formant.	18	F2	2 Formant
7	H1H2	Difference H1 - H2	19	F3	3 Formant
8	H2H4	Difference H2 - H4	20	F4	4 Formant
9	H1A1	Difference H1 - A1	21	B1	1 Bandwidth
10	H1A2	Difference H1 - A2	22	B2	2 Bandwidth
11	H1A3	Difference H1 - A3	23	B3	3 Bandwidth
12	CP	Cepstral Peak Prominence	24	B4	4 Bandwidth

TABLE 1 – Recupulatif des descripteurs choisis

4 Classification

Après avoir collecté les données, nous apprenons un modèle pour chaque mode phonatoire en utilisant des techniques d'apprentissage automatique. Le logiciel Weka et la bibliothèque Libsvm ont été utilisés dans ces expériences.

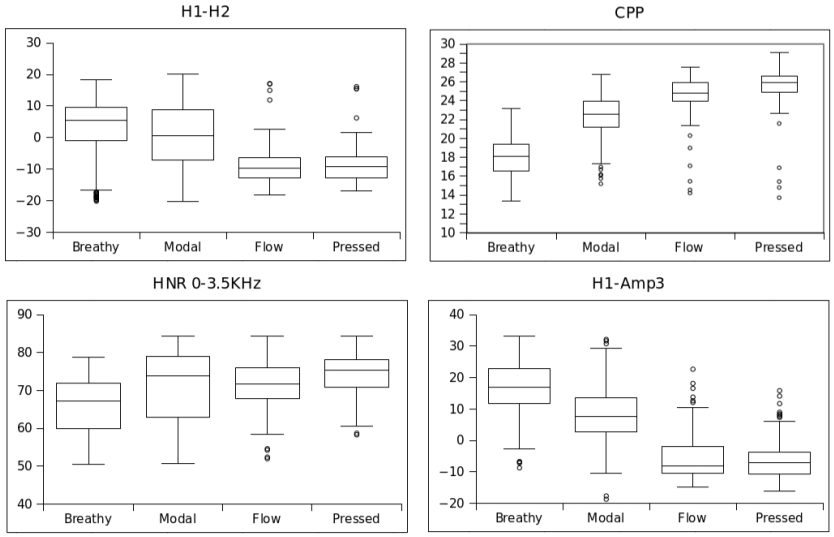


FIGURE 3 – Statistiques pour quatre paramètres. Toutes les valeurs sur l'axe Y sont en dB

4.1 Apprentissage automatique

Deux méthodes sont utilisées pour modéliser les modes phonatoires : les "Logistic Model Trees" (LMT) et les Machines à Vecteurs Support (SVM). Les SVM sont populaires pour la classification de motifs audio, notamment pour les tâches de classification supervisées comme dans le cas présent. Les SVM atteignent de bonnes performances pour la classification multiple et peuvent traiter des vecteurs de paramètres de dimensions importantes. La méthode consiste à recherche un hyperplan qui permet de séparer linéairement les classes avec un critère de marge maximale. Afin d'éviter les problèmes de généralisations, nous effectuons une validation croisée en 10 itérations permettant d'obtenir des résultats moins biaisés que l'approche classique apprentissage-test. Les performances des deux méthodes de classification sont évaluées par le biais des mesures de précision et rappel, ainsi que la F-mesure qui est une combinaison de ces deux mesures. Parmi les deux méthodes employées, ce sont les LMT qui permettent d'obtenir les meilleurs résultats en termes de précision et de F-mesure. Les LMT sont des arbres de classification qui utilisent la régression logistique sur leur feuilles. Les SVM obtiennent des résultats à peu près similaires, mais avec un temps de calcul moindre par rapport aux LMT, ce qui est intéressant d'un point de vue performance vs. temps de calcul. Les résultats détaillées sont présentés dans la section suivante.

SVM			**LM trees**			
Pr.	Re.	F-m.	Pr.	Re.	F-m.	Class
0.87	0.921	0.894	0.918	0.933	0.925	BR
0.857	0.845	0.851	0.903	0.907	0.905	NE
0.721	0.676	0.698	0.727	0.717	0.722	FL
0.712	0.702	0.707	0.725	0.709	0.717	PR
0.807	0.81	**0.808**	0.84	0.842	**0.841**	Av.

TABLE 2 – Resultats de classification automatique avec SVM et LMT

4.2 Results

La précision maximale (84.15%) est atteinte en utilisant les LMT. Les résultats analytiques sont données dans le tableau 2. Nous pouvons observer que les models soufflé et modal (NE) sont les mieux classés. D'autre part, le modes "flow" (FL) et serrée (PR) ne sont pas très bien distingués. La matrice de confusion obtenue avec la classification par LMT est donnée dans le tableau 3. Nous pouvons observer dans ce tableau que les vrais positifs pour les modes soufflé et modal sont nombreux et qu'une faible confusion est faite entre eux, alors qu'il sont bien distingués des autres modes. Le même phénomène peut être observé pour les modes "flow" et serré, cependant cette fois la confusion entre ces deux classes est plus grande.

BR	NE	FL	PR	←Classified as
223	16	0	0	BR
18	**205**	1	2	NE
0	5	**104**	36	FL
2	1	38	**100**	PR

TABLE 3 – Confusion matrix for the LMT classification experiment.

5 Discussion and Perspectives

La qualité vocale dans la voix chantée est un sujet de recherche actuel à cause de son importance dans la classification par le contenu de la musique. La phonation joue un rôle prépondérant dans la perception globale de la qualité vocale. Nous avons présenté ici une méthode pour la classification automatique des modes phonatoire soufflé, model, "flow" et serrée en employant des caractéristiques acoustiques extraites directement du signal. Les performances mesurées sont relativement élevées, particulièrement pour les deux premières classes qui semblent plus faciles à distinguer avec l'ensemble de paramètres que nous avons utilisé. Les confusions observées entre les modes "flow" et serré démontrent cependant que l'utilisation de méthodes plus sophistiquées est nécessaire pour améliorer les résultats. C'est possible que cette confusion soit à cause de l'énergie spectrale qui est plus élevée que dans les autres classes. Notre ensemble de descripteurs n'est pas capable de capter des différences qui sont davantage détectables dans le signal de la source. Nous envisageons d'étendre ce travail en ajoutant à notre jeu de paramètres des mesures d'instants d'ouverture et de fermeture glottique, ainsi que leur quotient. Les signaux glottaux semblent en effet utiliser des motifs différents selon la phonation. De plus, nous devons examiner la pertinence de notre méthode, mise au point sur une base de voyelles prononcées isolément, pour classer et caractériser la voix chantée lors d'enregistrements continus. Il est aussi important de faire ce travail sur d'autres sujets et pas seulement chez les sopranos. Enfin, il est important d'évaluer dans quelle mesure des sujets humains sont capables de distinguer les différents modes phonatoires, en utilisant la même base de données que pour notre expérience de classification automatique. Il a en effet été démontré que selon le niveau d'expérience ou d'entraînement, nous ne possédons pas tous de manière égale la faculté de distinguer la qualité vocale (Henrich et Sotiropoulos, 2007), (Sofranko et Prosek, 2012). Il sera intéressant de comparer les résultats de classification automatique avec la perception humaine.

6 Remerciements

Ce travail est partiellement soutenu par le projet ANR DIADEMS ANR-12-CORD-0022.

Références

ALKU, P. (2011). Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(October):623–650.

de KROM, G. (1993). A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals. *Journal of Speech, Language, and Hearing Research*, 36(2):254–266.

FANT, G. (1981). The source filter concept in voice production. *Speech, Music and Hearing Quarterly Progress and Status Report*, 22(1):021–037.

GRILLO, E. U. et VERDOLINI, K. (2008). Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects. *Journal of voice : official journal of the Voice Foundation*, 22(5):546–52.

HENRICH, N. et SOTIROPOULOS, D. (2007). Characterisation of Voice Quality in Western Lyrical Singing : from Teachers ' Judgements to Acoustic Descriptions. *Journal of Interdisciplinary Music Studies*, 1(2):62–91.

HILLENBRAND, J., CLEVELAND, R. A. et ERICKSON, R. L. (1994). Acoustic Correlates of Breathy Vocal Quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778.

KANE, J. et GOBL, C. (2011). Identifying regions of non-modal phonation using features of the wavelet transform. *INTERSPEECH*, pages 177–180.

KAWAHARA, H., MASUDA-KATSUSE, I. et CHEVEIGN, A. D. (1998). Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds.

MCAULAY, R. et QUATIERI, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 34(4):744–754.

PROUTSKOVA, P., RHODES, C. et CRAWFORD, T. (2012a). Breathy or Resonant a Controlled and Curated Database for Phonation Mode Detecion in Singing. *Internasional Symposium on Music Information Retrieval*, (Ismir):589–594.

PROUTSKOVA, P., RHODES, C., CRAWFORD, T. et WIGGINS, G. (2012b). Breathy , Resonant , Pressed - Automatic Detection Of Phonation Mode From Audio Recordings of Singing. *Journal of New Music Research*, pages 1–28.

SOFRANKO, J. L. et PROSEK, R. a. (2012). The effect of experience on classification of voice quality. *Journal of voice : official journal of the Voice Foundation*, 26(3):299–303.

SUNDBERG, J. (1987). *The science of the singing voice*. Northern Illinois University Press.

SUNDBERG, J. (1995). Vocal fold vibration patterns and phonatory modes. *Folia Phoniatrica Logopedica*, 47:218–228.

TITZE, I. R. (2001). Acoustic interpretation of resonant voice. *Journal of voice : official journal of the Voice Foundation*, 15(4):519–28.

Reconnaissance automatique de la parole à l'aide de colonies de fourmis

Benjamin Lecouteux Didier Schwab
Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole
Laboratoire d'Informatique de Grenoble
Univ. Grenoble Alpes
benjamin.lecouteux@imag.fr, didier.schwab@imag.fr

RÉSUMÉ

Cet article présente une approche originale permettant de décoder des graphes issus d'un Système de Reconnaissance Automatique de la Parole (SRAP) à l'aide d'un algorithme constructif : les colonies de fourmis. L'application d'un modèle de langage d'ordre supérieur à un graphe nécessite son extension afin de construire des historiques correspondants à chaque nouvel n-gramme observé. Cette extension peut rapidement engendrer des calculs lourds et une consommation de mémoire conséquente. Nous proposons une alternative où des colonies de fourmis explorent le graphe avec un nouveau modèle de langage sans la nécessité d'effectuer une extension. Nous présentons des premiers résultats encourageants se basant sur le corpus anglais TED où des bi-grammes sont réévalués en quadri-grammes. Finalement, nous discutons sur les atouts de cette approche qui permet d'envisager des décodages massivement parallélisables pour un même graphe ainsi que des contrôles stricts du temps de calcul et de la mémoire.

ABSTRACT

Automatic speech recognition using ant colony algorithm

In this paper we propose an original approach allowing to decode Automatic Speech Recognition Graphs by using a constructive algorithm based on ant colonies. When a graph is decoded with a language model using a higher order ; classical algorithm must to extend graph in order to develop each new observed n-gram. This extension increase time and memory consumption. We propose to use ant colony algorithm in order to explore ASR graphs with a new language model, without the necessity to extend it. We present first results based on TED english corpora where graphs are decoded with 4-grams. Finally, we show that our approach allows one to provide higly threaded decoding and a control of time and memory consumption.

MOTS-CLÉS : Décodage de graphe, colonies de fourmis, modèle de langage.

KEYWORDS: Graph decoding, ant colony algorithm, language model.

1 Introduction

Dans cet article nous présentons un travail préliminaire introduisant un nouveau paradigme permettant d'explorer les graphes issus d'un système de reconnaissance automatique de la parole (SRAP). Un problème rencontré lors de l'exploration des graphes de mots est l'application sur ces derniers d'un modèle de langage d'ordre supérieur. En effet, le nombre de chemins croît avec l'ordre du modèle de langage. Ainsi un graphe de mots initialement décodé avec un modèle bi-gramme qui est étendu avec un modèle quadrigramme peut voire sa taille multipliée par plus de 10. Plusieurs techniques ont été appliquées avec succès pour approcher (*compact-expansion*) ou effectuer des coupures dans la recherche (*beam-search*). Nous proposons une méthode alternative se basant sur un algorithme constructif très utilisé dans le domaine de la recherche opérationnelle : les colonies de fourmis. Cet algorithme a notamment été appliqué à la résolution du voyageur de commerce ou encore à la désambiguisation lexicale.

L'article s'organise de la manière suivante : dans la section 2 nous faisons un rapide tour d'horizon des techniques de décodage pour les SRAP grand vocabulaire ainsi que de l'algorithme des colonies de fourmis. La section 3 présente l'ensemble du système qui a été utilisé pour effectuer nos travaux. Ensuite nous abordons l'ensemble des expériences menées dans la section 5 avec leur analyse, suivies par une discussion dans la section 6. Enfin nous concluons en proposant quelques perspectives intéressantes liées à l'utilisation de cet algorithme.

2 État de l'art

2.1 Approches classiques

Le principe d'un SRAP est de trouver dans un graphe l'hypothèse qui maximisera les probabilités des modèles linguistiques et acoustiques. Généralement des heuristiques sont appliquées afin de limiter l'espace de recherche. On trouve ainsi dans la littérature plusieurs approches d'exploration dynamique pour générer les hypothèses :

— Les algorithmes de graphes synchrones qui génèrent une copie virtuelle du graphe pour l'ensemble des hypothèses se finissant à un temps t. Le mot suivant est partagé, mais associé à plusieurs historiques. Dans cette approche l'algorithme dynamique d'alignement est appliqué pour tout nouveau mot sans avoir à conserver d'historique. La mise en oeuvre de cet algorithme s'avère rare (Ortmanns et Ney, 2000).

— Les algorithmes de graphes réentrants où une copie virtuelle du graphe est explorée pour chaque contexte linguistique. L'information relative à chaque contexte est enregistrée pour chaque chemin et combinée avec la nouvelle racine virtuelle dépendant de l'historique relatif au modèle de langage (Ney *et al.*, 1992). Ces algorithmes sont très répandus et leur implémentation la plus courante est un Viterbi en faisceau (*beam-search*). Leur conception se rapproche de la programmation dynamique.

— Les algorithmes asynchrones à pile (Jelinek, 1969) dont le principe est d'explorer en profondeur et en priorité les hypothèses qui semblent prometteuses. Ceci se fait en étendant mot par mot l'hypothèse sélectionnée. Les réalisation de ce type d'algorithme se base sur des piles ordonnant les hypothèses à explorer. La mise en oeuvre la plus courante est l'implémentation de l'algorithme A^* (Paul, 1991; Nocera *et al.*, 2002).

Les documents (Aubert, 2002; Lecouteux, 2008) présentent en détail les principales techniques de décodage des systèmes de reconnaissance de la parole grand vocabulaire.

2.2 Algorithmes à colonies de fourmis

Les algorithmes à base de fourmis font partie des approches constructives. Dans ces approches, de nouvelles configurations sont générées par ajout itératif d'éléments de solution aux configurations en cours de construction. Ils ont pour origine la biologie et les observations réalisées sur le comportement social des fourmis. En effet, ces insectes ont collectivement la capacité de trouver le plus court chemin entre leur fourmilière et une source d'énergie. Il a pu être démontré que la coopération au sein de la colonie est auto-organisée et résulte d'interactions entre individus autonomes. Ces interactions, souvent très simples, permettent à la colonie de résoudre des problèmes complexes. Ce phénomène est appelé intelligence en essaim (Bonabeau et Théraulaz, 2000).

En 1989, Jean-Louis Deneubourg étudie le comportement des fourmis biologiques dans le but de comprendre la méthode avec laquelle elles choisissent le plus court chemin et le retrouvent en cas d'obstacle. Il élabore ainsi le modèle stochastique dit de Deneubourg (Deneubourg *et al.*, 1989), conforme à ce qui est observé statistiquement sur les fourmis réelles quant à leur partage entre les chemins. Ce modèle stochastique est à l'origine des travaux sur les algorithmes à fourmis.

Le concept principal de l'intelligence en essaim est la stygmergie, c'est-à-dire l'interaction entre agents par modification de l'environnement. Une des premières méthodes que l'on peut apparenter aux algorithmes à fourmis est l'écorésolution qui a montré la puissance d'une heuristique de résolution collective basée sur la perception locale, évitant tout parcours explicite de graphe d'états (Drogoul, 1993).

En 1992, Marco Dorigo et Luca Maria Gambardella conçoivent le premier algorithme basé sur ce paradigme pour le célèbre problème combinatoire du voyageur de commerce (Dorigo et Gambardella, 1997). Dans les algorithmes à base de fourmis artificielles, l'environnement est généralement représenté par un graphe et les fourmis virtuelles utilisent l'information accumulée sous la forme de chemins de phéromone déposée sur les arcs du graphe. De façon simple, une fourmi se contente de suivre les traces de phéromones déposées précédemment ou explore au hasard dans le but de trouver un chemin optimal, fonction du problème posé, dans le graphe. Le lecteur trouvera dans (Dorigo et Stützle, 2004) ou (Monmarche *et al.*, 2009) de bons états de l'art sur la question.

Ce type d'algorithme a déjà été appliqué à des problèmes en traitement automatique du langage, notamment dans le cadre de la désambiguisation lexicale (Schwab *et al.*, 2013). Nos travaux visent à étendre ce paradigme à la reconnaissance automatique de la parole.

3 Le système SRAP

Le système SRAP utilisé est basé sur le *toolkit* KALDI (Povey *et al.*, 2011). KALDI propose la mise à disposition des outils état de l'art pour la RAP. Les modèles acoustiques et les phonétisations de notre système ont été appris avec les données mises à disposition par le LIUM (Rousseau *et al.*, 2012) qui représentent 118 heures de données annotées. L'enrichissement des paramètres par les coefficients delta et delta-delta, ainsi que plusieurs transformations telles que LDA et MLLT

Données	TEM 2-gr	TEM 3-gr	TEM 4-gr
dev 2010 (4h12)	22.01 %	18.86%	17.6%
test 2010 (7h30)	21.67%	17.98%	17.0%

Ordre ML	Nb n-grammes	Perplexité
1g	41K	-
2g	+ 35M	234
3g	+ 238M	159
4g	+ 524M	150

TABLE 1 – Perplexité des différents modèles de langage utilisés par notre système, ainsi que les taux d'erreur mots (TEM) obtenus par notre système de référence.

(Gopinath, 1998) sont appliquées. Enfin, un apprentissage adapté au locuteur (*Speaker Adaptive Training* - SAT) est réalisé et combiné avec une adaptation dans l'espace des paramètres de type fMLLR (Gales, 1998).

Le modèle de langage a quant à lui été appris sur l'ensemble des données fournies pour la campagne IWSLT 2012 ainsi que sur la partie d'apprentissage fournie par le LIUM. Un modèle a été appris pour chaque sous-corpus et l'ensemble des modèles a été interpolé en minimisant la perplexité sur les données de développement, sans appliquer de sélections.

Au final la perplexité obtenue sur le dev est de 159 pour le modèle trigramme et 150 pour le quadrigramme. Les quantités de n-grammes et la perplexité des différents modèles utilisés pour nos expériences sont présentées dans le tableau 1.

L'ensemble des expériences ont été effectuées sur un corpus mis à disposition par la communauté pour la campagne d'évaluation IWSLT 2010 : le corpus TED qui correspond à un ensemble de conférences enregistrées en anglais (Paul *et al.*, 2010). Les résultats du système de référence sont également présentés dans le tableau 1.

3.1 Conversion des graphes

Il est important de noter que KALDI est conçu autour d'une modélisation basée sur des automates à états finis. Dans un premier temps, nos expériences sont menées sur des graphes plus classiques utilisant le formalisme HTK. Nous avons donc converti l'ensemble des graphes issus de KALDI en supprimant les epsilon-transitions afin de travailler avec une représentation HTK.

4 Approches proposées

Dans un premier temps, nous souhaitons valider l'approche en simplifiant certaines étapes du décodage. Pour cette raison, l'ensemble des travaux présenté est réalisé sur des graphes de mots extraits du SRAP KALDI. Nous avons volontairement utilisé un modèle de langage initial bi-gramme afin de ne pas introduire d'information supplémentaire au niveau linguistique. Nous utilisons donc l'algorithme de colonies de fourmis pour étendre le graphe à des ordres de n-gramme supérieurs. Notre implémentation générale de l'algorithme est la suivante : À chaque arc du graphe est associé une variable de type phéromone (un compteur) initialisée à 1 au début de chaque tour. Nos fourmis sont lancées à partir du noeud d'entrée du graphe. Pour sortir du noeud, elles vont devoir choisir un arc. Ce choix est effectué via un tirage aléatoire dont la fonction de distribution dépend de la présence de phéromones sur chaque arc (et éventuellement de facteurs locaux). À l'arrivée de la fourmi sur le noeud final, si la solution trouvée – c'est-à-dire le produit des probabilités

acoustiques et linguistiques du chemin parcouru : $\prod_{p=debutChemin}^{finChemin} P_n(W_p|w_{p-n}..w_{p-1}).P(X_p)$ – est meilleure que la précédente meilleure solution trouvée ; alors des phéromones sont déposées sur le chemin. Dans la littérature, il a été montré qu'avec une version aussi simple des colonies de fourmis, ces dernières ont tendance à converger sur un chemin qui ne représente pas forcément la solution optimale. Il est donc nécessaire d'utiliser des heuristiques locales ou globales pour éviter ce type de phénomène. La sous-section suivante présente quelques heuristiques locales biaisant le choix des fourmis.

4.1 Fourmis anamorphiques

Nous proposons différentes fourmis, comme proposé dans (Bontoux, 2008), qui utilisent des heuristiques d'exploration biaisées par l'information déjà présente dans le graphe initial.

Des probabilités *a posteriori* sont ainsi calculées à partir du graphe bi-gramme. Cette probabilité est calculée pour chaque arc a étant donné le graphe de mot courant G : $p(a|G)$. Elle correspond au ratio entre l'ensemble des probabilités de tous les chemins passant par l'arc a et l'ensemble de tous les chemins dans le graphe G : $p(a|G) = \frac{\sum_{C \in G, a \supset C} p(C|G)}{\sum_{C \in G} p(C|G)}$. Où $C \in G$ désigne un chemin C complet dans le graphe G et $a \supset C$ un arc qui est compris dans ce chemin. Cette probabilité *a posteriori* est calculée classiquement par via un algorithme *forward-backward* comme présenté dans (Wessel *et al.*, 2000).

Nous proposons différents types de fourmis qui sont testés indépendamment :

— Les fourmis à l'écoute (1) : celles-ci sont influencées par la probabilité *a posteriori* acoustique du noeud suivant, sans prendre en compte l'aspect linguistique.
— Les fourmis verbeuses (2) : à l'inverse des fourmis précédentes, ces dernières sont uniquement influencées par la probabilité *a posteriori* linguistique du noeud suivant.
— Les fourmis oracles (3) : c'est un modèle hybride qui combine les deux précédentes et qui représente la probabilité *a posteriori* classique du noeud suivant.

Les différentes probabilités *a posteriori* sont calculées via un algorithme Viterbi appliqué sur le graphe bi-gramme en annulant respectivement les modèles acoustiques (fourmi (1)) et linguistiques (fourmi (2)) ou en conservant le score initial pour la fourmi (3).

Ces différents types de fourmis permettent d'orienter l'exploration en fonction d'informations linguistiques ou acoustiques. Dans la prochaine section, nous présentons les performances de ces fourmis.

Enfin, nous proposons de combiner l'exploration des différentes fourmis en fusionnant leurs phéromones déposées. Pour ce faire, nous faisons travailler chaque type de fourmi sur une copie du graphe. Une fois les graphes explorés, ils sont fusionnés en un seul où les quantités de phéromones sont additionnées. Quelques fourmis sont alors relancées sur ce nouveau graphe et influencées par toutes les phéromones déposées.

Pour l'ensemble des expériences présentées, le nombre de périodes est initialisé à 30 et le nombre de fourmis à 2000, ce qui correspond à 60000 chemins explorés pour chaque graphe. La durée moyenne d'un décodage avec ces paramètres est d'environ 5 secondes par graphe (sur un Intel Xeon 2.5 Ghz). Lors de l'utilisation de plusieurs fourmis, nous avons réalisé 10 périodes avec chacune des fourmis avant de fusionner leurs graphes.

L'algorithme détaillé est présenté ci-dessous.

Algorithme 1 : Algorithme des colonies de fourmis appliqué au décodage de graphes de RAP

```
// 2000 fourmis et 30 périodes dans nos expériences;
Initialiser N_fourmis;
Initialiser N_periodes;
periode ← 0;
fourmis ← 0;
MeilleurScore ← −inf;
MeilleursChemins ← {};
pour tous les Noeuds N ∈ Graphe faire
    Initialiser probabilité a posteriori φ(N);
fin
tant que periode < N_periodes faire
    // Initialisation de l'ensemble des dépôts de phéromones, puis prise en
    considération des précédents meilleurs chemins;
    pour tous les Noeuds N ∈ Graphe faire
        Ph_N ← 1;
    fin
    pour tous les Noeuds N ∈ MeilleursChemins[0...periode] faire
        Ph_N ← Ph_N + 1;
    fin
    fourmi ← 0;
    tant que fourmi < N_fourmis faire
        Chemin ← {};
        Chemin=TrouverChemin(PremierNoeud);
        si Evaluer(Chemin) > MeilleurScore alors
            MeilleurScore ← Evaluer(Chemin);
            MeilleursChemins[periode] ← Chemin;
            pour tous les Noeuds N ∈ MeilleursChemins[periode] faire
                Ph_N ← Ph_N + 1;
            fin
        fin
        fourmi ← fourmi + 1;
    fin
fin
TrouverChemin(Noeud)
    NoeudSuivant = SelectionnerNoeudSuivant(Noeud) ;
    si NoeudSuivant = NoeudFinal alors
        Retourner NoeudSuivant
    fin
    Retourner {NoeudSuivant + TrouverChemin(NoeudSuivant)};
SelectionnerNoeudSuivant(Noeud)
    // Un arc de sortie est sélectionné aléatoirement en fonction d'une
    distribution discrète dépendant de la quantité de phéromones Ph et des
    probabilités a posteriori φ();
    Retourner TirageAleatoire({Liens_Noeud}, {Ph_NoeudSuivant * φ(NoeudSuivant)});
Evaluer(Chemin)
    // P_n(W) est un modèle de langage d'ordre n et P(X) le modèle acoustique;
    Score = Σ_{p=debutChemin}^{finChemin} (logP_n(W_p|w_{p−n}..w_{p−1}) + logP(X_p));
    Retourner Score ;
```

Données	Système de référence	Fourmi (3)	Fourmi (2)	Fourmi (1)	Fourmis (1)+(2)+(3)	Temps de Calcul référence/fourmis
Dev 2g	22.0%	22.0%	31.0%	29.3%	22.0%	1h/1h
Dev 3g	18.8%	20.8%	29.2%	26.4%	20.2%	13h/1h
Dev 4g	17.6%	19.5%	28.7%	26.0%	19.0%	18h/1h
Test 2g	21.6%	21.6%	30.1%	28.7%	21.6%	1h30/1h30
Test 3g	18.0%	20.0%	28.2%	25.9%	19.4%	19h/1h30
Test 4g	17.0%	19.1%	27.6%	25.0%	18.7%	27h/1h30

TABLE 2 – Résultats, en terme de TEM, des décodages avec l'algorithme des fourmis, les décodages avec les fourmis présentent des oscillations de +/− 0.2 point d'une exécution à l'autre.

5 Expériences préliminaires et résultats

Le tableau 2 présente les résultats (en terme de TEM) obtenus avec l'algorithme présenté précédemment. L'expérimentation confirme qu'orienter au départ les fourmis sur les chemins optimaux influence grandement le résultat final. Les fourmis lancées sur des chemins privilégiant l'acoustique (fourmis (1)) ou la linguistique (fourmis (2)) obtiennent de piètres résultats, mais fusionnées avec les fourmis plus hétérogènes (fourmis (3)) elles apportent une information significative.

Ces premières expériences montrent que l'algorithme de colonies de fourmis arrive à étendre le graphe, sans toutefois atteindre les résultats étalons. Par contre, d'un point de vue temps de calcul, l'algorithme basé sur les fourmis est près de 15 fois plus rapide qu'une méthode d'extension classique : la configuration présentée dans le tableau est celle atteignant les scores maximaux ; si des itérations sont rajoutées, les colonies continuent de converger sur le même chemin.

Les résultats en demi-teinte sont normaux étant donné la simplicité de l'algorithme utilisé actuellement. Dans la littérature, il a été montré que les heuristiques permettant au départ de distribuer les phéromones sur les noeuds avaient une importance capitale sur l'évolution de la fourmilière.

Par ailleurs (Dorigo et Gambardella, 1997) montre qu'un des soucis majeur de ce type d'algorithme est sa convergence vers des optimums locaux. Ce type de comportement peut être évité en utilisant plusieurs dimensions qui seront fusionnées *a posteriori*, comme présenté dans (Bontoux, 2008). La fusion des trois types de fourmis introduit cette heuristique et montre son efficacité.

6 Discussion sur les colonies de fourmis

6.1 Avantages

La parallélisation de l'algorithme des colonies de fourmis est triviale. En effet, les fourmis peuvent évoluer indépendamment les unes des autres à la recherche de nouveaux chemins. La parallélisation est également envisageable à un plus haut niveau : celui des colonies de fourmis. Différentes colonies peuvent évoluer sur des copies d'un même graphe puis échanger leurs informations en fusionnant les copies (c'est-à-dire en additionnant la quantité de phéromone présente sur chaque noeud). Un autre point non négligeable est la quantité de mémoire nécessaire. Elle est limitée à l'espace requis pour stocker le graphe en mémoire. Étant donné qu'il n'y a pas d'extension physique, la mémoire utilisée est constante. Dans nos expériences, l'algorithme utilise

en moyenne 100Mo contre plusieurs Go lors d'une extension classique. L'évaluation de la solution se fait lorsqu'une fourmi arrive sur le noeud final du graphe. Le nombre d'évaluations étant relativement faible (2000 par graphe dans notre cas), il est envisageable de rajouter d'autres informations affinant la solution (sémantique, modèle de langage à base de réseau de neurones etc.) Enfin, le temps de calcul peut être parfaitement maîtrisé : le nombre d'évaluations est connu à l'avance et son temps dépend uniquement de la profondeur du graphe.

6.2 Inconvénients

À moins de mémoriser la graine de génération aléatoire, les résultats de l'algorithme ne sont pas déterministes. Au cours de nos expériences, nous avons observé des différences de 0.2 point entre différentes explorations. Rien ne garantit l'optimalité de l'algorithme et ce dernier peut avoir tendance à se laisser piéger par des optimums locaux. Cependant, la littérature propose des solutions pour palier ces problèmes.

6.3 Évolutions envisagées

Le travail présenté ici est préliminaire. Nous souhaitons explorer et analyser plus en profondeur le paradigme des colonies de fourmis appliqué aux SRAP. Les prochaines évolutions se traduiront par l'utilisation en parallèle de fourmis ayant des comportements différents. Ainsi, une sélection des individus performants se fera au fur et à mesure de l'exploration. Il serait également intéressant d'analyser en détail de quelle manière sont explorés les graphes : cet aspect permettra d'optimiser les stratégies d'exploration. Enfin, à plus long terme, nous souhaiterions appliquer les colonies de fourmis dès le début de la chaîne du SRAP : du treillis de phonèmes à l'extension linguistique.

7 Conclusion

Dans cet article, nous avons présenté un nouveau paradigme permettant d'étendre des graphes de mots issus d'un système de reconnaissance automatique de la parole. Ce paradigme, basé sur les colonies de fourmis, permet d'explorer le graphe afin d'y trouver un chemin optimum sans avoir à le construire dynamiquement en fonction des historiques. À l'heure actuelle, notre méthode ne permet pas d'atteindre les résultats de notre système étalon où les graphes ont été étendus avec des méthodes classiques. Cependant, nous montrons que le paradigme fonctionne : l'application d'un modèle de langage d'ordre supérieur améliore systématiquement le décodage initial. L'implémentation présentée est très basique, nous souhaitons à court terme affiner les heuristiques d'exploration afin d'obtenir une approche compétitive avec celles existantes.

Références

AUBERT, X. L. (2002). An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech & Language*, 16(1):89–114.

BONABEAU, É. et THÉRAULAZ, G. (2000). L'intelligence en essaim. *Pour la science*, (271):66–73.

BONTOUX, B. (2008). *Techniques hybrides de recherche exacte et approchée : application à des problèmes de transport*. Thèse de doctorat, Université d'Avignon et des pays de Vaucluse.

DENEUBOURG, J.-L., GROSS, S., FRANKS, N. et PASTEELS, J.-M. (1989). The blind leading the blind : Modeling chemically mediated army ant raid patterns. *Journal of Insect Behavior*, 2:719–725.

DORIGO et STÜTZLE (2004). *Ant Colony Optimization*. MIT-Press.

DORIGO, M. et GAMBARDELLA, L. (1997). Ant colony system : A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, 1:53–66.

DROGOUL, A. (1993). When ants play chess (or can strategies emerge from tactical behaviors). *In Maamaw'1993*.

GALES, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12:75–98.

GOPINATH, R. A. (1998). Maximum likelihood modeling with gaussian distributions for classification. *In Proceedings of ICASSP*, pages 661–664.

JELINEK, F. (1969). A fast sequential decoding algorithm using a stack. *IBM J. Res. Develop.*, 13.

LECOUTEUX, B. (2008). *Reconnaissance automatique de la parole guidé par des transcriptions a priori*. Thèse de doctorat, Université d'Avignon et des pays de Vaucluse.

MONMARCHE, N., GUINAND, F. et SIARRY, P., éditeurs (2009). *Fourmis artificielles et traitement de la langue naturelle*. Lavoisier, Prague, Czech Republic.

NEY, H., HAEB-UMBACH, R., TRAN, B.-H. et OERDER, M. (1992). Improvements in beam search for 10000-word continuous speech recognition. *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-92*, volume 1, pages 9–12 vol.1.

NOCERA, P., LINARES, G. et MASSONIÉ, D. (2002). Principes et performances du décodeur parole continue speeral. *In XXIVées journées d'étude sur la parole*. Laboratoire Informatique d'Avignon.

ORTMANNS, S. et NEY, H. (2000). The time-conditioned approach in dynamic programming search for. 8(6):676–687.

PAUL, D. (1991). Algorithms for an optimal a* search and linearizing the search in the stack decoder. *In Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP-91*, pages 693–696 vol. 1.

PAUL, M., FEDERICO, M. et SEBASTIAN (2010). Overview of the iwslt 2010 evaluation campaign. *In IWSLT 2010*.

POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., SCHWARZ, O. G. N. G. M. H. P., SILOVSKY, J., STEMMER, G. et VESELY, K. (2011). The kaldi speech recognition toolkit. *In* SOCIETY, I. S. P., éditeur : *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, volume IEEE Catalog No. : CFP11SRW-USB.

ROUSSEAU, A., DELÉGLISE, P. et ESTÈVE, Y. (2012). TED-LIUM : an automatic speech recognition dedicated corpus. *In LREC 2012*.

SCHWAB, D., GOULIAN, J. et TCHECHMEDJIEV, A. (2013). Désambiguïsation lexicale de textes : éfficacité qualitative et temporelle d ?un algorithme à colonies de fourmis. *journal Traitement Automatique des Langues, vol. 54-1*.

WESSEL, F., SCHLÜTER, R. et NEY, H. (2000). Using posterior word probabilities for improved speech recognition. *In* SCHLUTER, R., éditeur : *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '00*, volume 3, pages 1587–1590 vol.3.

Multidimensionnalité de la liaison variable et difficultés de classification. Le cas des adverbes monosyllabiques.

Giulia Barreca
Université Paris Ouest Nanterre La Défense, MoDyCo, CNRS, 200, avenue de la République, 92001 Nanterre
giulia.barreca@u-paris10.fr

RESUME _____

La combinaison des multiples facteurs qui influencent la réalisation de liaison variable constitue une source de complexité qui rend difficile son analyse et son traitement classificatoire. Le critère morphosyntaxique a le plus souvent guidé les classifications de la liaison variable. Dans cette première étude menée sur des séquences d'adverbes monosyllabiques suivis d'adjectifs, nous avons essayé de démontrer que la prise en compte du seul critère d'appartenance aux catégories morphosyntaxiques s'avère insuffisante pour décrire l'hétérogénéité des comportements associés aux éléments appartenant à une même classe. Ensuite, notre but à été celui de décrire de façon plus précise la variabilité de ce phénomène, analysant, à l'aide d'un apprentissage automatique, les nombreuses dimensions linguistiques (phonologique, syntaxique et lexicale) en jeu dans la réalisation de la liaison variable.

ABSTRACT _____

Multidimensionality of the variable liaison and difficulties in classification. The case of monosyllabic adverbs.

The combination of multiple factors that influence the realization of variable liaison is a source of complexity that makes its analysis and classificatory treatment difficult. The morphosyntactic criterion has often guided the classifications of variable liaison. In this first research of sequences of monosyllabic adverbs followed by adjectives, we have tried to prove that considering only the criterion of belonging to morphosyntactic categories is insufficient to describe the heterogeneity of behaviors associated with elements of the same class. The aim of this research was to describe more accurately the variability of this phenomenon, using machine learning, in order to analyze the largest number of linguistic dimensions (phonological, syntactic and lexical) involved in the realisation of variable.

MOTS-CLES : Liaison variable, hétérogénéité, analyse multidimensionnelle, classifications.
KEYWORDS: Variable liaison, heterogeneity, multidimensional analysis, classifications.

1 Introduction

La liaison a fait l'objet de nombreux travaux qui ont dû faire face à l'extraordinaire idiosyncrasie de ce phénomène et par conséquent à la difficulté de dégager et décrire ses nombreuses facettes. En effet, la description de la liaison variable oblige à prendre en compte de plusieurs dimensions d'analyse (phonologique, syntaxique et lexicale), ce qui la rend inévitablement complexe.

Dans les pages qui suivent, après avoir présenté un bref parcours des études menées sur liaison, nous allons mettre en évidence les limites des classifications de la liaison établies, le plus souvent, sur le seul facteur morphosyntaxique[1]. Pour ce faire, nous avons développé une première étude de la catégorie de liaison variable constituée par les adverbes monosyllabiques suivis d'un adjectif (désormais Adv+Adj) à partir des données extraites des corpus PFC (Phonologie du Français Contemporain) (Durand et al., 2002) et CFPP2000 (Corpus de Français Parlé Parisien des années 2000) (Branca-Rosoff et al., 2012). Nous essayerons de démontrer que la prise en compte du seul critère d'appartenance à la classe morphosyntaxique permet difficilement de rendre compte de la complexité et de l'hétérogénéité des comportements de la liaison associés aux adverbes monosyllabiques.

2 Les études précédentes: plusieurs facettes du même phénomène

Les premières études sur la liaison, contrairement aux études génératives successives (Schane, 1967), ont le mérite d'avoir développé une problématisation de l'étude de la liaison et d'avoir mis en avant sa nature multidimensionnelle. En effet, Delattre (Delattre, 1966) considère que la syntaxe ne peut pas représenter le seul critère de classification des liaisons. Malgré la grille descriptive des trois catégories de liaison qu'il établit en 1947, il précise que ces catégorisations ne sont pas absolues, et peuvent varier sur la base de plusieurs facteurs, parmi lesquels le style, la prosodie et le degré phonématique de la consonne de liaison (désormais CL) (Delattre, 1966). En revanche, à partir des études génératives de Schane (Schane, 1967) la liaison ne sera traitée que comme étant un phénomène strictement phonologique et unitaire sans jamais parvenir à expliquer son caractère variationnel. Les études successives se contenteront de décrire de façon partielle l'influence de l'une ou l'autre de ces dimensions, attribuant aux autres un rôle secondaire, finissant par se heurter à l'incapacité d'expliquer la variabilité de la liaison. Parmi celles-ci, par exemple, les études centrées sur la dimension phonologique qui considèrent la liaison comme le résultat d'un évitement du hiatus (Encrevé, 1988) ou de l'émergence des *consonnes latentes* (Tranel, 2000) ne s'appliquent qu'à des contextes de liaison catégorique. Le niveau stylistique est également considéré par de nombreuses études comme le facteur le plus déterminant. Bien que le choix du registre puisse faire varier la fréquence des liaisons, comme le souligne Mallet (Mallet, 2008), cela ne suffit pas à expliquer la variabilité des réalisations pour un même locuteur et la même catégorie morphosyntaxique. Une autre série de travaux se base sur l'analyse des facteurs lexicaux tels que la fréquence du premier et deuxième mot appartenant à la séquence liaisonnée (désormais respectivement M1 et M2) et de leur cooccurrence (Bybee, 2001).

[1] Nous faisons référence aux tripartitions classiques de la liaison en: liaison obligatoire/catégorique, facultative/variable et interdite/erratique. Parmi lesquelles celles qui ont été proposées par Delattre (Delattre, 1947) et Encrevé (Encrevé, 1988).

3 Les limites de la classification de la liaison variable sur base morphosyntaxique. Hétérogénéité des comportements des adverbes monosyllabiques

Comme nous l'avons vu dans les paragraphes précédents, la multidimensionnalité de l'analyse de la liaison, constitue un obstacle à son traitement classificatoire, car on reconnaît que les dimensions sont interdépendantes et on ne peut pas les traiter séparément (Mallet, 2008).

L'appartenance à la classe morphosyntaxique constitue le moyen pour opérer cette classification qui, encore qu'elle soit nécessaire sur le plan métalinguistique, résulte insuffisante pour décrire la réalité des usages linguistiques. Dans ce contexte, à travers cette première étude nous souhaitons décrire de façon plus précise l'hétérogénéité des fonctionnements de la liaison à l'intérieur des séquences Adv+Adj appartenant à la classe des adverbes monosyllabiques suivi d'un M2. Cette classe, comme d'autres, a été dans un premier temps traitée comme contexte de liaison obligatoire (Delattre, 1947). Des études successives (Mallet, 2008) ont intégré ce contexte à la classe des liaisons variables. Toutefois, malgré cette révision, la classification actuelle ne parvient pas à expliquer et donc à décrire la présence d'un grand écart des réalisations de la liaison existant entre les différents adverbes monosyllabiques. En effet, selon l'adverbe qu'on considère, soit la liaison est réalisée très souvent (96,55 % *très*, 64% *plus*) soit presque jamais (1,36 % *pas*)[2]. Une telle variabilité de comportements de la liaison peut être mise en relation, de façon plus générale, aux caractéristiques propres des adverbes. En effet, plusieurs études (Wilmet 2010) ont déjà souligné que les adverbes constituent une classe résiduelle regroupant des éléments invariables qui ne rentrent pas dans d'autres catégories. Par conséquent, il s'agit d'un ensemble très hétérogène qui est difficilement classifiable. Différents critères de classement ont été évoqués par les nombreux travaux classificatoires: critère fonctionnel, sémantique, degré d'intégration de l'adverbe dans la phrase (Melis, 1983) ainsi que la portée de l'adverbe et sa fonction (Blumethal, 1990). Dans les pages suivantes, nous essayerons de mettre en évidence la nécessité d'étendre la description de la variation de la liaison à d'autres champs d'analyse (phonologique, syntaxique et lexicale) que celui d'appartenance à la catégorie morphosyntaxique.

4 Méthodologie

Pour étudier un tel sujet, l'utilisation d'un corpus annoté s'impose. Dans cette première étude, nous allons analyser les séquences Adv+Adj des adverbes *pas*, *très*, *plus*, *moins*, *trop*[3] issues du corpus PFC (Durand et al., 2002) et du corpus CFPP2000 (Branca-Rosoff et al., 2012). En ce qui concerne le corpus PFC, les données prises en compte ont été extraites des transcriptions des entretiens guidés[4] et des entretiens libres[5] (tâches orales) de 134[6] locuteurs appartenant à 34

[2] Données extraites de Mallet (Mallet, 2008).

[3] Nous n'avons pas pris en compte, dans cette première étude, les séquences contenant les adverbes *tout*, *fort*, *mieux*, *bien* à cause de leur faible nombre d'occurrences dans notre échantillon.

[4] Il s'agit d'un entretien semi-dirigé par des questions posées par l'enquêteur d'un registre de langue plutôt soutenu.

[5] L'entretien libre consiste en une interaction plus riche (de registre plus familier) entre plus de deux personnes.

[6] Bien que nous ayons mené la recherche sur l'ensemble des locuteurs du corpus PFC (394 locuteurs pour un total de 36 enquêtes), nous n'avons obtenu qu'un échantillon très limité où seulement 134 locuteurs appartenant à 34 enquêtes différentes sont représentés. Cela est probablement lié au fait que l'annotation morphosyntaxique actuelle du corpus PFC, réalisée à l'aide de Treetagger, présente de nombreuses erreurs d'étiquetage. Actuellement, nous sommes en train de ré-annoter en morphosyntaxe (POS et lemmes) le corpus PFC à l'aide de l'étiqueteur DisMo (Christodoulides et al., 2014) en collaboration avec le laboratoire Valibel de l'Université Catholique de Louvain et l'Université de Genève.

enquêtes différentes. Les données issues du corpus CFPP2000 ont été extraites de l'ensemble des entretiens semi-directifs. L'échantillon des données que nous avons recueilli correspond à 505 occurrences totales de séquences Adv+Adj susceptibles d'être liaisonnées, ainsi réparties parmi les adverbes monosyllabiques suivants: *pas*, *très*, *plus*, *moins*, *trop*.

	pas	très	plus	moins	trop
PFC	81	67	44	8	8
CFPP2000	57	131	82	18	9
TOT	138	198	126	26	17

TABLE 1 – Occurrences totales des séquences Adv+Adj reparties par adverbe

En ce qui concerne les adjectifs nous avons enregistré un ensemble assez hétérogène correspondant à 70 lemmes différents pour le corpus PFC et 118 lemmes pour CFPP2000[7]. Ensuite, dans le but d'assurer une analyse multidimensionnelle de la liaison, pour chaque séquence nous avons recueilli des données relevant de trois principales dimensions d'analyse: phonologique, syntaxique et lexicale. De l'ensemble de ces dimensions résultent les variables dépendantes suivantes[8]:

- Variables phonologiques: la nature de la consonne de liaison (CL), le nombre de syllabes du M2, à savoir de l'adjectif (NSYLL M2);
- Variables syntaxiques: degré de cohésion syntagmatique (DCS) et rapports de dépendance entre l'adverbe et l'adjectif. Notamment si l'adverbe porte sur le verbe et est régi par le verbe (DV) ou s'il porte sur l'adjectif et donc il dépend de l'adjectif (DA);
- Variables lexicales[9]: la fréquence du M1, à savoir de l'adverbe (FREQ M1), la fréquence du M2, à savoir de l'adjectif (FREQ M2) et la fréquence de cooccurrence de la séquence Adv+Adj (COOC)[10].

Pour ce qui est de la variable indépendante, la réalisation de la liaison (REALISATION), afin d'assurer des résultats plus fiables sur un échantillon restreint, nous avons décidé de ne considérer que deux cas possibles: liaison réalisée (désormais LR), liaison non réalisée (désormais LNR).

L'ensemble des données ainsi organisées a été soumis au traitement de deux différents logiciels d'apprentissage automatique: *Sipina* (Zighed, 1992) et *Weka* (Holmes et al., 1994). Il s'agit de deux logiciels qui, dans le cadre de l'apprentissage supervisé, génèrent des arbres de décision. Ceux-ci permettent d'effectuer des classifications de données et d'aboutir au traitement d'un grand nombre de variables explicatives à partir de l'observation d'un échantillonnage. Autrement dit,

[7] Il s'agit plus précisément de 223 formes fléchies totales d'adjectifs différents possédant des indices de fréquence compris entre 0 et 248,45, extraits de la fréquence *freqfilm2* de la base de donnée lexicale *Lexique 3* (New, 2006).
[8] Nous n'avons pas analysé le facteur de variabilité inter/intra-individuelle étant donné le nombre faible d'occurrences par locuteur présent dans notre échantillon.
[9] Ces derniers facteurs de nature lexicale ont été extraits de la fréquence *freqfilm2* de la base de donnée lexicale *Lexique 3* (New, 2006) correspondant à la moyenne des fréquences par million des quatre sous-corpus de sous titres de films. Afin de vérifier la fiabilité des ces indices nous les avons comparé aux occurrences totales (tout contexte mélangé) des mêmes adverbes tirées du corpus CFPP2000. Les rapports de fréquence entre les adverbes correspondent.
[10] Cette valeur est représentée de manière absolue.

grâce à l'emploi de ces outils d'apprentissage automatique, nous espérons pouvoir représenter des tendances qui décrivent la variation de la liaison à partir de notre base de données d'apprentissage de séquences Adv+Adj. Il s'agira concrètement de prédire la valeur (LR ou LNR) de la variable indépendante (REALISATION) à partir de l'ensemble des variables prédictives (CL, NSYLL M2, DCS, FREQ M1, FREQ M2, COOC) afin d'établir une hiérarchie parmi chacune des différentes variables.

Cette méthode nous permettra, tout en gardant les propriétés d'économicité d'un outil descriptif, d'étudier conjointement les multiples facteurs en jeu dans la réalisation de la liaison variable des séquences Adv+Adj. Cela ne signifie pas pour autant que la catégorie morphosyntaxique d'appartenance ne joue pas un rôle important, mais simplement nous supposons qu'elle ne constitue pas un critère descriptif suffisant.

5 Résultats et discussion

Si nous analysons les résultats totaux de notre échantillon ne prenant en compte que le critère d'appartenance à la catégorie Adv+Adj nous observons 66% de LR et 34% de LNR. Cependant, en analysant les taux de réalisations de chaque adverbe, nous pouvons constater un fort déséquilibre interne à la catégorie. La liaison est beaucoup réalisée avec les adverbes *très* (95% LR) et *plus* (94% LR) alors que elle n'est jamais réalisée avec *pas* (0% LR). L'adverbe *pas*, en effet, représente 82% des LNR totales.

	pas	très	plus	moins	trop
LR	0%	95%	94%	65%	59%
LNR	100%	5%	6%	35%	41%

TABLE 2 – Résultats des LR et des LNR des séquences Adv+Adj reparties par adverbe

Nous avons enregistré des comportements différents même dans des séquences Adv+Adj où *pas*[11] et *très*[12] étaient associés au même adjectif.

Ces données révèlent une anomalie existant à l'intérieur de la classification sur base morphosyntaxique. En effet, l'un des inconvénients de cette classification réside dans la tendance à regrouper sous une même classe des adverbes qui peuvent se différencier fortement à la fois selon leur distribution et leur fréquence. Les adverbes monosyllabique *pas*, *plus, moins*, *trop* sont des adverbes qui peuvent porter sur le verbe, sur l'adjectif ou sur un adverbe et possèdent des rapports de dépendance différents (DA ou DV) selon leur portée. De plus il faut souligner que dans des séquences Adv+Adj, *plus* (en tant qu'adverbe comparatif et superlatif) se trouve bien plus souvent que l'adverbe *pas* dans un rapport de dépendance de l'adjectif que du verbe. Dans notre échantillon nous avons pu observer que seulement dans 10% des séquences du type *pas*+Adj, l'adverbe *pas* dépend de l'adjectif. Dans ces cas il se rapproche de la fonction de *non*

[11] «Je sais pas+ je sais pas si euh+ c'est vrai que je trouve que le+ le mode de vie à Pris est assez stressant y a+ les gens sont *pas agréables*» (LNR - CFPP2000_Corpus Killian Belamy et Lucas Hermano).
[12] «et puis après rue du Faubourg Saint–Antoine c'est très vivant c'est vraiment *très agréable*» (LR - CFPP2000_Corpus Pierre Beysson).

en tant que préfixe adjectival[13]. En revanche, nous n'avons enregistré aucune séquence susceptible d'être liaisonnée, contenant *plus* régi par le verbe. Concernant l'adverbe *très*, il ne peut porter que sur l'adjectif ou sur l'adverbe. Par conséquent il sera toujours régi par l'adjectif ou l'adverbe duquel il dépend.

Nous pourrions imaginer que la dimension syntaxique permet à elle seule de décrire la variabilité de la liaison associée aux séquences Adv+Adj. Cette différence fonctionnelle et distributionnelle est sans doute l'un des facteurs principaux qui permet d'expliquer l'écart des réalisations de la liaison entre les adverbes monosyllabiques. Toutefois, la variable DCS n'explique ni l'absence totale de LR dans les cas où *pas* est régi par l'adjectif ni l'écart des réalisations entre *très*, *plus*, *moins* et *trop*. En effet, nous supposons que d'autres facteurs jouent un rôle important dans la réalisation de la liaison variable à l'intérieur de la catégorie Adv+Adj. Pour cette raison, dans une deuxième étape de ce travail, nous avons procédé à la construction d'arbres de décision à l'aide des logiciels *Sipina* et *Weka*. Nous avons utilisé les arbres de décision afin d'établir une hiérarchisation des variables explicatives à partir de l'observation de notre échantillonnage. Pendant l'utilisation de *Sipina*, nous avons subdivisé l'ensemble de données en deux échantillons: un échantillon apprentissage correspondant à 338 occurrences (67%) et un échantillon test correspondant à 167 occurrences (33%).

Dès nos premiers résultats sur cet échantillon le descripteur FREQ M1 semble constituer la première variable utilisée dans le partitionnement (variable de segmentation). De fait il possède le meilleur coefficient de Tschuprow[14] T=0,6794 et constitue la variable de segmentation la plus pertinente parmi tous les descripteurs discrets et continus.

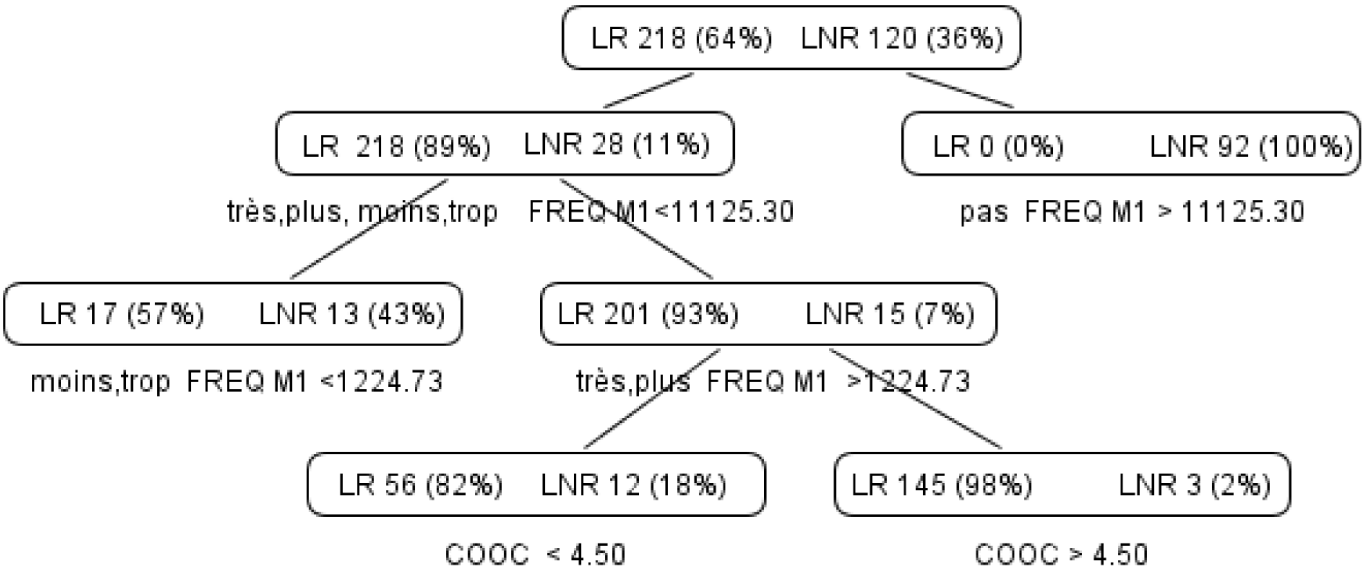


FIGURE 1 – Arbre de décision à partir de l'échantillon d'apprentissage des séquences Adv+Adj

La variable de segmentation FREQ M1 permet de classifier les données relatives à la variable

[13] « ah non ironique j+ *pas* ironique non mais euh j'sais que quand j'étais petit euh+euh j'étais très question réponse quoi » (Corpus CFPP2000_Corpus Mathieu Rosier et Elisa Rysnik).

[14] Indice de corrélation qui permet d'évaluer la qualité de la segmentation et donc de sélectionner le meilleur parmi les descripteurs, évaluant sa pertinence dans le partitionnement de l'arbre de décision.

REALISATION en deux groupes (arêtes). La valeur (seuil de coupure)[15] choisie pour opérer telle segmentation correspond à l'indice de fréquence 11.125,30. Ce point de coupure permet plus précisément de discrétiser, au premier niveau de l'arbre, entre deux groupes d'adverbes et donc de séquences liaisonnées qui leur sont associées. D'une part, l'adverbe pas, dont l'indice de fréquence est supérieur à 11.125,30. D'autre part, les adverbes monosyllabiques dont la fréquence est inférieure à 11.125,30, à savoir très, plus, moins et trop. Cette première classification repose aussi bien sur une dimension syntaxique dépendante de la variable DCS. Bien que celle-ci possède une indice de corrélation très élevé (Tschuprow T=0,6008) elle ne figure pas dans la représentation de l'arbre. En effet, cette variable ne parvient pas entièrement à classer la variable REALISATION. Il n'existe pas de correspondance précise entre la valeur LNR de la variable REALISATION et la valeur DV du descripteur DCS, puisque comme on l'a déjà observé, 10% des LNR associées à l'adverbe pas correspondent aux séquences où pas dépend de l'adjectif suivant (à savoir DA). En outre la variable syntaxique rend difficilement compte des différences de réalisation entre très, plus, moins, trop. En revanche, au deuxième niveau de l'arbre, la variable FREQ M1 permet de segmenter ce dernier ensemble en deux sous-groupes à travers un deuxième seuil de coupure correspondant à l'indice 1224,73. Le premier sous-groupe est constitué par les adverbes monosyllabiques dont l'indice de fréquence est inférieur à 1224,73, notamment moins et trop. Le deuxième est formé par les adverbes monosyllabiques dont la fréquence est supérieure à 1224, à savoir très et plus. Ce deuxième niveau de segmentation représente l'écart des réalisations existant entre d'une part très (95%) et plus (94%) et d'autre part moins (65%) et trop (59%). Autrement dit ce seuil de discrétisation permet de faire la différence parmi les adverbes monosyllabiques possédant la même distribution syntaxique. La variable de segmentation lexicale COOC (au troisième niveau de l'arbre) (Tschuprow T=0,0739) semblerait permettre d'expliquer les LNR des séquences associées aux adverbes très et plus comme possédant une fréquence de cooccurrence inférieure à 4,5 occurrences. Enfin, la variable FREQ M2 (Tschuprow T=0,05) semble exercer une faible influence sur la réalisation de la liaison ainsi que les variables phonologiques (NSYLL M2 et CL) qui ne possèdent presque aucune corrélation[16].

En résumant, la variable lexicale FREQ M1 semble être la variable la plus importante dans la description de la réalisation de la liaison permettant d'opérer des classifications des données à deux niveaux différents de l'arbre. Cependant, il faut préciser que bien que le descripteur FREQ M1 constitue la variable de segmentation la plus pertinente pour notre échantillon, notre classification ne tient compte que des adverbes monosyllabiques. Par conséquent cette variable ne possède que quatre valeurs correspondant aux indices des fréquences de *très*, *plus*, *moins*, *trop*. Il résulte que classifier les réalisations de la liaison sur la base du descripteur FREQ M1 revient à les classer à partir de l'adverbe monosyllabique lui-même. Autrement dit, il semblerait que la réalisation de la liaison des séquence Adv+Adj soit majoritairement influencée par la nature lexicale de l'adverbe même, strictement liée à sa fréquence. Afin d'évaluer la qualité de l'apprentissage, nous avons réalisé le *test set validation*[17] comparant la prédiction du modèle avec les résultats obtenus dans l'échantillon test. Nous constatons un taux d'erreur égal à 3,6%.

[15] Seuil établit par l'apprentissage supervisé à partir duquel la variable de segmentation permet de séparer les données du corpus d'apprentissage en deux arêtes.

[16] Cependant, nous supposons qu'une analyse menée sur un échantillon d'apprentissage plus grand pourrait mettre en évidence des tendances déjà observées par des études précédentes d'après lesquelles la fréquence du M2 peut jouer un rôle dans la réalisation de la liaison.

[17] Méthode de validation croisée utilisée pour estimer la fiabilité d'un modèle fondé sur une technique d'échantillonnage.

Dans un second temps, afin de vérifier la fiabilité de ces résultats, nous avons soumis le même corpus au logiciel *Weka*, utilisant une différente méthode de validation croisée (*K-fold cross-validation*). Cette deuxième analyse a confirmé nos résultats, sélectionnant la variable FREQ M1 comme variable de segmentation dont les meilleurs points de coupure sont les mêmes mis en évidence dans l'analyse précédente. Dans ce deuxième apprentissage, nous constatons un taux d'erreur sur l'ensemble de l'échantillon (505 instances) égal à 8,5% correspondant à 91,48% d'instances classifiées correctement.

6 Conclusions

Dans cette étude, nous avons présenté un premier travail exploratoire mené sur le corpus PFC et sur le corpus CFPP2000 qui nous a permis d'analyser conjointement l'incidence de plusieurs facteurs en jeu dans la production de la liaison. À partir des résultats de l'étude de cet échantillon nous avons relevé que la classe Adv+Adj possède une forte complexité interne liée à l'hétérogénéité qui caractérise les distributions syntaxiques et les indices de fréquence des adverbes monosyllabiques. Dans ce contexte, il nous apparaît nécessaire, tout en gardant l'économicité des classifications existantes, de rendre compte de façon plus précise, (à travers des outils relevant de la linguistique computationnelle), des comportements différents par rapport autres adverbes monosyllabiques. L'ensemble des résultats de cette analyse nous suggère que les classifications établies sur base morphosyntaxique, dépourvues de toute description interne, ne parviennent qu'à élaborer des généralisations qui permettent difficilement de saisir le fonctionnement variable de la liaison.

Dans cette première étude, malgré l'extension limitée de notre échantillon, nous avons essayé de donner un aperçu d'une analyse plus détaillée de la liaison variable en essayant d'établir une connexion entre les multiples facteurs en jeu dans la réalisation de la liaison et en soulevant des problèmes qui suggèrent de nouvelles pistes de recherche. Actuellement nous sommes en train d'élargir cette analyse multidimensionnelle aux autres contextes de liaison du corpus PFC. L'analyse d'un tel ensemble de données nous permettra de valider les considérations faites dans cette première étude et de vérifier l'incidence des facteurs phonologiques, syntaxiques et lexicaux dans la réalisation de la liaison en analysant toute occurrence de séquences liaisonnées à l'intérieur des classifications traditionnelles. Autrement dit, il s'agira de remettre en discussion la légitimité de l'appartenance aux catégories morphosyntaxiques en tant que seul critère de classification. À travers cette multiplication des plans d'analyse nous espérons parvenir à une description de la liaison qui se rapproche le plus possible de la variabilité des usages langagiers. Par conséquent, dans les études futures, il faudra établir une plus grande articulation entres les différentes dimensions, et réfléchir sur les liens existant entre les variables phonologiques et les facteurs lexicaux et entre ces derniers et la dimension syntaxique. De plus, il s'agira d'enrichir la réflexion sur la dimension syntaxique, de dépasser l'analyse superficielle des catégories morphosyntaxiques d'appartenance en étudiant les relations syntaxiques spécifiques à chaque séquence qui établissent les degrés de cohésion interne aux séquences liaisonnées[18].

[18] Laks (Laks, 2005).

Références

BLUMENTHAL, P. (1990). Classement des adverbes: Pas la Couleur, rien que la nuance? *Langue française*, 88(1):41-50.

BRANCA-ROSOFF, S., FLEURY, S., LEFEUVRE, F., PIRES, M. (2012). Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000). http://cfpp2000.univ-paris3.fr/CFPP2000.pdf. [consulté le 27/01/2014].

BYBEE, J. (2001). Frequency effects on French liaison. *In* (Bybee et Hopper, 2001), 337-359.

BYBEE, J. et HOPPER, P., éditeurs (2001). *Frequency and the Emergence of Linguistic Structure*. John Benjamins.

CHRISTODOULIDES, G., AVANZI, M., GOLDMAN, J.P. (2014) DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. *In IX Language Resources and Evaluation* (*LREC 2014),* Reykjavik, Inslande. À paraître.

DELATTRE, P. (1947). La liaison en français, tendances et classification. *The French Review*, 22(2):148-157.

DELATTRE, P. (1966). *Studies in French and comparative Linguistics*. Mouton.

DURAND, J., LAKS, B. et LYCHE, C. (2002). La phonologie du français contemporain: usages, variétés et structure. *In* (Pusch et Raible, 2002), 93-106.

ENCREVÉ, P. (1988). *La liaison avec et sans enchainement*. Seuil

HOLMES, G, DONKIN, A et WITTEN, H.I. (1994). Weka: A machine learning workbench. *In Proceedings of Second Australia and New Zealand Conference on Intelligent Information Systems*. Brisbane, Australie.

LAKS, B. (2005). Phonologie et construction syntaxique: la liaison, un test de figement et de cohésion. *Linx*, 53:155-171.

MALLET, G. (2008). *La liaison en français: description et analyses dans le corpus PFC*. Thèse de Doctorat non publiée. Laboratoire MoDyCo, Université Paris Ouest Nanterre la Défense.

MELIS, L. (1983). *Les circonstants et la phrase*. Presses Universitaires de Louvain.

NEW, B.(2006). Lexique 3: Une nouvelle base de données lexicales. *In Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain, Belgique.

PUSCH, C. et RAIBLE, W., éditeurs (2002). *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*. Gunter Narr Verlag.

SCHANE, S. A. (1967). La liaison et l'élision en français. *Langages*, 8:37-59.

TRANEL, B. (2000). Aspects de la phonologie du français et la théorie de l'optimalité. *Langue française*, 126:39-72.

WILMET, M. (2010). *Grammaire critique du français*. Duculot.

ZIGHED, D., AURAY, J., DURU, G. (1992). *Sipina: Méthode et Logiciel*. Lacassagne.

Segmentation et Regroupement en Locuteur pour le traitement incrémental des collections volumineuses

Grégor Dupuy, Sylvain Meignier et Yannick Estève
LUNAM Université, LIUM, Le Mans, France
prenom.nom@lium.univ-lemans.fr

RÉSUMÉ

Les systèmes de Segmentation et Regroupement en Locuteurs *cross-show* actuels reposent principalement sur un processus de regroupement global qui traite collectivement chaque émission d'une collection. Cette approche a déjà été étudiée dans diverses situations et semble être le meilleur moyen à ce jour pour atteindre des taux d'erreur satisfaisants, dans une durée de traitement raisonnable. Néanmoins, ce processus montre ses limites dans un contexte applicatif réaliste où de grandes et dynamiques collections doivent être traitées. Dans cet article, nous étudions l'utilisation d'un regroupement cross-show incrémental pour traiter de manière itérative des émissions devant être insérées dans une collection existante. Les nouvelles émissions à insérer sont traitées les unes après les autres, selon l'ordre chronologique de diffusion. Les expériences ont été menées sur les enregistrements LCP et BFMTV distribués au cours des campagnes d'évaluation françaises ETAPE et REPERE. L'ensemble représente 67 heures de données annotées, réparties sur 310 enregistrements, couvrant une période d'environ deux ans (de septembre 2010 à octobre 2012).

ABSTRACT

Cross-show Speaker Diarization to Incrementally Process Large Volume of Data

Current cross-show diarization systems are mainly based on an overall clustering process that handles collectively each show of a collection. This approach has already been studied in various situations and seems to be the best way so far to achieve low error rates. Nevertheless, that process shows its limits in a realistic applicative context where large and dynamically increasing collections have to be processed. In this paper we investigate the use of an incremental clustering cross-show speaker diarization architecture to iteratively process new shows within an existing collection. The new shows to be inserted are processed one after another, according to the chronological order of broadcasting. Experiments were conducted on the LCP and the BFMTV show recordings distributed among the ETAPE and the REPERE French evaluation campaigns. It represents 67 hours of annotated data, distributed among 310 shows, and covering a 2-years period (from Sept. 2010 to Oct. 2012).

MOTS-CLÉS : SRL, architecture incrémentale, regroupement PLNE global, i-vecteurs.

KEYWORDS: speaker diarization, incremental architecture, cross-show, ILP clustering, i-vectors.

1 Introduction

La tâche de Segmentation et Regroupement en Locuteurs (SRL) joue un rôle essentiel dans de nombreuses applications de traitement automatique de la parole, comme la transcription automatique, la détection des entités nommées et du rôle des locuteurs. Cette tâche a été définie par le NIST dans le cadre des campagnes d'évaluation *Rich Transcription* comme le partitionnement d'un flux audio en segments, et le regroupement de ces segments en fonction de l'identité des locuteurs. Ce processus traite individuellement les enregistrements audio, et les locuteurs détectés sont identifiés par des étiquettes anonymes spécifiques à chaque enregistrement : un même locuteur impliqué dans deux émissions différentes est identifié par deux étiquettes différentes. Un inconvénient à cette approche est de ne pas prendre en compte les interventions de certains locuteurs récurrents dans plusieurs émissions, bien que cette situation se produise fréquemment dans les programmes audiovisuels d'information. En conséquence, la tâche de SRL a récemment été considérée dans un contexte plus large, dans lequel les émissions qui composent un corpus sont traitées collectivement pour regrouper les locuteurs impliqués dans plusieurs émissions. Ainsi, un locuteur cross-show est toujours identifié par la même étiquette dans chacune des émissions de la collection.

Cette extension de la tâche "Qui parle ? Quand ?" a conduit à l'émergence du concept de *collection*, qui se réfère à un ensemble de données avec une ou plusieurs caractéristiques communes entre les émissions. L'un des questionnements actuels concerne les collections dont le volume augmente dynamiquement au cours du temps. Des études antérieures ont montré que les architectures de SRL cross-show les mieux adaptées pour faire face à ce genre de situation sont de type *incrémentales* (Tran *et al.*, 2011)(Yang *et al.*, 2011), c'est à dire, traiter les émissions les unes après les autres en utilisant les modèles de locuteur des émissions déjà traitées dans le but d'aider la SRL de l'émission courante. L'objet de cet article est d'étudier l'utilisation d'un système de SRL dans un contexte applicatif réaliste : nous avons observé l'effet de l'insertion de nouveaux enregistrements dans une collection existante, en ajoutant et traitant les émissions les unes après les autres en fonction de leur date de diffusion.

Nous donnons un aperçu de l'état de l'art de la SRL incrémentale dans la partie 2, puis nous décrivons l'architecture incrémentale que nous avons expérimentée dans la section 3. Nous présentons ensuite les données, les métriques d'évaluation et les résultats expérimentaux dans la partie 4. Enfin, nous concluons dans la partie 5.

2 Travaux connexes

La tâche de SRL cross-show vise à étendre la tâche de SRL à un contexte plus large, où les locuteurs qui apparaissent dans plusieurs enregistrements d'une même collection (les *locuteurs cross-show*) sont toujours identifiés par la même étiquette. Les systèmes de SRL qui travaillent au niveau "enregistrement", et ne se préoccupent pas des relations entre les locuteurs impliqués dans plusieurs émissions, sont nommés, de manière contrastive, systèmes *single-show*. La tâche de SRL cross-show partage le même objectif que la tâche de "liaison des locuteurs" (speaker linking) (Ferràs et Bourlard, 2013)(Bourlard *et al.*, 2013)(Ghaemmaghami *et al.*, 2013). Cependant, dans ce procédé l'appariement global des étiquettes de locuteur est indépendant de la tâche SRL.

Les expériences menées précédemment sur le regroupement incrémental ont été réalisées sur

des collections de petite taille (23 enregistrements pour une durée totale d'environ 4h) (Tran *et al.*, 2011)(Yang *et al.*, 2011). L'architecture de SRL cross-show par regroupement incrémental proposé par les auteurs (Tran *et al.*, 2011)(Yang *et al.*, 2011), illustrée en Figure 1, fonctionne de la manière suivante : chaque enregistrement de la collection est d'abord traité individuellement jusqu'à un regroupement agglomératif hiérarchique BIC. Dans ce premier regroupement, les locuteurs sont modélisés par des Gaussiennes à matrice de covariance pleine, afin de maximiser la pureté des classes.

Puisqu'aucune information *a priori* n'est disponible pour traiter le premier enregistrement de la collection, la segmentation résultant du regroupement BIC est utilisée telle quelle pour procéder au regroupement agglomératif hiérarchique CLR. Dans ce second regroupement, les locuteurs sont désormais modélisés par des Modèles de Mixture Gaussiens (GMMs). Un module d'identification en locuteur (OSI module) a été inséré entre les regroupements BIC et CLR. Ce module mémorise les modèles de locuteur des enregistrements déjà traités. À partir du second enregistrement, les GMMs déjà extraits des enregistrements précédemment traités peuvent être utilisés dans le regroupement CLR courant. Après avoir traité intégralement une émission, le module d'identification en locuteur est mis à jour avec les nouveaux modèles de locuteurs, et les modèles de locuteur existant sont actualisés avec les données additionnelles que la dernière émission traitée peut fournir.

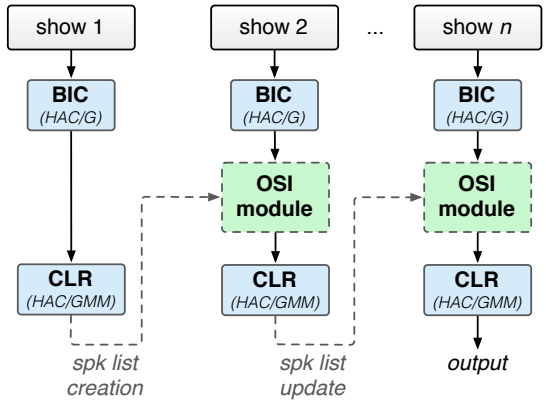


FIGURE 1 – Architecture incrémentale basée sur des regroupements agglomératifs hiérarchiques et des GMMs, incluant un module d'identification en locuteur (OSI module).

Cette approche cross-show de type incrémentale permet de traiter l'ensemble des enregistrements d'une collection bien plus rapidement que les architectures basées sur un regroupement de "global", qui traiteraient l'ensemble des enregistrements simultanément. Cette architecture est également adaptée aux situations où de nouveaux enregistrements viendraient régulièrement compléter une collection existante. Elle présente cependant deux inconvénients : le taux d'erreur DER (Diarization Error Rate) est plus élevé que celui obtenu par une approche cross-show basée sur un regroupement global, et l'ordre dans lequel les émissions sont traitées influence les résultats.

3 Regroupement incrémental global par PLNE

Motivés par l'approche incrémentale présentée dans la partie précédente, ainsi que par les récents résultats en SRL cross-show que nous obtenons avec le regroupement global par Programmation Linéaire en Nombres Entiers (PLNE), nous avons expérimenté un système de SRL regroupant ces deux aspects : architecture incrémentale et regroupement global par PLNE.

Nous avons considéré une collection initiale dans laquelle de nouveaux enregistrements ont été insérés et traités les uns après les autres (Figure 2). L'ordre dans lequel les émissions sont insérées dans la collection initiale suit l'ordre chronologique de diffusion. En considérant cet ordre chronologique, les n premières émissions ont été sélectionnées pour constituer la

collection initiale. Cette collection initiale est traitée avec une approche cross-show basée sur un regroupement global par PLNE (Dupuy *et al.*, 2012). La collection initiale a pour but d'amorcer le processus de SRL cross-show incrémental en considérant la segmentation cross-show et les modèles de locuteur de la collection initiale.

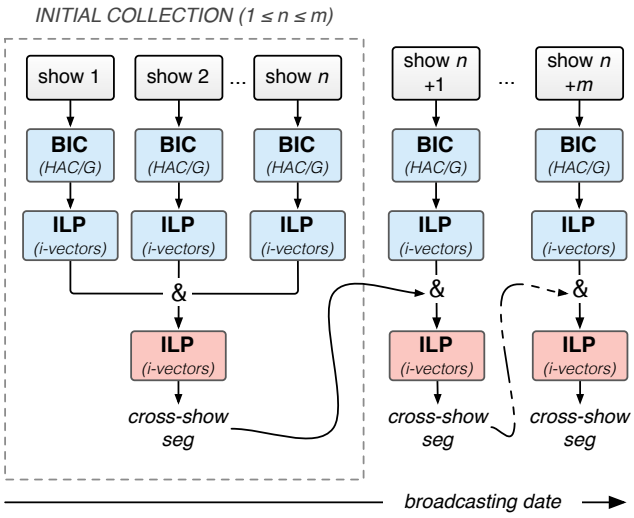


FIGURE 2 – Architecture cross-show incrémentale reposant sur un regroupement global par PLNE.

Le processus incrémental est réalisé comme suit : d'abord, chaque enregistrement de la collection est traité individuellement avec système de SRL single-show (*cf.* partie 3.1). Le traitement cross-show de la collection initiale consiste en un regroupement global par PLNE sur la concaténation des n segmentations single-show de la collection initiale (*cf.* section 3.2).

Pour chaque insertion d'une nouvelle émission, sa segmentation single-show est concaténée à la segmentation cross-show courante (qui résulte du traitement incrémental de l'émission précédemment insérée dans la collection). Un regroupement global par PLNE est alors appliqué, produisant une nouvelle segmentation cross-show dans laquelle la nouvelle émission a été prise compte. Le processus est ainsi répété jusqu'au traitement de la dernière émission de la collection.

3.1 SRL single-show

Le système de SRL single-show a été mis en place à l'aide de la boîte à outils *LIUM_SpkDiarization* [1] (Meignier et Merlin, 2009)(Rouvier *et al.*, 2013). Les systèmes conçus à partir de cette boîte à outils ont réalisé d'excellentes performances dans les tâches de SRL single- et cross-show lors des dernières campagnes d'évaluation francophones sur des émissions journalistiques, radio et TV, comme ETAPE (2011) (Gravier *et al.*, 2012) et REPERE (Jan. 2012, 2013 et 2014) (Galibert et Kahn, 2013).

Le système de SRL single-show est basé sur une segmentation acoustique et un regroupement agglomératif hiérarchique utilisant le critère BIC (Bayesian Information Criterion), avec des distributions Gaussiennes à matrices de covariance pleines. Les frontières des segments ont été ajustées par un décodage Viterbi utilisant des GMMs à 8 composantes appris sur les données de chaque locuteur via l'algorithme EM (Expectation-Maximization). Un autre décodage Viterbi, avec 8 HMMs (Hidden Markov Models) mono-état, représentés par des GMMs à 64 composantes entraînés par EM-ML (Expectation-Maximization Maximum Likelihood), est utilisé pour retirer les zones de non-parole. Segmentations, regroupements et décodages sont réalisés avec 12 paramètres MFCC (Mel-Frequency Cepstral Coefficients), complétés par l'énergie.

À cet instant, chaque locuteur n'est pas nécessairement représenté par une seule et unique classe. Un autre regroupement est donc effectué, exprimé sous la forme d'un problème de PLNE. Dans ce regroupement, les classes (les locuteurs) sont modélisées par des i-vecteurs et la similarité

1. `http://www-lium.univ-lemans.fr/en/content/liumspkdiarization`

entre les classes est estimée avec une distance de Mahalanobis (Rouvier et Meignier, 2012). Les i-vecteurs sont devenus état de l'art dans le domaine de la vérification du locuteur (Dehak *et al.*, 2011). Ils permettent de réduire les données acoustiques d'un locuteur en un vecteur de faible dimension, ne retenant que les informations dîtes "pertinentes" à son propos. L'approche à base de i-vecteurs a d'abord été adaptée à la SRL par l'intermédiaire de l'algorithme *K-means*, appliqué aux distances entre les i-vecteurs ,dans un corpus où le nombre de locuteurs est connu *a priori* (Shum *et al.*, 2011). Dans nos travaux, le nombre de locuteurs est inconnu.

Un i-vecteur de dimension 50 est extrait pour chacun des regroupements de la segmentation BIC, à l'aide d'un GMM-UBM de 256 composantes. Les i-vecteurs sont normalisés par leur longueur dans un processus itératif (Garcia-Romero et Espy-Wilson, 2011)(Bousquet *et al.*, 2011). Les paramètres acoustiques utilisés pour l'extraction des i-vecteurs correspondent à 12 MFCCs, complétés de l'énergie et des dérivées premières. Les paramètres acoustiques sont normalisés par la moyenne et la variance. Le problème de regroupement consiste, d'une part, à minimiser le nombre de classes centrales C choisi parmi les N i-vecteurs de la segmentation BIC, et d'autre part, à minimiser la dispersion des i-vecteurs au sein des classes. La valeur $C \in \{1, \ldots, N\}$ devant être déterminée automatiquement. Ce problème de classification est exprimé sous la forme d'un problème de PLNE (eq. 1), où la fonction objective de résolution est minimisée en vérifiant les contraintes suivantes :

$$\text{let } C \in \{1 \ldots N\}, \text{ let } K_{j \in C} = \{k / d(k,j) < \delta\}$$

$$\text{Minimize :} \qquad \sum_{k \in C} x_{k,k} + \frac{1}{N\delta} \sum_{j \in C} \sum_{k \in K_j} d(k,j) x_{k,j} \qquad (1)$$

$$\text{Subject to :} \qquad x_{k,j} \in \{0,1\} \qquad k \in K_j, j \in C \qquad (1.2)$$

$$\sum_{k \in K_j} x_{k,j} = 1 \qquad j \in C \qquad (1.3)$$

$$x_{k,j} - x_{k,k} < 0 \qquad k \in K_j, j \in C \qquad (1.4)$$

Où $x_{k,k}$ (eq. 1) est une variable binaire égale à 1 lorsque le i-vecteur k est un centre. Le nombre de centres C est implicitement inclus dans l'équation 1 ($C = \sum_{k=1}^{N} x_{k,k}$). La distance $d(k,j)$ est estimée par une distance de *Mahalanobis* entre les i-vecteurs k et j. $1/N\delta$ est un facteur de normalisation. La variable binaire $x_{k,j}$ est égale à 1 quand le i-vecteur j est assigné au centre k. Chaque i-vecteur j doit être associé à un seul centre k (eq. 1.3). L'équation 1.4 assure la sélection du centre k si un i-vecteur lui a été assigné. Les distances sont implicitement prises en compte dans les équations 1.2, 1.3 et 1.4, en utilisant l'ensemble K_j à la place de C. Pour une valeur j donnée, l'ensemble K_j représente l'ensemble des valeurs possibles de k (prise en 1 et N), pour lesquelles la distance entre les classes k et j sont plus petites que le seuil δ.

Le corpus de test fourni lors de la campagne d'évaluation française ESTER1(Galliano *et al.*, 2009) a été utilisé pour l'apprentissage du GMM-UBM, et le corpus d'apprentissage ESTER1 a été utilisé pour de l'apprentissage des i-vecteurs (étape de normalisation). Le problème de PLNE est résolu par l'algorithme *Branch and Bound* implémenté dans le solveur *glpsol* de la suite *GNU Linear Programming Toolkit* [2].

2. `http://www.gnu.org/software/glpk/`

3.2 SRL cross-show

La SRL cross-show consiste en un regroupement global par PLNE, dans lequel l'ensemble des émissions est traité simultanément. Ce regroupement global par PLNE est réalisé de la même manière que décrit dans la sous-partie précédente, cependant, la configuration à changé : le GMM-UBM de 256 composantes a été remplacé par un GMM-UBM à 1024 composantes, et les paramètres acoustiques correspondent désormais à 19 MFCCs, complétés par l'énergie ainsi que les dérivées premières et secondes. Ces paramètres restent normalisés par la moyenne et la variance.

L'aspect "cross-show" de ce regroupement PLNE repose sur la concaténation des segmentations single-show de plusieurs émissions. Le système de SRL cross-show présenté dans cet article diffère quelque peu de celui décrit dans (Dupuy *et al.*, 2012).

- Plutôt que de ré estimer les modèles de locuteur à partir de la concaténation des segmentations single-show, nous avons "recyclé" les modèles de locuteur (i-vecteurs) correspondant aux classes identifiées comme des centres, dans les segmentations single-show. Cette approche n'est possible que dans la mesure où les segmentations dont la concaténation est donnée en entrée ont été obtenues à partir de la même formulation du problème de PLNE.

- D'autre part, l'équation 4 du problème de PLNE, formulée dans la sous-partie précédente, est utilisée en amont : la matrice associant les distances aux couples de i-vecteurs peut être interprétée comme un graphe complètement connecté, dans lequel les nœdus représentent les locuteurs, et les arêtes représentent les distances entre les i-vecteurs. Il est possible d'identifier les sous-composantes connexes de ce graphe en éliminant les arêtes correspondant aux distances inférieures au seuil δ. Chaque sous-composante connexe représente un sous-ensemble indépendant du problème de regroupement. La recherche de sous-composantes connexes permet ainsi de découper le problème de PLNE, tel qu'originalement formulé, en un certain nombre de sous-problèmes élémentaires, qui ne consistent qu'en un faible nombre de variables et contraintes. Considérer le problème de regroupement en locuteur comme la recherche des sous-composantes connexes d'un graphe totalement connecté a permis de diviser par 5 le temps de calcul de ce regroupement global par PLNE sur les 142h de données que nous introduisons dans la partie 4.1 (2'47 minutes au lieu de 13'23 minutes).

Il est important de noter que nous avons exprimé notre problème de regroupement global par PLNE de manière à manipuler globalement la concaténation des segmentations single-show : le système de SRL cross-show est libre de regrouper les classes issues d'une même émission. En effet, nous avons constaté que le regroupement global par PLNE permet souvent d'améliorer, quelque peu, les segmentations single-show. Cette particularité explique pourquoi le DER single-show est amené à évoluer au cours des itérations de l'approche cross-show incrémentale.

4 Expériences

4.1 Données

Les données considérées pour réaliser les expériences consistent en un regroupement de toutes les émissions BFMTV et LCP distribuées dans les corpus d'apprentissage, de développement et

de test des campagnes françaises ETAPE et REPERE Phase 1. Nous avons également inclus les données d'apprentissage fournies pour la phase 2 du défi REPERE. L'ensemble représente un total de 142 heures de données audio, réparties sur 310 fichiers, enregistrés entre septembre 2010 et octobre 2012. La distribution des données en termes de nombre d'émissions et de durées est présentée dans le Tableau 1. Seule une partie de ces données a été annotée et évaluée : les durées présentées dans le Tableau 1 (exprimées en heures et minutes) ont été calculées d'après les fichiers UEM (Unpartitioned Evaluation Map) spécifiant les régions à évaluer.

Show name	Whole data		Initial collection	
	# Show	Duration	# Show	Duration
BFMStory	37	21 :21	7	05 :01
CultureEtVous	54	01 :42	0	-
PlaneteShowbiz	73	02 :24	1	02 :07
All BFMTV	*164*	*25 :27*	*8*	*07 :08*
CaVousRegarde	20	09 :02	9	04 :29
EntreLesLignes	24	08 :20	10	04 :05
LCPInfo13h30	31	07 :57	0	-
LCPInfo20h30	4	00 :40	1	00 :10
PileEtFace	32	08 :15	12	04 :07
TopQuestions	35	07 :41	9	02 :03
All LCP	*146*	*41 :55*	*41*	*15 :44*
BFMTV+LCP	*310*	*67 :22*	*49*	*22 :52*

TABLE 1 – Distribution des données, en nombre d'émissions et en durées (h :min), pour la totalité des données BFMTV+LCP, ainsi que pour la collection initiale.

Les 49 premières émissions, selon l'ordre chronologique de diffusion, ont été sélectionnées pour constituer la collection initiale. Les émissions restantes ont été itérativement ajoutées à la collection initiale. Cette collection initiale, qui représente environ la moitié des données en termes de durée, a été utilisée pour déterminer les seuils des systèmes de SRL single- et cross-show.

4.2 Résultats et discussion

Le DER (Diarization Error Rate) est la métrique utilisée pour mesurer les performances en SRL. Le DER fut introduit par le NIST comme la fraction du temps de parole qui n'est pas attribué au bon locuteur, en utilisant la meilleure correspondance entre les étiquettes des références et des hypothèses. L'outil d'évaluation a été développé par le LNE [3] dans le cadre des campagnes ETAPE et REPERE. Cet outil d'évaluation a été conçu de manière à calculer les DER single-show et cross-show.

Les DER single- et cross-show obtenus lors de chaque itération de l'expérience incrémentale sont présentés dans la Figure 3. Les droites horizontales représentent les DER single- et cross-show obtenus par regroupement global PLNE sur l'ensemble des 310 émissions de la collection. La collection est composée d'émissions enregistrées sur une période de deux ans, avec quelques ruptures temporelles notables entre deux enregistrements consécutifs. Les droites verticales interrompues, sur la Figure 3, représentent les interruptions temporelles de plus de 12 jours consécutifs. En moyenne, la durée écoulée entre deux enregistrements consécutifs est d'une journée.

3. Laboratoire National de métrologie et d'Éssais

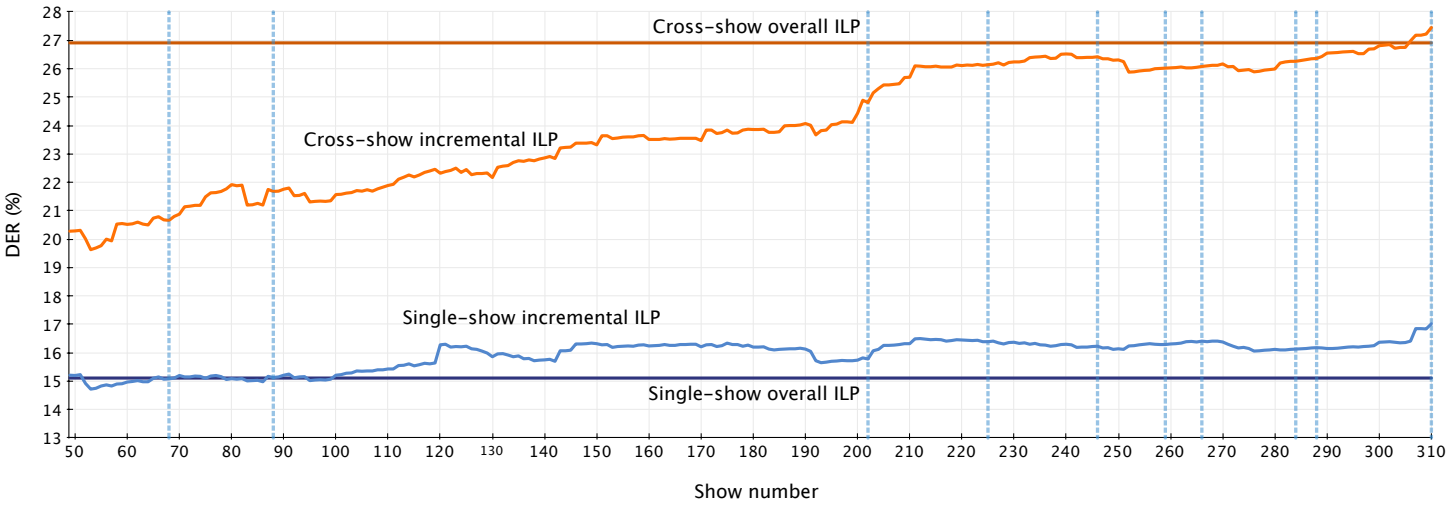


FIGURE 3 – DER single- et cross-show pour chaque itération de l'expérience incrémentale

Le DER cross-show obtenu avec le regroupement global par PLNE sur les 310 émissions de la collection est élevé (26,91%), comparé à celui obtenu sur les 49 enregistrements de la collection initiale (20,28%). Le DER obtenu avec l'approche incrémentale, après que la dernière émission ait été traitée, est de 27,45%. L'évaluation single-show des segmentations cross-show donne des DER plus élevés avec l'approche incrémentale (17,02%), comparé aux DER obtenus par regroupement global PLNE sur l'ensemble des données (15,11%). Une explication par rapport à la brutale augmentation des DER entre les émissions 200 et 210 pourrait être l'interruption temporelle entre les émissions 201 et 202, qui est de 24 jours. Cependant, ce comportement étrange ne s'observe pas au niveau des autres interruptions temporelles. Retirer ce bloc d'émissions de la collection permet d'obtenir des résultats globalement meilleurs, il convient donc de penser que certains locuteurs posent problème dans ces émissions. 55,8% des locuteurs de la collection initiale sont également présents dans le reste des données, qui sont traitées itérativement. De plus, 12 locuteurs sont particulièrement présents (la plupart d'entre eux correspondent à des présentateurs et des personnalités fortement médiatisées). Le temps de parole de ces 12 locuteurs représente environ 30,6% du temps de parole évalué sur les 310 émissions.

5 Conclusion

La complexité de la tâche de Segmentation et Regroupement en Locuteurs cross-show, en ce qui concerne les programmes audiovisuels d'information, augmente en fonction de la quantité de données à traiter. L'approche incrémentale peut convenir au traitement des collections volumineuses, ou susceptibles de prendre du volume au cours du temps. Les résultats sont très proches de ceux obtenus par regroupement global sur l'ensemble des données, mais ce qui rend l'approche incrémentale vraiment intéressante est le rapport résultats/temps de calcul. Considérer le problème de regroupement en locuteur comme la recherche des sous-composantes connexes d'un graphe totalement connecté a permis de diviser par 5 le temps de calcul. Le "recyclage" des modèles de locuteurs en fonction les regroupements effectués en amont permet d'économiser les ressources de manière non négligeable et d'effectuer les regroupements à la chaîne. Traiter une collection de 142 heures de données avec un regroupement global par PLNE reste une solution envisageable, cependant, cette approche globale risque de montrer ses limites avec des collections bien plus conséquentes en volume.

Références

Bourlard, H., Ferras, M., Pappas, N., Popescu-Belis, A., Renals, S., McInnes, F., Bell, P. et Guillemot, M. (2013). Processing and Linking Audio Events in Large Multimedia Archives : The EU inEvent Project. *In Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France.

Bousquet, P.-M., Matrouf, D. et Bonastre, J.-F. (2011). Intersession Compensation and Scoring Methods in the I-vectors Space for Speaker Recognition. *In Proceedings of Interspeech*, Florence, Italie.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. et Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *In Proceedings of IEEE TASLP*, volume 19, pages 788–798.

Dupuy, G., Rouvier, M., Meignier, S. et Estève, Y. (2012). I-vectors and ILP Clustering Adapted to Cross-Show Speaker Diarization. *In Proceedings of Interspeech*, Portland, Oregon (USA).

Ferràs, M. et Bourlard, H. (2013). Speaker Diarization and Linking of Large Corpora. *In Proceedings of IEEE Spoken Language Technology Workshop*, Miami, Floride (USA).

Galibert, O. et Kahn, J. (2013). The first official REPERE evaluation. *In Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. *In Proceedings of Interspeech*, Brighton, UK.

Garcia-Romero, D. et Espy-Wilson, C. Y. (2011). Analysis of I-vector Length Normalization in Speaker Recognition Systems. *In Proceedings of Interspeech*, Florence, Italie.

Ghaemmaghami, H., Dean, D. et Sridha, S. (2013). Speaker Attribution of Australian Broadcast News Data. *In Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A. et Galibert, O. (2012). The ETAPE Corpus for the Evaluation of Speech-based TV Content Processing in the French Language. *In Proceedings of LREC*, Istanbul, Turkey.

Meignier, S. et Merlin, T. (2009). LIUM SpkDiarization : An Open-Source Toolkit for Diarization. *In CMU SPUD Workshop*, Dallas, Texas (USA).

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T. et Meignier, S. (2013). An Open-source State-of-the-art Toolbox for Broadcast News Diarization. *In Proceedings of Interspeech*, Lyon, France.

Rouvier, M. et Meignier, S. (2012). A Global Optimization Framework For Speaker Diarization. *In Odyssey Workshop*, Singapore.

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. et Glass, J. (2011). Exploiting Intra-Conversation Variability for Speaker Diarization. *In Proceedings of Interspeech*, Florence, Italie.

Tran, V.-A., Le, V. B., Barras, C. et Lamel, L. (2011). Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization. *In Proceedings of Interspeech*, Florence, Italie.

Yang, Q., Jin, Q. et Schultz, T. (2011). Investigation of Cross-show Speaker Diarization. *In Proceedings of Interspeech*, Florence, Italie.

L'ombre d'un doute? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles

Lucie Scarbel[1], Denis Beautemps [1], Jean-Luc Schwartz [1],
Sébastien Schmerber[2], Marc Sato [1]
(1) GIPSA-LAB, Département Parole & Cognition ; CNRS UMR 5216 - Université de Grenoble-Alpes ;
(2) Service ORL du CHU Grenoble
lucie.scarbel@gipsa-lab.grenoble-inp.fr

RESUME

Un argument classique en faveur des théories motrices de la perception de la parole provient du paradigme de « close-shadowing » (répétition rapide). Le fait que cette tâche de close-shadowing entraîne des réponses orales bien plus rapides qu'en réponses manuelles suggère en effet un codage des représentations perceptives dans un format moteur, compatible avec une réponse orale. Un autre argument est apporté par les interactions audio-visuelles lors de la perception de parole, souvent interprétées en référence à un couplage fonctionnel entre audition, vision et motricité. Dans cette étude, nous avons combiné ces deux paradigmes de manière à tester si la modalité visuelle pouvait induire des réponses motrices plus rapides lors d'une tâche de close-shadowing. Pour ce faire, différentes tâches de catégorisation orale et manuelle de stimuli de parole présentés auditivement ou audio-visuellement, en présence ou non d'un bruit blanc, ont été réalisées. De manière générale, les réponses orales ont été plus rapides que les réponses manuelles, mais aussi moins précises, notamment dans le bruit, ce qui suggère que la représentation motrice induite par la stimulation pourrait être peu précise dans un premier niveau de traitement. En présence d'un bruit acoustique, la modalité audiovisuelle s'est avérée à la fois plus rapide et plus précise que la modalité auditive. Aucune interaction entre le mode de réponse et la modalité de présentation des stimuli n'a cependant été observée. Nous interprétons l'ensemble de ces résultats dans un cadre théorique proposant l'existence de boucles perceptivo-motrices, dans lesquelles les entrées auditives et visuelles seraient intégrées et reliées à la génération interne de représentations motrices préalablement au processus final de décision.

ABSTRACT

The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close shadowing

One classical argument in favor of a functional role of the motor system in speech perception comes from the close shadowing task in which a subject has to identify and to repeat as quickly as possible an auditory speech stimulus. The fact that close shadowing can occur very rapidly and much faster than manual identification of the speech target is taken to suggest that perceptually-induced speech representations are already shaped in a motor-compatible format. Another argument is provided by audiovisual interactions often interpreted as referring to a multisensory-motor framework. In this study, we attempted to

combine these two paradigms by testing whether the visual modality could speed motor response in a close-shadowing task. To this aim, both oral and manual responses were evaluated during the perception of auditory and audio-visual speech stimuli, clear or embedded in white noise. Overall, oral responses were much faster than manual ones, but it also appeared that they were less accurate in noise, which suggests that motor representations evoked by the speech input could be rough at a first processing stage. In the presence of acoustic noise, the audiovisual modality led to both faster and more accurate responses than the auditory modality. No interaction was however observed between modality and response. Altogether, these results are interpreted within a two-stage sensory-motor framework, in which the auditory and visual streams are integrated together and with internally generated motor representations before a final decision may be available.

MOTS-CLES : perception de la parole, production de la parole, perception de la parole audio-visuelle, close-shadowing, interaction sensorimotrice

KEYWORDS: speech perception, speech production, audio visual speech perception, close-shadowing, sensory motor interaction

1 Introduction

Un débat classique dans le domaine de la perception de la parole concerne l'implication du système moteur et du lien fonctionnel entre représentations auditives et motrices. Les théories auditives de la perception de parole réfutent l'implication du système moteur et proposent l'existence de processus auditifs de décodage phonétique à partir du signal acoustique (Diehl et al., 2004). A contrario, Liberman et al. (1985) et Fowler (1986) supposent que pour percevoir la parole, nous utilisons des représentations procédurales motrices basées sur notre expérience de locuteur. Enfin, les théories perceptivo-motrices postulent que les représentations motrices soient utilisées en lien avec les représentations auditives dans le traitement et décodage des informations phonétiques (Skipper et al., 2007, Schwartz et al. 2012).

Dans une récente revue sur les théories motrices, Galantucci et al. (2006) rappellent les principaux arguments expérimentaux et en mentionnent notamment deux qui fournissent le cœur du présent travail. Un premier argument en faveur d'un couplage perceptivo-moteur en parole provient du paradigme de close-shadowing. Ce paradigme (Porter et al., 1980 ; Fowler et al., 2003) consiste à répéter le plus rapidement possible des stimuli de parole. L'analyse des réponses orales permet de mesurer la rapidité et le taux de répétitions correctes. Galantucci (2006) compare ces résultats avec les temps de réaction obtenus par Luce (1986) dans une tâche de réponse manuelle dans laquelle les sujets utilisaient une touche. Les résultats montrent une augmentation des temps de réaction dans le cas de la réponse manuelle. Cette différence ne pouvant être expliquée par la difficulté du choix des touches dans la tâche de décision manuelle, Galantucci et al. l'interprètent par le fait que si percevoir la parole c'est percevoir des gestes, alors la perception des gestes préparerait la réponse orale et la rendrait ainsi plus rapide.

Le second paradigme est celui de la multisensorialité. De nombreuses études ont montré que l'entrée visuelle améliore la compréhension de parole, tant pour les sujets malentendants que normo-entendants. Sumby & Pollack (1954) ont été parmi les premiers à démontrer l'apport de la modalité visuelle pour percevoir et comprendre la parole dans des conditions

bruitées. L'apport de la modalité visuelle a été également démontré dans des études de répétition (shadowing) où la tâche consistait à répéter oralement les stimuli, et à mettre ainsi en évidence une meilleure compréhension par des sujets de matériaux linguistiques complexes ou produits dans une langue étrangère ou avec accent étranger (Reisberg, 1987 ; Davis and Kim, 2001). Cependant, ces tâches de répétition se sont faites sans pression de temps (shadowing et non close-shadowing). Or la tâche de répétition rapide fournit une fenêtre riche sur la dynamique temporelle du processus de décision. A l'inverse les expériences de close-shadowing n'ont jamais incorporé la modalité visuelle, se privant d'une connaissance sur le rôle des interactions audiovisuelles en lien avec les relations perceptuo-motrices. La présente étude se propose précisément d'étudier, pour la première fois, quel est l'apport de la modalité visuelle dans une tâche de close-shadowing.

Cette étude est composée de deux expériences, toutes deux focalisées sur une évaluation conjointe de la précision et de la rapidité de réponses orales ou manuelles à des stimuli auditifs ou audio-visuels. Les deux expériences ont été réalisées sur des stimuli de parole non-lexicaux (logatomes), présentés sans bruit dans la première expérience (Expérience A) ou avec bruit acoustique dans la seconde expérience (Expérience B). Les hypothèses sous-jacentes sont que (1) les réponses orales devraient être plus rapides que les réponses manuelles, en accord avec les études précédentes sur le close-shadowing, et que (2) les réponses aux stimuli audio-visuels devraient être plus rapides et plus précises que celles aux stimuli auditifs, au moins dans l'Expérience B impliquant des stimuli bruités. Une question supplémentaire concerne la possibilité d'une interaction entre ces deux effets, qui permettrait d'évaluer si l'effet de la vision est différent entre une modalité de réponse (orale) et l'autre (manuelle).

2 Méthodologie

2.1 Participants

Deux groupes de quinze et quatorze adultes sains, de langue maternelle française, ont participé aux Expériences A et B (Expérience A: 10 femmes; moyenne d'âge: 29 ans, entre 20 et 38 ans - Expérience B: 11 femmes; moyenne d'âge: 24 ans, entre 19-34 ans). Tous les participants ont une vision normale ou corrigée à la normale et ont rapporté n'avoir jamais eu des troubles moteurs, de la parole ou de l'audition.

2.2 Procédure expérimentale

Chaque expérience consistait en deux tâches de catégorisation : une tâche de close-shadowing où les réponses étaient données oralement, en répétant le plus vite possible la séquence présentée, et une tâche de décision manuelle, où les réponses étaient données manuellement, en appuyant le plus vite possible sur la touche appropriée. Les stimuli à catégoriser correspondaient aux séquences /apa/, /ata/ et /aka/ (voir ci-dessous). Les participants étaient informés qu'on allait leur présenter des séquences /apa/, /ata/ ou /aka/, soit de manière auditive soit de manière audio-visuelle. Dans la tâche de close-shadowing, on leur demandait de catégoriser et répéter chaque séquence le plus vite possible. Pour ce faire, ils devaient produire la voyelle initiale /a/ puis répéter immédiatement la syllabe CV perçue (/pa/, /ta/ ou /ka/). Lors de la tâche de décision manuelle, les participants devaient catégoriser chaque énoncé en appuyant le plus vite

possible avec leur main dominante sur une des trois touches correspondant respectivement à /apa/, /ata/ ou /aka/. L'ordre des touches était contrebalancé entre les participants. Pour chaque tâche (avec réponses manuelles ou orales) et chaque modalité (auditive ou audio-visuelle), 16 répétitions de chacune des séquences /apa/, /ata/ et /aka/ étaient présentées de manière randomisée. Les ordres de présentation des tâches et des modalités étaient contrebalancés entre les participants. Les deux expériences ont été réalisées dans une chambre sourde. Les participants étaient assis en face d'un ordinateur à une distance d'approximativement 50 cm. Les stimuli acoustiques étaient présentés à un niveau sonore confortable, celui-ci étant le même pour tous les participants. Le logiciel Presentation (Neurobehavioral Systems, Albany, CA) a été utilisé pour contrôler la présentation des stimuli et pour enregistrer les réponses manuelles. Toutes les productions des participants ont été enregistrées grâce à un microphone AKG 1000S pour les analyses offline, avec un système assurant la synchronisation entre les stimuli présentés et la réponse des participants. Une courte session d'entrainement précédait chaque tâche. La durée totale de chaque expérience était d'environ 30 minutes.

2.3 Stimuli

Les séquences /apa/ /ata/ et /aka/, produites dans une chambre sourde par un homme de langue maternelle française ont été enregistrées audio-visuellement au moyen d'un microphone AKG 1000S (44.1 kHz) et d'une caméra haute qualité au format PAL placée en face du locuteur (images détramées de 572 par 520 pixels, 50 Hz). Le corpus a été enregistré avec pour objectif d'obtenir pour chaque séquence 4 productions distinctes impliquant quatre durées de la voyelle initiale /a/ (0.5s, 1s, 1.5s et 2 s) dans le but de réduire toute prédiction temporelle de la séquence à catégoriser). Les durées des 12 stimuli ainsi sélectionnés ont été égalisées (3 séquences x 4 productions). Les stimuli étaient présentés sans bruit additionnel dans l'Expérience A, un bruit blanc (filtré à -6 dB/oct) a été ajouté à chaque stimulus dans l'Expérience B (rapport signal sur bruit de -3 dB).

2.4 Analyses acoustiques

Afin de calculer les temps de réaction (RTs) et la proportion de réponses correctes dans la tâche à réponses orales, des analyses acoustiques de la production des participants ont été réalisées en utilisant le logiciel Praat (Boersma et Weenink, 2013). Les temps de réaction ont été calculés uniquement pour les réponses correctes : les omissions ou tout autre type d'erreurs (c'est-à-dire le remplacement d'une consonne par une autre ou la production de deux consonnes ou de deux syllabes pour un même stimulus dans la tâche de close-shadowing) ont été exclues. Les temps de réaction pour les réponses orales ont été mesurés entre le début du burst consonantique du stimulus et de la réponse.

2.5 Analyse des données

Dans chaque expérience, la proportion de réponses correctes et la médiane des temps de réaction étaient déterminées individuellement pour chaque participant, chaque tâche et chaque modalité, séparément pour /apa/, /ata/ et /aka/. Deux ANOVA à mesures répétées ont été réalisées sur ces données, avec le groupe (Expérience A avec stimuli non-bruités vs. Expérience B avec stimuli bruités) comme variable inter-sujets et la tâche (close shadowing

vs. décision manuelle), la modalité (audio vs. audio-visuelle), et le type de stimulus (/apa/ vs /ata/ vs /aka/) comme variables intra-sujets.

3 Résultats

Pour toutes les analyses suivantes, le niveau de significativité était fixé à $p = .05$ et corrigé en cas de violation de l'hypothèse de sphéricité. Les analyses post-hoc ont été effectuées en utilisant des tests de Bonferroni.

3.1 Temps de réaction

Comme attendu, l'effet principal du groupe est significatif ($F(1,27)=24,38$; $p<0.001$), avec des temps de réaction pour les stimuli non bruités de l'expérience A plus courts que ceux des stimuli bruités de l'expérience B (351 ms vs 484 ms). Les effets principaux de la tâche ($F(1,27)=151,70$; $p<0.001$) et de la modalité ($F(1,27)=14,79$; $p<0.001$) sont aussi significatifs. Pour la tâche, les réponses orales étaient plus rapides que les réponses manuelles (286ms vs 545ms). Par rapport à la modalité, les temps de réaction étaient plus courts dans la modalité audio-visuelle par rapport à la modalité auditive (405 ms vs. 425 ms). Une interaction significative entre groupe et modalité ($F(1,27)=21,74$; $p<0.001$) montre que l'effet bénéfique de la présentation audio-visuelle est présent avec les stimuli bruités dans l'expérience B (461 ms vs. 507 ms) mais non avec les stimuli non-bruités de l'expérience A (354 ms vs. 349 ms). Par contre, l'interaction entre modalité et réponse n'est pas significative.

Ces effets semblent être dépendants des syllabes perçues. Notamment, une interaction à trois facteurs 'tâche x modalité x syllabe' a été trouvée ($F(2,54)=6,49$; $p<0.005$). Dans la modalité auditive, aucune différence significative des temps de réaction n'a été observée entre les syllabes à la fois pour les réponses orales et pour les réponses manuelles. Par contre, dans la modalité audio-visuelle, les temps de réaction pour les réponses orales étaient plus rapides pour la syllabe /pa/ par rapport aux syllabes /ta/ et /ka/, alors que les temps de réaction pour les réponses manuelles étaient plus rapides pour /pa/ par rapport à /ka/ et pour /ka/ par rapport à /ta/.

Ainsi, globalement, on obtient un patron de résultats attendus : des temps de réponse plus courts en réponse orale, un effet du bruit ralentissant les temps de réponse, une accélération de la réponse en modalité audiovisuelle par rapport à la modalité auditive en présence de bruit. Il n'apparaît cependant pas d'interaction entre modalité et type de réponse.

3.2 Proportion de réponses correctes

L'effet principal du groupe est significatif ($F(1,27)=266,28$; $p<0.001$) avec une proportion de réponses correctes plus élevée pour les stimuli non-bruités de l'expérience A (95%) par rapport aux stimuli bruités de l'expérience B (61%). D'autres effets principaux ont été significatifs, à la fois pour la tâche ($F1(1,27)=69,40$; $p<0.001$) et pour la modalité ($F(1,27)=52,39$; $p<0.001$). Concernant la tâche, une baisse importante des réponses correctes a été observée pour les réponses orales par rapport aux réponses manuelles (73% vs. 85%). Comme l'interaction significative du groupe par la tâche ($F(1,27)=38,67$; $p<0.001$) l'indique, cet effet n'apparait que pour les stimuli bruités de l'expérience B (71%

vs. 50%) alors qu'aucune différence n'a été observée entre les réponses orales et manuelles pour les stimuli non-bruités de l'expérience A (93% vs. 98%). Concernant la modalité, la modalité audio-visuelle apporte plus de réponses correctes que la modalité auditive (82% vs. 75%). Par contre, comme l'indique l'interaction significative 'groupe x modalité' ($F(1,27)=72,36$; $p<0.001$) aucune différence n'apparaît entre les deux modalités avec les stimuli non-bruités de l'expérience A (96% vs. 95%) alors qu'avec les stimuli bruités de l'expérience B, la modalité audio-visuelle apporte plus de réponses correctes (68%) que la modalité auditive (53%).

Là encore, les résultats dépendent de la syllabe présentée. Si les 3 syllabes sont parfaitement identifiées dans l'Expérience A en l'absence de bruit, quelle que soit la tâche et la modalité, en condition bruitée (Expérience B) la syllabe « pa » apparaît la plus saillante à la fois auditivement et visuellement. Si on obtient ici encore une confirmation d'un patron attendu (réponses plus précises en l'absence de bruit et, dans le cas de stimuli bruités, en présence de la modalité visuelle), un résultat fort et inattendu doit être relevé : le fait que la réponse orale dégrade la précision de la réponse en cas de stimuli bruités (Expérience B). Une nouvelle fois, il n'apparaît pas d'interaction entre modalité et type de réponse.

Tableau 1: Moyenne des temps de réaction (en ms.) et des % de réussite

modalité	**A**	**A**	**A**	**A**	**A**	**A**	**AV**	**AV**	**AV**	**AV**	**AV**	**AV**
mode	**oral**	**oral**	**oral**	**manuel**	**manuel**	**manuel**	**oral**	**oral**	**oral**	**manuel**	**manuel**	**manuel**
syllabe	**ka**	**pa**	**ta**	**ka**	**pa**	**ta**	**ka**	**pa**	**ta**	**ka**	**pa**	**ta**
RTs sans bruit	250	208	259	471	442	465	261	197	268	474	416	506
RTs avec bruit	348	335	373	614	666	632	277	296	318	531	653	601
% sans bruit	90%	99%	93%	98%	99%	98%	89%	100%	88%	98%	96%	97%
% avec bruit	56%	36%	40%	67%	42%	79%	89%	44%	38%	96%	58%	85%

4 Discussion

4.1 Effet de la tâche : mode de réponse oral ou manuel

Dans la condition non bruitée (Expérience A), les réponses orales sont plus rapide que les réponses manuelles (240ms vs. 462ms), mais aussi marginalement moins précises (93% vs. 98%, effet non significatif). Les temps de réaction sont en adéquation avec ceux obtenus par Fowler et al. (2003) et par Porter & Castellanos (1980). Ces auteurs interprètent la rapidité de la réponse dans le mode oral en référence avec les théories motrices. Le système orofacial serait ainsi favorisé pour répondre de manière très rapide, puisque le percept serait déjà en adéquation avec le format moteur ; le système manuel, nécessitant une étape transitoire entre la décision et l'action, serait ralenti d'autant. Cependant, les données de la condition bruitée (Expérience B) apportent des précisions importantes et inattendues sur ce raisonnement. En effet, alors que les temps de réaction restent plus courts dans la tâche de close-shadowing (334ms vs. 633ms), la précision dans la tâche de réponse orale diminue considérablement par rapport à la tâche de décision manuelle (50% vs. 71%).

Ces nouvelles données nécessitent de modifier l'interprétation des tenants des théories motrices jusqu'à un certain point. Nous allons proposer une tentative d'explication dans le

cadre du modèle proposé par Skipper et al. (2007) pour intégrer les interactions perceptuo-motrices dans la perception de parole. Ces auteurs proposent un modèle inspiré de « l'analyse par la synthèse » (Stevens & Halle 1967 ; Bever & Poeppel, 2010). Le modèle de Skipper et al. implique une boucle corticale entre les aires auditives et motrices. Après un stade initial de traitement auditif dans le cortex temporal (cortex auditif primaire, secondaire et aires associatives : stade 1), le cortex frontal générerait des hypothèses phonémiques associées avec les buts articulatoires puis des commandes motrices correspondant à cette prédiction initiale (Pars opercularis, cortex pré moteur ventral et cortex moteur primaire : stade 2), afin d'émettre des copies d'efférence qui seraient ensuite renvoyées dans le cortex auditif afin d'être comparées avec l'input auditif (stade 3). Ce modèle peut être utilisé comme une base pour tenter d'interpréter nos propres données. Pour ce faire, nous supposons que les réponses manuelles et orales sont générées à deux stades différents dans cette boucle de traitement. Les réponses orales seraient générées au stade 2, en accord avec les postulats de Porter & Castellanos ou de Fowler et al. Quand l'information provenant du cortex auditif aurait généré des commandes motrices dans le cortex moteur, le système orofacial, pré-activé depuis le début de l'expérience de close-shadowing pour permettre aux participants de répondre le plus vite possible, générerait une réponse orale produite par ces commandes motrices. Les réponses orales étant traitées à un stade précoce, elles seraient donc plus rapides mais sont aussi, selon nos résultats, moins précises, ce qui correspond au modèle de Skipper et al. qui considère qu'il ne s'agirait que d'une première hypothèse de réponse qui devraient être affinées à un stade ultérieur. Au stade 2, par contre, le système manuel ne reçoit pas de stimulations spécifiques permettant de générer une réponse. Par contre, au stade suivant (stade 3), le transfert de l'information auditive au cortex auditif, grâce à la copie d'efférence, fournit, en intégrant cette hypothèse motrice avec l'entrée acoustique, une information plus précise qui peut alors être transférée au système manuel pour réponse. Dans ce raisonnement, les réponses orales et manuelles seraient émises à deux instants différents du processus d'analyse par synthèse. Par conséquent, les temps de réaction pour les réponses manuelles seraient plus lents que ceux des réponses orales, mais les réponses seraient plus précises puisque, contrairement aux réponses orales, dans la tâche de décision manuelle les prédictions auraient été confirmées et ajustées avec le cortex auditif avant la décision finale envoyée aux commandes motrices manuelles pour appui de la touche adéquate. Bien sûr cette explication est probablement trop simple pour tenir compte de tous les aspects de nos données. Néanmoins, l'aspect crucial de nos résultats est qu'une hypothèse de pur processus d'identification motrice, certes compatible avec des temps de réaction plus courts en réponse orale, ne semble pas pouvoir rendre compte de la diminution de la précision de réponse orale pour des stimuli bruités. Ces résultats semblent donc réfuter une version stricte des théories motrices au profit de théories perceptuo-motrices de la perception de parole telles que celle de Skipper et al. (2007).

4.2 Effets de la modalité : auditive vs. audio-visuelle

Les effets de la modalité n'apparaissent dans notre étude que dans l'Expérience B avec les stimuli bruités. Dans la modalité auditive, les temps de réaction sont alors plus longs que dans la modalité audio-visuelle, et les proportions de réponses correctes sont plus faibles. Pris ensemble, ces résultats montrent un bénéfice clair de l'apport de la modalité visuelle à l'input auditif, ce qui est en accord avec toutes les études depuis Sumby et Pollack (1954).

Dans notre étude, l'avantage audio-visuel apparaît essentiellement pour la syllabe /pa/ ce qui est classique et en lien avec la haute visibilité des mouvements des lèvres associées à la bilabiale /p/, et le fort degré de confusion entre les mouvements visuels associés à /t/ ou /k/, généralement considérés comme appartenant à la même classe visémique. Ces effets de la modalité ne sont pas présents dans l'Expérience A avec des stimuli non-bruités, très probablement parce que les temps de réaction dans la modalité auditive sont déjà trop courts et les proportions de réponses correctes trop élevées pour être améliorés par l'entrée visuelle (effet plafond).

Un point intéressant est qu'il n'y a pas d'interaction significative entre la modalité et la tâche c'est-à-dire que la diminution des temps de réaction et l'amélioration des proportions de réponses correctes de la modalité auditive à la modalité audiovisuelle sont similaires dans les tâches orales ou manuelles. Nous allons là encore tenter d'interpréter cette absence d'interaction dans le cadre du modèle proposé par Skipper et collègues. Dans ce modèle, les informations auditives et visuelles, après prétraitement dans les aires auditives et visuelles, convergeraient dans une aire multi-sensorielle dans le cortex temporal postéro-supérieur (stade 1). Ensuite, dans le cas d'un input multi-sensoriel, les premières hypothèses seraient donc plutôt multi-sensorielles plutôt qu'uniquement auditives. A partir de là comme précédemment génération d'une hypothèse phonémique associée avec des buts articulatoires puis des commandes motrices orofaciales (stade 2), et émission d'une copie d'efférence fournissant une prédiction multisensorielle comparée avec l'input (stade 3). Dans notre étude, les interactions audio-visuelles du stade 1 affineraient les procédés sensori-moteurs et produiraient des hypothèses phonémiques plus rapides et plus précises au stade 2, qui est le stade où, dans notre interprétation, les réponses orales seraient générées. Ensuite le même gain en rapidité et en précision serait rétro-propagé vers les aires auditives pour génération des réponses manuelles (stade 3). Il n'y aurait ainsi pas de raison d'attendre des différences de gain visuel entre les tâches orales ou manuelles, le gain étant essentiellement déterminé dès le stade 1 dans le modèle.

5 Conclusion

En résumé, les résultats de la présente étude suggèrent que les réponses manuelles et orales sont générées à deux stades différents dans la chaine de perception de la parole. Dans la théorie du modèle « d'analyse par synthèse », les réponses manuelles serait fournies seulement à la fin de la boucle complète, incluant un processus feedworward de génération de prédictions phonémiques associées avec les buts articulatoires puis des commandes motrices émettant en feedback des hypothèses multi-sensorielles comparées au flux d'évènements multi-sensoriels. Contrairement aux réponses manuelles, les réponses orales seraient produites à un stade plus précoce (fin du processus feedforward) où les commandes motrices sont générées, produisant des réponses plus rapides mais moins précises. L'apport visuel améliorerait la rapidité et la précision pour les phonèmes suffisamment visibles (comme par exemple /p/) et ce lorsque le système auditif est en difficulté (conditions adverses, par exemple en présence de bruit). Bien évidemment, d'autres interprétations ou d'autres théories pourraient être confrontées aux présentes données expérimentales. Mais globalement, l'ensemble des résultats de cette étude semble nécessiter une théorie perceptuo-motrice de la perception de parole dans laquelle les flux auditifs et visuels sont intégrés à des représentations motrices auto-générées, avant d'aboutir à une décision finale.

Remerciements

Les auteurs tiennent à remercier l'ANR Plasmody qui a permis de financer cette étude, ainsi que les participants qui se sont portés volontaires pour les deux expériences.

Bibliographie

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics,* 4.2-.3, 174-200.

Davis, C. & Kim, J. (2001), 'Repeating and remembering foreign language words: Implications for language teaching system', *Artificial Intelligence Review*, vol 16, no 1 , pp 37 - 47.

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149-179.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, 49, 296–314.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3–28.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13, 361-377.

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

Luce RD. (1986). *Response times: Their role in inferring elementary mental organization.* Oxford University Press; New York.

Porter, R., & Castellanos, F. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America,* 67, 1349–1356.

Reisberg, D, McLean, J. & Goldfield, A. (1987). Easy to Hear to Understand: A Lip-Reading Advantage with Intact Auditory Stimuli. *In Dodd, B. and Cambell, R. (eds.) Hearing by Eye: The Psychology of Lip-Reading*, 97-113. London: Lawrence Erlbaum.

Schwartz, J.L., Basirat, A., Ménard, L., & Sato, M. (2010). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics,* 25, 336-354.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*. 17 : 2387-2399.

Stevens, K.N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. *In: Wathem-Dunn, W. (ed), Models for the Perception of Speech and Visual Form. Cambridge, MA: MIT Press.*

Sumby W.H., & Pollack I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26 (2), 212-215.

Modélisation de trajectoires et de classes de locuteurs pour la reconnaissance de voix d'enfants et d'adultes

Arseniy Gorin[1,2,3] Denis Jouvet[1,2,3]
Equipe Parole, LORIA
(1) Inria, 615 rue du Jardin Botanique, F-54600, Villers-lès-Nancy, France
(2) Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
(3) CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{arseniy.gorin, denis.jouvet}@inria.fr

RÉSUMÉ

Lorsque l'on considère de la parole produite par des enfants et des adultes, la variabilité acoustique de chaque unité phonétique devient grande, ce qui dégrade les performances de reconnaissance. Un moyen d'aller au-delà des modèles de Markov traditionnels, est de prendre en considération des classes de locuteurs. Les classes de locuteurs peuvent être obtenues automatiquement. Elles servent à fabriquer des modèles acoustiques spécifiques de chaque classe. Ce papier propose une structuration des composantes des densités multigaussiennes (GMMs) en relation avec des classes de locuteurs. Dans une première approche, cette structuration des densités est complétée par des pondérations des composantes gaussiennes dépendantes des classes de locuteurs, et dans une deuxième approche, par des matrices de transition entre les composantes gaussiennes des densités (comme dans les *stranded GMMs*). Ces deux approches apportent des gains substantiels pour la reconnaissance de voix d'enfants et d'adultes. La structuration des composantes gaussiennes complétée par des matrices de transition entre composantes réduit de plus d'un tiers le taux d'erreur mot sur le corpus TIDIGIT.

ABSTRACT

Explicit trajectories and speaker class modeling for child and adult speech recognition

When the speech data is produced by speakers of different age and gender, the acoustic variability of any given phonetic unit becomes large, which degrades speech recognition performance. One way to go beyond conventional Hidden Markov Model is to explicitly include speaker class information in the modeling. Speaker classes can be obtained automatically, and they are used for building speaker class-specific acoustic models. This paper introduces a structuring of the Gaussian components of the GMM densities with respect to the speaker classes. In a first approach, this structuring of the Gaussian components is completed with speaker class-dependent mixture weights, and in a second approach, with transition matrices, which add dependencies between Gaussian components of mixture densities (as in stranded GMMs). The two approaches bring substantial performance improvements when recognizing adult and child speech. Using class-structured components plus mixture transition matrices reduces by more than one third the word error rate on the TIDIGIT corpus.

MOTS-CLÉS : Reconnaissance de la parole ; classification non supervisée ; modèles de classes de locuteurs ; modèles stochastiques de trajectoire.

KEYWORDS: Speech recognition ; unsupervised clustering ; speaker class modeling ; stochastic trajectory modeling.

1 Introduction

Le traitement de la variabilité interlocuteur reste un problème pour les systèmes de reconnaissance automatique de la parole (RAP) ; elle est due entre autres au sexe du locuteur, à son âge et à son accent (Benzeghiba *et al.*, 2007; Stern et Morgan, 2012). La variabilité résultante des paramètres acoustiques doit être prise en compte par les modèles acoustiques indépendants du locuteur ; or les modèles de Markov cachés avec densités multigaussiennes (HMM-GMM : *Hidden Markov Model with Gaussian Mixture Model observation densities*) ne sont pas capables de représenter précisément des distributions très hétérogènes de paramètres acoustiques. Les techniques d'adaptation permettent d'améliorer la précision, en modifiant les paramètres acoustiques (par exemple, VTLN (Zhan et Waibel, 1997), fMLLR (Gales, 1998)) ou les paramètres des modèles (par exemple, MLLR, MAP (Gauvain et Lee, 1994)).

Ce papier traite de la situation générale où les données proviennent de locuteurs d'âge et de sexe différents, et dont la classe d'appartenance est inconnue pour le système de reconnaissance. Une classification automatique peut alors être appliquée sur chaque phrase, en supposant que le locuteur ne change pas au cours de la phrase (Beaufays *et al.*, 2010). La quantité de données disponibles pour l'apprentissage du modèle de chaque classe diminue lorsque le nombre de classes augmente, et cela peut rendre la modélisation moins fiable. Le manque de données peut être partiellement compensé par l'utilisation d'approches comme les voix propres, où les paramètres pour un locuteur sont déterminés comme une combinaison de modèles de classes (Kuhn *et al.*, 1998), ou en élargissant les classes de données pour l'apprentissage en autorisant une donnée d'apprentissage à appartenir à plusieurs classes (Jouvet *et al.*, 2012; Gorin et Jouvet, 2012).

Alors que dans les approches HMM (*Hidden Markov Model*) conventionnelles, les composantes des GMM (*Gaussian Mixture Models*) sont apprises indépendamment pour chaque densité, la structuration des composantes des densités en fonction de classes de locuteurs conduit à des GMM pour lesquels la $k^{ème}$ composante de chaque densité est associée à une même classe de données. La structuration en classes des HMM-GMM, résulte de l'initialisation des GMM à partir de GMM de plus faible dimension (i.e. ayant moins de composantes), et a été initialement proposée et étudiée pour la transcription d'émissions radio (Gorin et Jouvet, 2013). Pour cette modélisation HMM-SWGMM (*HMM with Speaker class-dependent Weights*), les composantes gaussiennes des mélanges GMM sont structurées en fonction des classes et partagées entre toutes les classes, tandis que les pondérations des composantes des densités sont dépendantes de la classe.

Dans ce papier, nous proposons de combiner la structuration des composantes des densités GMM en fonction des classes avec l'utilisation de matrices de transition entre composantes (MTM : *Mixture Transition Matrices*) comme dans les SGMM (*Stranded GMM*). Les SGMM sont similaires aux modèles Gaussiens conditionnels (Wellekens, 1987) qui ont été récemment étendus, reformulés et étudiés pour la reconnaissance robuste (Zhao et Juang, 2012). Dans les SGMM, les matrices de transition MTM définissent les dépendances entre les composantes gaussiennes des densités GMM adjacentes. Alors que les SGMM étaient originellement initialisés à partir de HMM-GMM conventionnels, la modélisation SSGMM (*class-Structured SGMM*) proposée dans ce papier combine SGMM et structuration des composantes en classes. Les matrices MTM modélisent alors la probabilité de rester sur une composante de la même classe au cours du temps, ou de passer vers une composante d'une autre classe. Cette modélisation explicite de trajectoires améliore la précision des modèles acoustiques. De plus, cette approche ne requiert pas d'étape additionnelle

de classification pour estimer la classe à laquelle appartient la phrase à décoder.

Le papier est organisé de la manière suivante. La section 2 présente le problème de la reconnaissance de voix d'enfants et d'adultes. La section 3 commente l'utilisation de modèles de classes. La section 4 introduit la structuration en classes pour les HMM-SWGMM. La section 5 rappelle le principe des SGMM et détaille la modélisation SSGMM proposée qui combine la structuration en classes des composantes et l'utilisation de matrices de transition entre composantes.

2 Système de référence et formulation du problème

Les expériences présentées dans ce papier ont été menées sur le corpus TIDIGIT (Leonard et Doddington, 1993) de chiffres connectés en anglais ; l'un des rares corpus disponibles à contenir à la fois des voix d'adultes et des voix d'enfants. L'ensemble d'apprentissage contient 41224 occurrences de chiffres (28329 prononcés par des adultes et 12895 par des enfants). L'ensemble de test contient 41087 occurrences de chiffres (28554 par des adultes et 12533 par des enfants). Les outils Sphinx3 (CMU, 2014) ont été utilisés, et enrichis pour traiter les SGMM (matrices de transition entre composantes - cf. Section 5). Les chiffres sont modélisés par des séquences de phones dépendants du mot. Chaque phonème est modélisé par un HMM à 3 états, sans saut. Chaque densité d'émission est une densité multigaussienne à 32 composantes. L'analyse acoutique calcule 13 coefficients (12 MFCC et le logarithme de l'énergie), auxquels on adjoint les dérivées premières et secondes. Le signal est sous-échantillonné à 8kHz, comme indiqué dans d'autres publications traitant ce corpus TIDIGITS (ex. (Burnett et Fanty, 1996)). Les taux d'erreur mot (WER : *Word Error Rate*) du système de référence sont indiqués dans la Table 1. Deux modèles indépendants du locuteur (SI : *Speaker Independent*) ont été utilisés, l'un appris sur les données provenant des adultes, et l'autre appris sur l'ensemble d'apprentissage complet (adultes et enfants). Les deux dernières lignes indiquent les performances obtenues avec des modèles adaptés par MLLR+MAP, en fonction de l'âge, ou en fonction du sexe et de l'âge.

	Adulte	Hom.	Fem.	**Enfant**	Garç.	Fille
Apprentissage sur données adultes	**0,64**	0,79	0,49	**9,92**	6,51	13,33
Apprentissage sur données adultes+enfants	**1,66**	1,86	1,46	**1,88**	1,69	2,08
+ adapt. âge (classe connue pour décodage)	**1,42**	1,56	1,28	**1,56**	1,52	1,54
+ adapt. sexe+âge (classe connue pour décodage)	**1,31**	1,57	1,04	**1,31**	1,14	1,49

TABLE 1: Taux d'erreur mot avec des HMM conventionnels pour les modèles indépendants du locuteur, et pour les modèles adaptés avec évaluation oracle (i.e. classe connue pour décodage)

L'apprentissage sur les données adulte seules fournit les meilleures performances pour les locuteurs adultes, mais de piètres performances pour les données des enfants. Quand des données d'enfants sont incluses dans le corpus d'apprentissage, les performances s'améliorent pour les enfants, mais se dégradent pour les adultes. L'utilisation de modèles adaptés, soit à l'âge du locuteur (i.e. adultes v.s. enfants) soit au sexe et à l'âge du locuteur (hommes / femmes / garçons / filles) améliore les performances. Dans la suite, l'apprentissage des modèles sera effectué en utilisant l'ensemble d'apprentissage complet (adultes + enfants), et l'information *a priori* (âge & sexe) ne sera pas utilisée (i.e. les classes de données seront estimées automatiquement).

3 Classification non supervisée et modélisation multiple

L'objectif d'une classification non supervisée est de regrouper automatiquement les données d'apprentissage en classes correspondant à des données acoustiquement similaires. Une approche de

classification automatique reposant sur des modèles multigaussiens a été présentée dans (Jouvet *et al.*, 2012). Un GMM générique est d'abord appris sur l'ensemble des données d'apprentissage. Puis, ce GMM est dupliqué, et les moyennes des gaussiennes légèrement modifiées. Ensuite les données d'apprentissage sont classifiées avec ces deux GMM, déterminant ainsi deux classes de données à partir desquelles les paramètres des deux GMM sont réestimés. Ces phases de classification et d'apprentissage sont itérées jusqu'à convergence. Ensuite, on itère si nécessaire à partir de l'étape de duplication des GMM, jusqu'à obtenir le nombre désiré de classes. Les modèles acoustiques indépendants du locuteur sont alors adaptés (MLLR+MAP) pour chaque classe en utilisant les données correspondantes.

Analyse de la classification automatique sur les données adultes et enfants avec des GMM ayant 256 composantes. La figure 1 présente la répartition des données d'apprentissage de chaque classe par rapport à l'âge et au sexe des locuteurs. Les deux premières classes séparent les données entre hommes d'un coté, et femmes et enfants de l'autre. Avec 4 classes, on voit apparaître une séparation entre les voix de femmes et les voix d'enfants. Par contre, il semble impossible de séparer les voix de garçons des voix de filles, même en augmentant le nombre de classes.

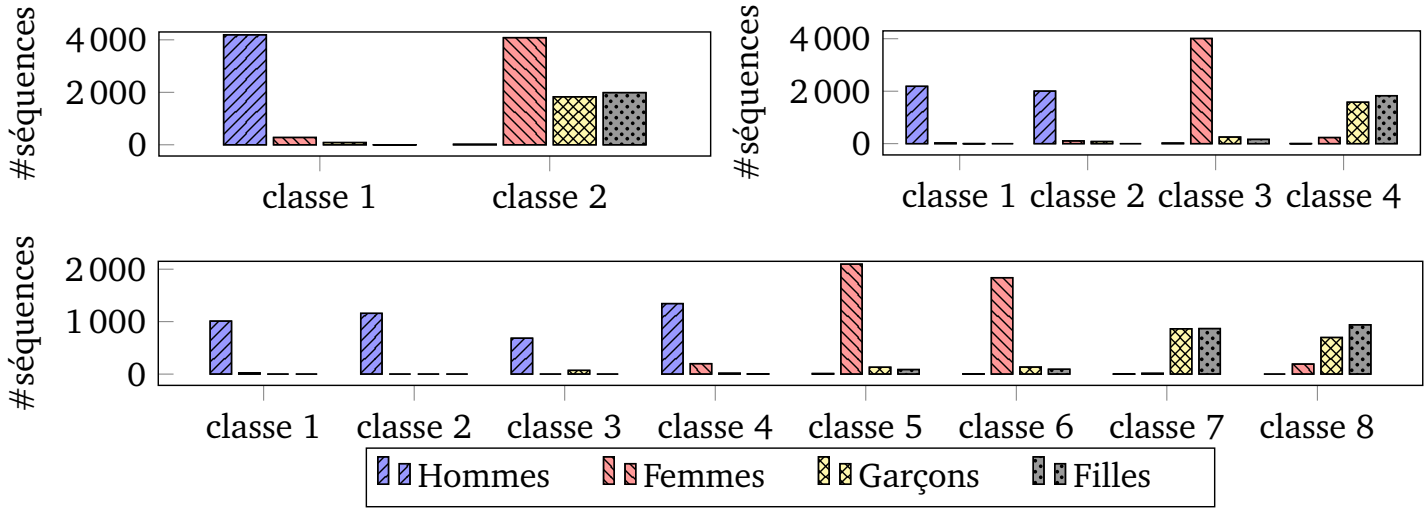


FIGURE 1: Nombre de séquences par catégorie de locuteurs en fonction du nombre de classes

Lors du décodage, chaque phrase est automatiquement classifiée et le décodage est effectué avec les modèles acoustiques correspondant à la classe identifiée. Les taux d'erreur mot, avec les intervalles de confiance à 95% associés, sont représentés par les barres "*CA-GMM*" de la figure 4. Les meilleures performances sont obtenues avec 4 classes (cf. ligne "*4 cla. CA-GMM*" de la table 2) et elles sont similaires à celles obtenues avec les modèles adaptés en fonction de l'âge et du sexe des locuteurs (cf. table 1). Au delà de 4 classes, les performances se détériorent en raison de la diminution de la quantité de données disponibles pour l'adaptation des modèles des classes.

4 Structuration des composantes pour les HMM-SWGMM

Au lieu d'adapter tous les paramètres des HMM pour chaque classe de données, l'approche HMM-SWGMM (*HMM-GMM with Speaker class-dependent Weights*) récemment proposée (Gorin et Jouvet, 2013) repose sur la structuration en classes des composantes gaussiennes, le partage des composantes gaussiennes entre les différentes classes, et la spécialisation des pondérations des gaussiennes à chaque classe. Ainsi, dans un HMM-SWGMM, la densité b_j pour une classe de locuteurs c est définie par $b_j^{(c)}(o_t) = \sum_{k=1}^{M} w_{jk}^{(c)} \mathcal{N}(o_t, \mu_{jk}, U_{jk})$, où M est le nombre de composantes par densité, o_t est le vecteur d'observation au temps t et $\mathcal{N}(o_t, \mu_{jk}, U_{jk})$ est la

composante gaussienne de moyenne μ_{jk} et de covariance U_{jk}. Lors du décodage, chaque phrase à reconnaître est d'abord automatiquement classifiée, et assignée à une classe c. Ensuite, le décodage est effectué avec le jeu de pondérations des gaussiennes qui correspond à cette classe.

La structuration des composantes gaussiennes est obtenue en initialisant les GMM par concaténation des composantes gaussiennes de GMM de plus faible dimensionalité appris sur les différentes classes. Par exemple, pour fabriquer un modèle avec M composantes gaussiennes associées à Z classes, Z modèles ayant chacun $L = M/Z$ composantes par densité sont appris. Puis, ces composantes sont regroupées dans un mélange global (cf. figure 2-a pour les moyennes).

(a)

$$\begin{bmatrix} \mu_{j1}^{(c_1)} \\ \dots \\ \mu_{jL}^{(c_1)} \end{bmatrix} \dots \begin{bmatrix} \mu_{j1}^{(c_Z)} \\ \dots \\ \mu_{jL}^{(c_Z)} \end{bmatrix} \Rightarrow \begin{bmatrix} \mu_{j1} \\ \dots \\ \mu_{jL} \\ \dots \\ \dots \\ \mu_{M-L+1} \\ \dots \\ \mu_{M} \end{bmatrix}$$

(b)

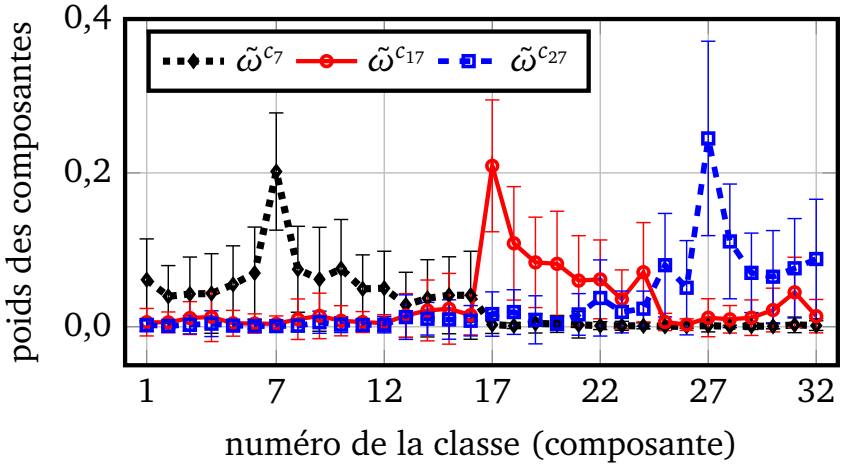


FIGURE 2: (a) Initialisation du SW-GMM à partir de plusieurs modèles associés aux classes et (b) statistiques des pondérations des composantes pour les classes c_7, c_{17} et c_{27} après ré-estimation (moyennes et écarts-types calculés sur toutes les densités ; Z=32, M=32)

Lors de l'initialisation du modèle structuré SW-GMM, les pondérations des modèles des classes sont également concaténées, puis renormalisées. Ensuite, les moyennes, variances et pondérations sont ré-estimées. Les pondérations, spécifiques à chaque classe, sont apprises à partir des données de la classe correspondante, tandis que toutes les données sont utilisées pour ré-estimer les moyennes et les variances qui sont partagées entre les classes :

$$\omega_{jk}^{(c_i)} = \frac{\sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)}{\sum_{l=1}^{M} \sum_{t=1}^{T} \gamma_{jl}^{(c_i)}(t)} \qquad \mu_{jk} = \frac{\sum_{i=1}^{Z} \sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t) o_t}{\sum_{i=1}^{Z} \sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)} \tag{1}$$

où $\gamma_{jk}^{(c_i)}(t)$ est le compteur Baum-Welch de la $k^{ème}$ composante de la densité b_j, générant l'observation o_t de la classe c_i. Les sommes sur t portent sur toutes les trames des phrases d'apprentissage de la classe concernée. Les variances sont ré-estimées de manière similaire aux moyennes.

Après ré-estimation, les pondérations dépendantes de la classe sont plus grandes pour les composantes associées à la même classe de données ; la figure 2-b montre l'exemple des pondérations des classes c_7, c_{17} et c_{27} , moyennés sur les toutes les densités.

Expériences avec les modèles HMM-SWGMM structurés en fonction des classes. Les modèles HMM-SWGMM à 32 gaussiennes par densité sont initialisés à partir de modèles de classes ayant moins de composantes, par exemple 2 modèles de classes à 16 gaussiennes, ou 4 modèles à 8 gaussiennes, et ainsi de suite jusqu'à 32 modèles de classes monogaussiens. Les paramètres sont ré-estimés au maximum de vraisemblance (MLE) pour les pondérations des gaussiennes dépendantes de chaque classe, et par maximum *a posteriori* (MAP) pour les moyennes et les variances partagées. Les performances sont indiquées par les barres "*SW-GMM*" dans la figure 4. Cette paramétrisation permet une estimation robuste des moyennes et des variances partagées, tout en gardant une dépendance vis-à-vis des classes avec les pondérations des composantes.

Avec un nombre limité de paramètres, le taux d'erreur est significativement réduit ; 0,80% sur les données des adultes et 1,05% sur les données enfants (cf. ligne *32 cla. SW-GMM* de la table 2).

5 Structuration des composantes pour les SGMM

Les modèles SGMM (*Stranded GMM*), proposés pour la reconnaissance robuste (Zhao et Juang, 2012), reposent sur un enrichissement des densités d'émission des HMM-GMM par l'introduction de dépendances explicites entre les composantes des densités d'états adjacents. Tandis que dans la version proposée par Zhao, les SGMM sont initialisés à partir de HMM-GMM conventionnels, cette section propose d'exploiter la structuration des composantes en fonction de classes, pour obtenir des SSGMM (*class-Structured SGMM*).

5.1 SGMM Conventionnels

Les modèles SGMM conventionnels font intervenir la suite des états $\mathscr{Q} = \{q_1, ..., q_T\}$, la suite des vecteurs d'observation $\mathscr{O} = \{o_1, ..., o_T\}$, et la suite des composantes des densités $\mathscr{M} = \{m_1, ..., m_T\}$, où $m_t \in \{1, ..., M\}$ désigne la composante gaussienne utilisée à l'instant t, et M précise le nombre de composantes par densité.

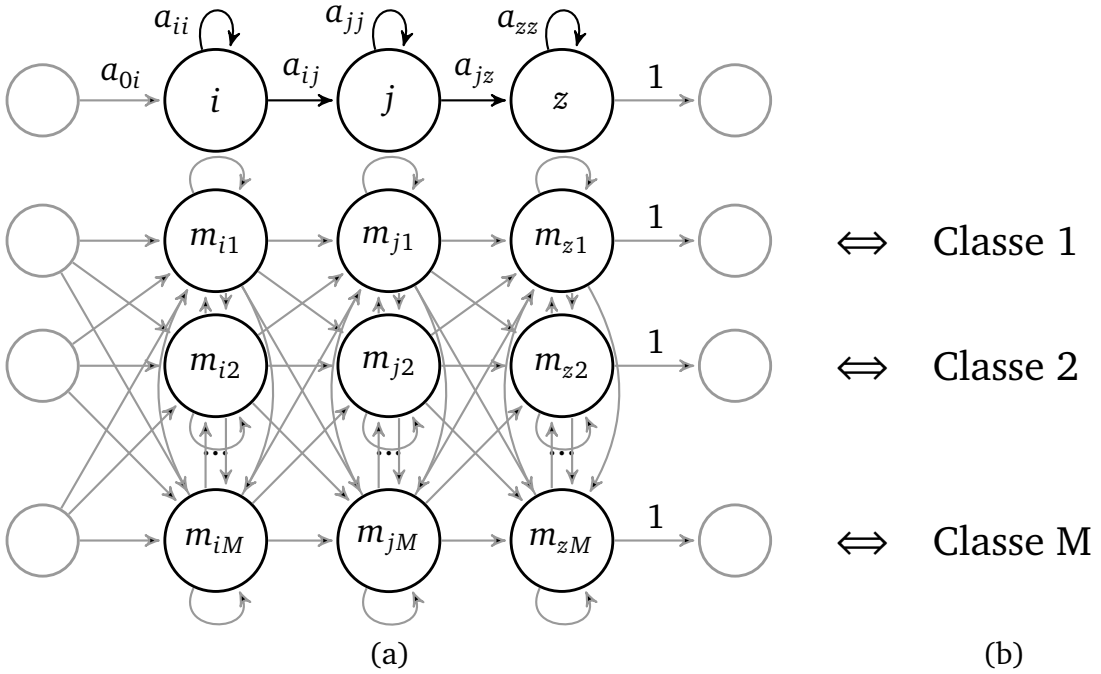


FIGURE 3: (a) *Stranded* GMM avec représentation des dépendances entre composantes gaussiennes ; (b) associations entre classes et composantes gaussiennes pour les SSGMM.

En comparaison des HMM-GMM conventionnels, les SGMM modélisent les dépendances entre la composante gaussienne m_t utilisée à l'instant t et la composante m_{t-1} utilisée à la trame précédente (cf. figure 3-a). La vraisemblance conjointe de la séquence d'observations, de la séquence d'états, et de la séquence de composantes est donnée par :

$$P(\mathscr{O}, \mathscr{Q}, \mathscr{M}|\lambda) = \prod_{t=1}^{T} P(o_t|m_t, q_t) P(m_t|m_{t-1}, q_t, q_{t-1}) P(q_t|q_{t-1}) \quad (2)$$

où $P(q_t = j|q_{t-1} = i) = a_{ij}$ est la probabilité de transition entre états, $P(o_t|m_t = l, q_t = j) = b_{jl}(o_t)$ est la probabilité de l'observation o_t pour la composante gaussienne $m_t = l$ de la densité associée à l'état $q_t = j$ et $P(m_t = l|m_{t-1} = k, q_t = j, q_{t-1} = i) = c_{kl}^{(ij)}$ est la probabilité de transition entre les composantes gaussiennes. L'ensemble des probabilités de transition

entre les composantes détermine les matrices de transition MTM (*Mixture Transition Matrices*) $C^{(ij)} = \{c_{kl}^{(ij)}\}$, avec les contraintes $\sum_{l=1}^{M} c_{kl}^{(ij)} = 1, \forall i, j, k$.

Expériences avec les SGMM conventionnels. Afin de réduire le nombre de paramètres, 2 matrices MTM seulement sont utilisées pour chaque état, l'une correspondant au bouclage sur l'état, et l'autre au passage à l'état suivant (les matrices MTM associées au dernier état d'un phonème et correspondant aux transitions vers un autre phonème sont partagées). Les taux d'erreur mot obtenus sont indiqués par les barres "*SGMM*" dans la figure 4 et dans la ligne correspondante de la table 2. Par rapport aux HMM-GMM conventionnels, les SGMM améliorent les performances, de 1,66 % à 1,11 % sur les données adultes et de 1,88 % à 1,27 % sur les données enfants. Ces améliorations sont significatives, compte tenu de l'intervalle de confiance à 95 %. L'approche SGMM, qui fonctionne en une seule passe, aboutit même à des performances meilleures que la modélisation de référence adaptée en fonction de l'âge et du sexe des locuteurs, sans toutefois dépasser les performances de l'approche HMM-SWGMM proposée dans la section précédente.

5.2 SGMM avec structuration en classes des composantes

La modélisation SSGMM (*class-Structured SGMM*) proposée repose sur la structuration en classes des composantes des densités, de telle manière que initialement, la $k^{ème}$ composante de chaque densité corresponde à une même classe de données (cf. figure 3-b). Pour obtenir cette structuration, le SSGMM est initialisé à partir du HMM-SWGMM décrit dans la section 4. Les moyennes et variances des gaussiennes sont obtenues directement à partir du HMM-SWGMM et les matrices MTMs sont initialisées avec des distributions uniformes. Les pondérations des gaussiennes du HMM-SWGMM, qui sont dépendantes des classes, ne sont pas utilisées.

Quand les HMM-SWGMM, qui servent à fabriquer les SSGMM, sont initialisés à partir ce modèles de classes monogaussiens, chaque composante correspond à une classe. Après ré-estimation, les éléments diagonaux des matrices MTM dominent, ce qui conduit à favoriser la consistance de la classe lors du décodage d'une phrase. Cependant, les éléments non-diagonaux non nuls rendent possibles les contributions d'autres composantes gaussiennes dans le calcul des scores acoustiques. L'avantage des SSGMM est qu'ils modélisent explicitement les trajectoires, tout en autorisant des changements de composantes (ou de classes). De plus, le décodage d'une phrase fonctionne en une seule passe ; il n'y a pas de classification préalable à faire.

Modèle	Decodage	Paramètres par état	**Adult.**	Hom.	Fem.	**Enf.**	Garç.	Fil.
SI GMM	1 passe	78*32+32=2528	1,66	1,86	1,46	1,88	1,69	2,08
4 cla. CA-GMM	2 passes	4*(78*32+32)=10112	**1,32**	1,47	1,17	**1,57**	1,61	1,52
8 cla. HMM-SWGMM	2 passes	78*32+8*32=2752	**0,75**	0,81	0,70	**1,21**	1,25	1,17
32 cla. HMM-SWGMM	2 passes	78*32+32*32=3520	**0,80**	0,96	0,64	**1,05**	1,03	1,08
SGMM	1 passe	78*32+2*32*32=4544	**1,11**	1,26	0,96	**1,27**	1,19	1,36
SSGMM	1 passe	78*32+2*32*32=4544	**0,52**	0,64	0,40	**0,86**	0,74	0,98

TABLE 2: Taux d'erreur mot et indication du nombre de paramètres par état. La classe du locuteur est inconnue lors du décodage, et estimée par GMM pour les versions "2 passes"

Expériences avec les SGMM structurés en classes. Pour cette expérience, les SSGMM sont fabriqués à partir du HMM-SWGMM à 32 classes (i.e. modèle *32 cla. HMM-SWGMM* dans la table 2). De même que précédemment, deux matrices MTM sont alors définies pour chaque état, et

initialisées avec des distributions uniformes, puis les paramètres des SSGMM sont ré-estimés par maximum de vraisemblance. Les performances sont indiquées avec les barres "*SSGMM* " dans la figure 4 et dans la ligne correspondante de la table 2. L'approche SSGMM proposée, qui repose sur la structuration en classes des composantes, améliore encore plus que les SGMM conventionnels, et permet d'obtenir un taux d'erreur mot de 0,52% sur les données adultes, et de 0,86% sur les données enfants. A noter que la fabrication de SSGMM à partir de HMM-SWGMM construits sur la base de différents nombres de classes (2, 4, 8 et 16) conduit aussi à des améliorations de performances par rapport au SGMM ; seul le résultat correspondant à 32 classes est indiqué.

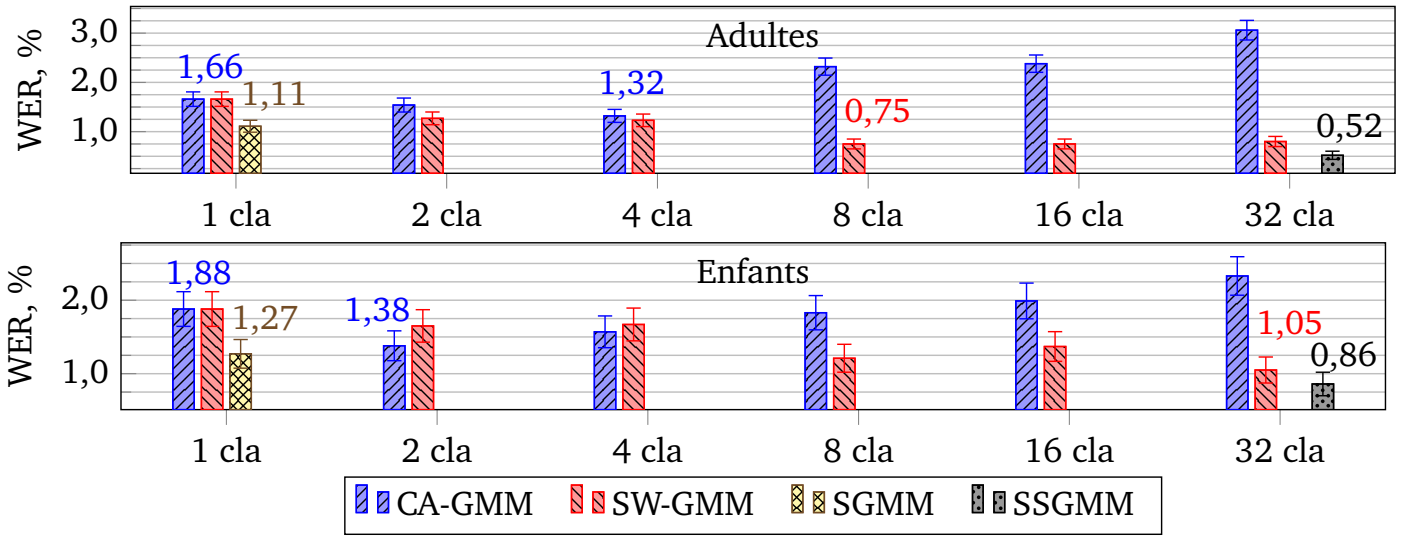


FIGURE 4: Taux d'erreur mot (WER) sur les données adultes et enfants, obtenus avec les CA-GMM (*Class-Adapted models*), les HMM-SWGMM (*HMM with Speaker class-dependent Weights*), les SGMM (*Stranded GMM*) et les SSGMM (*class-Structured SGMM*) contruits à partir de 32 classes

6 Conclusion et perspectives

Plusieurs approches pour améliorer la reconnaissance de données hétérogènes ont été étudiées. La taille limitée des corpus d'apprentissage ne permet pas de fabriquer un grand nombre de modèles de classes. Cependant, ces modèles de classes permettent de structurer les composantes gaussiennes des densités GMM en associant la $k^{ème}$ composante de chacune des densités à une même classe de locuteurs.

Deux types de modélisations exploitant cette structuration des composantes ont été proposées et approfondies, et ont montré un gain significatif en performances. Le premier modèle, HMM-SWGMM (*Speaker class-dependent Weights GMM*) repose sur un partage des composantes gaussiennes entre les différentes classes, et des pondérations des composants gaussiennes spécifiques à chaque classe. Ce partage des paramètres rend la modélisation robuste. La structuration des composantes a ensuite été appliquée aux *Stranded* GMM qui modélisent explicitement les trajectoires grâce à l'adjonction de matrices de transition entre les composantes des densités GMM. La modélisation résultante, SSGMM (*class-Structured SGMM*), conduit à un gain significatif en performances par rapport aux autres approches, pour la reconnaissance de données hétérogènes (voix d'adultes et voix d'enfants). De plus cette approche ne requiert pas de phase préalable de classification des données à reconnaître.

Une amélioration possible consistera à combiner les modélisations proposées avec une adaptation des paramètres reposant sur l'ajustement de l'axe fréquentiel, comme par exemple la normalisation VTLN (*Vocat Tract Length Normalization*).

Références

BEAUFAYS, F., VANHOUCKE, V. et STROPE, B. (2010). Unsupervised Discovery and Training of Maximally Dissimilar Cluster Models. *In Proc. INTERSPEECH*, pages 66–69, Makuhari, Japan.

BENZEGHIBA, M., DE MORI, R., DEROO, O., DUPONT, S., ERBES, T., JOUVET, D., FISSORE, L., LAFACE, P., MERTINS, A., RIS, C., TYAGI, V. et WELLEKENS, C. (2007). Automatic speech recognition and speech variability : A review. *Speech Communication*, 49(10):763–786.

BURNETT, D. C. et FANTY, M. (1996). Rapid unsupervised adaptation to children's speech on a connected-digit task. *In Proc. ICSLP*, volume 2, pages 1145–1148. IEEE.

CMU (2014). Sphinx toolkit http ://cmusphinx.sourceforge.net.

GALES, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2):75–98.

GAUVAIN, J.-L. et LEE, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and audio processing, IEEE transactions on*, 2(2):291–298.

GORIN, A. et JOUVET, D. (2012). Class-based speech recognition using a maximum dissimilarity criterion and a tolerance classification margin. *In Proc. Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 91–96. IEEE.

GORIN, A. et JOUVET, D. (2013). Efficient constrained parametrization of GMM with class-based mixture weights for Automatic Speech Recognition. *In Proc. LTC-6th Language & Technologies Conference*, pages 550–554.

JOUVET, D., GORIN, A. et VINUESA, N. (2012). Exploitation d'une marge de tolérance de classification pour améliorer l'apprentissage de modèles acoustiques de classes en reconnaissance de la parole. *In JEP-TALN-RECITAL*, pages 763–770.

KUHN, R., NGUYEN, P., JUNQUA, J.-C., GOLDWASSER, L., NIEDZIELSKI, N., FINCKE, S., FIELD, K. et CONTOLINI, M. (1998). Eigenvoices for speaker adaptation. *In Proc. ICSLP*, volume 98, pages 1774–1777.

LEONARD, R. G. et DODDINGTON, G. (1993). Tidigits speech corpus. *Texas Instruments, Inc*.

STERN, R. M. et MORGAN, N. (2012). Hearing is believing : Biologically inspired methods for robust automatic speech recognition. *Signal Processing Magazine, IEEE*, 29(6):34–43.

WELLEKENS, C. J. (1987). Explicit time correlation in hidden Markov models for speech recognition. *In Proc. ICASSP*, pages 384–386.

ZHAN, P. et WAIBEL, A. (1997). Vocal tract length normalization for large vocabulary continuous speech recognition. *In Technical report*, DTIC Document.

ZHAO, Y. et JUANG, B.-H. (2012). Stranded Gaussian mixture hidden Markov models for robust speech recognition. *In Proc. ICASSP*, page 4301–4304.

Étude des productions des consonnes du français par des apprenants vietnamiens du Français Langue Etrangère

Thi Thuy Hien TRAN[1], Nathalie VALLEE[1], Quang Thuan NGUYEN[2]
[1] GIPSA-lab DPC, UMR 5216, CNRS & Université de Grenoble Alpes, BP25 38040 Grenoble Cedex 9, France
[2] École Supérieure des Langues Étrangères, Université Nationale de Hanoi, Vietnam
thi-thuy-hien.tran@gipsa-lab.grenoble-inp.fr,
nathalie.vallee@gipsa-lab.grenoble-inp.fr, ngquangthuan@yahoo.fr

RESUME

Ce travail s'inscrit dans les recherches sur l'apprentissage des langues étrangères et traite plus particulièrement de l'acquisition des consonnes par des apprenants vietnamiens de deux niveaux, intermédiaire et avancé. Il s'agit d'une étude expérimentale sur la réalisation des consonnes en pseudo-mots et mots en tenant compte de leur position dans le mot, dans la syllabe. Dans une tâche de répétition, les stratégies de production des consonnes du français sont analysées typologiquement et quantitativement. L'objectif global de ce travail est de situer et comprendre les difficultés rencontrées par les étudiants vietnamiens, même de niveau avancé, à réaliser les consonnes simples et/ou les consonnes en séquences du français. Les résultats sont analysés par rapport aux éléments du crible phonologique de la L1 et par rapport aux tendances universelles des langues.

ABSTRACT

Acoustical study of French consonant productions by Vietnamese learners of French as a foreign language

This work addresses foreign language learning and specifically acquisition of French consonants by Vietnamese learners in both intermediate and advanced levels. It is an experimental study on production of French consonants, taking into account their position in the word and syllable. In a repetition task, production strategies of French consonants are analyzed typologically and quantitatively. The general goal of this work is to locate and understand difficulties encountered by Vietnamese learners, even advanced learners, when pronouncing French consonants and clusters. Results are compared to factors relating to L1 phonological background. They are also analyzed relative to universal properties of languages, and thus of language in general.

MOTS-CLES : acquisition, langue seconde, perception et production, consonnes, FLE, vietnamien
KEYWORDS: second language acquisition, perception and production, consonants, FFL, Vietnamese

1 Introduction

Les études en acquisition de la phonologie d'une langue seconde (L2) cherchent à expliquer et comprendre les processus d'apprentissage, les éléments favorisants comme les difficultés rencontrées par les apprenants. Les premières recherches en acquisition de la L2 se sont centrées sur le rôle de la langue première (L1). De nombreux travaux ont suggéré que phonétique et phonologie de la L1 ont une influence importante sur la prononciation des sons de la L2 (Polivanov, 1931 ; Troubetzkoy, 1939). Les recherches sur les aspects phonologiques de l'interlangue ont connu des avancées intéressantes avec les travaux qui ont considéré le rôle de la L1 à partir des différences et similarités par rapport à la L2. De ces travaux sont nés des modèles

d'explication et de prédiction des erreurs commises par les apprenants d'une L2 (Lado, 1957 ; Best, 1995 ; Flege, 1995 entre autres). Si un certain nombre d'erreurs de perception et/ou de prononciation des sons de la L2 sont attribuables à la L1 de l'apprenant, d'autres ne le sont pourtant pas. Les recherches liées aux propriétés typologiques de la L1 et de la L2 révèlent que les différences de perception et production de la parole entre locuteurs natifs et non natifs ne peuvent pas toutes être attribuées au transfert de la L1 et que des principes universels qui structurent les systèmes sonores et l'utilisation de leurs unités dans la chaîne parlée influencent considérablement l'apprentissage d'une L2 (voir par ex. Eckman, 1977 ; Weinberger, 1987 ; Carlisle, 1998). L'Hypothèse de la Différence de Marquage formulée par Eckman en 1977 prédit que les formes de la L2 qui sont différentes et plus marquées que celles de la L1 seront les plus difficiles à produire et que les formes moins marquées sont acquises avant les plus marquées.

Le travail présenté dans ces pages s'inscrit dans ce cadre de recherche sur l'apprentissage des langues étrangères et traite plus particulièrement de l'acquisition des consonnes en langue seconde en tentant de poser des éléments de la structure phonologique de l'interlangue. Les deux langues observées sont le vietnamien (langue source) et le français (langue cible). Vietnamien et français sont deux langues pour lesquelles de nombreuses dissemblances phonético-phonologiques existent dont une qui nous intéresse en particulier ici : les séquences consonantiques. Elles peuvent être hétérosyllabiques, c'est-à-dire réparties de part et d'autre d'une frontière syllabique : c'est le cas en vietnamien et en français. Elles peuvent être tautosyllabiques, c'est-à-dire situées dans un constituant sub-syllabique, soit l'attaque (initiale), soit la coda (finale) : c'est le cas en français mais pas en vietnamien. Langue tonale, le vietnamien est monosyllabique sur le plan phonologique mais aussi polysyllabique en partie sur le plan lexical. La différence entre mots simple et composé n'existe que par le nombre de syllabes (Doan, 1999). Les patrons syllabiques possibles du vietnamien sont de structures C_1(w)V(C_2) (entre parenthèses, les constituants facultatifs) où l'inventaire des C_2 est très restreint (/p t k m n ŋ/). La diversité des patrons syllabiques est plus large en français : (C_1)(C_2)(C_3)V(C_4)(C_5)(C_6)(C_7) (Rousset, 2004), attaque et coda pouvant être occupées par des groupes de consonnes. Les enseignants du Français Langue Étrangère (FLE) relatent souvent que les étudiants vietnamiens font partie de ces apprenants qui éprouvent des difficultés récurrentes dans la réalisation des consonnes finales et en particulier des groupes de consonnes (Nguyen, 2000 ; Nguyen, 2001). Les suites de deux ou plusieurs consonnes n'étant pas permises en vietnamien à l'intérieur d'une même syllabe, elles sont la plupart du temps réalisées déformées par rapport à la cible, entraînant chez l'auditeur incompréhension ou malentendu (Nguyen, 2000).

Bien que les deux langues présentent des phonèmes identiques, plusieurs études ont montré que leurs réalisations peuvent différer d'une langue à l'autre (Ladefoged et Maddieson, 1996 ; Ladefoged, 2001). C'est le cas, par exemple en vietnamien, des plosives /p t k/ non relâchées en position de coda mais réalisées en français avec un bruit d'explosion en général audible. Avant d'étudier la perception et production des séquences de consonnes du français par des apprenants vietnamiens, une étape préliminaire s'est imposée : l'analyse de la perception et de la production des consonnes simples en fonction de leur position dans la syllabe. L'objectif de cette présente étude est double : (1) tester chez des apprenants vietnamiens, la récupération des consonnes du français et leurs caractéristiques articulatoires à partir du signal acoustique ; (2) analyser typologiquement et quantitativement les stratégies de production des consonnes du français dans une tâche de répétition, et plus particulièrement des consonnes /p t k m n ŋ/, les seules permises en coda en vietnamien et par conséquent, les seules à initier, dans cette langue, les séquences de plusieurs consonnes.

2 Méthodologie

2.1 Constitution du corpus

Dix-huit consonnes du français /p t k b d g m n ɲ ŋ f s ʃ v z ʒ l ʁ/ ont été sélectionnées pour figurer dans le corpus. Des pseudo-mots et mots contenant ces consonnes en contexte /a/ ont été construits en prenant en compte les différentes distributions : position initiale, inter-vocalique et finale dans la syllabe (cf. table 1). L'utilisation des pseudo-mots vise à éviter un éventuel effet de la fréquence lexicale des stimuli à produire et à percevoir. Pour un contraste CV maximal facilitant le traitement des données, le contexte /a/ a été choisi. Etant donné que /ŋ/ ne figure jamais ni en initiale ni en inter-vocalique dans le lexique du français, un seul pseudo-mot contenant cette consonne en position finale figure par conséquent dans la liste des items construits.

Initiale	Inter-vocalique	Finale	Initiale	Inter-vocalique	Finale
p a	a p a	a p	f a	a f a	a f
t a	a t a	a t	s a	a s a	a s
k a	a k a	a k	ʃ a	a ʃ a	a ʃ
b a	a b a	a b	v a	a v a	a v
d a	a d a	a d	z a	a z a	a z
g a	a g a	a g	ʒ a	a ʒ a	a ʒ
m a	a m a	a m	l a	a l a	a l
n a	a n a	a n	ʁ a	a ʁ a	a ʁ
ɲ a	a ɲ a	a ɲ			a ŋ

TABLE 1. Liste des mots et pseudo-mots (API) utilisés pour l'expérience.

Ces pseudo-mots ont été insérés dans la phrase porteuse : « Tu prononces … trois fois ». Le corpus est constitué de 3 répétitions des 52 pseudo-mots insérés dans la phrase porteuse et mises en ordre aléatoire pour le locuteur. Les 156 phrases ont été lues à voix haute par un sujet masculin, 40 ans, de langue maternelle française, origine des Deux-Sèvres, habitant depuis 20 ans à Grenoble. L'enregistrement s'est déroulé dans la chambre sourde du Département Parole et Cognition du GIPSA-lab, Grenoble, avec un enregistreur Marantz PMD 670, micro AKG C1000S à directivité cardioïde. Le corpus a été numérisé à 44.1 kHz sous format .wav. Le locuteur a été invité à lire le corpus à un débit normal. Les 156 pseudo-mots ont ensuite été segmentés manuellement sous Praat et extraits sous forme de liste de sons. Le matériel expérimental est donc de 3 répétitions de 17 pseudo-mots de structure CV, 17 de structure VCV et 18 de type syllabique VC.

2.2 Participants

26 sujets vietnamiens (5 hommes et 21 femmes), tous locuteurs natifs du dialecte du Nord, âgés de 18 à 20 ans ont participé au test. Les participants sont des étudiants de deux niveaux (intermédiaire et avancé) du Département de français de l'Université Nationale de Hanoi. L'équilibre des sujets quant au paramètre sexe n'a pas pu être obtenu car la majorité des inscrits dans ces formations sont de sexe féminin. Leur entourage parle le même dialecte. Les participants vivent tous pour l'essentiel de leur temps dans les régions du Nord. Ils ont tous l'anglais comme deuxième langue étrangère.

2.3 Déroulement du test

Avant de passer le test, chaque étudiant devait remplir un questionnaire sur leur identité et leur expérience en français. L'enregistrement pour chaque participant s'est déroulé dans un studio

calme de l'Université Nationale de Hanoi. L'écoute se faisait au casque binaural Head Set HS-95. Les 156 stimuli ont été présentés dans un ordre aléatoire unique pour tous les participants qui étaient informés au préalable qu'un signal de parole, pas nécessairement un mot, allait être présenté dans le casque et qu'ils devaient répéter, sans attendre, ce qu'ils avaient entendu de manière claire et à un débit normal. Un entraînement avec cinq pseudo-mots différents de ceux du test leur était préalablement proposé et la séance de test durait environ 10 minutes pour chaque participant, avec une pause à mi-parcours. L'enregistrement a été numérisé à 16 kHz sous format .wav.

3 Résultats

Les 156 stimuli sonores produits par chacun des 26 apprenants ont été segmentés manuellement et analysés acoustiquement avec Praat©. Leurs répétitions, comparées avec les modèles acoustiques en entrée, ont été codées en 1, si la consonne était reproduite correctement ou en 0, dans le cas contraire. Ce codage permet de calculer le score moyen de réussite par sujet et par item. À la fois analysée à l'aide du signal acoustique et écoutée par un expert, la répétition est jugée incorrecte s'il y a confusion avec un autre phonème (en mode et/ou en lieu d'articulation) (voir plus de détails section 3.4.3). Les données ainsi obtenues pour tous les items et pour tous les participants ont été analysées statistiquement avec le logiciel SPSS©. Nous présentons ci-dessous les résultats pour l'ensemble des consonnes du français avant d'expliquer plus spécifiquement les résultats concernant les 6 consonnes /p t k m n ŋ/ existant également en coda en vietnamien.

3.1 Effet de la distribution

Une ANOVA à mesures répétées a été effectuée pour comparer les moyennes de scores de bonnes répétitions correctes obtenues pour chacune des positions des consonnes dans le pseudo-mot (initiale, inter-vocalique, finale). Le test met en évidence un effet de position sur la production des consonnes [F(2,48) = 21,218 ; p = 0]. Les résultats montrent en effet que les sujets produisent mieux les consonnes en position inter-vocalique (81,8 %) et initiale (76,2 %) qu'en position finale (67,6 %), quel que soit leur niveau de français. La différence est significative quand on compare les contrastes en intra-sujet entre position initiale et finale [F(1,24) = 14,199 ; p = 0,001], entre la position inter-vocalique et finale [F(1,24) = 44,931 ; p = 0], entre position initiale et inter-vocalique [F(1,24) = 6,534 ; p = 0,017]. Aucune interaction entre la position et le niveau des sujets n'est détectée [F(2,48) = 0,349 ; p = 0,705].

3.2 Effet du mode d'articulation

Si on regroupe les consonnes selon leur mode d'articulation, les résultats statistiques montrent un effet significatif global sur le taux de réussite selon le mode [F(3,61) = 45,09 ; p = 0]. Les étudiants avancés ne sont pas plus performants que les intermédiaires (p = 0,614).

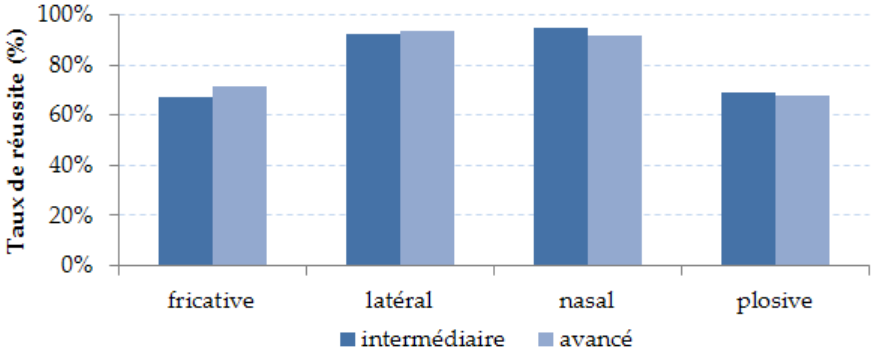


FIGURE 1. Taux de bonnes productions des consonnes du français selon leur mode d'articulation et le niveau des apprenants.

Quel que soit leur niveau, de manière générale, les étudiants produisent mieux les nasales (93,6 %) et les latérales (93,2 %) que les plosives (68,5 %) et les fricatives (69,2 %) (cf. figure 1). La différence entre moyennes est significative entre les productions fricatives et latérales (p = 0), fricatives et nasales (p = 0), plosives et nasales (p = 0), plosives et latérales (p = 0) ; mais ce n'est pas le cas entre productions plosives et fricatives (p = 0,656), nasales et latérales (p = 0,976).

Les consonnes /ʁ ɡ ʒ/ sont les moins bien reproduites par tous les participants, quel que soit leur niveau, alors que les trois nasales /m n ɲ/ sont les mieux réussies, avec des scores proches de 100 % par l'ensemble des apprenants des deux niveaux. Le score moyen obtenu pour /ŋ/ n'est pas très bon (57,6 % pour les apprenants de niveau intermédiaire et 48,5 % pour ceux de niveau avancé), peut-être dû à ce qu'il n'apparait qu'en position finale, la moins performante.

3.3 Effet du lieu d'articulation

Un effet significatif du score de bonne production a été également trouvé selon le lieu des consonnes [F(3,68) = 51,57 ; p = 0]. Aucune interaction entre le lieu et le niveau des apprenants en français n'a été détectée (p = 0,303). La figure 2 montrent que, quel que soit leur niveau, les étudiants réussissent mieux à reproduire la palatale /ɲ/ (98,3 %), devant les labio-dentales /f v/ (90 %) et les labiales /p b m/ (82,4 %). Ils rencontrent plus de difficultés à reproduire les vélaires /k ɡ ŋ/ (51,6 %) et l'uvulaire /ʁ/ (40,6 %).

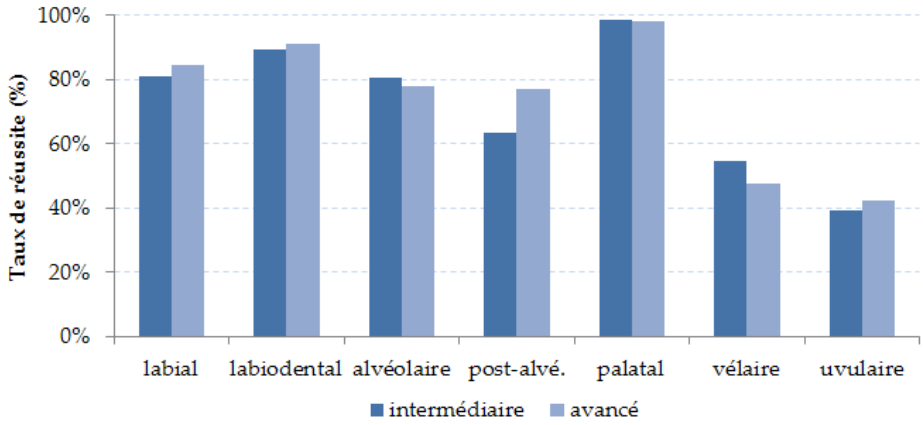


FIGURE 2. Taux de bonnes productions des consonnes selon leur lieu d'articulation et le niveau des étudiants.

3.4 Effet de similarités entre les deux langues : le cas des consonnes /p t k m n ŋ/

Le cadre global de cette étude concerne la production des consonnes faisant partie des séquences communes aux deux langues. Le vietnamien possède des séquences de consonnes uniquement aux frontières syllabiques de mot ou à l'intérieur de mots composés. Ces séquences sont donc obligatoirement constituées d'une des six plosives /p t k m n ŋ/, qui peuvent apparaître en coda.

Nous remarquons que globalement, les taux de production correcte de ces six consonnes sont au-dessus de 50 %, voire jusqu'à 100 % pour /p/, /t/ dans le groupe de niveau avancé ou /m/, /n/ pour les deux niveaux. Le niveau d'apprentissage du français des apprenants n'influence pas le taux de répétition correcte des plosives /p t k/ [F(1,24) = 0,035 ; p = 0,852], ni des nasales /m n/ [F(1,24) = 1,17 ; p = 0,29] dans toutes les positions, ni des nasales /m n ŋ/ en position finale [F(1,24) = 0,303 ; p = 0,587]. Aucune interaction entre le niveau des apprenants et leur performance à restituer des consonnes entendues n'est trouvée, quel que soit le facteur pris en considération (position, mode ou lieu).

3.4.1 Effet de la distribution

Les ANOVA à mesures répétées pour les plosives /p t k/ selon la position montrent le même effet déjà observé pour les dix-huit consonnes testées [F(2,48) = 7,49 ; p = 0,001]. Les sujets, quel que soit leur niveau, produisent mieux, de manière significative, les plosives /p t k/ en initiale (82 % en moyenne) qu'en finale (72 %) (p = 0,031) ; mieux en position inter-vocalique (87 %) qu'en finale (72 %) (p = 0,001). Il n'y a pas de différence significative entre les consonnes en position initiale et inter-vocalique (p = 0,148). L'effet de position n'est pas observé pour les nasales /m n/ : les étudiants des deux niveaux les réussissent aussi bien en initiale qu'en inter-vocalique ou en finale (97 %, 98 % et 98 % respectivement). Aucune différence significative de score d'identification n'est donc observée en fonction de leur position [F(2,48) = 0,376 ; p = 0,59].

3.4.2 Effet du lieu d'articulation

Les bilabiales et alvéolaires sont mieux produites (92 % et 93 % respectivement) que les vélaires (65 %) toutes positions confondues. En effet, /k/ est reproduit correctement seulement dans 57 % des cas, par rapport à /p/ (78 %) et /t/ (80,8 %). La différence est significative dans les deux cas (p = 0,015 et p = 0,005 respectivement). De même, la production de /ŋ/ est beaucoup moins réussie (53,8 %) par rapport à /m/ (100 %) et /n/ (96 %) (p = 0 pour les deux cas).

3.4.3 Analyse des erreurs

Une analyse quantitative des erreurs n'étant pas possible du fait du faible nombre d'erreurs pour certaines consonnes, une analyse qualitative des erreurs est présentée table 2.

Cons.	Réussi	Mode	Lieu	Mode + Lieu	Épen-thèse	Non relâch.	Nasalité	Aspira-tion	Total
p	197		17	3	5	3		8	233
t	209	11	5		3	2		4	234
k	161		10	53	3	3		4	234
m	233								233
n	224	10							234
ŋ	42	2	16		16		2		78
Total	1066	23	48	56	27	8	2	16	1246

TABLE 2. Traits d'erreurs et leurs occurrences dans la production des consonnes /p t k m n ŋ/, niveaux d'apprentissage confondus. À gauche, dans la colonne « Réussi », le nombre de réalisations correctes.

Les erreurs concernent principalement le mode et/ou le lieu d'articulation. Par exemple, les apprenants (avancés comme intermédiaires) ont réalisé le plus souvent /k/ comme une affriquée [tɕ], la variante du phonème /ʈ/ dans le parler du Nord (63 % des erreurs observées dans la production de /k/) ou comme [t] (12 %). /t/ est aussi produite le plus souvent comme [tɕ] (dans 46 % des réalisations de /t/) ou comme la rétroflexe [ʈ] (burst plus long et plus intense, F3 moins élevé : 23 % des erreurs de [t]). Alors que [p] en initiale comme en finale est réalisée comme [t] dans 47 % des cas, la nasale /m/ est produite correctement dans toutes les positions, par tous les étudiants des deux niveaux. La confusion entre /n/ et une autre consonne est relativement peu fréquente (seulement 10 cas sur 234 avec la latérale de même lieu d'articulation /l/). /ŋ/ en finale est réalisé plus souvent comme la palatale [ɲ] (33 % des cas). Sinon, dans 44 % des réalisations incorrectes de [ŋ], les étudiants produisent un bruit d'explosion vélaire très audible de type [k] ou [g] à la fin de ce segment répété [ŋk], [ŋg]. À signaler quelques cas dans lesquels les formes répétées des consonnes sont accompagnées d'une forte aspiration, davantage en inter-vocalique

que dans les autres positions. Quelques cas de non relâchement ont aussi été observés mais uniquement pour les plosives en coda. L'omission (6 % des cas) ne concerne que les plosives en coda.

4 Discussion

Cette expérience a mis en évidence des effets de la position dans la syllabe sur les performances de production des consonnes simples du français par des apprenants vietnamiens. Des analyses globales ont été menées prenant en compte toutes les consonnes avant d'approfondir les analyses des plosives orales et nasales /p t k m n ŋ/ qui entrent dans la composition des séquences de consonnes en vietnamien.

Les analyses sur l'ensemble des consonnes montrent que les étudiants ont de grandes difficultés à restituer les consonnes en position finale par rapport aux positions initiale ou inter-vocalique. Ce résultat, en conformité avec les résultats d'études antérieures (Major & Faudree, 1996 ; Anderson, 1987 ; Major, 1996 ; Sato, 1984 ; Weinberger, 1987 entre autres), valide l'Hypothèse de la Différence de Marquage (Eckman, 1977) et montre que l'acquisition des consonnes finales d'une langue seconde reste relativement problématique par rapport à d'autres positions. La même observation a été faite dans des études développementales qui ont révélé que chez les enfants, les consonnes en coda sont acquises plus tardivement que les consonnes en attaque (Stoel-Gammon, 1985 ; Demuth & McCullough, 2009). Les analyses par mode et lieu d'articulation montrent que les nasales et les latérales sont mieux récupérées et produites que les obstruentes par les apprenants des deux niveaux. Cependant, ces apprenants, quel que soit leur niveau, rencontrent plus de difficultés pour les uvulaires et les vélaires (/ʁ g ʒ/ sont les moins bien produites). Ces consonnes, n'existant pas en vietnamien, elles sont la plupart du temps réalisées sans voisement et produites comme leur homologue sourd (respectivement /χ/, /k/, /ʃ/), soit comme un phonème proche faisant partie du répertoire phonologique de la langue source (/ʁ/ > [x], /g/ > [ɣ], /ʒ/ > [z]). Ainsi, la majorité des erreurs dans les productions des étudiants, quel que soit leur niveau, trouvent leurs sources dans le système phonologique de la langue de départ. Il semblerait que ces nouveaux sons sont filtrés par le crible phonologique (Troubetzkoy, 1939), lequel perturbe la répétition correcte chez les apprenants, même après plusieurs années de pratique. Il se pourrait que ces nouveaux phonèmes du français ne soient pas « suffisamment » différents des sons de la langue première pour être assimilés en tant que nouvelle catégorie phonologique. Ce phénomène peut être décrit comme le mécanisme de *classification par équivalence* par lequel les catégories phonétiques de la L1 sont susceptibles de bloquer la formation de nouvelles catégories pour les sons proches de la L2, selon le *Speech Learning Model* de Flege (1995). Ces nouveaux sons ont été le plus souvent assimilés à des catégories natives et reproduits comme un bon exemplaire de ces catégories, figurant comme une des possibilités d'assimilation de segments non natifs décrites dans le *Perceptual Assimilation Model* de Best (1995).

L'analyse des données de production des six consonnes /p t k m n ŋ/ qui correspondent aux codas possibles en vietnamien et qui constituent toujours le premier segment des séquences de consonnes dans cette langue, montre un effet de la position pour les orales /p t k/. Les meilleures performances sont toujours observées en position initiale et inter-vocalique, alors qu'en finale, les plosives sont moins bien répétées. Par contre, le taux de répétition correcte est identique pour les nasales /m n/ quelle que soit la position. /ŋ/ apparaissant en coda dans les pseudo-mots du corpus présente des scores bien plus faibles par rapport à /m/ et /n/. La raison pour laquelle, dans les répétitions des apprenants, la vélaire /ŋ/ est plus souvent remplacée par la palatale /ɲ/ ou accompagnée d'un burst final, semble résider dans l'input (cf. figure 3). En effet, les transitions

convergentes de F2 et F3 dans les échantillons (2) et (3) présentés figure 3 indiquent que le segment post-vocalique est de type vélaire, alors que les trajectoires montantes de F2 et F3 dans le premier échantillon (1) montrent que le lieu d'articulation du segment qui suit la voyelle est plus antérieur que vélaire, ce qui pourrait expliquer le fait que les apprenants ont produit souvent [aŋ] à l'écoute du stimulus (1). En outre, le sonagramme de l'échantillon (2) montre que le locuteur a prononcé une double consonne post-vocalique [ŋg], expliquant un taux d'épenthèse consonantique en final à la répétition de ce stimuli (2) par les apprenants des deux niveaux.

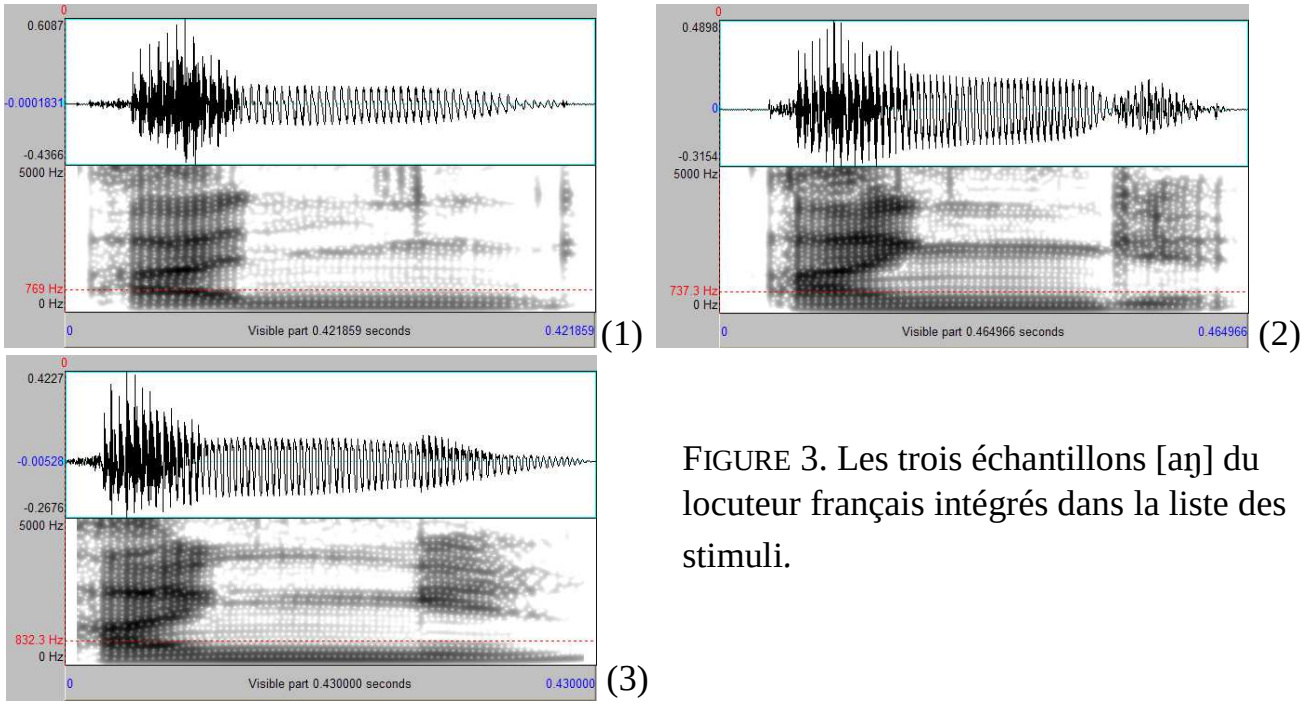


FIGURE 3. Les trois échantillons [aŋ] du locuteur français intégrés dans la liste des stimuli.

Un effet du lieu d'articulation sur le taux de bonne répétition a également été observé : les bilabiales et alvéolaires présentent les meilleurs scores (92 % et 92 %) par rapport aux vélaires (61 %) et ce, quel que soit le niveau d'apprentissage.

L'analyse qualitative des erreurs en fonction des positions des consonnes montre que celles-ci sont souvent liées à une confusion du mode et/ou du lieu d'articulation (plosif [t], [k] > affriquée [tɕ] ; nasal [n] > latéral [l]). Certains apprenants venant des provinces Hải Phòng, Thái Bình, confondent [n] avec [l] dans les répétitions. Cette confusion [n] et [l] est en effet très fréquemment rencontrée dans les parlers de ces provinces (ex. « nón » [nɔn^5] *chapeau* > [lɔn^5] *pas de signification* ; « lá » [la^5] *feuille* > [na^5] *arbalète*). Cette confusion entraîne des homophonies, la compréhension doit se faire en contexte. Parmi les qualités de production, les apprenants utilisent un trait phonétique caractéristique des codas de la langue première : le non relâchement qui constitue, selon les tenants de l'Hypothèse de l'Analyse Contrastive (Lado, 1957), un transfert négatif de la langue maternelle.

Cette première expérience met en évidence l'importance de la position (initiale, inter-vocalique, finale) dans la récupération et la production des consonnes simples en français par des apprenants vietnamiens dans une tâche de répétition. Dans le prolongement de cette première expérience, d'autres tests ont été effectués sur les séquences de consonnes (pseudo-mots et mots) communes aux deux langues.

Remerciements

Cette recherche a bénéficié d'un financement de l'Agence Universitaire de la Francophonie (PC – 411/2460). Un grand merci à Lionel Granjon et Christophe Savariaux pour l'aide matérielle qu'ils ont apporté à ce projet ainsi que pour les échanges constructifs et discussions très fructueuses.

Références

ANDERSON, J. (1987). The markedness differential hypothesis and syllable structure difficulty. In *Interlanguage Phonology: The Acquisition of a Second Language Sound System*, 279-291 (Eds G. Ioup and S. Weinberger). New York: Newbury House/Harper and Row.

BEST, C. T. (1995). A direct realist view of cross-language speech perception. In *Speech perception and linguistic experience: Issues in cross-language research*, 171-204 (Ed W. Strange). Timonium: MD: York Press.

CARLISLE, R. (1998). The acquisition of onsets in a markedness relationship. A longitudinal study. *Studies in Second Language Acquisition* 20: 245-260.

DEMUTH, K. & MCCULLOUGH, E. (2009). The longitudinal development of clusters in French. *Journal of Child Language* 36: 425-448.

ĐOÀN T. T. (1999). *Ngữ âm tiếng Việt (Tr. La phonétique du vietnamien).* Hanoi: Nhà xuất bản Đại học Quốc gia Hà Nội (Maison d'édition de l'Université Nationale de Hanoi).

ECKMAN, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning* 27: 315-330.

FLEGE, J. E. (1995). Second Language Speech Learning. Theory, Findings and Problems. In *Speech Perception and Linguistic Experience: Issues in Cross-language research.*, 233-277 (Ed W. Strange). Timonium: MD : Yord Press.

LADEFOGED, P. (2001). *Vowels and consonants: an introduction to the sounds of languages.* Oxford: Blackwell.

LADEFOGED, P. & MADDIESON, I. (1996). *The Sounds of World's Languages.* Oxford: Blackwell Publishers Ltd.

LADO, R. (1957). *Linguistics across cultures: Applied lunguistics for language teachers.* University of Michigan: Press: Ann Arbor.

MAJOR, R. C. & FAUDREE, M. C. (1996). Markedness Universals and the Acquisition of Voicing Contrasts by Korean Speakers of English. *Studies in Second Language Acquisition* 18(1): 69-90.

NGUYỄN T. B. M. (2000). Regards sur l'enseignement de la phonétique dans la formation des étudiants en F.L.E. à l'Université Pédagogique de Ho Chi Minh ville. Université de Rouen.

NGUYỄN T. N. H. (2001). Interférences phonologiques de la langue maternelle dans l'apprentissage du français, Etude du cas des étudiants de l'école des sciences de Hue et proposition pédagogiques. Université de Rouen.

POLIVANOV, E. (1931). La perception des sons d'une langue étrangère. *Travaux du Cercle linguistique de Prague* 4.

ROUSSET, I. (2004). Structures syllabiques et lexicales des langues du monde. Données, typologiques, tendances universelles et contraintes substantielles. Thèse de Doctorat en Sciences du Langage, Grenoble: Université Stendhal.

SATO, C. (1984). Phonological processes in second language acquisition: Another look at interlanguage syllable structure. *Language Learning* 34: 43-57.

STOEL-GAMMON, C. (1985). Phonetic inventories, 15-24 months: a longitudinal study. *Journal of Speech and Hearing Research* 28: 505-512.

TROUBETZKOY, N. S. (1939/2005). *Principes de phonologie (Grundzüge der Phonologie).* Klincksieck.

WEINBERGER, S. H. (1987). The influence of linguistic context on syllable structure simplification. In *Interlanguage phonology: The acquisition of a second language sound system*, 401–417 (Eds G. Ioup and S. H. Weinberger). Rowley, MA: Newbury House.

Analyse du corpus MATRICE-INA *: exploration et classification automatique d'archives audiovisuelles de 1930 à 2012

Antoine Laurent(2) Camille Guinaudeau(1,2) Anindya Roy(2)
(1) Université Paris-Sud, Rue du Château, 91400 Orsay
(2) LIMSI-CNRS, Rue John Von Neumann, 91400 Orsay
laurent@limsi.fr, guinaudeau@limsi.fr, roy@limsi.fr

RÉSUMÉ
Cet article décrit les méthodes mises en place pour permettre l'analyse d'un corpus composé de documents audiovisuels diffusés au cours des 80 dernières années : le corpus MATRICE-INA. Nous proposons une exploration des données permettant de mettre en évidence les différents thèmes et évènements abordés dans le corpus. Cette exploration consiste dans un premier temps à effectuer une analyse temporelle sur les notices documentaires produites manuellement par les documentalistes de l'Institut National de l'Audiovisuel et sur les transcriptions automatiques des documents. Puis, nous montrons, grâce à une technique de clustering automatique, que les transcriptions automatiques permettent également d'effectuer une analyse du corpus faisant émerger des thèmes cohérents avec les données traitées.

ABSTRACT
Analysis of the MATRICE corpus : exploration and automatic clustering of audiovisual archives from 1930 to 2012

This paper presents different approaches developed for the analysis of an audiovisual collection of documents broadcasted during the last 80 years, namely the MATRICE-INA corpus. We propose an exploration of the documents to highlight the different topics and events discussed in the corpus. This exploration consists first in a temporal analysis carried out on documentary records manually produced by the archivists of the French National Audiovisual Institute and on automatic transcripts of the documents. Then, we show, through an automatic clustering technique, that the automatic transcripts also allow an accurate topics detection that seems to be coherent.

MOTS-CLÉS : Fouille de données, clustering, transcription automatique de la parole, document audiovisuel.

KEYWORDS: Data mining, clustering, automatic speech recognition, audiovisual document.

1 Introduction

L'Institut National de l'Audiovisuel (INA), créé en 1974, est responsable du dépôt légal de la radio et de la télévision en France depuis la loi du 20 juin 1992. Ce travail donne lieu à une quantité de données très importante qui rend possible la mise en place de différentes études permettant

∗. Ce projet bénéficie d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'investissements d'avenir EQUIPEX MATRICE.

notamment d'analyser la façon dont ont été traités des évènements historiques importants mais également les évolutions possibles des discours associés à ces évènements. Dans ce cadre, le projet transdiciplinaire MATRICE, regroupant 24 partenaires, a été mis en place dans l'objectif d'une part, de repérer et d'analyser les grands récits nationaux sur deux grands événements (la seconde guerre mondiale et les attentats du *World Trade Center* à New York le 11 septembre 2001) et d'autre part d'étudier les comportements des publics de mémoriaux.

Afin d'analyser le corpus MATRICE-INA, l'une des possibilités consiste à étudier les notices documentaires associées à chacun des documents par les documentalistes de l'INA. Cependant ces notices ne sont, d'une part, pas toujours disponibles [1] et ne contiennent, d'autre part, pas toujours d'informations sémantiques exploitables. Certaines d'entre elles ne contiennent, en effet, que des informations sur la structure de montage du document, par exemple « Plan large sur le général de Gaulle » ou « Panoramique sur le maréchal Pétain », *etc.* Pour pallier ces deux difficultés, il semble naturel de chercher à employer les transcriptions automatiques de la parole. Cette solution, cependant, reste problématique car le modèle de langue et le modèle acoustique du système de reconnaissance automatique de la parole employé ne sont pas toujours adaptés aux données à transcrire qui s'étalent sur 80 ans (évolution de la qualité des enregistrements, évolution linguistique en terme de discours et de diction, *etc.*). L'objectif de cet article consiste donc à présenter les premières expérimentations effectuées afin d'estimer la faisabilité d'une analyse basée uniquement sur des transcriptions automatiques de la parole prononcée dans des archives portant sur les huit dernières décennies.

Dans cet article, nous présentons tout d'abord le corpus MATRICE-INA ainsi que l'approche employée lors de la transcription automatique des documents qui le composent. Dans un second temps, nous proposons une analyse temporelle du corpus permettant la mise en évidence d'évènements marquants dans les documents étudiés. Finalement, nous montrons dans une troisième partie qu'il est possible de fournir une analyse pertinente des archives du projet en se basant uniquement sur les transcriptions automatiques et en utilisant des techniques de clustering automatique. Nous montrons notamment que les documents audiovisuels fournis dans le cadre du projet abordent des thèmes plus variés que ceux originellement prévus.

2 Le corpus MATRICE-INA

Le corpus, qui a terme doit contenir environ 100 000 documents, est composé pour l'instant de 30 000 documents audiovisuels. L'écoute d'une partie des documents (sélectionnée de façon aléatoire), nous a permis de constater que certains ne contenaient pas de parole francophone. Nous avons donc utilisé un système de détection automatique des langues sur chaque document et conservé ceux détectés comme ne contenant que du français, soit 23 000 documents. Ces 23 000 documents ont pour thème la seconde guerre mondiale et les attentats du 11 septembre 2001. Par ailleurs, tous les documents audiovisuels du corpus ne contiennent pas de la parole (présence de films muets). Le tableau 1 fournit des détails sur le corpus et montre notamment que la majorité des données étudiées dans le cadre du projet a été diffusée avant 1995. Pour chacun de ces 23 000 documents, il nous a été fournit à la fois des notices documentaires manuelles – composées d'une partie description et d'une liste de mots-clés appelés *descripteurs* – et l'audio correspondant aux canaux gauches et droits. En effet, à l'époque où les vidéos les plus anciennes ont été enregistrées, elles ne disposaient pas de timecode permettant de naviguer à l'intérieur

1. Le corpus actuel ne contient pour l'instant aucune notice traitant des attentats du 11 septembre 2001.

TABLE 1: Détails sur le corpus traité

	Durée	Durée de parole	Nombre de mots	Nombre de mots distincts
Avant 1995	10840h	8274h	96 millions	168k
Après 1995	2814h	2396h	29 millions	124k
Total	13654h	10670h	124 millions	179k

de l'enregistrement. Pour pallier cela, l'un des canaux servait à enregistrer l'horloge parlante, ce qui a nécéssité la mise en place de techniques adaptées pour transcrire efficacement le corpus MATRICE-INA. Dans cette section nous présentons donc, dans un premier temps, le système de reconnaissance automatique de la parole employé dans le cadre du projet puis nous décrivons les particularités de nos données ainsi que les stratégies mises en place pour gérer les spécificités de notre corpus.

2.1 Transcription automatique de la parole

Le travail du LIMSI dans ce projet consiste à adapter son Système de Reconnaissance Automatique de la Parole (SRAP) pour améliorer les transcriptions de ces 80 années d'enregistrements, réalisés dans des conditions très hétérogènes en terme de vocabulaire et d'acoustique.

Dans cette étude préliminaire, nous avons utilisé tel quel le SRAP du LIMSI (développé conjointement par le LIMSI et la société Vocapia [2]) pour obtenir un premier jeu de transcriptions. Ne disposant pas encore d'un corpus de développement exploitable, il nous est en effet impossible de procéder à l'adaptation des différents modèles.

Le système est formé d'un partitioneur parole/non-parole et d'un décodeur de mots. Le décodeur de mots utilise un modèle acoustique HMM et un modèle de langage bi-gramme pour construire un treillis de mots qui est ensuite réévalué à l'aide d'un modèle de langue 4-gram. L'apprentissage des modèles acoustiques est réalisé en alignant une transcription orthographique exacte sur le signal de parole au moyen d'un jeu de modèles phonétiques et d'un lexique de prononciation. Ces modèles sont entraînés à partir de paramètres MLP (Multi-Layer Perceptron) concaténés avec des paramètres cepstraux standards (Perceptual Linear Prediction – PLP) et la fréquence fondamentale (F0). Ils ont été estimés à l'aide d'un apprentissage discriminant (MMIE) avec un apprentissage adaptatif aux locuteurs (SAT). Une adaptation par blocs diagonaux a été utilisée pour les paramètres MLP et PLP+F0. Le vecteurs de paramètres cepstraux est composé de 81 paramètres ($39MLP + 39PLP + F0 + \Delta F0 + \Delta\Delta F0$). Les modèles acoustiques sont des HMMs (Hidden Markov Models) composés d'un mélange de gaussiennes. Ils sont à états liés, dépendants du contexte et du sexe du locuteur. Le silence est modélisé par un seul état composé de 1024 gaussiennes. Le corpus d'apprentissage est composé de 866 heures de paroles, à partir desquelles 19 679 modèles contextuels ont été entraînés, partageant 11 517 états, pour un total de 370k gaussiennes. Les modèles de langue génériques ont été estimés sur un corpus d'environ 2.9 milliards de mots. Les modèles sont construits autour d'un vocabulaire de 200K mots. Les prononciations du dictionnaire sont obtenues par une phonétisation automatique à base de règles.

2.2 Quantification des données exploitables

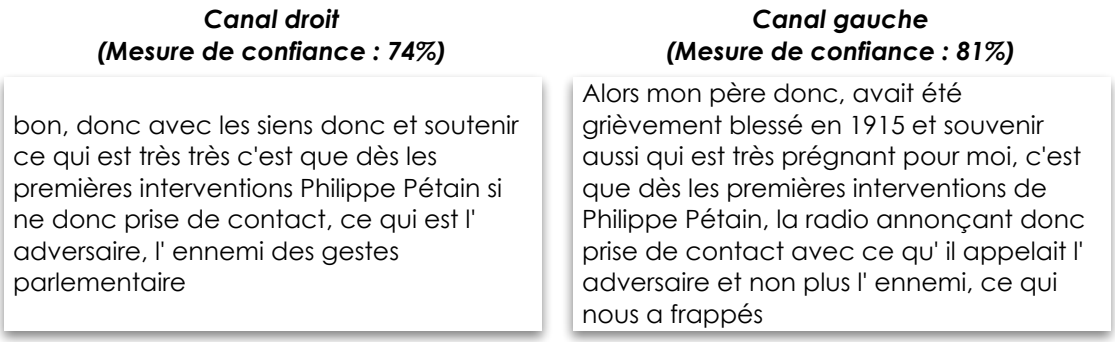
À partir de la sortie du système de reconnaissance automatique de la parole, nous avons cherché à éliminer le canal contenant l'enregistrement de l'horloge parlante. Pour ce faire, une recherche de

2. http ://www.vocapia.com/

TABLE 2: Nombre de documents par canal (G=Gauche, D=Droit)

	G et D identiques	Canal G et D différents		G ou D vide	G et D vide	G et D horloge
		Horloge parlante	Autres			
Nombre	19871	3853	2479	366	2638	6

FIGURE 1: Différence de transcription entre canal droit et canal gauche



différents mots clés dans la transcription automatique (par exemple « au top il sera exactement ») et une analyse des durées des segments de parole a été réalisée (les segments des fichiers contenant l'horloge parlante durent entre 3 et 4 secondes).

Lorsque l'audio est enregistré sur les deux canaux, il existe pour certains documents une importante différence de qualité d'enregistrement entre le canal droit et le canal gauche. Dans ce cas nous avons considéré le canal permettant de maximiser la mesure de confiance du système de reconnaissance automatique de la parole. Pour chacun des documents un taux d'erreur mot a été calculé entre le canal droit et le canal gauche pour déterminer s'ils étaient identiques ou non. La figure 1 présente un exemple où la transcription sur les deux canaux est différente. Il s'agit d'un extrait du décodage de la même zone audio sur les deux canaux. Il apparaît clairement que la transcription sur le canal gauche est plus proche de ce qui a réellement été prononcé. A l'écoute, le canal droit est vraiment très bruité. En tout, 237h d'audio (films muets, horloge parlante sur les deux canaux et documents pas en Français) ne sont pas utilisables. Le nombre de documents correspondant à chaque configuration est présenté dans le tableau 2.

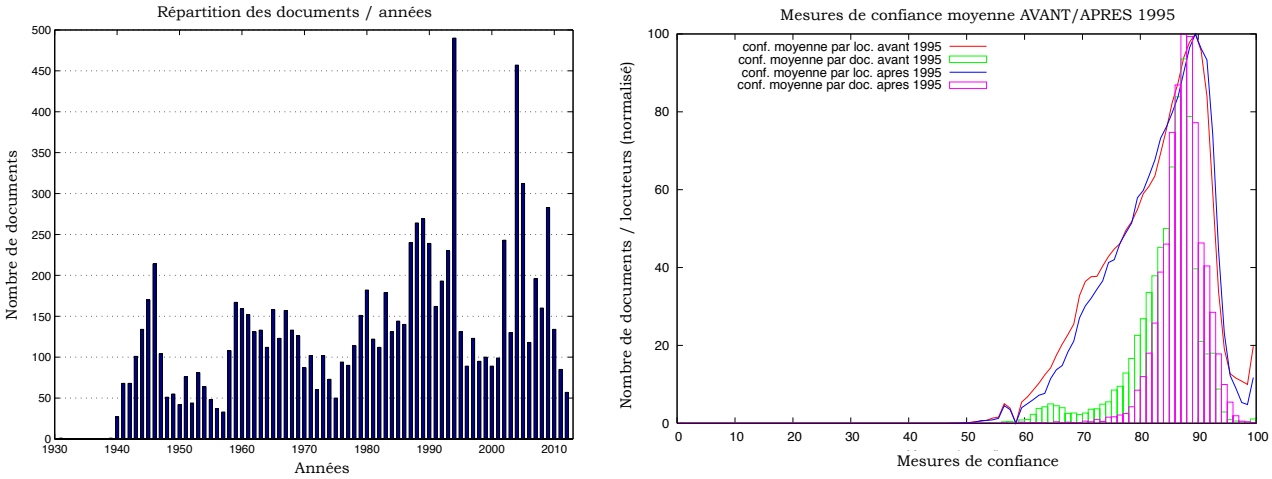
Un système de détection de la langue (LID) (BenZeghiba *et al.*, 2012) a été utilisé pour détecter les enregistrements contenant des données non francophones. En tout nous avons détecté 7 langues différentes en plus du Français.

La figure 2 présente la mesure de confiance moyenne des transcriptions par documents et par locuteur pour les enregistrements d'avant et après 1995[3]. On constate que la plupart des mesures de confiance se situent aux alentours de 85% pour les données d'avant et après 1995, le nombre de documents ayant des mesures de confiance faible étant cependant plus important pour les enregistrements d'avant 1995. Cette observation s'explique par le fait que les conditions acoustiques des enregistrements et le vocabulaire utilisé sont plus éloignés des données d'apprentissage du SRAP du LIMSI. Cette figure présente également le nombre de documents dont nous disposons pour chaque année.

Finalement, le lien entre les notices documentaires et les documents audiovisuels est souvent

3. La normalisation a été faite en ramenant à 100 le nombre maximum de documents/locuteurs rencontrés pour une classe de mesure de confiance.

FIGURE 2: Mesures de confiance moyenne avant/après 1995, répartition des documents par années



difficile à faire car les notices ne contiennent pas toujours le nom du fichier audiovisuel qu'elles décrivent. De plus, bien que nous ayons le même nombre de documents audio et de notices, aucune notice ne semble être associée aux documents abordant le thème du 11 septembre, alors même qu'il existe des documents abordant ce thème dans notre corpus. Pour ces raisons, nous avons choisi de nous concentrer, dans la suite de cet article, sur un sous-corpus composé uniquement de documents portant sur la seconde guerre mondiale et pour lesquels chaque transcription est associée de manière certaine à une notice (même titre de programme et d'émission et même date de diffusion)

Sur ce sous-corpus, composé de 12 592 documents, nous cherchons à évaluer dans quelle mesure les transcriptions automatiques imparfaites peuvent nous permettre de manipuler le corpus MATRICE-INA pour en faire ressortir des informations intéressantes. Pour cela, nous étudions dans un premier temps (Section 3) la qualité de l'analyse temporelle effectuée à partir des transcriptions automatiques, en comparant les mots apparaissant les plus fréquemment dans les documents traitant du général de Gaulle avec ceux obtenus à partir des notices documentaires. Puis nous estimons dans un second temps, la capacité de techniques de clustering appliquées sur les transcriptions automatiques à retrouver des thèmes extraits manuellement à partir des notices documentaires de qualité (*cf.* Section 4).

3 Analyse temporelle des documents

Dans cette partie, nous analysons les cooccurrences de termes les plus fréquentes au sein d'un thème particulier. Nous cherchons notamment à retrouver le vocabulaire couramment associé au Général de Gaulle afin d'évaluer si les transcriptions permettent de faire ressortir les mêmes évènements marquant que ceux observés grâce aux notices documentaires.

Pour mener à bien cette comparaison, nous sélectionnons tout d'abord dans notre corpus uniquement les couples notices-transcriptions pour lesquels les notices contiennent le descripteur *de Gaulle* et nous créons pour chaque type de données cinq sous-corpus correspondant aux cinq décennies entre la fin des années 1950 et les années 2000. Puis, nous traitons de la même façon le contenu des notices (c'est-à-dire la description du document, privée des descripteurs) et les transcriptions. Le contenu des notices et les transcriptions sont lemmatisés grâce à l'outil *TreeTagger*(Schmid, 1994) et une liste de *stop words* est définie afin de supprimer les termes

TABLE 3: Cooccurrences par décénnie

Notices					Transcriptions				
60s	70s	80s	90s	00s	60s	70s	80s	90s	00s
visite	intérieur	1er	ensemble	reportage	traverser	forces	épisode	madame	recherche
foule	vie	estaing	image	presse	auteur	cercueil	basque	écrire	milan
algérie	pompidou	différent	porter	mairie	nord	découvrir	bâton	fédéral	dénicher
arrivé	dernier	plan	président	églises	parlement	exister	formidable	église	oiseau
madame	chaban	libération	rue	csa	ouverture	préfet	larme	reprocher	malade
chrono	paris	foule	jour	geneviève	carrière	main	pension	recours	gueule
enfant	jean	appel	colombey	déplorer	ligne	taille	allemagne	anticyclone	76e
ville	président	image	rpr	académicien	algériens	valable	lecture	explosion	patrimonial
homme	cimetière	archives	jacques	consécration	discuter	littérature	ramener	cellule	hastings
gaulle	colombey	anniversaire	émission	tenir	puissant	génie	grève	défenseur	laure

vide de sens qui peuvent apparaître dans notre collection de documents. Certains de ces *stop words* sont très spécifiques à nos données, c'est le cas par exemple des termes « pano », « gp », « ps » (abbréviations respectives de « panoramique », « gros plan » et « plan serré »), présents dans les notices et employés pour décrire la structure de montage du document. Chaque mot des transcriptions et des notices est ensuite associé à son score *tf-idf*[4] et les 10 termes associés aux valeurs *tf-idf* les plus élevées sont conservés.

Le tableau 3 liste pour chaque décennie les termes les plus représentatifs des documents traitant du général de Gaulle. On peut voir dans ce tableau que les mots clés retrouvés pour les notices et pour les transcriptions sont à peu près équivalents pour les années 1960 et 1970. Pour les années 1960, le thème de la guerre d'Algérie est présent à la fois dans les notices, avec le mot « algérie », et dans les transcriptions, avec le mot « algérien ». De la même manière, on voit apparaître le thème de la mort du général de Gaulle dans la décénnie 70 à la fois grâce aux notices (« cimetière », « colombey ») et grâce aux transcriptions (« cercueil »). Cependant, à partir des années 1980, on peut constater que les mots clés obtenus grâce aux transcriptions sont plus flous et il devient moins facile de faire le lien entre ces mots clés et le général de Gaulle alors que ceux extraits des notices gardent un lien avec la politique et le décès du général (« églises » dans les années 2000 fait ainsi référence à Colombey les deux églises, lieu de décès de Charles de Gaulle). Selon nous, cette observation peut s'expliquer par le fait que les documents audiovisuels fournis par l'INA ne sont peut être pas thématiquement homogènes et qu'ils abordent probablement d'autres thèmes que celui de la seconde guerre mondiale. Afin de valider cette hypothèse, nous effectuons dans la section suivante un clustering automatique des transcriptions des documents afin de faire émerger les thèmes les plus importants de notre corpus.

4 Détection automatique de thèmes

Afin de faire émerger les thèmes de notre corpus, nous appliquons des méthodes de clustering automatique sur le corpus composé des 12 592 couples notice/document abordant tous les thèmes de la seconde guerre mondiale.

4. La notation *tf-idf* fait référence à la multiplication de la valeur de la fréquence d'apparition d'un terme dans un document, *tf*, par l'inverse de sa fréquence d'apparition dans l'ensemble des documents, *idf*. Ici, la valeur de l'*idf* est estimée sur chaque sous-corpus (transcriptions *vs* notices) séparemment.

Dans un premier temps, nous utilisons l'algorithme *Mean Shift* – proposé dans (Fukunaga et Hostetler, 1975) et qui ne nécessite pas de connaissances a priori sur le nombre de clusters à définir – afin de détecter automatiquement le nombre de thèmes présents dans notre corpus. Appliqué sur les notices documentaires, il nous permet de fixer à 6 le nombre de thèmes principaux présents dans nos données.

Dans un second temps, afin de caractériser automatiquement les différents thèmes présents dans les transcriptions du corpus, nous avons employé une factorisation de matrices non négatives NMF (Lee et Seung, 2001) qui consiste à factoriser une matrice V en deux sous-matrices W et H telles que les trois sous matrices soient composées uniquement d'éléments positifs (ou égaux à 0). Cette technique, couramment utilisée dans l'extraction automatique d'informations à partir de données textuelles mais également dans le contexte du regroupement de documents en thèmes (Shahnaz *et al.*, 2006), permet, grâce au principe de non-négativité, de manipuler des matrices plus faciles à analyser. En effet, dans le cadre de la NMF, le nombre de colonnes de la matrice W et le nombre de lignes de H sont choisis de façon à ce que le produit WH soit une approximation de V. L'obtention de matrices W et H de plus petite taille que V les rend plus facilement manipulables et analysables. Cela permet également de représenter les éléments de W par une quantité de données beaucoup plus faible.

Dans notre cas, la matrice V, de taille N (égal au nombre de documents) $\times$ D (correspondant au nombre de lemmes considérés dans notre corpus), contient les valeurs des scores *tf-idf* de chacun des lemmes dans les documents. La matrice W, de taille d – le nombre de thèmes, fixé à priori à 6 dans notre cas – $\times$ D, va quant à elle contenir des valeurs non négatives traduisant l'appartenance de chaque lemme à un thème. C'est cette matrice qui nous permet de fournir la caractérisation des thèmes présentée dans le tableau 4. Finalement, la matrice H, de taille $N \times d$ contient, pour chaque document, les thèmes activés.

Cette approche nous permet d'extraire automatiquement du corpus les 6 thèmes reportés dans le tableau 4. Chaque thème est caractérisé par ses 5 mots les plus représentatifs grâce à la matrice W. Dans la partie droite du tableau, nous voyons apparaitre les thèmes et les caractérisations extraits des notices. Ces thèmes sont majoritairement corrélés avec la thématique globale du corpus. Seul le thème **B** diffère du thème de la seconde guerre. Il est très spécifique au corpus que constituent les notices documentaires et correspond aux notices qui décrivent précisément la structure de montage des émissions. Concernant le clustering effectué sur les transcriptions automatiques des documents audiovisuels, 4 thèmes sur 6 semblent aborder précisèment la seconde guerre mondiale, à travers le débarquement (thème **2**), le général de Gaulle (thème **4**), le procès de Maurice Papon (thème **5**) et l'holocauste (thème **6**). Les caractérisations proposées pour les 2 autres thèmes sont cependant plus floues et semblent aborder des thèmes plus variés. Cette différence entre notices et transcriptions peut être liée, d'une part, au fait que les notices, plus synthétiques, ne contiennent que des élèments en rapport avec le thème de la seconde guerre mondiale. D'autre part, nous pensons, comme nous l'avons mentionné dans la section 3, que les documents audiovisuels transcrits ne sont pas thématiquement homogènes.

Afin de vérifier la présence d'autres thèmes dans les documents transcrits, nous avons effectué un second clustering grâce à la NMF en fixant le nombre de thèmes à 20 afin d'étudier les caractérisations obtenues et présentées dans le tableau 5. Grâce à ce tableau, nous voyons apparaitre des thèmes relatifs à la seconde guerre mondiale, tel que les thèmes numéro **5**, **6** ou **13**, mais également des thèmes moins corrélés avec la thématique globale du corpus. Les thèmes numéro **8**, **14** et **15**, par exemple, semblent respectivement liés au football, à la météo et à la

TABLE 4: Caractérisation des 6 thèmes détectés automatiquement dans les notices et les transcriptions automatiques correspondantes

Transcriptions	**Notices**
1 - jean pourcent paris pays foi	**A** - gaulle général foule discours pompidou
2 - débarquement juin plage soldat vétéran	**B** - plan différent ensemble interview blanc
3 - gens penser livre vie film	**C** - ficher chrono fichier détectage juif
4 - gaulle général politique président république	**D** - mitterrand françois chirac cérémonie gerbe
5 - papon maurice procès avocat prison	**E** - soldat allemand débarquement char avion
6 - juif camp nazi guerre barbie	**F** - jean barbie interview guerre film

TABLE 5: Caractérisation des 20 thèmes détectés automatiquement dans les transcriptions

1 - problème politique gouvernement pays question	**11** - barbie klaus procès moulin lyon
2 - livre penser gens film écrire vie	**12** - famille passer maison femme arriver
3 - gaulle général colombey juin charles	**13** - résistance guerre paris allemands général
4 - président république ministre mitterrand état	**14** - paris pourcent ouest température nuage
5 - papon maurice procès avocat bordeaux	**15** - église pape catholique jean paul vatican
6 - juif camp auschwitz déporté nazi	**16** - soviétique moscou communiste staline pays
7 - touvier paul procès crime milice	**17** - sarkozy nicolas chirac cérémonie jacques
8- match équipe championnat champion monde	**18** - allemand allemagne hitler europe berlin
9 - parti politique socialiste candidat gauche	**19** - palestinien israélien israël irakien irak
10 - algérie algérien alger musulman algériens	**20** - débarquement vétéran juin plage normandie

religion. Nous concluons de cette observation que certains documents abordent plusieurs thèmes potentiellement très différents, ce qui n'apparaissait pas dans le notices documentaires fournies par l'INA.

5 Conclusion et perspectives

Dans cet article, nous avons montré qu'il était possible de faire une analyse pertinente des archives du corpus MATRICE-INA pour en extraire les thèmes mais également les principaux évènements en se basant uniquement sur les transcriptions automatiques obtenues par un système non optimisé pour ces données. Nous avons notamment mis en évidence les thèmes principaux abordés dans les archives grâce à une factorisation de matrices non négatives NMF. Au cours de notre analyse, nous avons également constaté l'émergence de certaines thématiques très différentes du thème de la seconde guerre mondiale. Dans la suite du projet MATRICE, il nous semble donc nécessaire de nous attacher dans un premier temps à la mise en place d'étapes de pré-traitement de notre corpus et notamment une étape de segmentation thématique des transcriptions des documents audiovisuels afin de pouvoir nous concentrer uniquement sur les parties du corpus les plus pertinentes pour les problématiques soulevées dans le cadre du projet. Cette segmentation thématique des documents doit également nous permettre de mettre en place une adaptation des modèles de langue. Par ailleurs, nous souhaitons adapter de façon non supervisée les modèles acoustique pour les enregistrements les plus anciens (comme présenté dans (Lamel *et al.*, 2002)), et envisageons d'ajouter de nouveaux « fillers » pour modéliser des bruits très présents dans les actualités diffusées en début de films au cinéma dans les années 50, par exemple le bruit des bombes.

Références

BenZeghiba, M., Gauvain, J. et Lamel, L. (2012). Phonotactic language recognition using mlp features. *In Interspeech*, Portland, USA.

Fukunaga, K. et Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40.

Lamel, L., Gauvain, J.-L. et Adda, G. (2002). Lightly supervised and unsupervised acoustic model trainings. *Computer Speech and Language*, 16:115–129.

Lee, D. D. et Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems 13 : Proceedings of the 2000 Conference*, pages 556–562.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Shahnaz, F., Berry, M. W., Pauca, V. et Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373 – 386.

Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques

Camille Dutrey[1, 2, 3, 4] Sophie Rosset[2] Martine Adda-Decker[2, 3] Chloé Clavel[5]
Ioana Vasilescu[2]
(1) EDF R&D, 92 141 Clamart, France
(2) LIMSI-CNRS, 91 403 Orsay, France
(3) LPP, Université Paris III – Sorbonne Nouvelle, 75 005 Paris, France
(4) Université Paris Sud, 91 405 Orsay, France
(5) Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, 75 014 Paris, France
{dutrey,madda,rosset,ioana}@limsi.fr, chloe.clavel@telecom-paristech.fr

RÉSUMÉ

La parole spontanée est caractérisée par la présence de nombreux éléments disfluents qui peuvent s'avérer très informatifs quant au déroulement et à la compréhension du dialogue, tout en étant un obstacle à l'application de certains traitements automatiques. Nous proposons une méthode pour la détection automatique de disfluences dans des données conversationnelles en français. Nous développons et évaluons une méthode d'étiquetage séquentiel à base de CRF, reconnus pour donner de très bons résultats sur des tâches similaires. Notre système s'appuie sur des indices à la fois lexicaux et prosodiques : la combinaison de traits et de patrons construits à partir des transcriptions manuelles et du signal acoustique donne les meilleurs résultats pour de la détection de frontière, mais les indices acoustiques semblent dégrader les résultats pour une tâche conjointe d'identification de frontière et de structuration de la disfluence.

ABSTRACT

Disfluencies in conversational spontaneous speech : automatic detection using lexical and acoustic features

Spontaneous speech may be characterized by a relatively high rate of disfluencies. Keeping disfluencies in transcriptions may help to better understand the relative progress in dialogue, however their presence may harm a smooth application of various automatic processings. In this contribution, we propose a method to automatic disfluency detection in conversational speech. To this aim, we implemented a CRF (Conditional Random Field) approach to tag our data into disfluent/non-disfluent regions. CRFs are known to perform quite well on this kind of segmentation/labeling tasks. Implemented features include both lexical, morpho-syntactic and acoustic-prosodic information. Contrastive experiments show that the addition of features including information extracted from the speech signal beyond the mere transcription level give the best overall results in automatic disfluency region detection but seem to degrate the results for both region detection and structuration task.

MOTS-CLÉS : champs aléatoires conditionnels, disfluences, parole conversationnelle, parole spontanée.

KEYWORDS: conditional random fields, disfluencies, conversational speech, spontaneous speech.

1 Introduction

La parole spontanée est caractérisée par la présence de nombreux éléments disfluents. Très variés du point de vue de leur forme et de leur fonction – hésitations vocaliques, marqueurs discursifs, auto-corrections, faux-départs, *etc.* –, ils sont souvent considérés comme des signaux du message oral en construction, à la manière d'un brouillon précédent la version finale d'un texte écrit.

Si ces phénomènes ont été longtemps considérés comme préjudiciables à la qualité du discours et à sa compréhension, ils sont depuis quelques années de plus en plus étudiés ; leur lien avec le discours est plus spécifiquement analysé (voir par exemple Shriberg (1994); Beliao et Lacheret (2013); Vasilescu *et al.* (2010)). Malgré ces nombreuses études sur la richesse informationnelle des marques de l'oral spontané, le terme plutôt connoté négativement de « phénomène disfluent » perdure. Nous l'adoptons selon la définition proposée par Blanche-Benveniste *et al.* (1990) : une disfluence intervient lorsque le déroulement syntagmatique de l'énoncé est brisé.

L'objectif de cette étude est de proposer une méthode permettant de détecter automatiquement les disfluences d'édition dans des données d'oral spontané conversationnel en français. Notre méthode est développée et évaluée sur des données conversationnelles issues de centres d'appels, fournies par EDF : l'information contenue dans ces conversations est très précieuse pour les industriels, car elle peut être utilisée pour améliorer la connaissance client et la gestion de la relation clients/entreprise. En particulier, elle peut contribuer à mettre en évidence les dysfonctionnements dans les pratiques relationnelles et les stratégies de communication clients/agents. Des données similaires ont fait l'objet de recherches approfondies en analyse de l'oral, particulièrement dans le projet DECODA (Béchet *et al.*, 2012).

Cependant, relativement peu de travaux ont été réalisés en détection automatique des disfluences sur des données orales en français : on peut citer le DISTAGGER, outil de détection des disfluences à base de patrons élaborés manuellement (Constant et Dister, 2010), ainsi que les travaux de Bouraoui et Vigouroux (2009) qui proposent dans une perspective phraséologique une méthode également à base de patrons pour la détection d'un certain nombre de phénomènes disfluents.

Notre contribution à la problématique de la détection automatique de disfluences sur des données orales spontanées est double : d'une part, concernant le phénomène à détecter, nous nous attachons au repérage de segments disfluents en vue de nettoyer les données pour des traitements avals mais surtout de les utiliser pour mieux comprendre et modéliser le discours ; d'autre part, nous faisons l'hypothèse qu'une approche hybride, basée à la fois sur le texte et sur des indices acoustiques, permet de mieux modéliser les disfluences, liées à ces deux dimensions.

Nous présentons en section 2 nos données d'expérimentation et le modèle de disfluences ayant servi à leur annotation ; la section 3 décrit la méthode de détection automatique de disfluences que nous avons développée dans cette article ; celle-ci est testée et évaluée en section 4 ; enfin, la section 5 présente nos conclusions et les perspectives de travail qui en découlent.

2 Corpus de dialogues : conversations en centres d'appels

Nous utilisons le corpus VOXFACTORY, développé au sein du projet éponyme (Clavel *et al.*, 2013), continuité du projet Infom@gic–Callsurf. L'entreprise d'électricité française EDF a mené une campagne d'enregistrement dans un centre d'appels, produisant le corpus de données VOXFACTORY.

Cet ensemble de données, composé de dialogues entre des agents EDF et des clients particuliers, couvre un grand nombre de sujets liés aux services proposés par l'entreprise, comme par exemple des ouvertures de contrats, des questions liées à la facturation, des problèmes techniques, *etc.*

Un sous-ensemble du corpus VOXFACTORY, le corpus VOXDISS (60 conversations entre des agents EDF et des clients individuels, pour environ 10 heures de parole), a été manuellement annoté en disfluences d'édition par l'entreprise Vecsys, partenaire du projet Vox Factory, à l'aide de l'outil d'annotation `Transcriber` (Barras *et al.*, 1998). Nous utilisons le corpus VOXDISS dans nos expérimentations pour évaluer notre méthode de détection automatique des disfluences d'édition en contexte de dialogue oral spontané homme/homme en français.

La stratégie suivie pour annoter ce corpus s'appuie sur le guide d'annotation de méta-données conçu par le Linguistic Data Consortium (Strassel, 2004). Quatre classes d'événements disfluents ont été annotées : les auto-corrections (d'un mot simple ou d'un groupe de mots, accompagnée d'un changement de sens), les répétitions (d'un mot simple ou d'un groupe de mots), les faux-départ (abandon de l'énoncé) et les disfluences complexes (succession ou imbrication de différents types de disfluences d'édition). L'annotation des disfluences consiste en une annotation séquentielle des formes de surface, avec l'identification des trois parties distinctes structurant la zone disfluente, comme formalisé dans le modèle qui suit (en mettant en avant un exemple de répétition), selon la structuration de Levelt (1983) et la terminologie de Shriberg (1994) :

(reparandum) * ⟨ phase d'édition ⟩ [repair]

« je voulais payer (par) * ⟨ euh ben ⟩ [par] carte »

FIGURE 1 – Structure interne des disfluences d'édition illustrée avec l'exemple d'une répétition avec phase d'édition issue du corpus VOXDISS.

Dans ce modèle, le *reparandum* réfère à l'énoncé que le locuteur souhaite modifier. La phase d'édition, qui est une phase optionnelle, fait référence au segment de préparation de la correction. Le symbole * représente le point d'interruption qui initie la phase d'édition (le moment où le locuteur décide de modifier son énoncé). Le *repair* représente la phase de correction de l'énoncé.

3 Détection automatique des disfluences d'édition

Nous avons choisi de développer une méthode de détection automatique à l'aide des champs aléatoires conditionnels. En effet, non seulement cette méthode d'apprentissage automatique est connue pour donner d'excellents résultats sur des tâches similaires à la nôtre (par exemple pour la détection d'entités nommées) mais surtout les recherches récentes ayant abordé la détection de disfluences ont montré que les CRF donnaient les meilleurs résultats par rapport à d'autres modèles comme les modèles de Markov cachés ou l'entropie maximale (Liu *et al.*, 2006).

Nous présentons en section 3.1 les étiquettes définies pour les deux expériences de détection actuellement menées. La section 3.2 décrit notre système à base de CRF, son principe et sa mise en œuvre. Enfin, en 3.3 nous présentons les traits lexico-acoustiques et les patrons d'apprentissage que nous avons utilisés dans nos expériences.

3.1 Définition de la tâche et étiquettes de disfluences

La majorité des études existantes sur la détection automatique de disfluences a pour objectif la suppression des événements disfluents dans l'idée notamment de nettoyer les transcriptions. Le but principal de ces recherches est d'améliorer la lisibilité des données orales ou d'améliorer les modules avals de traitement automatique du langage naturel (voir par exemple Qian et Liu (2013)). La suppression se fait parfois même directement dans le flux d'une autre tâche, conjointe, par exemple d'analyse syntaxique (Rasooli et Tetreault, 2013). En fonction des objectifs de détection, l'accent est mis sur les événements disfluents de type pauses remplies (*filler words*) et/ou disfluences d'édition, et plus précisément sur une partie seulement des éléments composant leur structure (comme le *reparandum* ou le *repair*). En supprimant les zones disfluentes, la plupart des études se concentrent donc exclusivement sur la détection du *reparandum* (Georgila, 2009; Rasooli et Tetreault, 2013), éventuellement associé à la phase d'édition.

En complément de cet aspect de suppression de la zone disfluente à proprement parler, qui nous semble également important à des fins d'amélioration de la lisibilité des données, nous nous intéressons tout autant à la manière dont le locuteur modifie son énoncé qu'à la partie qu'il souhaite modifier, considérée comme la partie erronée. Ce segment de correction, le *repair*, peut être précieux dans la compréhension d'un dialogue : dans le cadre d'une auto-correction, le lexique choisi pour se corriger n'est pas fortuit (est-ce une correction visant à préciser le sens du propos ? un choix de synonyme visant à rendre le propos plus neutre ? *etc.*).

Nous avons ainsi choisi, pour initier nos travaux sur la détection automatique de disfluences d'édition, de réaliser deux séries d'expérience : la première (*XP1*) consiste à repérer les frontières des zones disfluentes, sans chercher à identifier les différents éléments composant la structure interne de cette zone. Notre seconde expérience (*XP2*) consiste à repérer un premier niveau de structure de la zone disfluente, en cherchant à séparer la partie « corrigée » (soit l'agrégat du *reparandum* et de la phase d'édition) de la zone de « réparation » de la disfluence (soit le *repair*).

Étiquette définie pour l'*XP1* Dans notre première expérience nous cherchons à identifier la totalité d'un segment disfluent (*reparandum*, phase d'édition et *repair*) de manière à pouvoir l'isoler du reste de l'énoncé. Nous avons donc défini une étiquette unique, déclinée selon si le token étiqueté débute un segment ou se trouve à l'intérieur de ce dernier (respectivement ***B-Dis*** et ***I-Dis***). Dans les autres cas, les tokens sont étiquetés « NULL ». Ainsi, dans l'exemple présenté en section 2, l'étiquetage de référence est le suivant : « je$_{NULL}$ voulais$_{NULL}$ payer$_{NULL}$ par$_{B-Dis}$ euh$_{I-Dis}$ ben$_{I-Dis}$ par$_{I-Dis}$ carte$_{NULL}$ ».

Étiquettes définies pour l'*XP2* Pour cette deuxième expérience nous avons défini des étiquettes différenciant la partie réellement disfluente (*reparandum* et phase d'édition le cas échéant) de sa réparation (*repair*). Cette structuration plus fine nous permet d'isoler d'une part la zone de la disfluence considérée pour le nettoyage (rapprochement de l'énoncé à une forme d'énonciation « finalisée ») et d'autre part les éléments lexicaux choisis par le locuteur pour « corriger » son énoncé. Dans cette perspective, les étiquettes suivantes ont été définies : pour la partie disfluente les labels ***B-Rpd-Edp*** (token initiant le segment) et ***I-Rpd-Edp*** (token à l'intérieur du segment) ; pour la partie *repair* les étiquettes ***B-Rpr*** (token initiant le segment) et ***I-Rpr*** (token à l'intérieur du segment) ; enfin, pour les tokens hors disfluence d'édition, l'étiquette « NULL ». Pour le même exemple que précédemment, l'étiquetage de référence devient : « je$_{NULL}$ voulais$_{NULL}$ payer$_{NULL}$ par$_{B-Rpd-Edp}$ euh$_{I-Rpd-Edp}$ ben$_{I-Rpd-Edp}$ par$_{B-Rpr}$ carte$_{NULL}$ ».

3.2 Choix des CRF et principes de mise en œuvre

Les champs aléatoires conditionnels (ou CRF, *cf.* Lafferty *et al.* (2001)) constituent un modèle d'apprentissage supervisé utilisé pour l'étiquetage séquentiel. Une de leur particularité est de pouvoir prendre en compte de grands ensembles de traits potentiellement redondants et d'intégrer des dépendances structurelles entre les classes que l'on cherche à identifier.

Nous avons utilisé la mise en œuvre des CRF fournie par `Wapiti` [1] (Lavergne *et al.*, 2010). Nous utilisons l'algorithme `rprop`, qui non seulement a déjà donné de bons résultats sur des tâches de détection structurées et complexes proches de la notre (voir par exemple Dinarelli et Rosset (2011)) mais surtout donne les meilleurs résultats sur notre corpus de développement. Le critère d'arrêt de l'algorithme a été empiriquement fixé à 500 itérations maximum, compte tenu des résultats obtenus sur le corpus de développement.

3.3 Traits et patrons lexico-acoustiques utilisés par le système

Nous avons développé une version minimal de notre système (*baseline*) sur le modèle de celle utilisée par Qian et Liu (2013). Les patrons sont simplement basés sur la forme fléchie des mots et sont appliqués sur une fenêtre de -2/+2 mots. N'utiliser que les informations lexicales au niveau local permet de mieux mesurer l'apport des traits et des patrons lexico-acoustiques utilisés pour enrichir le système.

Nous distinguerons trois versions du système : CRF_L (basé sur des patrons lexicaux, *cf.* tableau 1), CRF_A (basé des patrons acoustiques, *cf.* tableau 2) et CRF_LA (basé sur des patrons lexico-acoustiques). Les patrons lexicaux considèrent des traits purement linguistiques, comme les parties du discours (générées automatiquement par une version adaptée au français du tagger `Brill` (Allauzen et Bonneau-Maynard, 2008)). Les patrons acoustiques sont des traits classiquement employés pour caractériser la parole d'un point de vue acoustique. Tous les attribus numériques continus ont été discrétisés à l'aide de l'outil `discretize4crf` [2].

Séquences lexicales : unigrammes/bigrammes, fenêtre [-2,+2]
- forme fléchie ;
- traits de partie du discours.
Expressions rationnelles : unigrammes, fenêtre mot courant
- préfixes/suffixes aux positions 1 : 4 (soit 8 traits en tout) ;
Patrons binaires (oui/non) en expressions rationnelles : unigrammes, fenêtre [-1,+1]
- est entièrement en capitales ? (fenêtre mot courant)
- commence par une capitale ? (fenêtre mot courant)
- comporte une ponctuation ? (fenêtre [-1,+1])
- est une ponctuation ? (fenêtre mot courant)
- comporte une ponctuation (premier/dernier caractère exclus) ? (fenêtre mot courant)
- comporte un chiffre ? (fenêtre [-1,+1])
- est un nombre ? (fenêtre mot courant)

TABLE 1 – Patrons utilisés par le modèle de CRF reposant sur des traits lexicaux (CRF_L).

1. http://wapiti.limsi.fr/.
2. https://gforge.inria.fr/projects/discretize4crf/.

Séquences acoustiques : unigrammes/bigrammes, fenêtre [-2,+2]
- Prononciation du mot ;
- Durée du mot ;
- Nombre de phonèmes du mot ;
- Durée moyenne des phonèmes du mot.

TABLE 2 – Patrons utilisés par le modèle de CRF reposant sur des traits acoustiques (CRF_A).

4 Expériences

Nous avons mené deux expériences pour la détection de disfluences, correspondant aux tâches et étiquettes définies précédemment (*XP1* et *XP2*). Nous détaillons en section 4.1 le corpus d'expérimentation VOXDISS. Enfin, nous décrivons et discutons les résultats en sections 4.2 (*XP1*) et 4.3 (*XP2*).

Les mesures utilisées dans nos expériences correspondent aux mesures d'évaluation classiques de Précision, Rappel, F-mesure et SER ou *Slot Error Rate* (Makhoul *et al.*, 1999) [3].

4.1 Données d'expérimentation

Le tableau 3 présente la subdivision du corpus VOXDISS opérée pour l'entraînement, le développement et le test de notre système de détection des disfluences d'édition. Nous avons sélectionné aléatoirement (au sein des quartiles du corpus) des conversations suivant une distribution homogène sur la durée des dialogues.

Corpus	**Nb. appels**	**Dur. tot.**	**Dur. moy.**	**Nb. moy. loc.**	**Nb. moy. dis.**	**Densité lex. dis.**
Train	48	9h19	12m04	2.1	62.54	10.91%
Dev	5	1h01	12m20	2.2	54.20	8.6%
Test	7	1h13	11m29	3	37.14	7.4%

TABLE 3 – Subdivision du corpus VOXDISS en données d'entraînement, de développement et de test. Densité lex. dis. : pourcentage des mots du corpus dans une zone disfluente.

4.2 Identification des frontières des segments disfluents (*XP1*)

Nous cherchons ici à identifier la frontière des disfluences d'édition, *reparandum*, phase d'édition et *repair* sans structuration interne. L'objectif de cette première tâche est d'identifier, localiser et quantifier les disfluences d'édition dans des données orales spontanées. Le tableau 4 présente les résultats obtenus par les différentes versions du système (Baseline, CRF_L, CRF_A et CRF_LA) sur les données de test.

3. Le SER mesure les erreurs en terme d'insertions, de substitutions (frontières et types) et de suppressions. Cette mesure s'apparente au WER (ou *Word Error Rate*), utilisé en Reconnaissance Automatique de la Parole.

Modèle	Précision	Rappel	F-mesure	SER
Baseline	0.490	0.187	0.270	0.823
CRF_L	0.523	0.254	0.342	0.776
CRF_A	0.422	0.183	0.255	0.810
CRF_LA	0.523	0.257	0.345	0.765

TABLE 4 – Performances de détection de disfluences par CRF sans isolement du *repair*.

L'observation la plus frappante est la dégradation des résultats dès lors que l'on se base sur des indices acoustiques uniquement. Les résultats obtenus par CRF_A sont en effet moins bons que ceux obtenus par la baseline, à la fois en Précision, Rappel, F-mesure et Slot Error Rate.

Toutefois, les meilleurs résultats sont obtenus par CRF_LA, lorsque l'on combine des patrons lexicaux et acoustiques, ce qui laisse présager de bonnes perspectives d'amélioration sur cette voie. En effet, cette configuration utilisant des indices mixtes obtient un gain de 7,5 points en F-mesure (F=27% pour la baseline *versus* 34,5% pour CRF_LA) et de 5,8 points pour le SER (SER=82,3% pour la baseline *versus* 76,5% pour CRF_LA).

Il est également intéressant de voir que la précision reste stable entre CRF_L et CRF_LA : l'ajout des indices acoustiques, qui utilisés seuls donnent les moins bons résultats, permet un gain sur toutes les autres mesures sans pour autant dégrader la précision.

4.3 Identification et structuration des segments disfluents (*XP2*)

Nous cherchons ici à identifier la frontière des disfluences d'édition, *reparandum*, phase d'édition et *repair* avec structuration interne. L'objectif de cette seconde tâche est similaire à celui poursuivi précédemment (*XP1*), mais cette fois le système cherche à isoler la partie réellement disfluente (*reparandum* + phase d'édition) de sa réparation (*repair)*. Les résultats obtenus par les différentes versions du système sur les donnée de test sont présentés dans le tableau 5.

Modèle	Précision			Rappel			F-mesure			SER
	Rpd-Edp	Rpr	All	Rpd-Edp	Rpr	Rpd-Edp	All	Rpr	All	
Baseline	0.596	0.723	0.654	0.225	0.283	0.251	0.327	0.407	0.363	0.790
CRF_L	0.550	0.626	0.586	0.233	0.292	0.259	0.327	0.399	0.360	0.794
CRF_A	0.482	0.557	0.517	0.206	0.255	0.228	0.289	0.350	0.316	0.816
CRF_LA	0.586	0.688	0.632	0.260	0.311	0.283	0.360	0.429	0.391	0.752

TABLE 5 – Performances de détection de disfluences par CRF avec isolement du *repair*. Rpd-Edp : *reparandum* + phase d'édition. Rpr : *repair*.

Considérant les résultats obtenus pour cette seconde tâche, notons la supériorité du modèle combinant des patrons lexicaux et acoustiques (CRF_LA). Excepté en précision, CRF_LA obtient en effet les meilleurs résultats pour toutes les mesures d'évaluation, avec un gain de 2,8 points en F-mesure pour l'ensemble des étiquettes (F=36,3% pour la baseline *versus* 39,1% pour CRF_LA) et de 3,8 points pour le SER (SER=79% pour la baseline *versus* 75,2% pour CRF_LA).

En prenant en compte les résultats de l'*XP1*, il est clair que cette deuxième tâche est davantage per-

formante. Ces résultats sont encourageants quant à la poursuite de nos travaux sur l'identification de la structure des disfluences d'édition.

Enfin, le *repair* semble plus évident à localiser, son identification donnant de meilleurs résultats dans toutes les configurations du système. Cela peut être expliqué par le fait que la zone actuellement étiquetée *Rpd-Edp* nécessite également une structuration plus fine, dans la mesure où elle combine actuellement *reparandum* et phase d'édition, deux zones aux propriétés lexicales relativement différentes.

5 Conclusion et Perspectives

Nous avons présenté dans cette étude deux expériences de détection automatique de disfluences d'édition sur des données orales conversationnelles en français. Dans l'objectif de nettoyer les données orales de ces zones disfluentes et d'utiliser ces phénomènes en compréhension du discours, nous avons développé une méthode à base de CRF utilisant à la fois des indices lexicaux et acoustiques. Nous avons développé un premier jeu de patrons pour notre système, que nous avons testés individuellement puis combinés. Nous avons envisagé une détection des zones disfluentes avec isolement du *repair* et sans structuration interne.

Les résultats préliminaires obtenus sont prometteurs, et nous confortent dans l'idée d'utiliser conjointement des indices lexicaux et acoustiques. Actuellement, les patrons acoustiques sont moins développés que les patrons lexicaux, ce qui explique la faiblesse des résultats obtenus avec cette méthode : nous envisageons une utilisation plus fine des traits déjà identifiés, et allons enrichir cette méthode de nouveaux traits et patrons acoustiques (fréquence fondamentale, formants, silences, *etc.*). De plus, à ce stade de nos recherches, nous travaillons exclusivement sur les transcriptions manuelles des conversations. Cela explique également la plus grande efficacité des patrons lexicaux. Nous projetons d'évaluer notre système de détection des disfluences en conditions réelles sur des sorties de reconnaissance automatique de la parole, et faisons l'hypothèse que dans cette configuration le modèle basé sur des patrons acoustiques gagne en performance.

Par ailleurs, les résultats obtenus en structurant les disfluences d'édition (isolement du *repair*) sont meilleurs et nous confortent également dans l'objectif de réaliser une identification structurée des disfluences d'édition. Ainsi, nous envisageons d'isoler la phase d'édition du *reparandum* dans la zone disfluente, et de tester l'utilisation d'étiquettes différenciant les quatre classes de disfluences présentes dans nos données (répétitions, auto-corrections, faux-départ, disfluences complexes).

Remerciements

Ce travail a été partiellement financé par la convention CIFRE 2011/0916.

Références

ALLAUZEN, A. et BONNEAU-MAYNARD, H. (2008). Training and Evaluation of POS Taggers on the French MULTITAG Corpus. *In Proceedings of LREC*.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a Free Tool for Segmenting, Labeling and Transcribing Speech. *In Proceedings of LREC*, pages 1373–1376.

BELIAO, J. et LACHERET, A. (2013). Disfluency and Discursive Markers : When Prosody and Syntax Plan Discourse. *In Proceedings of DiSS*, pages 5–8.

BLANCHE-BENVENISTE, C., BILGER, M., ROUGET, C. et VAN DEN EYNDE, K. (1990). *Le français parlé : Études grammaticales*. Éditions du CNRS.

BOURAOUI, J.-L. et VIGOUROUX, N. (2009). Traitement automatique de disfluences dans un corpus linguistiquement contraint. *In Actes de TALN*.

BÉCHET, F., MAZA, B., BIGOUROUX, N., EL-BÈZE, M., DE MORI, R. et ARBILLOT, E. (2012). DECODA : a Call-center Human-human Spoken Conversation Corpus. *In Proceedings of LREC*.

CLAVEL, C., ADDA, G., CAILLIAU, F., GARNIER-RIZET, M., CAVET, A., CHAPUIS, G., COURCINOUS, S., DANESI, C., DAQUO, A.-L., DELDOSSI, M., GUILLEMIN-LANNE, S., SEIZOU, M. et SUIGNARD, P. (2013). Spontaneous Speech and Opinion Detection : Mining Call-centre Transcripts. *Language Resources & Evaluation*, 1:40.

CONSTANT, M. et DISTER, A. (2010). *Spoken Communication*, chapitre Automatic Detection of Disfluencies in Speech Transcriptions, pages 259–272. Cambridge Scholars Publishing.

DINARELLI, M. et ROSSET, S. (2011). Models Cascade for Tree-Structured Named Entity Detection. *In Proceedings of IJCNLP*, pages 1269–1278.

GEORGILA, K. (2009). Using Integer Linear Programming for Detecting Speech Disfluencies. *In Proceedings of NAACL HLT*, pages 109–112.

LAFFERTY, J. D., MCCALLUM, A. et PEREIRA., F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of ICML*.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical Very Large Scale CRFs. *In Proceedings of ACL*, pages 504–513.

LEVELT, W. J. M. (1983). Monitoring and Self-repair in Speech. *Cognition*, 14:41–104.

LIU, Y., SHRIBERG, E. E., STOLCKE, A., HILLARD, D., OSTENDORF, M. et HARPER, M. (2006). Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1526 – 1540.

MAKHOUL, J., KUBALA, F., SCHWARTZ, R. et WEISCHEDEL, R. (1999). Performance Measures For Information Extraction. *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252.

QIAN, X. et LIU, Y. (2013). Disfluency Detection Using Multi-step Stacked Learning. *In Proceedings of NAACL HLT*, pages 820–825.

RASOOLI, M. S. et TETREAULT, J. (2013). Joint Parsing and Disfluency Detection in Linear Time. *In Proceedings of EMNLP*, pages 124–129.

SHRIBERG, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Thèse de doctorat, Berkeley University of California.

STRASSEL, S. (2004). *Simple Metadata Annotation Specification*. Linguistic Data Consortium.

VASILESCU, I., ROSSET, S. et ADDA-DECKER, M. (2010). On the Functions of the Vocalic Hesitation euh in Interactive Man-machine Question Answering Dialogs in French. *In Proceedings of DiSS/LPSS*.

Traduction de la parole dans le projet RAPMAT

Hélène Bonneau-Maynard[1] Natalia Segal[1] Eric Bilinski[1] Jean-Luc Gauvain[1]
Li Gong[1] Lori Lamel[1] Antoine Laurent[1] François Yvon[1]
Julien Despres[2] Yvan Josse[2] Viet Bac Le[2]
(1) LIMSI-CNRS B.P. 133 91403 Orsay Cedex
(2) Vocapia Research 28, rue Jean Rostand Parc Orsay Université 91400 Orsay
{maynard,segal,bilinski,laurent,gauvain,gong,lamel,yvon}@limsi.fr,
{despres,joss,levb}@vocapia.com

RÉSUMÉ

Le projet RAPMAT vise à développer des systèmes de traduction de la parole en s'intéressant aux deux traitements constitutifs de la chaîne complète : la reconnaissance de la parole (RAP) et la traduction (TA). Dans la situation classique, les modèles statistiques utilisés par les deux systèmes sont estimés indépendemment, à partir de données de différentes natures (transcriptions manuelles de données de parole pour la RAP et corpus bilingues issus de données textuelles pour la TA). Nous proposons une approche semi-supervisée pour l'adaptation des modèles de traduction à la traduction de parole, dans laquelle les modèles de TA sont entraînés en intégrant des transcriptions manuelles et automatiques de la parole traduites automatiquement. L'approche est expérimentée sur la direction de traduction français vers anglais. Un prototype de démonstration sur smartphones, incluant notamment la traduction de parole pour les paires de langues français/anglais et français/chinois a été développé pour permettre la collecte de données.

ABSTRACT

Speech translation in the RAPMAT project

The goal of the RAPMAT project is to develop automatic speech-to-speecj translation systems focusing on the interconnection of its two main processing steps : automatic speech recognition (ASR) and machine translation (MT). The basic approach consists of training independent statistical models for each of the systems, using data of different nature : manual transcriptions of speech are used for ASR, whereas the MT models are trained using bilingual corpora of written texts. In this paper we present a semi-supervised approach for adaptation of the translation models to data produced by ASR, where the training corpus includes machine translated automatic and manual transcriptions of speech. This approach was tested on translation from French to English. An application prototype for smartphones has been developed for speech translation of French/English and French/Chinese language pairs that will also serve to collect additional data.

MOTS-CLÉS : reconnaissance de la parole, traduction automatique.

KEYWORDS: speech recognition, automatic translation.

1 Introduction

La traduction de la parole est un des défis sociétaux majeurs dont l'objectif est de permettre la communication parlée entre personnes s'exprimant dans des langues différentes. Le défi est également scientifique puisqu'il se situe à la jonction de deux domaines de recherche que sont la reconnaissance automatique de la parole (RAP) (Gauvain et Lamel, 2008) et la traduction automatique (TA) (Allauzen et Yvon, 2012). Si ces deux domaines ont connu dans les dernières décennies des progrès considérables grâce à l'introduction de modèles statistiques, il n'existe pas encore de solution satisfaisante pour leur couplage (Stüker *et al.*, 2012).

Le projet DGA RAPMAT (Reconnaissance Automatique de la Parole Multilingue Accentuée pour la Traduction) vise la réalisation de systèmes de traduction de la parole pour rendre cette technologie accessible à des usagers en déplacement, via un smartphone. Un prototype de démonstration été développé sur les plate-formes iOS et Androïd permettant d'expérimenter les fonctionnalités proposées. Concernant la traduction, c'est l'anglais, le chinois et le vietnamien qui sont traités, l'objectif étant d'offrir leur traduction depuis et vers le français. Trois autres langues sont traitées dans le projet.

Après une brève présentation du projet RAPMAT, l'article se focalise sur les aspects traduction de la parole. La section 3 situe les travaux dans le contexte scientifique. Les caractéristiques du système de reconnaissance de la parole sont décrites dans la section 4. La section 5 décrit les approches proposées pour l'adaptation des modèles de traduction à la traduction de la parole. La description des données et les aspects expérimentaux sont décrits dans les parties 6.1 et 6.

2 Le projet RAPMAT

Le projet DGA RAPMAT (24 mois, mai 2012 - mai 2014), dont les partenaires sont le LIMSI-CNRS et la société Vocapia Research [1], vise à développer des systèmes de traduction de la parole en s'intéressant aux deux traitements constitutifs de la chaîne complète : la reconnaissance de la parole puis la traduction de transcriptions. Par ailleurs, la participation du LIMSI au consortium U-STAR [2] permet de bénéficier de ses infrastructures de collecte de données offrant l'accès à des corpus dans de nombreuses langues.

En reconnaissance de la parole, l'accent est mis sur le développement de systèmes en vue de la traduction. Les langues pour lesquelles il est prévu de réaliser un système de transcription automatique seront dans un premier temps le français et l'anglais (avec adaptation pour la prise en compte d'accents non natifs, essentiellement asiatiques). Au total 7 langues sont traitées dans ce projet, Concernant la traduction, l'anglais et le chinois ont été traités en priorité, l'objectif étant d'offrir leur traduction depuis et vers le français. La construction des systèmes dépend de la quantité de données disponibles. Afin de mettre à jour les modèles de transcription et de traduction, deux ensembles de corpus multilingues de type "guide de conversation et requêtes touristiques" ont été utilisés au sein du projet : U-STAR (corpus HIT) et BTEC. Ces données ont servi à l'adaptation des modèles existants. Pour les langues pour lesquelles Vocapia Research et le LIMSI ne disposent pas encore de système, des données d'apprentissage ont été collectées

1. http ://www.vocapia.com
2. http ://www.ustar-consortium.com

sur internet (transcriptions approximatives ou news). Ces données ont permis de faire un apprentissage peu supervisé des modèles, évitant ainsi des coûts prohibitifs de transcription manuelle.

3 Traduction de la parole : État de l'art

Il existe plusieurs façons de coupler le système de reconnaissance de la parole avec le système de traduction automatique pour faire la traduction de la parole. L'approche découplée la plus élémentaire consiste à donner en entrée du système de traduction la séquence de mots la plus probable issue du système de reconnaissance de la parole, sans se soucier de l'absence d'adéquation entre les modèles qu'utilisent ces deux systèmes.

Plusieurs manières d'améliorer ce couplage simpliste ont été explorées dans la littérature, notamment à travers les évaluations IWSLT (Mauro *et al.*, 2013) :

– L'approche semi-découplée : permet au module de RAP de produire plusieurs hypothèses (n-meilleures solutions, treillis, réseaux de confusions) laissant ainsi au système de traduction la possibilité de choisir l'hypothèse de transcription qui produira la traduction la plus probable (Matusov et Ney, 2011). Cette approche, qui permet souvent d'obtenir des améliorations globales de performances, est une extension de l'approche découplée, car le système de traduction et le système de reconnaissance sont entraînés indépendamment et prennent leurs décisions d'une façon indépendante.
– L'approche intégrée : compose le modèle acoustique et le modèle de traduction en un système de décodage à un seul passage (Perez *et al.*, 2012). Cette approche impose une contrainte sur le modèle de traduction qui doit être implanté sous la forme d'un transducteur fini pour pouvoir réaliser la composition des modèles.
– L'adaptation du modèle de traduction qui consiste à modifier la partie source des corpus d'entraînement des modèles de traduction pour la rendre compatible avec la sortie produite par le système de reconnaissance de la parole.

La dernière de ces solutions, s'est avérée efficace pour améliorer la qualité de la traduction (Peitz *et al.*, 2012) ; elle nécessite toutefois une grande quantité de données orales pour lesquelles on disposerait également de la traduction manuelle de référence. Ces données sont coûteuses à obtenir, surtout pour un type d'application particulier tel que le dialogue. En revanche, il est plus facile d'obtenir, d'un côté, des données de parole pour lesquelles la transcription manuelle de référence est disponible, et de l'autre côté, des traductions manuelles de textes écrits. Nous pouvons ainsi entraîner un système de traduction de base sur des textes écrits (apprentissage supervisé) et ensuite enrichir ce système en utilisant les traductions automatiques faites à partir des transcriptions de référence de la parole (apprentissage semi-supervisé). Des approches utilisant l'apprentissage semi-supervisé ont déjà été proposées pour adapter un système de traduction automatique à un domaine en utilisant comme corpus d'adaptation les traductions automatiques faites par le système de base (Schwenk, 2008; Gahbiche-Braham *et al.*, 2011) et pour la portabilité d'un système de dialogue d'une langue vers une autre (Jabaian *et al.*, 2013). Nous présentons une adaptation similaire pour les données issues de la reconnaissance de la parole.

4 Reconnaissance de la parole

Le système de transcription de la parole pour le français, développé conjointement par le LIMSI et Vocapia Research, repose sur deux composants principaux : un partitioneur parole/non-parole et un décodeur de mots. Le décodeur de mots utilise un modèle acoustique HMM et un modèle de langage bi-gramme pour construire un treillis de mots qui est ensuite réévalué par modèle de langue quadri-gramme. Le système produit alors une transcription automatique, alignée temporellement au niveau des mots, dans un fichier XML.

Modèles acoustiques : L'apprentissage des modèles acoustiques est réalisé en alignant une transcription orthographique exacte sur le signal de parole au moyen d'un jeu de modèles phonétiques et d'un lexique de prononciation. Ces modèles sont entraînés à partir de paramètres MLP (Multi-Layer Perceptron) concaténés avec des paramètres PLP (Perceptual Linear Prediction) et la fréquence fondamentale. Ils ont bénéficié d'un apprentissage discriminant (MMIE) et adapté aux locuteurs (SAT). Une adaptation par blocs diagonaux a été utilisé pour les paramètres MLP et PLP+F0. Le vecteur MLP+PLP+f0 est composé de 81 paramètres ($39MLP + 39PLP + F0 + \Delta F0 + \Delta\Delta F0$). Les modèles acoustiques sont des HMMs (Hidden Markov Models) composés d'un mélange de gaussiennes. Ils sont à états liés, dépendants du contexte et du sexe du locuteur. Le silence est modélisé par un seul état composé de 1024 Gaussiennes. Au final, à partir d'un corpus de 866 heures de parole, 19679 modèles contextuels ont été entraînés, partageant 11517 états, pour un total de 370k gaussiennes.

Modèles de langue : 2.9 milliards de mots ont été utilisés pour entraîner des modèles de langue génériques. Les modèles sont construits autour d'un vocabulaire de 200K mots. Les modèles génériques ont été adaptés pour le projet RAPMAT en les interpolant avec des modèles spécifiques construits à partir des données des corpora BTEC et HIT décrits dans le paragraphe 6.1.

Modèle de prononciations : Les prononciations du dictionnaire sont obtenues par une phonétisation automatique (plus de 1000 règles), plusieurs possibilités de phonétisation sont générées pour les noms propres. Une passe de correction manuelle a été faite pour réduire les cas de prononciations multiple ce qui permet d'obtenir un gain en vitesse lors du décodage. Les probabilités de prononciations, obtenues par l'observation des prononciations, sont ajoutées au dictionnaire lors de l'apprentissage acoustique (Gauvain *et al.*, 2005).

5 Adaptation des modèles de traduction

Comme exposé ci-dessus (section 3) la traduction de parole peut bénéficier d'un couplage plus étroit entre les deux modules, en particulier pour que la traduction prenne d'une part en compte certaines spécificités du langage parlé, et d'autre part les erreurs produites par la transcription automatique. Notre approche pour réaliser ce couplage s'inspire de techniques d'adaptation au domaine selon le principe suivant.

Dans un premier temps, nous rapprochons les données d'entraînement pour le modèle de traduction des données en sortie de reconnaissance en enlevant toute la ponctuation de la source

dans le corpus d'apprentissage. Par la suite, l'adaptation du modèle de traduction aux données RAP est effectuée de deux manières différentes :

– Un petit corpus de parole, pour lequel la traduction manuelle de référence est disponible, est utilisé pour faire le réglage du modèle de traduction de base.
– Un grand corpus de parole, pour lequel uniquement la transcription manuelle de référence est disponible, est traduit automatiquement avec le système de traduction de base. Les traductions automatiques ainsi obtenues, alignées à la fois avec la transcription de référence et avec les sorties de RAP, sont utilisées pour l'entraînement d'un nouveau modèle de traduction plus adapté aux données de la parole.

6 Expériences et résultats

6.1 Données

Les corpus parallèles du tableau 1 ont été utilisés pour entraîner et développer le modèle de traduction de base. Notons que ces corpus sont de natures très différentes, le corpus HIT avec des phrases de dialogues courtes étant le mieux adapté à l'application visée. Les corpus parallèles contiennent des signes de ponctuation produits manuellement.

– BTEC (Kikui *et al.*, 2006) : corpus de dialogues autour des transports et du tourisme.
– HIT (Yang *et al.*, 2006) (Harbin Institute of Technology) : corpus de dialogues couvrant 5 domaines (voyages, restauration, sports, transports et business). Le corpus initial trilingue (anglais, chinois, japonais) est enrichi dans le cadre du projet U-STAR. Le LIMSI en a produit la traduction en français. Les langues disponibles à ce jour sont : français, anglais, chinois, japonais, vietnamien et hindi.
– Tatoeba : corpus multilingue couvrant des domaines divers provenant de traductions collaboratives en ligne [3], utilisé uniquement pour la paire français/anglais.
– TedTalk [4] : corpus de transcriptions et de traductions d'exposés vidéos, utilisé uniquement pour la paire français/chinois.

Pour l'adaptation semi-supervisée des modèles de traduction, un corpus d'actualités radio-télévisées monolingue en français (BN (Lamel et al, 2011)) a été transcrit automatiquement par le système de RAP (dernière ligne du tableau 1), avec un taux d'erreur mots de 12%. Le corpus BN contient des signes de ponctuation produits automatiquement par le système de reconnaissance de la parole.

Deux types de corpus ont été utilisés pour le développement et les tests : les corpus dev 2004 (500 phrases) et test 2009 BTEC (469 phrases) des évaluations IWSLT, et les corpus dev (436 phrases) et test (437 phrases) Tatoeba pour laquels on dispose du signal de parole en français de sa transcription manuelle, et de sa traduction en anglais.

3. http ://tatoeba.org
4. https ://wit3.fbk.eu/archive/2012-02//texts/zh-cn/fr/zh-cn-fr.tgz

Corpus	Lignes	Tokens FR	Tokens EN	Tokens ZH
BTEC	19 972	202 559	177 370	N/A
HIT	62 226	671 694	597 528	588 501
Tatoeba	111 168	1 038 548	928 936	N/A
TedTalk	142 296	2 925 803	N/A	2 614 332
BN	487 856	9 843 627	N/A	N/A

TABLE 1 – Statistiques des corpus.

6.2 Performances des systèmes de traduction de base

Pour l'ensemble des systèmes développés, les données parallèles sont alignées mot à mot avec MGgiza++ [5]. Le décodeur Moses [6] est ensuite utilisé pour symétriser les alignements en utilisant l'heuristique grow-diag-final-and et pour extraire les phrases avec une longueur maximale de 7 mots. Les poids des paramètres sont optimisés par MERT sur le corpus de développement. La traduction est réalisée par le décodeur Moses fonctionnant en mode "serveur".

Les systèmes de traduction de base ont été développés pour les paires français/anglais et français/chinois. L'ensemble des textes parallèles a été tokenisé côté source pour les rendre totalement compatibles avec la tokenisation des sorties du système de RAP.

Dans les textes chinois, toutes les ponctuations sont normalisées en ponctuations chinoises. Les chiffres et les lettres alphabétiques sont codés en ASCII. Au niveau de l'écriture, le texte chinois n'est pas segmenté : il n'y a pas de délimiteur explicite entre les mots. Pour construire les modèles de langage et les modèles de traduction où le mot joue un rôle primordial, une segmentation du texte chinois en mots est utilisée (Luo *et al.*, 2009). Ce module est basé sur un modèle N-gramme de mots.

Le tableau 2 résume les performances des systèmes de traduction de base réglés et testés sur les dev/test BTEC. Les corpus HIT, Tatoeba et BTEC ont été utilisés pour l'entraînement des modèles de traduction pour la paire français/anglais et les corpus HIT et TED pour la paire français/chinois.

Direction de traduction	Score BLEU
FR->EN	62,8 (6 références)
EN->FR	53,0
ZH->FR	25,0
FR->ZH	10,2

TABLE 2 – Performances sur le test 2009 BTEC des systèmes de traduction de base avec ponctuation en source.

5. http ://sourceforge.net/projects/mgizapp/
6. http ://www.statmt.org/moses/

6.3 Adaptation de la traduction automatique aux données RAP

Les expériences d'adaptation des modèles de traduction à la traduction de parole présentées ci-dessous ont été menées uniquement pour la direction du français vers anglais pour le moment, l'objectif étant de définir les techniques d'adaptation les plus prometteuses qui pourront par la suite être appliquées aux autres directions de traduction pertinentes pour le projet RAPMAT.

Le système de RAP utilisé offre la possibilité de produire une ponctuation automatique de ses sorties. Le tableau 3 résume les résultats de comparaison des modèles de traduction de base avec et sans ponctuation dans la partie source des données d'entraînement et de test. La ponctuation en cible a été gardée pour pouvoir viser directement la traduction correcte et complète sans post-traitement. Les performances ont été évaluées à la fois sur le test BTEC 2009 et sur les transcriptions exactes et automatiques (WER 22,5%) du test Tatoeba. Si une dégradation du score BLEU peut être observée pour les textes propres, ce n'est pas le cas sur les transcriptions automatiques (légère amélioration), où la ponctuation est produite automatiquement et peut contenir des erreurs. Il est important de noter que le corpus test BTEC contenait uniquement 3% de ponctuations (ce qui peut expliquer les performances très proches avec et sans ponctuation), tandis que le corpus test de Tatoeba contenait approximativement 12% de ponctuations dans la transcription manuelle et dans la transcription automatique. Dans les expériences suivantes nous avons donc utilisé les versions des corpus sans ponctuation en source pour l'entraînement, le réglage et les tests des modèles de traduction.

	test BTEC (6 références)	test Tatoeba	
		manuel	auto
Base, avec ponctuation en source	62,8	45,7	32,3
Base, sans ponctuation en source	62,9	45,1	32,5

TABLE 3 – Comparaison des scores BLEU des modèles de traduction avec et sans ponctuation en source, sur le test BTEC et les transcriptions manuelles du test Tatoeba

Adaptation par réglage : La première tentative consiste à utiliser des transcriptions automatiques lors de la phase de réglage des systèmes. Nous comparons ainsi le modèle de base (réglé sur le corpus de développement BTEC), à un modèle réglé en utilisant en plus les transcriptions manuelles et les transcriptions automatiques du corpus de développement Tatoeba. Les tests sont menés sur le test BTEC et sur les transcriptions manuelles et automatiques du test Tatoeba (première et deuxième ligne du tableau 4). On observe, comme on pouvait s'y attendre, que les performances baissent sur les données écrites (corpus de test BTEC) avec le réglage sur les données transcrites automatiquement. En revanche, les résultats de cette technique d'adaptation montrent une amélioration du score BLEU sur le corpus de test Tatoeba. Cette amélioration est plus prononcée pour les transcriptions automatiques que pour les transcriptions manuelles du même corpus, et nous pouvons donc conclure que l'adaptation est plus liée aux particularités de données produites par la RAP qu'au corpus lui-même.

Adaptation par production de données parallèles : La seconde approche consiste à intégrer un corpus de données transcrites automatiquement dans les données parallèles d'entraînement du système de traduction. Par manque de données parallèles qui correspondraient à des transcriptions automatiques de parole et à leur traduction manuelle, l'approche consiste à produire

Corpus Entraînement	Corpus Développement	Corpus Tests		
		BTEC	Tatoeba manuel	Tatoeba auto
Base	BTEC	62,9	45,1	32,5
Base	BTEC + Tatoeba (manuel + auto.)	61,8	45,3	33,3
Base + adaptation BN	BTEC	62,3	45,3	33,0
Base + adaptation BN	BTEC + Tatoeba (manuel+auto)	61,2	46,4	34,0

TABLE 4 – Comparaison des scores BLEU entre le modèle de traduction de base et le modèle entraîné avec les traductions automatiques des données RAP, avec réglage sur les données écrits et sur les données provenant de RAP.

artificiellement ce type de données en traduisant automatiquement le corpus monolingue français BN avec le modèle de traduction de base. Pour éviter des dégradations liées à la qualité médiocre de certaines traductions automatiques, un filtrage des traductions selon leur score de confiance (normalisé par rapport à la longueur de chaque traduction en mots) a été réalisé en éliminant les traductions dont le score normalisé est inférieur à la moyenne pondérée. Les traductions automatiques sélectionnées ont été alignées parallèlement à la fois à leurs transcriptions manuelles et à leurs transcriptions automatiques. Ces données parallèles, adaptées au système de RAP, sont utilisées comme un corpus supplémentaire pour l'entraînement du modèle de traduction. Le tableau 4 donne la comparaison entre le modèle de base et le modèle entraîné en utilisant en plus les données adaptées au système de RAP (avec les deux types de réglage présentés ci-dessus : 3ème et 4ème lignes). Le meilleur résultat est obtenu avec le modèle de traduction entraîné et réglé en utilisant les données provenant de la reconnaissance automatique.

7 Conclusion

Le projet RAPMAT a permis, entre autres objectifs, d'explorer les spécificités de la traduction de la parole. Nous avons proposé une approche d'apprentissage semi-supervisé permettant d'adapter les systèmes de traduction automatique à la traduction de la parole sans utiliser de grandes quantités des données de parole traduites manuellement. Le système de traduction ainsi adapté améliore les performances de traduction sur les transcriptions automatiques par rapport au système de base non adapté. Nous envisageons par la suite d'étendre cette approche aux autres directions de traduction du projet et de l'affiner en utilisant des modèles de traduction plus sophistiqués (notamment à multiples tables de traduction). D'autres approches d'adaptation et de création de données artificielles pour la traduction de la parole, vont également être explorées à court terme, telles que la génération de pseudo-transcriptions automatiques à partir de données textuelles.

Remerciements

Ces travaux ont été menés grâce au soutien de la DGA (Direction Générale de L'Armement) et de la DGCIS (Direction Générale de la Compétitivité, de l'Industrie et des Services).

Références

ALLAUZEN, A. et YVON, F. (2012). Statistical methods for machine translation. *In* GAUSSIER, E. et YVON, F., éditeurs : *Textual Information Access*, pages 223–304. ISTE/Wiley, London.

GAHBICHE-BRAHAM, S., BONNEAU-MAYNARD, H. et YVON, F. (2011). Two ways to use a noisy parallel news corpus for improving statistical machine translation. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, Portland, Oregon, USA.

GAUVAIN, J.-L., ADDA, G., LAMEL, L., LEFÈVRE, F. et SCHWENK, H. (2005). Transcription de la parole conversationnelle. *Traitement Automatique des Langues (TAL)*, 45.

GAUVAIN, J.-L. et LAMEL, L. (2008). Speech recognition systems. *In* MARIANI, J., éditeur : *Spoken Language Processing*.

JABAIAN, B., BESACIER, L. et LEFEVRE, F. (2013). Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE Transactions on Audio, Speech and Language Processing*, 21:636–648.

KIKUI, G., YAMAMOTO, S., TAKEZAWA, T. et SUMITA, E. (2006). Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 14:1674–1682.

LAMEL, L. et AL (2011). Speech recognition for machine translation in quaero. *In Proceeding of the 11th International Workshop on Spoken Language Translation*, San Francisco, USA.

LUO, J., LAMEL, L. et GAUVAIN, J.-L. (2009). Modeling characters versus words for mandarin speech recognition. *In International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

MATUSOV, E. et NEY, H. (2011). Lattice-based ASR-MT interface for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 19:721–732.

MAURO, C., JAN, N., SEBASTIAN, S., LUISA, B. et MARCELLO, F. (2013). Report on the 10th iwslt evaluation campaign. *In IWSLT*, Eidelberg, Allemagne.

PEITZ, S., WIESLER, S., NUSSBAUM-THOM, M. et NEY, H. (2012). Spoken language translation using automatically transcribed text in training. *In Proceeding of the 9th International Workshop on Spoken Language Translation*, Hong Kong, China.

PEREZ, A., TORRES, I. et CASACUBERTA, F. (2012). Finite-state acoustic and translation model composition in statistical speech translation : empirical assessment. *In Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, Donostia-San Sebastian, Spain.

SCHWENK, H. (2008). Investigations on largescale lightly-supervised training for statistical machine translation. *In Proceedings of the International Workshop on Spoken Language Translation*, Hawaii, USA.

STÜKER, S., HERRMANN, T., KOLSS, M., NIEHUES, J. et WOLFEL, M. (2012). Research opportunities in automatic speech-to-speech translation. *Potentials, IEEEe*, 3:26–33.

YANG, M., JIANG, H., ZHAO, T. et LI, S. (2006). Construc trilingual parallel corpus on demand. *Chinese Spoken Language Processing, Lecture Notes in Computer Science*, 4274:760–767.

Capacités d'apprentissage phonétique chez les sujets âgés francophones

Delvaux Véronique[1] Cano-Chervel Julie[2] Huet Kathy[1] Piccaluga Myriam[1] Harmegnies Bernard[1]
(1) Laboratoire de Phonétique, Institut de Recherche en Sciences et Technologies du Langage, UMONS, Belgique
(2) Hôpital Universitaire Des Enfants Reine Fabiola, Bruxelles
delvaux@umons.ac.be

RESUME

Cette étude investigue les capacités d'apprentissage phonétique chez des personnes âgées francophones, afin de fournir la ligne de base nécessaire à des comparaisons ultérieures avec des sujets âgés atteints de troubles de la parole. 9 hommes et 9 femmes (55-80 ans) dont la langue maternelle est le français ont été soumis à diverses tâches de production (répétition « la plus fidèle possible ») et de perception (discrimination AX), de pseudo-mots CV dont le VOT varie entre 20ms (typique du français) et 100 ms (VOT extra-long). En moyenne, les performances en discrimination sont modérées mais s'améliorent au fil du temps, avec une importante variabilité inter-individuelle. L'analyse statistique des mesures de VOT montre que les VOT des reproductions sont positivement corrélés aux VOT des stimuli, et que les performances sont significativement meilleures après vs. avant la tâche de reproduction. L'interaction significative entre les variables "Sujet" et "Tâche" est discutée ici, notamment dans le cadre plus général du débat sur la flexibilité à l'âge adulte des mécanismes cognitifs impliqués dans la production et la perception de la parole.

ABSTRACT

Phonetic learning abilities in ageing francophone speakers

This study investigates phonetic learning abilities in ageing francophone speakers, in order to provide a baseline for further comparisons with ageing speakers affected by speech disorders. 9 male and 9 female Belgian French speakers aged 55-80 were submitted to a number of production ("as faithfull as possible" reproduction) and perception (AX discrimination) tasks involving CV pseudo-words of 20-ms (French typical) to 100-ms (extra-long) VOT. Overall, discrimination performances are moderate, exhibiting large inter-individual variability and general improvement over time. The statistical analysis of VOT measures reveals that overall, response VOT positively correlates with stimulus VOT and that performances are significantly better after vs. before the reproduction task. The significant interaction between Speaker and Task is discussed here, particularly in the context of the debate on the flexibility in adult life of the cognitive mechanisms involved in speech production and perception.

MOTS-CLES : Apprentissage phonétique, vieillissement cognitif, VOT, lien production-perception.
KEYWORDS: phonetic learning, cognitive ageing, VOT, production-perception link.

1 Introduction

Cette étude s'inscrit dans un projet de recherche de phonétique clinique visant à investiguer les capacités d'apprentissage phonétique chez les patients francophones ayant subi un accident vasculaire cérébral et/ou atteints de maladies dégénératives. L'objectif général du projet est d'utiliser le paradigme d'apprentissage phonétique afin de compléter le diagnostic sur ces pathologies, et ainsi de contribuer, à terme, au développement d'instruments de mesure de

qualité basés sur la parole et applicables au sujet francophone pathologique. Même si les pathologies visées ne sont pas directement liées au vieillissement, dans les faits, leur prévalence chez les personnes âgées est nettement supérieure à celle de la population générale.

L'étude présentée ici investigue les capacités de la population âgée 'tout-venant' à maîtriser des variantes phonétiques non familières en langue maternelle (ici: une occlusive initiale non voisée avec un VOT long pour des sujets francophones), afin de fournir la référence nécessaire à des comparaisons ultérieures avec des sujets pathologiques. Cette étude s'inscrit donc dans le cadre plus large du débat sur la flexibilité, à l'âge adulte, des mécanismes cognitifs liés à la production et à la perception de la parole, et sur leur déclin potentiel lors du vieillissement.

En langue maternelle, on s'accorde généralement sur le fait que pour des tâches "simples" (qui incluent des syllabes, des pseudo-mots ou des mots présentés un à la fois, à une intensité typique ou supérieure à celle d'une conversation quotidienne), les sujets âgés ne présentent pas de déficit de compréhension de la parole par rapport à une population plus jeune lorsque l'on contrôle les capacités auditives périphériques (Gordon-Salant S., 2005; Humes, 2008). Par contre, dans des conditions d'écoute dégradées (adjonction de bruit, parole accélérée, écoute dichotique), ou lorsque la tâche est plus complexe (p.ex. identification de mots dans des phrases), un effet de l'avancement en âge est observé, en interaction ou non avec les capacités auditives périphériques (Gordon-Salant, 2005; George et al., 2007). L'effet de l'âge est attribué à une dégradation de mécanismes dit "de haut niveau", liés au système auditif central et/ou à des fonctions cognitives globales telles que la mémoire de travail et l'attention sélective (Salthouse, 2010; Humes et al., 2012), p.ex. la capacité à intégrer de l'information partielle de provenances diverses (Krull et al., 2013). Dans le cas du VOT, Gordon-Salant et collègues (2008) ont montré que les auditeurs anglophones âgés ont besoin d'un VOT plus long que leurs homologues jeunes pour identifier comme ***pie*** un mot ambigu sur un continuum ***buy/pie*** en contexte de phrase, mais pas en isolation.

En ce qui concerne la production de la parole en langue maternelle, on a constaté de nombreuses modifications chez les sujets âgés par rapport à la population générale: altérations de la voix (modifications de la F0, instabilités de la F0 et/ou de l'amplitude résultant en une augmentation des tremblements, du souffle dans la voix, de l'enrouement) (Linville, 2001); altérations des formants vocaliques (diminution généralisée liée à des changements physiologiques structurels, voire centralisation due à une réduction de la précision articulatoire); altérations prosodiques, notamment un ralentissement général de la vitesse d'articulation et du débit de parole (Zelnner Keller, 2006); variabilité plus importante dans la cinématique articulatoire traduisant une altération du contrôle moteur (Kent & Rosen, 2007); erreurs phonologiques (MacKay & James, 2004); etc..

Malgré le nombre considérable d'études rapportant les résultats d'un entraînement phonétique en perception (voire, plus rarement, en production) de la parole dans le cadre de l'apprentissage d'une langue seconde, il n'existe à notre connaissance que très peu d'études investiguant les performances de sujets âgés à de telles tâches (Tajima et al., 2002; Kubo et al., 2012). Les quelques résultats disponibles suggèrent que des sujets adultes ayant entre 60 et 70 ans sont capables d'apprendre de nouvelles catégories phonétiques, même si leurs performances sont inférieures à celles d'adultes plus jeunes, peut-être parce qu'ils recourent davantage aux stratégies efficaces en langue maternelle (Kubo et al., 2012).

Dans l'étude présentée ici, nous utilisons un paradigme adapté de précédents travaux visant initialement à entraîner de jeunes adultes francophones afin qu'ils acquièrent une nouvelle variante phonétique, non familière en langue maternelle, à savoir une occlusive initiale non

voisée avec un VOT long [t^{h}], typique de l'anglais (Piccaluga et al., 2011; Delvaux et al., 2013). Le paradigme inclut des tâches de perception et de production de la parole (détaillées ci-dessous), dont les modalités de réponse ont été légèrement adaptées afin d'être accessibles aux caractéristiques culturelles du public âgé (p.ex. la réponse "même" ou "différent" à la tâche de discrimination est simplement énoncée oralement et non pas encodée directement par le participant via une interface logicielle).

2 Matériel et méthode

2.1 Stimuli

Les stimuli sont 5 pseudo-mots de forme $C_{[t]}V_{[a]}$, où le VOT est respectivement de 20, 40, 60, 80 et 100ms (toutes les autres propriétés étant strictement invariantes). Ces 5 stimuli ont été construits à partir d'un pseudo-mot [t^{h}a] produit en parole naturelle par une locutrice anglophone, qui se composait d'un burst de 20ms, suivi d'une "aspiration" de 20ms (soit d'un VOT, défini comme le délai entre l'attaque du burst et le début de l'oscillation glottique, de 40ms), et enfin d'une voyelle de 210ms. De cette production naturelle, le burst et la voyelle ont été conservés dans tous les stimuli. Seule a été manipulée la durée de l'aspiration, qui varie selon les stimuli entre 0 et 80ms, par pas de 20ms.

2.2 Paradigme expérimental

Le paradigme expérimental a été administré en deux jours consécutifs, à heure fixe. Il est constitué des étapes suivantes:

Le premier jour: 1° Questionnaire de biographie linguistique; 2° Production de la "cible" de l'apprentissage: répétition 'la plus fidèle possible' du stimulus VOT60 (5 répétitions); 3° Perception: discrimination AX de 40 paires de stimuli (Intervalle inter-stimuli (ISI): 1000ms) (VOT20-20; VOT40-40; VOT60-60; VOT80-80; VOT100-100; VOT20-60; VOT60-20; VOT 40-80; VOT80-40; VOT60-100; VOT 100-60): 5 blocs où 8 paires de stimuli sont présentées en ordre aléatoire, soient les 6 paires de stimuli différents ainsi que 2 paires de stimuli identiques); 4° Production: répétition du stimulus VOT60 (5 répétitions) (idem qu'à l'étape 2°); 5° (Re)production: répétition 'la plus fidèle possible' des stimuli présentés un par un, soit en ordre 'montant': VOT20, 40, 60, 80, 100, soit en ordre aléatoire (4 blocs en ordre montant; pause; 4 blocs en ordre aléatoire); afin de limiter la fatigue cognitive, une pause est introduite tous les 10 stimuli (tâche distractrice brève: commentaire libre d'images); 6° Production: répétition du stimulus VOT60 (5 répétitions) (idem qu'à l'étape 2°).

Le deuxième jour: 7° Anamnèse médicale; 8° 'Calibration': production des mots du français (via la dénomination d'images): ***pas, tas, k*** (5 répétitions; présentation en ordre aléatoire); 9° Audiométrie tonale en conduction aérienne.

Ce paradigme expérimental permet de récolter de l'information quant aux antécédents linguistiques (1°) et médicaux (7°) des participants, ainsi que de documenter leurs capacités auditives (9°). Il permet également d'évaluer la capacité des participants à produire des VOT typiques de leur L1 (8°), à produire des VOT typiques d'une L2 avant expérimentation (2°), à discriminer entre des stimuli dont les valeurs de VOT sont, dans leur majorité, non familières en L1 (3°), à produire des VOT typiques d'une L2 après 'entraînement' perceptuel uniquement (4°), à reproduire des VOT non familiers juste après écoute du modèle, en comparant les performances selon que l'ordre de présentation favorise ou non un effet cumulatif (ordre montant vs. aléatoire) (5°), enfin à produire des VOT typiques d'une L2 après un

'entraînement' en perception et en production (6°). La phase 8° a été planifiée le deuxième jour car il a été remarqué au cours d'une étude pilote antérieure que le paradigme de dénomination lexicale en L1 pouvait avoir un effet perturbateur sur la tâche de discrimination (Delvaux et al., 2014).

2.3 Participants

Les participants sont 18 francophones (ayant le français pour seule langue maternelle), 9 hommes et 9 femmes, âgés de 55 à 80 ans au moment de l'expérience.

	M/F	Age	Origine	Expertise musicale	Classification audiométrique	Perte Auditive
F3	F	80	Bruxelles	non	Surdité légère	30,80%
F6	F	78	Paris	non	Audition normale	5,90%
F7	F	72	Mons	non	Surdité moyenne 1er degré	33,30%
F8	F	55	Aix-en-Provence	non	Audition normale	0,45%
F9	F	65	Aix-en-Provence	oui	Audition normale	1,00%
F10	F	68	Mons	oui	Surdité légère	13,70%
F11	F	69	Mons	oui	Audition normale	1,00%
F12	F	65	Paris	oui	Audition normale	2,60%
F13	F	65	Mons	oui	Audition normale	1,20%
H1	M	75	Mons	oui	Surdité légère	32,90%
H2	M	64	Luxembourg	non	Audition normale	0,80%
H3	M	80	Paris	oui	Surdité légère	10,10%
H4	M	69	Aix-en-Provence	oui	Surdité moyenne 1er degré	34,00%
H5	M	62	Aix-en-Provence	oui	Audition normale	5,00%
H6	M	62	Mons	oui	Audition normale	0,50%
H7	M	68	Paris	non	Surdité légère	15,50%
H8	M	73	Mons	non	Audition normale	11,00%
H9	M	60	Mons	non	Audition normale	3,90%

TABLEAU 1. Résumé des caractéristiques des 18 sujets.

Les critères d'exclusion consistaient en tout antécédent rapporté de trouble du langage, ou plus généralement d'atteinte neurologique, de trauma crânien ou de maladie dégénérative. Quatre sujets ayant initialement pris part à l'expérimentation ont également été exclus sur la base des résultats obtenus à l'audiométrie tonale aérienne (presbyacousie non diagnostiquée: perte auditive moyenne supérieure à 35%). Nous avons conservé dans l'étude 5 sujets qui, selon la classification audiométrique, sont atteints de surdité légère, ainsi que 2 autres qui sont atteints de surdité moyenne du 1er degré, afin de tester l'hypothèse selon laquelle une déficience auditive légère à modérée peut influencer négativement les performances aux tâches de perception et de production. Le tableau 1 résume les principales caractéristiques des 18 sujets (genre, âge, origine géographique, expertise musicale, perte auditive mesurée et classification audiométrique associée: BIAP, 1997).

2.4 Mesures

Afin d'évaluer les performances des sujets en production de la parole (étapes 2°, 4°, 5°, 6°, 8° du paradigme), le VOT a été mesuré manuellement (en ms) sur l'oscillogramme entre le début du burst et le début du voisement, défini comme le passage par zéro précédant le premier cycle glottique. Les performances en perception (étape 3° du paradigme) ont été évaluées via

le calcul de la proportion de réponses correctes (tant 'même' que 'différent') au test de discrimination.

3 Résultats

3.1 Discrimination

Les figures 1 et 2 illustrent les résultats obtenus lors de la tâche de discrimination AX, tous sujets confondus. La figure 1 donne le score moyen (proportion moyenne de réponses exactes: "même" quand les 2 stimuli de la paire sont identiques, "différent" lorsqu'ils sont différents) obtenu en fonction du temps (40 paires présentées). Elle rend compte du fait que, lorsque l'on considère les 18 sujets comme un seul groupe, les performances sont globalement modérées, mais qu'elles s'améliorent au cours du temps (variabilité de paire à paire plus faible, score moyen plus élevé).

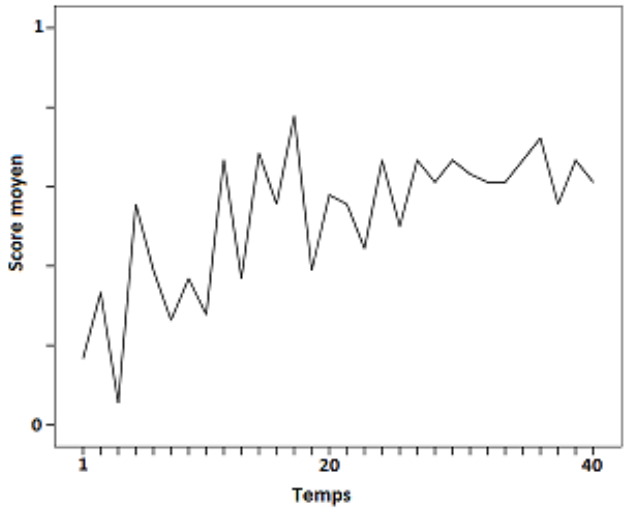


FIGURE 1. Discrimination. Evolution de la proportion de réponses correctes au cours du temps (18 sujets confondus).

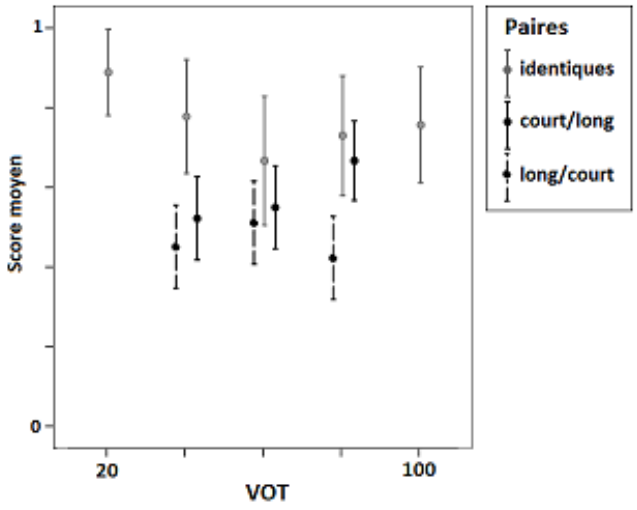


FIGURE 2. Discrimination. Proportion de réponses correctes en fonction du type de paire (18 sujets confondus).

La figure 2 détaille les scores moyens obtenus en fonction du type de paire à discriminer et du VOT (moyen) des stimuli concernés, en affichant les intervalles de confiance (95%) calculés autour de la moyenne générale (tous sujets confondus). On constate que les performances sont globalement meilleures pour les paires de stimuli identiques, et ce d'autant plus que leur VOT est très faible (paires VOT20-VOT20) ou très élevé (paires VOT100-VOT100). Les performances sont plus modérées pour les paires de stimuli différents (et avoisinent en moyenne le niveau du hasard: .5), ce qui indique que la détection de différences de VOT de 40ms était globalement une tâche difficile pour les 18 participants âgés. La présentation d'un stimulus court avant un stimulus plus long peut favoriser la détection (par rapport à une présentation long/court), en particulier dans le cas de la paire VOT60-VOT100.

Au-delà des performances globales du groupe de participants, il est utile de s'intéresser aux performances individuelles, en lien avec les caractéristiques des individus concernés. La figure 3 donne le score moyen obtenu par chacun des 18 participants, calculé ici uniquement à partir des réponses données aux stimuli différents afin d'éviter de surestimer les performances des sujets qui ont répondu "même" à toutes les paires présentées ou presque. Le calcul du ***d'***, qui permet de neutraliser ce type d'effet, n'est pas approprié ici étant donné le nombre réduit de réponses pour chaque paire (5). À l'examen de la figure 3, on constate une importante

variabilité inter-individuelle dans les performances en discrimination.

Une analyse de variance (UNIANOVA) a été réalisée avec comme variable dépendante le score obtenu (aux paires de stimuli différents) et comme variables indépendantes le genre du sujet (9 hommes; 9 femmes), le groupe d'âge auquel il appartient (55-64 ans: 5 sujets; 65-74 ans: 9 sujets; 75 ans et +: 4 sujets) ainsi que son niveau d'audition (audition normale: 11 sujets; surdité légère: 5 sujets; surdité moyenne du 1er degré: 2 sujets). En ce qui concerne les effets principaux[1], l'analyse révèle qu'il n'y a pas d'effet significatif du genre ($F_{1,535}=0,05$; $p=0,822$) sur la performance, alors qu'il y a un effet du groupe d'âge ($F_{2,535}=12,931$; $p<0,001$) et du niveau d'audition ($F_{2,535}=42,305$; $p<0,001$). Des tests Post Hoc (Tukey) révèlent par ailleurs que les performances sont significativement meilleures pour les sujets âgés de 55 à 64 ans, mais qu'elles ne diffèrent pas entre les 2 autres groupes d'âge, et que les performances sont significativement plus médiocres pour les sujets présentant une surdité moyenne du 1er degré (F7,H4, voir Figure 3), alors qu'elles sont équivalentes pour les sujets ayant une audition normale ou une surdité légère.

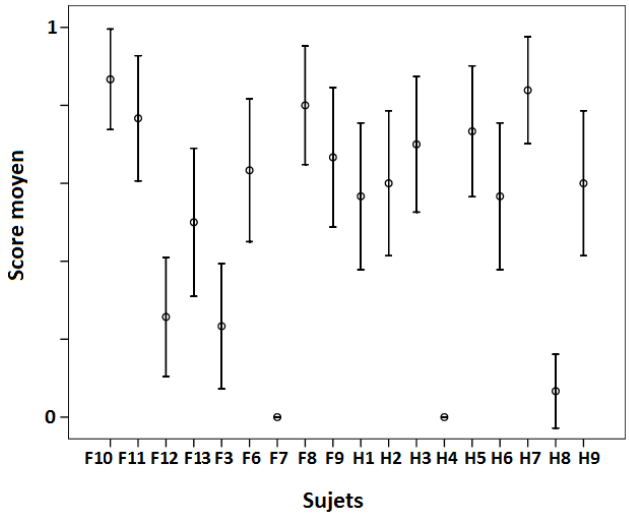


FIGURE 3. Discrimination. Score moyen (proportion de réponses correctes aux paires de stimuli différents) en fonction du sujet

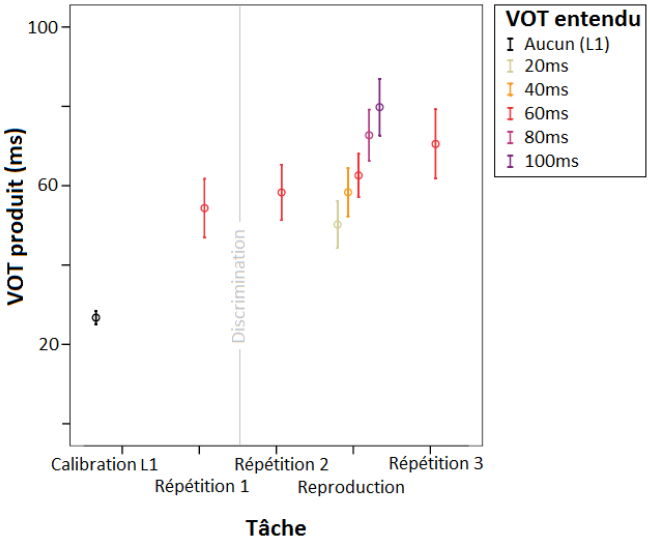


FIGURE 4. Production. VOT produit (ms) en fonction de la tâche et du VOT entendu (tous sujets confondus)

3.2 Production

Si l'on examine les performances globales (tous sujets confondus) aux diverses tâches de production de la parole impliquées dans le paradigme expérimental, il apparait que les 18 sujets âgés sont parvenus à accomplir les tâches demandées avec un certain succès. La figure 4 donne les intervalles de confiance (95%) calculés autour de la moyenne générale des VOT produits en fonction de la tâche demandée (18 sujets). Elle illustre le fait que les VOT produits en réponse aux stimuli proposés sont nettement plus élevés que les VOT mesurés lors de la tâche de calibration (8°), qui avait pour objectif d'établir la ligne de base des productions en langue maternelle (en l'occurence, la moyenne des VOT produits lors de la tâche de dénomination en français est de 28ms). Les moyennes des VOT produits lors de la tâche principale de (re)production (4°) augmentent à mesure que le VOT entendu augmente, même si l'on constate une réduction de l'empan des (moyennes des) VOT produits par rapport aux VOT entendus (moyennes s'étalant de 41ms à 77ms vs. 20ms à 100ms).

Une analyse de variance UNIANOVA a été effectuée avec pour variables indépendantes la

[1] L'analyse révèle également des effets d'interaction que nous n'analysons pas ici étant donné le déséquilibre des effectifs, les sous-groupes de participant ayant été constitués a posteriori.

tâche (5 niveaux: calibration; répétition 1 (2°); répétition 2 (4°); reproduction (5°); répétition 3 (6°)) et le sujet (18 niveaux), et pour variable dépendante le VOT produit (dans le cas de la tâche de reproduction on ne considère ici que les VOT produits en réponse aux stimuli VOT60). L'analyse statistique indique que la tâche, le sujet, ainsi que l'interaction entre la tâche et le sujet induisent des variations significatives de VOT produit. De plus, les tests Post Hoc (Tukey) sur la variable "tâche" révèle l'existence de 4 sous-ensembles: (1) calibration (Moyenne: 27ms); (2) répétition 1 (54ms) et répétition 2 (58ms); (3) répétition 2 (58ms) et production (63ms); (4) répétition 3 (70ms). On peut en conclure que: (i) les VOT produits en réponse aux stimuli VOT60 sont significativement différents des VOT produits en langue maternelle, et ce dès avant l'entraînement; (ii) qu'il n'y a pas de différence significative dans les performances en production après l'entraînement perceptuel via la tâche de discrimination (survenant entre la répétition 1 et la répétition 2); (iii) que lors de l'entraînement en production les VOT produits en réponse aux VOT60 sont significativement plus longs qu'au début de l'expérimentation, avant l'entraînement perceptuel; (iv) qu'en fin d'expérimentation, après entraînement en perception et en production, les VOT produits en réponse aux VOT60 sont significativement plus longs que lors de toutes les étapes antérieures.

L'interaction significative entre les variables "sujet" et "tâche" est illustrée dans la figure 5 pour les 9 sujets masculins. Comme on le constate à l'examen de la figure, les profils de réponse sont très variés d'un individu à l'autre.

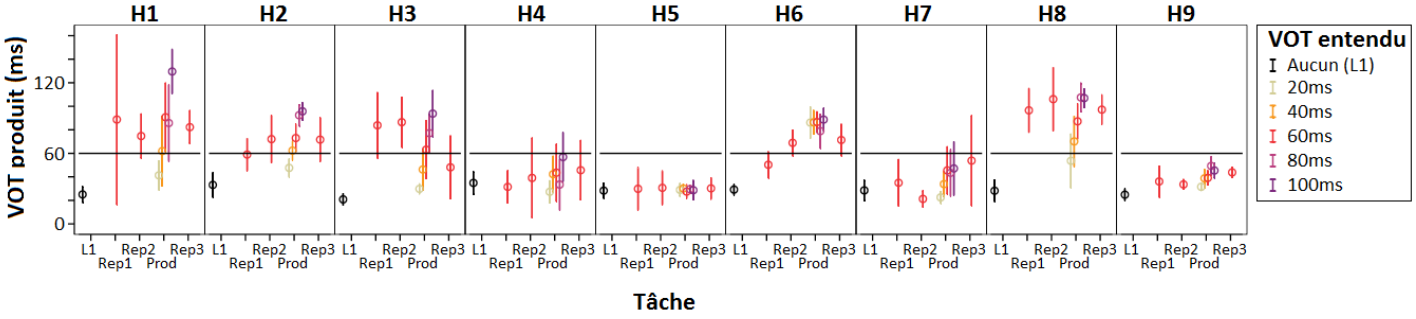


FIGURE 5. Production. VOT produit (ms) en fonction de la tâche (L1: calibration (1°); Rep1: répétition 1 (2°); Rep2: répétition 2 (4°); Prod: reproduction (5°); Rep3: répétition 3 (6°)) et du VOT entendu pour les 9 sujets masculins.

Une analyse statistique complémentaire (UNIANOVA) a été réalisée avec comme variable dépendante le VOT produit (en réponse aux VOT60) et commes variables indépendantes la tâche, le genre du sujet, son niveau d'audition et son groupe d'âge. L'analyse a révélé un grand nombre d'effets significatifs des principaux facteurs, seuls (tâche: $F_{4,627}=47,059$; $p<0,001$; genre: $F_{1,627}=4,049$; $p<0,05$; niveau d'audition: $F_{2,627}=62,402$; $p<0,001$; groupe d'âge: $F_{2,627}=14,283$; $p<0,001$) et en interaction. Néanmoins, les données sont difficiles à interpréter, dans la mesure où les performances (VOT produit par rapport au VOT entendu de 60ms) n'évoluent pas de façon ordinale par rapport aux différents niveaux des variables indépendantes. Ainsi, les tests Post Hoc (Tukey) réalisés montrent que le VOT est significativement plus court pour les 5 sujets atteints de surdité légère, que pour le groupe formé tant par les 11 sujets présentant une audition normale que par les 2 sujets atteints de surdité moyenne. De même, le VOT est significativement plus élevé chez les 65-74 ans qu'il ne l'est chez les plus jeunes (55-64ans) ou chez les plus âgés (75 ans et +).

Etant donné l'effectif réduit dans chacun des sous-groupes ici considérés, chaque individu représente une combinaison particulière des différents niveaux de variables telles que l'âge, le genre ou le niveau d'audition. Nous proposons dès lors ci-dessous une analyse exploratoire des

performances en production des 9 sujets masculins en lien avec les caractéristiques des individus et nos observations quant à leurs performances en discrimination.

Tout d'abord, un seul sujet (H5) ne manifeste aucune sensibilité, dans ses productions, aux variations de VOT dans les stimuli. H9 reste également cantonné à de faibles valeurs de VOT, mais semble néanmoins bénéficier de l'entraînement en production. Il est à noter que H5 est dans le tiers supérieur en ce qui concerne les performances en discrimination. Sa capacité à détecter des différences de VOT ne se traduit manifestement pas dans une aptitude à produire ensuite des VOT longs. Inversement, H2 est celui qui s'acquitte de la tâche de reproduction avec le plus de succès alors que ses performances préalables en discrimination sont très modérées. De façon générale, plusieurs sujets masculins présentent une tendance à "l'overshoot", c'est-à-dire qu'ils produisent des VOT beaucoup plus longs qu'attendus (H1, H6, H8). On peut considérer qu'ils ont dès lors détecté la nature du changement attendu par rapport à leur langue maternelle (allonger le VOT), mais qu'ils ne réussissent pas, soit à détecter les différents niveaux de VOT long, soit à ajuster finement leurs réalisations motrices à ces différentes cibles. H3, quant à lui, corrige sa tendance à l'overshoot (répétition 1, répétition 2) après un entraînement très réussi en production. Enfin, la figure 5 suggère que H4 (pourtant atteint de surdité moyenne et ne démontrant aucune aptitude à la discrimination) et H7 sont dans une certaine mesure sensibles aux variations de VOT des stimuli en production, et qu'un entraînement plus long leur aurait peut-être été profitable.

4 Conclusion

L'étude présentée ici investigue les capacités de sujets francophones âgés à produire et à percevoir l'occlusive initiale non voisée à VOT long [t^h]. Les performances lors de la tâche de discrimination sont globalement modérées mais s'améliorent au cours du temps, ce qui suggère qu'une expérimentation plus longue, avec un plus grand nombre de paires, aurait permis d'améliorer le niveau général des performances. Les résultats indiquent un effet lié au vieillissement (qui se manifeste principalement après 65 ans), ce qui confirme la nécessité d'établir des normes spécifiques pour les sujets âgés. Par ailleurs, le seuil à viser, si l'on veut éviter un effet parasite des capacités auditives sur les performances en discrimination, se situe à la frontière entre surdité légère et surdité moyenne du 1er degré.

Les résultats obtenus en production de la parole confirment que la tâche préalable de discrimination n'a pas permis, à elle seule, d'améliorer significativement la production de [t] à VOT long. Par contre, l'entraînement en production s'est avéré efficace pour une majorité de locuteurs. Malgré de grandes différences inter-individuelles, les sujets âgés francophones sont globalement efficients lorsqu'on leur demande de reproduire des stimuli dont les VOT varient entre 20 et 100ms. Pendant cet entraînement, et plus encore, immédiatement après, les participants produisent des VOT longs, significativement différents de ceux qu'ils réalisent dans leur langue maternelle, le français. Bien que le vieillissement induise de nombreuses modifications structurelles et fonctionnelles des différents sous-systèmes, périphériques et centraux, impliqués dans la production et la perception de la parole, l'étude présentée ici confirme dès lors que les sujets âgés sont capables d'apprentissage phonétique (Kubo et al., 2012). Une analyse exploratoire de nos données démontre l'utilité d'investiguer plus avant, au cours d'études ultérieures, les liens entre performances individuelles en discrimination et en production, ainsi que les effets sur la production de facteurs tels que le groupe d'âge et le niveau d'audition.

Cette recherche a été partiellement subventionnée par l'Action de Recherche Concertée (AUWB-2012- 12/17-UMONS- N°1).

Références

Bureau International d'Audiophonologie (Recommandations) 1997. Consulté le 09/02/2014. http://www.biap.org/index.php?option=com_content&view=article&id=5%3Arecommandation-biap-021-bis&catid=65%3Act-2-classification-des-surdites&Itemid=19&lang=fr.

Delvaux, V., Huet, K., Piccaluga, M., Harmegnies, B. (2013). Production training in Second Language Acquisition: A comparison between objective measures and subjective judgments. In ***Interspeech 2013***, 2375-2379, Lyon, France.

Delvaux, V., Huet, K., Piccaluga, M., Harmegnies, B. (2014). Capacité d'apprentissage phonique et troubles du langage à étiologie cérébrale. In Sock R., Vaxelaire B., Fauth C., (Eds). ***Travaux en phonétique clinique,*** Collection RPA, Mons: CIPA, 259-274.

George, E. L., Zekveld, A. A., Kramer, S. E., Goverts, S.T., Festen, J.M., Houtgast, T. (2007). Auditory and nonauditory factors affecting speech reception in noise by older listeners. ***Journal of the Acoustical Society of America***, 121: 2362–2375.

Gordon-Salant, S. (2005). Hearing loss and aging: New research findings and clinical implications. ***Journal of Rehabilitation Research and Development,*** 42: 9–24.

Gordon-Salant, S., Yeni-Komshian, G., Fitzgibbons, P. (2008). The role of temporal cues in word identification by younger and older adults: Effects of sentence context. ***Journal of the Acoustical Society of America***, 124(5): 3249–3260.

Humes, L. E. (2008). Aging and Speech Communication : Peripheral, Central-Auditory, and Cognitive Factors Affecting the Speech-Understanding Problems of Older Adults. ***The ASHA Leader***, 2008, April 15.

Humes, L. E., Dubno, J. R., Gordon-Salant, S., Lister, J.J., Cacace, A.T., Cruickshanks, K.J., Gates, G.A., Wilson, R.H., Wingfield A. (2012). Central presbycusis: A review and evaluation of the evidence. ***Journal of the American Academy of Audiology***, 23: 635–666.

Kent, R.D., Rosen, K. (2007). Motor control perspectives on motor speech disorders. In Maassen B., Kent R., Peters H. (Eds). ***Speech Motor Control: In Normal and Disordered Speech***, New York, NY: Oxford University Press, 285-312.

Krull, V., Humes, L.E., Kidd, G.R. (2013). Reconstructing Wholes From Parts: Effects of Modality, Age, and Hearing Loss on Word Recognition. ***Ear and Hearing***, 34(2): 14-23.

Kubo, R., Akahane-Yamada, R., Akagi, M. (2012). Effects of perceptual training on the ability of elderly adults of Japanese speakers to identify American English /r/-/l/ phonetic contrasts. ***Journal of the Acoustical Society of America***, 131(4): 3272.

Linville, S.E. (2001). ***Vocal Ageing***. San Diego: Singular Publishing Group.

MacKay, D.G., James, L.E. (2004). Sequencing, Speech Production, and Selective Effects of Aging on Phonological and Morphological Speech Errors. ***Psychology and Aging***, 19(1): 93–107.

Piccaluga, M., Clairet, S., Delvaux, V., Huet, K., Harmegnies, B. (2011). Guidage perceptuel de la production en L2 : Tendances générales et variabilité individuelle. ***Recherches Anglaises et Nord-Américaines***, Hors-série, 25-44.

Salthouse, T.A. (2010). ***Major Issues in Cognitive Aging (1st ed.).*** NY: Oxford Univ. Press.

Tajima, K., Akahane-Yamada, R., Yamada, T. (2002). Perceptual learning of English syllable rhythm by young and elderly Japanese listeners. In ***Proceedings of the 16th general meeting of the Phonetic Society of Japan***, 103-108, Tokyo, Japan.

Zellner Keller, B. (2006). Aging and Speech Prosody. In Rüdiger, H., Hansjörg, M. (Eds.) ***Speech Prosody 2006, Studientexte Zur Sprachkommunikation,*** 696-701.

Régression ordinale pour la prédiction de la qualité d'interaction

Layla El Asri[1, 2] Hatim Khouzaimi[1, 3] Romain Laroche[1] Olivier Pietquin[4]
(1) Orange Labs Issy-les-Moulineaux, France
(2) UMI 2958 GeorgiaTech-CNRS, France
(3) Laboratoire d'Informatique d'Avignon-CERI, France
(4) Université Lille 1, LIFL (UMR 8022 CNRS/Lille 1) - équipe SequeL, France
prénom.nom@orange.com olivier.pietquin@lifl.fr

RÉSUMÉ

La prédiction automatique de la qualité d'un dialogue permet d'assurer le suivi de la performance d'un système de dialogue et si nécessaire, adapter le comportement de celui-ci. Le modèle de prédiction est appris à partir d'un corpus de dialogues évalués par des utilisateurs ou des experts. Nous proposons dans cet article de modéliser cette tâche en tant que problème de régression ordinale. Nous appliquons les machines à vecteurs supports pour la régression ordinale sur un corpus de dialogues où un score allant de 1 à 5 mesurant la qualité d'interaction a été donné par des annotateurs experts après chaque échange système-utilisateur. Comparée aux modèles de classification et de régression précédemment proposés pour cette tâche, la régression ordinale produit des résultats significativement meilleurs selon les métriques suivantes : le coefficient κ de Cohen, le coefficient de corrélation de Spearman ainsi que les erreurs euclidienne et en valeur absolue.

ABSTRACT

Ordinal Regression For Interaction Quality Prediction

Automating the evaluation of the quality of a dialogue is useful to keep track of a spoken dialogue system's performance and, if necessary, adapt the behaviour of the system. Classifiers and regression models have been suggested to make this prediction. The parameters of these models are learnt from a corpus of dialogues evaluated by users or experts. In this paper, we propose to model this task as an ordinal regression problem. We apply support vector machines for ordinal regression on a corpus of dialogues where each system-user exchange was given a rate on a scale of 1 to 5 by experts. Compared to previous models proposed for this task, the ordinal regression predictor has significantly better results according to the following evaluation metrics : Cohen's agreement rate, Spearman's rank correlation coefficient, and the Euclidean and absolute value errors.

MOTS-CLÉS : Systèmes interactifs, apprentissage statistique, évaluation des performances.

KEYWORDS: Interactive Systems, statistical learning, performance evaluation.

1 Introduction

Un Système de Dialogue (SD) est une interface homme-machine associant différents modules : la reconnaissance vocale, l'analyse sémantique, la gestion du dialogue, la génération automatique de textes ainsi que la synthèse vocale. Pour être exhaustive, l'évaluation d'un tel système doit être double. Il faut en effet procéder à l'évaluation technique des différents modules mais aussi à une évaluation globale de l'utilisabilité du système (Dybkjaer *et al.*, 2004). Möller (Möller, 2000) définit l'utilisabilité comme étant la capacité d'un système ou d'un service à se conformer aux exigences des utilisateurs. Classiquement, deux types de campagnes d'évaluation de l'utilisabilité sont menés. Le premier consiste à demander à l'utilisateur de donner un score et/ou de remplir un questionnaire après chaque interaction (Walker *et al.*, 1997). Le second type fait appel à des experts (Evanini *et al.*, 2008; Schmitt *et al.*, 2011). Procéder ainsi permet de s'affranchir des biais de subjectivité liés à certains aspects indépendants de la gestion du dialogue comme la voix de la synthèse vocale (Möller, 2000; Schmitt *et al.*, 2011). Cela implique toutefois que les experts soient formés afin d'éviter une trop grande variabilité dans les annotations. Dans tout ce qui suit, nous désignerons les évaluations utilisateurs par Satisfaction Utilisateur (SU) et les scores experts par Qualité d'Interaction (QI) (Schmitt *et al.*, 2011).

Afin de limiter les coûts engagés pour ces campagnes d'évaluation, il a été proposé d'utiliser les corpus collectés afin d'apprendre un modèle de prédiction automatique de la performance du système. Les scores prédits peuvent alors servir à assurer le suivi en-ligne de l'efficacité du système et à apprendre un comportement optimal pour le système avec des techniques telles que l'apprentissage par renforcement (Levin *et al.*, 1997; Pietquin *et al.*, 2011; Daubigney *et al.*, 2011).

Ce problème de prédiction a été considéré tour à tour comme un problème de régression et de classification (Walker *et al.*, 1997; Engelbrecht *et al.*, 2009; Schmitt *et al.*, 2011). Cet article propose d'unifier ces deux vues en le modélisant comme un problème de régression ordinale. La régression ordinale renvoie des labels discrets tout en prenant en compte l'ordre entre les labels.

Nous comparons la Régression Ordinale avec Machines à Vecteurs Supports (SVOR) à différents modèles de régression et de classification sur le corpus LEGO (Schmitt *et al.*, 2012) qui contient 200 dialogues annotés collectés avec le système Let's Go de l'université Carnegie Mellon (Raux *et al.*, 2003). Des experts ont attribué un score compris entre 1 et 5 après chaque échange système-utilisateur. Ce score de QI évalue le déroulement du dialogue. Nous montrons sur ce corpus que les prédictions de SVOR sont significativement meilleures selon quatre métriques mesurant la distance et la corrélation.

La section 2 passe en revue les modèles proposés par le passé pour prédire les scores donnés par des utilisateurs ou des experts. La section 3 détaille les métriques utilisées pour comparer ces différents modèles et SVOR. Ensuite, la section 4 décrit le corpus LEGO, la section 5 détaille les modèles comparés et la section 6 présente les résultats du test des modèles.

2 Travaux précédents

L'évaluation automatique de la SU ou de la QI est utile pour le suivi des performances d'un SD et l'adaptation de la gestion du dialogue. Witt (Witt, 2011) a proposé d'effectuer cette adaptation

à l'aide d'une règle de calcul qui permet d'estimer la *Caller Experience Metric* (CEM, équation 1), un score allant de 1 à 5 et donné par un expert. Dans cette équation, les événements sont le rejet de la reconnaissance vocale, l'inactivité utilisateur, l'incompréhension de l'intention de l'utilisateur, la contradiction du système par l'utilisateur ou la demande de transfert. d est un facteur de réduction réduisant l'importance des événements les plus éloignés dans le temps.

$$\text{CEM}_t = d \times \text{CEM}_{t-1} + \text{ poids du dernier événement} \quad (1)$$

L'adaptation du comportement d'un SD peut aussi se faire en appliquant des méthodes d'apprentissage statistique telles que l'apprentissage par renforcement. Un agent apprenant par renforcement compare différentes stratégies sur la base d'une fonction de récompense. Dans le cas des SD, cette fonction doit évaluer la qualité de chaque dialogue, il est donc naturel d'inférer cette fonction à partir de scores SU ou QI (Walker *et al.*, 1998; El Asri *et al.*, 2012).

Enfin, une dernière possibilité est de demander à des experts de donner un score QI après chaque échange entre le système et l'utilisateur puis d'apprendre un modèle capable de prédire ce score au niveau du tour de dialogue. Le système peut alors adapter son comportement au cours du dialogue si jamais l'estimation de la QI se dégrade. Le SD peut choisir dans ce cas de jouer un message d'aide, de changer sa stratégie de dialogue ou en dernier recours de transférer l'appel à un opérateur humain.

La prédiction de la SU ou de la QI au niveau du dialogue ou du tour de dialogue s'est faite jusqu'ici à l'aide de modèles de régression et de classification. Walker *et. al* (Walker *et al.*, 1997) ont appliqué à ce problème une Régression Linéaire Multiple (MLR), modélisant la SU comme la réalisation de la tâche tout en minimisant les coûts du dialogue comme la durée ou le nombre de rejets de la reconnaissance vocale. Evanini *et. al* (Evanini *et al.*, 2008) ont construit un arbre de décision afin de prédire la QI. Engelbrecht *et. al* (Engelbrecht *et al.*, 2009) ont utilisé les modèles de Markov cachés pour représenter l'évolution de la SU au cours du dialogue. Hara *et. al* (Hara *et al.*, 2010), quant à eux, ont construit un modèle N-gram basé sur les actes de dialogues. Higashinaka *et. al* ont comparé les modèles de Markov cachés aux champs conditionnels aléatoires pour l'estimation de la QI au niveau du tour de dialogue. Pour le même problème, Schmitt *et. al* (Schmitt *et al.*, 2011) ont proposé les Machines à Vecteurs Supports (SVM).

La modélisation de la QI en tant que problème de classification pure ne permet pas de prendre en compte l'ordre naturel entre les labels. Nous proposons ici l'approche par SVOR pour prédire la QI au niveau du tour de dialogue (Shashua et Levin, 2002; Chu et Keerthi, 2007). La régression ordinale fait le pont entre la classification et la régression. Elle prend en entrée un ensemble d'exemples labellisés avec un ordre naturel entre les labels et construit un prédicteur qui minimise la distance en valeur absolue par rapport aux labels de référence tout en prenant en compte l'ordre entre ces labels.

3 Métriques

Pour les systèmes de dialogue apprenant par renforcement, il est important de classer correctement les dialogues. En effet, si l'on utilise les prédictions de QI au niveau du tour de dialogue comme récompenses pour le système, il est nécessaire que ces prédictions conservent l'ordre établi par les experts. Pour mesurer la performance des modèles sur cet aspect, nous utilisons le coefficient de corrélation ρ de Spearman (Spearman, 1904) qui mesure la corrélation entre

deux classements. Soit $y = \{y_1, ..., y_n\}$ et $\hat{y} = \{\hat{y}_1, ..., \hat{y}_n\}$ respectivement un vecteur de scores QI et les estimations de ces scores sur un ensemble de n échanges système-utilisateur. Soient $r(y) = \{r(y_1), ..., r(y_n)\}$ et $r(\hat{y}) = \{r(\hat{y}_1), ..., r(\hat{y}_n)\}$ les classements correspondant à ces vecteurs. Par exemple, avec $y = \{1, 15, 3, 12, 27\}$, $r(y) = \{1, 4, 2, 3, 5\}$. L'équation 2 rappelle le calcul de ρ.

$$\rho(y, \hat{y}) = \frac{\sum_i (r(y_i) - \text{moyenne}(r(y))) \sum_i (r(\hat{y}_i) - \text{moyenne}(r(\hat{y})))}{\sqrt{\sum_i (r(y_i) - \text{moyenne}(r(y)))^2} \sqrt{\sum_i (r(\hat{y}_i) - \text{moyenne}(r(\hat{y})))^2}} \quad (2)$$

Nous comparons aussi les modèles sur la base des erreurs euclidienne et en valeur absolue, rappelées dans les équations 3 et 4.

$$\text{erreur_euclidienne}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (3)$$

$$\text{erreur_en_valeur_absolue}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (4)$$

Le taux d'accord avec les experts est mesuré à l'aide du coefficient κ de Cohen (Cohen, 1960). Ce coefficient calcule la probabilité d'accord entre y et $\hat{y}$. Dans l'équation 5, $P(a)$ est la probabilité d'accord et $P(ca)$ la probabilité d'accord aléatoire.

$$\kappa(y, \hat{y}) = \frac{P(a) - P(ca)}{1 - P(ca)} \quad (5)$$

4 Le corpus LEGO

Le système Let's Go fournit des informations sur les horaires des bus de la ville de Pittsburgh (Raux *et al.*, 2003). Schmitt *et. al* (Schmitt *et al.*, 2012) ont fait annoter 200 dialogues avec Let's Go par trois experts. Ceux-ci devaient donner un score de QI compris entre 1 et 5 après chaque échange entre le système et l'utilisateur. Au total, 5282 échanges ont été évalués. Comme Schmitt *et. al* (Schmitt *et al.*, 2011), nous utiliserons pour notre test la médiane des trois scores donnés par les experts.

Dans le corpus LEGO, chaque échange système-utilisateur est décrit par un ensemble de variables caractéristiques obtenues automatiquement ou manuellement. Ces variables dépendent de la reconnaissance vocale (score de confiance, rejets...), de l'analyse sémantique (acte de dialogue utilisateur, *parsing* sémantique,...) et de la gestion du dialogue (acte de dialogue système). De même que Schmitt *et. al* (Schmitt *et al.*, 2011), nous utilisons les variables quantitatives sur trois niveaux : la valeur pour le dernier échange système-utilisateur, la valeur moyenne sur les trois derniers échanges ainsi que la valeur moyenne depuis le début du dialogue. La liste complète des variables du corpus se trouve dans (Schmitt *et al.*, 2012). Nous n'avons conservé que les variables caractéristiques obtenues automatiquement afin de traiter le cas où l'on souhaite pouvoir suivre les performances du SD en-ligne sans avoir recours à de nouvelles annotations. Parmi les variables retenues, deux sont qualitatives : l'Acte de Dialogue Utilisateur

(ADU) et l'Acte de Dialogue Système (ADS). Par exemple, l'ADU peut prendre des valeurs telles que ADU_CONFIRMER_DEPART, ADU_INFORMATION_LIGNE, *etc*. Dans ce cas, en notant n_a le nombre de labels possibles, nous avons séparé la variable en $n_a - 1$ variables booléennes. ADS a été séparé en 27 variables et ADU en 21. En tout, chaque échange système-utilisateur est constitué de 67 variables, toutes centrées et normalisées. La généralisabilité des méthodes décrites dans la section suivante a été testée par une validation croisée *10-fold*.

5 Modèles

5.1 Régression linéaire multiple et régression multivariée par spline adaptive

Les prédicteurs ont été sélectionnés un à un. Nous avons calculé le coefficient de régression simple pour chaque variable puis nous avons gardé la variable ϕ_i de plus haut coefficient en valeur absolue β_i [1]. Nous avons alors calculé les coefficients de régression simple pour chaque variable restante avec $QI - \beta_i \phi_i$ comme fonction objectif. Nous avons répété ce procédé jusqu'à ce que l'erreur soit minimisée.

La Régression Multivariée par Spline Adaptive (MARS) (Friedman, 1991) revient à modéliser la QI comme fonction linéaire de fonctions charnières de la forme : $B_i(x) = max(0, x-s)$ or $B_i(x) = max(0, s-x)$ où s est une instance du corpus. Nous avons utilisé l'outil Matlab ARES [2]. Cet algorithme requiert deux paramètres : *maxFuncs*, le nombre maximum de fonctions charnières ainsi que *cubic*, un booléen indiquant si l'utilisation de splines cubiques pour lisser les points où la fonction du modèle n'est pas dérivable est souhaitée. Nous avons testé différentes combinaisons sans observer d'impact significatif sur les métriques d'évaluation. Nous avons donc conservé les paramètres par défaut, à savoir un nombre *maxFuncs* de 21 et la non-utilisation de splines cubiques.

5.2 Processus Gaussiens

Un Processus Gaussien (GP) est un ensemble de variables aléatoires dont la distribution jointe est une Gaussienne (Rasmussen et Williams, 2005). Pour faire une régression par GP, y est modélisé par : $y = \hat{y} + \Delta\hat{y}$ avec $\Delta\hat{y}$ un bruit Gaussien centré et $\hat{y}$ un GP dont la moyenne et la variance sont à déterminer. L'un des avantages de la régression par GP est que celle-ci renvoie une distribution de probabilités sur toutes les valeurs possibles plutôt qu'une règle de décision ponctuelle. Un autre avantage est qu'elle permet un traitement Bayésien des données, inférant une distribution postérieure à partir d'un prior et d'observations. Comme Rasmussen et Williams (Rasmussen et Williams, 2005), nous définissons un prior centré $GP(0, k(x, x))$ où k est la fonction de covariance.

Nous avons calculé la distribution postérieure en effectuant une inférence exacte à partir des exemples du corpus. Nous avons utilisé la librairie Matlab GPML proposée par Rasmussen et

1. nous rappelons que les valeurs ont été au préalable centrées et normalisées

2. Jekabsons G., ARESLab : Adaptive Regression Splines toolbox pour Matlab/Octave, 2011, http ://www.cs.rtu.lv/jekabsons/

Nickisch [3]. Nous avons testé trois fonctions de covariance : une fonction linéaire, une rationnelle quadratique et une exponentielle et nous avons appris les paramètres de chacune par détermination automatique de la pertinence (*automatic relevance determination* (Jung et Stone, 2009)). Puisque ce calcul est très coûteux ($O(n^3)$ avec n le nombre de données), nous n'avons gardé en mémoire qu'un dictionnaire de points en suivant la méthode de Engel (Engel *et al.*, 2005). La complexité en est réduite à $O(m^2 n)$ avec m la taille du dictionnaire. La section 6 ne présente que les meilleurs résultats qui sont ceux obtenus avec la covariance exponentielle : $\exp(-\frac{1}{2}(x-x')^T \Lambda^{-2}(x-x'))$.

5.3 Machines à vecteurs supports

Originellement, les SVM (Vapnik, 1998) ont été créés pour la classification binaire. Ils prennent en entrée des exemples $\{(x_i, y_i)\}$ avec $y_i \in \{-1, 1\}$ et recherchent un hyperplan dans un espace de variables caractéristiques séparant les deux classes et de manière que la marge par rapport à l'hyperplan soit maximisée. La fonction de décision renvoyée est celle de l'équation 6 où $\phi(x)$ est la projection de x sur l'espace de variables caractéristiques.

$$f(x) = \text{sign}(\omega^T \phi(x) + b) \tag{6}$$

Les SVM pour la classification (SVMClass), pour la régression (SVR) et SVOR résolvent un problème d'optimisation quadratique afin de trouver le vecteur de poids ω et la constante b. La différence entre les trois approches réside dans les contraintes du problème et la forme de la fonction de décision. Nous détaillons ces trois approches dans les sections suivantes.

5.3.1 Machines à vecteurs supports pour la classification

Afin d'effectuer une classification sur 5 classes avec des SVM, nous avons appliqué une approche un-contre-un. Nous avons appris 10 classifieurs, chacun séparant deux classes. A chaque nouvel exemple, nous assignons la classe qui a reçu le plus de vote. En cas d'égalité, nous avons toujours favorisé le label minimal car il est moins critique de prédire un QI inférieur à la vraie valeur que l'inverse. La fonction de décision est celle de l'équation 6.

Nous avons utilisé le logiciel LIBSVM (Chang et Lin, 2011). Nous avons testé SVMClass avec deux fonctions noyau : une fonction linéaire et une exponentielle. Nous ne présentons encore une fois que les meilleurs résultats qui sont ceux obtenus avec le noyau exponentiel d'expression : $\exp(-\gamma(x-x')^T(x-x'))$.

5.3.2 Machines à vecteurs supports pour la régression

Dans le cadre des SVR (Drucker *et al.*, 1996), les poids ω sont définis de sorte à assurer que l'erreur en valeur absolue pour chaque donnée soit inférieure à un paramètre ϵ.

La valeur prédite pour une nouvelle donnée x est alors $f(x) = \omega^T \phi(x) + b$. Nous avons utilisé à nouveau LIBSVM avec une fonction noyau exponentielle. Le paramètre ϵ a été fixé à 0.1.

3. http ://wwww.gaussianprocesses.org/gpml

	Valeur absolue	Euclidienne	ρ de Spearman	κ de Cohen
MLR	0.600 ± 0.007	0.792 ± 0.009	0.831 ± 0.004	0.364 ± 0.008
MARS	0.608 ± 0.008	0.799 ± 0.01	0.821 ± 0.002	0.368 ± 0.009
GP	0.593 ± 0.007	0.853 ± 0.01	0.833 ± 0.005	0.408 ± 0.006
SVR	0.559 ± 0.007	$\mathbf{0.766 \pm 0.009}$	0.837 ± 0.004	0.414 ± 0.008
SVOR	$\mathbf{0.453 \pm 0.005}$	$\mathbf{0.763 \pm 0.008}$	$\mathbf{0.842 \pm 0.005}$	$\mathbf{0.489 \pm 0.006}$
SVMClass	0.580 ± 0.009	0.965 ± 0.01	0.776 ± 0.004	0.421 ± 0.005

TABLE 1 – Résultats de la validation croisée 10-fold sur le corpus LEGO. Nous calculons des intervalles de confiance à 95% pour chaque métrique.

5.3.3 Machines à vecteurs supports pour la régression ordinale

La régression ordinale construit un modèle qui prédit des classements ordinaux. SVOR (Chu et Keerthi, 2007) adjoint des contraintes à la formulation des SVM de manière à prendre en compte l'ordre entre les labels. L'idée est de trouver un vecteur de poids optimal ω comme dans SVMClass mais aussi $r-1$ seuils définissant b_j hyperplans discriminants pour les r labels. Pour chaque donnée x_i^j (l'exposant j signifie que x_i^j appartient à la j-ème catégorie), en notant b_j le seuil séparant les catégories j et $j+1$, on impose à la fonction de valeur $\omega^T \phi(x_i^j)$ d'être inférieure à la marge inférieure $(b_j - 1)$ et on impose à chaque exemple x_i^{j+1} d'être associé à une valeur $\omega^T \phi(x_i^{j+1})$ supérieure à la marge supérieure $(b_j + 1)$.

On associe à chaque nouvel exemple x le label $\text{argmin}_i\{i \mid \omega^T \phi(x) < b_i\}$. Nous avons utilisé un noyau Gaussien avec le programme C disponible à l'adresse http ://www.gatsby.ucl.ac.uk/˜chuwei/svor.htm.

6 Résultats

Le tableau 1 présentent les résultats du test. Pour les techniques de régression, le coefficient κ a été calculé sur les arrondis des prédictions.

De manière surprenante, les résultats de la MLR sur le corpus sont plutôt bons. Schmitt *et. al* ont appliqué la MLR sur le même corpus et ils ont obtenu un κ de 0.35 et un ρ de 0.46 sur les valeurs arrondies. Nous avons obtenu un ρ de 0.83. Cette différence de résultats s'explique par le fait que des ensembles différents de variables explicatives ont été utilisés et que nous avons fait le choix de séparer les variables qualitatives en variables booléennes.

Hormis l'erreur euclidienne, les erreurs sont moindres avec les GP qu'avec MLR et MARS. MLR et MARS minimisent la distance euclidienne donc ce résultat pouvait être anticipé. Les résultats avec SVR sont quant à eux meilleurs qu'avec ces trois modèles sur toutes les métriques, notamment les erreurs euclidiennes et en valeur absolue.

SVOR surpasse toutes ces méthodes sur l'ensemble des métriques utilisées. On observe une amélioration significative des erreurs en valeur absolue et du κ. L'erreur en valeur absolue est améliorée de 0.1 points comparée à SVR. Le coefficient κ est amélioré de 0.06 point par rapport à

SVMClass.

Le coefficient de corrélation le plus faible a été obtenu avec SVMClass. Cela corrobore le fait que la classification ne soit pas la technique la plus adaptée à la prédiction de la QI car elle ignore l'ordre des labels. SVMClass est toutefois, comme l'on pouvait s'y attendre, la seconde technique la plus performante concernant κ. Les performances de classification de SVMClass et SVOR peuvent être plus amplement comparées grâce à leurs matrices de confusion respectives (Tableau 2). Les résultats prédits par SVOR dévient moins des valeurs réelles que ceux prédits par SVMClass. Par exemple, une donnée avec le label 1 a reçu les labels 3, 4 ou 5 88 fois avec SVOR contre 236 avec SVMClass. Les rappel et précision pour les labels critiques 1 et 2 sont significativement meilleurs avec SVOR.

SVOR	V 1	V 2	V 3	V 4	V 5	Précision
P 1	572	99	32	4	1	0.808
P 2	124	145	118	33	1	0.344
P 3	69	236	378	228	23	0.405
P 4	18	76	347	716	359	0.472
P 5	1	1	20	269	1410	0.829
Rappel	0.730	0.260	0.422	0.573	0.786	
SVMClass	T 1	T 2	T 3	T 4	T 5	Précision
P 1	541	95	55	27	3	0.719
P 2	57	70	55	23	2	0.183
P 3	98	188	253	164	19	0.313
P 4	81	182	432	626	274	0.440
P 5	7	22	100	410	1496	0.781
Rappel	0.690	0.126	0.282	0.501	0.834	

TABLE 2 – Matrices de confusion pour SVOR et SVMClass. V *i* signifie que *i* est le vrai label et P *i* signifie que *i* est le label prédit.

7 Conclusion

Cet article propose de modéliser la prédiction de la qualité d'interaction avec un système de dialogue comme un problème de régression ordinale. Nous avons comparé différents modèles de classification et de régression à la régression ordinale sur un corpus de dialogues évalués et nous avons montré que la régression ordinale aboutissait aux meilleurs résultats en termes d'erreur euclidienne, d'erreur en valeur absolue, de coefficient de corrélation de Spearman et de coefficient κ de Cohen. Ces résultats s'expliquent par le fait que la régression ordinale prend en compte l'ordre naturel entre les labels discrets attribués à chaque dialogue. Nous utiliserons par la suite la régression ordinale afin d'inférer une fonction de récompense pour un système de dialogue apprenant par renforcement.

Références

CHANG, C.-C. et LIN, C.-J. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2.

CHU, W. et KEERTHI, S. S. (2007). Support vector ordinal regression. *Neural Computation*, 19.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

DAUBIGNEY, L., GAŠIĆ, M., CHANDRAMOHAN, S., GEIST, M., PIETQUIN, O. et YOUNG, S. (2011). Uncertainty management for on-line optimisation of a pomdp-based large-scale spoken dialogue system. *In Proc. of Interspeech*, pages 1301–1304.

DRUCKER, H., BURGES, C. J., KAUFMAN, L., SMOLA, A. et VAPNIK, V. (1996). Support Vector Regression Machines. *In Proc. of NIPS*.

DYBKJAER, L., BERNSEN, N. O. et MINKER, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43:33–54.

EL ASRI, L., LAROCHE, R. et PIETQUIN, O. (2012). Reward function learning for dialogue management. *In Proc. of STAIRS*.

ENGEL, Y., MANNOR, S. et MEIR, R. (2005). Reinforcement learning with gaussian processes. *In Proc. of ICML*, pages 201–208.

ENGELBRECHT, K.-P., GÖDDE, F., HARTARD, F., KETABDAR, H. et MÖLLER, S. (2009). Modeling user satisfaction with hidden markov model. *In Proc. SIGDIAL*.

EVANINI, K., HUNTER, P., LISCOMBE, J., SUENDERMANN, D., DAYANIDHI, K. et PIERACCINI, R. (2008). Caller experience : A method for evaluating dialog systems and its automatic prediction. *In Proc. of IEEE SLT*, pages 129–132.

FRIEDMAN, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67.

HARA, S., KITAOKA, N. et TAKEDA, K. (2010). Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system. *In Proc. of LREC*.

JUNG, T. et STONE, P. (2009). Feature selection for value function approximation using bayesian model selection. *In Proc. of ECML-PKDD*.

LEVIN, E., PIERACCINI, R. et ECKERT, W. (1997). Learning dialogue strategies within the markov decision process framework. *In Proc. of IEEE ASRU*.

MÖLLER, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer.

PIETQUIN, O., GEIST, M., CHANDRAMOHAN, S. et FREZZA-BUET, H. (2011). Sample-efficient batch reinforcement learning for dialogue management optimization. *ACM Transaction on Speech and Language Processing*, 7(3):1–21.

RASMUSSEN, C. E. et WILLIAMS, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

RAUX, A., LANGNER, B., BLACK, A. et ESKENAZI, M. (2003). LET'S GO : Improving Spoken Dialog Systems for the Elderly and Non-natives. *In Proc. of Eurospeech*.

SCHMITT, A., SCHATZ, B. et MINKER, W. (2011). Modeling and predicting quality in spoken human-computer interaction. *In Proc. of SIGDIAL*.

SCHMITT, A., ULTES, S. et MINKER, W. (2012). A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. *In Proc. of LREC*.

SHASHUA, A. et LEVIN, A. (2002). Ranking with large margin principle : Two approaches. *In Proc. of NIPS*.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

VAPNIK, V. N. (1998). *Statistical Learning Theory*.

WALKER, M. A., FROMER, J. C. et NARAYANAN, S. (1998). Learning optimal dialogue strategies : A case study of a spoken dialogue agent for email. *In Proc. of COLING/ACL*, pages 1345–1352.

WALKER, M. A., LITMAN, D. J., KAMM, C. A. et ABELLA, A. (1997). PARADISE : a framework for evaluating spoken dialogue agents. *In Proc. of EACL*, pages 271–280.

WITT, S. (2011). A global experience metric for dialog management in spoken dialog systems. *In Proc. of SemDial*, pages 158–166.

Développement d'un système de reconnaissance automatique de la parole en coréen avec peu de ressources annotées

Antoine Laurent Lori Lamel
Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
laurent@limsi.fr, lamel@limsi.fr

RÉSUMÉ

Ce papier décrit le développement d'un système de reconnaissance automatique de la parole pour le coréen. Le coréen est une langue alpha-syllabique, parlée par environ 78 millions de personnes dans le monde. Le développement de ce système a été mené en utilisant très peu de données annotées manuellement. Les modèles acoustiques ont été adaptés de manière non supervisée en utilisant des données provenant de différents sites d'actualités coréens. Le corpus de développement contient des transcriptions approximatives des documents audio : il s'agit d'un corpus transcrit automatiquement et aligné avec des données provenant des mêmes sites Internet. Nous comparons différentes approches dans ce travail, à savoir, des modèles de langue utilisant des unités différentes pour l'apprentissage non supervisé et pour le décodage (des caractères et des mots avec des vocabulaires de différentes tailles), l'utilisation de phonèmes et d'unités "demi-syllabiques" et deux approches différentes d'apprentissage non supervisé.

ABSTRACT

Development of a speech recognition system for Korean with little annotated training data

This paper describes the development of a speech-to-text transcription system for the Korean language. Korean is an alpha-syllabary language spoken by about 78 million people worldwide. System development was carried out with only a small amount of manually transcribed audio data from LDC. Additional audio data downloaded from several Korean were included in the training corpus and used in an unsupervised manner. Korean web texts were also downloaded and used for language modeling. The development corpus is associated with approximative transcripts obtained by aligning automatic transcripts with texts taken from the same web site. We compare several systems : the use of different LM (in terms of vocabulary size and using chars instead of words) for the unsupervised training and decoding, the use of phonemes and "half-syllable" acoustic units and two different approaches for unsupervised training.

MOTS-CLÉS : Reconnaissance automatique de la parole, apprentissage non supervisé, langue sous dotée.

KEYWORDS: Speech recognition system, unsupervised training, under-resourced language.

1 Introduction

Dans le cadre du projet RAPMAT [1], il était nécessaire de réaliser un système de reconnaissance automatique de la parole pour le coréen. Hors, ne disposant pas de données pour l'apprentissage d'un tel système, nous avons tout d'abord développé un système de base en utilisant les ressources (en quantité très limitée) disponible via le LDC [2], puis étudié une solution d'apprentissage non supervisé que nous présentons dans cet article.

Traditionnellement, les systèmes de reconnaissance automatique de la parole sont entrainés sur de grandes quantités de données audio transcrites et sur une importante quantité de textes écrits (Lamel et Vieru, 2010). Cependant, l'obtention d'un tel corpus est couteux et requiert des locuteurs natifs dans la langue pour laquelle le système est développé. L'un des coûts de développement les plus cités est celui du temps nécessaire à l'obtention des données audio transcrites, puisque cela nécessite à la fois de la main d'oeuvre et du temps. De nombreuses sources audio (radio, television, web, ...) peuvent être trouvées, cependant, dans la majorité des cas, les données audio ne sont pas associées à des transcriptions précises (Lamel *et al.*, 2002).

Plusieurs projets de recherche se sont intéressés à la réduction de ces coûts (Kimball *et al.*, 2004) et quelques données d'entrainement audio, comme dans le programme DARPA Gale sont associés à des transcriptions rapides (Cieri *et al.*, 2004). Quelques sources audio sont associées avec du texte, comme des sous-titres, des résumés ou des transcriptions plus éloignées (provenant par exemple de page Internet). Une variété d'approches a été étudiée, la plupart s'appuyant sur la supervision d'un modèle de langue. Les différentes approches varient en terme d'utilisation ou non des mesures de confiance (Collan *et al.*, 2005; Bisani et Ney, 2008; Wessel et Ney, 2005) et selon le degré de supervision (Lamel *et al.*, 2002). Les auteurs de (Ma et Schwartz, 2008) proposent un apprentissage itératif.

Dans cette étude, le développement du système est très légèrement supervisé. Nous avons utilisé un petit corpus annoté pour construire un premier modèle de langue et un modèle acoustique et ensuite seulement des données audio pour améliorer nos modèles. Les modèles de langue ont été construits en utilisant plusieurs sources textuelles différentes. Nous proposons de construire plusieurs systèmes en utilisant différents modèles de langue pour l'apprentissage non supervisé et pour le décodage (en terme de taille du vocabulaire, en utilisant des caractères à la place des mots), en utilisant les phonèmes et les "demi-syllables" comme unités acoustiques et en utilisant deux approches différentes pour l'apprentissage non supervisé des modèles acoustiques.

La section suivante donne un aperçu des caractéristiques de la langue coréenne, suivie par une description de l'approche et du corpus utilisé dans cette étude. Ensuite, les modèles de langue, la liste des phonèmes et les modèles acoustiques sont décrits, avant de présenter le contexte expérimental et les différents résultats obtenus.

2 La langue coréenne

Pendant plus d'un millénaire, le coréen était écrit avec des caractères Chinois adaptés, appelés Hanja. Au XVème siècle, un système d'écriture national appelé Hangul a été proposé et complète-

1. Ce travail a été en partie financé par le projet DGA RAPID RAPMAT : http ://www.limsi.fr/tlp/rapmat.html
2. https ://www.ldc.upenn.edu/

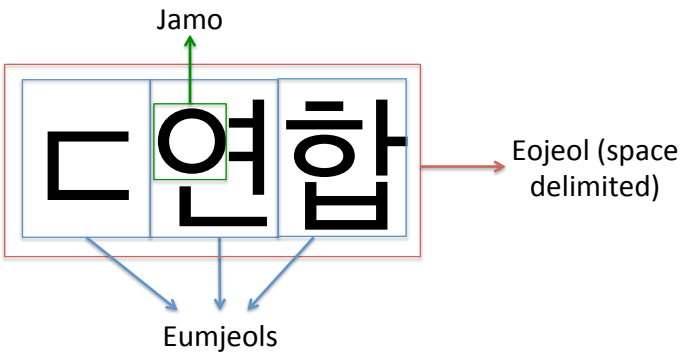


FIGURE 1 – Symbole coréen

ment adopté au XXème siècle. Comme présenté dans (Kwon et Park, 2003), en coréen, un espace est placé entre deux "mots-phrases" adjacents, qui représentent deux ou trois mots Anglais au sens sémantique. C'est un système alpha-syllabique (Taylor, 1979). Comme décrit dans (Lee et Lee, 2009), les ensembles de Jamo (représentant des phonèmes) sont groupés en Eumjeols (syllables), et les séquences d'Eumjeols sont regroupées en Eojeol (mots séparés par des espaces), voir figure 1. Chaque Eumjeols est composé de deux ou trois éléments : le Choseong (consonne initiale), le Jungseong (la voyelle) et le Jongseong (optionnel, la consonne finale).

La plupart des résultats des systèmes de reconnaissance automatique de la parole développés pour le coréen sont substantiellement moins bon que les résultats obtenus avec des langues plus dotées telles que le Français ou l'Anglais. Comme les Eojeol représentent généralement plus d'un mot, le vocabulaire du système de reconnaissance basé sur les mots (éléments séparés par des espaces) doit contenir un nombre élevé d'entrées (Kiecza *et al.*, 1999). Par exemple, un corpus Anglais contenant 40 millions de mots contient environ 190000 mots distincts (Lamel et Vieru, 2010), alors que notre corpus de 95 millions de mots en coréen contient environ 2 millions de mots différents. Une solution proposée pour ce problème est la décomposition des mots en morphèmes (comme présenté dans (Kwon et Park, 2003; Choi *et al.*, 2004)). Une autre explication des mauvais résultats des différents systèmes est le manque de données de parole et de texte pour l'apprentissage des systèmes. Seulement un nombre très restreint de corpus sont disponibles pour le coréen, et la plupart des travaux précédent se basent sur des corpus non redistribués.

3 Approche et corpus

L'approche général retenue pour ce travail est comparable à celle présentée dans (Ma et Schwartz, 2008; Lamel *et al.*, 2000, 2002; Lamel et Vieru, 2010) dans le sens ou un système de reconnaissance de la parole est utilisé pour produire des transcriptions approximatives pour l'apprentissage des modèles acoustiques. Dans cette étude, les données audio sont transcrites en lots et en procèdent à des itérations successives, les modèles sont estimés sur une quantité plus importante de données. Nous proposons également une autre approche dans laquelle à chaque itération, l'ensemble du corpus audio est transcrit pour l'adaptation des modèles acoustiques. Notre approche a été développée et testée sur un corpus pour lequel nous n'avons pas de référence exacte. Comme personne au laboratoire n'a de connaissances sur le coréen, nous ne savons pas à quel point notre corpus, extrait de sites web d'actualités coréen, est distant de l'audio provenant de

ces mêmes sites web.

Dans cette étude, nous explorons plusieurs approches, l'objectif étant de donner un aperçu des techniques qui semblent fonctionner et de celles qui semblent ne pas donner de bons résultats.

Nous avons utilisé un ensemble de documents audio et textuels pour l'apprentissage des différents modèles. Nous avons utilisé un corpus provenant de LDC qui contient des données d'émissions de journaux transcrites. Ce corpus ne contient qu'une petite quantité de données (9 heures). Nous avons également utilisé, comme données textuelles, le corpus LDC Korean newswire 2 et les transcriptions du corpus LDC de conversations téléphoniques pour l'apprentissage des modèles de langue (55M + 230k mots). Nous avons également utilisé des données provenant de trois différents sites d'actualités coréen : VOA [3], RFA [4] et NHK [5]. Pour NHK, 406 heures de données ont été téléchargées depuis novembre 2007 avec des transcriptions approximatives (5,5M mots). Ces transcriptions approximatives correspondent au contenu HTML des pages web associées à des données audio. Les deux autres sources de données ont été téléchargées depuis octobre 2013 : nous avons récupéré 5 heures de données de RFA et 4 heures de VOA. Les transcriptions pour RFA, VOA et NHK extraites des pages web, ne couvrent qu'une partie de l'audio et ne sont pas alignées. Une partie de RFA, VOA et NHK a été utilisé comme corpus de développement / test approximatif. Le corpus a été automatiquement transcrit en utilisant notre système d'amorçage utilisant uniquement les données de LDC, puis, en utilisant un algorithme basé sur la programmation dynamique (DTW), nous avons aligné les sorties automatiques du système avec les contenu HTML des pages et nous avons écarté les parties pour lesquels aucun alignement n'a pu être réalisé. Le même corpus a été utilisé pour le développement et le test de notre système.

4 Modèles de langue

Les données textuelles de LDC et les transcriptions approximatives de NHK, VOA et RFA ont été utilisées pour construire les différents modèles de langue. Nous avons construits des modèles 2-gram, 3-gram et 4-gram en utilisant l'ensemble des 2 millions de mots distincts de notre corpus d'apprentissage. Ces modèles de langue ngram ont été obtenus en interpolants les différents modèles entrainés sur les sous ensembles de données en utilisant la technique du lissage Kneser-Ney modifié. Les poids d'interpolations ont été automatiquement estimés en utilisant l'algorithme EM de façon à minimiser la perplexité du modèle final sur le corpus de développement/test. Ce corpus de développement est composé d'environ 10k mots. Nous avons également construit des modèles de langue contenant les 200k unigram les plus probables dans le modèle de langue 1-gram interpolé, et nous avons construit un modèle de langue à base de caractères.

La normalisation des textes coréens a été une grosse partie de ce travail. Comme nous n'avons pas de connaissance de la langue, nous avons utilisé la littérature disponible pour nettoyer les textes. Nous avons défini une liste de caractères obsolètes et ne respectant pas l'encodage EUC-KR (en utilisant les informations disponibles sur http ://en.wikipedia.org/wiki/Hangul), et retiré l'ensemble des phrases contenant des éléments de cette liste de notre corpus d'apprentissage. Nous avons également retiré les phrases contenant des séquences illégales de symboles, ou

3. http ://www.voakorea.com/
4. http ://www.rfa.org/korean/
5. http ://www.nhk.or.jp/korean/

FIGURE 2 – Exemple de découpage en "demi-syllabe"

contenant des symboles composés uniquement d'un seul Jamo. Les phrases contenant des mots anglais ont également été retirées. 122 caractères ont été identifiés comme étant des variantes du caractère espace, et donc substitués. Un dictionnaire contenant 25251 mots est livré avec le corpus de conversations téléphonique du LDC (ldc2003l02), ainsi qu'un outil pour phonétiser automatiquement les mots nouveaux. Nous avons également utilisé cet outil pour trouver les séquences illégales de symboles.

5 Modélisation acoustique

5.1 Liste des phonèmes et unités acoustiques

Les mots d'origine étrangère mis à part, le coréen est écrit en utilisant 14 consonnes simples et 5 consonnes doubles qui sont formées à partir des consonnes simples. Il y a 9 voyelles simples et 12 voyelles complexes. Ces voyelles complexes sont appelées diphtongues, elles sont la combinaison de 2 voyelles simples. Les mots coréens sont écrits de la gauche vers la droite. L'ensemble de phonèmes retenu pour ce travail comprend 25 phonèmes, 13 consonnes (les phonèmes K et k ont été regroupés en k, T et tt en t, P et pp en p, c et cc en C et S et ss en s – au format Sampa), 9 voyelles et 3 unités pour les évènements acoustiques (respiration, silence, filler).

Des modèles avec des "demi-syllabes" comme unité phonémique ont également été construits. Partant de la représentation phonétique de chaque symbole, les phonèmes ont été regroupés pour former ces "demi-syllabes" : chaque phonème jusqu'à rencontrer la première voyelle est considéré comme la première partie de la syllabe, et depuis cette voyelle jusqu'à la fin de la représentation phonétique du symbole est considéré comme la deuxième partie (voir figure 2).

5.2 Modèles acoustiques

Un modèle acoustique pour les phonèmes anglais a été utilisé pour initialiser notre modèle acoustique coréen. Une correspondance entre les phonèmes anglais et les phonèmes coréens a été faite et les modèles extraits ont été utilisés pour initialiser notre nouveau modèle. Les paramètres cepstraux standards (perceptual linear prediction - PLP) ont été utilisés. Les vecteurs de paramètres cepstraux sont composés de 39 paramètres : 12 coefficients cepstraux plus l'énergie sous forme logarithmique, associés à leur dérivées premières et secondes. Les modèles acoustiques sont des mélanges de gaussiennes à états partagés, gauche-droite dépendant du contexte. Les

modèles de phonèmes tri phones dépendants du contexte sont indépendants des mots, mais dépendants de leur position à l'intérieur des mots. Les états partagés sont obtenus au moyen d'un arbre de décision. Les modèles acoustiques sont indépendants du genre avec un apprentissage adaptatif (Speaker Adaptative Training - SAT). Le silence est modélisé par un seul état à 1024 gaussiennes.

6 Résultats expérimentaux

L'une des difficultés concernant le développement d'un système de reconnaissance de la parole basé sur les mots en coréen est le fait que le vocabulaire doit contenir un nombre élevé d'entrées. Comme présenté dans le tableau 1, en utilisant un modèle de langue contenant 200k mots (200k LM), le taux de mots hors vocabulaire (Out-Of-Vocabulary – OOV) sur notre corpus de développement est élevé (presque 10%), et est toujours de 3% avec un modèle de langue de 2 millions de mots (2M LM) contenant tous les mots du corpus d'apprentissage.

TABLE 1 – Perplexité et taux d'OOV des différents modèles de langue

Modèle	PPL	OOV (%)
2M LM (4g)	732	3.02
200k LM (4g) LM	1596	9.69

Le tableau 2 montre les résultats en termes de taux d'erreur mots (Word Error Rate – WER) et en termes de taux d'erreur sur les caractères (CER) en utilisant le modèle de langue 200k LM pour le décodage du corpus de développement.

TABLE 2 – WER et CER en utilisant 200k LM pour le décodage

Sources Audio	heures	200k LM		2M LM	
		WER	CER	WER	CER
LDC	9	50.6	32.1	50.6	32.1
Web	10	50.2	32.8	50.1	32.6
LDC+Web	19	49.1	31.5	48.8	31.0
LDC+Web	34	48.7	30.4	47.7	30.4
LDC+Web	79	48.0	29.9	47.6	29.7

Dans ce tableau, 200k LM et 2M LM correspondent au modèle de langue utilisé pour l'apprentissage non supervisé. Nous pouvons remarquer que le WER est relativement élevé. Cela peut être expliqué en partie par le fait que notre corpus de développement est un corpus approximatif et par le fait que, comme dit précédemment, chaque mot correspond à deux ou trois mots anglais au sens sémantique (Kwon et Park, 2003). En utilisant seulement un modèle acoustique estimé de façon non supervisé sur 10 heures de données audio provenant d'Internet (appelé "Web"), le WER du système est meilleur qu'en utilisant le corpus de LDC estimé sur un corpus annoté de 9 heures. Cela peut être expliqué par le fait que notre corpus de développement sur lequel le système est testé provient lui aussi de différents sites Web. L'ajout de nouvelles données permet d'observer une diminution en termes de WER et de CER.

Le tableau 3 montre les résultats obtenus en utilisant le modèle de langue de 2 millions de mots pour le décodage du corpus de développement. Nous pouvons remarquer que les résultats obtenus avec ce modèle sont meilleurs en terme de WER et de CER que ceux obtenus avec le modèle 200k et que la même tendance est visible lors de l'adaptation non supervisée du modèle acoustique.

TABLE 3 – WER et CER en utilisant 2M LM pour le décodage

Sources Audio	heures	200k LM WER	200k LM CER	2M LM WER	2M LM CER
LDC	9	47.5	29.4	47.5	29.4
Web	10	48.1	30.5	46.6	29.5
LDC+Web	19	45.2	28.2	42.3	27.0
LDC+Web	34	44.7	27.7	41.1	25.7
LDC+Web	79	44.0	27.0	40.2	25.4
LDC+Web	150	43.1	26.4	39.9	25.2

D'autres expériences, dans lesquelles un modèle de langue à base de caractères (char LM) pour l'apprentissage non supervisé, ont également été menées. Les modèles acoustiques résultant de ces expériences ont été utilisés pour décoder le corpus de développement avec le modèle 2M LM et avec le modèle char LM. Le décodage avec le modèle 2M LM ne donne pas de bons résultats en terme de WER (44,5% avec notre meilleur système), mais les résultats sont comparables, en terme de CER, en utilisant les modèles char LM ou 2M LM. Les résultats sont présentés dans le tableau 4.

TABLE 4 – WER et CER en utilisant char LM pour l'apprentissage non supervisé

Sources Audio	Heures	2M LM WER	2M LM CER	Char LM CER
LDC+Web	79	44.5	26.7	27.0
LDC+Web	150	44.0	26.4	26.5

Le tableau 5 montre les résultats obtenus en utilisant la deuxième technique d'apprentissage non supervisée : l'ensemble des données audio est transcrit, le modèle adapté, puis l'opération de transcription recommence. La diminution en termes de WER et de CER est plus rapide qu'en utilisant l'autre méthode : dès la première itération les résultats sont meilleurs que lorsqu'on ajoute les données au fur et à mesure.

TABLE 5 – WER et CER en utilisant toutes les données à chaque itération

Sources Audio	Iteration	2M LM WER	2M LM CER
LDC+Web 150h	1	39.7	25.2
LDC+Web 150h	2	39.7	25.3
LDC+Web 150h	3	39.6	25.3
LDC+Web 150h	4	39.5	25.1

Les expériences dans d'autres langues montrent une diminution plus rapide et plus importante des WER et CER en utilisant les deux méthodes non supervisées présentées. Nous pensons

que cette plus petite amélioration (WER de 47,5 à 39,5% et CER de 29,5 à 25,1%) observée pour le coréen est due à l'utilisation de notre corpus approximatif. Un corpus annoté devrait être disponible prochainement (un coréen natif a été recruté pour réaliser les transcriptions manuellement).

Nous pouvons remarquer dans le tableau 6 que les résultats en utilisant des "demi-syllabes" ne sont pas encourageants. Nous espérons également que l'expertise du transcripteur coréen va nous permettre d'améliorer la représentation utilisée.

TABLE 6 – WER et CER en utilisant 200k LM pour le décodage en unités "demi-syllabiques" et pour l'apprentissage

Sources Audio	Heures	200k LM WER	200k LM CER
LDC+Web	19	59.3	41.8
LDC+Web	79	53.8	34.5
LDC+Web	150	52.4	32

7 Conclusion

La description du développement d'un système de reconnaissance automatique de la parole pour le coréen, avec peu de données annotées à été présentée dans ce papier. Uniquement des données audio provenant de sites Web ont été utilisées pour l'adaptation non supervisées des modèles acoustiques. Pour cette étude, un corpus de développement/test approximatif a été utilisé. Ce corpus a été construit à partir de données (audio et textes) provenant de sites d'actualités coréens.

Les résultats montrent une diminution d'environ 17% en terme de WER et de 15% en terme de CER (relatif). Deux méthodes d'apprentissage non supervisées ont été proposées avec différents types de modèles de langue pour l'apprentissage et le décodage.

Un coréen natif va rejoindre l'équipe, il sera en mesure de transcrire notre corpus de développement approximatif et pourra nous aider pour la normalisation du texte et sa représentation acoustique.

La normalisation des données textuelles a été améliorée depuis le début de ce travail, nous aurons donc à recommencer l'apprentissage non supervisé du début pour construire de nouveaux modèles. Enfin, nous mettrons en place des modèles MLP (Multi-Layer Perceptron) pour améliorer nos modèles.

En utilisant peu de données annotées, nous avons développé un système de reconnaissance automatique de la parole en coréen. La méthode, nécessitant uniquement des données audio pour adapter le modèle acoustique, pourrait être appliquée à d'autres langues peu dotées, ou pour adapter des systèmes pour lesquels des ressources sont disponibles, à des conditions audio spécifiques.

Références

Bisani, M. et Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *In Speech Communication*, volume 50, pages 434–451.

Choi, I.-J., Kim, N.-H. et Yoon, S. Y. (2004). Large vocabulary continuous speech recognition based on cross-morpheme phonetic information. *In Interspeech 2004*, pages 453–456.

Cieri, C., Miller, D. et K., W. (2004). The fisher corpus : a resource for the next generations of speech-to-text. *In LREC 2004*, pages 69–71.

Collan, C., Bisani, M., Kanthak, S., Schlüter, R. et Ney, H. (2005). Cross domain automatic transcription on the tc-star epps corpus. *In ICASSP 2005*, volume 1, pages 825–828.

Kiecza, D., Schultz, T. et Waibel, A. (1999). Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR. *In Proceedings of International Conference on Speech Processing (ICSP'99)*, pages 323–327, Seoul, Korea.

Kimball, O., Kao, C., Iyer, R., Arvizo, T. et Makhoul, J. (2004). Using quick transcriptions to improve conversational speech models. *In Interspeech 2004*, pages 2265–2268.

Kwon, O.-W. et Park, J. (2003). Korean large vocabulary continous speech recognition with morpheme-based recognition units. *Speech Communication*, 39:287–300.

Lamel, L., Gauvain, J.-L. et Adda, G. (2000). Lightly supervised acoustic model training. *ITRW ASR*, 1:150–154.

Lamel, L., Gauvain, J.-L. et Adda, G. (2002). Lightly supervised and unsupervised acoustic model trainings. *Computer Speech and Language*, 16:115–129.

Lamel, L. et Vieru, B. (2010). Development of a speech-to-text transcription system for finnish. *In Workshop on Spoken Languages Technologies for Under Resourced Languages (SLTU 2010)*, pages 62–67, Penang, Malaysia.

Lee, J. et Lee, G. G. (2009). A data-driven grapheme-to-phoneme conversion method using dynamic contextual converting rules for korean tts systems. *Computer Speech and Language*, 23:423–434.

Ma, J. et Schwartz, R. (2008). Unsupervised versus supervised training of acoustic models. *In Interspeech 2008*, pages 2374–2377.

Taylor, I. (1979). The korean writing system : An alphabet ? a syllabary ? a logography ? *In Proceeding of Visible Language*, volume 2, pages 67–82, New York, USA.

Wessel, F. et Ney, H. (2005). Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *In IEEE Transactions on Speech and Audio Processing*, pages 23–31.

Train&Align : un outil d'alignement phonétique automatique disponible en ligne

Sandrine Brognaux[1, 2] Sophie Roekhaut[1] Thomas Drugman[2] Richard Beaufort[3]

(1) Cental - Université catholique de Louvain (UCL), Belgium

(2) TCTS - UMons, Belgium (3) Nuance Communications, Inc., Belgium *

sandrine.brognaux@uclouvain.be, sophie.roekhaut@uclouvain.be, thomas.drugman@umons.ac.be, richard.beaufort@nuance.com

RÉSUMÉ

Plusieurs outils d'alignement phonétique automatique de corpus oraux sont actuellement disponibles. Ils exploitent généralement des modèles indépendants du locuteur pour aligner de nouveaux corpus. Leur désavantage est qu'ils couvrent un nombre très limité de langues et fournissent parfois un alignement de qualité réduite quand ils sont appliqués sur différents styles de parole. Cet article présente Train&Align, un nouvel outil d'alignement phonétique automatique disponible en ligne (http ://cental.fltr.ucl.ac.be/train_and_align). Sa spécificité est qu'il entraîne les modèles directement sur le corpus à aligner, ce qui le rend applicable à toutes les langues et à tous les styles de parole. Des tests effectués sur six corpus dans plusieurs langues et styles de parole montrent qu'il produit un alignement de qualité comparable aux autres outils d'alignement. Train&Align permet également d'optimiser certains paramètres d'entraînement. Ainsi, une partie manuellement alignée du corpus peut notamment être utilisée afin d'améliorer la qualité des modèles. Les tests montrent une amélioration du taux d'alignement dépassant les 15%, quand 30 secondes de corpus aligné manuellement sont utilisées.

ABSTRACT

Train&Align : An automatic phonetic alignment tool available online

Several automatic phonetic alignment tools have been proposed in the literature. They usually rely on pre-trained speaker-independent models to align new corpora. Their drawback is that they cover a very limited number of languages and might not perform properly for different speaking styles. This paper presents Train&Align, a new tool for automatic phonetic alignment available online. Its specificity is that it trains the models directly on the corpus to align, which makes it applicable to any language and speaking style. Experiments on six corpora in different languages and speaking styles show that it provides results comparable to other existing tools. Train&Align also allows the tuning of some training parameters. A manually-aligned part of the corpus can, for instance, be used as bootstrap to improve the model quality. Alignment rates were found to significantly increase, by more than 15%, using only 30 seconds of bootstrapping data.

MOTS-CLÉS : Phonétique, alignement, HMM, corpus oraux, annotation.

KEYWORDS: Phonetics, alignment, HMM, speech corpora, annotation.

∗. Cette étude a été réalisée alors que Richard Beaufort travaillait au CENTAL (UCL, Belgium).

1 Introduction

Les corpus oraux de taille importante jouent un rôle crucial tant en recherche linguistique que dans les technologies de la parole. Une particularité de ces corpus est que le son est rarement étudié seul. Ses transcriptions orthographique et phonétique sont généralement nécessaires. Les phonèmes, en particulier, doivent être alignés avec le son. Des outils d'annotation tels que Praat (Boersma et Weenink, 2009), EXMARaLDA (Schmidt, 2012) ou ELAN (Wittenburg *et al.*, 2006) permettent de définir plusieurs couches d'annotation. Ils offrent également la possibilité d'aligner manuellement ces différentes annotations avec le son. Cependant, l'alignement manuel présente deux problèmes majeurs. D'une part, un tel traitement requiert un temps considérable : de 130 à 800 fois la durée du son (Kawai et Toda, 2004; Schiel et Draxler, 2003). D'autre part, ce travail nécessite des phonéticiens entraînés et une bonne concertation entre les différents annotateurs.

Pour tenter de résoudre ce problème, des outils d'alignement automatique ont vu le jour. Ils permettent de fournir un alignement à la fois cohérent et reproductible. Les méthodes fondées sur l'utilisation de modèles de Markov cachés (HMM) sont reconnues comme fournissant les meilleurs résultats d'alignement phonétique (Adell *et al.*, 2005; van Niekerk et Barnard, 2009). Comme en reconnaissance vocale, des modèles acoustiques de chaque (groupe de) phonème(s) sont appris sur un corpus de taille plus ou moins importante. Ces modèles servent ensuite à aligner d'autres corpus avec leur transcription phonétique.

Le toolkit HTK (Young *et al.*, 1995) offre les outils nécessaires à l'alignement phonétique automatique. Utilisable en ligne de commande, il nécessite cependant un certain niveau de programmation de la part de l'utilisateur. Plusieurs outils, avec ou sans interface graphique, tels que EasyAlign (Goldman, 2011), SPPAS (Bigi, 2012) ou P2FA (Yuan et Liberman, 2008) sont donc venus se greffer sur HTK ou sur des toolkits similaires. Ils fournissent des modèles indépendants du locuteur entraînés sur de grandes bases de données multi-locuteurs ainsi que les méthodes pour aligner d'autres corpus à l'aide de ces modèles. Ces outils présentent trois limites majeures. Premièrement, les modèles ne sont proposés que dans un nombre très limité de langues (p. ex. quatre langues pour EasyAlign). Il est alors impossible d'aligner un corpus dans une langue non couverte. Une seconde limitation concerne le fait que la phase d'entraînement n'est pas accessible. Ces outils proposant des modèles pré-entraînés, il n'est pas possible d'optimiser les paramètres d'entraînement. Enfin, les modèles fournis sont censés offrir une représentation globale de la langue, non spécifique à un locuteur ou un style de parole particulier, on parle ainsi de modèles génériques. Cependant, les modèles dépendent fortement du corpus d'entraînement, généralement un corpus de parole neutre et lue. Leur utilisation pour l'alignement d'autres styles de parole (spontanée, expressive, etc.) produit donc souvent un alignement de qualité réduite. Cette dernière limitation concerne également les phonèmes utilisés lors de l'annotation : seuls les phonèmes présents dans le modèle peuvent être alignés. Ainsi, EasyAlign ne propose pas de modèle pour le phonème /ŋ/, prononcé notamment en fin du mot "camping", tandis que SPPAS n'effectue pas la distinction entre les phonèmes /ɛ̃/ et /œ̃/ (de "brin" et "brun"). Enfin, si certains phonèmes étaient rares ou sous-représentés dans le corpus d'entraînement, ils seront plus enclins à être mal alignés, comme le soulignent Goldman et Schwab (2011).

Train&Align [1] est un nouvel outil d'alignement phonétique automatique qui offre une solution pour pallier ces problèmes. Il se distingue des outils existants par le fait qu'il entraîne directement les modèles acoustiques sur le corpus à aligner. Il ne nécessite donc pas de modèle préalablement

1. /http ://cental.fltr.ucl.ac.be/train_and_align

entraîné sur un corpus multi-locuteurs et peut ainsi être utilisé pour aligner n'importe quelle langue ou n'importe quel style de parole. De plus, Train&Align implémente différentes options d'entraînement qui permettent d'améliorer la qualité de l'alignement. Contrairement à EasyAlign qui ne propose qu'une version sous Windows, notre outil est directement accessible en ligne et peut ainsi être utilisé sur tous les systèmes d'exploitation. Enfin, le système offre la possibilité d'évaluer automatiquement la qualité de l'alignment produit si une partie manuellement alignée du corpus est fournie. Il faut cependant noter que notre outil présuppose actuellement l'utilisation d'une transcription phonétique correcte et ne permet donc pas l'introduction de variantes phonétiques lors de l'alignement comme proposé par d'autres études et outils (Goldman, 2011; Boula de Mareuil *et al.*, 2008; Bigi, 2012).

L'objectif de cet article est de présenter ce nouvel outil d'alignement ainsi que les techniques sur lesquelles il repose. Une description de l'outil et de son fonctionnement est donnée en section 2. La section 3 présente le protocole expérimental établi pour évaluer les performances de Train&Align, suivi de la comparaison des taux d'alignement obtenus avec ceux atteints par les modèles fournis par d'autres outils existants. Enfin, la section 5 conclut l'article.

2 Présentation de Train&Align

Train&Align est un nouvel outil d'alignement phonétique automatique. Sa particularité est qu'il entraîne directement les modèles acoustiques sur le corpus à aligner, ce qui le rend applicable à toutes les langues et styles de parole. L'intérêt de cette technique a déjà été mise en évidence par de nombreuses recherches en reconnaissance vocale (Leggetter et Woodland, 1995) qui montrent que les modèles dépendant du locuteur permettent d'atteindre de meilleurs résultats.

Dans un premier temps, l'entièreté du corpus (non aligné) et sa transcription phonétique sont utilisés afin d'entraîner les modèles acoustiques de chaque phonème (monophones à 5 états) [1]. Notons que le modèle de silence repose sur une configuration particulière qui permet une modélisation plus adéquate de sa durée. Des modèles de courtes pauses ('sp') sont également automatiquement insérés entre les mots. Ils correspondent à des modèles de silence spécifiques et permettent la détection de silences non annotés dans la transcription phonétique. Ces modèles sont également implémentés dans EasyAlign et P2FA. SPPAS, au contraire, ne propose pas de modèle pour les silences qui sont traités dans une phase indépendante de l'alignement.

Les modèles acoustiques sont initialisés à l'aide d'un alignement uniforme du son et de sa transcription phonétique. Chaque phonème est ainsi modélisé par des paramètres moyens, communs à tous les phonèmes. Les paramètres sont ensuite adaptés itérativement par l'algorithme de Baum-Welsh (Young *et al.*, 1995). Dans un second temps, ces modèles sont utilisés pour aligner le corpus d'entraînement lui-même. HTK est utilisé pour l'entraînement et l'alignement.

Deux options principales sont proposées par Train&Align et permettent de modifier certains paramètres de l'entraînement des modèles.

Tout d'abord, la configuration des modèles peut être modifiée. Au lieu d'entraîner un modèle par phonème (monophone), il est possible d'entraîner un modèle par triphone, c'est-à-dire par séquence de trois phonèmes. L'utilisation de tels modèles nécessite cependant un corpus de

1. Les paramètres des modèles sont 12 Mel Frequency Cepstral Coefficients (MFCC) et un coefficient d'énergie, ainsi que leur première et seconde dérivée.

taille importante afin de modéliser correctement chaque triphone. De plus, l'augmentation du nombre de modèles s'accompagne d'une augmentation du temps de traitement. Afin de pallier ce problème, il est également possible d'utiliser des tied-state triphones. Dans ce cas, le contexte gauche et droit des triphones est remplacé par des classes phonétiques (fricatives, voyelles, plosives, etc.) ce qui réduit considérablement le nombre de modèles. Notons que la liste des classes phonétiques est ouverte et peut être définie librement pour chaque nouvelle langue ou nouveau style de parole à aligner. L'utilisation de triphones ou de tied-triphones est une option que peuvent plus difficilement offrir les outils reposant sur des modèles pré-entraînés. Idéalement, l'outil devrait alors proposer des modèles pour tous les triphones possibles de la langue, ce qui demande l'utilisation d'un corpus d'entraînement de taille très importante. Dans le cas contraire, l'alignement échoue lorsqu'il rencontre des triphones inconnus. Naturellement, ce problème ne survient pas lorsque l'on entraîne les modèles sur le corpus à aligner.

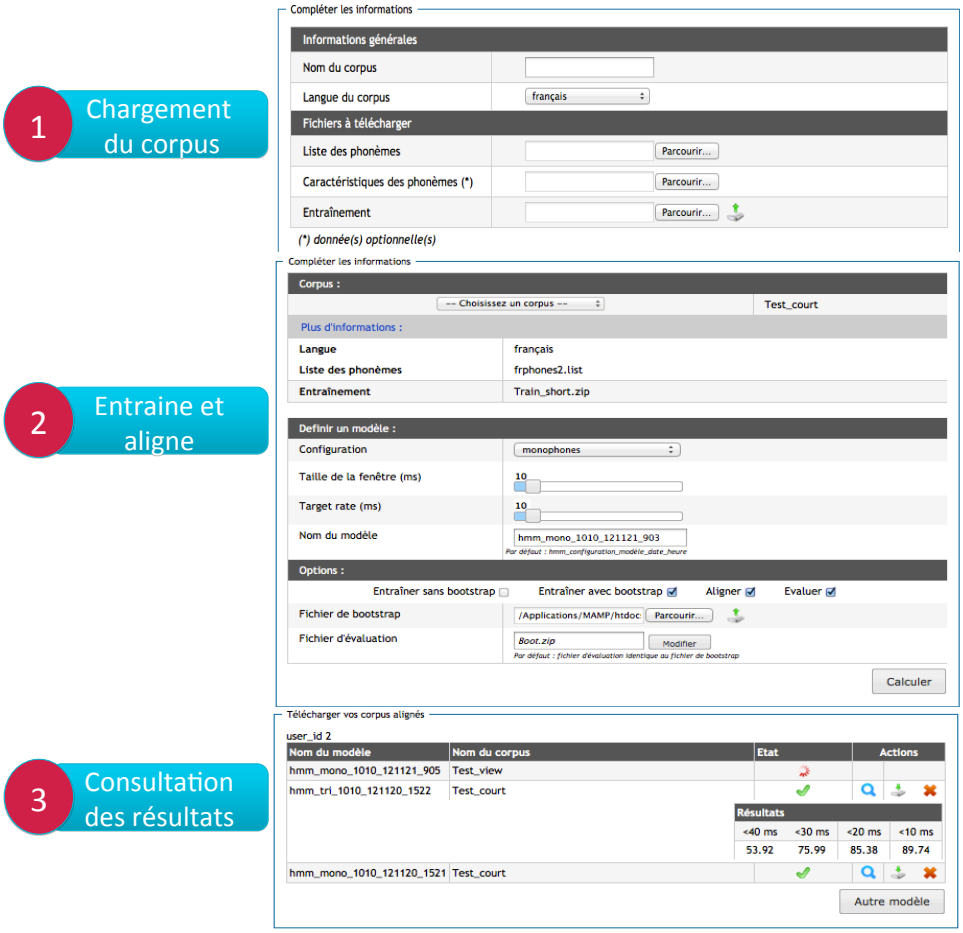


FIGURE 1 – Aperçu de Train&Align

Une seconde option est le bootstrap. Si une petite partie du corpus a été alignée manuellement, elle peut être exploitée afin d'améliorer l'initialisation des modèles et l'alignement qui en résulte. Dans ce cas, l'initialisation uniforme est remplacée par une procédure itérative qui détermine les paramètres d'après l'alignement manuel.

Train&Align permet également d'évaluer automatiquement la qualité de l'alignement. Si une partie du corpus alignée manuellement est disponible, elle peut être comparée à l'alignement automatique. On calcule ainsi le pourcentage de frontières de phonèmes qui sont similaires dans les deux alignements, avec un certain seuil de tolérance. Nous considérons donc la proportion de

frontières pour lesquelles l'erreur temporelle est inférieure à un seuil qui varie de 10 à 40 ms, par pas de 10 ms. Le bootstrap est associé à une évaluation particulière. Si le même sous-corpus est fourni pour l'évaluation et le bootstrap, une 'évaluation croisée à 5 plis' est réalisée : 4/5 du sous-corpus sont, tour à tour, exploités pour le bootstrap, le cinquième restant servant à l'évaluation. Les taux d'alignement correct obtenus à chaque itération sont ensuite moyennés. Il est également possible d'utiliser deux sous-corpus différents pour l'évaluation et le bootstrap.

Notons enfin que seul le son et sa transcription phonétique sont nécessaires pour entraîner un nouveau modèle et ainsi produire l'alignement du corpus. Cette transcription peut être fournie sous la forme d'un simple fichier .txt. Train&Align est également compatible avec le format Praat, acceptant des TextGrids en entrée et produisant des TextGrids contenant une couche phonétique alignée. Il permet aussi d'exclure de l'alignement des parties du TextGrid qui ne doivent pas être alignées ou n'ont pas été transcrites. L'interface de l'outil peut être observée en Figure 1.

3 Évaluation de Train&Align

3.1 Protocole expérimental

Nous proposons ici une évaluation de Train&Align sur différents corpus, de langues et styles de parole différents. Dans le cadre de cette étude, six corpus, manuellement alignés, sont utilisés :
– Pour le français, un corpus neutre lu de 110 minutes utilisé en synthèse vocale (*Marie* (Colotte et Beaufort, 2005)) et un corpus expressif spontané de commentaires sportifs de 15 minutes (*Sportic* (Brognaux *et al.*, 2013)) ;
– Pour l'anglais, un corpus neutre lu de 12 minutes utilisé en synthèse vocale (*Will*[3]) et un corpus expressif lu de 51 minutes composé de 5 émotions et 5 locuteurs (*Woggle* (Dellaert *et al.*, 1996)) ;
– Deux corpus neutres de langues peu dotées : le *féroïen*[3] (21 min) et le *gaélique*[4] (8 min).
Les corpus expressifs sont caractérisés par un degré important de variabilité.

En Section 3.2, l'alignement par Train&Align est comparé à celui obtenu en utilisant les modèles fournis par des outils d'alignement existants : P2FA (Yuan et Liberman, 2008) pour l'anglais, SPPAS (Bigi, 2012) pour le français et l'anglais et EasyAlign (Goldman, 2011) pour le français. Les outils existants ne fournissant pas de modèles pour le féroïen et le gaélique, ces corpus sont donc uniquement alignés par Train&Align. Les résultats obtenus en exploitant le contexte phonétique sont ensuite analysés en Section 3.3. Enfin, la Section 3.4 étudie l'amélioration obtenue lors de l'utilisation d'une partie manuellement alignée du corpus comme bootstrap.

Pour tous les corpus, la transcription phonétique correcte est fournie pour l'alignement.

3.2 Version de base

L'alignement des corpus par Train&Align (T&A) est ici comparé à celui obtenu par les modèles fournis par les outils existants. Les taux d'alignement sont calculés sur un corpus de chaque

3. gracieusement fourni par Acapela Group SA
4. gracieusement fourni par le Phonetics & Speech Laboratory, Trinity College, Dublin

langue (neutre pour le français et expressif pour l'anglais). Deux précisions méthodologiques doivent être faites. Tout d'abord, SPPAS fournit des modèles triphones pour l'anglais. Cela signifie que l'alignement à l'aide de ce modèle échoue lorsqu'il présente des triphones inexistants. Alors que le programme SPPAS a été développé afin de permettre de résoudre ce problème, nous avons décidé ici d'exclure toutes les phrases contenant des triphones inconnus et ce afin de permettre une comparaison des modèles entre eux. Seule une petite moitié des fichiers en anglais a donc pu être évaluée par les modèles de SPPAS. Il faut ensuite noter que P2FA fournit des modèles différents selon l'accent que reçoit la voyelle (non accentuée, accents primaire et secondaire). Afin de mesurer la pleine potentialité de ce modèle, notre corpus a donc également été annoté accentuellement pour l'alignement avec ce modèle seulement. Notons cependant que les résultats obtenus sans distinguer les différents niveaux d'accentuation avec le modèle de P2FA sont nettement inférieurs (diminution de plus de 6 % à 20 ms près).

TABLE 1 – Taux d'alignement (%) avec Train&Align et les modèles fournis par d'autres outils

	Modèle	Correct <10 ms	Correct <20 ms	Correct <30 ms	Correct <40 ms
Français (*Marie*)	SPPAS	43.04	68.91	81.5	88.63
	EasyAlign	52.54	77.54	87.72	92.11
	T&A	60.67	84.56	92.88	96.69
Anglais (*Woggle*)	T&A	43.44	63.9	78.34	87.18
	SPPAS	37.3	65.2	82.07	88.69
	P2FA	46.68	69.74	81.2	87.46
Féroïen	T&A	47.1	74.09	87.24	93.61
Gaélique	T&A	75	88.78	93.59	95.98

Les résultats sont présentés en Table 1. Pour le corpus en français, nous observons que Train&Align dépasse largement les performances des modèles des autres aligneurs. Le taux d'alignement obtenu est d'ailleurs relativement élevé et comparable aux taux d'accord interannotateurs rapportés par Goldman (2011) qui étaient d'environ 80 % à 20 ms près. En ce qui concerne le corpus en anglais, les résultats de Train&Align sont sensiblement comparables à ceux obtenus par le modèle de SPPAS avec une précision supérieure de 6 % à 10 ms près mais des taux légèrement inférieurs pour les seuils plus élevés. Seul P2FA permet d'atteindre de meilleurs résultats. Ceci s'explique notamment par son utilisation d'un modèle différent pour chaque type d'accent, ce qui permet une meilleure modélisation de la langue. De plus, les modèles utilisés par P2FA sont entraînés sur plus de 25 heures de parole alignées manuellement au mot. Afin de fournir de tels résultats pour chaque langue, un travail manuel important devrait donc être effectué, ce qui est économiquement peu réalisable. La qualité globalement plus faible de l'alignement, comparé au corpus français, est probablement dûe à la plus grande variabilité de ce corpus qui contient plusieurs émotions et locuteurs. Il nous faut tout de même noter que Train&Align offre des résultats relativement comparables à P2FA, surtout pour des seuils de tolérance élevés, bien qu'il ne nécessite pas de modèle pré-entraîné, ni d'annotation de l'accentuation. Enfin, les taux d'alignement correct obtenus pour le féroïen et le gaélique sont assez proches de ceux obtenus sur le corpus français, ce qui indique une bonne adaptation de notre outil à des langues peu dotées. Notons enfin que l'application de Train&Align sur un corpus en néerlandais lors d'une précédente étude nous a également permis d'obtenir de bons taux d'alignement (Michaux *et al.*, 2014).

Train&Align entraînant les modèles acoustiques directement sur le corpus à aligner, la taille de ce corpus joue un rôle évident dans la qualité des modèles. Les taux d'alignement pour le corpus en français sont ici assez élevés mais les tests sont effectués sur un corpus de plus de 100 minutes de parole. On peut donc s'interroger sur la qualité de l'alignement lorsque la taille du corpus est plus réduite. L'influence de la taille du corpus sur la qualité de l'alignement a été étudiée. Nous avons ainsi pu observer que les taux d'alignement restent relativement stables jusqu'à l'utilisation d'un corpus de 5 minutes. Pour des corpus neutres avec peu de variabilité (p. ex. *Marie*), un taux correct d'alignement peut être conservé jusqu'à 2 minutes. En deçà de ce seuil, la qualité de l'alignement décroit rapidement.

3.3 Considération du contexte phonétique

Comme mentionné précédemment, Train&Align offre différentes options d'entraînement. Une première modification possible concerne la configuration des modèles. Au lieu d'un modèle par phonème (mono), un contexte phonétique plus large peut être considéré. Les résultats obtenus lors de l'utilisation de triphones (tri) et de tied-state triphones (tied) sont présentés en Table 3. Ces options ne sont ici testées que sur les corpus anglais et français, langues pour lesquelles nous disposons d'une table des caractéristiques acoustiques de chaque phonème utilisé. On observe que l'utilisation de tels modèles permet d'atteindre, globalement, un meilleur alignement des corpus. Ainsi, pour *Sportic*, l'utilisation de tied-state triphones est recommandée et permet d'obtenir une augmentation de près de 3 % à 20 ms près. L'utilisation des tied-state triphones sur le corpus expressif anglais montre cependant une amélioration pour les seuils de 30 ms ou plus seulement. Il est donc important de sélectionner ou non cette option en fonction du corpus à aligner. Si une petite partie manuellement alignée du corpus est disponible, une évaluation avec les différentes options d'entraînement permettra de choisir la meilleure configuration.

TABLE 2 – Taux d'alignement (%) avec Train&Align-mono, -tri et -tied

	Français				Anglais			
	<10 ms	<20 ms	<30 ms	<40 ms	<10 ms	<20 ms	<30 ms	<40 ms
NEUTRE			*Marie*				*Will*	
mono	60.67	**84.56**	92.8	96.69	49.86	**76.85**	87.14	92.68
tri	62.7	85.44	92.56	96.53	49.05	75.95	86.81	91.99
tied	63.64	**86.05**	92.91	96.63	50.79	**78.27**	87.68	93.06
EXPRESSIF			*Sportic*				*Woggle*	
mono	51.75	**72.6**	83.63	89.23	43.44	**63.9**	78.34	**87.18**
tri	51.66	72.71	84.15	89.89	42.85	63.38	78.78	88.17
tied	54.88	**75.57**	85.97	91.26	42.43	**62.63**	78.37	**88.24**

3.4 Bootstrap

Une seconde option proposée par Train&Align est le bootstrap. Une petite partie du corpus alignée manuellement peut être exploitée afin de permettre une meilleure initialisation des modèles et ainsi une meilleure qualité de l'alignement. Nos tests montrent qu'un sous-corpus manuellement aligné de 30 secondes (hors silences initiaux et finaux) permet d'atteindre une nette amélioration du taux d'alignement (voir Table 3). L'alignement obtenu est ici évalué sur le même sous-corpus de 2 minutes avec et sans bootstrap afin d'éviter tout biais introduit par le

sous-corpus d'évaluation. Au delà de 30 secondes de bootstrap, la qualité de l'alignement reste majoritairement stable, certains corpus témoignant encore d'une augmentation avec un bootstrap de 1 minute à 2 minutes. Comme mentionné précédemment, il a été montré que l'alignement manuel nécessite environ 130 fois la durée du son (Kawai et Toda, 2004). Pour 30 secondes, une heure de travail manuel serait donc nécessaire, ce qui est économiquement faisable. Pour *Woggle*, le corpus en anglais expressif qui obtenait des taux d'alignement assez bas avec la version de base, on observe ainsi une augmentation de près de 20 %, ce qui lui permet d'atteindre une qualité d'alignement proche de celle du corpus neutre. Ceci semble montrer que l'utilisation de bootstrap est particulièrement indiquée pour l'alignement de corpus caractérisés par un haut degré de variabilité.

TABLE 3 – Taux d'alignement des corpus sans et avec 30 secondes de bootstrap

		<10 ms	<20 ms	<30 ms	<40 ms
Marie	mono	58.84	**81.69**	91.43	95.13
	boot (30 sec.)	62.65	**84.76**	93.97	96.08
Woggle	mono	44.58	**65.34**	78.33	87.18
	boot (30 sec.)	64.27	**85.11**	92.31	95.53
Féroïen	mono	45.95	**74.83**	87.86	94.51
	boot (30 sec.)	52.66	**79.10**	90.52	95.68
Gaélique	mono	75.31	**89.10**	94.29	96.39
	boot (30 sec.)	78.33	**91.60**	95.67	97.11

4 Conclusion

Cet article a présenté un nouvel outil d'alignement phonétique automatique disponible en ligne : Train&Align . Contrairement aux outils existants, il entraîne les modèles acoustiques directement sur le corpus à aligner, ce qui lui permet d'aligner n'importe quelle langue ou n'importe quel style de parole. Les tests montrent que Train&Align fournit des résultats comparables aux outils existants et permet également d'obtenir de bons taux d'alignement pour des langues peu dotées. De plus, il offre la possibilité de modifier certains paramètres d'alignement. Ainsi, il a été montré que l'utilisation de 30 secondes de bootstrap, permet d'atteindre une amélioration allant jusqu'à 20 % pour un corpus caractérisé par un haut degré de variabilité. Enfin, l'outil fournit une évaluation automatique de l'alignement si une partie manuellement alignée est disponible.

Remerciements

S. Brognaux et T. Drugman sont soutenus par le « FNRS ». S. Roekhaut est financée par la convention ARC 12/17-044. Les auteurs tiennent également à remercier J.-P. Goldman et B. Bigi pour leur aide lors de l'utilisation de leur outil et pour leur enthousiasme concernant cette étude.

Références

ADELL, J., BONAFONTE, A., GOMEZ, J. A. et CASTRO, M. J. (2005). Comparative study of automatic phone segmentation methods for TTS. *In ICASSP*.

BIGI, B. (2012). SPPAS : a tool for the phonetic segmentation of speech. *In LREC*.

BOERSMA, P. et WEENINK, D. (2009). Praat : doing phonetics by computer (version 5.1.05) [computer program].

Boula de MAREUIL, P., VIERU-DIMULESCU, B., WOEHRLING, C. et ADDA-DECKER, M. (2008). Accents trangers et rgionaux en franç. *Traitement Automatique des Langues*, 49 (3):135–163.

BROGNAUX, S., PICART, B. et DRUGMAN, T. (2013). A new prosody annotation protocol for live sports commentaries. *In Interspeech*.

COLOTTE, V. et BEAUFORT, R. (2005). Linguistic features weighting for a text-to-speech system without prosody model. *In Interspeech*.

DELLAERT, F., POLZIN, T. et WAIBEL, A. (1996). Recognizing emotion in speech. *In ICSLP*.

GOLDMAN, J.-P. (2011). EasyAlign : an automatic phonetic alignment tool under Praat. *In Interspeech*.

GOLDMAN, J.-P. et SCHWAB, S. (2011). Easyalign spanish : An (semi-) automatic segmentation tool under Praat. *In 5th CFE*.

KAWAI, H. et TODA, T. (2004). An evaluation of automatic phone segmentation for concatenative speech synthesis. *In ICASSP*.

LEGGETTER, C. et WOODLAND, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185.

MICHAUX, M.-C., BROGNAUX, S. et CHRISTODOULIDES, G. (2014). The production and perception of L1 and L2 Dutch stress. *In Speech Prosody*.

SCHIEL, F. et DRAXLER, C. (2003). The production of speech corpora. Rapport technique, Bavarian Archive for Speech Signals.

SCHMIDT, T. (2012). Exmaralda and the folk tools. *In LREC*.

van NIEKERK, D. et BARNARD, E. (2009). Phonetic alignment for speech synthesis in under-resourced languages. *In Interspeech*.

WITTENBURG, P., BRUGMAN, H., RUSSEL, A., KLASSMANN, A. et SLOETJES, H. (2006). Elan : a professional framework for multimodality research. *In LREC*.

YOUNG, S., KERSHAW, D., ODELL, J., OLLASON, D., VALTCHEV, V. et WOODLAND, P. (1995). *The HTK Book (for HTK Version 3)*. Cambridge University.

YUAN, J. et LIBERMAN, M. (2008). Speaker identification on the SCOTUS corpus. *In Acoustics '08*.

Aspects de la coordination pneumophonique dans la dysarthrie parkinsonienne

Léa Cottavoz[1], Silvain Gerbers[1], Bernard Teston[2], François Viallet[2] et Didier Demolin[1]
(1) Gipsa-lab, Université de Grenoble Alpes
11 rue des Mathématiques, Grenoble Campus BP46, F-38402 Saint Martin d'Hères
(2) Laboratoire Parole et Langage, Université de Provence
CNRS - Université de Provence, BP 80975, 13604 Aix-en-Provence
Lea.cottavoz@e.u-grenoble3.fr, Didier.demolin@gipsa-lab.grenoble-inp.fr

RESUME

Cette étude cherche à évaluer les phénomènes dysarthriques résultant de la relation entre l'énergie pulmonaire et l'activité phonatoire chez des locuteurs parkinsoniens. Les comportements articulatoires des sujets sont aussi considérés. Les données proviennent de 23 sujets produisant 6 répétitions de courtes phrases dans l'état ON (avec dopamine) et OFF (sans dopamine). Les données ont été enregistrées avec la station EVA2. Les résultats des mesures aérodynamiques montrent que la pression intra-orale (PIO) est affectée par des fermetures courtes ou incomplètes des lèvres. La valeur de la PIO est en effet plus faible ou diminue progressivement chez les parkinsoniens.

ABSTRACT

Aspects of the pneumophonic coordination in Parkinsonian dysarthy

This study tries to evaluate dysarthric phenomena resulting from the relation between pulmonic energy and phonation among Parkinsonian speakers. Prosodic and articulatory behaviour are also considered. Data come from 23 subjects producing 6 repetitions of small sentences in ON (with Dopamine) and OFF conditions (without Dopamine). Data were recorded with the EVA2 station. Results of aerodynamic measurements show that intra-oral pressure (PIO) is affected by short or incomplete lips closures. PIO is weaker or diminishing gradually with Parkinsonian speakers.

MOTS-CLES : Maladie de Parkinson, phonétique expérimentale, dysarthrie hypokinétique.
KEYWORDS: Parkinson's disease, experimental phonetics, hypokinetic dysarthria.

1. Introduction

L'évaluation des troubles de la parole dans la maladie de Parkinson s'effectue par une analyse perceptive qui est la principale méthode d'évaluation à l'heure actuelle. Comme le souligne Viallet (2003), cette méthode repose sur la seule perception du clinicien qui effectue une appréciation qualitative de la parole. Les auditeurs peuvent juger la hauteur, l'intensité ou encore la qualité vocale. Afin de permettre l'évaluation exhaustive de la maladie de parkinson (MP), les neurologues disposent de quelques tests consistant à chronométrer la réalisation de certaines tâches et d'échelles quantitatives. Il en existe de plusieurs types, comme les échelles de qualité de la voix PDQ 39 ou PDQ 8. L'échelle la plus couramment utilisée dans le monde est globale : c'est l'échelle *UPDRS (Unified Parkinson Disease Rating Scale)*. Cette échelle est multidimensionnelle, elle recouvre, en effet, toutes les situations cliniques qui peuvent être observées. L'UPDRS est également utile au suivi des patients et à l'évaluation de l'effet des interventions thérapeutiques et/ou pharmacologiques. La comparaison des états OFF (sans dopamine) et ON (avec dopamine)

permet d'évaluer la réponse éventuelle à la L-Dopa (les scores de la parole et du tremblement sont compris entre 1 et 4).

L'étude des effets de la maladie de Parkinson sur la parole a déjà été abordée dans de nombreuses publications (Duez (2005), Ghio& Pinto (2008), Locco (2004), Robert &Spezza (2005), Sarr et al. (2007), Viallet (2007). Certains paramètres de la production de la parole sont cependant rarement abordés dans ces études. Les paramètres aérodynamiques comme la pression intra-orale (PIO) et le débit d'air buccal (DAB) permettent, lorsqu'il est possible de les obtenir, d'avoir des données précieuses sur les capacités phonatoires des locuteurs Parkinsoniens.

L'étude présentée ci-après se situe dans cette problématique. En particulier, nous cherchons à savoir comment évolue la PIO pendant la production de courtes phrases, également programmées pour évaluer indirectement la pression sous-glottique. Cette dernière donne une estimation de l'énergie d'un locuteur Parkinsonien en production de la parole. Des études antérieures (Giovanni et al. 2000) ont prouvé que cette dernière peut être estimée au pic de pression de PIO. Ces données, combinées avec des mesures de débit d'air buccal (DAB), permettent, en outre, d'évaluer la tonicité de la fermeture labiale, dans le protocole spécifique de cette étude. Nous présentons ici les premiers résultats d'une étude en cours.

2. Méthodologie

2.1 Matériel

Le matériel de ce travail provient de la base de données du CHU et du Laboratoire Parole et Langage (LPL) d'Aix en Provence. Cette base de données est centrée sur la maladie de Parkinson et ses effets sur la parole. Les patients étaient enregistrés avec l'appareil EVA2 qui permet de mesurer et d'enregistrer les paramètres aérodynamiques et acoustiques du signal sonore. Il s'agit du débit d'air buccal (DAB), de la pression intra-orale (PIO) et du signal sonore qui sont enregistrés de manière synchrone.

2.2 Corpus et participants

Il était demandé aux patients de répéter 6 fois la petite phrase : " *Papa ne m'a pas parlé de bon papa* ". Un groupe de 23 locuteurs d'âge, de sexe et de degré d'atteinte différent ont participé aux enregistrements. Cette phrase contient plusieurs répétitions de la consonne occlusive bilabiale [p] qui permet d'obtenir une mesure de pression sous-glottique estimée au moment du pic de PIO (Giovanni et al. 2000). Chaque locuteur a été enregistré sous dopamine (ON) et sans dopamine (OFF). Lors de l'enregistrement des données aérodynamiques, les patients étaient assis, la bouche placée dans un masque, face aux transducteurs de l'appareil EVA2, à 5 cm du micro. Pour les données acoustiques le micro était placé à 30 cm de la bouche. L'enregistrement était précédé d'une discussion de quelques minutes avec le patient et de l'explication du déroulement de l'examen. L'examen proprement dit durait de 30 à 45 minutes. Les sujets parkinsoniens retenus pour la présente étude sont des sujets dénués de tout autre trouble neurologique ou psychologique. Ils font l'objet de troubles dysarthriques identifiables au moyen d'une étude perceptuelle. Ce sont des locuteurs francophones natifs, âgés de 52 à 81 ans et présentent, d'un point de vue perceptuel, une altération de la parole analysée par l'équipe de médecins et de chercheurs du laboratoire d'Aix en Provence (LPL et CHU). 10 sujets contrôles ont

effectués les enregistrements dans les mêmes conditions que les patients. L'ensemble des données a été traité avec le logiciel *phonédit*.

2.3 Mesures

Afin d'estimer les ressources d'énergie des locuteurs et la tonicité articulatoire de la fermeture bilabiale, trois paramètres aérodynamiques ont étés mesurés de manière synchrone avec le signal acoustique. Il s'agit du pic du maximum de la PIO, de la durée de la PIO mesurées pendant la fermeture bilabiale. La durée du DAB de la voyelle qui suit les occlusives bilabiales sourdes [p] a également été mesurée. La figure 1 montre un exemple d'une réalisation de la phrase *papa ne m'a pas parlé de bon papa*, où sont représentés le signal acoustique, la PIO et le DAB.

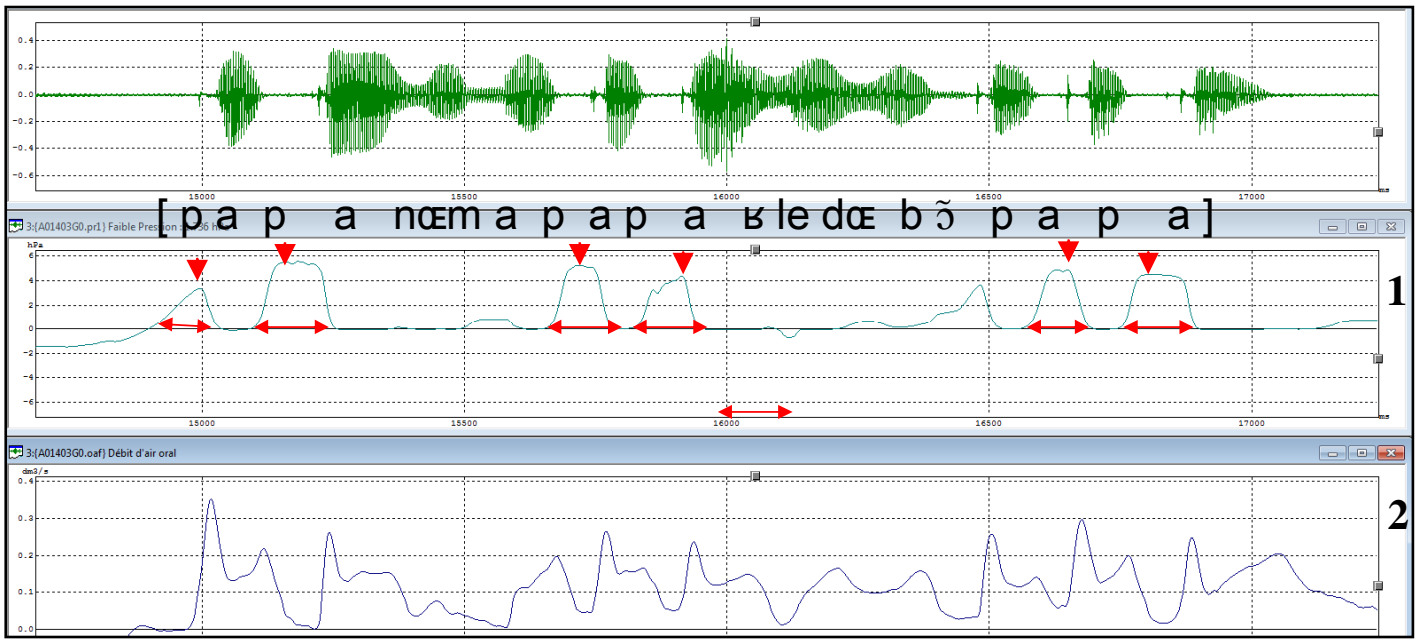


FIGURE1.Signal acoustique, tracé de PIO (1) et du DAB(2) pendant la phrase 'Papa ne m'a pas parlé de beau papa', [papanœmapapaʁledœbɔ̃papa] produite par un locuteur Parkinsonien. Les flèches sur le tracé de PIO indiquent les maxima. La durée de la PIO est indiquée par les doubles flèches sur les tracés de PIO.

2.4 Analyse statistique

L'objectif est d'étudier les variations d'une variable réponse et l'influence de deux facteurs (condition et consonne) sur celle-ci.

Dans notre expérience, nous recueillons plusieurs valeurs de la variable réponse pour un même sujet. Cette situation ne garantit plus l'indépendance des observations qui est une condition d'application des méthodes classiques d'analyse de la variance. Certaines différences sur la variable réponse peuvent donc avoir pour origine cette dépendance. C'est pourquoi nous choisissons d'utiliser le modèle linéaire à effets mixtes. C'est une extension du modèle linéaire dans lequel les effets aléatoires se rajoutent aux effets fixes. Les effets fixes sont les effets des facteurs et sont communs à l'ensemble des individus. Les effets aléatoires représentent quant à eux les effets des individus. Cette méthode permet alors d'être plus précis sur l'explication des variations de la variable réponse.

Les modèles mixtes reposent entre autre sur l'hypothèse selon laquelle les résidus suivent une loi normale. Pour permettre le respect des conditions d'applications des modèles mixtes, nous avons choisi de transformer la variable réponse en son logarithme.

$$\log Y_{ijkl} = \mu + \alpha_k + \beta_l + (\alpha\beta)_{kl} + \tau_i + \varepsilon_{ijkl}$$

avec:

Y_{ijkl} : j-ème valeur du sujet i prise par la variable réponse pour la modalité k du facteur A et l du facteur B

α_k : effet de la modalité k du facteur A sur la variable réponse

β_l : effet de la modalité l du facteur B sur la variable réponse

$(\alpha\beta)_{kl}$: effet de l'interaction de la modalité k du facteur A et de la modalité l du facteur B sur la variable réponse.

τ_i : effet du sujet i sur la variable réponse. C'est un effet aléatoire qui suit une loi normale, $\tau_i \sim \mathcal{N}(0, \sigma_\tau^2)$ avec σ_τ^2 qui représente la variance de la variable réponse entre sujets

ε_{ijkl} représente le terme d'erreur. Il suit une loi normale, $\varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Sur la base du modèle, nous souhaitons réaliser plusieurs comparaisons et regarder la différence entre les deux modalités du facteur A à l'intérieur de chaque modalité du facteur B. Pour cela, nous appliquons la méthode présentée dans l'article de Hothorn& al. (2008) et nous utilisons la fonction *glht* du package *multcomp* du logiciel R ainsi que la fonction *lsmeans* du package *lsmeans*.

Cette méthode nous permet de réaliser des comparaisons multiples de moyennes en tenant compte du fait que nous nous utilisons un modèle mixte**.** D'autre part, cette méthode garantit que le risque de première espèce lié à la prise simultanée de toutes les décisions ne dépasse un seuil que l'on se fixe à l'avance (généralement 5%) en ajustant les p-values.

3. Résultats

3.1 Pics de PIO

Les comparaisons multiples des durées moyennes entre les conditions ON et OFF pour chacune des consonnes relevées dans la phrase *papa ne m'a pas parlé de bon papa* montrent que les différences entre les phases ne sont pas toujours significatives (Table 1.Figures 2 et 4). La table 1 montre l'estimation des différences moyennes des pics de PIO (log) entre conditions ON et OFF pour la réalisation des six consonnes bilabiales [p] (C1, C2, C3, C4, C5, C6) avec écarts-types des différences, valeur de la statistique et p-value (l'hypothèse du test est une différence ON-OFF nulle).

	Estimate	Std. Error	z	value Pr(>|z|)
OFF - ON | C1 == 0	-0.24160	0.09159	-2.638	0.04901 *
OFF - ON | C2 == 0	-0.16123	0.09128	-1.766	0.38307
OFF - ON | C3 == 0	-0.22054	0.09128	-2.416	0.09054 .
OFF - ON | C4 == 0	-0.03910	0.09203	-0.425	0.99873
OFF - ON | C5 == 0	-0.30943	0.09128	-3.390	0.00419 **
OFF - ON | C6 == 0	-0.16700	0.09128	-1.829	0.34176

TABLE1. La première colonne du tableau indique l'hypothèse du test. La colonne « estimate » contient les estimations ponctuelles des différences de moyennes. La colonne «Std. Error » contient les écarts-type des différences de moyennes. La colonne « Z » donne la valeur de la statistique du test. La colonne « valuePr(>|z|) » donne la p-value du test. Si la p-value est inférieur à 0,05, on

rejette l'hypothèse du test et on considère que la différence entre les moyennes est significative.

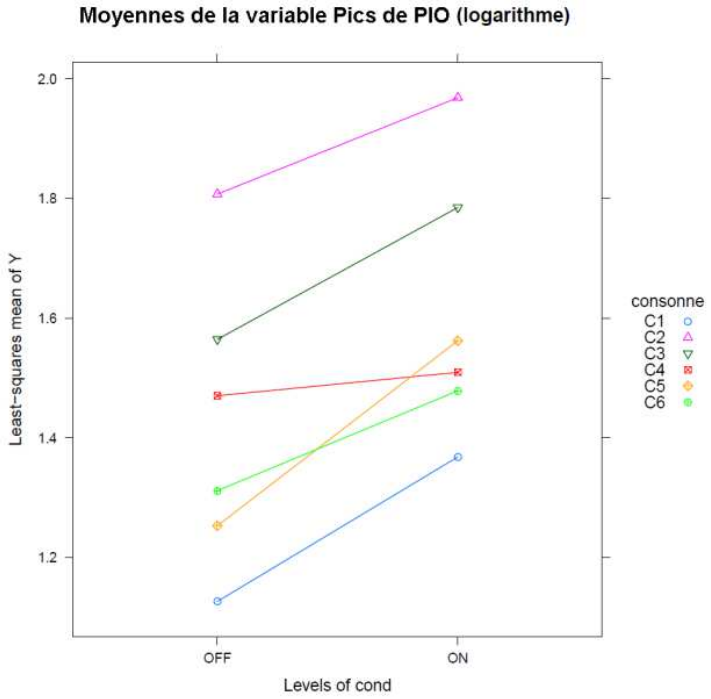


FIGURE2. Moyennes de la variable pics de PIO. En abscisse : condition ON et OFF. En ordonnée : la valeur de la variable estimée pour toute la population.

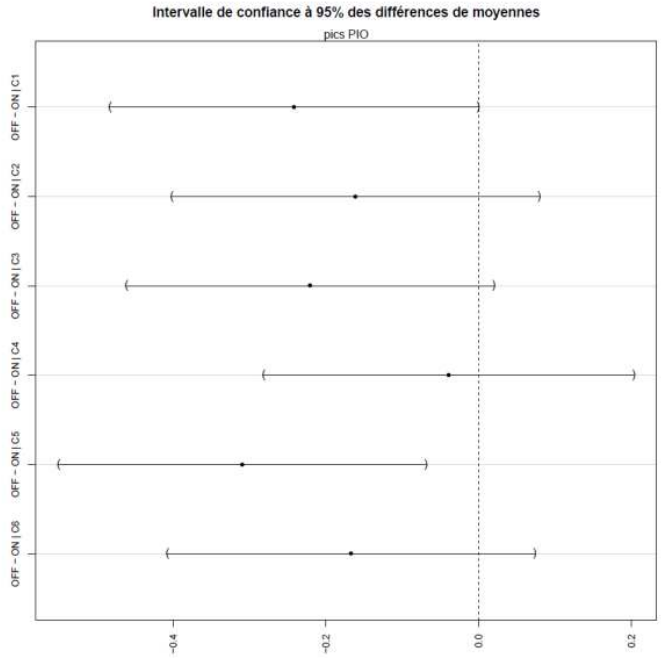


FIGURE3. Intervalle de confiance des différences moyennes. Rapport ON-OFF pour chaque consonne. Le point est une estimation de la différence observée. Les parenthèses sont les bornes de l'intervalle de confiance.

L'analyse statistique montre une différence entre la condition ON et OFF pour la réalisation des pics de PIO. Cette différence met en exergue l'effet de la MP sur la PIO et témoigne de l'impact de la L-dopa sur la production. Nous pouvons constater que la L-dopa provoque une augmentation des valeurs de pics de PIO. La différence entre la condition ON et OFF pour la réalisation des pics de PIO n'est pas significative pour toutes les consonnes. Seules les consonnes en première (C1) et en cinquième position (C5) dans la phrase observent une différence significative entre la condition ON et OFF. La figure 2 montre également que les consonnes en deuxième et troisième position (C2 et C3) dans la phrase obtiennent des valeurs plus élevées que les autres. Notamment par rapport aux consonnes C1 et C6 qui ont les valeurs les plus basses. On peut aussi constater une équivalence entre ON et OFF pour C4.

3.2 Durée de la PIO

	Estimate	Std. Error	z	value Pr(>|z|)
OFF - ON | C1 == 0	-0.19881	0.09449	-2.104	0.194
OFF - ON | C2 == 0	0.07721	0.09418	0.820	0.959
OFF - ON | C3 == 0	0.10150	0.09418	1.078	0.862
OFF - ON | C4 == 0	-0.13562	0.09497	-1.428	0.631
OFF - ON | C5 == 0	-0.13324	0.09872	-1.350	0.690
OFF - ON | C6 == 0	-0.22221	0.09872	-2.251	0.138

TABLE2. (cf. table 1 pour explication)

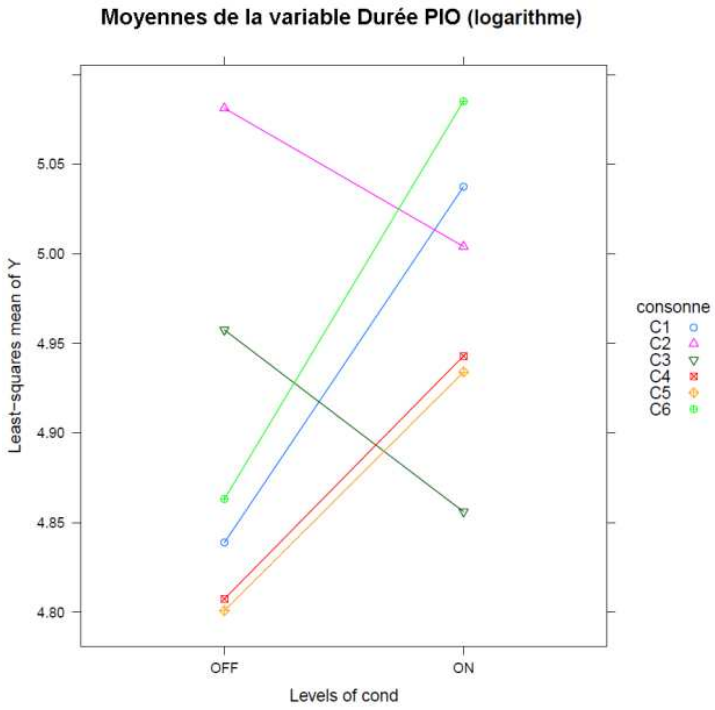


FIGURE 4. Moyennes de la variable durée de PIO. En abscisse : condition ON et OFF. En ordonnée : la valeur de la variable estimée pour toute la population.

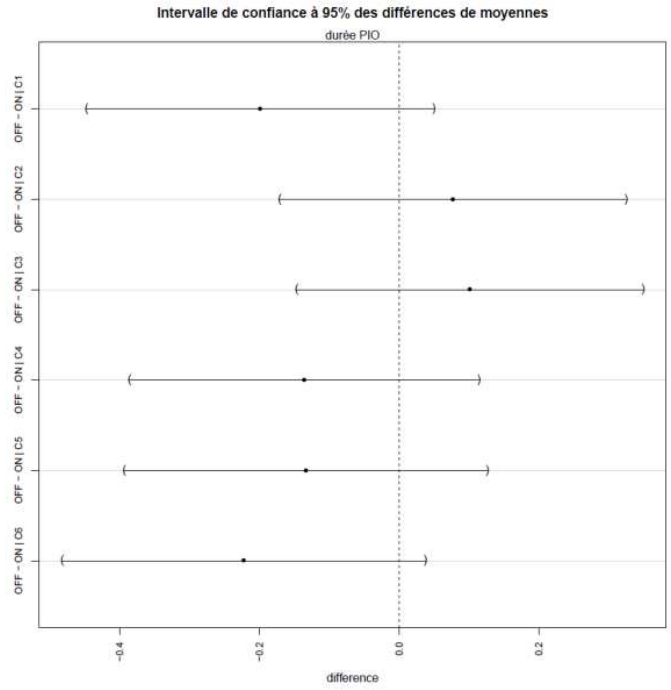


FIGURE5. Intervalle de confiance des différences moyennes. Rapport ON-OFF pour chaque consonne. Le point est une estimation de la différence observée. Les parenthèses sont les bornes de l'intervalle de confiance.

L'analyse statistique montre une différence entre la condition ON et OFF pour le maintien (durée) de PIO. Aucun résultat n'est significatif. Les consonnes C1, C4, C5 et C6 montrent une augmentation dans le maintien de PIO entre les conditions OFF et ON. Cependant, les consonnes C2 et C3 sont plus élevées en OFF qu'en ON. Egalement, nous remarquons que les consonnes ayant les valeurs les plus élevées en condition ON sont C6 et C1.

3. 3 Durée du DAB

L'analyse statistique montre que les valeurs des consonnes C1, C2 et C3 sont en régression (de manière non significative) entre la condition OFF et ON alors que les valeurs de consonnes C4 et C6 sont en augmentation (non significatif) entre les deux conditions. Nous remarquons aussi une quasi-équivalence entre les valeurs de C5 en OFF et ON.

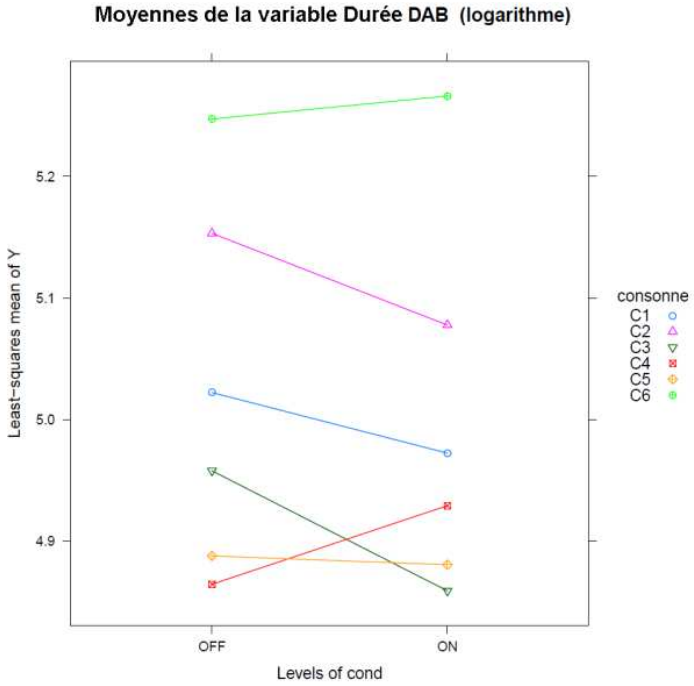


FIGURE6. Moyennes de la variable durée du DAB. En abscisse : condition ON et OFF. En ordonnée : la valeur de la variable estimée pour toute la population.

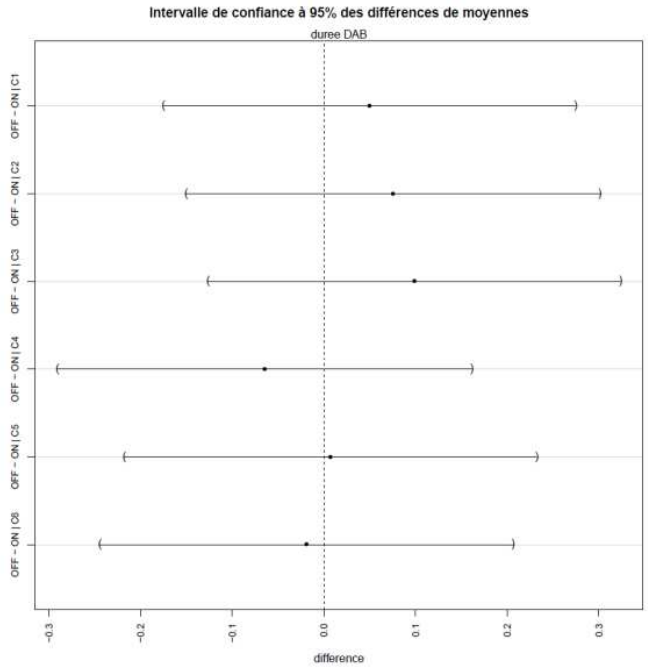


FIGURE7. Intervalle de confiance des différences moyennes. Rapport ON-OFF pour chaque consonne. Le point est une estimation de la différence observée. Les parenthèses sont les bornes de l'intervalle de confiance.

3.4 Durée du DAB et de la PIO

La Figure 7 montre une comparaison entre la durée moyenne de la pression intra-orale (PIO) et la durée moyenne du débit d'air buccal (DAB) en condition ON. Nous pouvons observer que l'ordre des consonnes est préservé malgré une inversion entre C1 et C2. Il est important de souligner qu'il n'y a de relation proportionnelle entre la durée moyenne de la PIO et celle du DAB. Néanmoins, le rang de chaque consonne reste équivalent.

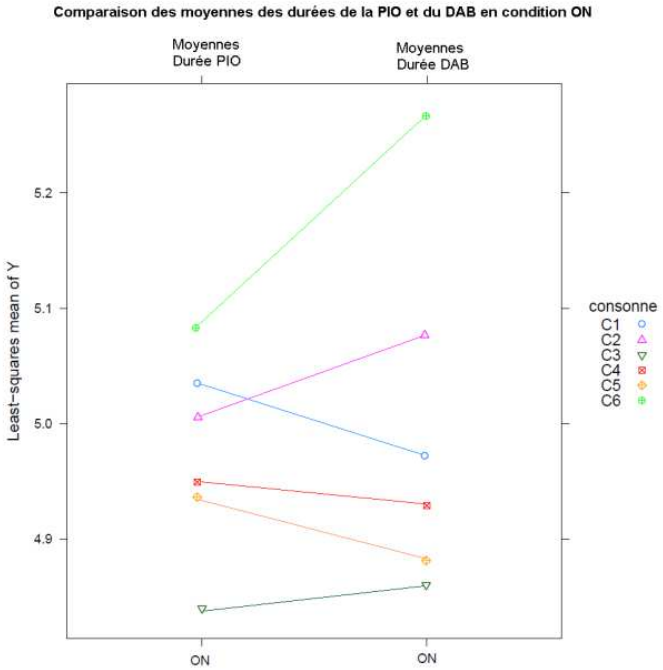


FIGURE8. Moyennes des variables durée de PIO et durée de DAB. En abscisse : condition ON et OFF. En ordonnée : la valeur (Log) de la variable estimée pour toute la population.

Ces résultats nous permettent de formuler des hypothèses concernant la réalisation des

consonnes. Etant donné que les données des sujets contrôles sont en cours de réalisation, nous nous focaliserons, pour le moment, sur l'ordre des valeurs de chaque consonne.

4. Discussion.

L'objectif de cette étude est de montrer les effets de la L-dopa sur la parole dysarthrique de patients parkinsoniens. Les résultats présentés ci-dessus permettent de formuler des hypothèses concernant le degré de contrôle articulatoire qu'un sujet acquiert par rapport à l'énergie et à la force avec le traitement. Nous postulons que la L-dopa ne permet pas un meilleur contrôle articulatoire de la fermeture labiale. La comparaison des figures 2 et 4 nous permet de déduire un effet des pics de PIO sur la durée de la PIO. L'ordre des consonnes ON de la figure 2 s'inverse sur la figure 4, ce qui suggère la relation suivante : plus le pic de PIO d'une consonne est élevé en ON, plus la durée de la PIO de cette consonne en ON sera court. Par exemple, les consonnes C3 et C2 ont en moyenne des pics de PIO plus élevés en ON et obtiennent les valeurs faibles au niveau de la durée du DAB. A l'inverse, les consonnes C1 et C6 ont en moyenne des pics de PIO plus faibles en ON et obtiennent les valeurs élevées au niveau de la durée du DAB. Cette comparaison nous permet d'observer qu'il n'y a pas de corrélation entre les pics de PIO en OFF et la durée de la PIO en OFF. Ceci suggère que la maladie de Parkinson entraîne une fluctuation, une instabilité entre les pics de PIO et la durée de la PIO. Cette instabilité serait rectifiée par la prise de L-dopa sans indiquer une amélioration de la réalisation. La durée du DAB suppose une corrélation avec la durée de la PIO en condition ON. Cette corrélation ne se vérifie pas en OFF (figure 8). Malgré la variation des valeurs, l'ordre des consonnes est préservé traduisant, l'influence suivante : plus la durée de la PIO est longue, plus la durée du DAB sera longue. De manière générale, il y a une augmentation des valeurs entre l'état OFF et ON mais cette augmentation ne révèle pas une amélioration, c'est-à-dire une meilleure réalisation de la cible articulatoire.

5. Conclusion

Les résultats préliminaires que nous obtenons vont dans le sens d'un effet de la L-dopa sur la parole des locuteurs parkinsoniens. Il est possible de montrer qu'il y a une augmentation de l'énergie globale sous L-dopa, dans la mesure où les pics de PIO mesurés (qui donnent une valeur estimée de la pression sous-glottique) sont plus élevés dans la condition ON par rapport à la condition OFF. Ceci ne montre cependant pas que les locuteurs parkinsoniens ont un meilleur contrôle des articulations en condition ON, par exemple celle de la fermeture des lèvres. Pour mieux évaluer ces aspects il sera nécessaire de comparer ces données avec des sujets contrôle et d'effectuer des corrélations précises entre les paramètres aérodynamiques et les paramètres acoustiques.

Références

Berendse H. W, Ponsen M. M. (2009). *Diagnosing premotor Parkinson's disease using a two-step approach combining olfactory testing and DAT SPECT imaging*, Parkinsonism Relat.Disord., vol. 15 Suppl 3. 26-30.

Defebvre, L. (2007). *La maladie de Parkinson et les syndromes parkinsoniens*, Les dysarthries, Solal, Marseille.

Duez, D. (2005). *Organisation temporelle de la parole et dysarthrie parkinsonienne,* In Özsancak C, Auzou P (ed). Les troubles de la parole et de la déglutition dans la maladie de Parkinson. Solal. Marseille. 195-211.

Giovanni, A., Demolin, D., Heim, C. and Triglia, J.-M. (2000). Estimated subglottic pressure in normal and dysphonic subjects. *Ann OtolRhinolLaryngol,* 109. 500-504.

Ghio, A., Pinto S. (2008). *Troubles du contrôle moteur de la parole : contribution de l'étude des dysarthries et dysphonies à la compréhension de la parole normale.*

Ho, A.K., Iansek, R., Bradshaw, J.L. (1999). *Regulation of parkinsonian speech volume: the effect of interlocuter distance.* J NeurolNeurosurgPsychiatry. 199-202.

Hothorn, F., Bertz, F. & Westfall, P.(2008), *Simultaneous inference in general parametric models*, Biometrical journal 50(3), 346-363.

Jankowski, L., Purson, A., Teston, B., Viallet, F. (2004). *Effets de la L-Dopasur la dysprosodie et le fonctionnement laryngien de patients parkinsoniens*. Actes des XXVèmes journées d'études sur la parole, Fès, Maroc

Lagrue, B., Meynadier, Y., Mignard, P. (1998). *Voix et maladie de Parkinson : Étude de la hauteur, de l'étendue et de la dynamique tonale*, In: S. Santi, I. Guaitella, C. Cavé, G. Konopczynski (Eds), Oralité et gestualité. Communication multimodale, interaction. Actes du colloque Orange 98, L'Harmattan, Paris. 235-240.

Locco, J. (2004). *Aspects aérodynamiques de la production des occlusives dans la maladie de Parkinson.* Actes des XXVèmes journées d'études sur la parole, Fès, Maroc.

Robert, D., Spezza, C. (2005). *La dysphonie parkinsonienne et les troubles articulatoires dans la dysarthrie parkinsonienne,* In Özsancak C, Auzou P (eds). Les troubles de la parole et de la déglutition dans la maladie de Parkinson. Solal. Marseille. 131-159.

Sarr, MM., Viallet, F., Testoon, B., Jankowski, L. Et Purson, A. (2007). *Effets de la stimulation du noyau subthalamique sur la production de parole dans la maladie de parkinson : Etude des coordinations pneumophoniques*, 2JPC, Grenoble.

Viallet, F. (2007). *La dysarthrie dans la maladie de Parkinson*, Les dysarthries, Solal, Marseille.

Comment évaluer la qualité des transcriptions automatiques pour la détection d'entités nommées ?

Mohamed Ameur Ben Jannet[α,β,γ,δ] Martine Adda-Decker[δ,α] Olivier Galibert[γ]
Juliette Kahn[γ] Sophie Rosset[α]
[α]LIMSI–CNRS [β]Université Paris-Sud [γ]LNE
[δ]LPP, Université Sorbonne Nouvelle
{first.last}@limsi.fr
{first.last}@lne.fr

RÉSUMÉ

La métrique standard pour l'évaluation des performances des systèmes de reconnaisance automatique de la parole est le taux d'erreur mot (Word Error Rate). Cette métrique est très efficace quand il s'agit d'évaluer les systèmes de transcription seules, mais de nos jours les systèmes RAP sont très souvent combinés avec d'autres briques technologiques de Traitement Automatique des Langues telles que la traduction de la parole, le dialogue homme-machine, ou l'extraction d'informations. Cette situation exacerbe le besoin exprimé par la communauté pour une métrique qui permet d'estimer la qualité des transcriptions automatiques en tenant compte du contexte applicatif. Nous proposons une métrique alternative au WER qui permet d'évaluer la qualité des transcriptions automatiques dans un contexte d'extraction d'entités nommées à partir de la parole. Notre métrique utilise des probabilités a posteriori pour estimer le risque d'erreur induit par les erreurs de transcription sur un système de détection d'entités nommées appliqué en aval.

ABSTRACT

How to assess the quality of automatic transcriptions for the extraction of named entities ?

The standard metric used to evaluate the performance of Automatic Speech Recognition (ASR) systems is the Word Error Rate (WER). This well known metric is very useful when ASR systems are used in isolation, but nowadays these systems are often combined with other natural language processing (NLP) systems to perform tasks like speech translation, man-machine dialogue, or information retrieval from speech. That situation exacerbates the need expressed by the speech processing community for a new evaluation metric which can estimate the quality of automatic transcription taking into account their applicative context. We propose an alternative evaluation metric to the WER in the context of Named Entity detection in spoken documents. Our proposed metric makes use of posterior probability to estimate the risk of error that ASR errors can induce to the Named Entity detection system applied downstream.

MOTS-CLÉS : Métrique, Évaluation, WER, ETER, Reconnaissance de la Parole, Extraction d'Entités Nommées.

KEYWORDS: Metrics, Evaluation, WER, ETER, Automatic Speech Recognition, Named Entities Recognition.

1 Introduction

Les performances des systèmes de Reconnaissance Automatique de la Parole (RAP) sont mesurées en taux d'erreur de mot : WER (Word Error Rate). Cette métrique est utilisée depuis 1981 (Pallett, 2003), l'époque où la transcription automatique de la parole était une fin en soi. Aujourd'hui la RAP est arrivée à maturité et des nouvelles perspectives d'utilisation sont envisageables. La communauté de traitement automatique des langues au sens large ne s'intéresse plus seulement à reconnaître et transcrire de la parole, mais vise aujourd'hui à exploiter la RAP comme brique technologique pour extraire de l'information à partir de la parole, pour traduire de la parole, pour résumer automatiquement des documents audio... L'extraction d'entités nommées à partir de documents transcrits automatiquement a fait l'objet de nombreuses campagnes d'évaluations internationales dans des langues différentes. Pour le français, cette tâche a été adoptée durant les évaluations ESTER2 (Galliano *et al.*, 2009), ETAPE (Galibert *et al.*, 2014) et QUAERO (Galibert *et al.*, 2011). Dans cet article nous exploitons les résultats de la campagne ETAPE pour étudier l'évolution des performances des systèmes de détection d'entités nommées (NER) en fonction du WER et nous proposons une mesure alternative au WER qui permet d'estimer le risque de dégradation de performance que peuvent causer les erreurs d'un système de RAP à un système de NER traitant des transcriptions automatiques.

La section suivante présente la problématique. Le modèle que nous proposons est décrit à la section 3 tandis que les corpus et expériences sont décrits dans les sections 4 et 5. Enfin la section 6 présente conclusions et perspectives.

2 Etat de l'art et problématique

Pendant longtemps, les travaux en traitement automatique de la parole se sont concentrés sur la réduction des taux d'erreur des systèmes RAP. Les progrès réalisés dans ce domaine ont abouti à la construction de systèmes plus robustes aux variations des locuteurs et pouvant traiter de la parole continue à grand vocabulaire. Toutefois, malgré les progrès, les erreurs de transcriptions persistent et elles sont plus ou moins gênantes selon le cadre applicatif.

Les travaux de (Munteanu *et al.*, 2006) sur l'archivage de données web à partir de sur leurs transcriptions automatiques ont montré qu'une augmentation importante du taux d'erreur (WER) se traduit par une chute de performances des systèmes. Par contre, d'autres travaux sur la recherche d'information à partir de la parole (Garofolo *et al.*, 2000) et sur l'analyse sémantique de documents audio (Wang *et al.*, 2003) et (Favre *et al.*, 2013) déplorent un désaccord entre le WER et les performances qu'ils obtiennent sur les sorties des systèmes RAP. Ces auteurs souhaiteraient disposer de mesures plus adaptées à leurs besoin permettant une meilleure prédiction de la dégradation des performances induites par les erreurs des transcripteurs automatiques. En effet, les erreurs de transcription peuvent avoir un impact important si elle modifient des informations pertinentes pour la tâche globale. En revanche, certaines erreurs ont une influence marginale et sont sans conséquences pour le résultat final.

Dans cet article nous étudions l'impact de la qualité des transcriptions sur la détection d'entités nommées dans des documents audio. Cette tâche a fait l'objet de nombreuses campagnes d'évaluation, dont récemment la campagne ETAPE. Un des objectifs était de détecter les entités nommées sur des transcriptions manuelles (références) et aussi des transcriptions automatiques

(hypothèses) produites par différents systèmes RAP. Les résultats de l'évaluation ETAPE (voir figure 1) montrent les performances obtenues par les différents systèmes NER mesurées en ETER (Entity Tree Error Rate) (Ben Jannet *et al.*, 2014) sur des transcriptions automatiques.

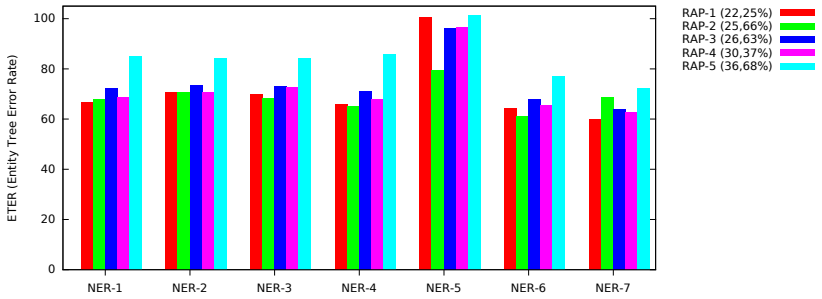


FIGURE 1 – évolution des erreurs de détection d'entités nommées mesurées en ETER (Entity Tree Error Rate) en fonction du WER, résultats de la campagnes ETAPE.

L'évolution des performances des systèmes de NER en fonction du WER montrent que ces derniers n'obtiennent pas forcément leurs meilleurs résultats sur les transcriptions ayant le WER le moins élevé.

Ceci met en évidence le besoin émergeant de mettre en place des procédures d'évaluation qui tiennent compte du cadre applicatif dans lequel les systèmes de RAP sont utilisés. Dans cet article, nous proposons une nouvelle approche permettant d'estimer la qualité des sorties RAP pour la détection d'entités nommées.

3 Modèle et proposition

Nous proposons ici une nouvelle méthodologie pour évaluer la qualité des transcriptions automatiques (TA) dans le contexte d'extraction d'entités nommées à partir de documents transcrits automatiquement. Les erreurs de transcription des systèmes RAP déforment les observations en entrée des systèmes NER en modifiant des mots et/ou leurs contextes d'apparition, provoquant ainsi des erreurs supplémentaires. Les systèmes NER visent à détecter et classifier les entités nommées qui se trouvent dans ces transcriptions bruitées en utilisant différentes méthodologies et en se basant sur divers critères. Les critères peuvent être des observations de n-gramme de mots, de POS, des dictionnaire de noms propres ou autres. Il est diffcile de se prononcer sur l'impact de la modification d'un ou plusieurs critères puisque la décision est prise par rapport à une observation globale.

L'existence de systèmes de NER statistiques raisonnablement performants montre qu'il est possible de construire des modèles probabilistes de présence d'entités nommées à un endroit donné d'un texte. Cela nous mène à l'hypothèse sous-jacente à notre travail : *il est possible de mesurer l'impact des erreurs de transcription sur les systèmes NER en mesurant le changement de probabilité a posteriori de présence des entités nommées*.

La première étape est de construire un modèle de présence des entités dans le texte. Nous avons choisi l'approche standard du modèle à maximum d'entropie, que nous avons entraîné sur de grandes quantités de données transcrites manuellement et annotées en entités nommées (cf.

section 4).

Nous pouvons ensuite utiliser notre modèle pour calculer les probabilités de présence d'entité sur les transcriptions automatiques et sur les transcriptions manuelles et faire une comparaison afin d'estimer l'impact des erreurs de transcription sur la détection de la bonne réponse.

Nous avons mis en place et comparé : calculer la différence des probabilités de la réponse recherchée ($\hat{y}$) obtenues sur transcription manuelle et transcription automatique (équation 1), ou calculer la différence des marges. La marge (voir équation 2) est définie comme étant l'écart entre la probabilité que le mot a de recevoir l'étiquette de référence ($\hat{y}$) et l'étiquette incorrecte la plus probable ($y' \neq \hat{y}$). Plus la marge est élevée, plus la réponse recherchée est mise en avant par rapport aux autres possibles. Plus la marge est faible, plus la discrimination de la bonne réponse est difficile.

$$Delta_{Prob} = P(\hat{y}|TA) - P(\hat{y}|TM) \quad (1)$$

$$Marge = P(\hat{y}|T) - max_{y' \neq \hat{y}}(P(y'|T)) \;\; avec \; y' \in Y \;\; T = les \; transcriptions \quad (2)$$

Calculer la différence entre la marge calculée sur TM ($Marge_M$) et celle calculée sur TA ($Marge_A$) nous permet d'avoir une estimation des difficultés de discriminer la bonne réponse introduite par les erreurs de transcription (voir équation 3).

$$Delta_{Marge} = Marge_A - Marge_M \quad (3)$$

4 Corpus

4.1 Description des données

Nous avons utilisé les données de la campagne d'évaluation ETAPE et en partie celles de la campagne d'évaluation QUAERO. Le tableau 1 contient une description des corpus en terme de nombre de mots et d'entités nommées.

Les données ETAPE sont constituées de 41 heures et 20 minutes d'enregistrement d'émissions radiophoniques et télévisées, sélectionnées de manière à contenir principalement de la parole non planifiée et une portion raisonnable de données multi-locuteur (1,5 heures). Le jeu d'entraînement ETAPE comporte 33 heures, alors que le jeu de test comporte 8 heures et 20 minutes (Gravier *et al.*, 2012). Nous avons également utilisé le jeu d'entraînement du corpus QUAERO qui est constitué de 100 heures d'émissions radiophoniques et télévisées de type Broadcast News et Broadcast Conversations provenant de la campagne ESTER 2 (Galibert *et al.*, 2011; Galliano *et al.*, 2009).

Toutes les données ont été transcrites manuellement et annotées en entités nommées par des experts. Les annotations ont été réalisées conformément au guide d'annotation développé durant le projet QUAERO (Rosset *et al.*, 2011).

	Train-ETAPE	Train-QUAERO	Test-ETAPE
Nombre de mots	335 387	1 251 586	115 803
Nombre d'entités	19 270	113 885	5 933
Nombre de composants	27 656	146 405	8 609

TABLE 1 – Description des corpus ETAPE (apprentissage et test) et QUAERO (apprentissage) en terme de nombre de mots et d'entités/composants.

4.2 Annotation en entités nommées

Nous présentons ici brièvement la structure et les classes d'entités telles que définies dans le guide d'annotation QUAERO (voir (Rosset *et al.*, 2011) pour plus de détails). Les entités nommées QUAERO possèdent une structure hiérarchique comportant 7 types et 32 sous-types qui permettent de spécifier le type général. Les types généraux sont relativement classiques et couvrent les *personnes*, les *lieux*, les *organisation*, les *montants* et *expressions temporelles* ainsi que les *productions* et les *professions et fonctions*. La définition implique également que chaque mot composant l'entité soit annoté à l'aide d'un composant qui peut être transverse, c'est à dire commun à différentes entités (comme *name* et *kind*) ou spécifique à un type donné (comme les composants *name.first* et *name.last* qui sont spécifiques aux entités de type *personne*). Les prépositions et déterminants qui sont intégrés dans une entités ne sont annotés en composants. Les entités peuvent être inclusives et les phénomènes de métonymie et d'antonomase sont également annotés. La figure 2 illustre l'application de ce schéma d'annotation.

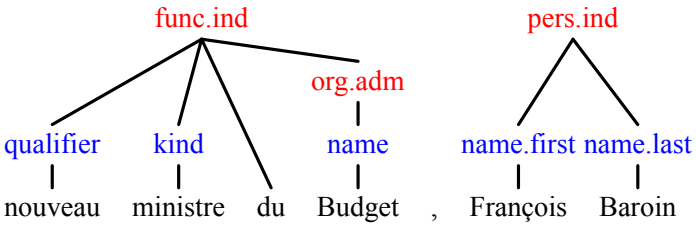


FIGURE 2 – Exemple d'annotation en composants (en bleu) et en types (en rouge).

5 Expériences et résultats

5.1 Expériences

Pour suivre l'approche proposée section 3 nous devons construire un modèle probabiliste de présence d'entités nommées, projeter les entités de la référence sur la transcription automatique, et enfin comparer les probabilités de présence obtenues.

Le type de modèle choisi est un modèle à maximum d'entropie, s'appuyant sur des traits classiques :

— Les mots et bigrammes de mots se trouvant dans une fenêtre locale de [-2,+2] autour du mot.

— Les préfixes et suffixes des mots se trouvant dans la fenêtre [-1,1] autour du mot.

— Des tests portant sur la présence d'une majuscule, d'un caractère numérique ou d'un signe de ponctuation dans la fenêtre [-1,1].

Ce modèle MaxEnt a été construit avec l'outil Wapiti (Lavergne *et al.*, 2010) sur les corpus d'entraînement QUAERO et ETAPE (voir section 4).

La projection suit et étend l'approche définie dans (Galibert *et al.*, 2011). La position temporelle des mots de la référence est obtenue par alignement forcé, et de là est déduite la position temporelle des frontières des entités. Un intervalle autour de ces instants donne une courte liste de mots possibles comme début et fin de chaque entité. Enfin un choix est fait dans ces listes en retenant dans toutes les possibilités celle diminuant la distance phonétique entre les deux chaînes. La phonétisation est obtenue par simple référence à un dictionnaire de prononciation.

Il est important de noter que cette projection ne donne comme position précise que le début et la fin des entités. Les autres mots ne sont pas mis en relation. De cela découle notre choix de construire deux modèles : un modèle de début d'entité, s'appliquant sur le premier mot de l'entité, et un modèle de fin s'appliquant sur le dernier. Les labels prédits ont la forme B-*type* (B : beginning) respectivement E-*type* (E : ending) et NULL hors entité.

À partir de là, on peut comparer les probabilités obtenues sur transcription manuelle et automatique. Le processus global est montré figure 3. Nous notons ici que cette expérience nous permet de mesurer seulement la probabilité d'augmentation d'erreurs de délétions et de substitutions d'entités nommées, puisque les probabilités sont calculée uniquement sur les frontières de début et de fin d'entité.

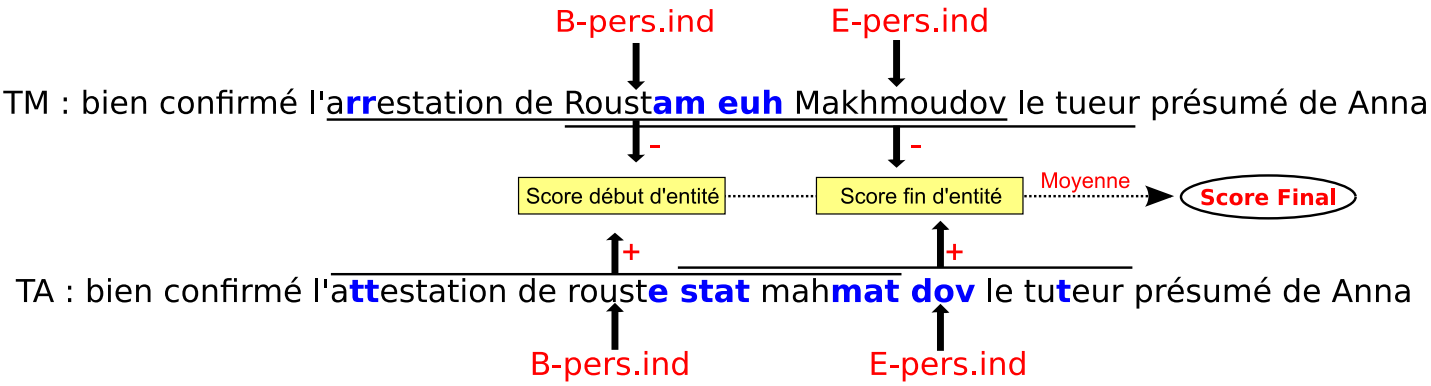


FIGURE 3 – Processus général du calcul du score.

Les modèles permettent de calculer la probabilité et/ou la marge pour l'étiquette attendue en début et fin d'entité sur transcriptions manuelles et automatiques. La différence automatique/manuel donne les scores de début et fin d'entité. Celui-ci est multiplié par -100 pour des raisons de lisibilité. Plus le score (positif) est élevé, plus l'impact des erreurs de reconnaissance sur les systèmes NER est négatif.

5.2 Résultats

Nous souhaitons observer la capacité de notre mesure à prédire la baisse de performance pour les systèmes NER induite par les erreurs de RAP. Afin de la mesurer nous faisons la différence entre les taux d'erreurs obtenus sur les sorties RAP et celles obtenus sur les transcription manuelles

(ETER sur TA - ETER sur TM). Le tableau 2 présente les mesures de la qualité des différentes sorties RAP en terme de WER, différences des marges et différences de probabilités. Nous nous intéressons à comparer la capacité de chacune des trois mesures à estimer l'augmentation des taux d'erreurs pour les systèmes NER quand il sont appliqués sur chacune des sortie RAP. Les pertes de performance des systèmes NER sont affichées dans la partie basse du tableau 2.

	Les Transcriptions					
Mesure qualité de transcriptions	Manuel	RAP-1	RAP-2	RAP-3	RAP-4	RAP-5
WER	0	22,3	25,7	26,6	30,4	36,7
Score différence des marges	-	31,8	25,7	35,1	35,7	37,0
Score différence de probabilité	-	17,0	10,1	18,4	19,0	20,7
	ETER_M	Perte ETER_A - ETER_M				
NER-1	37,7	27,2	27,6	31,5	29,1	45,2
NER-2	33,8	31,1	30,9	33,2	31,6	45,1
NER-3	33,9	23,0	21,8	25,6	26,3	37,9
NER-4	36,1	26,7	24,2	29,1	28,3	45,3
NER-5	51,3	47,4	24,5	40,8	42,9	46,8
NER-6	32,9	28,5	24,9	30,2	29,6	41,1
NER-7	37,9	19,5	27,6	22,5	22,2	31,4

TABLE 2 – Perte des performances (ETER sur TA - ETER sur TM) en fonction du WER, des scores différences des marges et des différences de probabilités pour les systèmes NER, calculées sur le corpus test de la campagne ETAPE.

Nous pouvons remarquer que les moyennes des différences de marges et des différences de probabilités prédisent que les sorties du RAP-2 sont de meilleure qualité que celles de RAP-1 pour la détection d'entités nommées bien que RAP-1 affiche un WER inférieur à celui de RAP-2 (22,3 < 25,7). Cette observation est en accord avec les résultats obtenus par les systèmes NER. En effet, cinq systèmes NER sur les sept présents obtiennent de meilleurs résultats sur les sorties RAP-2.

Nous remarquons aussi que la différence entre RAP-2 et RAP-3 en terme de WER n'est pas très grande (25,7 et 26,6), alors que les scores obtenus avec les moyennes des différences de marges et des différences de probabilités prévoient que les sorties de RAP-2 sont nettement mieux que celles de RAP-3 pour la détection des entités nommées. Ce résultat est aussi en accord avec les résultats de l'évaluation ETAPE puisque 6 systèmes NER obtiennent de meilleurs résultats sur les sorties de RAP-2 par rapport à RAP-3. De plus, malgré l'écart assez important en terme de WER entre les sorties de RAP-3 (26,6) et RAP-4 (30,4), quatre systèmes NER sur sept obtiennent de meilleurs résultats sur les sorties de RAP-4 par rapport à celles de RAP-3. Néanmoins, la différence en terme de performance pour les systèmes NER n'est pas très grande et ne dépassent pas 1% pour la plupart des systèmes. Les scores calculés avec notre méthode sont en accord avec cette observation et affichent des scores très similaires pour ces deux systèmes RAP.

Nous avons aussi mesuré la corrélation des rangs offerte par le tau de Kendall, entre les classement des sorties RAP suivant chacune des trois mesures (WER, différences des marges et différences des probabilités) et le classement des sorties RAP suivant l'augmentation du taux d'erreur qu'ils ont causé pour les systèmes NER. Les indices de corrélation obtenus sont les suivants :

— 0,36 pour la WER avec une P-value de 0.001
— 0,39 pour la différence de marge avec une P-value de 0.0007

— 0,39 pour la différence de probabilité avec une P-valuer de 0.0007

Les indices de corrélation obtenus sont faibles et suggèrent donc que nos mesures ne sont pas à l'heure actuelle très efficaces. Néanmoins l'obtention d'un meilleur indice de corrélation avec nos mesures qu'avec le WER valide notre approche préliminaire.

6 Conclusions et perspectives

Cette étude propose une nouvelle méthodologie pour évaluer la qualité des sorties RAP dans un contexte applicatif particulier qui est la détection d'entités nommées à partir de la parole transcrite automatiquement. Elle s'appuie sur un modèle statistique pour mesurer le changement de probabilité a posteriori de présence des entités. D'une certaine façon elle tente d'évaluer la perte informationnelle due aux erreurs de reconnaissance du système RAP autour des entités nommées. Les résultats obtenus sont encourageants et montrent que les mesures proposées ont le potentiel d'offrir de meilleures estimations que le WER dans ce cadre applicatif.

Un premier point d'amélioration va être de prendre en compte la totalité des entités, et spécifiquement les composants, au lieu de se limiter aux types. Il y aura cependant besoin de suivre la structure de la métrique ETER pour les prendre en compte correctement sachant qu'ils sont bien plus nombreux que les types ce qui peut poser un problème de prévalence.

Le second point sera de généraliser autant que possible la projection pour permettre de comparer tous les mots et non uniquement ceux en frontière d'entité. Cela permettra de prendre en compte les cas d'erreurs de transcription provoquant des insertions.

Enfin un dernier point sera de rajouter des indices à plus longue distance dans les modèles. En particulier les parties du discours sont un indice couramment utilisé par les systèmes d'annotation en entités nommées. Une telle annotation, supervisée ou non, pourrait permettre une meilleure estimation des probabilités et donc des scores finaux.

Remerciements

Nous remercions les relecteurs pour nous avoir recommandé l'utilisation du tau de kendall pour mesurer la corrélation des rangs et pour tous leurs conseils fructueux.

Ce travail a été financé partiellement par le projet VERA - ANR 12 BS02 006 04 et par la convention CIFRE No 2012/0771.

Références

BEN JANNET, M. A., ADDA-DECKER, M., GALIBERT, O., KAHN, J. et ROSSET, S. (2014). Eter : a new metric for the evaluation of hierarchical named entity recognition. *In Proc of LREC*, Reykjavik, Iceland. ELRA.

FAVRE, B., CHEUNG, K., KAZEMIAN, S., LEE, A., LIU, Y., MUNTEANU, C., NENKOVA, A., OCHEI, D., PENN, G., TRATZ, S., VOSS, C. et ZELLER, F. (2013). Automatic Human Utility Evaluation of ASR Systems : Does WER Really Predict Performance ? *In Interspeech, Lyon (France)*.

GALIBERT, O., LEIXA, J., ADDA, G., CHOUKRI, K. et GRAVIER, G. (2014). The ETAPE speech processing evaluation. *In Proc of LREC*, Reykjavik, Iceland. ELRA.

GALIBERT, O., ROSSET, S., GROUIN, C., ZWEIGENBAUM, P. et QUINTARD, L. (2011). Structured and extended named entity evaluation in automatic speech transcriptions. *In Proc of IJCNLP*, Chiang Mai, Thailand.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *In Proc of Interspeech 2009*.

GAROFOLO, J. S., AUZANNE, C. G. et VOORHEES, E. M. (2000). The TREC spoken document retrieval track : A success story. *NIST SPECIAL PUBLICATION SP*, 500(246):107–130.

GRAVIER, G., ADDA, G., PAULSSON, N., CARRÉ, M., GIRAUDEL, A. et GALIBERT, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. *In* CHAIR), N. C. C., CHOUKRI, K., DECLERCK, T., ur DO AN, M. U., MAEGAARD, B., MARIANI, J., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical Very Large Scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513.

MUNTEANU, C., BAECKER, R., PENN, G., TOMS, E. et JAMES, D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. *In Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 493–502. ACM.

PALLETT, D. S. (2003). A look at nist's benchmark asr tests : past, present, and future. *In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 483–488. IEEE.

ROSSET, S., GROUIN, C. et ZWEIGENBAUM, P. (2011). Entités nommées structurées : guide d'annotation quaero. limsi–cnrs, orsay, france.

WANG, Y.-Y., ACERO, A. et CHELBA, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. *In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 577–582. IEEE.

Exploitation de la distribution des locuteurs pour la segmentation thématique de journaux télévisés

Abdessalam Bouchekif[1,2] Géraldine Damnati[1] Delphine Charlet[1]
(1) Orange Labs, 2 Avenue Pierre Marzin 22307 Lannion - France
(2) Laboratoire d'Informatique de l'université du Maine, LIUM - France
{abdessalam.bouchekif,geraldine.damnati,delphine.charlet}@orange.com

RÉSUMÉ

Dans cet article, nous étudions l'apport de la distribution des locuteurs dans le cadre de la segmentation thématique de journaux télévisés. A partir d'un système fondé sur la cohésion lexicale, nous proposons une méthode qui intègre directement la distribution des locuteurs dans le calcul de la cohésion. Une frontière potentielle est valide si la distribution conjointe des mots et des locuteurs diffère suffisament de part et d'autre de la frontière. Des expériences menées sur deux corpus couvrant 8 chaînes françaises montrent une amélioration du processus global de segmentation.

ABSTRACT

In this paper, we study the contribution of speaker distribution in topic segmentation of TV Broadcasts. Based on a lexical cohesion system, we propose an appraoch that directly integrates the speaker distribution when processing the cohesion. A potential boundary is effective if the joint distribution of terms and speakers is different enough from one side of the boundary to the other. Experiments run on two corpora of various Broadcasts News formats collected from 8 French TV channels, show a significant improvement in the overall topic segmentation process.

MOTS-CLÉS : Segmentation thématique, Distribution en locuteurs, Cohésion lexicale.

KEYWORDS: Topic segmentation, Speakers distribution, Lexical cohesion.

1 Introduction

La segmentation thématique (ST) est une tâche qui consiste à trouver des segments thématiquement homogènes : *i.e.* parlant d'un seul sujet. Disposer des instants de passage d'un thème à un autre peut être utilisé dans de nombreux domaines comme la recherche d'information ou le résumé automatique. Dans la littérature, trois catégories d'indices ont été exploitées : des indices *lexicaux, acoustiques et visuels*. Les **indices lexicaux** reposent sur l'exploitation de la transcription automatique. Les principales approches reposent sur des calculs de similarité à partir des *distributions de mots* ou de façon indirectes à partir de *chaînes lexicales*. Les **indices acoustiques** peuvent être la *durée de pause* : un nouveau sujet est généralement précédé par une pause de durée importante. Le passage à d'autres informations comme la météo et le sport sont souvent précédés par un *jingle*. Le *repérage du présentateur principal* peut s'averer également un bon indicateur pour repèrer les frontières (Bouchekif *et al.*, 2013). Les **indices visuels** peuvent inclure

la détection du visage du présentateur principal, de logos ou de textes incrustés. Cependant, les indices visuels sont fortement liés aux règles éditoriales de chaque chaîne télévisée (Xie *et al.*, 2010). Par exemple, le système développé dans (Rosenberg et Hirschberg, 2006) nécessite des informations sur la source des émissions et plus précisément sur la provenance (CNN, ABC, ...). Dans ce papier, nous exploitons uniquement l'audio sans aucune information visuelle.

Plusieurs algorithmes de segmentation thématique fondés sur la cohésion lexicale ont été proposés dans la littérature. L'algorithme de TextTiling (Hearst, 1997) calcule la similarité entre chaque paire de blocs adjacents. Une valeur élevée signifie qu'il y a des termes en communs entre ces deux blocs et donc qu'ils appartiennent au même sujet. Une faible valeur indique une rupture de cohésion et un changement de sujet. Une courbe de cohésion est générée à partir de laquelle sont extraites les hypothèses de frontières. L'algorithme C99 (Utiyama et Isahara, 2001) et MinCut(Malioutov et Barzilay, 2006) calculent la similarité pour n'importe quel couple de phrases et ces valeurs sont exploitées de façon indirecte (à travers une matrice de rang et un graphe non orienté respectivement). Notre approche s'inspire de l'algorithme de TextTiling, avec un mécanisme original de pondération des termes préalablement au calcul de similarité et un algorithme d'extraction de frontières à partir de la courbe de cohésion (Bouchekif *et al.*, 2014).

Dans cet article nous nous intéressons à l'information véhiculée par la structuration en locuteurs d'un journal télévisé. Le rôle du locuteur a été exploité dans plusieurs travaux. Dans de précédents travaux, nous avons montré que la détection du présentateur principal (*anchor*) pouvait améliorer le processus de segmentation, que ce soit en l'intégrant à l'algorithme d'extraction de frontières (Bouchekif *et al.*, 2013) ou dans le calcul de la pondération des termes (Bouchekif *et al.*, 2014). Cet indice permettait d'améliorer nettement les performances des systèmes, ce qui incombe essentiellement à la structure des JTs : le présentateur principal donne un aperçu du sujet qui est ensuite détaillé par un reportage ou une interview. Au delà du présentateur principal, la distribution de l'ensemble des locuteurs du JT représente une source d'information qui doit pouvoir être exploitée de façon plus approfondie. Les systèmes de ST exploitant l'information liée aux locuteurs le font essentiellement dans un cadre supervisée (i.e. à l'aide d'un classifieur). Par exemple, (Wang *et al.*, 2012) exploitent la segmentation en locuteurs à travers un descripteur binaire (changement de locuteur / pas de changement) pour désigner la prise de parole. Dumont et al. (Dumont et Quénot, 2012) exploitent le résultat d'un regroupement en locuteurs en ajoutant l'indice du locuteur dans chaque observation (frontière potentielle). Dans ce cas, les valeurs potentielles de ce descripteur correspondent au nombre de locuteurs présent dans l'émission. Dans cet article, nous proposons une méthode d'intégration directe de la répartition des locuteurs dans le calcul de la cohésion avec une approche non supervisée. En effet, si l'anchor intervient généralement tout au long du JT, il est plus probable que les autres locuteurs (reporters ou intervenants externes) n'interviennent que dans un seul sujet. La notion de cohésion appliquée jusqu'ici à la distribution des mots, s'étend à la distribution des locuteurs.

Ce papier est structuré de la façon suivante : la section 2, est consacrée à l'algorithme de segmentation thématique utilisé dans la suite de notre travail. La section 3 décrit la méthode d'intégration de la distribution des locuteurs dans le calcul de la cohésion lexicale. Les expériences sont présentées dans la section 4.

2 Systèmes de segmentation thématique

Notre système de segmentation repose sur l'analyse de la distribution des mots. Une similarité est calculée entre chaque paire de blocs adjacents. Les segments unitaires considérés sont des groupes de souffle (GS), c'est-à-dire des séquences de mots entre deux pauses dans un tour de parole. Les pauses et les changements de locuteur sont détectés automatiquement par le système de transcription automatique. La similarité est donc calculée tout au long de l'émission à l'aide d'une fenêtre glissante de taille K, entre des blocs adjacents de K GS de part et d'autre de la frontière potentielle. Il en résulte une courbe de cohésion lexicale à partir de laquelle sont extraites les hypothèses de frontières. Relever les changements de sujets directement à partir des vallées de la courbe (comme dans l'algorithme de TextTiling (Hearst, 1997)) n'est pas toujours optimale (Bouchekif *et al.*, 2013). Nous utilisons une combinaison linéaire entre les valeurs de la cohésion et la profondeur des vallées. Cette combinaison permet de favoriser les valeurs faibles de la cohésion lexicale qui sont en même temps des minimums locaux.

Pour une frontière potentielle j, le score suivant doit être maximisé :

$$score(j) = \lambda(1 - cohesion(j)) + (1 - \lambda) depth(j) \quad (1)$$

La pondération des termes permet de favoriser les termes les plus pertinents lors du calcul de la cohésion. Nous avons proposé dans (Bouchekif *et al.*, 2014) un mécanisme itératif permettant d'estimer efficacement une pondération intra-contenu, sans avoir recours à des ressources externes.

1. Initialement, l'émission est coupée en N morceaux uniformes (chunks) qui sont utilisés pour estimer les pondérations $TF-IDF$. L'indice du premier GS de chaque chunk uniforme est considéré comme l'ensemble initial de frontières et est placé dans le vecteur hyp_0.
2. A l'itération i :
   – les hypothèses de l'itération $i-1$, notées (hyp_{i-1}) sont utilisées pour estimer les pondérations $TF-IDF(i)$. Le poids associé au terme t dans le groupe de souffle x est associé au poids $w(c(x), t)$ qui dépend du chunk $c(x)$ dans lequel se trouve x.

$$w_{TfIdf}(c(x), t) = TF(c(x), t) \times IDF(t) \quad (2)$$

   où $\text{TF}_{(c(x),t)}$ est la fréquence du terme t dans le morceau $c(x)$.
   $IDF(t) = \log(\frac{N}{n_t})$
   n_t est le nombre de chunks dans lequel le terme t apparaît.
   Cette approche permet de faire ressortir les mots discriminants dans un passage de l'émission relativement aux autres passages.
   – La mesure cosinus est utilisée pour mesurer la proximité entre la représentation vectorielle de deux blocs adjacents b_j et b_{j+1}. Le coefficient associé au terme t dans la représentation vectorielle d'un bloc b est une valeur pondérée $v(b, t)$

$$v(b, t) = \sum_{x \in b} \left(f_{(x,t)} \times w(c(x), t) \right) \quad (3)$$

   où $f_{(x,t)}$ est la fréquence du terme t dans le GS x. Pour une frontière potentielle entre deux blocs b_j et $b_{(j+1)}$, la cohésion est donnée par $cohesion(j) = cosine(V_j, V_{j+1})$.

– La combinaison linéaire entre la cohésion lexicale et la profondeur des vallées est recalculée. L'algorithme de division récursive (splitting) est appliqué pour déterminer les hypothèses à la $i^{ème}$ itération (hyp_i).

3. Les hypothèses obtenues sont présentées à l'étape 2.
4. L'algorithme s'arrête lorsque la segmentation se stabilise (pas de changement significatif entre les hypothèses de deux itérations successives hyp_{i-1} et hyp_i).

La stabilisation de l'algorithme est calculée à l'aide de la métrique p_k (Beeferman *et al.*, 1999). La mesure p_k compare une segmentation de référence R et une hypothèse de segmentation H. Elle est fondée sur le principe d'une fenêtre glissante de taille k parcourant la totalité de l'émission. Dans notre implémentation, la valeur de k est fixée à 6. L'algorithme s'arrête lorsque la valeur de p_k entre hyp_{i-1} et hyp_i est proche de 1 $(1 - p_k(hyp_{i-1}, hyp_i)) \leq \epsilon$.

Entrée :
- $file$: transcription automatique de l'émission
- Pot_Bound : L'ensemble des frontières potentielles

Sortiée : hyp_i : vecteur contenant les hypothèses

begin
  $p_k^0 \leftarrow 1$; $i \leftarrow 0$; $N \leftarrow$ NumberChunks(file_Size) ; $hyp_0 \leftarrow$ UniformCut(N) ;
  while *(stopCriterion not true)* **do**
    i ← i + 1 ;
    $MatrixWeights \leftarrow$w($file, hyp_{i-1}$)
    foreach Pot_Bound_j *in* $file$ **do**
      $Cohesion(j) \leftarrow cosine(b_j, b_{j+1})$
    end
    for $j = 0$ **to** n **do**
      score(j)=$\lambda(1 - Cohesion(j)) + (1 - \lambda)depth(j)$
    end
    $hyp_i \leftarrow$Splitting (score)
    $p_k^i \leftarrow p_k(hyp_i, hyp_{i-1})$
  end
  return hyp_i
end

Algorithme 1 : Algorithme de segmentation itérative

3 Exploitation de la distribution des locuteurs dans la segmentation thématique

3.1 Structuration en locuteurs (speaker diarization)

La structuration en locuteurs a été effectuée en deux étapes : l'étape de segmentation suivie d'un regroupement hiérarchique ascendant basés tous deux sur le critère BIC. Elle permet d'obtenir une structuration initiale dans laquelle les clusters sont assez purs et contiennent suffisamment de données pour permettre de modéliser le locuteur par un mélange de gaussiennes. Ensuite,

avec une telle modélisation, un processus itératif de segmentation via un décodage de Viterbi, et de regroupement par le critère CLR (Barras *et al.*, 2006) est réalisé. La mise en oeuvre et les performances du système sur des données de type broadcast news (environ 10% de Diarization Error Rate (DER)) figurent dans (Charlet *et al.*, 2013).

3.2 Intégration de la distribution des locuteurs dans le calcul de la cohésion lexicale

La figure 1 illustre la corrélation entre la distribution des locuteurs et les changements de sujets. Il peut arriver que notre ST propose des frontières au milieu d'un reportage ou bien avant l'achèvement d'une interview. Dans ces cas, l'exploitation de cette information peut être une solution pour limiter le nombre d'hypothèses.

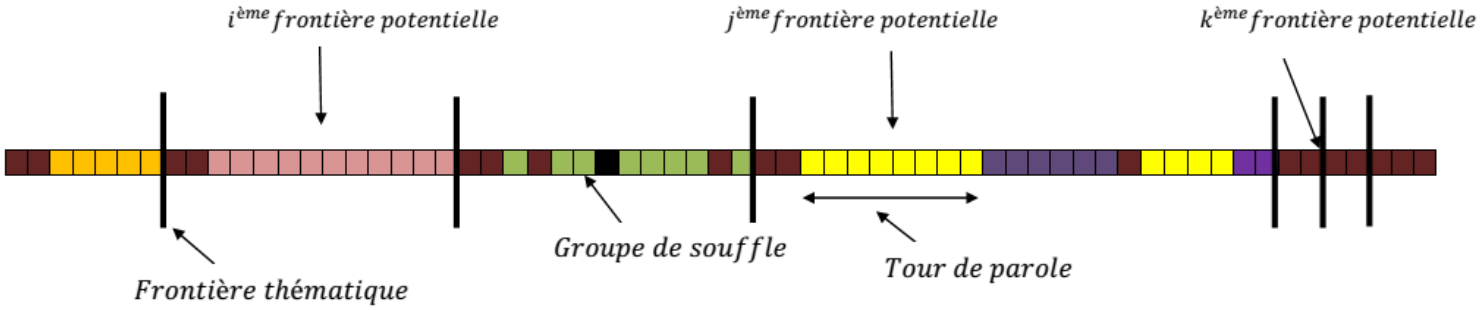


FIGURE 1 – Corrélation entre la distribution des locuteurs et les changements de sujets

Nous proposons une approche pour intégrer la structuration en locuteurs directement dans le calcul de la cohésion lexicale. Pour cela, nous procédons au préalable à une étape de synchronisation entre les groupes de souffles déterminés par le système de transcription et les tours de parole obtenus par le système de structuration en locuteur. L'idée est ensuite de considérer l'identifiant du locuteur comme un terme, c'est-à-dire que chaque groupe de souffle (GS) est représenté non seulement par les mots prononcés mais également par l'indice du locuteur correspondant.

Dans notre algorithme de segmentation l'émission est transformée en une matrice de taille $n \times m$ où n est le nombre de groupes de souffle et m la taille du vocabulaire. L'élément $e_{i,j}$ représente le nombre d'occurrences du $j^{\text{ème}}$ terme dans le $i^{\text{ème}}$ groupe de souffle. Donc le vecteur de représentation associé à un GS était jusqu'ici constitué d'autant de composants que le JT comportait de termes. Ce vecteur est désormais augmenté d'autant d'éléments que le JT contient de locuteurs différents. Cette représentation permet de faire cohabiter la distribution des mots et des locuteurs dans le calcul de la cohésion.

Afin de tenir compte indirectement du temps de parole associé aux locuteurs, la valeur $f_{(x,l)}$ associée au locuteur l dans le groupe de souffle GS_x est définie à partir du nombre de mots nb_x retenus dans le groupe de souffle à l'issue de l'étape de pré-traitement (filtrage par score de confiance et par catégorie syntaxique). Un simple comptage du nombre de mots pour le locuteur associé au groupe de souffle n'est pas approprié car il génère une contribution trop importante de l'information liée au locuteur à l'échelle de la fenêtre sur laquelle est calculée la cohésion. Afin d'équilibrer la contribution des mots et des locuteurs, une normalisation est réalisée relativement au nombre maximum de mots (max) dans l'ensemble des groupes de souffle du JT. Enfin, afin

d'éviter une valeur nulle dans le cas où un groupe de souffle ne contient aucun terme à l'issue des pré-traitements, nous appliquons un lissage qui conduit à la formulation suivante.

$$f_{(x,l)} = \frac{1 + nb_x}{1 + max} \tag{4}$$

Similairement à l'équation 3, le coefficient associé au locuteur l dans la représentation vectorielle d'un bloc b est une valeur pondérée $v(b,l)$.

$$v(b,l) = \sum_{x \in b} \left(f_{(x,l)} \times w(c(x),l) \right) \tag{5}$$

Le mécanisme de pondération itérative présenté précédemment s'appliquera de la même manière aux valeurs associées aux termes et aux valeurs associées aux locuteurs. Pour ces derniers, la pondération permettra de pénaliser les locuteurs présents durant toute l'émission (comme l'anchor par exemple) et de favoriser les locuteurs qui interviennent localement pour un sujet donné. Cette approche permet de définir une mesure de cohésion qui prend en compte conjointement le changement sémantique et le changement d'intervenants.

4 Expériences et résultats

Les expériences sont menées sur deux corpus d'émissions. Le premier nommé Multi-Chaînes (MC) est constitué de 33 journaux télévisés de 7 chaînes françaises (TF1, France2, France3, LCI, France24, Arte, M6) d'une durée moyennes de 22 minutes. Le deuxième est composé de 23 JT de chaîne D8, d'une durée moyenne de 13 minutes. Le corpus MC contient 379 frontières thématiques, et la durée moyenne des sujets est de 115*s* et le corpus D8 comporte 140 frontières pour une durée moyenne des sujets de 128s. Le corpus MC est utilisé pour régler les paramètres libres de l'algorithme et le corpus *D*8 permet de valider la robustesse de l'approche sur des données issues d'une autre chaîne. Les émissions ont été transcrites par le biais du système de reconnaissance de la parole de Vocapia (Gauvain *et al.*, 2002). Le taux d'erreurs mots sur le corpus MC est de 16,1%. Nous ne pouvons donner les performances obtenues sur le corpus D8 qui n'a pas été transcrit manuellement. Les mots qui ont un score de confiance inférieur à 0.5 sont écartés. Les pré-traitements classiques ont été appliqués : lemmatisation (à l'aide de l'outil Lia_tagg [1])), suppression des mots outils. Par ailleurs, de façon similaire à (Guinaudeau *et al.*, 2010) nous avons écarté la première partie d'un journal lorsqu'elle ne contient que des titres et la dernière partie lorsqu'il s'agit du rappel des titres.

La taille de la fenêtre pour la détermination de la courbe de cohésion lexicale a été optimisée sur le corpus MC et a été fixée à 16 groupes de souffles. Le coefficient λ d'interpolation pour le calcul du score a été fixé à $0,75$. Le seuil ϵ sur la mesure p_k pour le critère d'arrêt de l'algorithme itératif vaut $0,09$. Les performances sont mesurées en termes de rappel et de précision, en comparant la segmentation de référence avec celle de l'hypothèse. De façon similaire à plusieurs travaux dans la littérature, une tolérance de 10*s* a été autorisée entre les frontières d'hypothèses et de références. Le tableau 1 décrit les performances de système de segmentation avec et sans prise en compte de la distribution des locuteurs. Le système a obtenu les meilleurs résultats en tenant compte des indices de locuteurs dans le calcul de la cohésion. En effet, la prise en compte conjointe de ces deux informations augmente la F-mesure de +2.4 et de +7.1 points pour le corpus MC et D8

1. http ://pageperso.lif.univ-mrs.fr/ frederic.bechet/download.html

Corpus	Sans locuteur			Avec locuteur		
	Rappel	Precision	Fmesure	Rappel	Precision	Fmesure
MC	63.6	63.2	63.4	66.5	65.1	**65.8**
$D8$	77.4	58.0	66.2	77.1	69.9	**73.3**

TABLE 1 – Impact de la distribution des locuteurs

respectivement. Après analyse des résultats, nous avons constaté que l'insertion des étiquettes de locuteurs dans le calcul de la cohésion permet d'affiner la segmentation même si le JT contient peu de répétitions. En effet, il peut arriver que notre système de segmentation propose une frontière au milieu d'une intervention ou au milieu d'un reportage au moment où un intervenant prend la parole (par exemple, la $i^{ème}$ et la $j^{ème}$ frontières potentielles de la figure 1). Dans ce dernier cas, si le reporter reprend ultérieurement la parole, la redondance au niveau locuteurs renforce la cohésion, ce qui rend les deux blocs adjacents liés thématiquement.
Les résultats montrent que la combinaison favorise le corpus D8 dans une plus grande mesure. De l'analyse de ces deux corpus, nous avons constaté que cette différence revient essentiellement à la taille des segments. Deux types de segments thématiques sont considérés : *long* si la durée est supérieure à 30*s* sinon il est jugé comme *court*. Ainsi, les frontières peuvent être de trois types : *long-long*, *court-court*, *court-long*. Une frontière portant un label *court-long* signifie qu'elle est précédée par un court segment puis suivie par un long ou inversement. Le tableau 2 donne une description détaillée sur la nature des segments considérés. De l'analyse du tableau, il ressort

Corpus	MC	$D8$
Nb frontières long-long	227	121
Nb frontières court-court	29	1
Nb frontières court-long	123	18
Total	379	140

TABLE 2 – Description des corpus en termes de longueurs de segments

que le corpus MC possède un nombre plus important de segments de courte durée. Ce type de segments n'est pas favorable à notre système fondé sur la redondance. En effet, des sujets de moins de 30*s* correspondent généralement à des brèves, lues par le présentateur principal ou par un autre journaliste, qui sont difficiles à isoler. Ce qui explique que notre système de segmentation met en évidence peu de changement de thèmes (voir la $k^{ème}$ frontière de la figure 1). Notre hypothèse a été confirmée dans le tableau 3 où les résultats sont donnés en terme de rappel pour chaque type de frontières. Les journaux télévisés de la chaîne D8 sont construits selon des

Corpus	Sans locuteur			Avec locuteur		
	Long-Long	Long-court	court - court	Long-Long	Long-court	court - court
MC	64.3	53.7	17.2	67.8	51.2	24.1
$D8$	55.4	50.0	0.0	69.4	38.9	0.0

TABLE 3 – L'infuence de la taille des segments sur la qualité de segmentation en terme de F-mesure

schémas plus traditionels avec des reportages suffisament longs ce qui rend d'autant plus visible l'apport de la méthode fondée sur la cohésion. L'amélioration est nettement plus importante dans le corpus *D*8 car la plupart des frontières sont de type long-long (86.4%), contrairement au corpus MC qui contient uniquement 59.9% de cette catégorie de frontières.

5 Conclusion

Cet article propose une approche pour intégrer la distribution des locuteurs au même niveau que la distribution lexicale. La cohésion lexicale et la cohésion en locuteur sont ainsi regroupées en une seule notion : la cohésion de la parole (*speech cohesion*). Nous avons donné deux dimensions aux valeurs de la similarité : la répartition des mots et des intervenants. L'utilisation simultanée de ces deux informations améliore nettement les résulats en particulier pour détecter des frontières entre deux segments longs. En revanche, les performances du système restent à améliorer lorsqu'il s'agit de segments courts. Ces dernières devraient faire l'objet d'un traitement spécifique.

Références

BARRAS, C., ZHU, X., MEIGNIER, S. et GAUVAIN, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1505–1512.

BEEFERMAN, D., BERGER, A. et LAFFERTY, J. (1999). Statistical models for text segmentation. *Mach. Learn.*, 34(1-3).

BOUCHEKIF, A., DAMNATI, G. et CHARLET, D. (2013). Complementarity of lexical cohesion and speaker role information for story segmentation of french tv broadcast news. *In SLSP*.

BOUCHEKIF, A., DAMNATI, G. et CHARLET, D. (2014). Intra-content term weighting for topic segmentation. *In ICASSP*, Florence, Italie.

CHARLET, D., BARRAS, C. et LIÉNARD, J.-S. (2013). Impact of overlapping speech detection on speaker diarization for broadcast news and debates. *In ICASSP*, pages 7707–7711.

DUMONT, E. et QUÉNOT, G. (2012). Automatic story segmentation for tv news video using multiple modalities. *Int. J. Digital Multimedia Broadcasting*, 2012.

GAUVAIN, J.-L., LAMEL, L. et ADDA, G. (2002). The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89–108.

GUINAUDEAU, C., GRAVIER, G. et SÉBILLOT, P. (2010). Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels. *In 17e conférence sur le traitement automatique des langues naturelles, TALN'10*, Montréal, Québec, Canada.

HEARST, M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

MALIOUTOV, I. et BARZILAY, R. (2006). Minimum cut model for spoken lecture segmentation. *In International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 25–32.

ROSENBERG, A. et HIRSCHBERG, J. (2006). Story Segmentation of Broadcast News in English, Mandarin and Arabic. *In NAACL*, pages 125–128.

UTIYAMA, M. et ISAHARA, H. (2001). A statistical model for domain-independent text segmentation. *In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 499–506.

WANG, X., XIE, L., LU, M., MA, B., CHNG, E. et LI, H. (2012). Broadcast news story segmentation using conditional random fields and multimodal features. *IEICE Transactions*, 95-D(5):1206–1215.

XIE, L., YANG, Y., LIU, Z., FENG, W. et LIU, Z. (2010). Integrating acoustic and lexical features in topic segmentation of chinese broadcast news using maximum entropy approach. *In International Conference on Audio, Language and Image Processing*, pages 407–413.

Comparaison des stratégies articulatoires d'un locuteur bilingue anglais-français: Données et modèles préliminaires

Pierre Badin [1] Thomas R. Sawallis [2] Laurent Lamalle [3]
(1) GIPSA-Lab / DPC, UMR 5216, CNRS – Université de Grenoble Alpes, France
(2) New College, University of Alabama, Tuscaloosa, AL 35487, USA
(3) Inserm US 17 — UMS IRMaGe – Université de Grenoble Alpes, France
pierre.badin@gipsa-lab.grenoble-inp.fr, tom.sawallis@ua.edu

RESUME

Notre objectif est de comparer les stratégies articulatoires utilisées par les bilingues, afin de mieux comprendre et le bilinguisme et les différences entre langues. Nous avons donc acquis des images IRM médiosagittales pour tous les phonèmes produits par un locuteur bilingue dans sa langue maternelle, l'anglais américain (AE), et en français (FR) qu'il maitrise à très haut niveau. Nous avons comparé les contours de ses articulateurs entre les deux langues : (1) comparaisons directes des contours de phonèmes analogues, (2) comparaison des nomogrammes articulatoires générés par les modèles des deux langues, et (3) reconstruction croisée des articulations d'une langue par un modèle basé sur l'autre. Il est apparu, entre autres observations intéressantes, que la protrusion labiale vocalique s'étend davantage en FR qu'en AE, et que les mouvements du corps de la langue sont plus grands et plus orientés vers l'avant pour l'AE.

ABSTRACT

Comparing articulatory strategies in an English/French bilingual speaker: Data and preliminary models

Our goal is to compare the articulatory strategies used by bilinguals, to better understand both bilingualism and the differences between languages. We have therefore obtained mid-sagittal MRI scans of a bilingual talker producing all the phonemes in both his native American English (AE) and in French (FR), a second language in which he is highly fluent. We have compared his articulatory contours in the two languages with: 1) direct comparison of the contours of analogous phonemes, 2) comparison of articulatory nomograms, and 3) crossed reconstructions of one language's articulations by a model based on the other language. Various interesting observations were made, including that labial protrusion in vowels is greater in FR than in AE, and that tongue body movements were greater and more oriented toward fronting in AE than in FR.

MOTS-CLES : Bilinguisme, français, anglais américain, articulation, IRM, mesure articulatoire, modèle articulatoire.
KEYWORDS: Bilingualism, French, American English, articulation, MRI, articulatory measurement, articulatory model.

1 Introduction

Une meilleure connaissance comparative des comportements phonétiques des bilingues dans leurs diverses langues est d'un grand intérêt en soi, mais aussi dans la perspective d'applications pour l'apprentissage des langues secondes (L2). La plupart des études liées à ces questions portent soit sur le bilinguisme soit sur l'articulation, mais la littérature associant les deux aspects semble très pauvre, et vise en général deux thèmes.

Une première série d'études porte sur la *posture articulatoire* (PA) (en anglais *articulatory setting*), définie par Honikman (1964) (cité par Mennen, Scobbie, de Leeuw, Schaeffler & Schaeffle, (2010) comme la configuration phonétique – spécifique à une langue – vers laquelle le conduit vocal a tendance à retourner habituellement lors des pauses. Mennen, Scobbie et al. (2010) mentionnent la difficulté à séparer l'influence de la PA d'autres aspects du langage tels que l'inventaire phonémique, et indiquent que la variabilité interlocuteurs qui affecte la plupart des mesures peut masquer les différences de PA entre langues. Wilson (2006) s'est servi de l'Optotrak et de l'ultrason pour caractériser respectivement la position de la mandibule et la forme de la langue dans la *posture inter-énoncés* (PIE) (*Inter-speech posture*) comme estimation de la PA. Il a trouvé que la PIE permet de distinguer parmi ses sujets canadiens les groupes francophone et anglophone, ce dernier ayant en moyenne une pointe de langue plus élevée, des lèvres plus protruses et une aperture labiale horizontale plus étroite. En outre, il a trouvé que les bilingues les meilleurs (jugés natifs par des monolingues des deux langues) ont bien des PIE différentes et proches des monolingues pour leurs deux langues. Il conseille l'enseignement de la PA aux apprenants en langues secondes, ce qui est peut-être prématuré, vu l'absence de démonstration que la PA soit la source d'accents étrangers ou d'erreurs segmentales précises.

L'autre thème, classique en recherche sur le bilinguisme, est l'étude du transfert, c'est-à-dire l'apport chez un locuteur de sa langue maternelle dans l'utilisation d'une langue seconde, ici aux niveaux phonologique et articulatoire. Tepperman, Bresch, Kim, Lee, Goldstein & Narayanan (2009) cherchent à mettre en évidence des traces de *transfert phonologique* mesurables au niveau articulatoire. A l'aide de vidéos IRM ils comparent les contrastes /v-w/ et /d-ð/ en anglais chez 3 américains et 3 allemands en L2. Les allemands affichent certaines différences articulatoires entre les phonèmes, mais sensiblement moins que les américains natifs. En outre, là où manquent les différences, les substituts ne sont pas différenciés de leur allemand natif, ce qui implique le transfert attendu de leur L1 vers leur L2.

Oh (2002) étudie par échographie les différences de timing et d'articulation entre les productions de /l/ en anglais et en coréen pour des enfants monolingues et bilingues, et montre que les bilingues mélangent les gestes spécifiques aux deux langues.

Zerling (1992) a mesuré les dimensions de l'ouverture aux lèvres des voyelles, pour le français à partir d'images que lui-même a prise, et pour l'anglais américain à partir de deux études publiées par Fromkin (1964) et Linker (1982). Son analyse contraste non seulement les deux langues, mais aussi les deux études de l'anglais : il trouve que la simple dimension de labialisation phonologique ne suffit pas pour expliquer les différences. Il ajoute des dimensions de *type*, *degré* et *stratégie de labialisation* pour tout prendre en compte et bien différencier les langues. Malheureusement, ces études ne s'étendent pas aux bilingues dans leurs langues secondes, donc on ne peut rien savoir sur leur comportement articulatoire.

Montagu (2002) compare les formes labiales et les productions acoustiques des voyelles orales, nasales, et nasalisées du français chez des locuteurs natifs et des américains anglophones dans leur français, langue seconde. Elle trouve que les français distinguent les nasales arrières /ɑ̃ ɔ̃/ de leurs équivalents orales non seulement par la simple nasalisation phonétique mais aussi par l'arrondissement (réduction de l'aire aux lèvres) et par la protrusion. Pour /ɑ̃/, les américains bilingues reproduisent à peu près les articulations des français natifs. Par contre, ils ne reproduisent pas le *sur-arrondissement* des français dans leur /ɔ̃/, version nasale d'une voyelle déjà arrondie. Ceci semble montrer que ni la PA, ni le transfert phonologique ne suffisent à expliquer l'articulation des bilingues d'une façon globale, et qu'il faille les deux, plus une analyse propre à chaque phonème, pour comprendre la globalité des stratégies articulatoires.

Les limites de cette littérature ont motivé notre étude expérimentale et modélisatrice de l'articulation d'un locuteur anglais américain natif ayant acquis tardivement une très haute maîtrise du français. Cette étude vise à mettre en évidence les similitudes et les différences générales de stratégies articulatoires entre langues, un domaine plus large que la PA ou le transfert. Les principales étapes de l'étude ont été : (1) l'acquisition et le traitement des données IRM, (2) la comparaison directe de phonèmes analogues, (3) le développement de modèles articulatoires, et (4) la comparaison globale des deux systèmes articulatoires basée sur les modèles ainsi développés.

2 Acquisition et traitement des données articulatoires

Le locuteur, né et élevé en Floride du Sud, région alors monolingue mais avec un mélange de dialectes, a suivi quatre années de cours de français dans le secondaire, et quatre années à l'université. A 24 ans, il s'installe à Grenoble, où il fait 3 années d'études en linguistique, et plus tard 2 autres séjours sabbatiques, cumulant plus de 6 ans en milieu francophone. Pour la présente étude, il a été enregistré en deux sessions à six semaines d'intervalle, une fois en anglais américain (AE) et l'autre en français (FR), suivant un protocole standard (*cf.* Badin & Serrurier (2006) pour les détails). Lors de chaque session, seule la langue concernée était utilisée pour les interactions avec les opérateurs, de façon à maintenir le locuteur dans un mode monolingue pour chaque langue à enregistrer.

Les données IRM ont été acquises à l'aide d'un imageur Philips Achieva 3.0T TX, avec une antenne neurovasculaire SENSE, en mode d'écho de spin turbo (TSE, TE = 10,7 ms, Flip angle = 80°, TR = 426 ms). L'acquisition de chaque image médiosagittale (épaisseur de 4 mm, résolution à l'acquisition de 1 mm / pixel, champ de vue de 256×256 mm^2) prenait 8,1 secondes, pendant lesquelles le locuteur devait maintenir une articulation fixe, condition que l'on pourrait comparer à un « arrêt sur image ».

Le corpus anglais comprenait 189 éléments : d'une part les voyelles /i ɪ e ɛ æ ɑ[1] ɔ o ʊ u ʌ ə ɻ/, prononcées en isolation, dans quatre contextes CVC /m ɹ ɬ/ et [ɻ][2], dans une liste de 17 mots naturels, et suivies par /ɹ/ dans une liste de 14 mots naturels ; d'autre part toutes les consonnes /p t k f θ s ʃ m n ŋ ɹ ɬ j w/ et [ɻ] dans six contextes vocaliques symétriques VCV /i e æ a o u/; enfin les diphtongues isolées /au ɔi ai/ bloquées sur leur partie initiale. Le corpus français comprenait

[1] Dans ce travail préliminaire, nous considèrerons que les /ɑ/ AE et /a/ FR sont analogues. Il semblerait cependant que le locuteur le réalise comme [ɐ] dans son AE maternel, plus central que le [a] FR qu'il a appris à produire. Ce point fera l'objet d'une étude complémentaire.

[2] Nous avons profité de l'occasion d'enregistrer cette variante rétroflexe du /ɹ/, non-native à notre locuteur, mais qu'il sait articuler de par ses études phonétiques. Ceci fera aussi l'objet d'études complémentaires.

143 éléments : les voyelles orales /i e ɛ a ɔ o u y ø œ/ et nasales /ã ɛ̃ ɔ̃/, et les consonnes /p t k f s ʃ m n ʁ l/ dans les treize contextes vocaliques. Notons que seules les versions sourdes des occlusives et des fricatives ont été utilisées (le voisement ne pouvant pas être maintenu durant les 8,1 secondes de prise de vue IRM). Notons en complément que, dû à la limitation de la durée des séances d'enregistrement, chaque articulation n'a été prononcée qu'une fois.

Suivant la procédure décrite dans Badin & Serrurier (2006), les contours des divers organes déformables ont été manuellement créés et ajustés à partir des images médiosagitales, tandis que les contours des structures rigides ont été positionnés manuellement par translation et rotation. Toutes les articulations ont été finalement alignées entre elles grâce aux structures rigides crâniennes de façon à pouvoir être comparées par simple superposition et être utilisées pour la modélisation.

3 Comparaison deux à deux de phonèmes analogues

Nous avons comparé deux à deux les phonèmes analogues[3] entre les deux langues. Pour les voyelles isolées /i e ɛ a o u/, les principales différences se situent aux lèvres, plus protruses pour le /u/ FR et plus fermées pour le /o/ FR ; pour la langue, nous notons un recul pour le /u/ FR, un abaissement et un bombement pour le /o/ FR, et un abaissement pour le /a/ FR (voir des exemples à la Figure 1).

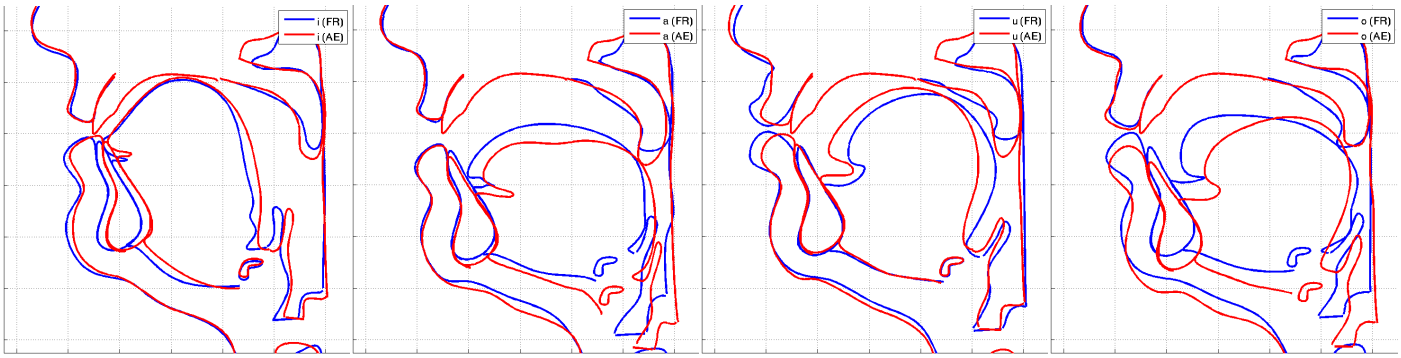


FIGURE 1– Superposition des voyelles isolées /i a u o/ en AE (rouge) et FR (bleu).

La comparaison de consonnes /p t k f s ʃ m n/, analogues dans les deux langues, dans les cinq contextes vocaliques /i e a o u/ analogues, montre à nouveau que les principales différences portent sur les lèvres en contexte /u/ pour toutes les consonnes, ainsi qu'en contexte /o/ pour /t f s m n/, la protrusion étant plus marquée en français qu'en anglais. On trouve également des différences linguales : la langue est plus reculée pour le français que pour l'anglais pour toutes les consonnes en contexte /u/ excepté pour /n/. On trouve par ailleurs de petites différences d'articulation linguale pour les consonnes labiales /p f m/, ce qui est logique puisque la langue n'est pas vraiment spécifiée pour ces consonnes.

[3] L'inventaire des phonèmes en L1 permet de répartir ceux de la L2 en différentes classes, par exemple *identique*, *similaire*, ou *nouveau*, termes qui ont des valeurs théoriques reposant en partie sur des critères perceptuels et phonologiques (Flege, 1987; Escudero, 2007). Comme nous ne cherchons dans le présent travail qu'à comparer les seules articulations, entre autres pour quantifier leur similitude, nous avons choisi le terme *analogue* comme terme pré-théorique pour les paires de sons L1-L2 qui se trouvent vers le côté *identique* de l'échelle, plutôt que le côté *dissimilaire*.

De manière attendue, les comparaisons des paires, analogues du point de vue phonologique, /ɹ–ʁ/ et /ɬ–l/ dans les cinq contextes vocaliques /i e a o u/ fait apparaître d'importantes différences. En particulier, nous avons trouvé un plus grand abaissement du corps de la langue pour le /ɬ/ AE sombre vélarisé que pour le /l/ FR clair laminal (voir des exemples à la Figure 2).

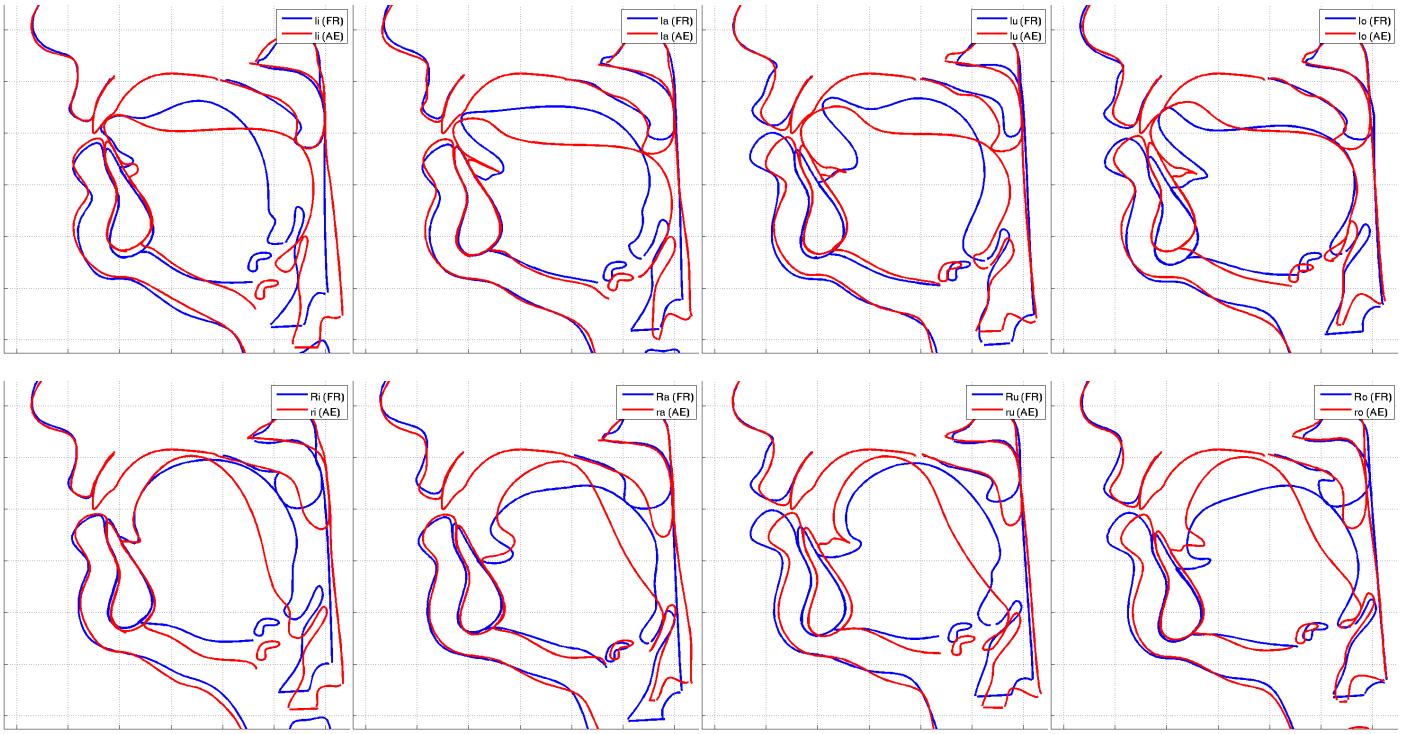


FIGURE 2 – Superposition des consonnes /ɬ–l/ (haut) et /ɹ–ʁ/ (bas) en AE (rouge) et en FR (bleu) en contexte /i a u o/.

4 Modèles articulatoires

L'étape suivante a consisté à comparer les espaces articulatoires des deux langues à un niveau plus global, par modélisation articulatoire. Nous avons donc construit des modèles articulatoires à partir de différents sous-ensembles de la base de données des contours, en utilisant les méthodes de décomposition linéaire décrites par Badin, Valdés Vargas, Koncki, Lamalle & Savariaux (2013). Le principe général consiste à modéliser les coordonnées X et Y de chaque point de contour comme une combinaison linéaire de paramètres de contrôle, en addition à une constante obtenue comme la moyenne de la coordonnée sur le corpus. Pour la langue, les composantes contrôlées par les paramètres *JH*, *TB*, *TD*, et *TTV* correspondent respectivement à l'influence de l'ouverture de la mâchoire, au déplacement du corps de la langue, à la forme bombée ou plate du corps, et aux mouvements verticaux de l'apex. Pour les lèvres, les mesures de protrusion (*UL_pro* et *LL_pro*) et de hauteur (*UL_hei* et *LL_hei*) de chaque lèvre ont été utilisées comme paramètres complémentaires au paramètre de hauteur de mâchoire *JH*. Nous avons construit ces modèles pour six corpus différents : les voyelles orales pour FR (10) et AE (13), les voyelles et consonnes analogues pour FR et AE (45), et les corpus complets FR (143) et AE (189). Pour les corpus complets les erreurs RMS de reconstruction vont de 0,5 mm pour la lèvre supérieure à 1,7 mm pour la langue pour l'anglais, et de 0,5 mm à 1,4 mm pour le français. Ces résultats sont tout à fait comparables avec ceux documentés dans d'autres études de ce type.

Pour illustrer le comportement de ces modèles, nous avons généré des nomogrammes, *i.e.* affiché la variation de la forme des articulateurs pour un balayage systématique des paramètres de contrôle du modèle dans les gammes qui résultent de l'analyse fondatrice du modèle. Chaque figure montre les variations d'un ou deux paramètres.

La Figure 3 illustre le fait que la protrusion labiale pour les voyelles a une plus grande étendue de mouvement en FR qu'en AE, ce qui généralise les observations faites à la section 3. La différence est moins claire lorsqu'elle est masquée par la coarticulation consonantique (d'où notre choix d'un sous-ensemble réduit aux voyelles du corpus). La Figure 3 illustre aussi les différences pour la composante corps de la langue TB entre FR and AE: le mouvement vertical est clairement plus grand et plus oblique vers l'avant en AE qu'en FR, surtout lorsque les corpus complets sont considérés. Cela correspond en particulier aux différences d'articulation /ɹ–ʁ/ que l'on peut observer sur la Figure 2.

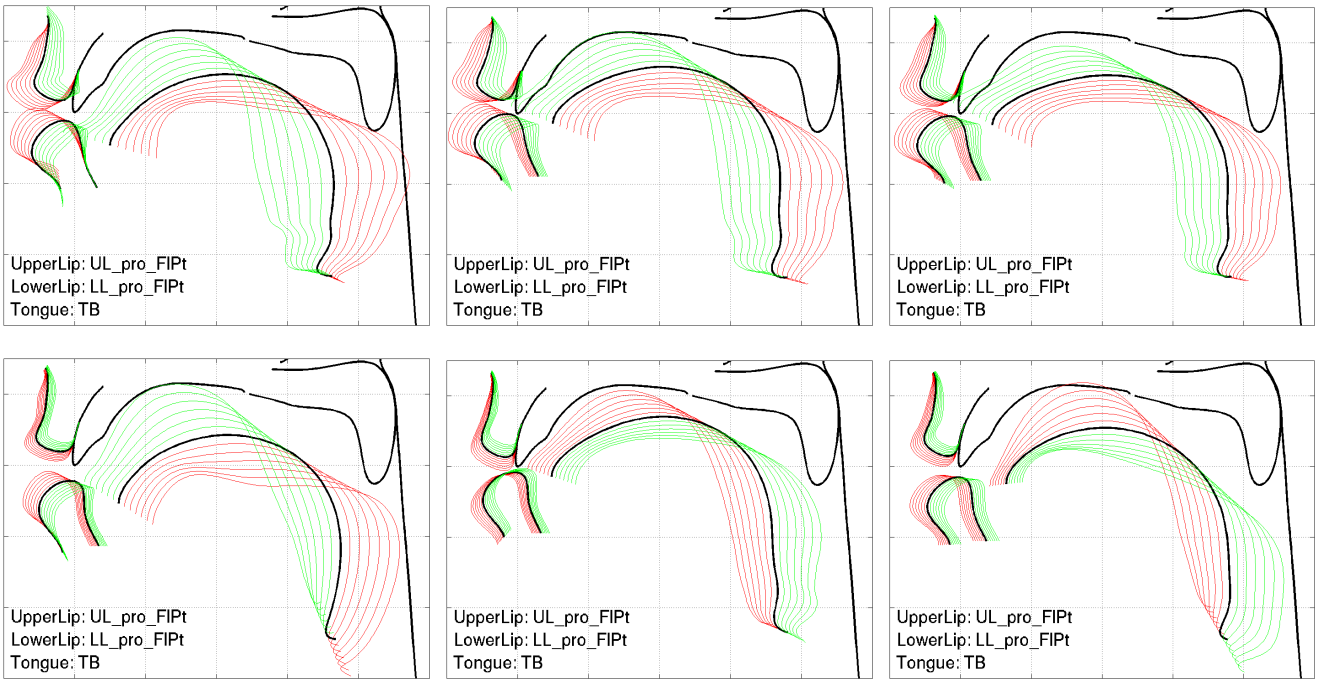


Figure 3 – Nomogrammes de lèvres pour les paramètres de protrusion UL_pro et LL_pro, et de langue pour le paramètre de corps de langue TB variant de -3 à +3 par pas de 0,5 écarts-type (haut : FR, bas : AE ; gauche : voyelles, milieu : consonnes analogues, droite : toutes les articulations).

La Figure 4 montre un mouvement vertical de la lèvre inférieure légèrement plus marqué pour le français que pour l'anglais. La composante dos de la langue *TD* pour les voyelles de l'anglais comme pour le corpus complet anglais montre clairement une capacité plus grande à se bomber vers le velum que pour le français. Ceci est vraisemblablement lié à la forme particulière bombée (*bunched*) vers l'avant du /ɹ/ anglais.

La Figure 5 semble montrer que les mouvements de la pointe de la langue sont plus localisés vers l'extrémité en français qu'en anglais: ceci pourrait être dû, entre autres raisons, à la

vélarisation associée au /ɬ/ sombre anglais, dont on peut observer une trace sur les mouvements du corps de langue associés à cette composante pour l'anglais (mais pas pour le français), comme on le voit sur la colonne de droite de la Figure 5.

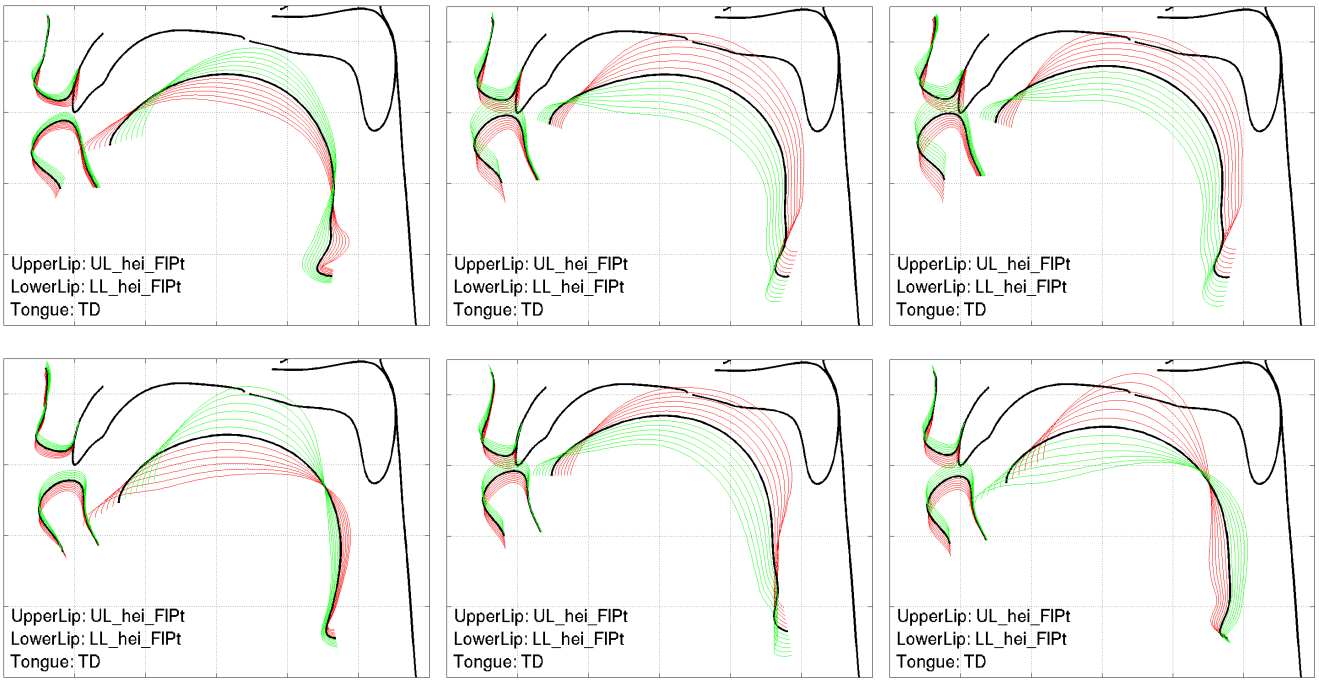


FIGURE 4 – Nomogrammes de lèvres pour les paramètres de hauteurs labiales *UL_hei* et *LL_hei*, et de langue pour le paramètre de dos de langue *TD* variant de -3 à +3 par pas de 0,5 écarts-type (haut : FR, bas : AE ; gauche : voyelles, milieu : consonnes analogues, droite : toutes les articulations).

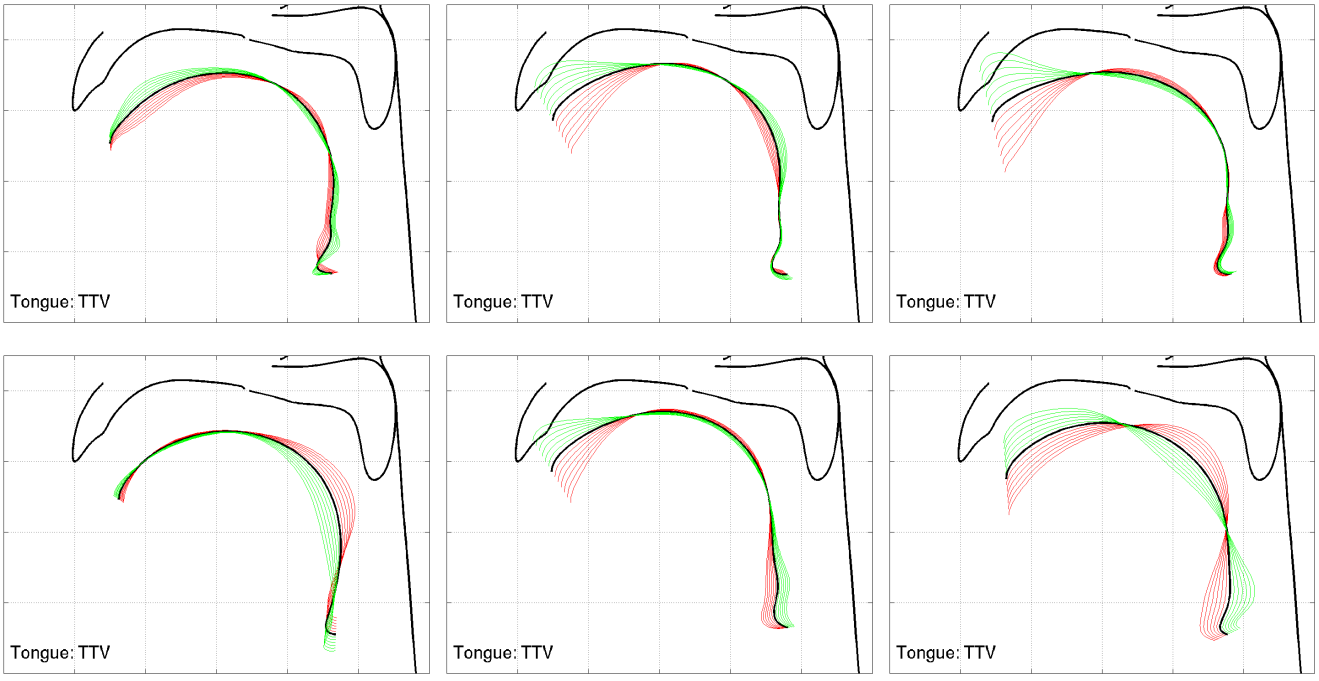


FIGURE 5 – Nomogrammes de langue pour le paramètre de pointe de langue vertical *TTV* variant de -3 à +3 par pas de 0,5 écarts-type (haut : FR, bas : AE ; gauche : voyelles, milieu : consonnes analogues, droite : toutes les articulations).

5 Perspectives d'évaluation par modélisation croisée

Afin d'évaluer le recouvrement entre les espaces articulatoires pour l'anglais et le français, nous avons suivi la méthodologie proposée par Serrurier, Badin, Barney, Boë & Savariaux (2012) pour comparer les articulations en parole et déglutition: nous avons construit un modèle pour chaque langue, basé chacun sur ses propres voyelles et consonnes. Nous avons ensuite utilisé chaque modèle pour reconstruire les articulations de l'autre langue : nous avons généré du français avec le modèle anglais et vice versa. Ces simulations n'ont pas encore permis de dégager des différences significatives de reconstruction croisées : il est possible que, même si leurs composantes possèdent des caractéristiques différentes, les deux modèles puissent servir à représenter les données aussi bien du français que de l'anglais.

6 Conclusions et autres perspectives

Nous avons acquis un large ensemble de contours articulatoires produits par un locuteur bilingue à la fois dans sa langue maternelle AE et une langue seconde FR. Nous avons comparé ses articulations en AE et FR, à savoir: (1) comparaisons directes deux à deux des contours, (2) comparaison des nomogrammes articulatoires obtenus à partir de divers modèles articulatoires élaborés à partir de différentes sous-ensembles de phonèmes, et (3) reconstruction articulatoire croisée des articulations d'une langue à partir du modèle élaboré avec l'autre langue. Malgré les contraintes imposées au locuteur par les limites de l'IRM statique, cette étude a généré une variété de résultats intéressants et utiles pour la recherche à l'intersection de la parole et du bilinguisme. Dans une prochaine étape, nous utiliserons des techniques d'IRM dynamique, et nous inclurons l'enregistrement simultané et l'analyse des sons associés avec les articulations. Outre les conditions plus naturelles d'élocution, cette méthode permettra d'obtenir une quantité de données beaucoup plus considérable et de pouvoir envisager des comparaisons dont on pourra évaluer plus rigoureusement le niveau de signification. Nous devrons également mettre en œuvre des techniques pour évaluer plus précisément le degré de bilinguisme du locuteur. A plus long terme, nous envisageons d'enregistrer plus de locuteurs / paires de langues avec des locuteurs bilingues à différents degrés. Des applications dans le domaine de l'apprentissage de la prononciation aidée par ordinateur (en anglais CAPT, *Computer Aided Pronunciation Training*) seront également explorées.

Remerciements

Nous remercions bien sincèrement Simon Crépel pour avoir tracé de manière soigneuse les contours articulatoires pour les deux corpus.

Références

BADIN, P. & SERRURIER, A. (2006). Three-dimensional modeling of speech organs: Articulatory data and models. In *IEICE Technical Report* pp. 29-34. Kanazawa, Japan: The Institute of Electronics, Information, and Communication Engineers.

BADIN, P., VALDÉS VARGAS, J.A., KONCKI, A., LAMALLE, L. & SAVARIAUX, C. (2013). Development and implementation of fiducial markers for vocal tract MRI imaging and speech articulatory modelling. In *Proccedings of Interspeech 2013,* pp. 1321-1325. Lyon, France.

ESCUDERO, P. (2007). Second-Language phonology: The role of perception. In *Phonology in context* (M.C. Pennington Ed.), pp. 109-134. Basingstoke, England: Palgrave Macmillan.

FLEGE, J.E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*, 47-65.

FROMKIN, V. (1964). Lip positions in American English vowels. *Language and Speech, 7*, 215-225.

HONIKMAN, B. (1964). Articulatory settings. In *In honour of Daniel Jones* (D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott & T.J.L. M, Eds.), pp. 73-84. London: Longman.

LINKER, W.J. (1982). *Articulatory and acoustic correlates of labial activity: a cross-linguistic study.* UCLA, Los Angeles, USA.

MENNEN, I., SCOBBIE, J.M., DE LEEUW, E., SCHAEFFLER, S. & SCHAEFFLER, F. (2010). Measuring language-specific phonetic settings. *Second Language Research, 26*(1), 13-41.

MONTAGU, J. (2002). L'articulation labiale des voyelles nasales postérieures du français : comparaison entre locuteurs français et anglo-américains. In *Actes des XXIV^èmes Journées d'études sur la Parole,* pp. 253-256. Nancy, France.

OH, S. (2002). Cross-language blending of /l/ gestures by Korean-English bilingual children. In *Proceedings of the 26^th annual Boston University Conference on Language Development* (B. Skarabela, S. Fish & A.H.-J. Do, Eds.), pp. 473-484.

SERRURIER, A., BADIN, P., BARNEY, A., BOË, L.-J. & SAVARIAUX, C. (2012). The tongue in speech and feeding: Comparative articulatory modelling. *Journal of Phonetics, 40*, 745-763.

TEPPERMAN, J., BRESCH, E., KIM, Y.-C., LEE, S., GOLDSTEIN, L. & NARAYANAN, S.S. (2009). An articulatory analysis of phonological transfer using real-ime MRI. In *Proccedings of Interspeech 2009*, pp. 700-703. Brighton, UK.

WILSON, I.L. (2006). *Articulatory settings of French and English monolingual and bilingual speakers.* University of British Columbia, Vancouver, Canada.

ZERLING, J.-P. (1992). Frontal lip shape for French and English vowels. *Journal of Phonetics, 20*, 3-14.

Implémentation d'un nouvel algorithme de codage de la parole basé sur la transformation en paquets de framelettes

Souhir Bousselmi Kaïs Ouni
Unité de Recherche Systèmes Mécatroniques et Signaux, SMS
Ecole Supérieure de Technologie et d'Informatique (ESTI)
Université de Carthage. Tunis, Tunisie.
souhir.bousselmi@laposte.net, kais.ouni@esti.rnu.tn

RÉSUMÉ

Dans ce papier nous proposons d'étudier un algorithme de codage de la parole en bande étroite et à débit fixe basé sur la transformation en paquets de framelettes. Cette nouvelle transformation permet d'obtenir une bonne localisation temps-fréquence et une décomposition en sous-bande selon un critère donné ou une application désirée. Une étude comparative avec la transformation en paquets d'ondelettes a été menée. Les résultats obtenus montrent que la qualité du codage basé sur la transformation en paquets de framelettes est supérieure à celle basé sur la transformation en paquets d'ondelettes.

ABSTRACT

Implementation of a new speech coding algorithm based on the Framelet Packet Transform

In this paper we propose to study a narrow band speech coding algorithm at a fixed bit rate based on the framelet packet transform. This new transformation allows powerful time-frequency localization and sub-band decomposition according to a given criterion or a desired application. A comparative study with wavelet packet transform has been conducted. The obtained results show that the speech coding by framelet packet transform provides a better quality compared to wavelet packet transform.

MOTS-CLÉS : Codage de la parole, frame d'ondelette, transformation en paquets de framelette, transformation en paquets d'ondelette.

KEYWORDS: Speech coding, wavelet frame, framelet packets transform , wavelet packets transform.

1 Introduction

Le codage par transformée des signaux de la parole est une application du traitement de la parole en pleine expansion. La particularité essentielle de ce type de codage est la modification de la représentation temporelle du signal d'entrée par une représentation temps-fréquence. Ce changement d'espace de représentation réduit la redondance due à la corrélation du signal ce qui rend la quantification plus efficace que la quantification directe des échantillons du signal (Dia, 1993) (Mariani, 2002). La transformation en paquets d'ondelettes est une représentation

temps-fréquence qui a été très utilisée en littérature dans l'élaboration des codeurs de la parole et audio (Kastantin, 1996) (Sinha et Tewfik, 1993). Toutefois, cette transformation présente des inconvénients qui limitent ses performances dans le codage de la parole. En effet, les ondelettes orthogonales à support compact ne sont pas symétriques et ne possèdent pas un déphasage linéaire, à l'exception de l'ondelette triviale de Haar. De plus, vu l'échantillonnage critique les ondelettes orthogonales ne sont pas invariantes par translation (Abdelnour et Selesnick, 2005) (Selesnick, 2001). Pour remédier à ces inconvénients il est intéressant d'intégrer la théorie des frames d'ondelettes. Les frames d'ondelettes permettent aussi d'avoir une reconstruction parfaite et robuste des signaux et une forte résistance aux bruits de quantification (Goyal et Vetterli, 1998). La construction d'une frame d'ondelette est basée sur l'analyse multirésolution et le principe d'extension unitaire (Benedetto et Li, 1998)(Daubechies et Shen, 2003). Les représentations temps-fréquences issues de la théorie des frames d'ondelettes telles que la transformée en framelette et la transformation en paquets de framelettes sont devenues des outils primordiaux dans l'analyse, le codage et le débruitage des signaux (Bousselmi et Ouni, 2010) (Bousselmi et Ouni, 2011) (Bousselmi et Ouni, 2012) (Parker, 2005). Malgré l'importance de la transformation en framelette, elle ne permet pas d'avoir des bonnes résolutions fréquentielles en hautes fréquences ainsi qu'en basses fréquences. En effet, dans cette transformation on ne décompose que les sorties du filtre passe-bas. Une bonne résolution est nécessaire pour améliorer la qualité de la parole perçue. Les performances de la transformation en framelette dans le codage de la parole sont limitées même en augmentant le nombre de niveau de décomposition. La transformée en paquets de framelettes a été introduite pour pallier ces problèmes (Parker, 2005) (Lu et Fan, 2011)(SUQI, 2009). Dans (Bousselmi et Ouni, 2012) nous avons étudié les performances de la transformation en paquets de framelettes dans le codage de la parole, où nous avons considéré un schéma de codage simplifié et une décomposition en paquets de framelettes au niveau 3. Dans ce papier, nous proposons d'étendre cette étude en considérant un codeur de parole plus développé basé sur la transformation en paquets de framelettes. La particularité essentielle de cet algorithme réside dans l'utilisation d'un arbre de décomposition en 17 sous-bandes les plus proches des bandes critiques, pour mieux s'adapter aux propriétés de l'oreille humaine. Du fait que le spectre du signal évolue constamment au cours du temps, nous envisageons une procédure d'allocation dynamique des bits codant les coefficients de chaque trame du signal. Il est à noter que notre objectif n'est pas de concurrencer les meilleurs codeurs de parole de la bande étroite (téléphonique) mais de montrer l'avantage de l'utilisation de la transformation en paquets de framelettes. Une étude comparative avec l'algorithme de codage basé sur la transformation classique en paquets d'ondelettes a été menée. Ce papier est organisé comme suit : dans la deuxième section nous présentons la méthode que nous avons adoptée pour obtenir la transformation en paquets de framelettes. Dans la troisième section nous présentons les différentes étapes d'implémentation du codeur. Dans la quatrième section nous présentons les résultats de simulation du codeur. Ce papier se termine par une conclusion.

2 Transformation en paquets de framelette

La transformation en paquets de framelettes est construite à partir d'un traitement répété dans la bande passe-bas ainsi que dans les deux bandes passe-haut du banc de filtres sur-échantillonné de la figure 1 (Lu et Fan, 2011). Dans cette figure, h_0 est un filtre passe-bas et les deux filtres h_1 et h_2 sont de type passe-haut. Chaque bande du banc de filtres est décimée par 2. Du fait que la frame est ajustée, les filtres de synthèse sont donnés par l'inverse des filtres d'analyse. Nous présentons dans la figure 2 l'arbre de décomposition dyadique correspondant à une analyse en

paquets de framelettes au niveau 3. Les indices 0, 1 et 2 associés à chaque feuille/nœud de l'arbre correspondent respectivement aux filtres d'analyses h_0, h_1, h_2 décimés par 2. Dans cet arbre, les nœuds auxiliaires associés au filtre de bande passante étroite h_1 seront considérés comme des nœuds terminaux qui ne seront pas considérés dans l'étape d'analyse. Les nœuds associés aux filtres h_0 et h_2 sont appelés les nœuds dyadiques. La condition principale pour qu'un arbre en paquets de framelettes soit admissible est que chaque nœud dyadique possède 0 ou 3 enfants, et chaque nœud auxiliaire possède 0 enfant. Pour un niveau de décomposition donné, chaque nœud dyadique représente une étape d'analyse qui divise chaque sous-signal en deux bandes de fréquences séparées, où la largeur de chaque bande de fréquence est égale à la moitié de la largeur de la bande du niveau précédant (Parker, 2005).

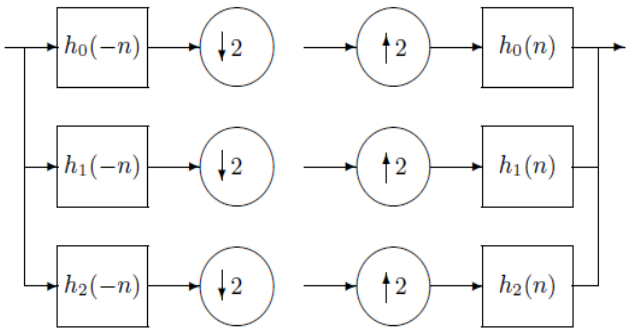


FIGURE 1: Banc de filtres sur-échantillonné à trois sous-bandes

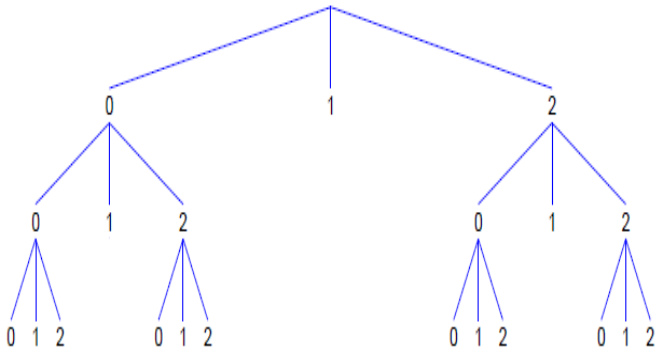


FIGURE 2: Arbre de décomposition au niveau 3 correspondant à la transformation en paquets de framelettes

3 Implémentation du codeur

Notre travail concerne l'introduction d'une transformation en paquets de framelettes dans la conception d'un codeur de parole en bande étroite et à débit fixe dont le schéma de principe est présenté dans la figure 3. La première étape de l'encodeur consiste à déterminer les coefficients issus de cette transformée pour chaque trame du signal. Après normalisation des coefficients et allocation dynamique des bits, on réalise une quantification scalaire optimale des coefficients normalisés avant leur transmission au récepteur. Le décodeur réalise les étapes inverses de celles de l'encodeur. En premier lieu, le décodeur déduit les résultats de la procédure d'allocation des bits. Ensuite, il fait la quantification inverse, qui consiste à recalculer les valeurs des coefficients à partir de leurs codes. Puis à l'aide d'une transformation inverse une trame du signal est reconstruite. Pour former le signal reconstruit on somme les trames voisines.

3.1 Calcul des coefficients de la transformation en paquets de framelettes

Le calcul des coefficients de la transformation en paquets de framelettes est effectuée en se basant sur le banc de filtres sur-échantillonnés de la figure 1. Il dépend de la sélection des filtres $\{h_0; h_1; h_2\}$ et du choix de l'arbre de décomposition. Dans le cas du codage de la parole l'évaluation subjective est la plus significative, de ce fait il est indispensable de tenir compte des propriétés psychoacoustiques du système auditif humain dans le choix de la meilleure base des paquets de framelettes. Le modèle de perception auditive est basé sur l'analyse en bandes critiques, effectuée par l'oreille interne (Zwicker et Feldtkelleri, 1981). De ce fait, le spectre du

signal doit être découpé selon ces bandes critiques, afin de prendre en considération la résolution perceptuelle de l'oreille. A cause des limites en résolution fréquentielle donnée par l'algorithme de paquets de framelettes, nous choisissons comme base celle qui conduit au découpage en fréquence le plus proche des bandes critiques. Pour cela nous avons découpé le signal en 17 sous-bandes. L'arbre de décomposition en 17 sous-bandes est présenté dans la figure 4. Nous présentons dans le tableau 1 l'intervale fréquentielle correspondant aux 17 sous-bandes. Nous avons considéré les filtres $\{h_0; h_1; h_2\}$ dont les coefficients sont donnés par Selesnick (Selesnick, 2004).

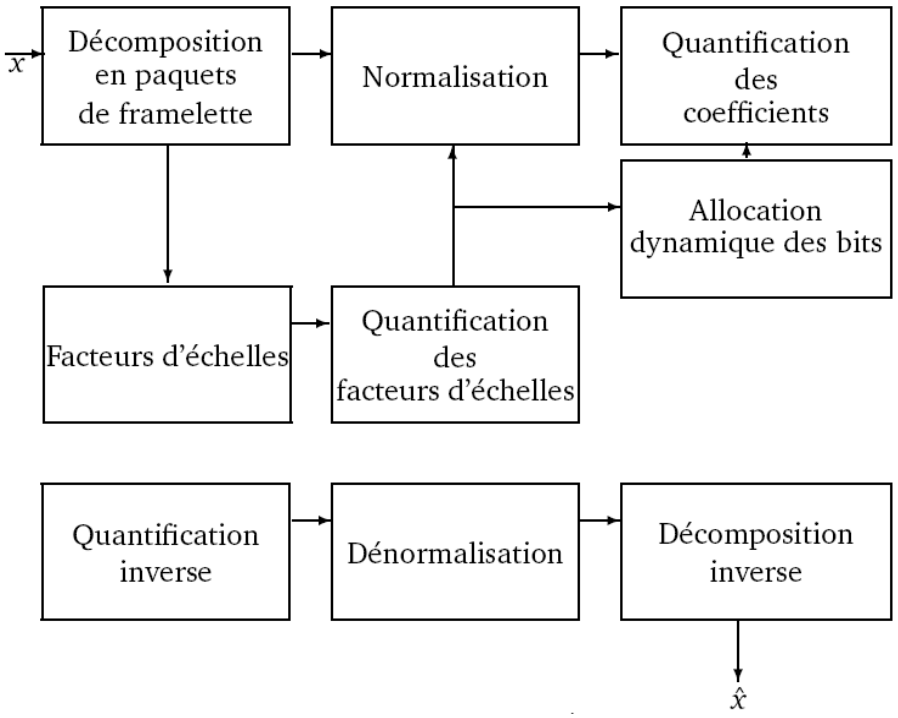


FIGURE 3: Structure du codeur/décodeur

3.2 Facteur d'échelles et allocation dynamique des bits

Les facteurs d'échelles servent à normaliser les coefficients avant leur quantification, afin de diminuer l'intervalle dynamique des quantificateurs. Ils sont donnés par la variance de chaque sous-bande de l'arbre de décomposition en paquets de framelettes et quantifiés uniformément selon une loi logarithmique avec quatre bits. La procédure d'allocation dynamique des bits permet de répartir d'une manière récursive le débit disponible sur tous les coefficients d'une trame du signal. À chaque itération, un bit de plus est assigné au groupe de coefficients qui possède la variance maximale, puis cette variance est divisée par un facteur f égal à 4 et le nombre total de bits diminue d'une valeur correspondant à la taille du groupe. Chaque groupe comprend la totalité des coefficients d'une sous-bande. Cette opération est répétée tant que les bits ne sont pas tous alloués.

3.3 Quantification scalaire optimale

La quantification scalaire optimale des coefficients consiste à représenter tout vecteur x par un vecteur y appartenant à un ensemble fini appelé dictionnaire. La création du dictionnaire est faite à l'aide de l'algorithme de Lloyd-Max dont le principe est d'améliorer successivement un dictionnaire initial formé à partir d'une séquence d'apprentissage par application de deux

conditions d'optimalité : la condition du plus proche voisin et la condition du centroïde. Le critère d'arrêt est la stabilisation de la distorsion moyenne. Dans le cadre de notre travail, nous avons utilisé la base TIMIT pour former la séquence d'apprentissage, les vecteurs constituants cette séquence correspondent à un ensemble de coefficients de la transformation en paquets de framelettes.

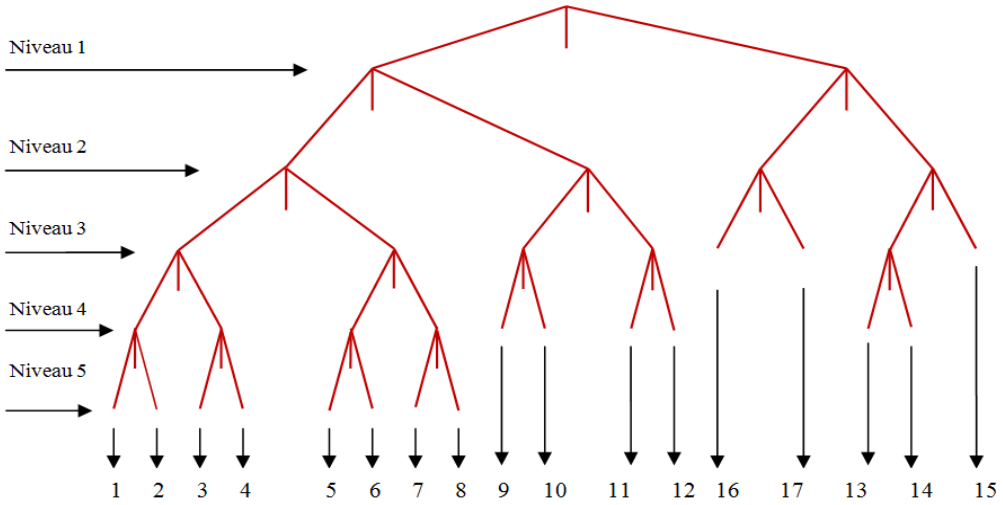


FIGURE 4: Arbre de décomposition en 17 sous-bandes correspondant à la transformation en paquets de framelettes couvrant la bande fréquentielle [0 - 4000 Hz].

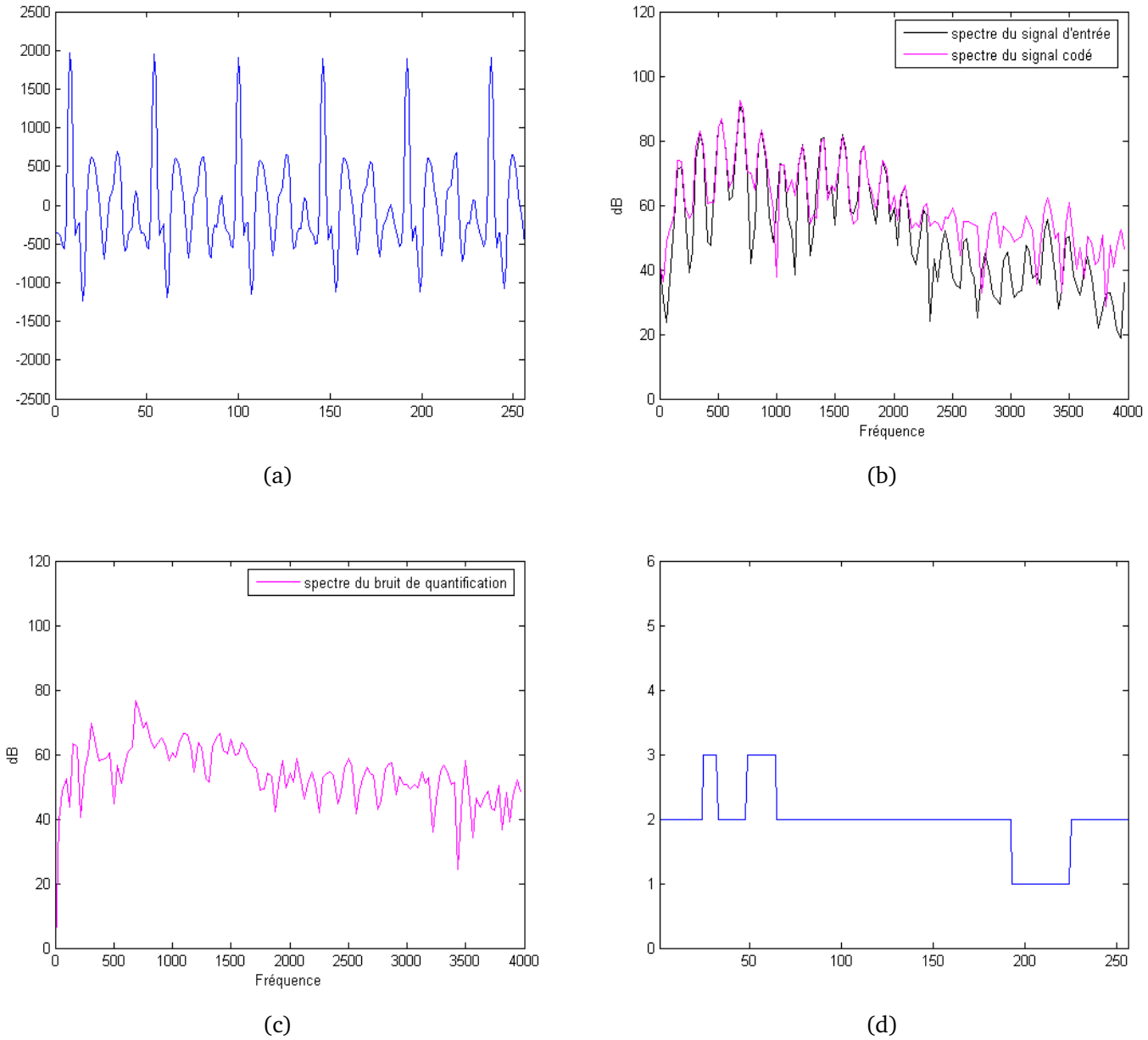
TABLE 1: L'intervale fréquentiel correspondant aux 17 sous-bandes.

$N°$ sous-bande	l'intervalle fréquentiel (Hz)
1	0-125
2	125-250
3	250-375
4	375-500
5	500-625
6	625-750
7	750-875
8	875-1000
9	1000-1250
10	1250-1500
11	1500-1750
12	1750-2000
13	2000-2250
14	2250-2500
15	2500-3000
16	3000-3500
17	3500-4000

4 Simulation et résultats

Dans cette étude, nous avons effectué des simulations d'un codeur de parole en bande téléphonique basé sur la transformation en paquets de framelettes fonctionnant avec un débit fixe, nous avons

étudié des débits entre 8 Kbits/s et 16 Kbits/s. Nous avons mené notre étude sur des signaux issus de la base TIMIT, échantillonnés à 16 kHz. Compte tenu du fait que le codeur proposé est de type bande téléphonique, nous avons été amenés à faire un sous-échantillonnage pour avoir une fréquence de 8 kHz (Fisher et Goudie-Marshall, 1986). Chaque signal est segmenté en trames de 256 échantillons avec recouvrement de 20 échantillons et chaque segment est pondéré par la fenêtre de Hanning avant transformation.



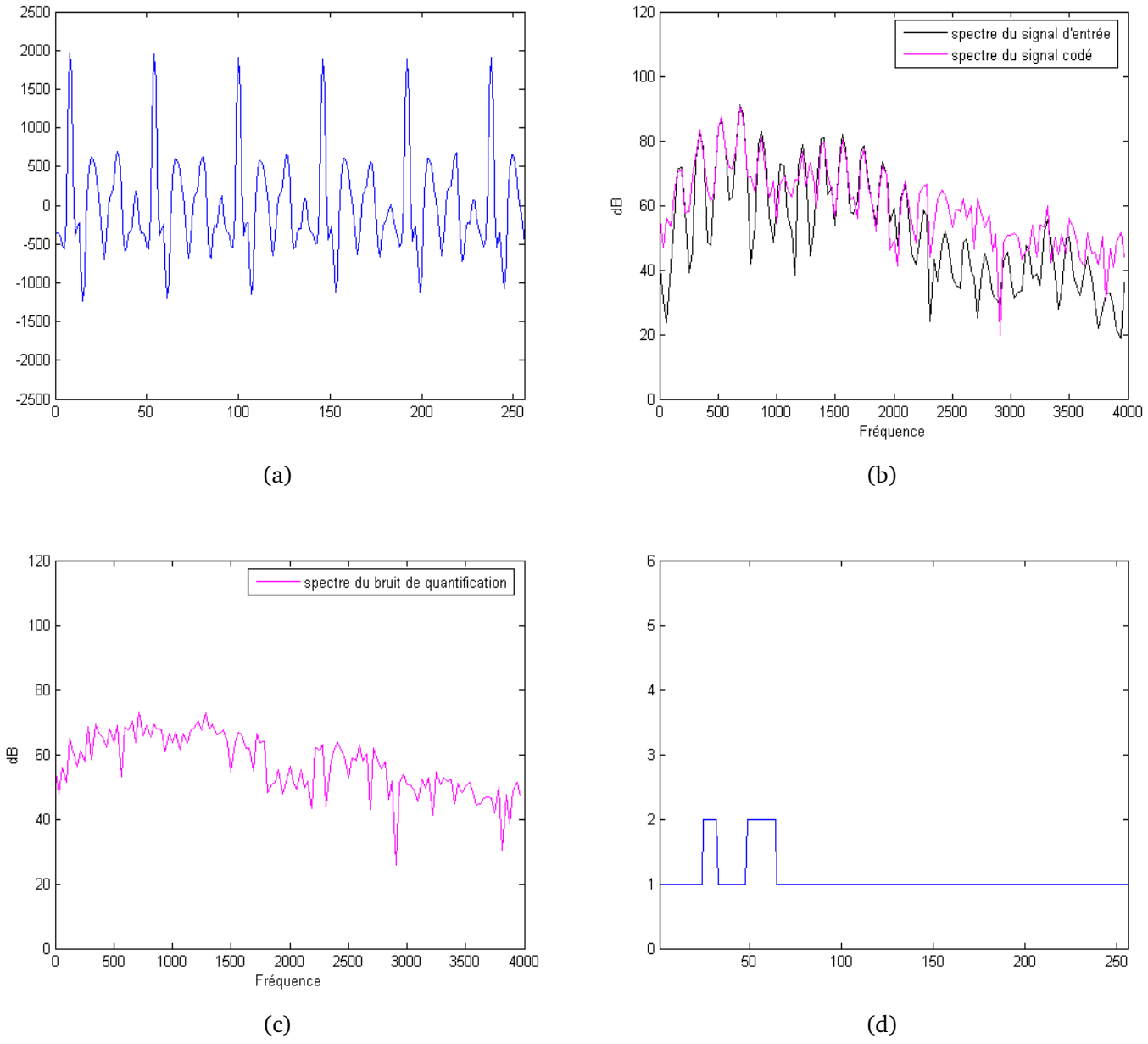
(a) (b)

(c) (d)

FIGURE 5: Exemple d'un signal codé avec un débit de 16 kbits/s : a) Signal original b) Spectres des signaux d'entrée et codé c) Spectre du bruit de quantification d) Résultat de l'allocation dynamique des bits.

Nous présentons dans les figures 5 et 6 un exemple d'une trame voisée codée respectivement avec les débits de 16 Kbits/s et 8 kbits/s. Nous avons ainsi visualisé les spectres de la trame d'origine, la trame codée, le spectre du bruit de quantification correspondant et le résultat de l'allocation dynamique des bits. Nous remarquons l'existence d'un bruit de codage entre les harmoniques et ce bruit augmente si on diminue le débit. Afin d'étudier les performances de l'algorithme de codage par transformation en paquets de framelettes nous avons codé 20 phrases issues du corpus

TIMIT, puis nous avons calculé les valeurs du rapport signal sur bruit segmental RSBseg, les valeurs du PESQ (*perceptual evaluation of speech quality*) et les valeurs du rapport signal sur bruit segmental RSBseg à fréquence pondérée (fwSNRseg) pour des débits compris entre 8 Kbits/s et 16 Kbits/s. Ces critères sont les plus utilisés en codage de la parole, ils sont définis dans les références (Loizou, 2007)(UIT-T, 2000). Une étude comparative avec l'algorithme de codage basé sur la transformation en paquets d'ondelettes a été menée. Les résultats obtenus sont présentés respectivement dans les figures 7,8 et 9.



(a) (b)

(c) (d)

FIGURE 6: Exemple d'un signal codé avec un débit de 8 kbits/s : a) Signal original b) Spectres des signaux d'entrée et codé c) Spectre du bruit de quantification d) Résultat de l'allocation dynamique des bits.

Il est à noter que les valeurs du RSBseg, du PESQ et du fwSNRseg dans le cas de la transformation en paquets de framelettes sont supérieures à celles obtenues dans le cas de la transformation en paquets d'ondelettes pour un même débit. D'autre part, les tests informels du codeur montrent que la qualité de la parole codée par la transformation en paquets de framelettes est plus fidèle que celle codée par la transformation en paquets d'ondelettes. Une bonne qualité de la parole codée est obtenue pour un débit de 16 Kbits/s. Certaines dégradations sur la qualité sont perçues. Ils sont de plus en plus remarquables lorsqu'on diminue le débit au-dessous de 12 Kbits/s.

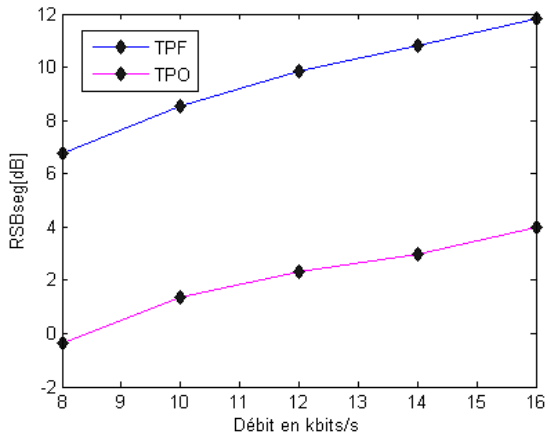


FIGURE 7: Variation du RSBseg en fonction du débit

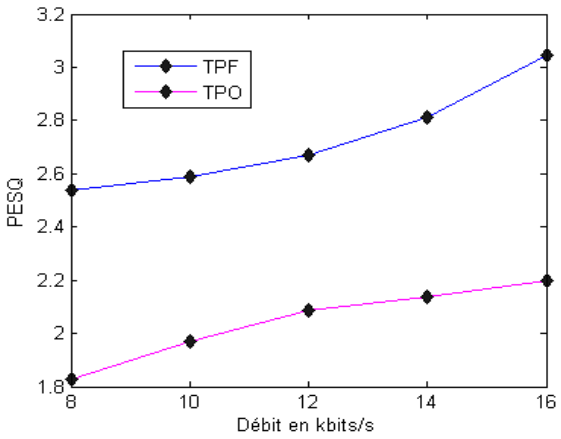


FIGURE 8: Variation du PESQ en fonction du débit

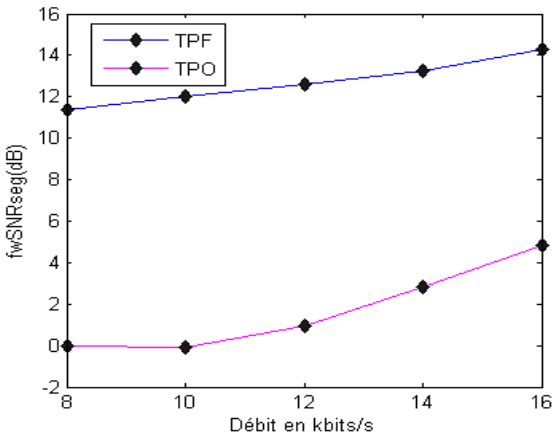


FIGURE 9: Variation du fwSNRseg en fonction du débit

5 Conclusion

Dans cet article, nous avons présenté une étude portant sur le codage de la parole en bande étroite et à débit fixe. Sa particularité réside dans l'utilisation de la transformation en paquets de framelettes. Une étude comparative avec la transformation en paquets d'ondelettes a été menée. Les résultats obtenus montrent l'intérêt de la transformation en paquets de framelettes dans le codage de la parole.

Références

ABDELNOUR, A. F. et SELESNICK, I. W. (2005). Symmetric nearly shift-invariant tight frame. *In IEEE Transactions on Signal Processing*, volume 53, pages 231–239.

BENEDETTO, J. et LI, S. (1998). The theory of multiresolution analysis frames and applications to filter banks. *In Applied and Computational Harmonic Analysis*, volume 5, pages 389–427.

BOUSSELMI, S. et OUNI, K. (2010). Speech signal reconstruction based on the symmetric tight wavelet frame decomposition. *In International Congress on Image and Signal Processing (CISP)*, volume 17, pages 3453–3456.

BOUSSELMI, S. et OUNI, K. (2011). Performance evaluation of the symmetric tight wavelet frame in speech coding. *In International Multi-Conference on Systems Signals and Devices (SSD)*, pages 1–5.

BOUSSELMI, S. et OUNI, K. (2012). Etude pour l'amélioration de la parole codée par transformation en paquets de framelette serrée. *In Actes de la conférence conjointe JEP-TALN-RECITAL*, volume 1, page 473–480.

DAUBECHIES, I. Han, B. R. A. et SHEN, Z. (2003). Framelets : Mra-based constructions of wavelet frames. *In Applied and Computational Harmonic Analysis*, volume 14, pages 1–46.

DAUBECHIES, I. (1992). Ten lectures on wavelets. *In CBMS Conference Series in Applied Mathematics*, volume 61.

DIA, H. (1993). Codage par transformée de la parole à bande élargie (0-7khz). *In Thèse de Doctorat, Institut National Polytechnique de Grenoble*.

FISHER, W. Dodington, G. et GOUDIE-MARSHALL, K. (1986). The timit-darpa speech recognition research database : Specification and status. *In DARPA Workshop on Speech Recognition*.

GOYAL, V.K Thao, N. et VETTERLI, M. (1998). Quantized overcomplete expansions in r^n : analysis synthesis, and algorithms. *In IEEE Transaction on Information Theory*, volume 44, pages 16–31.

KASTANTIN, R. (1996). Codage de la parole basé sur la transformation en ondelettes. *In Thèse de Doctorat, Institut National Polytechnique de Grenoble*.

LOIZOU, P. (2007). Speech enhancement : Theory and practice. *In Edition CRC*.

LU, D. et FAN, Q. (2011). A class of tight framelet packets. *In Czechoslovak Mathematical Journal*, volume 61, pages 623–639.

MARIANI, J. (2002). Analyse, synthèse et codage de la parole. *In Edition Hermes*.

PARKER, S. (2005). Cfs : Time-frequency representations of acoustic signals based on redundant wavelet methodologies. *In Thesis, University of Wisconsin Madison*.

PETUKHOV, A. (2003). Symmetric framelets. *In Constructive Approximation*, volume 19, pages 309–328.

SELESNICK, I. W. (2001). Smooth wavelet tight frames with zero moments. *In Applied and Computational Harmonic Analysis*, volume 10, pages 163–181.

SELESNICK, I. W. (2004). Symmetric wavelet tight frames with two generators. *In Applied and Computational Harmonic Analysis*, volume 17, pages 211–225.

SINHA, D. et TEWFIK, A. H. (1993). Low bit rate transparent audio compression using adapted wavelets. *In IEEE Transactions on Signal Processing*, volume 41, pages 3464–3479.

SUQI, P. (2009). Tight wavelet frame packet. *In Thesis, Departement of Mathematics, National University of Singapore*.

UIT-T (2000). Perceptual evaluation of speech quality (pesq) : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *In ITU-T Recommendation P.862*.

ZWICKER, E. et FELDTKELLERI, R. (1981). Psychoacoustique : L'oreille, récepteur d'information. *In Edition Masson*.

Etude des HMM paramétriques pour la reconnaissance de parole en environnement bruité

Mathieu Radenen[1] Pierrick Milhorat [3] Thierry Artières [1] Jerôme Boudy [2]
Gérard Chollet [3]
(1) Université Paris 6, 4 place Jussieu 75005 Paris
(2) Télécom-SudParis, 9 Rue Charles Fourier, 91000 à Evry
(3) Télcom-ParisTech, 37-39 rue Dareau 75014 Paris
mathieu.radenen@lip6.fr, pierrick.milhorat@telecom-paristech.fr,
thierry.artieres@lip6.fr, jerome.boudy@telecom-sudparis.eu,
gerard.chollet@telecom-paristech.fr

RÉSUMÉ

Nous décrivons une étude de la pertinence de HMM paramétriques, dans lesquels les paramètres des lois gaussiennes dépendent de variables externes dites contextuelles, pour une tâche de reconnaissance en environnement bruité. Les résultats montrent d'une part l'intérêt de ce type de modélisation pour des variables contextuelles de différentes natures, des variables calculées à partir du signal lui-même ou bien correspondant à des informations additionnelles sur le signal.

ABSTRACT

A study of Parameterized HMM on noisy speech recognition

We describe an investigation of Parametric HMM, whose Gaussian distribution depend on external variables named contextual variables, for a noisy speech recognition task. We show the relevance of few contextual variables for the task, either computed from the signal itself or consisting of additional information on the signal.

MOTS-CLÉS : HMM Paramétrique, MRHMM, reconnaissance de parole en milieu bruité.

KEYWORDS: Parametric HMM, PHMM, MRHMM, noisy speech recognition.

1 Introduction

Les travaux en reconnaissance vocale traitent usuellement le bruit de deux manières complémentaires. D'un coté le pré-traitement a pour but de normaliser ou filtrer un signal pour le rapprocher de ceux sur lesquels les modèles ont été appris. Parmi ces traitements, on compte la séparation bruit/voix (Álvarez *et al.*, 2011) qui permet soit de filtrer le signal utile, soit de labéliser des segments à rejeter. Ces méthodes se rattachent à celles de rehaussement de la parole (Mohammadiha *et al.*, 2013) et de normalisation (Variani et Schaaf, 2011) qui s'attachent à projeter les signaux vers des espace communs inter-locuteurs et indépendants de l'environnement sonore. Les techniques de séparation de sources (Geiger *et al.*, 2013), communes à plusieurs systèmes du challenge CHiME traitent un signal pour en extraire les différentes composantes décorrélables. Pour finir, certains articles présentent l'utilisation de nouveaux coefficients de description acoustique comme une source d'amélioration des performances de reconnaissance en

milieu bruité. D'un autre côté, les modèles eux-mêmes peuvent aussi être adaptés. L'adaptation de type MAP (Maximum A Posteriori) ou MLLR (Maximum Likelihood Linear Regression) modifie les densités de probabilité d'émission des observations pour chaque état à partir de quantités réduites de données. Les modèles acoustiques peuvent également être initialement entraînés sur des données bruitées, artificiellement ou non. Cela implique cependant la nécessité de conditions similaires lors du décodage.

Nous proposons ici d'attaquer le problème de la reconnaissance de parole bruitée pour cette seconde approche, en construisant un HMM dont la distribution de probabilité dépend de variables contextuelles (i.e. le contexte noté $\boldsymbol{\theta}$), et en explorant divers types de ces variables contextuelles. Dans les HMM paramétriques proposés initialement par Bobick et al (Wilson et Bobick, 1999) la moyenne des distribution Gaussiennes varie linéairement en fonction de variables contextuelles. La loi de probabilité d'émission dépend donc non seulement de l'état mais aussi du contexte. Ainsi un modèle peut exprimer de nombreuses distributions avec un nombre limité de paramètres. La motivation de ce type de modèles proposés pour l'analyse et la reconnaissance de gestes dans les vidéos réside dans l'idée qu'une part importante de la variabilité dans les séquences d'observation peut être due à quelques variables explicatives, observables ou non, fixes ou variant dans le temps. Une phrase peut par exemple être pronnoncée très différemment en fonction de l'émotion du locuteur. Un geste peut être éxécuté avec plus ou moins d'amplitude suivant la morphologie ou l'age de la personne le réalisant...

Le formalisme des PHMM est similaire à d'autres approches de la littérature proposées pour le signal de parole, telles que (Fujinaga *et al.*, 2001), (Yu *et al.*, 2009), (Cui et Gong, 2007), il a été étendu dans (Radenen et Artières, 2014) à des modèles discriminants. Les approches de cette famille se différencient par la nature de la dépendance entre les paramètres du HMM et les variables contextuelles, par la faculté à utiliser des variables de contexte dynamiques, (i.e. évoluant dans le temps), par leur faculté à inférer les variables de contexte. Nous utilisons ici des PHMM simples, équivalents dans cette forme à des MRHMM (multiple regression HMM) (Fujinaga *et al.*, 2001), pour apprendre des modèles dont les paramètres dépendent du rapport signal à bruit, de deux autres mesures dérivées du signal, l'applatissement spectral et la vitesse d'élocution, mais aussi de l'identité du locuteur.

Les travaux de (Yu *et al.*, 2009), (Cui et Gong, 2007), (Radenen et Artieres, 2012), (Radenen et Artières, 2014) suggèrant que la paramétrisation de la covariance apporte des gains beaucoup moins significatifs que la paramétrisation de la moyenne, nous basons notre étude sur l'utilisation de (Wilson et Bobick, 1999) dont seules les moyennes dépendent des variables contextuelles, d'autant qu'il s'agit d'une approche plus simple offrant un meilleur compromis temps de calcul/performance.

2 HMM paramétriques

Les HMM paramétriques (PHMM) ont été introduit par (Wilson et Bobick, 1999) pour la reconnaissance de gestes. Ce sont des HMM dont les distributions de probabilité Gaussiennes dépendent d'une ou plusieurs variables contextuelles que nous notons θ (un réel ou un vecteur de réels). Dans sa forme la plus simple, aussi connue sous le nom de MRHMM (Fujinaga *et al.*, 2001), un PHMM est un HMM dans lequel la moyenne de la probabilité d'émission Gaussienne dans l'état j est défini comme une fonction linéaire des variables de contexte :

$$\hat{\mu}_j(\boldsymbol{\theta}) = W_j \times \theta + \bar{\mu}_j$$

avec W_j une matrice de transformation linéaire et $\bar{\mu}_j$ un vecteur que l'on peut imaginer comme la moyenne de la Gaussienne dans le cas de HMM Gaussiens conventionnels. $\boldsymbol{\theta}$ est un vecteur de variables. Potentiellement, n'importe quelles variables contenant de l'information sur la séquence observée peuvent être utilisées. Dans la suite du document, nous étudierons différents types de paramétrisation (i.e variables contextuelles) qui s'avèreront utile en reconnaissance de parole. Lorsque la dimension des variables contextuelles θ ne dépasse pas la dimension des observations (ce qui est le cas pour nos expériences), la complexité des HMM paramétriques en apprentissage et en inférence est du même ordre de grandeur qu'un HMM à covariance pleines.

3 Variables contextuelles

Nous nous présentons ici les variables contextuelles que nous avons sélectionné pour nos expériences (section 5). Nous avons retenus de types de variables. Les variables intra signal, concernent des quantités dérivées ou estimables directement à partir du signal. Les variables extra signal, représentent des quantités non calculables à partir du signal seul.

3.1 Variables dérivées des signaux

3.1.1 Signal sur bruit

Yifan Gong (Gong, 1995) a montré que la distribution des paramètres d'un HMM (qui caractérise un signal de parole) varie grandement en fonction des conditions d'enregistrement, ce qui peut causer une perte importante de performance en reconnaissance. Xiaogond Cui et Yifan (Cui et Gong, 2007) ont par ailleurs entrainé un modèle acoustique artificiellement corrompu par différents niveaux de bruits (SNR) et ont expérimentalement observé la variation de la moyenne obtenue suivant les conditions de bruit, et la forte corrélation de la moyenne des gaussiennes avec le SNR. Nous avons donc choisi d'expérimenter l'utilisation de cette mesure de signal sur bruit (SNR) comme variable de paramétrisation. La figure 1 illustre ce phénomène et montre l'évolution de la moyenne en fonction du SNR dans les états d'un HMM appris sur les réalisations du phonème /c/ de la base CHIME.

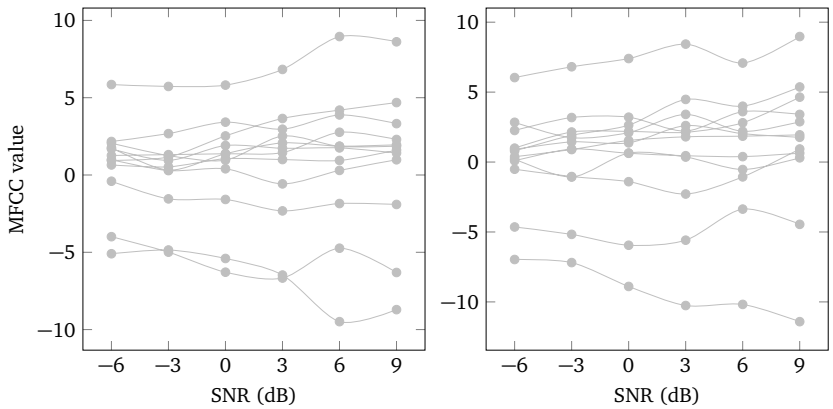


FIGURE 1: Moyenne des Gaussiennes des 2 premiers états d'un HMM appris sur la lettre 'c' (séquence phonétique /se/) du corpus CHiME en fonction du SNR. Ces moyennes représentent les 12 coefficients cepstraux (MFCC).

3.1.2 Spectral flatness

Nous avons également exploré l'intérêt de la Spectral Flatness Measure (SFM) qui représente le taux de composantes sinusoidales dans chaque bande de fréquences qui est calculé directement à partir du signal. Cette mesure est donnée par le rapport entre la moyenne géométrique et la moyenne arithmétique du spectre :

$$SFM = \frac{\left(\prod_{k\in \mathrm{K}} a(k)\right)^{1/k}}{\sum_{k\in \mathrm{K}} a(k)}$$

où $a(k)$ est l'amplitude du spectre à la fréquence k. Pour un signal tonal, la SFM est proche de 0. Pour un signal bruité, la SFM est proche de 1. Elle est calculée au niveau de chaque phrase.

3.2 Variables externes

Une variable externe peut apporter une information forte orthogonale ou du moins complémentaire à l'information extractible du signal et peut donc s'avérer pertinente. Dans le cadre de signaux de parole, il peut s'agir du sexe de la personne, de son âge, de son état émotionnel, de fatigue ou de stress, ou encore de l'identité du locuteur.

Dans le cadre de cette étude nous nous sommes donc focalisés sur les variables contextuelles disponibles pour le jeu de données considéré (compétition CHIME). Il s'agit tout d'abord de la vitesse d'élocution, qui peut se mesurer par exemple en nombre de syllabe par seconde, estimée ici par la longueur de l'élocution divisée par le nombre de mots. Nous disponsons et utilisons également l'identité du locuteur en test, car cela entre dans le cadre des conditions de la compétition CHiME.

4 Protocole expérimental

Les données viennent du corpus CHiME (Vincent *et al.*, 2013) (Barker *et al.*, 2013). Elles contiennent des enregistrements dans un environnement domestique. Le corpus est composé de 34 locuteurs (18 hommes et 16 femmes) lisant des phrases de 6 mots dont la grammaire est définie par : <couleur :4> <préposition :4> <lettre :25> <chiffre :10> <adverbe :4> où le nombre entre chevrons indique l'ensemble des choix possibles. Chaque phrase est ensuite mélangée à 6 différents niveaux de bruits (SNR = -6, -3, 0, 9 dB). Les détails sur le bruit de fond et les protocoles d'enregistrement sont disponibles dans (Christensen *et al.*, 2010). Toutes les données sont acquises au format wav 16 bits, échantillonnées à 16kHz.

4.1 Tâche

La tâche est de reconnaître la lettre et le chiffre de chaque phrase pour chaque niveau de bruit. Le critère de performance est le taux de reconnaissance de mot. L'ensemble d'apprentissage contient 500 phrases prononcées par 34 locuteurs. L'ensemble de test contient 3600 phrases (600 phrases

× 6 SNR). Ces ensembles sont tous deux annotés au niveau de chaque mot par le système de référence (un HMM par mot) en alignement forcé. L'ensemble de validation contient également 3600 phrases que nous avons annotées de la même façon.

4.2 Système de référence

Le système de référence du challenge CHiME est un HMM gauche droite avec 7 Gaussiennes à covariance diagonale. Il est entrainé avec HTK (Young *et al.*, 2006) sur le corpus bruité (données bruitées à 6 différents SNRs). Un modele est appris par mot et par locuteur mais conjointement sur l'ensemble des différents SNRs. Chacun des 51 mots du vocabulaire est modélisé par un HMM dont la taille est définie par la règle de 2 états par phonème. Ces $51x34$ (nbmots × nblocuteurs) modèles, sont d'abord initialisés par un HMM de même topologie (7 Gaussiennes diagonale, meilleur choix parmi 3-5-7 Gaussiennes diagonales) appris sur l'ensemble des locuteurs.

Les modèles sont appris en utilisant une réprésentation classique de la parole : 39 coefficients MFCC constitués de 12 coefficients cepstraux et d'un 1 terme d'énergie augmentés par leur dérivées temporelles de premier et de second ordre. Le signal de parole est au préalable réduit à 1 canal en prenant la moyenne des canaux gauche et droite. Les coefficients sont ensuite extraits par fenêtres de 25ms décalées par pas de 10ms.

5 Expériences

Nous avons d'abord recherché la topologie la plus performante entre des modèles à 4, 8, 16 ou 32 Gaussiennes, à matrices de covariance pleines ou diagonales, et en utilisant des modèles à 3 états par mot pour réduire le coût algorithmique. Par ailleurs, un HMM à 3 états par mot nous donne des performances similaires à la baseline du challenge (qui utilise 2 états par phonème). Pour des raisons de place, nous ne mentionnons dans la suite que les résultats obtenus avec ce modèle selectionné, à 8 Gaussiennes par état, pour des matrices de covariance pleines ou diagonales. A noter que les matrices de covariance pleines apportent un gain clair ici avec des caractéristiques de type MFCC (voir table 1).

SNR (dB)	Baseline	PHMM 8 Gaussiennes diag	PHMM 8 Gaussiennes full
-6	49.33	55.6	58.1
-3	58.67	64.8	67.6
0	67.5	76.2	76.2
3	75.08	81.3	82.6
6	78.83	85.5	86.4
9	82.92	87.6	89.1

TABLE 1: Comparaison des taux de reconnaissance lettre+chiffre de HMM standards et de PHMM exploitant des matrices de covariance pleines ou diagonales et exploitant une combinaison de variables contextuelles : Identité du locuteur, SNR, SFM, et Vitesse d'élocution. Tous les modèles ont la même topologie.

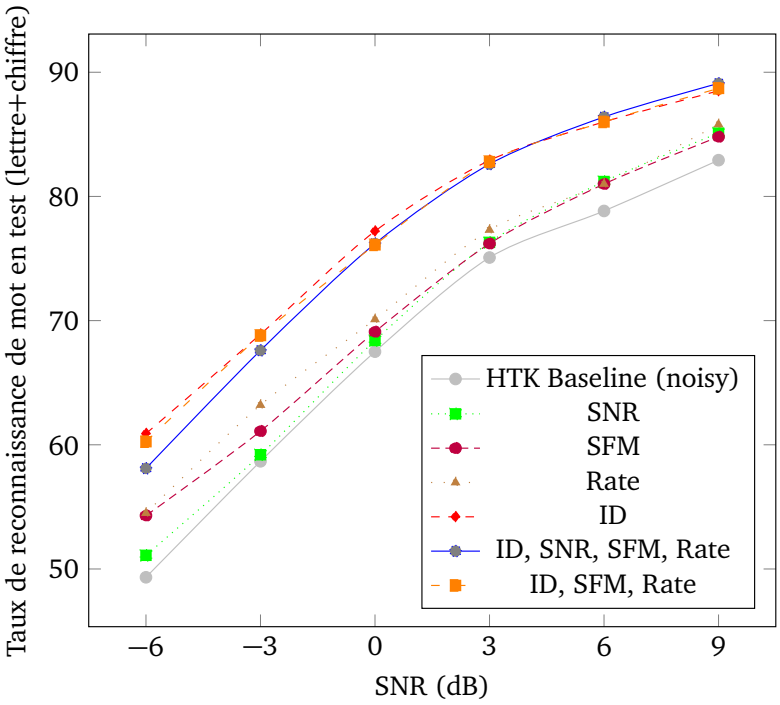


FIGURE 2: PHMM à 8 Gaussiennes pleines vs Baseline en utilisant différentes variables de paramétrisation (SNR, SFM, Rate, ID locuteur ou une combinaison)

Comme on le voit, les HMM paramétriques améliorent significativement la performance du système HMM de référence. Nous avons pu faire ce même constat quelles que soient les variables contextuelles employées (SNR, SFM, Rate, ID). La figure 2 compare les performances obtenues par des PHMM utilisant seules ou en les combinant, les quatres variables contextuelles envisagées. On voit qu'efffectivement l'usage de variables contextuelles améliore systématiquement les performances, mais il est également surprenant de voir que leur combinaison est parfois moins performante qu'utiliser uniquement l'information sur l'identité du locuteur, à l'exception de certaines situations à SNR élevés (6 et 9dB). En effet il parait naturel de penser que l'identité du locuteur et le SNR par exemple, étant deux informations totalement décorélées et chacune apportant isolément un gain significatif de performance, devraient pouvoir être combinées avec succès, ce qui n'est pas le cas ici. Il s'agit probablement d'un problème d'estimation des modèles, que nous cherchons à résoudre aujourd'hui.

Il faut noter ici que comparativement au système de référence qui apprend un HMM par locuteur et par mot (34×51 modèles), le PHMM n'estime que 51 modèles, 1 par mot, quel que soit le niveau de bruit, et quel que soit le locuteur. Le PHMM devient un modèle dépendant du locuteur uniquement lors de l'inférence, lorsqu'un HMM est instancié avec l'identifiant du locuteur comme variable de contexte. Ces modèles peuvent donc exploiter conjointement l'ensemble des données de tous les locuteurs pour l'estimation des paramètres contrairement aux HMM qui les sépare. Ce type d'estimation, qui paraît plus robuste, explique probablement la différence importante entre les performances des HMM et des PHMM particulièrement sur des données fortement bruitées.

La figure 3 compare les résultats obtenus avec les PHMM (en exploitant comme variables contextuelles la combinaison de l'identité du locuteur, du SFM, et de la vitesse d'élocution) avec les systèmes ayant participé à la compétition CHIME. Le système présenté est très simple et ne peut prétendre tel quel à battre les meilleurs systèmes de la compétition, très complexes et utilisant

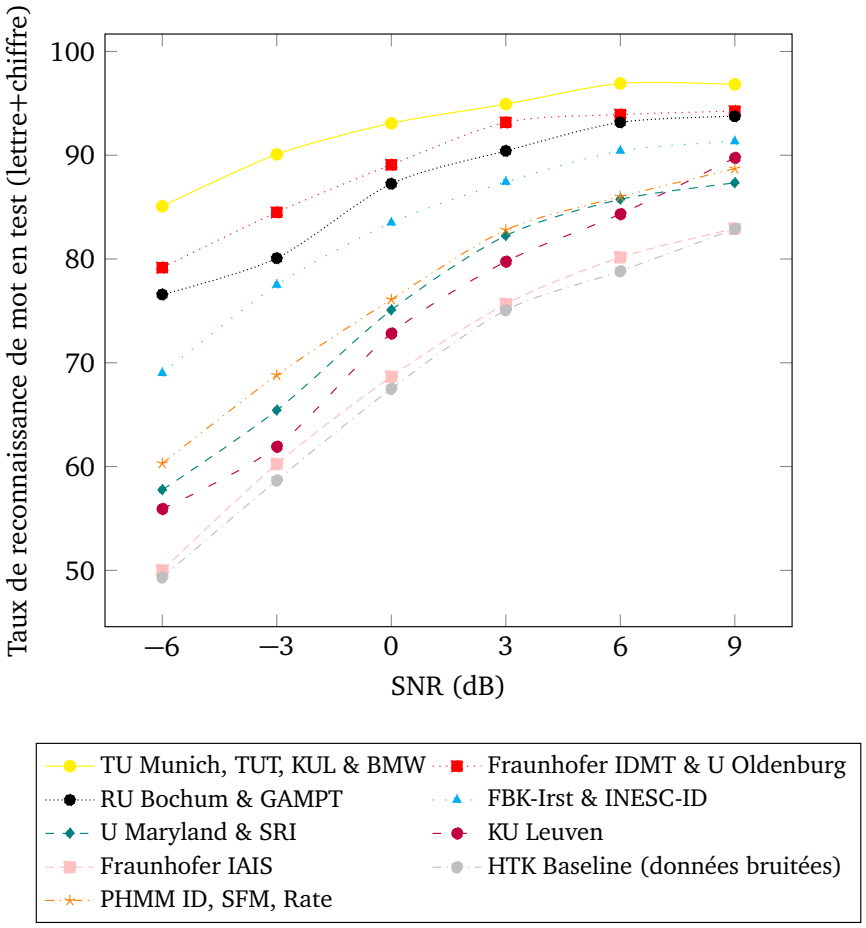


FIGURE 3: Résultats des PHMM parmi quelques résultats sélectionnés pour la représentativité de leur performance sur l'ensemble des soumissions au challenge CHiME (track 1). Y figurent la baseline et la performance du meilleur système.

des transformations de caractéristiques, des données d'apprentissage étendues ou encore une modélisation multi flux. Cela étant, les premiers résultats obtenus montrent que ce type de modélisation se comporte bien malgré sa simplicité et peut sans doute être largement amélioré en optimisant la prise en compte jointe de plusieurs variables contextuelles.

6 Conclusion

Nous avons proposé une première étude sur l'utilisation de modèles paramétriques pour la reconnaissance de parole bruitée. Nous avons exploré l'utilisation de plusieurs variables contextuelles et montré que les systèmes obtenus sont significativement plus performants que le système de référence basé sur des HMM gaussiens standards. Les expériences ont montré que chacune des variables proposées est utile à la reconnaisance mais que l'identité du locuteur est de loin la variable la plus explicative et que la combinaison de variables contextuelles, si elle peut apporter un gain, peut sans doute être amélioré. Ces résultats préliminaires, même si ils sont inférieurs aux meilleurs résultats de la compétition CHIME, sont encourageants car obtenus avec une modélisation très simple et offrent donc probablement une grande marge de progression.

Remerciements

Nous remercions Ayoub Ben Hassine, Ecole supérieure des communications de Tunis, pour la réalisation de premières expérimentations.

Références

ÁLVAREZ, L., LLERENA, C. et ALEXANDRE, E. (2011). Application of neural networks to speech/music/noise classification in digital hearing aids. *In Proceedings of the 11th WSEAS International Conference on Signal Processing, Computational Geometry and Artificial Vision, and Proceedings of the 11th WSEAS International Conference on Systems Theory and Scientific Computation*, GAVTASC'11, pages 97–102, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

BARKER, J., VINCENT, E., MA, N., CHRISTENSEN, H. et GREEN, P. (2013). The {PASCAL} {CHiME} speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621 – 633. Special Issue on Speech Separation and Recognition in Multisource Environments.

CHRISTENSEN, H., BARKER, J., MA, N. et GREEN, P. (2010). The chime corpus : a resource and a challenge for computational hearing in multisource environments. *In in Proc. Interspeech'10 Makuhari*.

CUI, X. et GONG, Y. (2007). A study of variable-parameter gaussian mixture hidden markov modeling for noisy speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1366–1376.

FUJINAGA, K., NAKAI, M., SHIMODAIRA, H. et SAGAYAMA, S. (2001). Multiple-regression (hmm). *In Acoustics, Speech and Signal Processing (ICASSP), 2001 IEEE International Conference on*.

GEIGER, J. T., WENINGER, F., HURMALAINEN, A., GEMMEKE, J., WÖLLMER, M., SCHULLER, B., RIGOLL, G. et VIRTANEN, T. (2013). The TUM+TUT+KUL Approach to the CHiME Challenge 2013 : Multi-Stream ASR Exploiting BLSTM Networks and Sparse NMF. *In proceedings of the 2nd CHiME workshop*, pages 25–30.

GONG, Y. (1995). Speech recognition in noisy environments : A survey. *Speech Communication*, 16(3):261–291.

MOHAMMADIHA, N., SMARAGDIS, P. et LEIJON, A. (2013). Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(10):2140–2151.

RADENEN, M. et ARTIERES, T. (2012). Contextual hidden markov models. *In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2113–2116.

RADENEN, M. et ARTIÈRES, T. (2014). Handling signal variability with contextual markovian models. *Pattern Recogn. Lett.*, 35:236–245.

VARIANI, E. et SCHAAF, T. (2011). Vtln in the mfcc domain : Band-limited versus local interpolation. *In INTERSPEECH*, pages 1273–1276.

VINCENT, E., BARKER, J., WATANABE, S., LE ROUX, J., NESTA, F. et MATASSONI, M. (2013). The second 'CHiME' Speech Separation and Recognition Challenge : Datasets, tasks and baselines. *In ICASSP - 38th International Conference on Acoustics, Speech, and Signal Processing - 2013*, pages 126–130, Vancouver, Canada.

WILSON, A. et BOBICK, A. (1999). Parametric hidden markov models for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(9):884 –900.

YOUNG, S. J., EVERMANN, G., GALES, M. J. F., HAIN, T., KERSHAW, D., MOORE, G., ODELL, J., OLLASON, D., POVEY, D., VALTCHEV, V. et WOODLAND, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

YU, D., DENG, L., GONG, Y. et ACERO, A. (2009). A novel framework and training algorithm for variable-parameter hmms. *IEEE Trans. on Audio, Speech, and Language Processing*, 17(7):1348 –1360.

Équations de locus et temporalité de la parole : Une étude de l'impact de la durée vocalique sur les paramètres des équations de locus en Arabe Jordanien.

Mohammad Abuoudeh Olivier Crouzet
Laboratoire de Linguistique de Nantes – LLING / EA3827
Université de Nantes, Chemin de la Censive du Tertre
44312 Nantes Cedex
mohammad.abuoudeh@univ-nantes.fr | olivier.crouzet@univ-nantes.fr

RÉSUMÉ

Les équations de locus sont des modèles de la relation Consonne-Voyelle qui procèdent à une abstraction de la dimension temporelle de la parole en se focalisant sur la relation de fréquence entre les deux composantes d'une séquence Consonne-Voyelle vues comme des segments. Nous présentons un travail dans lequel nous faisons l'hypothèse que ce point de vue est insuffisant. Afin d'évaluer les effets potentiels de la dimension temporelle de la parole sur les paramètres des équations de locus, nous avons procédé à l'enregistrement de séquences Consonne-Voyelle en contexte de phrase auprès de 4 locuteurs de l'Arabe Jordanien et avons évalué l'effet de la longueur vocalique sur les paramètres des équations de locus associées à 5 occlusives non-voisées. Les données recueillies conduisent à observer un effet significatif de la longueur vocalique sur les paramètres de pente et d'ordonnée à l'origine des équations de locus. Ces résultats nous conduisent à nous interroger sur l'aspect non-continu des mesures impliquées dans le estimations des équations de locus mais des analyses complémentaires sont suggérées afin d'évaluer les causes possibles des effets observés.

ABSTRACT

On locus equations and speech timing: An investigation of the impact of vowel length on locus equation parameters in Jordanian Arabic.

Locus equations are a model of Consonant-Vowel relationships based on an abstraction of the temporal dimension of speech. They focus on the relation between acoustic frequencies issued from two speech "segments". In this communiation, we argue that this hypothesis is erroneous and provide data from an experiment in which the importance of the time dimension on the computation of locus equations was investigated. We have recorded Consonant-Vowel sequences produced in a fixed setence context by 4 Jordanian Arabic speakers and compared locus equations parameters between long and short vowels for 5 unvoiced stops. According to our analyses, there is a significant impact of vowel length on locus equations, which provides arguments against the non-continuous bases of locus equations parameters in terms of place of articulation, though more thouroughfull analyses are needed in order to get a better understanding of these effects.

MOTS-CLÉS : Transitions formantiques, équations de locus, durées vocalique, arabe jordanien.

KEYWORDS: Formant transitions, locus equations, vowel duration, Jordanian Arabic.

1 Introduction

Les équations de locus sont des modèles linéaires exprimant, dans des séquences CV ou VC (avec C = Consonne et V = Voyelle), la relation de dépendance entre les configurations articulatoires réalisées dans la production d'une séquence consonne – voyelle coarticulée (Lindblöm, 1963). Ces équations sont de la forme :

$$F2_{onset} = k \times F2_{mid} + c \quad (1)$$

où les coefficients k et c (respectivement la pente et l'ordonnée à l'origine) sont issus d'une régression linéaire appliquée sur la relation entre la fréquence du second formant mesurée au milieu de la voyelle ($F2_{mid}$, conventionnellement placée en abcisse) et celle de ce même formant mesurée au moment du relâchement de la consonne ($F2_{onset}$, placée en ordonnée).

1.1 Équations de locus, coarticulation et catégories phonologiques

Afin d'obtenir une équation de locus (LE pour *Locus Equation*) on extrait des paires de fréquences $F2_{mid}$–$F2_{onset}$ sur de nombreuses réalisations individuelles impliquant une catégorie consonantique unique produite en contexte avec différentes voyelles. La plupart des travaux ont porté sur les occlusives (cf. par exemple Sussman *et al.*, 1991) même si certains se sont intéressés à des propriétés qui seraient généralisées à différents modes d'articulation (Sussman et Shore, 1996), voire aux propriétés qu'elles pourraient présenter en lien avec les voyelles (Al-Tamimi, 2007).

Deux approches parallèles cherchent à rendre compte de l'interprétation théorique des équations de locus : dans la lignée du travail séminal de Lindblöm (1963), les équations de locus permettent d'exprimer la « force » de la coarticulation (le degré de coarticulation) entre la consonne et la voyelle (Krull, 1989; Duez, 1992; Iskarous *et al.*, 2010). En théorie, s'il n'y a aucune coarticulation entre la consonne et la voyelle, la réalisation de la consonne est totalement indépendante de celle de la voyelle. Changer la réalisation de la voyelle qui est coarticulée avec la consonne ne devrait donc avoir aucun impact sur le locus consonantique (Delattre *et al.*, 1955), lequel devrait alors être stable quelle que soit la fréquence du formant vocalique. La pente exprimant la relation entre les deux valeurs devrait alors être égale à 0. À l'inverse, si les configurations du conduit vocal associées à la réalisation de la consonne et de la voyelle sont dans une interdépendance massive, on observerait une relation « parfaite » entre ces deux valeurs (p. ex. un changement de 100 Hz dans le formant de la voyelle correspondrait à un changement de 100 Hz dans le locus de la consonne) et la pente de l'équation de locus serait alors égale à 1. Les valeurs des pentes des équations de locus mesurées se distribuent donc en théorie entre ces deux valeurs (0 et 1) : plus la pente est élevée plus le degré de coarticulation est important.

Un autre ensemble de recherches s'est attaché à évaluer la relation entre les propriétés des équations de locus (pente mais aussi ordonnée à l'origine) et les catégories phonologiques (place d'articulation notamment). Sussman *et al.* (1991) montrent que les coefficients de l'équation de locus ne sont pas seulement un indicateur de degré de coarticulation mais aussi un bon descripteur du lieu d'articulation des occlusives.

Ainsi, sur la base de la valeur de pente de la droite de régression obtenue sur les 3 places d'articulation les plus étudiées, les vélaires montrent un degré de coarticulation relativement fort (pente élevée autour de 0.9), les alvéolaires ont quant à elles un degré de coarticulation relative-

ment faible (autour de 0.4) et celui des bilabiales est intermédiaire (environ 0.7). Ils formulent alors l'hypothèse que la relation linéaire entre $F2_{onset}$ et $F2_{mid}$ telle qu'elle est exprimée par les équations de locus représente, sinon un invariant, un indice pour décrire le lieu d'articulation des occlusives. Si les travaux associés à cette problématique se réfèrent également à la question du taux de coarticulation CV, ils orientent plus fortement le débat sur la « fonction » linguistique de ces équations. Alors que la relation entre classes sonores et équations de locus soulève des débats (voir notamment Sussman *et al.*, 1998; Iskarous *et al.*, 2010; Lindblom et Sussman, 2012), cette possibilité d'une catégorisation des données acoustiques associée aux places d'articulation suscite également des interrogations sur leur possible fonction perceptive (Branzacio et Fowler, 1998, par exemple). En effet, les séquences individuelles se distribuent dans des espaces acoustiques relativement distincts et des analyses discriminantes mettent en évidence l'impact des mesures associées sur la classification statistique des séquences (Sussman *et al.*, 1991) et si ces paramètres ne sont ni « invariants » ni suffisants (Branzacio et Fowler, 1998), ils contribueraient de manière non négligeable à la classification (Sussman *et al.*, 1998).

Selon Fowler (1994, voir aussi Iskarous *et al.*, 2010), les équations de locus ne fournissent pas d'informations permettant de spécifier le lieu d'articulation. Par contre, elles caractériseraient les propriétés de chevauchement coarticulatoire, qui peut varier non seulement avec la résistance coarticulatoire due au lieu ou mode d'articulation, mais aussi avec l'accent tonique. Selon Fowler (1994) donc, les équations de locus ne peuvent pas être utilisées comme des descripteurs du lieu d'articulation. À l'inverse, Sussman et Shore (1996) ont estimé des équations de locus dérivées de séquences CV/t/ dans lesquelles le mode d'articulation et le voisement de la consonne initiale variait (C ∈ {d, t, n, z, s} ; cette consonne étant associée à 10 voyelles différentes) chez 22 locuteurs américains et observent que les pentes des équations de locus mesurées pour des séquences contenant des consonnes qui ont le même lieu d'articulation (alvéolaire uniquement) et différents modes d'articulation (voisement, nasalité, et friction) seraient similaires.

1.2 Équations de locus et temporalité de la parole

Il apparaît que la représentation qui est fournie par les équations de locus fait abstraction d'un paramètre central de la parole : la temporalité. En effet, les valeurs qui servent à construire le nuage de points servant de support à la régression linéaire (voir l'illustration Fig. 1) sont des mesures de fréquence prises à deux instants dans la réalisation : le moment initial correspondant à la consonne (bruit de friction résultant du relâchement ou premiers cycles de vibration laryngale faisant suite à l'aspiration consonantique) et le milieu (ou zone de stabilité maximale) de la voyelle. Le fait que ces deux « événements » puissent se répartir différemment dans le temps, et son impact possible sur les trajectoires formantiques, ne semble donc pas crucial dans l'expression des équations de locus. Récemment, Rhone et Jongman (2012) ont calculé les équations de locus avec un $F2_{mid}$ extrait de trois positions temporelles différentes (entre $F2_{onset}$ et $F2_{mid}$) du F2 et observent que la position temporelle du $F2_{mid}$ n'a pas d'impact sur les équation de locus. Cela pourrait mettre en évidence une relative indépendance entre les paramètres des équations de locus et les propriétés temporelles des signaux de parole.

Ce mode de calcul impose aux trajectoires formantiques (O'Shaughnessy, 1986; Al-Tamimi, 2007) une forme de « discrétisation » ignorant les propriétés de continuité du signal de parole. Pourtant, Krull (1989), dans son travail sur la comparaison des degrés de coarticulation entre parole spontanée et contrôlée observe un effet lié au style de parole. En comparant, pour 5 locuteurs

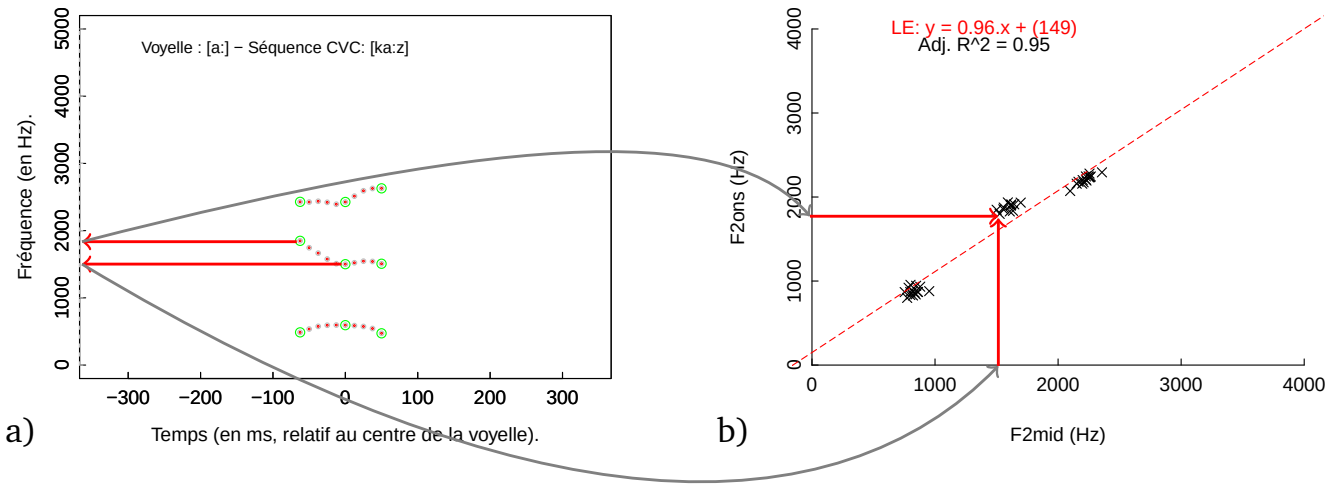


FIGURE 1 – Illustration de la méthode de construction des données pour le calcul des équations de locus. À gauche –Fig. 1a–, un spectrogramme schématique dérivé automatiquement par une analyse LPC permet d'identifier les points correspondant à la mesure de $F2_{mid}$ et $F2_{onset}$ dans une séquence CV. À droite –Fig. 1b–, les deux valeurs de fréquence prélevées permettent de tracer un point unique dans un espace $F2_{mid}$ / $F2_{onset}$.

suédois, la production de mots isolés et en conversation spontanée pour les consonnes {b, d, m, n, l}, elle montre que les pentes des équations de locus sont plus élevées pour la parole spontanée que les mots en isolation. Selon elle, cet effet pourrait être dû aux différences de débit ainsi qu'à des phénomènes de réduction vocalique. Parallèlement, Duez (1992) a répliqué le travail de Krull (1989) sur des locuteurs natifs français et ses résultats sont en accord avec les observations de Krull (1989). Pourtant, Sussman *et al.* (1998) semble considérer que les effets mis en évidence par Krull (1989) sont relativement limités :

> « mean slope difference between speaking styles across all consonants was only .06. Most important, speaking style variation did not perturb locus equation slopes in their role as phonetic descriptors of consonant place. » (Sussman *et al.*, 1998, p.248) [1].

Or, s'il apparaît effectivement que les valeurs moyennes observées diffèrent peu en termes de taille observée de l'effet et que les valeurs moyennes des pentes restent ordonnées relativement les unes aux autres entre labiales et dentales sans changement majeur, il n'en reste pas moins que ce que met en évidence le travail de Krull (1989), c'est bien l'impact statistiquement significatif que peuvent avoir des influences stylistiques sur les paramètres des équations de locus. Par ailleurs, le travail de Krull (1989), s'il illustre bien un maintien de la différence entre labiales et dentales quel que soit le style de parole du point de vue de la pente des équations, n'apporte pas d'évidence concernant l'impact potentiel de cette variation en termes perceptifs ; pour cela il conviendrait d'analyser la contribution de ces paramètres acoustiques à leur classification.

Krull (1989) fait référence à deux phénomènes qui pourraient être à l'origine de ces variations de pente : l'un est un phénomène d'origine principalement spectrale concentré sur la voyelle, la réduction vocalique plus marquée en parole spontanée peut conduire à des changements de fréquence des formants plus importants au niveau de la cible vocalique ; l'autre est un phénomène impactant la relation entre les segments : les changements de pente pourraient s'expliquer par

1. « la différence moyenne de pente entre styles de parole sur l'ensemble des consonnes était seulement de .06. Plus important, les variations de style de parole ne perturbaient pas les pentes des équations de locus dans leur rôle de descripteurs phonétiques de la place [d'articulation] des consonnes. »

des variations dans la relation temporelle qui s'établit entre la consonne et la voyelle. Cet impact du style de parole pourrait par exemple influencer le degré de coarticulation et contribuer alors à modifier simultanément les propriétés spectrales de la consonne et de la voyelle.

L'objet du présent travail est de s'intéresser à l'impact de modifications temporelles de la relation consonne-voyelle sur les équations de locus en étudiant l'influence d'une opposition de durée vocalique sur les propriétés acoustiques des consonnes occlusives en Arabe Jordanien.

2 Méthode

2.1 Participants

Quatre locuteurs jordaniens masculins (étudiants à l'Université de Nantes au moment de la passation) ont participé à cette expérience. L'un des 4 locuteurs est le premier auteur. Ils sont originaires de trois régions différentes de Jordanie (Ma'an, Amman et Irbid), ne présentent aucun trouble du langage et sont âgés de 24 à 38 ans.

2.2 Stimuli

Les locuteurs devaient lire, à un rythme soutenu mais confortable, des séquences CVC enchâssées dans une phrase porteuse (/ħaka CVC marteːn/ : « Il a dit CVC deux fois »). Les séquences apparaissaient l'une après l'autre en alphabet arabe vocalisé (les diacritiques représentant les voyelles étaient affichés) sur un écran d'ordinateur.

Nous nous sommes intéressés aux 5 occlusives non-voisées de l'arabe /t, tˤ, k, q, ʔ/ (alvéolaire, alvéolaire pharyngalisée, vélaire, uvulaire, glottale). Chacune de ces occlusives apparaissait dans un mot de la langue ayant une structure CVC. Ces 5 occlusives cibles ont été combinées avec 8 voyelles (3 courtes –{i, a, u}– et 5 longues –{iː, aː, uː, eː, oː}/) dans deux positions lexicales (initiale $C_{cible}VC$ vs. finale CVC_{cible}). L'autre consonne du mot était sélectionnée au hasard en fonction des mots existant dans le vocabulaire de l'arabe. Ce qui fait au total 80 mots différents, tous en usage en Arabe Jordanien. La fréquence des mots n'a pas été contrôlée.

2.3 Procédure

Nous avons demandé aux locuteurs de lire les séquences « naturellement » sans chercher à « parler correctement ». Un programme Python contrôlait le décours de l'expérience (durée d'affichage, ordre de présentation des séquences, interruption et reprise lors des pauses). Les séquences (mot en contexte avec la phrase porteuse) étaient présentées sur l'écran d'ordinateur dans un ordre aléatoire et à un rythme relativement soutenu mais confortable (une phrase toutes les 2500 ms). Les participants ne connaissaient pas les séquences avant le début de l'enregistrement. Chaque mot a été répété 15 fois. Nous avons donc enregistré 1200 séquences par locuteur (5 consonnes × 2 positions syllabiques × 8 voyelles × 15 répétitions).

Les tracés de fréquence des 3 premiers formants de chaque séquence CV cible et leurs coordonnées temporelles ont été extraites avec le logiciel *Praat* (analyse LPC). Ces données ont ensuite été

traitées par un script R chargé de procéder à l'analyse des données. Les tracés formantiques ont été lissés par un moyennage glissant sur toutes les suites de 2 valeurs successives afin de compenser des variations trop rapides des fréquences formantiques (filtrage passe-bas). Les valeurs initiale ou finale (en fonction de la position de l'occlusive dans le mot ; F_2onset) et médiane (F_2mid) de chaque tracé formantique ont ensuite été sélectionnées et associées aux variables contrôlées dans l'expérience (occlusive cible, longueur vocalique, voyelle, position de l'occlusive dans le mot, locuteur). Afin de ne pas déséquilibrer l'analyse en raison de la présence de voyelles longues dont le timbre n'est pas présent dans l'inventaire des voyelles courtes de la langue (/eː/ et /oː/), les données obtenues pour ces voyelles n'ont pas été retenues pour l'analyse présentée ici, qui porte donc sur un sous-ensemble de 900 séquences CVC par locuteur.

Des régressions linéaires ont ensuite été produites à partir de ces points afin d'estimer les paramètres de chacune des 80 équations de locus (5 consonnes × 2 longueurs vocaliques × 2 positions syllabiques × 4 locuteurs). Afin d'éviter l'impact de valeurs extrêmes sur les pentes des équations de locus, toute mesure dont le résidu (l'erreur) entre la valeur observée et la valeur prédite était supérieure à la moyenne des valeurs absolues des résidus plus 1.5 fois l'écart-type de ces résidus était identifiée comme un *outlier* et retirée de l'analyse. Suite à cette première phase de l'analyse, une seconde phase de régressions linéaires était produite sans ces valeurs. Ces données extrêmes représentent 7% des données.

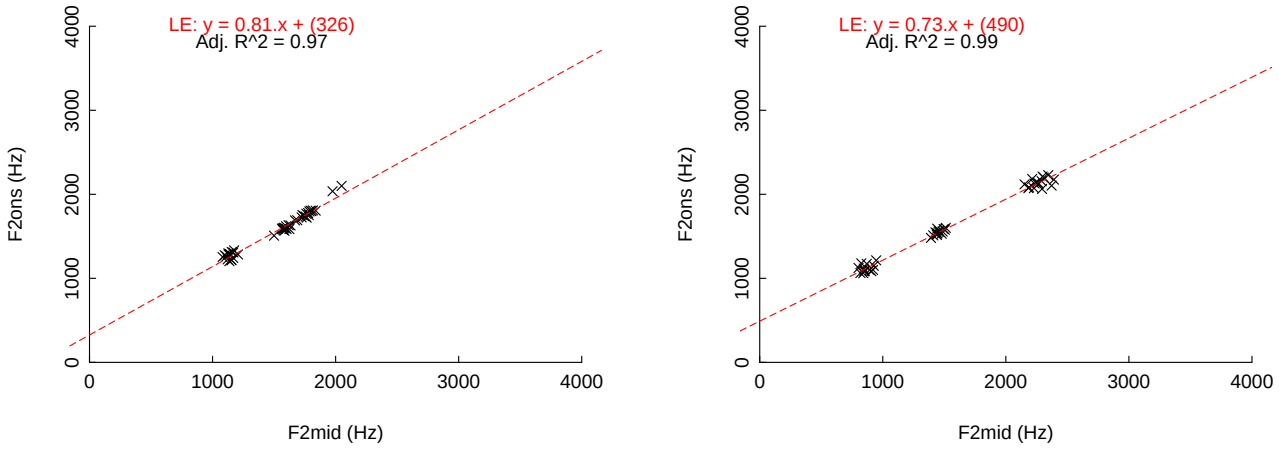
	Pente		Ordonnée à l'origine		R^2	
C	V courtes	V longues	V courtes	V longues	V courtes	V longues
t	0.78	0.67	375	549	0.94	0.95
tˤ	0.77	0.54	198	384	0.95	0.86
k	0.96	0.94	-33	3	0.95	0.94
q	0.84	0.78	43	158	0.93	0.92
ʔ	0.93	0.83	-62	119	0.98	0.94

TABLE 1 – Valeurs moyennes des pentes, des ordonnées à l'origine et des R^2 des équations de locus calculées pour chaque consonne coarticulée avec les voyelles courtes et les voyelles longues.

2.4 Résultats

2.4.1 Analyse descriptive

Les valeurs moyennes de pente et d'ordonnée à l'origine des régressions linéaires (équations de locus) associées aux 5 places d'articulation et à l'opposition de longueur vocalique sont présentées Table 1 (cf. Fig. 2 pour une illustration). Les pentes des voyelles longues sont abaissées par rapport à celles des voyelles courtes pour toutes les consonnes. Néanmoins, ce phénomène semble surtout marqué pour les occlusives antérieures (alvéolaire et alvéolaire pharyngalisée) et beaucoup moins pour les occlusives postérieures. On peut aussi remarquer que les ordonnées à l'origine estimées semblent être plus élevées en contexte de voyelles longues qu'en contexte de voyelles courtes. Contrairement aux mesures de pente, l'effet de longueur vocalique semble plus généralisé sur les ordonnées à l'origine mais est moins marqué pour la vélaire. Les valeurs de R^2 représentent le degré d'adéquation entre les données observées et le modèle linéaire estimé ; elles sont relativement élevées pour toutes les conditions étudiées.



(a) (b)

FIGURE 2 – Illustration de deux ensembles de données servant de support à l'estimation des modèles linéaires. Ici, on obtient deux équations de locus issues des mesures réalisées sur l'occlusive alvéolaire /t/ en contexte de voyelle courte (gauche) vs. voyelle longue (droite) – position initiale de mot –CV–, locuteur L2.

2.4.2 Analyse inférentielle

Deux ANOVA de Type I à mesures répétées (avec trois facteurs intra-locuteur –Longueur, Consonne, Position–, et les locuteurs comme facteur aléatoire) ont été conduites pour comparer les effets de la longueur vocalique respectivement sur les pentes et ordonnées à l'origine des équation de locus obtenues[2]. La plupart des interactions sont non significatives à l'exception de l'interaction Consonne × Position (resp. $F_{(4,12)} = 3.37, p < .05$ pour les pentes, $F_{(4,12)} = 8.55, p < .01$ pour les ordonnées à l'origine). Par contre aucune des interactions impliquant la longueur vocalique n'est significative.

En ce qui concerne les effets principaux, les deux analyses de variance confirment les observations issues des tableaux de moyennes (cf. Table 1). Les pentes des voyelles longues diffèrent significativement des pentes des voyelles courtes à travers toutes les consonnes ($F_{(1,3)} = 53.35, p < .01$) ; de-même l'effet de longueur sur les ordonnées à l'origine est significatif ($F_{(1,3)} = 19.27, p < .05$). On observe également, conformément aux travaux classiques sur les équations de locus un effet principal de la place d'articulation des occlusives sur les pentes et ordonnées à l'origine (resp. $F_{(4,12)} = 23.64, p < .001; F_{(4,12)} = 19.64, p < .001$).

2.4.3 Analyse acoustique des voyelles

Les analyses de données effectuées dans le cadre de ce travail semblent confirmer l'hypothèse d'un effet systématique de la longueur vocalique sur les propriétés des équations de locus. Afin d'affiner nos interprétations, nous avons en premier lieu cherché à savoir si, effectivement, l'opposition de longueur vocalique se manifestait bien par une différenciation temporelle puisque notre hypothèse portait avant tout sur l'impact que des variations de durée auraient sur les équations de locus. Nous avons donc procédé à l'analyse de la relation entre opposition de

2. Le plan d'analyse est équilibré. Ces ANOVAs ont été effectuées sur des données de pente et d'ordonnée à l'origine transformées numériquement (carré des pentes, racine carrée des ordonnées à l'origine transformées en valeurs positives) afin de s'assurer qu'elles respectaient bien les conditions de validité de l'ANOVA (distribution normale, homogénéïté des variances). Les statistiques descriptives portent sur les valeurs réelles.

longueur et propriétés acoustiques des voyelles. Comme on pouvait s'y attendre, les voyelles longues ont bien des durées plus longues (129 ms, SE=0.85) que les voyelles courtes (68 ms, SE=1.29). Dans une ANOVA à un seul facteur à mesures répétées avec les locuteurs comme facteur aléatoire, cet effet est significatif ($F_{(1,3)} = 145.6, p < .01$).

Cependant, on observe aussi des influences notables sur les propriétés spectrales des voyelles (Table 2). La même analyse réalisée sur les fréquences des formants aboutit à des différences significatives pour les 3 premiers formants. Nous nous concentrons ici sur les résultats observés concernant le second formant puisque c'est sur lui que nos analyses d'équations de locus sont concentrées. Le F_2 des /i/ est élevé en voyelle longue alors qu'il s'abaisse pour /a/ et /u/. Cette interaction est significative ($F_{(2,6)} = 65.36, p < .001$).

	i	a	u
long	2157 (20)	1353 (27)	963 (26)
court	1711 (34)	1388 (28)	1090 (25)

TABLE 2 – Moyenne (et erreur standard moyenne) des valeurs des fréquences du F2 en Hz mesurées au milieu du formant pour les voyelles longues comparées aux voyelles courtes.

3 Discussion

Sans aller directement à l'encontre du modèle conventionnel des équations de locus comme descripteurs de la place d'articulation (Sussman et Shore, 1996), les données présentées ici fournissent des éléments de critique contribuant à reconsidérer leur aspect a-temporel. Il conviendra d'approfondir les analyses afin de mieux comprendre le rôle que peut avoir joué une variation de timbre spectral dans la réalisation des voyelles de longueur distincte sur l'effet observé dans cette étude. Ainsi, il est possible que l'effet observé ne corresponde à aucune modification inhérente à la consonne ou à la relation entre consonne et voyelle mais n'émerge que sous l'influence des changements vocaliques. Les données présentées ici semblent compatibles avec le travail récent de Berry et Weismer (2013) qui montre que les équations de locus sont sensibles à des modifications de débit de parole. Par ailleurs, des analyses complémentaires sont en cours afin de procéder à des vérifications des critères de segmentation utilisés pour l'extraction des valeurs de fréquence du second formant, notamment en ce qui concerne la localisation du moment correspondant au $F2_{onset}$ qui pose plus de difficultés avec des occlusives non-voisées qu'avec des voisées. Enfin, si les résultats présentés ici fournissent des arguments en faveur d'une prise en compte des variations temporelles dans les études sur les équations de locus, ils ne disent rien sur l'impact que peut avoir ce phénomène en termes de perception de la place d'articulation. Une étude fondée sur les techniques d'analyse discriminante linéaire permettrait d'évaluer l'impact de cet effet sur la classification perceptive basée sur les propriétés constituant les équations de locus.

Remerciements

Ce travail a été rendu possible grâce au soutien financier de l'Université Al-Hussein Bin Talal (Jordanie). Nous tenons à remercier deux relecteurs anonymes pour leurs commentaires éclairés et fructueux.

Références

Al-Tamimi, J.-E. (2007). *Indices dynamiques et perception des voyelles : Étude translinguistique en arabe dialectal et en français*. Phd thesis, Université Louis Lumière - Lyon 2.

Berry, J. et Weismer, G. (2013). Speaking rate effects on locus equation slope. *Journal of Phonetics*, 41:468–478.

Branzacio, L. et Fowler, C. A. (1998). On the relevance of locus equations for production and perception of stop consonants. *Perception & Psychophysics*, 60(1):24–50.

Delattre, P., Liberman, A. et Cooper, F. (1955). Acoustical loci and transitional cues for consonants. *The Journal of the Acoustic Society of America*, 27(4):769–773.

Duez, D. (1992). Second formant locus-nucleus patterns : An investigation of spontaneous french speech. *Speech Communication*, 11(4-5):471–427.

Fowler, C. (1994). Invariants, specifiers, cues : An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, 55:597–610.

Iskarous, K., Fowler, C. a. et Whalen, D. H. (2010). Locus equations are an acoustic expression of articulator synergy. *The Journal of the Acoustical Society of America*, 128(4):2021–32.

Krull, D. (1989). Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *Phonetic experimental research at the institute of linguistics university of Stockholm-PERILUS*, 10:87–108.

Lindblom, B. et Sussman, H. M. (2012). Dissecting coarticulation : How locus equations happen. *Journal of Phonetics*, 40:1–19.

Lindblöm, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35(11):1773–1781.

O'Shaughnessy, D. (1986). The effects of speaking rate on formant transitions in French synthesis-by-rule. *In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, volume 11, pages 2027–2030.

Rhone, A. et Jongman, A. (2012). Modified locus equations categorize stop place in a perceptually realistic time frame. *The Journal of Acoustical Society of America*, 131(6):EL487–EL491.

Sussman, H. M., Fruchter, D., Hilbert, J. et Sirosh, J. (1998). Linear correlates in the speech signal : The orderly output constraint. *Behavioral and Brain Siences*, 21:241–299.

Sussman, H. M., McCaffrey, H. A. et Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop consonant place categorization. *Journal of the Acoustical Society of America*, 90:1309–1325.

Sussman, H. M. et Shore, J. (1996). Locus equation as phonetic descriptors of consonantal place of articulation. *Psychonomic Society, Inc.*, 58(6):936–946.

Lien entre acquisition des premiers mots chez des enfants francophones monolingues et leur environnement langagier

Ludivine Glas[1, 2] Sophie Kern[1]
(1) Dynamique du langage – UMR 5596 / CNRS – Université Lyon 2
(2) Labex ASLAN
Ludivine.glas@wanadoo.fr, Sophie.Kern@cnrs.fr

RESUME

Il existe une grande variabilité dans l'acquisition du vocabulaire chez le jeune enfant en termes quantitatif et qualitatif. Cette variabilité est expliquée par des caractéristiques individuelles, sociales et/ou linguistiques.
L'objectif de ce travail est d'observer l'influence de la quantité et la qualité d'exposition et de certaines caractéristiques interactionnelles des parents et enfants sur le développement précoce du vocabulaire chez l'enfant. Pour ce faire, nous avons analysé la production spontanée de 4 enfants français monolingues aux stades de 1-5 et 50 mots au regard de leur propre comportement linguistique et interactionnel ainsi qu'au regard de ceux de leurs parents. Les premiers résultats montrent une corrélation positive entre le comportement interactionnel des parents et le comportement linguistique des enfants, ainsi qu'une corrélation positive entre le comportement interactionnel des enfants et leur comportement linguistique futur. D'autres enfants et d'autres stades linguistiques seront analysés afin de vérifier la validité des résultats et leur stabilité au cours du développement.

ABSTRACT

Quantitative as well qualitative high level of variability exists in young children vocabulary development in quantitative and qualitative term. This variability can be explained by individual, social and linguistics features.
The goal of this study is to show the influence of quantity and quality of input as well as parent and children interactional characteristics on early lexical development in children. We have observed the spontaneous production of 4 French monolingual children at the linguistic stages of 1-5 and 50 different produced words, taking parental and children linguistic and interactional behaviors into consideration. The first results show a positive correlation between children linguistic behavior and parental interactional behavior as well as between children's interactional behavior and their forthcoming production.
To check the validity and stability of the results during development, we will extend our work to a more important number of children and to more linguistic stages.

MOTS-CLES : acquisition du lexique, environnement langagier, fréquence d'exposition.
KEYWORDS: lexicon development, input, frequency.

1 Rôle de l'exposition au langage dans l'acquisition

La production du lexique des jeunes enfants est quantitativement et qualitativement différente de celle que l'on trouve chez les adultes (Fenson et al. 1993). Au début de l'acquisition, le lexique est quantitativement limité et les premières productions de l'enfant peuvent être phonologiquement et sémantiquement très éloignées des formes adultes qui sont alors considérées comme des cibles à atteindre au cours du développement. Par ailleurs, bien qu'il existe un certain consensus sur l'existence de trajectoires développementales composées d'étapes précises par lesquelles passent tous les enfants, on observe également d'importantes variations interindividuelles entre des enfants du même âge. Pour citer quelques exemples : les premiers mots peuvent apparaitre vers 10 mois comme vers 24 mois (Bates 1997) et les vitesses d'acquisition, c'est-à-dire la rapidité de passage d'une étape à l'autre peuvent être différentes selon les enfants).

Ces variations peuvent s'expliquer par des caractéristiques propres à l'enfant comme son sexe ou son rang dans la fratrie (Fenson et al. 1994). Mais elles peuvent aussi être expliquées par des expositions différentes au langage que reçoit le jeune enfant. En effet, il existe un lien entre la qualité du discours adressé à l'enfant et ses acquisitions (Huttenlocher et al. 1991). Il en est de même si on prend en compte la fréquence d'exposition (Huttenlocher et al. 1991). Cependant, cette exposition à la langue qui joue un rôle dans l'acquisition du vocabulaire, est elle-même influencée par des facteurs sociaux comme la catégorie socioprofessionnelle des parents et le sexe des interlocuteurs. Ainsi, les mères qui ont fait des études supérieures ont un vocabulaire plus étendu, une syntaxe plus complexe et sollicitent plus leur enfant que les mères qui n'ont pas poursuivi d'études supérieures (Hoff et Naigles 2002).

Concernant les interactions, de nombreuses études cherchent des corrélations entre les variables du discours maternel et le développement du langage de l'enfant. Ainsi, les comportements communicatifs tels que les séquences de répétitions, de clarifications et les reformulations lors des échanges parents/enfants peuvent expliquer également certaines acquisitions. Yoder et Kaiser 1989 montrent notamment que les demandes de confirmation sont positivement corrélées au développement rapide du langage chez l'enfant, lien qui n'est pas trouvé pour les demandes de clarification. Dans ce domaine également, des différences dans le style conversationnel maternel sont liées à des différences sociales (Hoff-Ginsberg 1991) et culturelles (Tamis-LeMonda, Baumwell, et Cristofaro 2012).

Cette exposition à la langue est donc elle-même sujette à fluctuation. Il alors donc intéressant de voir quel lien il existe entre le discours du parent et les acquisitions de l'enfant, en termes de fréquence d'exposition, mais aussi en fonction des différents comportements interactionnels au sein de la dyade parent/enfant.

2 Méthodologie

2.1 Méthode d'enregistrement et sélection des sessions d'enregistrements

Nous avons suivi dans le cadre du projet PREMS (ANR blanche 2011, porteur S. Kern) 10 dyades parent/enfant depuis la production des premiers mots des enfants jusqu'à quelques mois après l'accélération du rythme d'acquisition du vocabulaire productif. Nous avons effectué des enregistrements audio-vidéo de leurs interactions en contexte spontané et naturel pendant 1 heure

toutes les 2 semaines. En parallèle, les Inventaires Français du Développement Communicatif (Kern et Gayraud 2010) adaptations françaises des MacArthur-Bates Communicative Development Inventories (Fenson et al. 1993), sont documentés tous les mois par les parents.

Nous avons choisi d'analyser 4 dyades. Elles sont monolingues du français, 2 d'entre elles sont unisexes (mère/fille et père/fils) et 2 autres sont hétérogènes au niveau du sexe. Toutes les familles sont issues du même niveau socio-professionnel (niveau d'études supérieures) et tous les enfants sont gardés par leur parent. Nous avons restreint nos analyses à deux stades linguistiques : le stade de 1-5 mots différents produits (stade 1-5) et celui de 50 mots (stade 50), l'objectif étant de comparer les comportements linguistiques et interactionnels des enfants et des parents au moment des premiers mots avec leurs comportements au début de l'explosion lexicale. Les stades des enfants ont été déterminés par le biais des résultats obtenus aux comptes rendus parentaux. Parmi l'ensemble des données collectées, nous avons travaillé sur une session d'enregistrement égalisée à 45 minutes pour toutes les dyades. L'âge des enfants varie de 12;5 à 13;3 mois (âge moyen 12;8). À l'exception d'une session où les deux parents étaient présents, toutes les sessions comprennent l'enfant et le parent qui s'occupe le plus de l'enfant avec une prédominance dans toutes les sessions des activités de jeux (puzzle, dessin, nommer les objets), de routines (changer la couche, soulager les caprices), et des explorations (déplacement, découverte des objets).

Dyade	Sexe parent	Sexe enfant	Âge stade 1-5	Âge stade 50	Nombre de minutes analysées par session	Total analysé (en min)
1	féminin	féminin	12;8	19;23	45	90
2	féminin	masculin	12;27	18;12	45	90
3	féminin	masculin	13;3	19;12	45	90
4	masculin	masculin	13;14	16;27	45	90

TABLEAU 1 : échantillon analysé

2.2 Transcription et analyse des vidéos

Pour notre analyse, seul le langage de l'enfant considéré par l'étude (exclusion des frères et sœurs par exemple) et du parent le plus présent sont pris en compte et donc transcrits, codés et analysés.

La transcription est faite en alphabet phonétique international et sous forme orthographique grâce au logiciel Phon 2.6 beta© conçu par Yvan Rose (Rose et al. 2006). Les productions ont été segmentées en énoncés. Pour ce faire, nous avons adopté les critères explicités dans Parisse et Le Normand 2007. Un deuxième niveau de segmentation considéré est le mot. Nous avons considéré comme un mot produit, une forme linguistique qui est produite spontanément (hors imitation) dans un contexte approprié et qui se rapproche ou est identique à la cible adulte (Vihman et McCune 1994).

Pour chacun des participants nous avons mesuré des comportements linguistiques et interactionnels. Font partie des comportements linguistiques (cf colonne 1 du tableau 1) : le nombre total de mots (occurrences), le nombre de mots différents (types), et l'indice D de diversité lexicale. Pour évaluer la diversité lexicale, nous avons choisi de calculer l'indice de diversité lexicale selon la méthode de McKee, Malvern, et Richards (2000) qui correspond au

ratio des occurrences sur les types de mots produits pour un nombre déterminés d'énoncés. Cette méthode permet de lisser les différences entre les enfants en termes de productivité (plus ou moins d'énoncés dans le même temps). Plus l'indice D est grand plus la diversité est grande.

Pour les comportements interactionnels (cf colonne 2 du tableau 1), au vu des résultats de la littérature (Yoder et Kaiser 1989; Bernicot et Clark 2010; Bloom, Rocissano, et Hood 1976) nous nous sommes focalisées sur les répétitions et les clarifications ainsi que sur le nombre total d'énoncés. Au sein des répétitions et des clarifications nous avons considéré différemment les auto-répétitions / clarifications des hétéro-répétitions / clarifications ainsi que la taille des éléments répétés (Énoncés vs. Mots).

Comportements linguistiques		Comportements interactionnels	
Occurrence :	nombre total de mot	Nombre d'énoncés hétéro-répétés :	répétition de l'énoncé d'un autre locuteur
Type :	nombre de mots différents	Nombre de mots hétéro-répétés :	répétition d'un mot d'un autre locuteur
Indice D de diversité lexicale :	Type de mots / occurrence de mots pour un nombre d'énoncés déterminés	Nombre d'énoncés auto-répétés :	le locuteur répète son énoncé
		Nombre d'énoncés hétéro-clarifiés :	clarification et/ou ajout d'information de l'énoncé d'un autre locuteur
		Nombre d'énoncés auto-clarifiés :	le locuteur clarifie et/ou ajoute une/des information(s) à son énoncé précédent
		Nombre d'énoncés total :	nombre de prises de parole lors de la session

TABLEAU 2 –variables analysées

Nous avons observé les comportements linguistiques des enfants et des parents pour les deux stades linguistiques. Ensuite, nous avons cherché à établir des corrélations entre les comportements linguistiques des enfants et leurs propres comportements interactionnels mais également entre leurs comportements linguistiques et les comportements linguistiques et interactionnels des parents. Enfin, pour mesurer un lien à plus long terme les comportements linguistiques et interactionnels des parents ainsi que les comportements interactionnels des enfants au stade linguistique de 1-5 mots ont été corrélés aux comportements linguistiques des enfants à 50 mots. L'analyse des valeurs de corrélation propres à chaque cellule de la matrice de Spearman a été effectuée ainsi que des tests de corrélation de Spearman sur les variables les plus cohérentes à la suite de l'analyse de la matrice. Toutes les statistiques sont calculées sous le logiciel R.

2.3 Résultats et discussion

2.3.1 Comportements linguistiques des dyades parent / enfant

Concernant le stade 1-5 mots, on observe de grandes différences entre les enfants dans le nombre de mots produits spontanément qui va de 7 à 27 mots avec une moyenne de 18 mots (écart-type de 8,27) (figure 1). La diversité lexicale n'a pas été calculée compte tenu de la taille restreinte du vocabulaire et du nombre encore plus restreint de types (3, 1, 2 et 2).

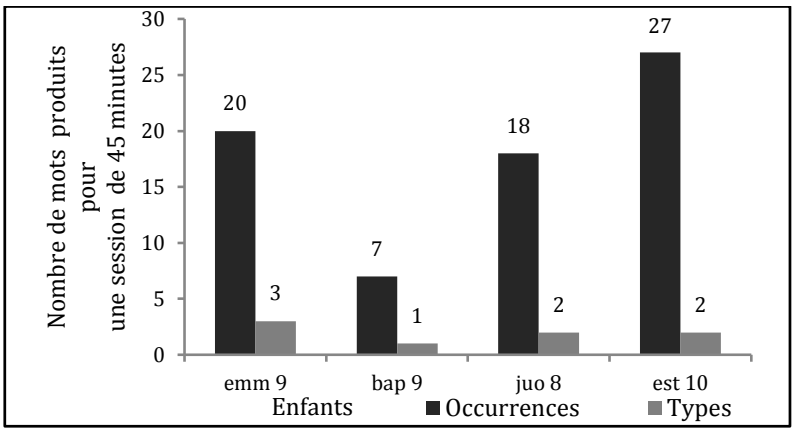


FIGURE 1 : Comportements linguistiques de 4 enfants français au stade 1-5 mots

Chez les parents, on observe de grandes différences interindividuelles en termes d'occurrences de mots produits, allant de 367 à 2251 avec une moyenne de 1495 mots et un écart-type de 824,44, et de types de mots produits allant de 93 à 335 avec une moyenne de 236,25 mots et un écart-type de 106,34. Toutefois 3 parents sur 4 ont plus de 1000 occurrences de mots et plus de 200 types (figure 2a). On observe des résultats similaires dans la diversité lexicale (moyenne de 84,9 ; écart-type de 23,74) avec 3 parents qui présentent une diversité lexicale de plus de 90. Le 4[ème] parent, qui a également le nombre d'occurrences et le nombre de types le plus faible, a une diversité lexicale deux fois moins importantes (49,47).

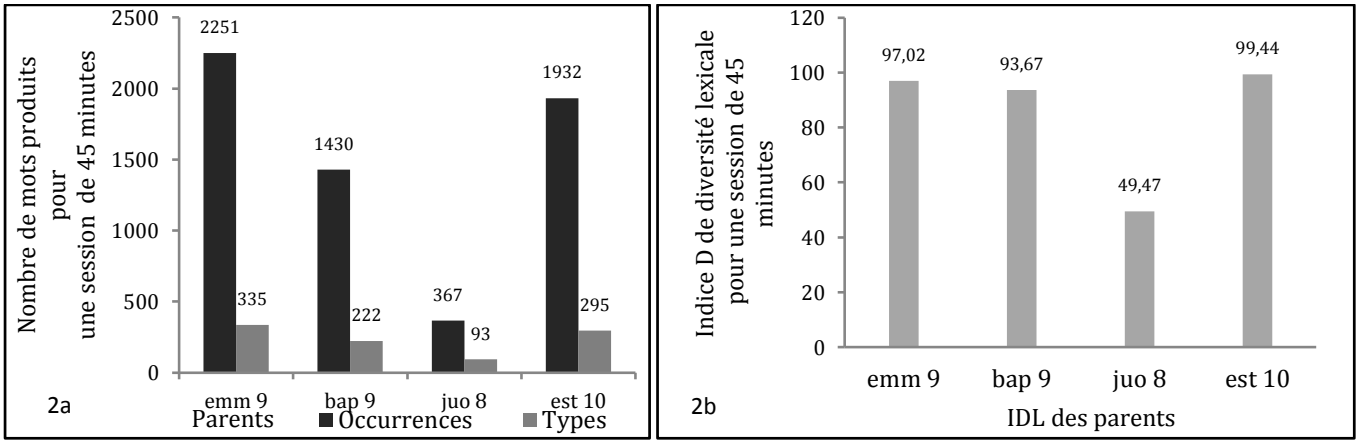


FIGURE 2 : Comportements linguistiques de 4 parents français au stade 1-5 mots (2a occurrences et types de mots produits et 2b diversité lexicale)

Des différences interindividuelles sont aussi présentes chez les enfants au stade linguistique de 50 mots (figure 3a), que ce soit en termes d'occurrences (allant de 25 à 255 mots ; moyenne de 113,75 ; écart-type de 104,59), de types (allant de 8 à 42 types ; moyenne de 23 ; écart-type de 17,63), ou de diversité lexicale (les valeurs de l'IDL allant de 0 à 12,4 moyenne de 8,19 ; écart-type de 6,03). Les enfants qui ont le plus d'occurrences (255 et 130 *vs.* 25 et 45) ont aussi le plus de types (42 et 34 *vs.* 8) et ont aussi la plus grande diversité lexicale (10,9 et 12,4 *vs.*0 et 1,28).

Si l'on compare le stade linguistique de 1-5 mots à celui de 50 mots, on remarque une discontinuité de comportements chez deux enfants sur quatre enfants. En effet, l'enfant qui avait le plus de mots au stade précédent (27 *vs.* moins de 20) est maintenant celui qui en produit le moins (45). Inversement, l'enfant qui avait le moins de production de mot (7 *vs.*plus de 15) a à ce stade un plus grand nombre de mots (130).

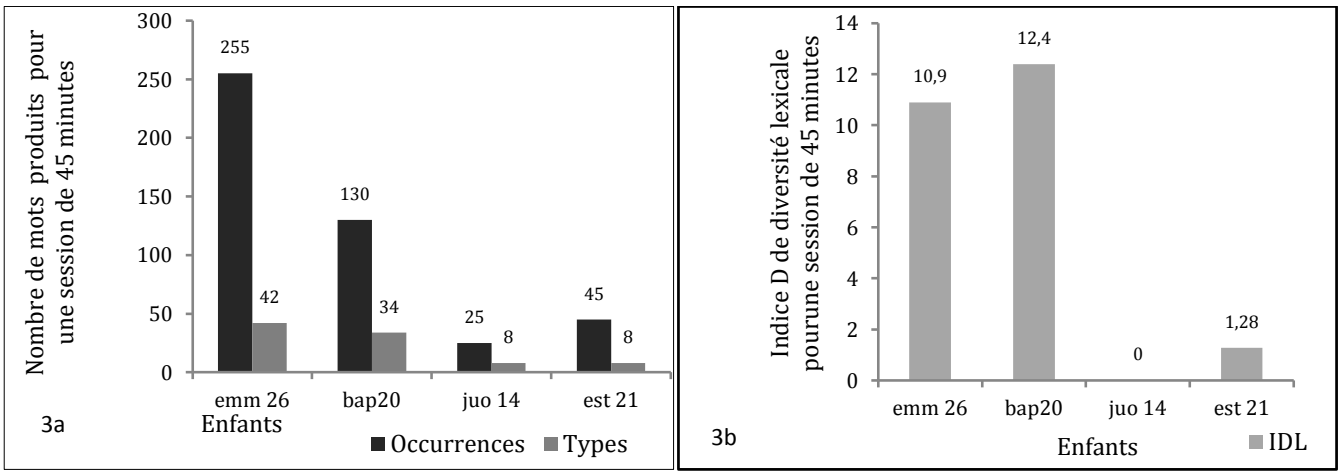


FIGURE3 : Comportements linguistiques de 4 enfants français au stade 50 mots (3a occurrences et types de mots produits 3b diversité lexicale)

Du côté des parents, d'importantes différences interindividuelles sont notées pour le nombre d'occurrences (moyenne de 1973 ; écart-type de 1081,40), de types (moyenne de 273,75 ; écart-type de 111,24) et la diversité lexicale (moyenne de 74,1 ; écart-type de 24,96) (Figure 4). Trois parents ont une production de mots supérieure à 1800 mots, un nombre de types supérieur à 250 et une diversité lexicale supérieure à 80, tandis qu'un parent a moins de 600 mots, moins de 150 types et une diversité lexicale inférieure à 50. Le parent qui a produit le plus de mots (2978) et de types (359) est le parent de l'enfant qui a produit le plus de mots (255) et de types (42). De même, le parent qui a produit le moins de mots (525) et le moins de types (120) est le parent de l'enfant qui a produit le moins de mots (25) et de types (8). Dans les deux cas, aucune correspondance n'est relevée entre la diversité lexicale de l'enfant et du parent.

Enfin, nous observons la même distribution des comportements linguistiques des parents d'un stade linguistique à l'autre (Figure 2 et 4). Le nombre d'occurrences et de types augmente chez tous les parents (occurrences : allant de 158 à 727 mots avec une moyenne de 478 mots et un écart-type de 261,59 ; types : allant de 24 à 57 types avec une moyenne de 37,5 et un écart-type de 15,20) tandis que la diversité lexicale diminue pour 3 parents sur 4 (moyenne de 5,55 ; écart-type de 6,62). Un seul parent (est21) voit sa diversité lexicale augmenter de 99,44 à 102,88.

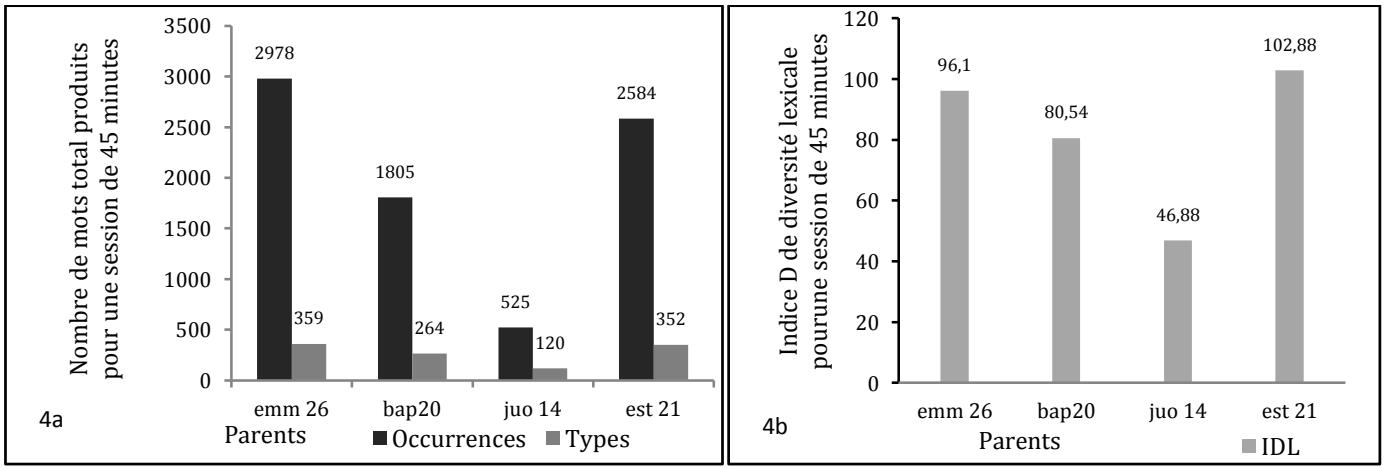


FIGURE 4 : Comportements linguistiques chez 4 parents français au stade 50 mots (4a : occurrences et types de mots produits et 4b : diversité lexicale)

2.3.2 Lien entre comportements linguistiques des enfants et comportements linguistiques et interactionnels des parents

Compte-tenu de ces différents comportements linguistiques observés chez les enfants et leurs parents selon les stades linguistiques, nous avons poursuivi nos analyses pour observer un lien possible entre les comportements linguistiques des enfants et les comportements linguistiques et interactionnels des parents. Nous rappelons que le comportement linguistique est défini à partir du nombre d'occurrences, du nombre de types de mots et de l'indice D de diversité lexicale, et les comportements interactionnels à partir du nombre d'énoncés, des hétéro- et auto-répétitions, et des hétéro- ou auto-clarifications.

Au stade de 1-5 mots il n'existe aucune corrélation entre les comportements linguistiques des enfants et ceux des parents. Par contre, des corrélations positives entre les comportements linguistiques des enfants et les comportements interactionnels des parents apparaissent (cf. tableau 3 ci-dessous). Un effet tendanciel ressort uniquement entre les occurrences de l'enfant et le nombre d'énoncés hétéro-clarifiés du parent (S=0,51 ; r=0,94 ; p=0,05). En d'autres termes, plus un parent a d'énoncés comprenant des hétéro-clarifications, plus l'enfant a un nombre important d'occurrences. Le nombre de types des enfants est positivement corrélé avec le nombre d'énoncés hétéro-répétés (S=0 ; r=1 ; p<2,2e-16) et tend à la significativité pour le nombre de mots hétéro-répétés (S=0,51 ; r=0,95 ; p=0,05). Ainsi, plus un parent utilise les hétéro-répétitions (énoncé et mot), et plus l'enfant a tendance à utiliser des mots différents (types).

	Comportement linguistique enfant	Comportement interactionnel parent		
		Nombre d'énoncés hétéro-répétés	Nombre de mots hétéro-répétés	Nombre d'énoncés hétéro-clarifiés
Stade 1-5 mots	Occurrences			*
	Types	**	*	

TABLEAU 3 –Résultats des tests de corrélation de Spearman (** : p<0.05 ; * : p > 0.05 et <0.08)

Ces résultats rejoignent ceux de la littérature qui montre l'importance des répétitions qui ont pour but de corriger la forme de l'énoncé et aussi de vérifier l'intention des enfants (Bernicot et Clark 2010). Les répétitions peuvent aussi faire progresser le thème de l'échange par l'ajout d'informations (Bernicot et Clark 2010).

On obtient des résultats très différents pour le stade linguistique de 50 mots. Aucun lien n'est présent entre les comportements linguistiques des enfants et les comportements linguistiques et interactionnels des parents.

2.3.3 Lien entre comportements linguistiques et interactionnels des enfants

Compte-tenu de ces différents comportements interactionnels observés chez les parents nous avons poursuivi nos analyses pour observer un lien possible entre les comportements linguistiques des enfants et leurs propres comportements interactionnels.

Au stade de 1-5 mots il n'existe aucune corrélation entre les comportements linguistiques des enfants et leurs propres comportements interactionnels.

On obtient des résultats très différents pour le stade linguistique de 50 mots. On observe des tendances intéressantes entre les comportements linguistiques des enfants et leurs propres

comportements interactionnels (cf. tableau 4). Le nombre d'occurrences tend à augmenter avec le nombre d'énoncés auto-répétés (S=0 ; r=1 ; p=0,08). De plus, le nombre de types des enfants tend à augmenter également avec le nombre d'énoncés auto-répétés des enfants (S=0,5 ; r=0,95 ; p=0,05).

	Comportement linguistique enfant	Comportement interactionnel enfant Nombre d'énoncés auto-répétés
STADE 50	Occurrences	*
	Types	*

TABLEAU 4 –Résultats des tests de corrélation de Spearman

Pour ce stade, ce sont donc les énoncés auto-répétés de l'enfant qui sont liés à un nombre d'occurrences et de types plus grand. Il serait intéressant de savoir si c'est le fait d'avoir plus d'occurrences dans la session qui fait que l'enfant a plus d'auto-répétitions ou si au travers des auto-répétitions que les enfants acquièrent des mots nouveaux.

2.3.4 Lien entre le comportement linguistique de l'enfant au stade 50 mots et les autres comportements au stade 0-5 mots

Aucun lien n'est observé entre les comportements linguistiques des enfants au stade de 50 mots et les comportements linguistiques et interactionnels des parents au stade de 1-5 mots. Cependant, on observe certains liens entre les comportements linguistiques des enfants au stade de 50 mots et leurs comportements interactionnels à 1-5 mots (cf. tableau 5). Les occurrences des enfants au stade 50 mots présentent une corrélation positive tendancielle avec leur nombre d'énoncés hétéro-répétés au stade 1-5 mots (S=0,5 ; r=0,95 ; p=0.05). Les types des enfants au stade 50 mots augmentent eux aussi avec le nombre d'énoncés hétéro-répétés au stade 1-5 mots (S=0, r=1 p< 2,2e-16).

Comportement linguistique enfant Stade 50	Comportement interactionnel enfant Stade 1-5 Nombre d'énoncés hétéro-répétés
Occurrences	*
Types	**

TABLE 5 –Résultats des tests de corrélation de Spearman

La taille du vocabulaire des enfants au stade linguistique de 50 mots est donc prédite par les caractéristiques interactionnelles des enfants à un stade antérieur. Ainsi plus un enfant produira des énoncés auto- et hétéro-répétés au stade linguistique de 1-5 mots, plus il aura d'occurrences et de types de mots au stade linguistique de 50 mots. En effet, plusieurs études dont celles de Bloom, Rocissano, et Hood 1976, ont montré l'importance des imitations dont celles des enfants dans les interactions parents/enfants. Les évènements d'imitations permettent selon les auteurs d'apprendre à utiliser l'information antérieure dans le message des adultes.

2.3.5 Conclusion

Ces résultats préliminaires montrent que le vocabulaire de l'enfant (comportement linguistique) est lié à celui du parent (celui qui s'occupe de l'enfant au quotidien) mais aussi aux comportements interactionnels des parents et des enfants eux-mêmes. De plus, les liens entre les

comportements sont différents en fonction des stades linguistiques. Ainsi, pour le stade linguistique 1-5 mots, on observe un lien entre le comportement interactionnel du parent et le comportement linguistique de l'enfant. Pour le stade linguistique de 50 mots, ce sont les énoncés auto-répétés de l'enfant qui sont liés à leur comportement linguistique. Enfin, uniquement le comportement interactionnel et plus précisément les hétéro-répétitions (énoncés) de l'enfant à 1-5 mots est lié au comportement linguistique de l'enfant au stade linguistique de 50 mots. Cette étude montre la diversité des phénomènes mis en jeux dans l'acquisition du vocabulaire du jeune enfant. Concernant la diversité lexicale de l'enfant, nous n'avons réussi à montrer aucun lien entre parent et enfant, ce qui peut s'expliquer par une quantité trop restreinte de données.

D'une manière générale, l'analyse de l'ensemble des données, (plus de sessions par stade, plus de stades linguistiques et plus d'enfants) pourra confirmer ou infirmer les tendances observées. De nombreuses études ont montré la diminution des hétéro-imitations du langage de l'enfant au profit d'autres structures, (Bloom, Rocissano, et Hood 1976). Un autre prolongement consistera à prendre en compte d'autres variables linguistiques au sein de la dyade (fréquence des énoncés avec babillage vs. énoncés avec mots, proportion de noms et de verbes, etc.) et plus de variables concernant la structuration des échanges par les interactants (Hoff et Naigles 2002). Par ailleurs, la quantité et la qualité d'exposition à la langue environnante dépendent aussi du contexte de production, et plus particulièrement des types de situations (lecture et jeux : Choi 2000). Aussi tiendrons-nous compte de ces différences contextuelles dans de futures analyses.

En conclusion, les caractéristiques de l'input et les caractéristiques interactionnelles de l'enfant lui-même influencent de manière différenciée l'acquisition du vocabulaire. Pour développer son vocabulaire de manière harmonieuse, l'enfant a non seulement besoin d'un input fréquent qui favorise l'apprentissage, mais également d'un propre style interactif qui lui permette de profiter au mieux de la situation d'interaction et des données entendues.

Références

Bates, Elizabeth. 1997. « Origins of language disorders: A comparative approach ». *Developmental Neuropsychology* 13 (3): 447‑476.

Bernicot, Josie, et Eve V. Clark. 2010. « La fonction des répétitions dans les interactions entre parent et jeune enfant: Une étude sur un grand corpus ». *Psychologie de l'Interaction* 26/26: 221‑51.

Bloom, Lois, Lorraine Rocissano, et Lois Hood. 1976. « Adult-child discourse: Developmental interaction between information processing and linguistic knowledge ». *Cognitive Psychology* 8 (4): 521‑552.

Choi, Soonja. 2000. « Caregiver input in English and Korean: use of nouns and verbs in book-reading and toy-play contexts ». *Journal of Child Language* 27 (01): 69‑96.

Fenson, Larry, Philip S. Dale, J. Steven Reznick, Elizabeth Bates, Donna J. Thal, Stephen J. Pethick, Michael Tomasello, Carolyn B. Mervis, et Joan Stiles. 1994. « Variability in Early Communicative Development ». *Monographs of the Society for Research in Child Development* 59 (5): i.

Fenson, Larry, Phillip S. Dale, J. Steven Reznick, Donna Thal, Elizabeth Bates, Jeffrey P. Hartung, Stephen Pethick, et Judy S. Reilly. 1993. MacArthur Communicative Development Inventories: User's guide and technical manual. Singular Publishing Group San Diego

Hoff, Erika, et Letitia Naigles. 2002. « How Children Use Input to Acquire a Lexicon ». *Child Development* 73 (2): 418‑33.

Hoff-Ginsberg, Erika. 1991. « Mother-Child Conversation in Different Social Classes and Communicative Settings ». *Child Development* 62 (4): 782‑96.

Huttenlocher, Janellen, Wendy Haight, Anthony Bryk, Michael Seltzer, et Thomas Lyons. 1991. « Early vocabulary growth: Relation to language input and gender. » *Developmental Psychology* 27 (2): 236‑248.

Kern, S., et F. Gayraud. 2010. *L'inventaire français du développement communicatif (IFDC)*. La Cigale. Grenoble.

McKee, G., D. Malvern, et B. Richards. 2000. « Measuring Vocabulary Diversity Using Dedicated Software ». *Literary and Linguistic Computing* 15 (3): 323‑338.

Nelson, Katherine. 1981. « Individual differences in language development: Implications for development and language ». *Developmental Psychology* 17 (2): 170‑187.

Parisse, Christophe, et Marie-Thérèse Le Le Normand. 2007. « Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans ». *Glossa* 97: 10‑30.

Rose, Yvan, Brian MacWhinney, Rodrigue Byrne, Gregory Hedlund, Keith Maddocks, Philip O'Brien, et Todd Wareham. 2006. « Introducing Phon: A software solution for the study of phonological acquisition ». In *Proceedings of the 30th Boston University Conference on Language Development*, 489‑500.

Tamis-LeMonda, Catherine S., Lisa Baumwell, et Tonia Cristofaro. 2012. « Parent–child Conversations during Play ». *First Language* 32 (4): 413‑438.

Vihman, Marilyn May, et Lorraine McCune. 1994. « When is a word a word? » *Journal of Child Language* 21 (03): 517‑542.

Yoder, Paul J., et Ann P. Kaiser. 1989. « Alternative explanations for the relationship between maternal verbal interaction style and child language development ». *Journal of Child Language* 16 (01): 141‑160.

Les occlusives de l'alsacien : une étude temporelle

Steiblé Lucie[1], Sock Rudolph[1,2]
Université de Strasbourg
E.A. 1339 Linguistique, Langues et Parole – LiLPa
E.R. Parole et Cognition & Institut de Phonétique de Strasbourg - IPS
(1) Le PATIO : 22 rue René Descartes 67084 Strasbourg Cedex, Bureau 4418.
(2) Université Paul Josef Safarika, Kosice, Slovaquie
steiblelucie@gmail.com, sock@unistra.fr

RESUME

Cet article traite des consonnes occlusives de l'alsacien. Les résultats présentés proviennent des enregistrements de trois locutrices de l'alsacien, également francophones. Elles ont été enregistrées dans leur langue maternelle, l'alsacien. Le corpus d'étude comprends toutes les consonnes occlusives de l'alsacien (transcrites à l'écrit comme en français : **p t k** et **b d g**). L'analyse de ces enregistrements montre d'une part que les deux séries de consonnes occlusives de l'alsacien ne sont jamais sonores, mais que leur distribution correspond à celles de paires minimales oppositives. La distinction entre les deux séries repose donc sur d'autres indices, tels que les durées intra-segmentales, notamment les durées des silences au sein même des deux séries de consonnes.

ABSTRACT

This article deals with Alsatian plosives, and French plosives as pronounced in French by Alsatian speakers. Results are based on data from 3 Alsatian female speakers, who also speak French. They were recorded in their native language, Alsatian. The corpus includes all Alsatian stops (which are written **b d g** and **p t k**, like in French). Analyses of the data show that there are no voiced stops in Alsatian, and that despite absence of the voiced / voiceless contrast, these stops may be phonological. . Hence, the distinction between the series is based on various cues, such as intra-segmental phases, and especially the duration of the acoustic silent phase.

MOTS-CLES : alsacien, lecture événementielle, occlusives, voisement.
KEYWORDS: Alsatian, event-based approach, , stops, voicing.

1 Introduction

Langue dialectale en usage dans l'est de la France, l'alsacien est parlé en Alsace par 39 % des adultes, soit environ 500 000 personnes (données INSEE, 1999). Pourtant, l'alsacien a rarement fait l'objet d'études en phonétique expérimentale, bien que son système phonologique soit saillant lorsque les alsaçophones parlent français, avec leur « accent » caractéristique. L'assourdissement de certaines occlusives normalement sonores du français est perçu par les autres locuteurs non alsaciens, donnant lieu à des jeux de mots, blagues et sketchs fondés sur les modifications du trait de sonorité. Ces modifications en français sont probablement liées à la nature des consonnes occlusives de l'alsacien. Ces consonnes ne semblent pas partager le trait de sonorité avec les consonnes occlusives du français.

Quelles sont alors les caractéristiques acoustiques des consonnes de l'alsacien observables sur un signal sonore ? Cette question, point de départ de notre recherche, permet de soulever plusieurs problématiques intéressantes du point de vue de la phonétique acoustique et expérimentale.

1.1 A propos des occlusives alsaciennes

La revue de littérature concernant la phonétique des occlusives alsaciennes est rapide : les recherches sont, à notre connaissance, presque inexistantes. Il existe cependant un questionnement chez les sociolinguistes concernant les corrélats phonétiques de l'accent alsacien, comme on peut le lire chez Bothorel-Witz (2007) :

« Dans le cas de l'Alsace, il se pose la question de savoir si ce qui est catégorisé par le sociolinguiste comme la variété normée ne doit pas inclure certains corrélats phonétiques ou prosodiques de l'« accent alsacien ». Cette interrogation nous paraît d'autant plus fondée que les productions de bien des locuteurs dialectophones (dont le français peut être la langue dominante) présentent souvent des marqueurs phonétiques et prosodiques qui échappent à la correction. Les descriptions, comme le souligne très justement Tabouret-Keller (1988 : 88), restent« relativement coupées de l'observation et de l'établissement des données empiriques réactualisées » »

C'est pourquoi notre étude a pour but de tenter de discerner la structure sous-jacente à ce système consonantique, différent de celui du français puisque spécifiquement vecteur de modifications sur les consonnes du français, lorsque des dialectophones parlent la langue nationale, mais disposant de deux séries graphiques de consonnes occlusives.

1.2 Quels traits et quels indices ?

« Il est reconnu que la notion phonologique de trait s'appuie sur des bases phonétiques, faute de quoi l'analyse phonologique se réduit à une construction intellectuelle, un simple jeu de l'esprit. »

(Bothorel-Witz & Pétursson, 1972)

Notre objectif sera donc de discerner parmi les réalités mesurables sur le signal de parole, quels événements sont mis en jeu pour séparer les consonnes de l'alsacien en deux séries : celles qui sont graphiées **p t k**, et celles qui sont graphiées **b d g**. Nous renvoyons à Abry et al. (1985) pour plus de détails sur la lecture événementielle des signaux de parole. Ces deux séries apparaissent en distribution complémentaire, dans des paires de mots comme ***Gass*** (la rue) et ***Kass*** (la caisse), ***Trottel*** (la gourde) et ***Troddel*** (le gland), etc.

Notre hypothèse est que les consonnes de l'alsacien sont en fait distinguées non pas par le trait de voisement, mais par le trait [fortis/lenis], terme que nous empruntons aux chercheurs ayant travaillé sur l'allemand, notamment Kohler (1984). L'allemand, langue cousine de l'alsacien et issue de la famille des langues germaniques, partage certaines caractéristiques avec l'alsacien. En l'absence de travaux sur ce dialecte spécifique, nous soutenons que l'opposition admise en allemand, qui repose sur une opposition de tension plutôt que de voisement est applicable à l'alsacien. En effet, sur aucun signal les consonnes enregistrées ne montrent de périodicité, révélatrices de vibration des plis vocaux, pas plus qu'une activité dans les basses fréquences du spectre pendant la totalité de la phase d'occlusion. Il s'agit toujours d'un silence acoustique, comme en témoignent la Figure 1, ci-dessous :

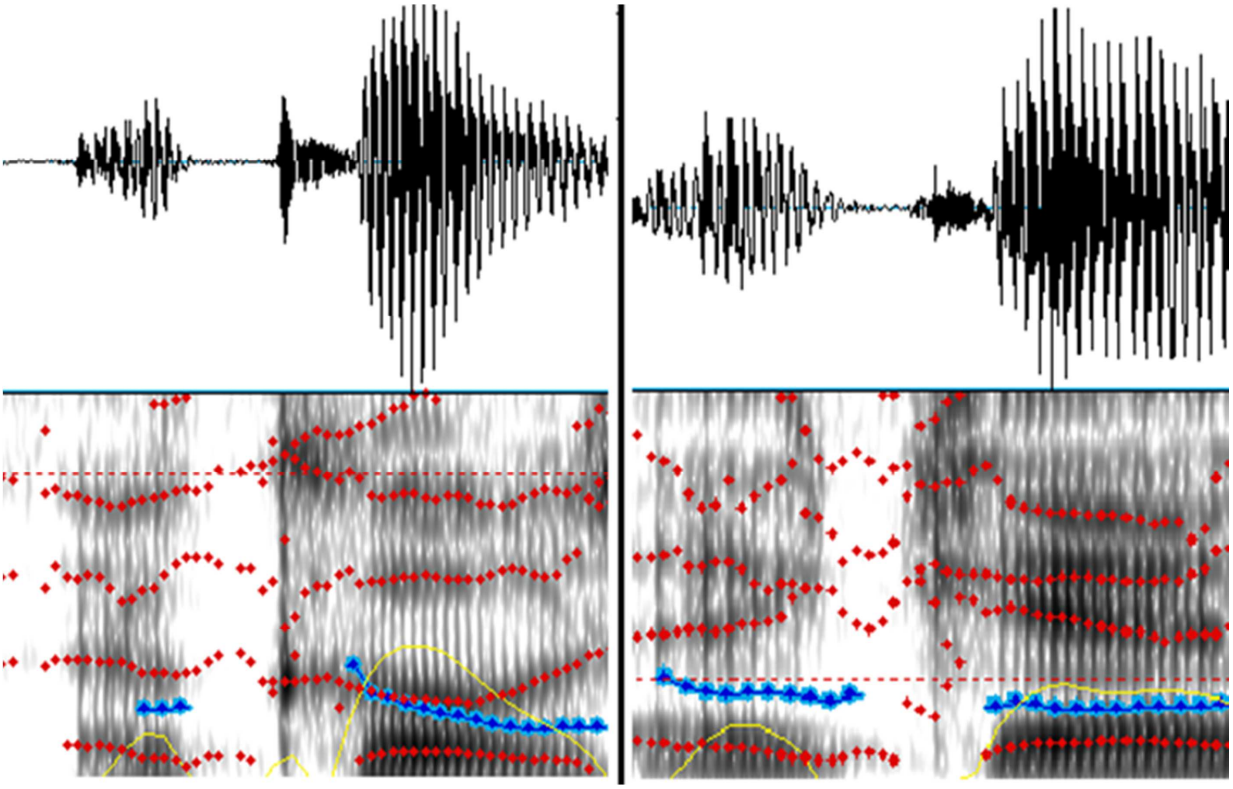


FIGURE 1 – signaux sur PRAAT de /aka/ (à droite) dans « a Kass » de /aga/ (à gauche) dans « a Gass »

Parmi les indices du trait [fortis/lenis], on cite souvent différents types de durée, incluant la durée de la voyelle précédente, la durée de l'occlusion, la durée de l'aspiration, ou encore la durée totale de la consonne (cf. Jakobson, Fant et Halle 1952, Ladefoged & Maddieson 1996, Jessen 1998). La durée, selon Jakobson, Fant & Halle (1952), est le corrélat primaire du trait [fortis/lenis], ou « tendu » :

"In consonants, tenseness is manifested primarily by the length of their sounding period, and in stops, in addition, by the greater strength of the explosion".

Jakobson & Halle (1961) vont même plus loin en considérant la durée non seulement comme le corrélat primaire mais probablement comme le seul indice du trait de tension :

"The relative duration of the consonant and the antecedent phoneme may remain for certain contextual or optional variants of tense and lax consonants the chief or even the only cue to their distinction."

Fischer-Jørgensen (1976), à partir des données de l'allemand du Nord, a trouvé que la durée d'occlusion des segments p, t, k est en moyenne de 44 ms plus longue que celle de b, d, g De telles différences significatives ont aussi été observées par Haag (1979) et Kohler (1979, 1982). On attribue généralement la durée plus importante des occlusives sourdes, comparées à leurs contreparties sonores, à une plus grande force d'articulation les caractérisant.

Un autre corrélat du trait [fortis/lenis], fréquemment cité dans la littérature est la durée de la voyelle précédant la consonne : les voyelles sont plus courtes quand elles précèdent les consonnes tendues. Pour Malécot, c'est l'indice le plus important : *« Vowels […] are short before fortes and long before lenes »* (1966).

Le trait de tension étant particulièrement difficile à observer à l'aide d'indices acoustiques, il serait

possible de s'aider ainsi des voyelles devant et derrière les consonnes occlusives de l'alsacien pour déterminer si l'on peut les classer en deux groupe grâce à la tension. Comme on peut le lire chez Ridouane (2003) : *« Il a été largement observé que les voyelles sont plus brèves devant les sourdes que devant les sonores (House & Fairbanks 1953 ; Delattre 1962 ; Chen 1970 ; Klatt 1973 ; Port et al. 1980 ; Port 1981 ; Keating 1985) »* Ce corrélat joue probablement son rôle dans la distinction des consonnes occlusives alsaciennes, qui sont toutes sourdes.

Les voyelles suivantes peuvent aussi être influencées par la tension, comme c'est le cas pour les géminées et leurs conséquences sur les voyelles qui les suivent. Selon Local & Simpson (1999), à partir des données sur le malayalam, la gémination a une influence non seulement sur la durée des voyelles précédentes mais aussi sur les voyelles suivantes. Ainsi, la classification des consonnes pourrait aussi être éclaircie par l'analyse de l'entourage des phonèmes qui nous occupent, prouvant si besoin en était l'importance des effets de coarticulation.

2 Corpus et méthodes

Les trois locutrices, J, C et M sont âgées de 59, 63 et 67 ans. Toutes sont de langue maternelle alsacienne, et ont appris le français à l'école, entre l'âge de 3 et 5 ans. Toutes parlent le français tout comme l'alsacien au quotidien. Elles ont étés enregistrées dans une chambre insonorisée, à l'aide d'un magnétophone Marantz Professional© PMD661, lié à un microphone directif Sennheiser© e845S.

Le corpus est en alsacien, comportant 6 mots à prononcer dans de courtes phrases porteuses :

Phrase montrée	Traduction	Logatome
a Bumb	Une bombe	/abu/
a Pump	Une pompe	/apu/
a Dritt	Un coup	/adri/
a Tritt	Un troisième	/atri/
a Gass	Une ruelle	/aga/
a Kass	Une caisse	/aka/

TABLE 1 – corpus

Le corpus a été répété douze fois, les locutrices recevant comme consigne de s'assoir confortablement à 20 centimètres du microphone, et de lire à haute voix les phrases présentées sur des cartons, en ordre aléatoire.

L'analyse acoustique à proprement parler a été conduite à l'aide du logiciel PRAAT. Nous avons segmenté la première voyelle du logatome, de formants stables à formants stables. Le Voice Termination Time (VTT), ou délai d'arrêt du voisement, a été mesuré, pour les consonnes non voisées évidemment. Agnello (1975) décrit le VTT comme étant l'intervalle allant de la disparition de la structure formantique clairement définie de la voyelle précédant la consonne (disparition de cette structure provoquée par une closion articulatoire) jusqu'à la dernière vibration glottique dans

la tenue consonantique. Selon certains auteurs, le VTT est le délai nécessaire à l'arrêt de vibration des plis vocaux après la fermeture complète du conduit vocal, comme chez Sock (1993) ou Abry et al (1985) par exemple.

Le Voice Onset Time (VOT), ou délai d'établissement du voisement, a été mesuré à la manière de Klatt (1975), c'est-à-dire du relâchement consonantique à l'établissement d'une structure formantique stable pour la voyelle suivante.

Enfin, le silence acoustique a été mesuré entre le VTT et le VOT pour les consonnes sourdes. Pour les consonnes sonores, la tenue consonantique s'étend de la disparition de la structure formantiquement stable de la voyelle précédente jusqu'au VOT.

Après segmentation, nous avons utilisé un script (celui de Gendrot [1]) afin d'extraire automatiquement nos durées. Toutes nos mesures ont été vérifiées par la suite.

3 Résultats

3.1 Analyses des durées

Aucune des consonnes mesurées ne disposait de vibrations laryngées caractéristiques du voisement. On peut simplement dire qu'en alsacien, pour autant que nous le sachions, toutes les consonnes occlusives sont sourdes, elles disposent donc toutes d'un VTT mesurable.

Voici, dans les histogrammes ci-dessous (Figures 2 et 3), les durées des différents segments mesurés. Les données sont montrées ici par paires : la première voyelle de /abu/ (V1b) avec la première voyelle de /apu/ (V1p), et de même pour les Voice Termination Time (VTT), les silences acoustiques (sil), les Voice Onset Time (VOT) et les secondes voyelles (V2).

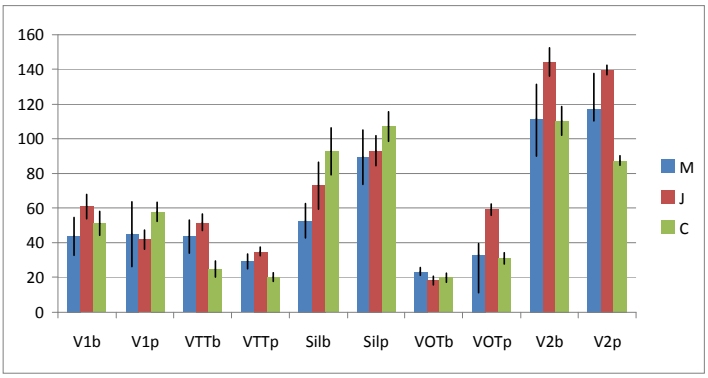


FIGURE 2 – histogramme des durées (en ms) des logatomes /aba/ et /apa/ en fonction des locuteurs

[1] Nous tenons à remercier Cédric Gendrot pour ce script : http://ed268.univ-paris3.fr/lpp/pages/EQUIPE/gendrot/page_web/scripts.htm

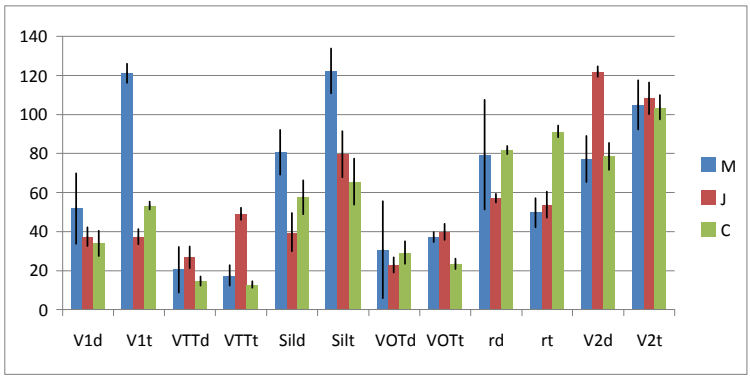


FIGURE 3 – histogramme des durées (en ms) des logatomes /ada/ et /ata/ en fonction des locuteurs

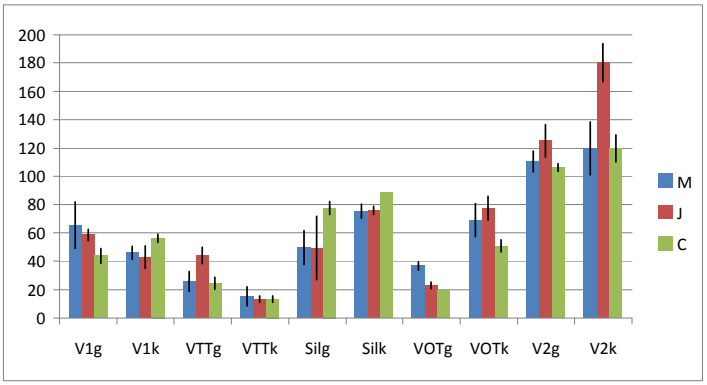


FIGURE 4 – histogramme des durées (en ms) des logatomes /aga/ et /aka/ en fonction des locuteurs

Ces données absolues montrent que les durées des silences des consonnes graphiées p,t ou k sont toujours plus longs que ceux de b, d ou g. Les VOT suivent la même règle.

Les VTT ont une nette tendance à être plus longs quand les consonnes proviennent de la série **b,d,g** comparativement à leurs homologues de la série **p,t,k.**

Les voyelles ont un comportement qui semble plus difficile à expliquer par la présence de l'un ou l'autre type de consonne ; nous y reviendrons plus en détail.

3.2 Etude du comportement des paramètres inter-segmentaux vocaliques et intra-segmentaux des consonnes : approche statistique

Les données analysées comportent 10 items dans chaque série. Dans le cadre de l'analyse statistique, les données de cette expérience représentent donc 30 items (10x3 locutrices). Les données ont en premier lieu été représentées en histogrammes de fréquence et sous forme de « boîtes à moustaches » à l'aide du logiciel R[2], afin de déterminer leur répartition. Les résultats montrent que les moyennes en valeurs absolues ne suivent pas complètement une distribution normale ; nous avons donc décidé d'utiliser, pour travailler sur la variance des données, un test

[2] R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

statistique adapté aux données non-paramétriques. Les variables que nous analysons sont de deux nature : une nominale (le type graphique de consonne, de la série **b,d,g** ou **p,t,k**) et une variable numérique (la durée, en milliseconde, de chaque segment.) Le test choisi est donc *le test de Wilcoxon.*

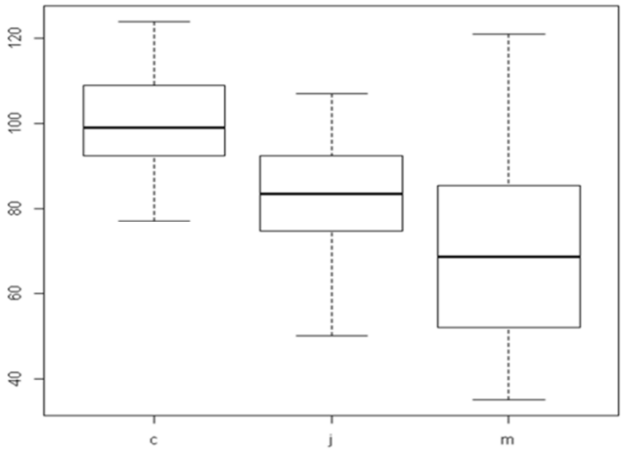


Figure 5 : Exemple de boîtes à moustaches des durées des silences des /p/ en fonction des locutrices. Les données ne suivent pas tout à fait la loi normale.

3.2.1 Les voyelles

Nous avons déjà évoqué le comportement instable des voyelles, déjà visible sur les histogrammes des durées. Dans l'ensemble, les voyelles initiales ne montrent pas de tendance claire à être courtes ou longues selon la consonne qui les suit. Les voyelles après les consonnes, quant à elles, montrent une tendance à être plus longue après les consonnes graphiées **ptk**, mais selon les locutrices, le phénomène est inversé. Après analyse, les tests de Wilcoxon permettent d'affirmer qu'il n'y a pas toujours de signifiance de la durée des voyelles en fonction du type de consonne. Ces résultats sont en partie explicables par la variabilité inhérente aux réalisations chez certaines locutrices, surtout pour la voyelle initiale.

- les voyelles initiales (V1) :

Pour la paire p/b : p-value = 0.1998, pour la paire t/d : p-value = 0,0000003514, pour la paire k/g : p-value = 0.01423

- les voyelles finales (V2) :

Pour p/b : p-value = 0.09523, pour t/d : p-value = 0.008611, pour k/g : p-value = 0.002139.

3.2.2 Le Voice Termination Time

Les VTT ont une tendance, visible sur l'histogramme, à être plus longs après les consonnes **graphiées bdg**, comme attendu dans la littérature pour des VTT de consonnes [lenis]. Les tests montrent, pour certaines paires, un lien significatif entre le type de consonne et la durée absolue du VTT. La différence entre t/d n'est en revanche pas corrélée significativement à la durée du Voice Termination Time. Il est possible que ce résultat soit dû aux VTT très longs produits par l'une des locutrices (J) lors de la réalisation des /t/ du corpus, conduisant à une moyenne nettement plus élevée pour cette consonne, à l'inverse des attentes et des autres cas de l'étude.

Dans le cas du rapport entre les graphies p/b et leurs durées, nous obtenons une p-value = 0.0004566. Pour t/d et leurs durées, nous enregistrons une p-value = 0.899. Enfin, pour la paire k/g, p-value = 0,0000000001309

3.2.3 Le silence

Les silences de toutes les paires montrent une tendance assez marquée à être plus longs pour la série **ptk**. Bien entendu, cela correspond aux descriptions des consonnes dotées du trait [fortis]. Les tests statistiques révèlent une grande corrélation entre la durée de ce silence, et le type de consonne. Ces résultats statistiques, très stables quels que soit le lieu d'articulation ou la locutrice, confirment ce qui est visible dès l'analyse segmentale : sur les signaux déjà, les durées des silences au sein des consonnes sont visiblement plus longues pour les consonnes **graphiées ptk**.

Pour les durées des silences de p/b en fonction de la graphie : p-value = 0,0000031056.

Pour t/d : p-value = 0,00005395

Pour k/g : p-value = 0,00002277

Comme cela a été décrit dans la littérature concernant le trait [fortis/lenis], les durées sont des indices particulièrement efficaces pour catégoriser les deux types. Nous pouvons ajouter ici que la durée du silence au sein de ces consonnes sourdes est un indice très efficace pour prédire la catégorie de la consonne, ainsi, en l'occurrence, que la façon dont elle sera graphiée.

3.2.4 Le Voice Onset Time

Enfin, les VOT eux aussi sont habituellement plus long dans le cas des consonnes de la série **ptk**.

Pour la paire p/b, les durées des VOT sont effectivement fortement conditionnées par le type de consonne : p-value = 0,000000001759. Pour la paire t/d, le type de consonne est également lié à la durée du VOT : p-value = 0.009681. Pour k/g : p-value = 0,0000000002235

Le Voice Onset Time est également un paramètre fiable, qui permet de corréler statistiquement le type de consonne avec une durée.

4 Discussion

Nous avons observé des problèmes de production des alsaçophones en français, une langue qui exploite le trait de sonorité pour opposer ses consonnes occlusives. Nous nous sommes alors demandé quelles étaient les caractéristiques acoustiques des consonnes occlusives de l'alsacien, puisque si ces caractéristiques étaient les mêmes qu'en français, les locuteurs de l'alsacien ne transformeraient pas ces sons. Il doit cependant exister une différence entre ces deux séries de consonnes, qui disposent de deux séries de graphèmes pour les représenter, et qui apparaissent dans des contextes identiques mais en induisant un changement de sens.

En premier lieu, nous avons noté que les consonnes de l'alsacien n'étaient jamais sonores : à aucun moment, il n'est possible d'observer des vibrations laryngées qui se maintiennent sur toute la durée de la consonne, ou même, sur seulement une moitié de la consonne. Les occlusives de l'alsacien ne sont jamais caractérisées par une vibration des plis vocaux : au contraires, les durées des VTT sont comparables, pour les plus longues, à celles des occlusives sourdes du français.

Les signaux, visiblement (cf. Figure 1), montrent des différences de durées, qui dénotent une opposition plus quantitative (liée aux durées) que qualitative. À l'inverse, le français exploite en premier lieu la qualité de la consonne (vibration des plis vocaux, ou non) pour déterminer si la consonne relève de la série voisée ou non voisée. Ces phénomènes d'opposition de durée sont corroborés lors des analyses de variance appliquées aux données : les paramètres intra-

segmentaux, essentiellement, sont liés significativement au type de consonne.

Ainsi, deux langues proches (deux langues en contact, plus précisément) n'exploitent pas les mêmes relations temporelles entre événements acoustiques pour obtenir le même résultat phonologique : opposer six consonnes occlusives par paires et selon le lieu d'articulation. L'organisation temporelle de la parole dans les deux langues semble être donc distincte, comme le montre les données issues des productions des locutrices.

5 Conclusion

En alsacien, contrairement au français, les deux séries de consonnes graphiées ptk et bdg correspondent à des phonèmes non pas opposés phonologiquement selon le trait de voisement, mais plutôt selon le trait **[fortis/lenis]**.

Les consonnes graphiées ptk, **fortis**, sont dotées de VTT plus courts mais de silences et de VOT significativement plus longs que leurs homologues **lenis**. Ces dernières consonnes disposent de VTT plus longs, mais de silences et de VOT significativement plus courts.

Il nous reste maintenant à compléter les données absolues de cette investigation par une approche relative qui prendrait en considération l'élasticité bien connue du signal de parole, suivant les locuteurs et leur vitesse d'élocution qui leur est propre.

Références

ABRY, C., BENOIT, C., BOE, L.J., & SOCK, R. (1985). Un choix d'événements pour l'organisation temporelle du signal de parole. 14èmes Journées d'Etudes sur la Parole, Société Française d'Acoustique, 133-137.

BOTHOREL-WITZ, A. PHILIPP, M. SPINDLER, S. (1984). *Atlas linguistique et ethnographique de l'Alsace,* Volume II : Paris : Editions du C.N.R.S.

CHEN M. (1970). Vowel length variation as a function of the voicing of the consonant environment. Phonetica 22, 129-159.

DELATTRE P. (1962). Some factors of vowel duration and their cross-linguistic validity. Journal of the Acoustical Society of America 34 : 1141-1143.

FISCHER-JORGENSEN E. (1976). Some data on North German stops and affricates. Annual Report of the Institute of Phonetics, University of Copenhagen 10, 149-200.

HAAG W. (1979). An articulatory experiment on voice onset time in German stop consonants. Phonetica 36, 169-181.

HUCK, D. BOTHOREL-WITZ, A. et GEIGER-JAILLET, A. (2007). L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière. *In Aspects of Multilingualism in European Border Regions: Insights and Views of Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol, p. 13-100.*

JAKOBSON R., FANT, G., HALLE M. (1952). Preliminaries to speech analysis: The distinctive features and their correlates. Cambridge, Mass.: M.I.T. Press.

JAKOBSON R., HALLE M. (1961). Tension et laxité. Essais de Linguistique Générale, 150-157. Les Editions de Minuit, Paris.

JESSEN M. (1998). Phonetics and phonology of tense and lax obstruents in German. John Benjamins Publishing Company, Amsterdam/Philadelphia.

KIM C.W. (1965). On the autonomy of the tensity feature in stop classification. Word 21, 339-359.

KOHLER K. (1979). Dimensions in the perception of fortis and lenis plosives. Phonetica 36, 332-343.

KOHLER K. (1982). F0 in the production of lenis and fortis plosives. Phonetica, 39(4-5), 199-218.

LADEFOGED P., MADDIESON I. (1996). The sounds of the world's languages. Blackwell Publishers: Oxford.

MALECOT A. (1966). The effectiveness of intraoral pressure pulse parameters in distinguishing between stop cognates. Phonetica 14, 65-81.

RIDOUANE R, (2003) Suites de consonnes en berbère chleuh : phonétique et phonologie. Thèse de doctorat d'Etat, Université de la Sorbonne Nouvelle, Paris III. KOHLER, K. (1984) Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica,* 41, (3): 150-74.

Etude des voyelles et de la force de voix par analyse discriminante

Jean-Sylvain Liénard[1], Claude Barras[1,2]
(1) LIMSI-CNRS
(2) Université Paris Sud
{jean-sylvain.lienard, claude.barras}@limsi.fr

RESUME

L'effort vocal, représenté ici par une mesure d'intensité objective appelée force de voix, est à la fois un facteur de variabilité de la parole et une grandeur acoustique utilisée par les interlocuteurs pour échanger diverses informations dans une situation donnée. La présente étude s'intéresse aux indices acoustiques codant ces informations dans le spectre des voyelles. L'Analyse Discriminante est mise en œuvre d'une part pour identifier les voyelles et d'autre part pour estimer la force de voix en dépit de leurs variabilités mutuelles. Les résultats, établis sur deux bases de données différentes, montrent que la force de voix peut être estimée avec précision à partir du spectre des voyelles et que la connaissance préalable de la force de voix permet d'améliorer la classification des voyelles.

ABSTRACT

Vocal Effort, represented here by an objective intensity measurement called Voice Strength, is both a speech variability factor, and a physical quantity used by interlocutors in order to exchange some types of information in a given situation. The present study deals with the acoustical features coding this information in the vowel spectrum. Discriminant Analysis is used to identify the vowels as well as to estimate the voice strength despite their mutual variabilities. The results established on two different databases, show that voice strength may be precisely estimated from the vowel spectra, and that knowing the voice strength improves the classification of the vowels.

MOTS-CLES : Voix, parole, communication orale, voyelles françaises, effort vocal, analyse acoustique, analyse discriminante, interactions voix-parole.
KEYWORDS: Voice, speech, oral communication, French vowels, vocal effort, acoustic analysis, discriminant analysis, voice-speech interactions.

1 Introduction

La communication orale est ***située*** et ***interactive***. Lorsqu'on parle, on s'adresse toujours à un interlocuteur, réel ou imaginaire, proche ou lointain, familier ou non. Le signal vocal encode divers types d'information, de nature linguistique (parole) ou non (voix), intimement mélangés. L'effort vocal (EV) est ajusté par le locuteur, de manière consciente ou non, afin d'assurer la transmission de ces informations à l'interlocuteur, en fonction des caractéristiques de ce dernier, de leur situation mutuelle, du cours de l'échange oral et des conditions de la transmission acoustique (distance, bruit ambiant).

L'EV est une notion qualitative, issue du domaine de la phoniatrie. Nous faisons l'hypothèse qu'elle peut être indexée par l'intensité acoustique mesurée à une distance donnée de la bouche du locuteur, selon des modalités particulières (pics d'intensité mesurés sur les noyaux vocaliques selon une fenêtre temporelle de l'ordre de 50 ms). Dans la suite nous appelons cette grandeur physique "force de voix" (FDV). L'expérience quotidienne montre à l'évidence que la force de voix est distincte de l'intensité du signal reçu par l'auditeur: on ne change pas une voix faible en une voix forte simplement en l'amplifiant de quelques dB ou en s'approchant du locuteur.

L'effort vocal relève de plusieurs problématiques différentes, concernant l'intelligibilité de la voix criée (ROSTOLLAND, 1982), le rôle de la distance entre les interlocuteurs (TRAUNMULLER and ERIKSSON, 2000), l'écoute dans le bruit et l'effet Lombard (JUNQUA, 1992), (GARNIER, 2007), le fonctionnement de la source glottique (FANT et al, 1985), (DOVAL et al 2006), (HANSON,1997), l'incidence de l'EV sur les formants (LIENARD and DI BENEDETTO, 1999), (HUBER et al, 1999), sur la qualité de la voix chantée (HENRICH et al, 2005), ainsi que sur la prosodie (D'ALESSANDRO et al, 2006). Les travaux menés en reconnaissance de la parole (ZELINKA et al, 2012) ou du locuteur (BRUNGART et al, 2001) (HANSEN and VARADARAJAN, 2009) montrent une grande dégradation des résultats avec l'EV, catégorisé en un petit nombre de types (chuchoté, faible, normal, fort, crié ou Lombard), mais jamais envisagé comme repérable par une grandeur mesurable telle que la FDV. Les références données ici à titre indicatif n'ont rien d'exhaustif mais permettent de comprendre la diversité des points de vue sur l'EV.

Les différentes études convergent sur les observations suivantes. Le fait de parler plus fort entraîne le déplacement vers l'aigu du "formant glottique" variant avec f0 et avec le quotient ouvert (Oq) des cordes vocales. Le renforcement peut se manifester dans toute l'étendue du spectre. La pente spectrale à long terme ("spectral tilt") représente globalement le rehaussement des composantes spectrales élevées. Le fondamental f0 augmente avec l'EV, d'environ une octave entre voix très faible et voix très forte. La fréquence du 1er formant F1 augmente de manière significative avec l'EV. Enfin la différence d'amplitude entre le 1er harmonique (fondamental) et le second semble en relation avec le formant glottique, mais ceci dépend de f0 et de la voyelle.

Dans ce qui suit nous décrivons les deux bases de données utilisées ainsi que les outils d'analyse des voyelles françaises pour diverses valeurs de la FDV. Nous présentons l'application de l'Analyse Discriminante à l'identification des voyelles, puis à l'estimation de la FDV, et nous examinons en quoi l'une peut être utilisée pour améliorer l'autre.

2 Données

Pour étudier la force de voix de manière quantitative il faut des données étalonnées en niveau sonore, dont on puisse être sûr du niveau sonore émis (en dB SPL), et surtout de la constance des conditions de prise de son (distance bouche-microphone, chaîne de traitement électroacoustique). Il n'existe pas, à ce jour, de base de données de grande taille offrant toutes les garanties d'étalonnage et de constance de la prise de son. Nous avons donc dû nous contenter de deux bases de données de taille réduite disponibles dans notre laboratoire.

2.1 Données CRC

La base de données CRC, utilisée dans l'article (LIENARD et DI BENEDETTO, 1999) comporte 12 voyelles françaises (9 orales, 3 nasales) prononcées isolément par 6 locuteurs et 7 locutrices, selon 3 degrés d'effort vocal suscités par des variations de distance entre locuteur et opérateur. L'ensemble a été enregistré en 3 sessions, dont la 3e six mois après la première. Le locuteur était assis, sa bouche à 30 cm du microphone omnidirectionnel, et devait répéter une courte phrase puis des voyelles émises par l'opérateur, lui-même debout et placé à une distance choisie parmi trois: 0.40 m (condition "p" pour proximité), 1.50 m (condition "n" pour normal) ou 6 m (condition "l" pour loin). Les locuteurs n'ont pas tous participé aux 3 sessions; le corpus se compose de 720 sons ou "tokens", soit 20 séries de 36 tokens, chaque série comportant les 12 voyelles émises par un même locuteur selon les 3 conditions p, n et l. Les voix restent dans la dynamique usuelle de la conversation (30 à 35 dB entre les sons les plus faibles et les plus forts) et les conditions d'enregistrement sont maintenues constantes.

2.2 Données JAE

Ce corpus (AUGUSTE-ETIENNE, 1999) comporte des syllabes de type CV et des voyelles isolées /a, i, u/ produites par 7 hommes, 7 femmes et 3 enfants. Chaque syllabe est répétée ad libitum en partant d'une voix que le locuteur considère comme neutre. La consigne est de produire un effort vocal croissant par paliers jusqu'à un maximum proche de la voix criée, puis décroissant, et de recommencer cette fois en baissant la voix jusqu'à un minimum proche de la voix chuchotée, puis en retournant à la valeur neutre. La prise de son est effectuée avec deux microphones placés à distance fixe (20 cm) de la bouche du locuteur. L'un des deux est un microphone de mesure, étalonné de temps en temps au moyen d'une source étalon (1 kHz, 94 dB). La différence de niveau sonore entre les sons les plus faibles et les plus forts excède 60 dB. La présente étude ne prend en compte que les voyelles isolées, produites en nombre variant de 39 à 232 selon les locuteurs, pour un total de 1452 tokens.

3 Outils et méthodes

3.1 Sélection et prétraitement

Tous les calculs sont effectués en Praat (BOERSMA and WEENINK, 2012). L'intensité ***ax*** (en dB par rapport à une référence arbitraire) et le fondamental ***f0*** (en Hz) sont calculés sur une fenêtre gaussienne de 50 ms. Le signal est ensuite pondéré selon la courbe A des sonomètres (atténuation des fréquences inférieures à 1 kHz) et son intensité ***ap*** est calculée (en dBA). L'analyse utilise la méthode BarkFilter en 18 bandes de 1 Bark. Le spectre Bark est prélevé à l'instant d'intensité maximale. Le fondamental ***f0*** et les amplitudes du spectre Bark normalisé par rapport à son maximum (arbitrairement fixé à 50), notées ***b1*** à ***b18***, sont retenus comme indices acoustiques. Les intensités utilisées dans la suite ne correspondent pas directement à des niveaux sonores absolus (SPL) mesurables dans les conditions de référence (à 1m, fenêtre de 125 ms, environnement anéchoïque). Mais elles sont comparables entre elles pour les deux bases de données, centrées sur une valeur moyenne arbitraire de 50 dBA.

3.2 Analyse discriminante

L'Analyse Discriminante (AD) considère des observations définies chacune par un ensemble de valeurs numériques et une catégorie d'appartenance. L'AD calcule un nouvel espace orthogonal de représentation des données, qui simultanément minimise la variance des observations appartenant à une même catégorie et maximise la variance entre catégories (pour une introduction à l'AD on pourra se reporter à l'aide en ligne de Praat). Le résultat est un classifieur (discriminant), dont la qualité peut être mesurée par le coefficient lambda de Wilks. L'utilisation du classifieur est évaluée par le taux d'observations mal classées, les données étant soit celles qui ont servi pour apprendre le classifieur (mode classification, ou autocohérence), soit de nouvelles données (mode prédiction). Dans la présente étude nous nous limiterons au mode classification.

4 Discrimination des voyelles

Les données soumises à l'AD sont, pour chaque token, la catégorie phonétique (l'une des voyelles françaises représentées dans le corpus) et les 19 indices acoustiques f0+18Bk.

4.1 Données CRC

La figure 1 présente les ellipses de dispersion à 1 écart-type dans le plan factoriel principal, le profil des deux premiers vecteurs propres, les valeurs propres attachées à chaque axe (en % d'explication de la variance) et en cumul, le coefficient lambda de Wilks et les taux d'erreur obtenus en classification. On ne peut pas juger de la classification à partir du seul plan principal lorsque les vecteurs propres d'ordre supérieur à 2 contribuent notablement à l'explication de la variance, ce qui est le cas ici.

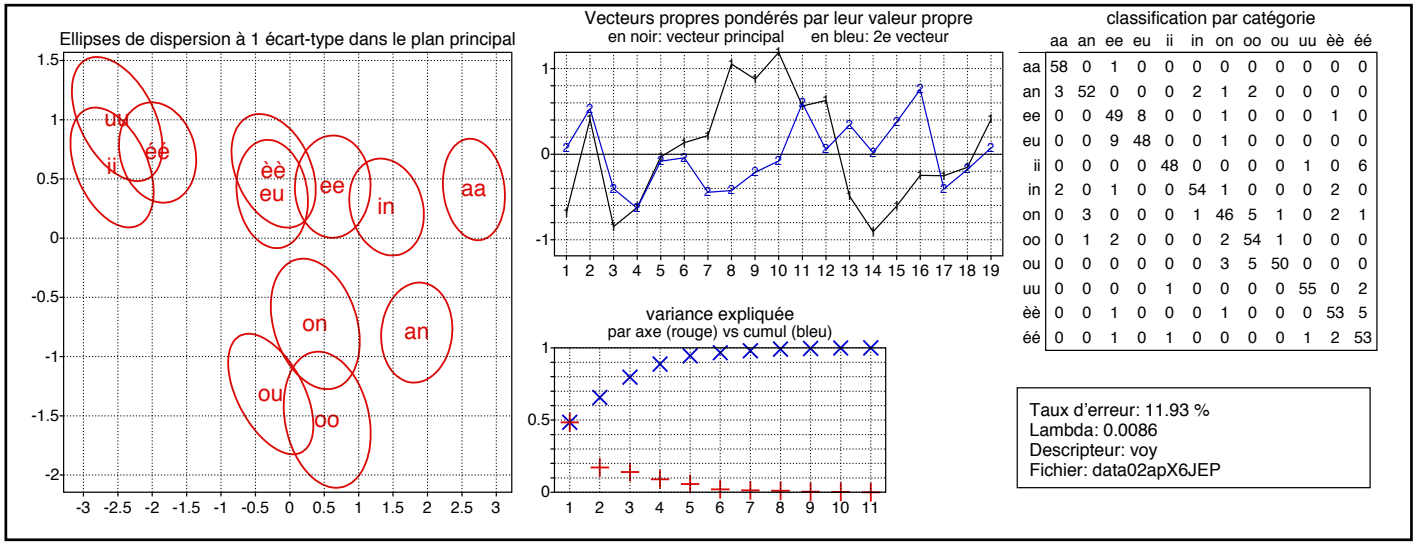


FIGURE 1 – Analyse Discriminante appliquée à la classification des 12 voyelles (9 orales et 3 nasales) du corpus CRC

Dans chaque groupe vocalique la variabilité est due à plusieurs causes: force de voix, locuteur, petites différences d'émission au cours de la session ou entre sessions. Malgré ces variations l'AD classifie les 12 voyelles de manière satisfaisante, avec un taux d'erreur de l'ordre de 12%. Les confusions constatées sont pour l'essentiel présentes dans les données elles-mêmes: un test perceptif (identification directe des voyelles par un groupe

d'auditeurs francophones) rapporté dans l'article (LIENARD and DI BENEDETTO, 1999) montrait un taux d'erreur de 9,3% sur les seules voyelles orales. Le test de classification ci-dessus donne sur celles-ci un taux d'erreur comparable (10,0%). Parmi les indices acoustiques dégagés par l'AD on observe que ***f0*** occupe une place secondaire.

4.2 Données JAE

Ces données ne comportent que 3 catégories vocaliques. Les résultats de l'AD sont présentés dans la figure 2.

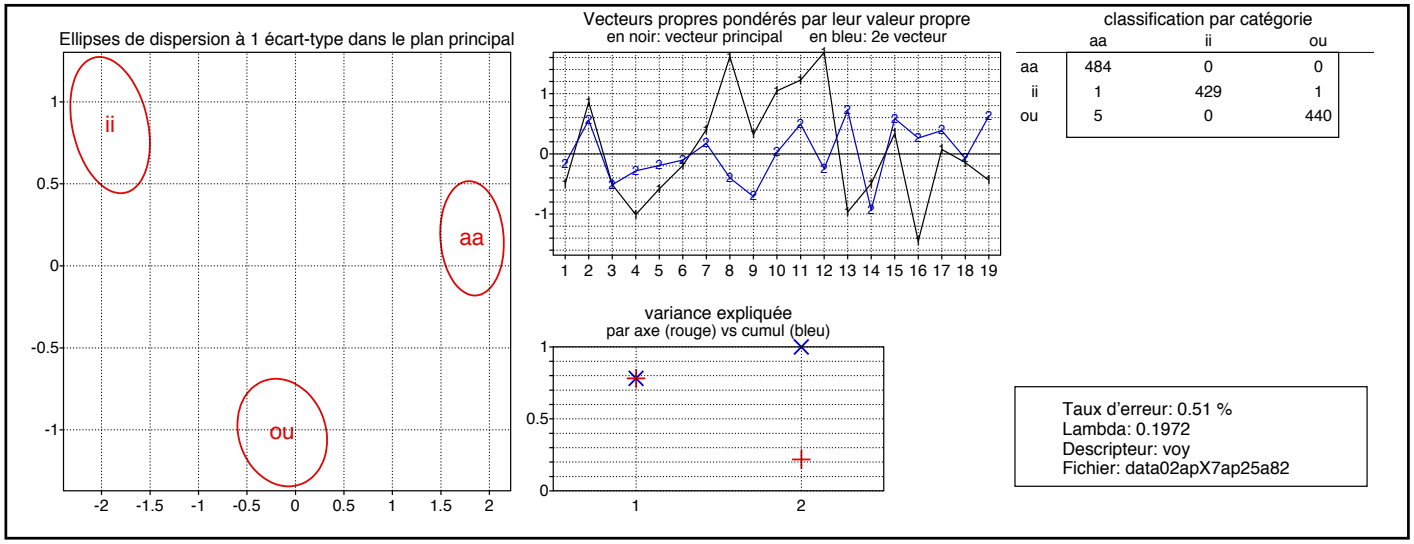


FIGURE 2 – Analyse Discriminante appliquée à la classification des 3 voyelles [a, i, u] du corpus JAE

Le taux d'erreur en classification est très faible, et identique à celui des données CRC réduites à ces mêmes voyelles cardinales. Lambda est nettement plus élevé, ce qui reflète une plus grande variance intragroupe par rapport à la variance intergroupe, due à la plus grande variabilité des données JAE selon la FDV et le locuteur.

4.3 Commentaire des résultats

Les essais menés sur les deux bases de données montrent que les voyelles sont correctement discriminées malgré les variations dues à la force de voix et au locuteur. Le nombre de dimensions nécessaires pour rendre compte d'au moins 95% de la variance est de 5 (sur 11) quand on considère 12 voyelles et de 2 (sur 2) quand on considère seulement les 3 voyelles cardinales.

5 Estimation de la force de voix

Nous utilisons le même outil et les mêmes données pour étudier la variabilité due à la force de voix. Celle-ci est une grandeur intrinsèquement continue, exprimée en dB, mais rien ne permet d'affirmer a priori qu'elle soit en relation linéaire avec une quelconque combinaison linéaire d'indices acoustiques. C'est pourquoi nous distinguons de manière quelque peu artificielle des paliers de FDV, espacés de 6 dB, qui seront considérés comme les centroïdes des catégories par rapport auxquelles l'AD regroupera les observations de manière optimale. Comme il y a continuité de la grandeur sous-jacente d'un palier à

l'autre, la valeur prise par lambda sera moins pertinente que dans le cas où les catégories sont par nature disjointes (comme dans le cas des voyelles). En revanche, lors de la phase de classification on pourra effectuer une interpolation entre centroïdes voisins pour retrouver une valeur continue. Ainsi, pour chaque token, l'écart entre la valeur initiale ***ap*** de la FDV et la valeur ***apc*** calculée par le classifieur et interpolée servira de base pour calculer une marge d'erreur statistique, qui constituera le véritable critère de réussite de l'analyse.

5.1 Données CRC

La figure 3 montre, de gauche à droite: i) le plan principal de l'analyse factorielle et les ellipses de dispersion à 1 écart-type, ii) la représentation des deux premiers vecteurs propres en fonction des 19 indices acoustiques retenus, et iii) la correspondance entre la valeur mesurée ***ap*** de la FDV (en abscisse) et la valeur ***apc*** (en ordonnée) estimée au moyen du classifieur établi à partir des 19 indices acoustiques (f0+18 Bk).

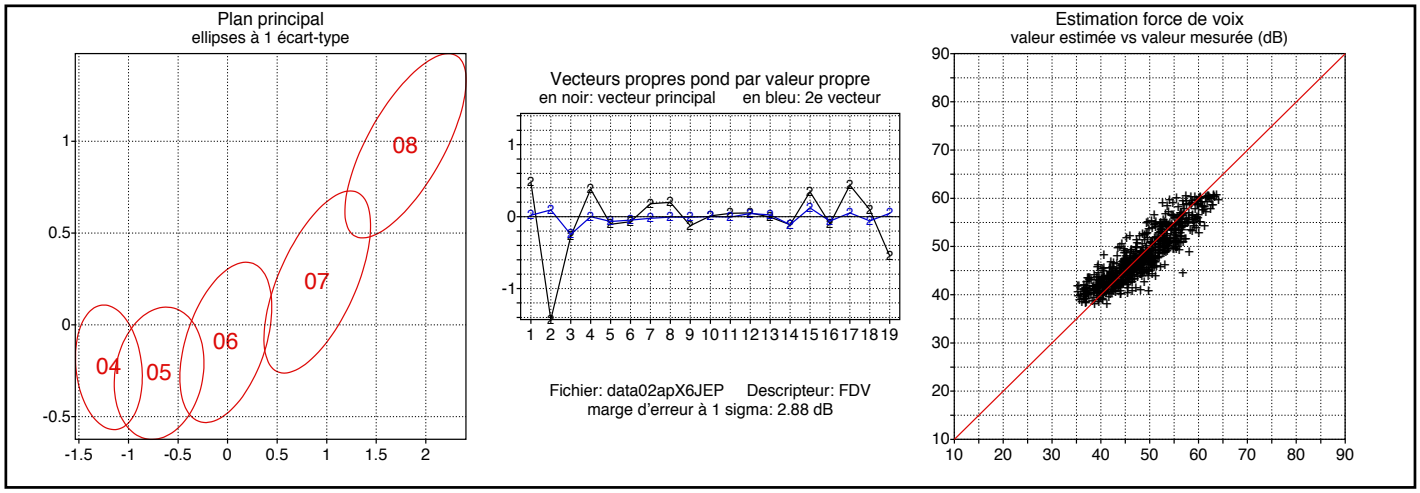


FIGURE 3 – Analyse Discriminante appliquée à l'estimation de la force de voix pour un sous-ensemble du corpus CRC (les classes incomplètes 03 et 09 ont été omises)

On constate une performance remarquable: la marge d'erreur (à ±1 écart-type de la valeur correcte) est de moins de 3 dB en classification, ce qui indique que l'information permettant de discerner 5 à 10 nuances de FDV dans la dynamique considérée (30 dB) est présente dans le spectre Bark du signal. L'axe 1, où intervient majoritairement l'indice 2 (***b1***) suffit presque à lui seul à estimer la FDV (explique 92.7% de la variance). On observe une forte non-linéarité dans le plan principal pour les faibles FDV, manifestée surtout sur l'axe 2 au niveau de ***b2***, ce qui justifie a posteriori le choix d'une technique de catégorisation par paliers. Le fondamental ***f0*** intervient peu dans les vecteurs propres, mais une partie de l'information qu'il porte se retrouve dans les bandes ***b1*** et ***b2***.

5.2 Données JAE

La moindre performance illustrée dans la figure 4 (marge d'erreur 6.2 dB), s'explique par la plus grande dynamique de ces données. Elle permet de discerner statistiquement 8 à 10 nuances de FDV. Les axes principaux sont semblables à ceux obtenus avec le corpus CRC, ainsi que la non-linéarité qui concerne surtout les voix faibles. Ces résultats confirment la pertinence des indices ***b1*** et ***b2*** pour l'estimation de la FDV.

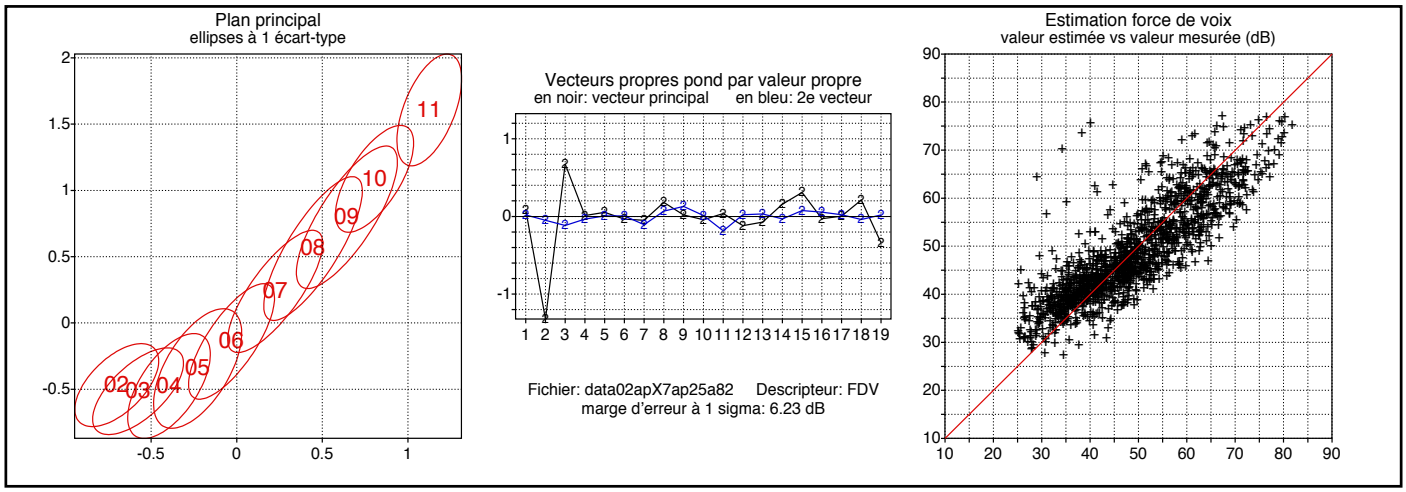


FIGURE 4 – Analyse Discriminante appliquée à l'estimation de la force de voix pour un sous-ensemble du corpus JAE (les classes incomplètes 01 et 12 ont été omises)

6 Interactions entre voyelle et FDV

Nous avons vu dans les sections précédentes que tant les voyelles que la force de voix pouvaient être estimées avec un faible taux d'erreur à partir des spectres Bark normalisés. Nous allons maintenant chercher s'il existe des interactions entre les deux descripteurs. Pour cela nous nous posons la double question: le fait de connaître la voyelle (par une première analyse ou par une information descendante) permet-il d'améliorer l'estimation de la force de voix ? Et, réciproquement, la connaissance de la FDV peut-elle être utilisée pour mieux discriminer les voyelles ?

6.1 Connaître la voyelle permet-il de mieux estimer la FDV ?

Pour tester cette proposition un classifieur de la FDV est calculé pour chaque catégorie vocalique du corpus, cad les 12 voyelles de CRC et les 3 de JAE, et dans un second temps ce classifieur est appliqué à ces mêmes tokens. Les résultats sont consignés dans la table 1, pour les deux bases de données. La colonne "voyelle inconnue" indique les valeurs sur l'ensemble des données, la classification étant effectuée sur toutes les voyelles. La colonne "voyelle connue" indique la moyenne des marges d'erreur trouvées pour chaque voyelle.

Marge d'erreur sur la FDV (dB)	voy inconnue	voy connue
CRC (12 voy, 720 tok, dyn 30 dB)	2,97	2,12
JAE (3 voy, 1395 tok, dyn 60 dB)	6,51	4,87

TABLE 1 – Amélioration de l'estimation de la FDV lorsque la voyelle est connue

Les chiffres précis diffèrent légèrement de ceux rapportés plus haut car les conditions de calcul ne sont pas exactement identiques. Mais le résultat est sans ambiguïté: dans les deux cas la connaissance préalable de la voyelle entraîne une amélioration notable de l'estimation de la FDV.

6.2 Connaître la FDV permet-il de mieux discriminer les voyelles ?

Pour tester cette proposition on ne peut guère prendre comme seul critère le taux d'erreur de classification, intrinsèque aux données dans le cas de CRC et très faible dans le cas de JAE. Il est donc accompagné entre parenthèses de la valeur de lambda (table 2). Dans le cas de CRC on a considéré les tokens produits selon la consigne p, n ou l, tout en sachant que les FDV moyennes dans ces trois tranches sont proches (resp. 62,7 pour p, 66,6 pour n et 74,1 dB pour l) et se recouvrent partiellement, ce qui explique les valeurs faibles de lambda. Dans le cas de JAE on a considéré 4 tranches de FDV espacées de 12 dB; le taux d'erreur affiché est la moyenne des taux obtenus pour chaque tranche.

Erreurs de classification voy % (lambda)	FDV inconnue	FDV connue
CRC (12 voy, moy consignes FDV: p,n,l)	12,5 (0,0089)	9,0 (0,0048)
JAE (3 voy, moy 4 tranches de FDV de 12 dB)	0,5 (0,2018)	0,3 (0,1311)

TABLE 2 – Amélioration de la discrimination des voyelles lorsque la FDV est connue

Tout en gardant à l'esprit les réserves formulées plus haut, on constate une amélioration sur les deux critères, pour les deux bases de données: la connaissance, même approchée, de la FDV permet d'améliorer la classification des voyelles.

7 Conclusion

Cette étude a mis en évidence le caractère quantifiable des variations spectrales dues à l'effort vocal, qui se distinguent ainsi d'un simple changement de timbre. Ce caractère systématique permet de retrouver la force de voix à partir du spectre dont a été retirée l'information directe de niveau sonore, et ceci avec une précision comprise entre 3 et 6 dB. On peut donc penser que les degrés d'effort vocal perceptibles d'emblée dans la voix conversationnelle sont plus nombreux que les 3 nuances habituellement mentionnées dans la littérature. Les résultats laissent entrevoir, à terme, la possibilité de retrouver dans tout enregistrement de parole l'intensité véritablement émise par le locuteur, information habituellement perdue dès la prise de son.

Sur un plan plus général, la démarche suivie s'inscrit dans la problématique des interactions entre voix et parole. Souvent considérés comme étrangers l'un à l'autre, ces deux aspects du signal oral apparaissent comme partiellement liés et complémentaires, au point que la connaissance même approximative de l'un peut faciliter l'estimation de l'autre.

Remerciements

Les auteurs remercient les instances scientifiques du LIMSI-CNRS, ainsi que leurs collègues Nicolas Audibert, Philippe Boula de Mareüil, Christophe d'Alessandro et Albert Rilliard qui ont apporté divers éclairages sur le problème de l'effort vocal.

Références

AUGUSTE-ETIENNE, J. (1999). "Etude d'un protocole d'enregistrement pour l'analyse du timbre de la voix", mémoire de recherche, ENS Louis Lumière, Noisy le grand.

BOERSMA, P. and WEENINK, D. (2012). "Praat: doing phonetics by computer" version 5.3.32, retrieved 17 October 2012 from http://www.praat.org/.

BRUNGART, D., SCOTT, K. and SIMPSON, B. (2001) "The influence of vocal effort on human speaker identification", ***Eurospeech***, Scandinavia.

D'ALESSANDRO, C. (2006). "Voice source parameters and prosodic analysis", in S. Sudhoff et al [Eds] ***Methods in Empirical Prosody Research***, 63-87, Walter de Gruyter.

DOVAL, B., D'ALESSANDRO, C. HENRICH, N. (2006). "The spectrum of glottal flow models", ***Acustica united with Acta Acustica***, 92:1026-1046, 2006.

FANT, G., LILJENCRANTS, J. and LIN, Q. (1985). "A four parameter model of glottal flow", STL-QPRS, 26(4):1-13, 1985.

GARNIER, M. (2007). "Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal", thèse de doctorat, université Paris VI.

HANSEN, J. and VARADARAJAN, V. (2009). "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition", IEEE tr. on ASLP, 17 (2), 366-378.

HANSON, H. (1997). "Glottal characteristics of female speakers: acoustic correlates", ***J. Acoust. Soc. Am.*** 101 (1), 466-481, 1997.

HENRICH, N., D'ALESSANDRO, C., DOVAL, B. and CASTELLENGO, M. (2005). "Glottal open quotient in singing: measurement and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency", ***J. Acoust. Soc. Am.*** 117 (3), 1417-1430.

HUBER, J.E., STATHOPOULOS, E.T., CURIONE, G.M., ASH T.A. and JOHNSON, K. (1999). "Formants of children, women, and men: the effects of vocal intensity variation", ***J. Acoust. Soc. Am.*** 106 (3), 1532-1542.

JUNQUA, J.-C. (1992). "The Lombard reflex and its role on human listeners and automatic speech recognizers", ***J. Acoust. Soc. Am.*** 93, 510-524.

LIENARD, J.S. and DI BENEDETTO, M.G. (1999). "Effect of vocal effort on spectral properties of vowels", ***J. Acoust. Soc. Am.*** 106 (1), 411-422.

LIENARD J.S. and BARRAS C. (2013). "Fine-grain voice strength estimation from vowel spectral cues", ***InterSpeech***, Lyon.

ROSTOLLAND, D. (1982). "Acoustic features of shouted voice", ***Acustica***, vol 50, 118-125.

TRAUNMULLER, H. and ERIKSSON, A. (2000). "Acoustic effects of variation in vocal effort by men, women and children", ***J. Acoust. Soc. Am.*** 107 (6), 3438-3451.

ZELINKA, P., SIGMUND, M. and SCHIMMEL, J. (2012). "Impact of vocal effort variability on automatic speech recognition", ***Speech Communication*** (54), 732-742.

Phonétisation automatique du dialecte tunisien

Abir Masmoudi[1,2], Mariem Ellouze Khemakhem[1], Yannick Estève[2] , Fethi Bougares[2], Sawssan Dabbar[1], Lamia Hadrich Belguith[1]

(1) ANLP group, MIRACL Lab., Université de Sfax, Tunisie
(2) LIUM, Université du Maine, France
masmoudiabir@gmail.com, Mariem.Ellouze@planet.tn, yannick.esteve@lium.univ-lemans.fr, fethi.bougares@lium.univ-lemans.fr, sawssendabbar@gmail.com, l.belguith@fsegs.rnu.tn

RÉSUMÉ

Un dictionnaire phonétique est un composant primordial d'un système de reconnaissance de la parole ou d'un système de synthèse de la parole. Notre travail cible la génération automatique d'un dictionnaire de prononciation du dialecte tunisien, en particulier dans le domaine du transport ferroviaire. Pour ce faire, nous avons créé deux outils de phonétisation de mots voyellés ou non voyellés en dialecte tunisien. La méthode proposée pour générer automatiquement des dictionnaires phonétiques est à base de règles et est présenté dans cet article. Cet article présente également les différentes étapes de création de notre propre corpus d'étude. Ensuite, il détaille les exceptions phonétiques et phonologiques du dialecte tunisien et donne des exemples de règles utilisées pour la construction des dictionnaires phonétiques.

ABSTRACT

A phonetic dictionary is an essential component of speech recognition system or speech synthesis system. Our work targets the generation of an automatic dictionary pronunciation of Tunisian Dialect, in particular in the field of railway transport. To do this, we created two tools of phonetic voweled and non voweled words in Tunisian Dialect. The proposed method to automatically generate phonetic dictionaries is based on rules and is presented in this article. This paper outlines the steps to create our own study corpus. Then it details the phonetic and phonological exceptions of Tunisian dialect and illustrates some rules used in the phonetic dictionary construction.

MOTS-CLÉS : Dialecte Tunisien, Phonétique, Phonologique, Dictionnaire phonétique, mot voyellé et non voyellé.
KEYWORDS: Tunisian Dialect, Phonetic, Phonologic, Phonetic Dictionary, voweled and non voweled words.

1 Introduction

Ce travail s'inscrit dans le cadre de la réalisation d'un système de reconnaissance automatique du dialecte tunisien utilisé dans le domaine du transport ferroviaire. Dans cet article, nous nous intéressons à l'un des composants clés des systèmes de reconnaissance de la parole, à savoir le dictionnaire phonétique. Le manque de ressources parlées et écrites est l'un des principaux problèmes rencontrés pour le traitement du dialecte tunisien. Ainsi, nous proposons dans cet article les efforts que nous avons déployés afin de créer nos propres ressources, ainsi que notre méthode de génération automatique d'un dictionnaire phonétique pour les mots voyellés et les mots non voyellés en dialecte tunisien.

Cet article comprend cinq sections. D'abord, nous présentons un aperçu de l'état de l'art. Puis, nous exposons quelques spécificités du dialecte tunisien. Ensuite, nous présentons les étapes de création de notre corpus d'étude. Nous détaillons aussi les exceptions phonologiques et phonétiques du dialecte tunisien. Dans la quatrième section de l'article, nous exposons le principe et les étapes de génération automatique des dictionnaires phonétiques. Enfin, nous donnons les résultats d'évaluation de nos outils de phonétisation.

2 Le Dialecte Tunisien

Le dialecte tunisien, généralement connu sous le nom de " Darija " ou " Tounsi ", est un sous-ensemble de dialectes arabes du groupe de l'Ouest. Il est utilisé dans la communication orale de la vie quotidienne des Tunisiens. On peut distinguer trois sous-dialectes suivant les appartenances sociales: les citadins, paysans/agriculteurs et bédouins. Ces sous-dialectes diffèrent considérablement les uns des autres. Ces différences affectent tous les niveaux de la langue, c'est-à-dire la prononciation, la phonologie, le vocabulaire, la morphologie et même la syntaxe. Malgré ces différences, le dialecte tunisien reste toujours compris par l'ensemble des Tunisiens. Parmi les différences phonologiques, nous pouvons noter que les Bédouins prononcent[1] la lettre " ق " /q/ comme " ڤ " /g/, mais la population urbaine prononce " ق " /q/. Au niveau du vocabulaire, les citadins utilisent le mot " يوقف " /i:u:qif/, mais les agriculteurs utilisent le mot "يحبس " /i:aħbis/. Ces deux mots ont le même sens : "arrêter".

Parmi les spécificités remarquables du dialecte tunisien, on note la présence de mots empruntés au français, au berbère, à l'italien, au turc et à l'espagnol. Aussi, on note la présence de mots dont l'origine est l'arabe standard moderne(MSA). La présence de ces mots est le résultat de nombreux facteurs et événements historiques qui ont eu lieu tout au long des siècles, tels que : les invasions islamiques, la colonisation française, les migrations, les échanges commerciaux, etc.

Aujourd'hui, le dialecte tunisien est de plus en plus souvent utilisé lors d'interviews, de conversations téléphoniques, dans les services publics, etc. et il est très présent aujourd'hui dans les blogs, les forums et les commentaires d'utilisateur/lecteur sur internet.

Il est donc important de considérer le dialecte tunisien dans le contexte des nouvelles technologies telles que les systèmes de reconnaissance de la parole, les systèmes de dialogue Humain-Machine, etc. Cependant, le dialecte souffre de l'absence d'outils et de ressources

[1] Pour représenter les phonèmes, nous utilisons les symboles de l'alphabet phonétique international (http://fr.wiktionary.org/wiki/Annexe:Prononciation/arabe)

linguistiques. Les outils pour traiter le MSA sont très difficilement utilisables, compte tenu de la grande différence entre le MSA et le dialecte tunisien. Ainsi, il convient de noter que notre travail étudie le tunisien alors que nous manquons d'outils linguistiques adaptés à ce dialecte.

3 Etat de l'art

Dans un contexte de reconnaissance automatique de la parole, pour faire le lien entre le niveau lexical et le niveau acoustique, il est nécessaire d'associer à chaque mot du vocabulaire une ou plusieurs séquences d'unités acoustiques de base. Dans la littérature plusieurs approches sont utilisées pour obtenir ces séquences de phonèmes. Nous pouvons distinguer l'approche guidée par les données qui utilise généralement des données d'apprentissage et l'approche à base de règles qui nécessite une expertise linguistique (Béchet, 2001).Dans ce qui suit, nous allons expliquer davantage chacune de ces approches.

3.1 Approche guidée par les données

L'idée générale de cette approche est d'utiliser un dictionnaire de mots phonétisés manuellement. Des approches probabilistes basées sur des modèles joints (Bisani et Ney, 2008) ou encore sur des approches s'appuyant sur des outils de traduction automatique (Laurent et al. 2009), permettent de capturer par apprentissage automatique des liens entre les graphèmes utilisés pour écrire le mots et les phonèmes, unités de sons utilisées pour représenter la prononciation d'un mot.

Ces techniques statistiques nécessitent de posséder des données d'apprentissage. Lorsque ces données sont en nombre suffisant, elles permettent d'obtenir des résultats intéressants, tout en minimisant l'expertise humaine. En l'absence de données, ces approches ne sont pas applicables.

3.2 Approche à base de règles

La construction d'un système de phonétisation automatique à base de règles nécessitent une bonne connaissance de la langue et de ses règles de phonétisation qui, par ailleurs, ne doivent pas contenir trop d'exceptions : la phonétisation automatique utilisant des connaissances linguistiques ne nécessite pas de techniques innovantes, mais plutôt une expertise importante sur la langue.

L'avantage de cette approche est qu'elle permet de mieux contrôler la qualité de la construction des dictionnaires de prononciation : en cas d'erreur il est possible d'ajouter une nouvelle règle. De plus, le coût de développement peut être comparé au coût nécessaire à la construction manuelle du corpus d'apprentissage utilisé par l'approche statistique.

4 Corpus d'étude

Outre le manque d'outils pour traiter le dialecte tunisien, le manque de ressources linguistiques est également patent. Pour construire un système de reconnaissance automatique de la parole, nous avons besoin, a minima, d'un corpus constitué

d'enregistrements audio et de leur transcription. Nous avons été dans l'obligation de créer notre propre corpus d'étude.

La création de corpus repose sur trois phases : 1) la production d'enregistrements audio, 2) la transcription manuelle de ces enregistrements et 3) la normalisation de ces transcriptions (Masmoudi et al.,2014).

Le Tableau 1présente quelques statistiques de notre corpus.

Nombre d'heures	Nombre de dialogues	Nombre d'énoncés	Nombre de mots
20 h	4662	18657	71684

TABLE 1 - Statistiques de notre corpus d'étude

Lors de l'analyse de notre corpus, nous avons remarqué que les mots étrangers représentent 20% du vocabulaire. Voici quelques exemples d'emprunts étrangers en Dialecte Tunisien : le mot "تكاي " "Ticket " /tika:i:/ d'origine française, ou le mot "ترينو " "Trinou "/tri:nu:/ d'origine italienne. On peut noter aussi la présence des mots d'origine anglaise comme le mot "أوكاي " "Ok" /ʔʊ:kɑ:i:/.

Aussi, nous avons remarqué que certains mots étrangers s'utilisent avec des modifications par rapport à leurs origines. Ainsi, un mot étranger subit un ajout d'enclitiques ou de proclitiques de la langue arabe. Voici quelques exemples de mots :

- Le mot "سيپرميها " /sIprimIha:/ " la supprimer " est un mot emprunté du français, il subit un ajout de l'enclitique arabe "ها" /ha:/ "la" qui est attaché au mot.
- Le mot "ريزرڤيلي " /rIzirvIlI/ "réserver moi" est un mot emprunté du français, il subit un ajout de l'enclitique arabe "لي " /lI/ "moi" qui est attaché au mot.

5 Phonologie et phonétique du dialecte tunisien

Dans cette section, nous présentons d'abord le système d'écriture du dialecte tunisien et ensuite les exceptions phonétiques et phonologiques sa phonologie.

5.1 Le système d'écriture

Comme le MSA, le dialecte tunisien s'écrit de droite à gauche. L'alphabet se compose de trente et une lettres. Vingt-cinq d'entre elles sont des consonnes provenant de MSA, trois consonnes sont le résultat de la présence de mots étrangers : 'ڤ' "v", 'ق' "g" et 'پ' "p" et trois lettres représentent les voyelles longues.

Chaque lettre possède au maximum quatre formes différentes, selon qu'elle apparaît au début, au milieu ou à la fin d'un mot, ou de manière isolée.

Les lettres sont la plupart du temps reliées les unes aux autres graphiquement et il n'y a pas de capitalisation. Une caractéristique distinctive du système d'écriture de l'arabe et ses dialectes est la présence de voyelles courtes appelées signes diacritiques. Ces signes diacritiques ne sont pas représentés par les lettres de l'alphabet, mais marqués par des traits courts placés au-dessus ou au-dessous de la consonne. On peut distinguer neuf signes

diacritiques: (a) trois signes diacritiques sont des voyelles courtes ; (b) trois signes diacritiques "de nunation" représentant une combinaison d'une voyelle courte et le marqueur / n / (ce type de signe diacritique est rarement utilisé dans le Tunisien), (c) un signe diacritique représente le doublement d'une consonne (appelé Shadda); (d) un signe diacritique pour marquer l'absence de voyelles courtes (appelé Sukun).

5.2 Quelques exceptions phonétiques et phonologiques

À partir de notre corpus d'étude, nous avons pu mettre en évidence qu'il existe plusieurs exceptions phonétiques et phonologiques dans le dialecte tunisien. Nous citons ci-dessous quelques-uns de ces traits phonétiques (Masmoudi et al.,2014) :

• Les voyelles courtes sont souvent omises. Par exemple, le verbe «écrire» est prononcé dans le MSA «كتب» /kataba/ mais en dialecte tunisien, nous disons /ktib/.

• L'intervention de la consonne «ش»/ʃ/ dans les régions rurales et dans les zones urbaines. Il semble en particulier dans la voix interrogative du dialecte.

• Il y a beaucoup de consonnes dans le dialecte tunisien qui peuvent être prononcées de différentes manières. Ci-dessous, nous présentons certaines d'entre elles :

- La consonne "ش" /ʃ/ peut être prononcée comme "ش" /ʃ/ ou "س" /s/.
- La consonne "ث" /θ/ peut être prononcée de deux manières : "ث" /θ/ ou "ف" /f/.
- La consonne "غ" /ɣ/ peut être prononcée de deux manières: "غ" /ɣ/ ou "خ" /x/.
- La consonne "ط" /tˤ/ peut être prononcée de deux manières : "ط" /tˤ/ ou "ت" /t/
- La consonne "أ" /ʔ/ au sein d'un mot peut être prononcé "أ" /ʔ/ ou "ه" /h/ .

• Le "Hamza" "أ " /ʔ/, accepte plusieurs variations de prononciations selon la position de cette lettre.

• "Ta-Marbouta" à la fin du mot est habituellement silencieux, mais il y a quelques exceptions.

• Il y a des lettres qui ne sont pas prises en compte. Le "alif" dans le mot "خرجوا" /xardʒuː/ "ils ont sorti" ne correspond pas à un son (silence).

• Les voyelles longues deviennent des voyelles courtes. Par exemple, "فِي التران"/fItran/ "dans le train" "فِ التران" / fi tran/.

• Nous avons remarqué l'élimination d'une consonne dans certains mots. Par exemple, " مانعرفش " /maːnaʕˤrafʃ/ "Je ne sais pas" peut être prononcé "مانعرش "/maːnaʕˤraʃ/. Dans cet exemple, la consonne "ف " /f/ est éliminée.

• Les nombres ont une caractéristique spécifique : les nombres entre "trois" et "neuf", acceptent une double prononciation dont l'une subit une élimination de quelques consonnes et une modification de la voyellation; les nombres à partir de "onze" acceptent aussi deux prononciations, dont l'une nécessite un ajout de la phonème "N" à la fin du nombre.

• Nous avons noté l'ajout de la phonème /E IH/ pour appuyer la prononciation de la première consonne muette d'un mot.

6 Construction automatique d'un dictionnaire phonétique

Nous avons créé deux outils de phonétisation : un pour les mots voyellés, et l'autre pour les mots non voyellés. Dans les sections suivantes, nous expliquons davantage le principe général de la phonétisation automatique puis nous détaillons les spécificités des deux outils de phonétisation.

6.1 Principe général de la phonétisation

Pour générer automatiquement un dictionnaire de prononciation, nous avons illustré, à partir de nos corpus d'étude, un ensemble de règles phonétiques et un lexique contenant des exceptions. Le dialecte tunisien se caractérise par la présence de nombreuses exceptions phonétiques. En effet, on peut trouver un mot qui peut être prononcé de deux ou plusieurs façons.

Le processus de phonétisation pour les mots voyellés ou non voyellés s'effectue en deux phases : la consultation de la base du lexique d'exceptions et l'application des règles de phonétisation.

6.1.1 Lexique d'exceptions

Il y a des mots qui ne peuvent pas suivre notre ensemble de règles phonétiques, il est donc nécessaire de définir un lexique des exceptions. Ce lexique est consulté avant que les règles soient appliquées. Si le mot est parmi les exceptions, il est encodé directement sous forme phonétique. Sinon, nous devons appliquer les règles pour générer sa forme phonétique. Dans notre lexique, nous avons plus de 30 exceptions. Notre base de lexique des exceptions a été validée par trois experts (de langue maternelle). Voici quelques exemples d'exceptions :

• Le mot "نصف" /nisˁf/ peut être prononcé de trois manières différentes: "نصف" /nisˁf/ , "نص" /nisˁ/ ou "نفص" /nifsˁ/.

• Le mot "هذا " /haða:/est prononcé comme ça "هاذا "/ha:ða:/.

6.1.2 Règles phonétiques

Nous avons développé un ensemble de règles phonétiques qui doivent être fournies pour chaque lettre. Chaque règle essaie de faire correspondre certaines conditions relatives au contexte de la lettre et de fournir un phonème ou séquences de phonèmes et parfois le silence . Nos règles ont également été validées par trois experts.

Chaque règle se lit de gauche à droite et suit ce format détaillé dans (Masmoudi et al.,2014):

{ Condition droite} + {Graphème} + { Condition gauche} ==> Phonétisation

Le nombre total de règles est d'environ quatre-vingts. Pour faciliter l'utilisation de ces règles et en se basant sur les exceptions phonétiques du dialecte tunisien, nous avons décidé de diviser les règles en cinq bases : une base pour les nombres, une base pour les mots qui commencent par la lettre "أ " /E/, une base pour les mots étrangers qui contiennent la lettre "R", une base pour les mots qui commencent par "Alif et Lam" et une base pour le reste des règles de phonétisation.

6.2 Phonétisation de mots voyellés

Lors de la transcription de notre corpus, nous avons choisi de voyeller les mots selon la prononciation du locuteur. La présence de ces voyelles permet une diminution substantielle du degré de l'ambiguïté phonétique.

Les mots voyellés en dialecte tunisien se terminent en général par un Sukun (consonne muette) ou bien par une Voyelle longue.

Pour générer automatiquement la forme phonétique d'un mot, l'application de règles phonétiques est faite dans le sens de lecture du mot, c'est à dire on commence par la premier lettre du mot et on respecte l'ordre des lettres.

Ci-dessous un exemple de phonétisation d'un mot voyellé :

Le mot «حْدَاشْ » /ħda: ʃ/ "Onze" appartient à la base des nombres, alors ce mot accepte une double prononciation dont l'une se termine par le phonème « N ».

Les règles utilisées sont les suivantes :

R1 : {C= ح /ħ/} + { Sukun } => {Phonème = HH }

Lorsque une consonne est suivie par "Sukun", alors on obtient toujours le phonème de la consonne.

R2 : {!= n'est pas pris en considération} + {C = د /d/ } + {VL = Fatha + alif} =>

{Phonème = D AE:}

Lorsque une consonne est suivie par une voyelle longue "Fatha et alif", alors on obtient les deux phonèmes : dont l'un correspondant du consonne "D" et l'autre du voyelle longue "AE:".

R3 : {!} + {ش/ʃ/ }+ { Sukun } => {Phonème = SH}

Lorsque une consonne est suivie par "Sukun", alors on obtient toujours le phonème du consonne.

La phonétique du mot est la suivante : «حْدَاشْ» /ħda: ʃ/ "Onze" 1. HH D AE: SH
2. HH D AE: SH N

6.3 Phonétisation de mots non voyellés

En partant de trois principes : (a) un mot en dialecte tunisien se termine soit par une consonne muette (avec Sukun) ou par une voyelle longue, (b) chaque voyelle longue est toujours suivie et précédée par une consonne muette et (c) on ne peut pas avoir deux consonnes successives qui portent une voyelle (longue ou courte) à l'exception des mots avec "Shadda" (dédoublement de consonnes), les règles de phonétisation pour les mots non voyellés peuvent être divisées en deux groupes : les règles d'appui et les règles secondaires.

Les règles d'appui sont appliquées dans un premier temps. La majorité de ces règles sont à l'origine de la production de phonèmes de voyelles longues dans le mot, ce qui facilite l'application des règles secondaires dans un deuxième temps. Les règles secondaires sont à

l'origine de la production de phonèmes de voyelles courtes ou "Sukun". Contrairement au mots voyellés, l'application de règles phonétiques ne respecte pas un ordre bien défini et donc ne suit pas le sens de lecture du mot.

En effet, la phonétisation commence par localiser dans le mot au moins l'une de ces quatre consonnes ("ا" /a:/ "Alif ", "ى" /a:/ "Alif Maksoura", "ي" /i:/ "Ya" et "وا" /u:/ "Waw et Alif") pour appliquer les règles d'appui. Ces règles permettent de donner la phonétisation de voyelles longues. Ensuite et en se basant sur le principe que "chaque voyelle longue est toujours suivie et précédée par une consonne muette", les règles secondaires sont appliquées afin d'ajouter les phonèmes des voyelles courtes ou "Sukun".

En absence de l'un de ces quatre consonnes, la phonétisation se repose sur deux principes : le mot se termine par une consonne muette (avec Sukun) et en deuxième lieu, on ne peut pas avoir deux consonnes successives qui portent une voyelle.

7 Evaluation

Le système de phonétisation a été testé sur un corpus de test constitué de 400 mots collectés à partir des blogs tunisiens de domaine variés (politique, sportifs, culture …). Le corpus a été dans un premier temps, normalisé selon la convention CODA (Zribi et al. 2014). Dans l'objectif d'évaluer les deux systèmes de phonétisation (avec et sans voyelles), nous avons créé une version voyellée du corpus de test (voyellation manuelle). Le tableau suivant montre les résultats obtenus pour la phonétisation du corpus de test avec et sans voyelles.

Corpus de test	Taux d'erreur phonème
Voyellé	0 %
Non voyellé	12.9 %

TABLE 2 - Résultat de l'évaluation

Comme présenté dans la Table 2, le système de phonétisation de dialecte tunisien est 100% performant pour les textes voyellés. En ce qui concerne le texte non voyellé la phonétisation est une tâche plus complexe où notre système peut proposer plusieurs variantes de phonétisation. Le taux d'erreur phonème de 12,9% présenté dans la table 2 est celui de la première proposition du système.

8 Conclusion

Ce papier décrit nos efforts pour créer deux outils de phonétisation automatique des mots voyellés et non voyellés en dialecte tunisien dans le domaine de transport ferroviaire. Les outils génèrent un dictionnaire de prononciation en se basant sur des règles phonétiques du tunisien. Chaque règle essaie de faire correspondre certaines conditions relatives au contexte de la lettre et de fournir un remplaçant. Le nombre total de règles est environ 80.

Pour faire face à l'absence des ressources linguistiques en dialecte tunisien, nous étions amenés à créer notre propre corpus d'étude. La création de corpus repose sur trois phases à savoir, la production d'enregistrements audio, la transcription manuelle de ces enregistrements et la normalisation de ces transcriptions.

Références

Afify, M., Nguyen, L., Xiang, B., Abdou, S. et Makhoul.J., (2005), Arabic broadcast news transcription using a one million word vocalized vocabulary. *In Proceedings of Interspeech 2005*, Portugal.

Algamdi, M. (2000), Arabic Phonetics. Riyadh, Saudi Arabia.

Algamdi, M., (2003), KACST Arabic Phonetics Database. Fifteenth International Congress of Phonetics Science, (KACST'2003), Barcelona.

Algamdi, M., Elshafei, M., et Almuhtasib, H., (2002),Speech Units for Arabic Text-to-speech. *Fourth Workshop on Computer and Information Sciences*.

Antoine Laurent, Paul Deléglise, Sylvain Meignier: Grapheme to phoneme conversion using an SMT system. Interspeech 2009: 708-711

Baccouche , T., (1994), L'emprunt en arabe moderne, Beit Elhikmaet IBLV, Tunis.

Biadsy, F., Hirschberg, J. et Habash. N., (2009), Spoken Arabic Dialect Identification Using Phonotactic Modeling. *In Proceedings of EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece.

Bisani M. and Ney. H. (2008) "Joint-Sequence Models for Grapheme-to-Phoneme Conversion". Speech Communication, Volume 50, Issue 5, , Pages 434-451

Diehl, F., Gales, M. J. F., Tomalin, M., et Woodland, P. C., (2008), Phonetic pronunciations for Arabic speech-to-text systems. IEEE International Conference on Acoustics, Speech and Signal Processing.

El-Imam.Y., (2004).Phonetization of Arabic: rules and algorithms. In Computer Speech and Language.

Frédéric Béchet, LIA_PHON (2001) ,: un système complet de phonétisation de textes, Traitement Automatique des Langues - TAL - volume 42 numéro 1 - pp 47-67.

Gales, M. J. F., Diehl, F., Raut, C. K., Tomalin, M., Woodland, P. C., et Yu, K., (2007), Development of a phonetic system for large vocabulary Arabic speech recognition. IEEE Workshop on Automatic Speech Recognition & Understanding.

Habash, N., Diab, M., Rambow, O., (2012), Conventional Orthography for Dialectal Arabic. In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Masmoudi, A., Ellouze Khmekhem, M., Estève, Y., Hadrich Belguith, L., et Habash, N.,(2014), A corpus and phonetic dictionary for Tunisian Arabic speech recognition, In *19th edition of the Language Resources and Evaluation Conference (LREC'2014)*, Reykjavik, Iceland.

Masmoudi, A., Estève, Y., Ellouze Khmekhem, M, Estève, Y., Bougares, F et Hadrich Belguith, L., (2014), Phonetic tool for the Tunisian Arabic, *The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14),* Saint Petersburg, Russia.

Zribi, I., Boujelben, R., Masmoudi, A., Ellouze Khmekhem M., Hadrich Belguith, L., et Habash, N.,(2014), A Conventional Orthography for Tunisian Arabic, *In 19th edition of the Language Resources and Evaluation Conference (LREC'2014*), Reykjavik, Iceland.

Correction interactive de transcriptions de parole par fusion de phrases

Mickael Rouvier, Benoit Favre, Frédéric Béchet *
Aix-Marseille Université, CNRS, LIF UMR 7279, 13000, Marseille, France
prenom.nom@lif.univ-mrs.fr

RÉSUMÉ

La clarification sous forme de dialogue permet d'aider à corriger les erreurs de RAP dans un système de traduction de parole automatique ainsi que dans d'autres applications interactives. Nous proposons d'utiliser des variantes de Levenshtein pour fusionner une phrase contenant une erreur et une phrase de clarification. Les erreurs de RAP qui pourraient nuire à l'alignement sont traitées par correspondance phonétique et une distance de plongement de mots est utilisée pour prendre en compte les synonymes en dehors des segments d'erreurs. Ces paramètres permettent une amélioration relative de 30 % du taux d'erreur de mots sur une sortie de RAP comparé l'absence de clarification. De plus, nous générons un ensemble de fusions potentielles et entraînons un réseau de neurones afin de sélectionner la meilleure fusion, permettant de sélectionner correctement 24 % de plus d'instance. Le système est utilisé dans le cadre du projet BOLT.

ABSTRACT

Interactive correction of speech transcripts

Clarification dialogs can help address ASR errors in speech-to-speech translation systems and other interactive applications. We propose to use variants of Levenshtein alignment for merging an errorful utterance with a targeted rephrase of an error segment. ASR errors that might harm the alignment are addressed through phonetic matching, and a word embedding distance is used to account for the use of synonyms outside targeted segments. These features lead to a relative improvement of 30 % of word error rate on ASR output compared to not performing the clarification. Twice as many utterance are completely corrected compared to using basic word alignment. Furthermore, we generate a set of potential merges and train a neural network on crowd-sourced rephrases in order to select the best merger, leading to 24 % more instances completely corrected. The system is deployed in the framework of the BOLT project.

MOTS-CLÉS : Correction d'erreur, Systèmes de dialogue, Détection d'erreurs, Transcription de la parole, Reranking, Fusion de phrases.

KEYWORDS: Error correction, Dialog systems, ASR error detection, Reranking, Levenshtein alignment.

Les systèmes de transcription automatique de la parole génèrent des transcriptions imparfaites à cause des conditions acoustiques, des mots hors vocabulaire ou des ambiguïtés venant d'un contexte lointain. Malgré les progrès récents sur la robustesse (Maas et al, 2012; Weninger et al, 2012), la reconnaissance en vocabulaire ouvert (Gerosa et Federico, 2009; Parada et al, 2011), la modélisation du contexte à long terme (Jonson, 2006; Mitchell et Lapata, 2009; Liu *et al.*, 2012), les systèmes de transcriptions font toujours des erreurs ayant un impact sur les tâches en aval.

Les systèmes interactifs offrent l'opportunité de corriger ces erreurs à l'aide de dialogues de clarifi-

∗. Ce travail est partiellement financé par le projet DARPA HR0011-12-C-0016 sous la forme d'un contrat entre AMU et SRI International.

cation, comme dans les systèmes de dialogues en domaine restreint qui utilisent des confirmations implicites ou explicites (Shin et al, 2002; López-Cózar *et al.*, 2010). Néanmoins, ce n'est pas possible en domaine ouvert car il faudrait une bien meilleure modélisation sémantique du contenu des messages à transcrire.

Les efforts récents d'estimation de mesures de confiance (Yu et Deng, 2010; Seigel *et al.*, 2011), de détection de mots hors vocabulaire (Marin et al, 2012; Kombrink *et al.*, 2012; Parada *et al.*, 2010), de détection et de caractérisation automatique des erreurs (Bechet et Favre, 2013; Dufour *et al.*, 2012) offrent l'opportunité de localiser les segments d'erreurs et d'initier un dialogue de clarification de manière à améliorer la transcription. Ces systèmes de clarification peuvent demander à l'utilisateur de désambiguïser des homophones, épeler des mots hors vocabulaire ou reformuler une partie de leur phrase d'origine afin de corriger les erreurs (Prasad et al, 2012; Stoyanchev *et al.*, 2012). Au lieu de demander une reformulation complète, ces stratégies ciblent les erreurs à corriger de la même manière que le font les humains, générant des interactions naturelles et intuitives. En particulier, il est possible d'obtenir une transcription parfaite à l'aide de plusieurs questions de clarification.

Dans cet article, nous nous intéressons à la fusion d'une phrase où une erreur a été détectée et de la réponse de l'utilisateur à une question de clarification ciblant cette erreur. Nos contributions sont les suivantes :

— nous proposons une modification de l'alignement de Levenshtein en termes de topologie et de fonction de coût (phonétique, sémantique) pour prendre en compte les difficultés de la tâche de fusion ;
— nous générons plusieurs variantes de fusion dont la meilleure est sélectionnée par un perceptron multicouches tirant parti de caractéristiques riches ;
— le système est évalué sur un jeu de dialogues de clarification collecté dans le cadre du projet BOLT de traduction simultanée automatique.

1 Travaux connexes

Améliorer les transcriptions automatiques à l'aide d'interactions est principalement traité par des confirmations dans les systèmes de dialogue et des commandes d'édition dans les systèmes multimodaux comme la dictée vocale.

Il existe de nombreux travaux sur la confirmation dans les systèmes de dialogue (Shin et al, 2002). Dans sa plus simple forme, le résultat de la transcription automatique est resynthétisé et il est demandé à l'utilisateur si ce dernier est correct ou non. Les systèmes de compréhension à base de concepts reposent sur des scores de confiance sémantique pour ne confirmer que les concepts mal compris (Jung *et al.*, 2008). Si la modélisation sémantique est plus profonde, il est possible d'évaluer la cohérence des croyances du système pour en déduire des questions ciblées (Bohus et Rudnicky, 2009). Ces approches n'étant pas envisageables en domaine ouvert, certains travaux tirent parti des spécificités de la tâche pour cibler les confirmations, comme ceux de (Misu *et al.*, 2004) pour la recherche d'information.

Lorsque des interactions multimodales sont possibles, on peut tirer parti des autres modalités pour améliorer la transcription. Par exemple les systèmes de dictée vocale comme *Dragon Naturally Speaking* permettent de voir les mots de la transcription et de les corriger par des commandes spécifiques. De plus, des candidats pour la correction peuvent être générés à partir des réseaux de confusion pour faciliter leur sélection (Ogata et Goto, 2005; Huggins-Daines et Rudnicky, 2008; Laurent et al, 2011). Dans (Hoste *et al.*, 2012), une estimation des mots regardés par l'utilisateur permet de sélectionner ceux qui sont incorrects.

Dans la communauté du traitement automatique du langage naturel, la fusion de phrase a été explorée dans le cadre du résumé automatique, pour créer une version plus courte d'un ensemble de phrases se chevauchant en termes de contenu. La tâche est effectuée par fusion d'arbres syntaxiques (Filippova et Strube, 2008; Barzilay et McKeown, 2005) ou à l'aide d'une distance d'édition apprise automatiquement sur des corrections manuelles (Elsner et Santhanam, 2011).

2 Tâche

Nos travaux s'intéressent à la tâche de transcription automatique pour la traduction simultanée. Cette tâche induit les points suivants : le domaine est ouvert, les interactions doivent être orales pour permettre à l'utilisateur de garder un contact visuel, les interactions doivent être intuitives. Formellement, étant donné la transcription erronée d'une phrase d'**origine** et de la **réponse** à une question de clarification relative à un **segment d'erreur** de l'original, la tâche consiste à construire une nouvelle transcription capturant au mieux l'intention du locuteur. Les mots à l'intérieur du segment d'erreur sont considérés comme erronés et doivent être remplacés ; toutefois les mots hors du segment d'erreur doivent parfois aussi être modifiés par exemple pour satisfaire un nouveau contexte syntaxique. Voici un exemple d'interaction :

Parole d'origine	Où est la chambre **hyperbare** ?
ASR	où est *l'a* chambre **hyper barre**
Erreur détectée	hyper barre
Question	Pouvez-vous reformuler AUDIO(hyperbare) ?
Réponse	la chambre de décompression
Phrase générée	où est la chambre de décompression

Dans cet exemple, les entrées du système de fusion sont : la phrase originale où l'erreur détectée est masquée *"où est l'a chambre* $\langle err \rangle$" et la réponse *"la chambre de décompression*". La fusion de référence est *"où est la chambre de décompression*". Notez qu'ici une erreur supplémentaire sur *"l'a*" n'a pas été détectée par le système et doit donc être corrigée lors de la fusion.

Nous avons étudié le comportement des utilisateurs face aux dialogues de clarification ciblés du système de traduction simultanée créé dans le contexte du projet BOLT (Bechet et Favre, 2013). Les comportements les plus courants sont :

— la réponse remplace exactement le segment d'erreur ;
— la réponse contient des mots supplémentaires qui contextualisent l'opération d'édition ;
— la réponse est une reformulation complète de la phrase ;
— le contexte syntaxique de l'original doit être modifié pour y insérer la réponse (relative au lieu d'adjectif, par exemple) ;
— des mots du contexte sont reformulés par concision (anaphore pronominale, par exemple) ;
— des expressions permettent d'introduire la reformulation ("j'ai dit ...").

Les sections suivantes proposent un système pour effectuer la tâche de fusion étant donné ces comportements.

3 Système

L'idée générale de notre système est d'aligner les mots de la réponse à ceux de l'original et de ne remplacer dans l'original que ceux qui sont alignés avec le segment d'erreur. Pour traiter les

différents comportements de l'utilisateur, nous créons des variantes de systèmes fondées sur cette stratégie puis nous utilisons un classifieur pour sélectionner la meilleure variante comme fusion finale.

Les deux premiers systèmes triviaux sont le remplacement de l'original par la réponse (ciblant une reformulation complète) et l'insertion de la réponse à la place du segment d'erreur. Les autres variantes sont basées sur l'alignement de Levenenshtein appliqué dans le cadre d'automates transducteurs pondérés. Dans ce cadre, l'original et la réponse sont représentés par des accepteurs linéaires reconnaissant leur séquence de mots. Dans la phrase originale, le segment d'erreur est remplacé par une boucle sur un token spécial $\langle err \rangle$ de manière à pouvoir y aligner plusieurs mots de la réponse. En introduisant un troisième automate effectuant toutes les opérations d'édition possibles (insertion, suppression, substitution), il est possible par composition des trois automates d'obtenir l'alignement de plus faible coût entre l'original et la réponse. Soit A_o, l'original, A_c, la réponse et T_e le transducteur d'édition, le plus court chemin dans la composition $A_o \circ T_e \circ A_c$ donne l'alignement. La figure 1 détaille la structure de ces automates.

La fonction de coût d'édition est structurée de manière à regrouper les insertions/suppressions au sein de segments grâce à un coût supplémentaire α pour commencer un segment, puis un coût β pour continuer ce segment (Altschul et Erickson, 1986). De plus nous positionnons le coût d'alignement d'un mot avec le segment d'erreur à zéro afin de pousser les mots à y être alignés en priorité. En dehors du segment d'erreur, le coût de substitution γ est calculé à partir de la similarité entre les mots alignés :

— similarité wordnet : $\gamma = 0.5$ si les deux mots partagent un synset dans Wordnet, sinon $\gamma = 1$;
— similarité dans un plongement : γ est la similarité cosine entre les vecteurs représentant les mots dans un plongements à 300 dimensions. Ce plongement a été entrainé sur un corpus textuel du domaine en prenant la couche cachée d'un réseau de neurones entraîné à prédire chaque mot sachant son contexte (Mikolov et al, 2013) ;
— similarité phonétique : γ est la distance d'édition minimale entre toutes les paires de phonétisations des deux mots dans un lexique de transcription automatique (Lenzo, 2007).

Les variantes de systèmes sont créées en activant ou non ces différents coûts afin d'obtenir différentes propriétés de fusion.

4 *Reranker*

Le but du reranker est de sélectionner parmi l'ensemble des hypothèses (données par les systèmes de fusion) celle qui va générer la meilleure transcription selon les intentions de l'utilisateur. Nous arborons le problème de reranker comme un classifieur binaire : étant donné un ensemble de paramètres qui caractérisent la sortie d'un système de fusion, cette sortie correspond-elle bien aux intentions de l'utilisateur ?

Nous proposons d'extraire 14 paramètres pour caractériser la sortie d'un système de fusion :

— Identité du système de fusion
— Similarité du cosine entre l'hypothèse de fusion et la phrase originale
— Similarité du cosine entre l'hypothèse de fusion et la phrase de clarification
— Minimum, maximum et moyenne du score au mot donné par un réseau de neurones récurrents (*Recurrent Neural Network*, RNN)
— Longueur de la phrase du système de fusion
— Nombre de systèmes de fusion d'accord avec l'hypothèse de fusion
— Distance de Levenshtein entre l'hypothèse de fusion et la phrase originale
— Distance de Levenshtein entre l'hypothèse de fusion et la phrase de clarification

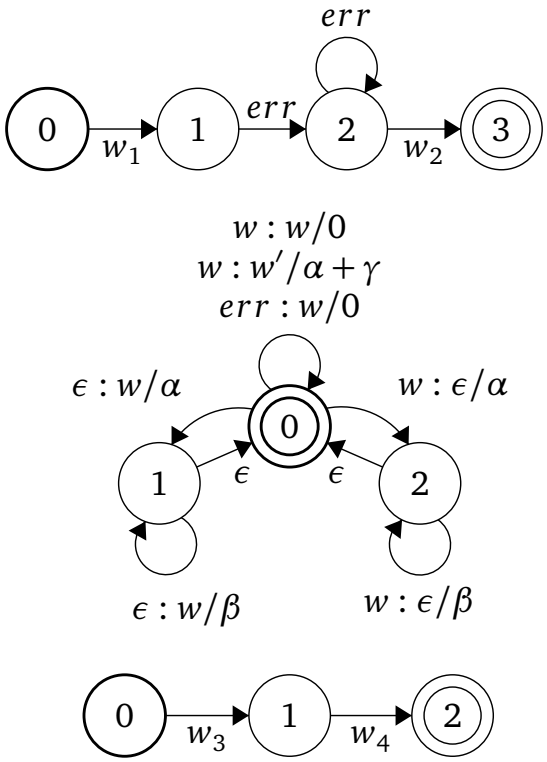


FIGURE 1 – Alignement par composition de transducteurs. L'accepteur du haut, A_o représente la phrase d'origine avec une boucle de symboles d'erreur à la place du segment d'erreur ; le transducteur du milieu T_e est un patron pour le transducteur d'édition pour une paire de mots (w, w') ; A_c est l'accepteur représentant la réponse de à la question de clarification. L'alignement est obtenu par le plus court chemin de la composition $A_o \circ T_e \circ A_c$. α est le coût de début d'une séquence d'insertions/supressions, β est le coût souffert pour ajouter un mot à cette séquence, et γ est le coût de substitution entre ces séquences.

— Différence de longueur entre l'hypothèse de fusion et la phrase originale
— Différence de longueur entre l'hypothèse de fusion et la phrase de clarification
— La phrase est-elle incluse dans l'originale ? Dans la clarification ?

Le reranker est un *Multi-Layer Perceptron* (MLP) doté de 6 couches (1 couche d'entrée, 4 couches cachées et 1 couche de sortie), entraîné via l'algorithme de *backpropagation*. Les couches d'entrée, cachées et de sortie sont composées respectivement de 14, 10 et 1 neurones. Dans la phase d'entraînement, le neurone de la couche de sortie est mis : à +1 si l'hypothèse donnée par le système de fusion correspond à la référence et à −1 dans l'autre cas. Nous utilisons un grand nombre de dialogues de clarification obtenus par *crowd-sourcing* (décrit dans la Section 5) comme corpus d'apprentissage, le RNN est entraîné sur le corpus TRANSTAC (corpus *in-domain*). Finalement, le système sélectionné parmi l'ensemble de systèmes de fusion donnés est celui qui obtient le meilleur score [1].

5 Expériences et résultats

L'approche développée dans ce papier est testée dans le cadre du projet BOLT Tâche b/c, conversation bilingue humaine-humaine par l'intermédiaire d'une machine. Le système joue le rôle d'un interprète qui peut prendre l'initiative de clarifier la saisie de l'utilisateur avant de la traduire.

1. Pour le MLP nous utilisons le toolkit FANN (http ://leenissen.dk/fann/wp) et pour le RNN nous utilisons le toolkit RNNLM (http ://www.fit.vutbr.cz/ imikolov/rnnlm).

Les systèmes de détection d'erreur de RAP (Reconnaissance Automatique de la Parole) et de TA (Transcription Automatique) sont gérés en amont pour détecter les segments d'erreur. Le module de dialogue demande alors à les reformuler. Le système de fusion, décrit dans ce papier, prend en compte la phrase originale et la phrase de clarification, dans le but de produire une meilleure transcription avant la traduction. Dans ce qui suit, seule la langue anglaise est traitée mais le traitement serait le même pour une autre langue.

Afin d'évaluer la qualité de notre système, nous avons produit deux corpus : un corpus de texte de reformulation collecté sur *Amazon Mechanical Turk* (AMT), principalement utilisé pour l'entraînement ; et un corpus de parole reproduisant le dialogue de clarification dans BOLT. Le premier corpus est composé d'un jeu de 900 phrases *in-domain* pour lesquelles nous demandons aux *Turkers* de reformuler aléatoirement une partie de la phrase (3 *Turkers* × 3 segments aléatoires × 100 phrases en entrée). Le jeu de données est ensuite sous-échantillonné de manière aléatoire en jouant sur les limites de clarification et d'erreur pour obtenir 11 175 phrases uniques. Le corpus de parole est composé de 70 phrases de dialogue pour lesquelles la phrase originale contient au moins un segment d'erreur lié au système de RAP. Ce corpus est très difficile pour le système de RAP car il contient un grand nombre de mots hors vocabulaire. La transcription automatique a été obtenue en utilisant le système de RAP du SRI basé sur les DNN. Ce système a été utilisé dans le cadre du projet BOLT et a obtenu sur les phrases originales un TEM de 30.6 % et 14.7 % sur les phrases de clarification. Dans la suite du papier, nous appelons *AMT* le corpus collecté par *crowd-sourcing* et *Parole* le corpus de parole audio.

Dans nos expériences, nous proposons 9 systèmes de fusions : *replace* et *insert* des systèmes *baseline*, un alignement de Levenshtein sans boucle d'erreur (*Align no-err-loop*), un alignement avec une boucle d'erreur (*Err-loop*), un alignement avec une pénalité affine d'insertion/suppression et une boucle d'erreur (*Err loop + affine gap* ; $\alpha = \beta = \gamma = 1$), le même système mais celui-ci aligne une séquence phonétique au lieu d'une séquence de mots (*Phonetic*), une pénalité affine d'insertion/suppression et une boucle d'erreur avec γ correspondant à la similarité phonétique entre les mots (*Phonetic + words*), le même système mais γ correspondant au minimum entre la similarité *Wordnet* et la similarité phonétique (*Phonetic + Wordnet*), le même système mais γ correspondant au minimum entre la similarité de plongement et la similarité phonétique (*Phonetic + embedding*).

Méthode	TCC	TEM
Replace (baseline)	45.94	20.24
Insert (baseline)	25.41	32.56
Align no err-loop	61.84	13.00
Err loop	69.76	07.30
Err loop + affine gap	75.02	05.37
Phonemes	51.86	10.51
Phonetic + words	77.60	04.74
Phonetic + Wordnet	80.02	03.99
Phonetic + embedding	78.70	04.42
Reranker	**88.36** ±0.72	**02.72** ±0.23
Oracle	96.18	00.56
(Bechet et Favre, 2013)	63.70	12.44

TABLE 1 – Les résultats des systèmes de fusion et du reranker sont reportés en termes de Taux de Classification Correcte (TCC) et Taux d'Erreur Mots (TEM) sur le corpus *AMT*. L'oracle est calculé systématiquement en sélectionnant la meilleure sortie des systèmes de fusion. Pour comparaison, nous donnons aussi les résultats de la méthode proposée dans (Bechet et Favre, 2013).

Les performances des systèmes de fusion sont évaluées selon deux métriques : (1) le Taux de

Classification Correcte (TCC) qui correspond au nombre de bonnes réponses données par le système de fusion sur le nombre de cas testés et (2) le taux d'erreur de mots (TEM) qui est le taux de mots incorrectement reconnus de l'hypothèse par rapport à la référence.

Le Tableau 1 donne les résultats obtenus par les différents systèmes de fusion sur le corpus *AMT*. On peut constater que tous les systèmes à base d'alignements améliorent de manière significative la *baseline*, conduisant à une réduction importante du TEM. Le système de fusion ayant obtenu les meilleurs résultats est : *Phonetic + Wordnet*. Il obtient un taux de classification correcte de 80.02 %. L'alignement sans aucune connaissance *a priori* sur la langue (sans lexique phonétique, ni représentation sémantique) obtient un taux de classification correcte de 75.02 %. Le reranker offre une amélioration supplémentaire de 10.4 % de précision et de 31.8 % de TEM par rapport au meilleur système de fusion. Nous notons que sur ce corpus, la méthode proposée dans (Bechet et Favre, 2013) permet d'améliorer uniquement la *baseline* et le système *Align no err-loop* (qui est un alignement basique).

Dans le Tableau 2, nous présentons les mêmes résultats sur le corpus *Parole*. Le comportement général des systèmes de fusion reste le même que sur le corpus de texte, excepté que le taux de classification correcte est plus bas sur les sorties de RAP. Pourtant, le TEM est grandement amélioré en utilisant la phrase de clarification puisqu'on observe un gain d'environ 10 points absolus. Sur le corpus *Parole*, le système *Phonetic + Wordnet* obtient de moins bons résultats que le système *Phonetic + embedding*, probablement parce que l'espace continu est plus robuste aux erreurs commises par le système de RAP. Les méthodes proposées dans (Bechet et Favre, 2013) donnent de bons résultats sur le texte de référence mais elles ne parviennent pas à améliorer le TEM sur la sortie de RAP, un facteur de motivation pour les futurs travaux.

Concernant le reranker, il obtient un taux de classification correcte proche de l'oracle mais le TEM n'est pas aussi élevé que sur le corpus *AMT*. Ceci suggère que (1) plusieurs stratégies de système de fusion doivent être explorées pour améliorer l'oracle, et (2) le reranker devrait être entraîné pour minimiser le TEM en plus de maximiser le taux de classification correcte.

	Ref.		RAP	
Méthode	TCC	TEM	TCC	TEM
Replace (baseline)	25.71	49.08	12.86	55.18
Insert (baseline)	37.14	28.09	08.57	46.52
Align no-err-loop	35.71	21.56	10.00	32.34
Err-loop	70.00	08.09	17.14	22.27
Err loop + affine gap	84.29	02.13	21.43	21.56
Phonemes only	74.29	04.68	15.71	21.56
Phonetic + words	85.71	**01.84**	21.43	21.13
Phonetic + Wordnet	82.86	02.70	18.57	21.28
Phonetic + embed.	85.71	**01.84**	21.43	20.99
Reranker	**87.57** ±1.36	01.93 ±0.58	**26.67** ±0.78	**20.97** ±0.58
Oracle	90.00	1.42	28.57	16.17
(Bechet et Favre, 2013)	84.29	02.27	15.71	30.64
No clarification	0.0	15.14	0.0	30.60

TABLE 2 – Les résultats des systèmes de fusion, du reranker, de l'oracle, de (Bechet et Favre, 2013) sont reportés en termes de Taux de Classification Correcte (TCC) et Taux d'Erreur Mots (TEM) sur le corpus *Parole*.

Le Tableau 3 donne le taux de classification correcte obtenu par le reranker sur le corpus *AMT* et *Parole*. Les résultats donnés sur le corpus *AMT* sont réalisés en utilisant une méthode d'évaluation croisée *2-fold* (la première partie du corpus est utilisée pour entraîner les modèles et l'autre partie pour les tester). Les résultats sur le corpus *Parole* sont donnés en utilisant un modèle entraîné sur toutes les données *AMT*. Comme les poids du MLP sont initialisés de manière aléatoire, nous

	AMT	Parole	
	Texte	Ref.	RAP
TCC	90.65 ±1.68	90.83 ±1.43	58.74 ±2.00

TABLE 3 – Taux de classification correcte sur les corpus *AMT* et *Parole*.

lançons 100 fois l'expérience et rapportons la moyenne et l'écart-type du taux de classification correcte.

Nous observons que sur le texte de référence le reranker obtient un taux de classification de 90.65 % et 90.83 % respectivement pour les transcriptions de référence des corpus *AMT* et *Parole*, tandis que sur la transcription automatique le taux de classification chute à 58 %. Cette différence peut être expliquée par le fait que le reranker est entraîné sur du texte de référence (*AMT*) et il est donc moins robuste aux erreurs commises par le système de RAP.

Le Tableau 4 donne le nombre de fois où chaque système de fusion a été classifié à +1 sur le corpus *Parole* par le reranker (le neurone de la couche cachée doit être > 0), ainsi que le ratio de bonnes réponses parmi eux. Nous notons que le nombre de bonnes réponses est lié à l'oracle, ce qui explique pourquoi les résultats sont plus bas sur la sortie de la RAP. Le reranker est entraîné sur du texte propre *AMT*, les systèmes ont tendance à être choisis moins souvent sur la sortie de la RAP en raison d'erreurs de mots qui réduisent le nombre de mots appariés entre la phrase originale et la phrase de clarification. Nous notons que même si plusieurs systèmes de fusion peuvent être classés à +1 pour une instance donnée, nous utilisons l'*argmax* comme sortie pour le reranker.

	Ref.		RAP	
Méthode	Sel.	TCC	Sel.	TCC
Replace (baseline)	27.14	94.74	22.86	50.00
Insert (baseline)	47.14	78.79	45.71	18.75
Align no-err-loop	35.71	100.00	22.86	43.75
Err-loop	74.29	94.23	51.43	33.33
Err-loop + affine-gap	94.29	87.88	80.00	26.79
Phonemes only	80.00	89.29	58.57	24.39
Phonetic + words	92.86	89.23	77.14	27.78
Phonetic + Wordnet	87.14	91.80	70.00	26.53
Phonetic + embedding	92.86	89.23	81.43	26.32

TABLE 4 – Les systèmes de fusion classés avec +1 par le reranker sur le corpus *Parole* en pourcentage d'instance et taux de classification correcte.

6 Conclusion

Dans ce papier, nous proposons un système pour résoudre le problème de fusion entre une phrase originale contenant une erreur et une phrase de clarification. Nous proposons plusieurs alignements qui sont des variantes de Levenshtein et un reranker pour sélectionner la meilleure hypothèse. Le système proposé permet d'améliorer le TEM d'environ 30 % et le reranker aide à éliminer complètement les erreurs dans 26 % des instances, atteignant ainsi les scores de l'oracle.

Dans les futurs travaux, nous allons étudier l'utilisation de l'alignement non-monotone qui sont des approches utilisées en traduction automatique, et faire usage de treillis de mots dans l'opération de fusion.

Références

ALTSCHUL, S. F. et ERICKSON, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bulletin of mathematical biology*, 48(5):603–616.

BARZILAY, R. et MCKEOWN, K. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

BECHET, F. et FAVRE, B. (2013). Asr Error Segment Localization for Spoken Recovery Strategy. *In ICASSP*.

BOHUS, D. et RUDNICKY, A. I. (2009). The ravenclaw dialog management framework : Architecture and systems. *Computer Speech & Language*, 23(3):332–361.

DUFOUR, R., DAMNATI, G. et CHARLET, D. (2012). Automatic error region detection and characterization in lvcsr transcriptions of tv news shows. *In ICASSP*, pages 4445–4448. IEEE.

ELSNER, M. et SANTHANAM, D. (2011). Learning to fuse disparate sentences. *In Monolingual Text-To-Text Generation*, pages 54–63. Association for Computational Linguistics.

FILIPPOVA, K. et STRUBE, M. (2008). Sentence fusion via dependency graph compression. *In EMNLP*, pages 177–185. Association for Computational Linguistics.

GEROSA, M. et FEDERICO, M. (2009). Coping with out-of-vocabulary words : open versus huge vocabulary asr. *In ICASSP*, pages 4313–4316. IEEE.

HOSTE, L., DUMAS, B. et SIGNER, B. (2012). Speeg : a multimodal speech-and gesture-based text input solution. *In International Working Conference on Advanced Visual Interfaces*, pages 156–163. ACM.

HUGGINS-DAINES, D. et RUDNICKY, A. (2008). Interactive asr error correction for touchscreen devices. *In HLT*, pages 17–19. Association for Computational Linguistics.

JONSON, R. (2006). Dialogue context-based re-ranking of asr hypotheses. *In SLT*, pages 174–177. IEEE.

JUNG, S., LEE, C. et LEE, G. G. (2008). Using utterance and semantic level confidence for interactive spoken dialog clarification. *JCSE*, 2(1):1–25.

KOMBRINK, S., HANNEMANN, M. et BURGET, L. (2012). Out-of-vocabulary word detection and beyond. *In Detection and Identification of Rare Audiovisual Cues*, pages 57–65. Springer.

LAURENT, A. et AL (2011). Computer-assisted transcription of speech based on confusion network reordering. *In ICASSP*, pages 4884–4887. IEEE.

LENZO, K. (2007). The cmu pronouncing dictionary.

LIU, X., GALES, M. J. et WOODLAND, P. C. (2012). Use of contexts in language model interpolation and adaptation. *Computer Speech & Language*.

LÓPEZ-CÓZAR, R., GRIOL, D. et QUESADA, J. F. (2010). New technique to enhance the performance of spoken dialogue systems by means of implicit recovery of asr errors. *In Spoken Dialogue Systems for Ambient Environments*, pages 96–109. Springer.

MAAS, A. L. et AL (2012). Recurrent neural networks for noise reduction in robust asr. *In Interspeech*.

MARIN, A. et AL (2012). Using syntactic and confusion network structure for out-of-vocabulary word detection. *In SLT*, pages 159–164. IEEE.

MIKOLOV, T. et AL (2013). Efficient estimation of word representations in vector space. *In ICLR*.

MISU, T., KAWAHARA, T. et KOMATANI, K. (2004). Confirmation strategy for document retrieval systems with spoken dialog interface. *In Interspeech*.

MITCHELL, J. et LAPATA, M. (2009). Language models based on semantic composition. *In EMNLP*, pages 430–439. Association for Computational Linguistics.

OGATA, J. et GOTO, M. (2005). Speech repair : quick error correction just by using selection operation for speech input interfaces. *In Interspeech*, pages 133–136. Citeseer.

PARADA, C. et AL (2011). Learning sub-word units for open vocabulary speech recognition. *In ACL*, pages 712–721.

PARADA, C., DREDZE, M., FILIMONOV, D. et JELINEK, F. (2010). Contextual information improves oov detection in speech. *In NAACL*.

PRASAD, R. et AL (2012). Active error detection and resolution for speech-to-speech translation. *IWSLT*.

SEIGEL, M. S., WOODLAND, P. C. et AL (2011). Combining information sources for confidence estimation with crf models. *In Interspeech*, pages 905–908.

SHIN, J. et AL (2002). Analysis of user behavior under error conditions in spoken dialogs. *In ICSLP*, volume 2.

STOYANCHEV, S., LIU, A. et HIRSCHBERG, J. (2012). Clarification questions with feedback. *In Feedback Behaviors in Dialog*.

WENINGER, F. et AL (2012). Non-negative matrix factorization for highly noise-robust asr : To enhance or to recognize ? *In ICASSP*, pages 4681–4684. IEEE.

YU, D. et DENG, L. (2010). Semantic confidence calibration for spoken dialog applications. *In ICASSP*, pages 4450–4453. IEEE.

Production et planification de la parole spontanée en contexte narratif chez l'enfant âgé de 3 à 11 ans

Jean-Marc Colletta[1], Catherine Pellenq[2], Ali Hadian Cefidekhanie[1]
(1) LIDILEM, , U. Stendhal, Grenoble, France
(2) ESPE, U. Joseph Fourier, Grenoble, France
jean-marc.colletta@u-grenoble3.fr, catherine.pellenq@ujf-grenoble.fr, ali.hadian-cefidekhanie@u-grenoble3.fr

RESUME

Le traitement de la parole en production chez l'enfant reste un phénomène mal connu. Cette étude met en évidence des évolutions en fonction de l'âge chez 85 enfants âgés de 3 à 11 ans en train de raconter à partir d'un extrait de film d'animation. Les productions linguistiques des sujets ont été transcrites et annotées sous *ELAN*. Pour chaque récit, nous avons évalué la quantité d'informations verbalisées et nous avons mesuré le débit de parole ainsi que la densité informationnelle des groupes phoniques (segments de parole séparés par des pauses). Les résultats obtenus montrent une augmentation de toutes les variables mesurées ainsi qu'une accélération du débit au fil de l'âge et suggèrent l'existence de deux patrons développementaux distincts, l'un portant sur les habiletés en matière d'encodage et de planification de la parole, l'autre sur les performances narratives. La discussion explore les points de rendez-vous possibles et fournit de nouvelles hypothèses.

ABSTRACT

On speech production and speech planning during narration in children aged 3 to 11 years.
The way the child processes speech in language production remains largely unstudied. This study examines age-related changes in oral narrative discourse of 85 French children aged 3 to11 years who tell a story from a short animated film. Language was transcribed and coded using *ELAN* as an annotation tool. We measured the informational content of each narrative, the speaker's speech flow, and the information density of phonic groups (e.g. speech segments marked by pauses). The results show an age related increase in all quantity and density measures including speech flow rate. They suggest the existence of two distinct developmental patterns as regards speech planning and narrative abilities. We discuss new lines of investigation on the developmental 'rendez-vous' of both types of abilities.

MOTS-CLES : parole, production, développement, planification, narration
KEYWORDS: speech, production, development, planning, narrating

1 Introduction

Au cours du dernier demi-siècle, la psycholinguistique a largement interrogé l'acquisition des connaissances phonologiques, lexico-sémantiques et morphosyntaxiques nécessaires à la parole, le développement précoce des conduites de communication et, dans une moindre mesure, la genèse des habiletés pragmatiques. Pourtant il est un domaine encore peu exploré, c'est celui de l'énonciation même de la parole et de son évolution chez l'enfant. Comment l'enfant de trois ans et plus produit-il du langage parlé ? Qu'est-ce qui évolue entre la fin de la petite enfance et le début de l'adolescence et que trahit cette évolution des habiletés sous-jacentes en matière de conceptualisation, de planification, de mémoire ou de

contrôle moteur ? La psycholinguistique de la production de la parole chez l'enfant reste encore largement à construire. Cette étude prolonge les travaux antérieurs menés sur le débit verbal et propose de nouvelles pistes de réflexion.

2 Questions à propos de la production linguistique chez l'enfant

En matière de production de la parole chez l'enfant, la principale variable qui a attiré l'attention des chercheurs est le débit de parole. Les études portant sur la relation entre le débit, la boucle articulatoire et la mémoire phonologique ont fait apparaître une accélération du débit de parole au cours de l'enfance (Adams & Gathercole, 1995 ; Hulme et al., 1984), et même si la régularité de cette évolution (Walker & Archibald, 2006) ou la constance de la relation entre débit et empan mnésique (Ferguson, Bowey & Tilley, 2002) reste en question, des données récentes viennent encore confirmer cette évolution (Pavao Martins et al., 2007). Dans les travaux que nous rapportons ci-après, le débit est calculé sur le seul temps de parole (durée des pauses exclue) et il correspond ainsi à la vitesse d'articulation mesurée au nombre de syllabes par seconde (Legendre et al. 2012).

Comparant le débit chez l'enfant avec le débit chez l'adulte, certains annoncent un débit moyen de 2 à 3 syll/sec. chez le jeune enfant, et un débit compris entre 5 et 7 syll/sec. chez l'adulte (den Os, 1990 ; Koopmans-van Beinum, 1993). Mais les estimations sont à prendre avec précaution étant donné l'ampleur des variations enregistrées. Ainsi, Kowal, O'Connell et Sabin (1975) enregistrent un débit de 2 syll/sec. à 4 ans alors que den Os (1990) enregistre ce même débit dès 2.4 ans et note une accélération de la vitesse de parole à partir de cet âge (3 syll/sec. à 3 ans). Par ailleurs, Kowal, O'Connell et Sabin (1975) rapportent un débit de 4 syll/sec. à 12 ans, alors que Ryan (2000) trouve déjà un débit de 3.8 syll.sec. à 4 ans et demi. De leur côté, à partir d'une étude longitudinale, Walker et Archibald (2006) ne notent pas d'accélération de la vitesse de parole entre 4 et 6 ans. Chez l'enfant dont la langue maternelle est le français, en extrapolant à partir des données fournies par Konopczynski et Vinter (1994) qui utilisent comme indice la durée des syllabes, on obtient un débit moyen de 2.5 à 3 syll/sec. à l'âge de 2 ans, âge où, d'après les auteures, le rythme de base de l'articulation du français est en place. Dans une étude préliminaire examinant les performances enfantines dans des tâches narrative et conversationnelle (Colletta, Pellenq & Rousset, 2008), nous avons relevé chez l'enfant plus âgé des valeurs moyennes qui vont de 3.6 syll/sec. à l'âge de la petite section de maternelle (3-4 ans) à 4.4 syll/sec. à l'âge du CM1 (9-10 ans) trahissant une augmentation réelle mais modeste de la vitesse d'articulation au fil de l'âge.

Lors de la même étude, nous avons mis en évidence un effet de l'âge sur un aspect moins documenté de la prosodie en mesurant la durée et le contenu informationnel des groupes de souffle (segments de parole ininterrompus entre deux pauses). Ainsi, pour la tâche narrative, nous avons relevé un allongement significatif de la durée de ces groupes (de 1.18 sec. en moyenne chez les plus jeunes à 1.83 sec. chez les plus âgés) et une augmentation significative de leur teneur en syllabes (de 4.59 chez les plus jeunes à 7.54 chez les plus âgés). Etant donné les relations (métriques, accentuelles et intonatives) qu'entretiennent les unités prosodiques avec la syntaxe, cet accroissement de la densité informationnelle des groupes de souffle a à voir avec le développement syntaxique (Diessel, 2004).

Le constat précédent a été mis en évidence à la faveur d'une tâche narrative. Or la planification de la parole se pose en des termes fondamentalement différents dans le

contexte dialogal et interactif de l'échange binaire du type question-réponse et dans le contexte de la parole monologuée qui aboutit (entre autres) à la narration d'un événement. Dans le cas du récit, la production verbale est contrainte à un niveau supplémentaire : le niveau du texte, et elle doit mettre en œuvre des propriétés de cohésion et de cohérence (Halliday & Hasan, 1976) qui sont sans équivalent dans le cours du dialogue construit à la faveur d'échanges rapides. De plus, l'activité de raconter nécessite des capacités cognitives d'évocation, de contextualisation des informations et de décentration (Karmiloff-Smith, 1979 ; Tolchinsky, 2004) dont la lente mise en place contribue à expliquer pourquoi le développement narratif s'étale sur des années, depuis les premiers scripts simples verbalisés par le jeune enfant jusqu'aux récits parlés complets et commentés des jeunes adolescents (Colletta, 2004). Les travaux sur le développement narratif en production ont d'ailleurs mis en évidence des périodes (vers 7-8 ans, puis entre 9 et 11 ans) au cours desquelles les performances narratives semblent évoluer de manière significative (Fayol, 1985, 2000 ; Hickmann, 2003). La narration est bien le lieu par excellence où il est possible de mettre en évidence des évolutions tant en matière d'habiletés textuelles qu'en matière de production et planification de la parole.

La présente étude vise à compléter les observations réalisées en 2008, limitées à 4 classes d'âge, en examinant les récits parlés d'enfants âgés de 3 à 11 ans appartenant à tous les niveaux scolaires de la maternelle et du primaire et collectés à partir d'un protocole unique. Tout en évaluant l'accomplissement de la tâche narrative à l'aide d'indices de quantité informationnelle, nous voulions vérifier la solidité des évolutions qui touchent les variables prosodiques (le débit de parole, la durée et le contenu informationnel des groupes de souffle). Nous souhaitions en outre préciser la forme (continue ou discontinue) des évolutions affectant chaque série d'observations, et les mettre en relation pour établir l'existence ou non d'une coévolution, en d'autres termes, vérifier si les habiletés narratives sont ou non liées directement aux capacités de production et planification de la parole, et de quelle manière.

3 Méthodologie de l'étude

Les récits parlés sur lequel reposent nos observations ont été enregistrés dans un groupe scolaire de l'agglomération grenobloise auprès d'enfants âgés de 3 ½ à 11 ½ans. Les récits devant, à terme, être analysés au plan de leur teneur en informations gestuelles, seuls 85 d'entre eux (sur 122 enregistrés au total) ont été incorporés au corpus de base.

	G1	G2	G3	G4	G5	G6
Age moyen (jours)	1607 (188.79)	2199 (115.71)	2586 (113.57)	3054 (186.33)	3579 (94.53)	3951 (138.62)
Age moyen (ans ; mois)	4;4	6;1	7 ;1	8 ;4	9 ;9	10 ;9
Classe d'âge (mois)	43-60	66-78	80-92	94-109	113-121	123-136
Ecart interclasses (mois)	17	12	12	15	8	13
Effectif	15	15	12	13	16	14
Filles/garçons	5/10	8/7	5/7	4/9	8/8	10/4

TABLEAU 1 – Caractéristiques des sujets

Afin de permettre une comparaison inter-âges plutôt qu'entre niveaux scolaires, les enfants ont été répartis dans 6 classes d'âge correspondant plus ou moins à ces niveaux (par

exemple, le G1 regroupe les enfants de petite et de moyenne section, le G2 les enfants de grande section et les plus jeunes du CP, etc.). Le tableau 1 indique pour chaque groupe l'âge moyen (lorsqu'il apparaît, l'écart-type est indiqué entre parenthèses) et l'empan d'âge ainsi que l'effectif. Tous les sujets ont passé le test standardisé ELO d'évaluation du langage (Khomsi, 2001) afin de s'assurer qu'aucun ne révélait un profil atypique au regard de ses capacités lexicales et syntaxiques. Les scores moyens obtenus sur l'ensemble de ces épreuves sont tous dans la norme des groupes d'âge.

Il était ensuite proposé une tâche narrative au cours de laquelle chaque enfant devait raconter un extrait d'un film d'animation qu'il venait de visionner (les trois premières minutes de « Rasé de près » de Nick Park, une aventure de *Wallace et Gromit*). Chaque récit a été filmé avec enregistrement sonore externe (microphone directionnel posé sur la table près de l'enfant), puis a été transcrit à l'aide du logiciel *ELAN* qui, grâce à l'alignement de la transcription sur la source audio, permet un repérage fin des unités prosodiques qui nous intéressent tout particulièrement : les groupes phoniques. Un « groupe phonique » (désormais GP) est un « […] groupe de syllabes formant une unité sonore pouvant avoir un sens ou non. » (Léon, 2011 : 141). Il désigne un segment de parole ininterrompu entre deux silences ou pauses (un GP correspond à un « groupe de souffle » lorsqu'il est encadré par des pauses avec inspiration ou expiration du locuteur). A l'instar de den Os (1990), nous avons considéré que toute interruption de parole supérieure à 200ms équivalait à une pause et permettait d'identifier la frontière (amont ou aval) d'un GP.

Parmi les mesures effectuées, la durée d'un récit correspond à la somme de la durée des GP qu'il comporte. En effet, les temps de pause et les silences étant exclus, la durée d'un récit équivaut au temps moyen (en secondes) mis par le sujet pour le verbaliser. Les indices de quantité informationnelle permettant d'évaluer l'accomplissement de la tâche sont obtenus en comptabilisant le nombre de syllabes effectivement produites par le narrateur, le nombre de mots (y compris les mots incomplets et les marques d'hésitation telles « heu ») et le nombre de propositions (au double sens logique et syntaxique du terme : une proposition = une prédication, c.à.d. une information linguistique composée d'arguments organisés autour d'un verbe-noyau) de chaque récit. Le débit verbal, ici calculé comme un indice de vitesse articulatoire, est obtenu en divisant le nombre de syllabes par la durée du récit en secondes. Enfin, des indices de densité informationnelle sont obtenus à partir des GP identifiés dans chaque récit, en mesurant leur teneur en syllabes, en mots et en propositions.

4 Résultats

Toutes les variables mesurées ont fait l'objet de comparaisons de variances qui ne montrent pas d'effet du genre ni de l'interaction entre genre et âge. En revanche, les tests réalisés montrent un net effet de l'âge, y compris sur le débit verbal.

Nous indiquons d'abord, dans le tableau 2, les valeurs moyennes obtenues pour la longueur du récit et les indices de quantité informationnelle, utiles pour apprécier le niveau d'accomplissement de la tâche narrative. Très court chez les enfant les plus jeunes, le récit s'allonge et voit sa durée doubler dès l'âge du CP (G3). Cet allongement de la durée va de pair avec un accroissement de tous les indices de quantité informationnelle, qu'on calcule celle-ci à partir du nombre de syllabes, du nombre de mots ou du nombre de propositions, avec un pic entre 9 ans ½ et 10 ans (G5).

	G1	G2	G3	G4	G5	G6
Récit (durée en sec.)	41.66 (19,97)	61.04 (34,78)	100.39 (52,76)	87.33 (38,42)	103.38 (48,25)	95.43 (33,87)
Syllabes (nb)	156,53	234,20	346,17	354,62	423,81	407.36
Mots (nb)	112,27	173,67	257,75	260,23	316,19	304.43
Propositions (nb)	19,13 (7,42)	28,87 (16,44)	45,67 (23,97)	44,62 (22,75)	56,88 (32,67)	53.64 18,24)

TABLEAU 2 – Longueur du récit et indices de quantité informationnelle

La comparaison des variances montre un effet de l'âge pour l'allongement de la durée du récit ($F(5, 79) = 5.910$, $p<.000$), et le test post-hoc de Fisher permet de localiser un écart significatif entre, d'une part, G1 et les quatre dernières classes d'âge (G3 à G6), et d'autre part G2 et les classes d'âge G3, G5 et G6, les deux premières classes d'âge (G1 et G2) ne se différenciant pas entre elles (voir la figure 1). Un effet analogue apparaît pour toutes les variables de quantité informationnelle. A titre d'exemple, la comparaison intergroupes pour le nombre de propositions donne $F(5, 79) = 6.788$, $p<.000$), avec toujours des écarts significatifs entre les mêmes groupes (voir la figure 2). La stabilité de l'effet est liée à la forte co-variation entre les variables (nombre de syllabes, de mots et de propositions).

Le tableau 3 donne les valeurs moyennes obtenues pour le débit verbal mesuré en syllabes par secondes. La comparaison des variances est également significative pour cette variable ($F(5,79)=3.950$, $p<.003$) pour laquelle on relève des valeurs plus faibles chez les enfants les plus jeunes (notamment 3,54 syll/sec. pour le G3) et des valeurs plus élevées chez les plus âgés (4,29 syll/sec. pour le G6).

	G1	G2	G3	G4	G5	G6
Débit verbal (syll/sec)	3.84 (0.42)	3.88 (0.40)	3.54 (0.50)	4.02 (0.41)	4.08 (0.47)	4.29 (0.50)

TABLEAU 3 – Débit verbal

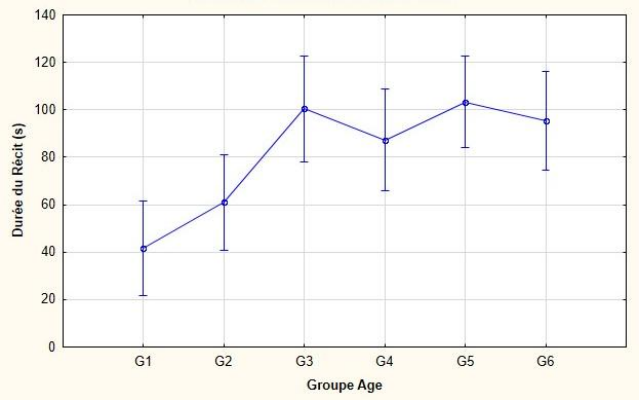


FIGURE 1 – Age et durée du récit

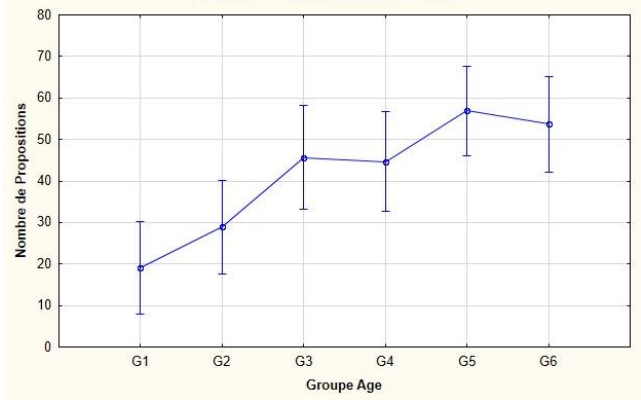


FIGURE 2 – Age et quantité informationnelle

Dans le tableau 4, nous présentons les résultats obtenus concernant les GP et les indices de densité informationnelle. Le nombre moyen de GP par récit suit une évolution similaire à celle des indices de quantité informationnelle, avec un accroissement significatif entre le G2 et le G3. La durée moyenne d'un GP est également sensible à l'âge puisqu'elle passe de 1,25 sec. (G1) à 1,97 sec. (G6), mais son évolution au fil de l'âge se présente de manière très différente.

	G1	G2	G3	G4	G5	G6
GP (nb)	33,13	37,47	63,00	55,62	62,19	49.85
GP (durée en sec.)	1.25 (0.28)	1.65 (0.26)	1.67 (0.48)	1.62 (0.29)	1.70 (0.44)	1.97 (0.39)
Syllabe/GP (nb)	4,73	6,44	5,81	6,51	6,91	8,32
Mots/GP (nb)	3,44	4,79	4,37	4,86	5,21	6,30
Propositions/GP (nb)	0.60 (0.13)	0.79 (0.23)	0.76 (0.26)	0.82 (0.19)	0.90 (0.27)	1.11 (0.24)

TABLEAU 4 – GP et indices de densité informationnelle

La comparaison des variances montre un effet de l'âge pour l'allongement de la durée du GP ($F(5,79) = 5.867$, $p<.000$), et le test post-hoc de Fisher permet cette fois de localiser deux écarts significatifs, le premier entre le G1 et tous les autres groupes, le second entre le G6 et tous les autres groupes tandis que les groupes G2 à G5 ne se distinguent pas entre eux (voir la figure 3). Le même effet apparaît pour toutes les variables de densité informationnelle. La comparaison intergroupes pour le nombre de propositions par GP (qui montre que l'enfant du G1 a besoin de deux GP pour formuler une proposition (0,60) tandis que l'enfant du G6 place plus d'une proposition dans un GP (1,11)), donne $F(5,79)=8,190$, $p<000$, avec toujours les mêmes écarts significatifs entre les classes d'âge G2 à G6 et les groupes G1 d'une part, G6 d'autre part (voir la figure 4). Là encore, la stabilité de l'effet est liée à la co-variation entre les variables (nombre de syllabes, de mots et de propositions).

Dans le prolongement de ces résultats, une analyse des corrélations (test de Pearson) entre l'âge et l'ensemble des variables mesurées fait apparaître des corrélations positives plus fortes pour les variables de densité informationnelle ($R = 0,571$ pour la relation entre l'âge et le nombre de syllabes par GP ; $R = 0,533$ pour celle entre l'âge et le nombre de propositions par GP) que pour les autres variables dont le débit verbal ($R = 0.325$ pour la relation entre l'âge et cette variable). En outre, l'analyse des régressions montre que l'âge explique 32% de la variance du nombre de syllabes par GP, et 28% de la variance du nombre de propositions par GP.

En guise de synthèse, il apparaît donc que plus l'enfant avance en âge, plus son récit s'allonge et contient d'informations (en syllabes, mots et propositions), plus il parle vite, et surtout, plus il répartit l'information linguistique dans des segments de parole qui s'allongent et gagnent en complexité syntaxique.

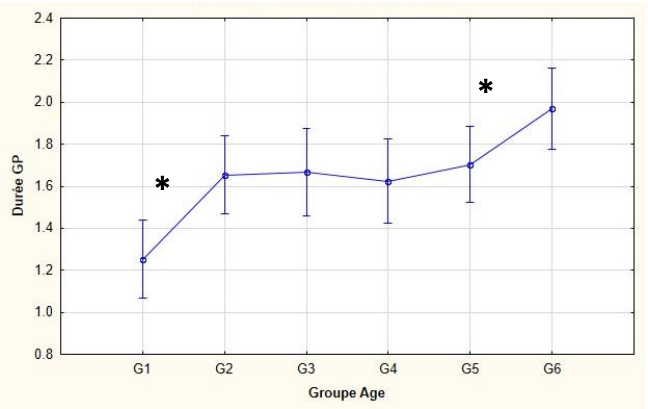


FIGURE 3 – Age et durée du GP

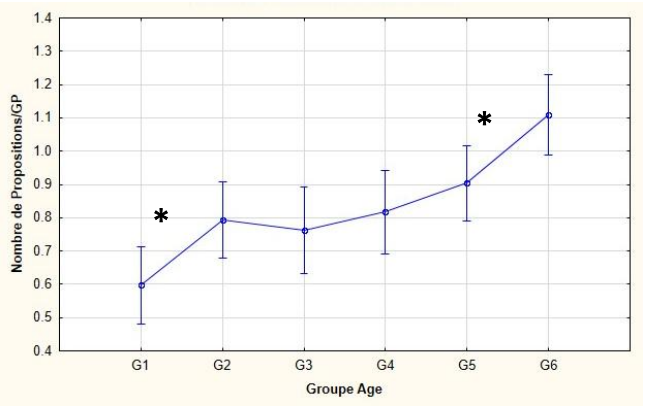


FIGURE 4 – Age et densité informationnelle

Discussion

Le fait que l'enfant plus âgé réalise des récits plus longs et plus riches en informations est sans surprise et nos résultats, obtenus auprès d'enfants âgés de 3 ans ½ à 11 ans ½, viennent compléter et conforter d'autres observations menées auprès d'enfant francophones à partir de récits spontanés (Colletta, 2004) et élicités (Colletta, Pellenq & Guidetti, 2010). Par ailleurs, le résultat concernant le débit de parole confirme le lent accroissement de ce dernier tout au long de l'enfance, comme le suggéraient des observations antérieures (Colletta, Pellenq & Rousset, 2008 ; Walker & Archibald, 2006).

Plus intéressants sont les résultats provenant de l'analyse de la durée et du contenu informationnel des GP. Ces résultats montrent un allongement significatif de la durée du GP au fil de l'âge, et surtout, un accroissement notable de son contenu informationnel qui s'observe aussi bien au niveau syntaxique de la proposition qu'au niveau de la syllabe et du mot. Le fait qu'à 11 ans ½ les GP de l'enfant contiennent en moyenne plus d'une proposition alors qu'à 3 ans ½ il ait besoin de deux GP pour en verbaliser une est particulièrement significatif et montre qu'avec l'âge, il devient de plus en plus habile à planifier sa production verbale, à tout le moins au plan local de l'organisation syntaxique de l'énoncé (en effet, il reste hasardeux d'établir un lien direct entre cette habileté et la gestion de la narration au plan de son organisation textuelle).

L'autre enseignement de cette étude porte précisément sur les relations entre les deux types d'habiletés. Les deux séries de résultats, celle qui porte sur les indices de quantité informationnelle et celle qui porte sur le contenu informationnel des GP, suggèrent l'existence de deux patrons développementaux distincts dans leur forme, le premier avec une augmentation significative de la durée et de la quantité d'informations du récit entre 6 et 7 ans, soit à l'âge du CP, et le second avec deux augmentations significatives de la durée et du contenu informationnel des GP, l'une entre 4 ans ½ et 6 ans (dernière année de maternelle) et l'autre entre 10 et 11 ans (dernière année de primaire).

La durée et la quantité informationnelle sont des indices parmi d'autres des habiletés narratives qui se mettent progressivement en place chez l'enfant (Colletta, 2004). Or le palier des 6-7 ans semble important tant au plan des connaissances textuelles qu'au plan des outils linguistiques assurant la cohésion/cohérence textuelle : l'enfant verbalise de moins en moins de scripts (brèves séquences d'actions prototypiques) au profit de chaînes d'événements manifestant une organisation narrative (Fayol, 2000), et son emploi des marques linguistiques à fonction discursive (pronoms, déterminants, localisateurs spatiaux et temporels) commence à se généraliser (Hickmann, 2000). On peut y voir un effet de la scolarisation et du contact prolongé avec la littéracie.

La durée et le contenu informationnel des GP attestent en revanche de l'organisation locale des informations linguistiques au sein des segments de parole entre pauses. Celle-ci met en jeu aussi bien la respiration (des contraintes biophysiologiques) que la mémoire de travail, l'accès lexical et l'encodage phonétique, ou encore, la nature des informations linguistiques à transmettre, comme le montrent les effets de fréquence (sur ces différents aspects, voir Meyer & Wheeldon, 2006). Les résultats obtenus au terme de cette étude développementale semblent pointer deux périodes clés au cours desquelles les GP s'allongent significativement et permettent de verbaliser des segments de parole plus riches en informations. Sans nier l'effet de l'âge sur la mémoire et les autres habiletés cognitives, on ne peut pas ne pas mettre

en relation cette évolution avec le développement syntaxique en production. Pistant l'émergence des constructions complexes entre 2 et 5 ans, Diessel (2004) a montré que les productions linguistiques évoluent à partir de structures syntaxiques simples (constructions figées à partir de schémas tels « vouloir faire + quelque chose», structures dialogués du type « P parce que Q » avec P verbalisé par l'interlocuteur) vers l'expression de relations complexes impliquant la coordination ou la subordination. Cette évolution se traduit par un allongement des énoncés dans le répertoire verbal de l'enfant aux âges considérés. Un prolongement direct de notre étude consistera à examiner en détail cette relation privilégiée entre prosodie et syntaxe à travers l'analyse syntaxique des GP du corpus. Il s'agira de voir comment l'organisation syntaxique se reflète dans la production des groupes phoniques, et s'il existe bien un rendez-vous développemental entre les deux entre 4 ans ½ et 6 ans.

Quant au palier des 10-11 ans, nous pensons qu'il doit être mis en relation non plus avec les habiletés syntaxiques, mais avec les nouvelles habiletés textuelles qui apparaissent vers cet âge, et en particulier la capacité à résumé, donc à extraire les informations pertinentes à partir d'un ensemble de faits, capacité qui requiert l'existence d'une représentation globale de l'histoire (Fayol, 1985). Le sujet capable de se construire une vision globale intégrant les principaux événements produit des récits plus courts assortis de commentaires (Colletta, Pellenq & Guidetti, 2010), et nous faisons l'hypothèse que cela se traduit au niveau de la planification locale par ce soudain allongement des GP assorti d'un gain informationnel consistant. Pour le vérifier, il nous faudra compléter nos données avec des récits collectés auprès de sujets plus âgés (12 ans et plus) et analyser plus en détail le contenu verbal des récits en même temps que l'évolution du débit de parole, lequel pourrait s'accroître plus rapidement à partir de là pour atteindre les valeurs enregistrés chez l'adulte (autour de 6 syllabes par seconde en moyenne).

Par ailleurs, pour avancer sur la question de la planification de la parole aux niveaux local et global chez l'enfant, on ne peut pas ne pas prendre en compte les indices d'énonciation pointés dans la littérature que sont les hésitations et les pauses d'une part (Campione & Véronis, 2004), et d'autre part, la gestualité coverbale dont on sait qu'elle fait partie intégrante du processus de production des énoncés (Colletta, Pellenq & Guideti, 2010 ; Mayberry & Jacques, 2000). Mais ceci est un autre travail.

Références

ADAMS, A.M., GATHERCOLE, S.E.. (1995). Phonological working memory and speech production in young children. *Journ. of Speech, Language, and Hearing Research*, 38 : 493-514.

CAMPIONE, E., VERONIS, J. (2004). Pauses et hésitations en français spontané. *JEP 2004. Actes des XXVe J.E.P., Fès, Maroc, 19-22 avril, 2004*.

COLLETTA J.-M. (2004). *Le développement de la parole chez l'enfant âgé de 6 à 11 ans*. *Corps, langage et cognition*. Hayen (Belgique), Mardaga.

COLLETTA J.-M., PELLENQ C et GUIDETTI M. (2010) : Age-related changes in co-speech gesture and narrative: Evidence from French children and adults. *Speech Communication*, 52 : 565-576.

COLLETTA, J.-M., PELLENQ, C., ROUSSET, I. (2008). Evolution du débit de parole chez l'enfant francophone dans des tâches narrative et conversationnelle. *XXVIIèmes J.E.P., Avignon, 9-13 juin 2008* .

DIESSEL, H. (2004). *The acquisition of complex sentences*. Cambridge Studies in Linguistics, 105.

FAYOL, M. (1985). *Le récit et sa construction, une approche de psychologie cognitive.* Neuchâtel, Delachaux et Niestlé.

FAYOL, M. (2000). Comprendre et produire des textes écrits : l'exemple du récit. In M. Kail & M. Fayol, *L'acquisition du langage T2*. Paris, P.U.F., p. 183-213.
FERGUSON, A.N., BOWEY, J.A., TILLEY, A. (2002). The association between auditory memory span and speech rate in children from kindergarten to sixth grade. *Journ. of Exp. Child Psychology*, 81 : 141-156.
HALLIDAY, M.A.K., HASAN, R. (1976). *Cohesion in English*. London, Longman.
HICKMANN, M. (2003). *Children's Discourse: Person, Space and Time across Languages*. Cambridge, C.U.P.
HULME, C., THOMSON, N., MUIR, C., LAWRENCE, A. (1984). Speech rate and the development of short-term memory span. *Journ. of Exp. Child Psychology*, 38 : 241-253.
KARMILOFF-SMITH, A. (1979). *A functional approach to child language*. Cambridge, C.U.P.
KHOMSI, A. (2001). *Evaluation du Langage Oral*. E.C.P.A., Paris.
KONOPCZYNSKI, G. VINTER, S. (1994). Acquisition du rythme français : comparaison enfant entendant/enfant sourd. *Cahiers d'Acquisition et de Pathologie du Langage*, 11, 1 : 119-134.
KOOPMANS-van BEINUM, F.J. (1993). Cyclic effects of infant speech perception, early sound production, and maternal speech. *IFA Proceedings*, 17 : 65-78.
KOWAL, S., O'CONNELL, D.G., SABIN, E.J. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journ. of Psycholing. Research*, 4 : 195-207.
LEGENDRE, C., VAISSIERE, J., PRANG, I., DE LAMAZE, A., GAILLARD, D., GARABEDIAN, N., LOUNDON, N. (2012). Étude du débit de parole et du débit d'articulation d'enfants normo-entendants et d'enfants sourds porteurs d'un implant cochléaire unilatéral. *A.F.O.R.L.*, 129, 4S : A52.
LEON, P. (2011). *Phonétisme et prononciations du français*. Paris, Armand Colin.
MAYBERRY, R.I., JACQUES, J. (2000). Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. In McNeill, D., *Language and gesture*. Cambridge, C.U.P.
OS (den), E. (1990). Development of temporal properties in the speech of one child between one and three years of age. *Proceedings of the Institute of Phonetic Sciences*, Amsterdam, 14 : 39-52.
PAVAO MARTINS, I., VIEIRA, R., LOUREIRO, C., SANTOS, M.E. (2007). Speech rate and fluency in children and adolescents. *Child Neuropsychology*, 13, 4 : 319-332.
RYAN, B.P. (2000). Speaking rate, conversational speech acts, interruption, and linguistic complexity of 20 pre-school stuttering and non-stuttering children and their mothers. *Clin. Ling.& Phon.*, 14, 1: 25-51.
TOLCHINSKY L. (2004). The nature and scope of later language development. In R.A. Berman (Ed.), *Language Development across Childhood and Adolescence*, Amsterdam, John Benjamins, p. 233-247.
WALKER, J.F., ARCHIBALD, L.M.D. (2006). Articulation rate in preschool children: A three years longitudinal study. *Intern. Journ. of Language and Communication Disorders*, 41, 5 : 541-565.

Modélisation de l'intonation interlangue : le cas des questions

Nicolas Ballier[1], Elisabeth Delais-Roussarie[2], Sophie Herment[3], Anne Tortel[4]
(1) Université Paris-Diderot, CLILLAC – ARP (EA 3967), France
(2) UMR 7110-Laboratoire de Linguistique Formelle, Université Paris-Diderot, France
(3) Aix-Marseille Université, UMR 7309 - Laboratoire Parole et Langage, France
(4) Université Nice-Sophia Antipolis, UMR 7320 - Bases, Corpus, Langage, France
nicolas.ballier@univ-paris-diderot.fr,
elisabeth.roussarie@wanadoo.fr, sophie.herment@univ-amu.fr,
atortel@unice.fr

RESUME

Cet article propose d'étudier les patrons prosodiques observés dans la production d'apprenants afin d'évaluer ce qui contraint les déviances : des différences systémiques entre la L1 des apprenants et la L2, des différences d'implémentation phonétique, etc. L'analyse consiste en une étude croisée de l'intonation des questions polaires en français, en anglais et en anglais L2. Cinq questions polaires ont été extraites du corpus AixOx, ensemble de 40 textes qui ont été lus par 10 locuteurs natifs de français, 10 locuteurs natifs d'anglais et 20 apprenants francophones d'anglais (10 de niveau indépendant et 10 de niveau expérimenté). L'analyse des données montre que les différences entre natifs et non natifs n'affectent pas la forme du contour nucléaire, mais que les apprenants francophones d'anglais tendent à assigner un mouvement montant à la fin des mots prosodiques, ce qui engendre une claire différence de rythme.

ABSTRACT

In this paper, we study the intonational patterns observed in learners' productions in order to evaluate what motivates the deviations observed: systemic differences between the learners' L1 and the L2, differences in phonetic implementation, etc. The analysis consists of a cross-comparison of the intonation of yes-no questions in French, English and English as an L2. It is based on five information-seeking yes-no questions that were extracted from the AixOx corpus, which contains a set of 40 texts that were read by 10 native French speakers, 10 Native English speakers and 20 French learners of English. The analysis of the data showed that the differences between native and non-native speakers do not affect the form of the nuclear contour. It mostly shows that French speakers of English have a tendency to assign a rising pitch movement at the end of prosodic words, which leads to a clear difference in rhythm.

MOTS-CLES : acquisition de la prosodie L2, intonation, phrasé prosodique, rythme, modélisation prosodique, corpus d'apprenants, interphonologies.
KEYWORDS: acquisition of prosody in L2, intonation, prosodic phrasing, rhythm, prosodic modelling, learner corpora, interphonologies.

1 Introduction

Les travaux antérieurs sur l'intonation interlangue ont montré que les patrons intonatifs observés dans les productions des apprenants sont souvent influencés par la L1 (cf., entre autres, Mennen, 2004 ; Rasier & Hiligsmann, 2007). La notion de transfert de la L1 vers la L2 est donc souvent évoquée pour rendre compte des productions. Cependant, Mennen (2007) précise que le transfert peut avoir lieu au niveau phonologique aussi bien que phonétique. Les transferts qui ont lieu au niveau phonologique résultent de différences dans la structure métrique ou dans l'inventaire tonal. Dans leur étude sur les questions tags en anglais, Ramírez & Romero (2005) montrent par exemple que les apprenants hispanophones utilisent des montées à la fin des questions tags de demande de confirmation en anglais, là où les natifs réalisent des chutes, interprétant donc ces patrons prosodiques comme le résultat d'un transfert phonologique. *A contrario*, les transferts au niveau phonétique ont lieu lorsqu'une forme phonologique identique diffère dans la façon dont elle est implémentée phonétiquement dans les deux langues. Des différences d'alignement temporel des accents mélodiques représentent un cas de transfert phonétique (voir Mennen, 2004 pour des exemples concrets).

Distinguer les différents types de transfert ou les diverses déviances est d'un grand intérêt dans l'étude de l'intonation interlangue, ainsi que l'explique Mennen (à paraître). Dans cet article, nous montrons cependant que la classification des déviances observées est loin d'être aisée car le type de déviance peut évoluer dans le temps et qu'une simple comparaison des formes de surface peut se révéler insuffisante. Notre analyse est fondée sur l'étude de questions polaires de demande d'information réalisées par des apprenants francophones d'anglais et extraites du corpus AixOx (Herment *et al.*, 2012 et à paraître). L'article est organisé comme suit : les caractéristiques prosodiques des questions polaires de demande d'information sont décrites dans la deuxième section ; la troisième section présente les données et la méthodologie employée pour l'analyse prosodique ; les résultats sont détaillés en section 4, puis discutés en section 5.

2 Intonation des questions polaires

2.1 Questions polaires en français

D'un point de vue morphosyntaxique, on peut employer en français trois constructions distinctes pour former des questions polaires : une structure déclarative similaire à celle observée dans les phrases assertives (1); une inversion sujet-verbe, que le sujet soit nominal ou pronominal (2); et la particule interrogative *est-ce que* peut être insérée en position initiale de phrase, le reste de l'énoncé étant comparable à ce que l'on a dans les assertions (3).

(1) *Vous avez appris des langues étrangères ?*
(2) *Pierre est-il venu ?*
(3) *Est-ce que c'est vrai ?*

En ce qui concerne l'intonation, les contours montants représentent la forme canonique associée aux questions déclaratives (cf., entre autres, Delattre, 1966). Par contre, pour les questions dans lesquelles la modalité est indiquée par un marqueur morphosyntaxique ou lexical (cf. exemples (2) et (3)), on trouve aussi bien des contours non-montants que montants (cf., entre autres, Delattre, 1966). Notons cependant que le contour montant est de loin le plus fréquent dans les questions polaires de demande d'information, toutes constructions

confondues (cf., entre autres, Santiago-Vargas & Delais-Roussarie, 2012).

2.2 Questions polaires en anglais

Dans les questions polaires en anglais, la modalité de la phrase est toujours indiquée par la morphosyntaxe: soit par l'inversion auxiliaire-sujet (4), soit par l'insertion de l'auxiliaire *do* (5). On trouve aussi en anglais des questions déclaratives, mais ce sont généralement des questions-écho, et elles ne sont pas représentées dans notre corpus.

(4) *Is Peter coming?* (*vs. Peter is coming.*)
(5) *Does he live in Paris?* (*vs. He lives in Paris.*)

Contrairement aux questions en *wh-*, pour lesquelles le contour par défaut est descendant en anglais, les questions polaires (*yes-no*) sont réalisées avec un contour montant, même si on trouve aussi parfois un schéma descendant, mais moins fréquemment (cf., entre autres, Cruttenden, 1997 ; Wells, 2006).

3 Corpus et méthodes

3.1 Corpus

3.1.1 Le corpus AixOx

Le corpus AixOx (Herment et al 2012) est un corpus d'apprenants multilingue qui comprend quarante enregistrements de passages de une minute lus en anglais et en français par des locuteurs natifs et des apprenants. Les passages lus, tirés du corpus EUROM1 (Chan *et al.* 1995), correspondent à de petites scènes contextualisées comparables dans chaque langue, mais qui ne constituent pas des traductions mot à mot. Les apprenants ont été divisés en deux groupes, B et C, selon leur niveau, d'après le Cadre Européen Commun de Référence des Langues (CECRL). Les locuteurs de niveau B sont autonomes dans leur expression (niveau B1 et B2 du CECRL) et les locuteurs du groupe C sont des locuteurs expérimentés (C1/C2). Le corpus est réparti en six groupes de locuteurs (tableau 1 ci-dessous). Pour chaque groupe, dix locuteurs (cinq femmes et cinq hommes) ont été enregistrés, le corpus comptant ainsi soixante locuteurs et environ trente heures de parole.

Langue	Locuteurs natifs	Apprenants L2 B1/B2	Apprenants L2 C1/C2
Anglais	ENEN	FRENB	FRENC
Français	FRFR	ENFRB	ENFRC

TABLEAU 1 – Groupes des locuteurs dans AixOx

3.1.2 Les questions extraites du corpus

Avant d'extraire les questions des passages pseudo-dialogués du corpus EUROM1, nous avons dû prendre en compte plusieurs facteurs :

- Les corpus anglais et français ne sont pas des traductions mutuelles, même s'ils sont proches puisque relevant de situations comparables, et le corpus anglais comprend 23 questions, là où le français n'en compte que 22 ;

- Les questions polaires tout comme les questions ouvertes en *Qu-* sont représentées dans le corpus ;

- Dans le corpus français, différents types de constructions sont utilisés pour les questions polaires, alors que dans le corpus anglais, elles sont toutes réalisées avec une inversion auxiliaire/sujet ;

- Dans les deux langues, certaines questions ne sont pas neutres, dans la mesure où il peut s'agir de questions rhétoriques que l'on s'adresse à soi-même (*'What will 1992 really mean to the person in the street?'*), d'ordres déguisés (*'Can you give me a firm date now?'*) ou de questions totales ayant la signification de questions partielles (*'Can you tell me what's on television tonight?'*).

Pour mener notre analyse comparée des questions polaires, nous avons pris en considération les différents types de questions, sur le plan pragmatique aussi bien que syntaxique. Nous nous sommes concentrés sur cinq questions polaires de demande d'information où la modalité de l'énoncé est indiquée dans les deux langues par des procédés morpho-syntaxiques. Elles figurent ci-dessous :

- Questions du corpus en français

Q1 Est-ce que vous pourriez me donner leur nouveau numéro de téléphone ?
Q2 Est-ce que c'est vrai ?
Q3 Est-ce que vous pourriez me donner la liste des restaurants de mon quartier ?
Q4 Est-ce que vous avez des tarifs spéciaux pour les collectivités ?
Q5 Est-ce qu'un organisme universitaire peut en bénéficier ?

- Questions du corpus en anglais

Q1 Can you give me their new number please ?
Q2 Could you please tell me the best connections to Sheffield from East Greenstead?
Q3 Do you take reservations by telephone?
Q4 Can you give me a list of the restaurants in the neighbourhood?
Q5 Do you have special corporate academic institutions?

3.2 Méthodes pour l'analyse prosodique

Afin de comparer l'intonation des questions réalisées par les locuteurs natifs et par les apprenants de niveau B et C, nous avons eu recours à deux approches distinctes pour coder les schémas intonatifs réalisés dans les différents énoncés : une méthode perceptive et une méthode semi-automatique.

3.2.1 L'approche perceptive

A la suite de (Halliday, 1967), et suivant la tradition britannique, notre analyse perceptive se fonde sur l'analyse en tonalité (la division en syntagmes intonatifs), en tonicité (le placement des syllabes nucléaires) et en contours tonals (les mouvements mélodiques distinctifs). A la manière de (Grabe et al., 2001), nous considérons qu'il n'y a qu'un seul niveau de frontière, qui est associé au syntagme intonatif (IP). Dans le cas des questions courtes, un seul IP sera réalisé, mais des questions plus longues telles que *Could you please tell me the best connections to Sheffield from East Greenstead ?* peuvent être subdivisées en deux IP par les natifs, et même davantage par les apprenants. Les frontières d'IP sont indiquées par une barre oblique.. La place de la syllabe tonique ou nucléaire est étudiée. Le principe adopté (cf., entre autres, Halliday 2008) est qu'il n'y a qu'une seule syllabe tonique ou nucléaire par IP : c'est celle qui porte le contour (le mouvement mélodique distinctif) de l'IP. La syllabe nucléaire est soulignée dans nos représentations. Enfin, les contours ont été codés. Le contour est le mouvement mélodique

distinctif, celui qui porte sur la syllabe nucléaire (ou qui en part) et se prolonge sur les syllabes post-nucléaires. Nous avons utilisé le codage suivant : F pour une chute, R pour une montée simple, HF pour une grande chute et FR pour un ton creusé, suivant en cela la tradition britannique, mais dans le cadre d'un inventaire des tons plus restreint (Gussenhoven 1984, Hirst 2005).

Les questions sont donc codées comme sous (6), (7) et (8), qui représentent respectivement un locuteur natif, un apprenant du groupe B et un du groupe C :

(6) *Can you give me their new num*ber, please F /
(7) *Can you give me F / their new number, please FR /*
(8) *Can you give me their new number, please R /*

Afin de permettre la comparaison, les questions du français ont été codées de la même manière, comme en (9), (10) et (11), qui représentent respectivement un locuteur natif, un apprenant du groupe B et un du groupe C. C'est le contour global de l'IP qui a été pris en compte, et non les contours des groupes accentuels.

(9) *Est-ce que vous pourriez me donner leur nouveau numéro de téléphone R /*
(10) *Est-ce que vous pourriez me donner F / leur nouveau numéro de téléphone R /*
(11) *Est-ce que vous pourriez me donner leur nouveau numéro de téléphone R /*

3.2.2 L'approche semi-automatique

Nous avons également procédé à une analyse semi-automatique. Les questions étudiées ont été transcrites sous Praat, alignées avec SPPAS (Bigi, 2012) et syllabifiées manuellement. La tire syllabique, encodée en SAMPA, a été utilisée pour l'extraction automatique des paramètres acoustiques à l'aide du script ProsodyPro (Xu, 2013). Les résultats obtenus de l'extraction ont ensuite permis de mieux appréhender et visualiser les points de convergences et de divergences entre groupes de locuteurs.

4 Résultats

4.1 L'intonation des questions totales de l'anglais L1 revisitée

Les résultats de l'analyse, présentés dans le tableau 2, montrent clairement que le contour montant n'est pas le plus fréquemment observé en anglais L1 (voir colonne 2), contrairement à ce qui est dit dans la littérature : 70 % des 50 questions analysées ont en effet été réalisées avec un contour descendant. En français (colonne 3), au contraire, les réalisations observées vont dans le sens de ce qui est dit dans la littérature : le contour montant est la forme la plus fréquemment observée pour les phrases de notre corpus.

4.2 L'intonation en anglais L2: contours

Les contours terminaux observés dans les productions des apprenants sont présentés dans les deux colonnes de droite du tableau 2. Les apprenants des groupes B et C utilisent très fréquemment un contour montant à la fin des questions, comme dans leur langue maternelle, mais cette tendance diminue à mesure que le niveau des apprenants augmente : 44% des questions sont réalisées avec un contour descendant chez les locuteurs de niveau C, ce qui est important comparé aux 20 % observés dans leur L1. Ces différences entre les locuteurs natifs et les apprenants apparaissent clairement si on compare les interpolations statistiques faites par R

à partir des valeurs moyennes de F0 estimées pour chaque syllabe lors des analyses semi-automatiques sous ProsodyPro (voir fig. 1). Il est possible d'avoir avec ce programme des représentations assez fines, basées sur les mesures prises sur 10 points successifs par syllabe. Celles-ci ont le grand avantage de neutraliser les différences de durées : les dix points sont en effet localisés tous les 10èmes de syllabe, et les valeurs qui leur sont associées peuvent être comparées entre plusieurs locuteurs.

Questions	English natives	French natives	FRENB	FRENC
Q1	F 80%	R 100%	R 90%	R 60%
Q2	F 100%	R 90%	R 80%	R 70%
Q3	R 70%	F60%	R 90%	R 60%
Q4	F 60%	R 90%	R 80%	F 60%
Q5	F 80%	R 80%	R 90%	R 80%
Total	**F 70%**	**R 80%**	**R 86%**	**R 66%**

TABLEAU 2. Contours pour les natifs anglophones et francophones et pour les apprenants francophones d'anglais niveaux B (FRENB) et C (FRENC).

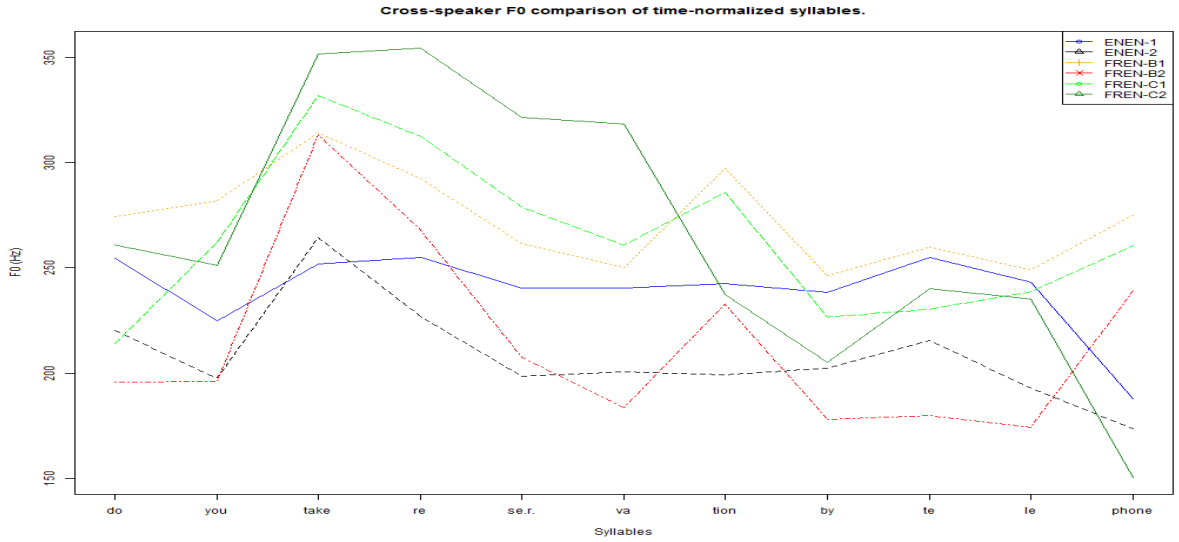


FIGURE 1. Visualisation des formes intonatives estimées pour l'énoncé '*Do you take the reservations by telephone*' pour deux natifs, deux apprenants de niveau B, et deux de niveau C.

Dans la figure 1, comme dans les autres visualisations générées à partir de l'analyse avec ProsdyPro, le nombre des locuteurs pris en considération a été limité à 6 pour une meilleure visibilité (deux anglophones natifs, deux apprenants de niveau B, et deux de niveau C).

4.3 L'intonation en anglais L2: les découpages prosodiques

La forme du contour nucléaire terminal n'est pas le trait le plus pertinent de la production des apprenants. Les découpages prosodiques sont le marqueur le plus net pour différencier les apprenants de niveau B de ceux de niveau C, et bien-sûr des natifs. De fait, les apprenants autonomes (niveau B) ont tendance à découper prosodiquement les énoncés comme ils le feraient en français : ils produisent un mouvement mélodique à la fin de chaque groupe de mots correspondant à un syntagme accentuel en français, comme on le voit dans l'exemple (12).

(12) *Could you please R / tell me R / the best connections R / to Sheffield R / from East Greenstead R /*

Chez les apprenants de niveau avancé (groupe C), cette tendance à construire des groupes prosodiques comparables à ceux du français disparaît. Pour la question 2, aucun locuteur ne fait plus de deux groupes prosodiques, comme on le voit en (13).

(13) *Could you please tell me the best connections to Sheffield F / from East Greenstead R /*

Ces différences dans les découpages sont clairement visibles dans les énoncés relativement longs. Ainsi, pour l'énoncé (14), dont une représentation dérivée de l'analyse semi-automatique faite avec ProsdyPro est proposée dans la figure 2, on voit que les apprenants de niveau B ont tendance à réaliser des mouvements mélodiques à la fin des mots *me* et *list*, structurant donc différemment des natifs la séquence prénucléaire (voir courbes bleue et noire). De leur côté, les natifs réalisent une montée initiale en début d'énoncé, puis un mouvement mélodique associé à la syllabe accentuée de *RESTaurant*.

Outre cette différence dans les découpages, les apprenants de niveau B diffèrent nettement des natifs, mais aussi des apprenants de niveau C, dans la réalisation des accents toniques. Sur la Figure 2, par exemple, les locuteurs natifs ainsi que ceux de niveau C réalisent majoritairement l'accent mélodique sur la syllabe initiale de *REStaurants,* (courbes verte et jaune). Les apprenants de niveau B, quant à eux, réalisent un mouvement mélodique montant sur la dernière syllabe de *restauRANTS* (courbes bleue et noire), déplaçant ainsi l'accent tonique vers la droite, comme en français. Notons également que pour ce mot, seulement 2 syllabes sont prononcées par les natifs et les apprenants du groupe C, alors que 3 syllabes sont produites pour le groupe B, reproduisant un découpage français. L'absence de compression de la syllabe intermédiaire et le déplacement d'accent sur la dernière syllabe sont deux phénomènes typiques de la prononciation d'un apprenant français.

5 Discussion et conclusion

En ce qui concerne la forme des contours nucléaires, les anglophones natifs ont tendance à préférer à la fin des questions polaires un contour descendant, là où les apprenants utilisent surtout des contours montants. L'utilisation de cette forme peut résulter d'un transfert du français, ce qui expliquerait que le choix entre les formes montantes et descendantes tend à s'équilibrer à mesure que le niveau de l'apprenant augmente : les apprenants de niveau B utilisent des formes montantes dans plus de 85% des cas, contre 66% chez les apprenants de niveau C. Ceci étant, comme cette forme est aussi acceptable chez les natifs, bien que dans une moindre mesure, elle ne constitue pas le trait essentiel pour différencier les apprenants des natifs. Les différences les plus importantes entre les différents groupes de locuteurs (anglophones natifs et apprenants) apparaissent surtout dans la structuration et la réalisation de la séquence prénucléaire. Les natifs ont en effet tendance à réaliser un seul IP pour la séquence *'can you give me a list of the restaurants'*, avec un accent mélodique associé à la syllabe tonique de *RESTaurants*. Au contraire, les apprenants de niveau B structurent la même séquence en groupes accentuels comparables à ceux du français, et réalisent des mouvements montants qui correspondent aux accents démarcatifs du français à la fin des mots *me*, *list* et *restaurant*. Cette tendance disparaît progressivement chez les apprenants de niveau C, dont les découpages ressemblent davantage à ceux des natifs (voir figure 2).

En s'appuyant sur une analyse semi-automatique (qui peut comporter quelques limites, voir Ballier & Martin, 2013), notre étude a permis de mettre au jour une idée essentielle pour les recherches sur corpus d'apprenants : les découpages prosodiques et la réalisation syllabique

sont des priorités pour la modélisation de l'intonation interlangue. La comparaison des productions des questions polaires de demande d'information par les natifs et les apprenants montre que la forme du contour nucléaire n'est en effet que peu affectée, contrairement au rythme et aux syllabes prénucléaires, qui subissent les différences les plus importantes. Les apprenants francophones de l'anglais ont tendance à produire un ton montant sur la fin des mots prosodiques, ce qui entraîne d'importantes différences de rythme. Alors que la plupart des études sur l'intonation en L2 examine la forme nucléaire phonologique, la présente étude encourage la poursuite de travaux sur la structuration des séquences prénucléaires et sur la façon dont se fait leur implémentation phonétique.

Cette analyse de la modélisation des strates d'interlangue reste limitée puisqu'elle n'a porté que sur un nombre restreint de locuteurs et sur un seul acte de parole, la question fermée. A l'avenir, nous envisageons donc de l'étendre à un nombre plus important de locuteurs, voire à d'autres actes de parole. En outre, les résultats des études pourront être confrontés à d'autres corpus d'apprenants intégrant de la parole spontanée. Par ailleurs, la méthodologie utilisée, notamment dans la phase semi-automatique, est centrée sur les variations mélodiques. Trouver des outils capables de comparer les variations entre natifs et non-natifs dans la gestion des durées syllabiques est aussi une piste importante, d'autant que la durée est un paramètre essentiel dans la construction du rythme, notamment dans des langues comme le français.

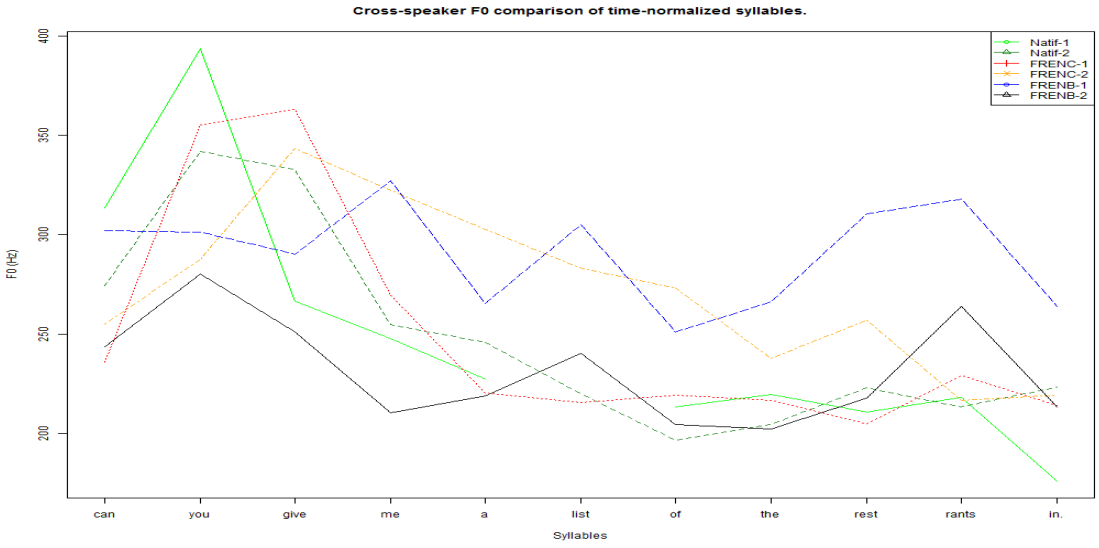


FIGURE 2: Formes intonatives estimées pour la séquence *'can you give me the list of the restaurants'* pour deux natifs, deux apprenants de niveau B, et deux de niveau C.

Remerciements

Le corpus AixOx a été enregistré dans le cadre d'un partenariat Hubert Curien Alliance entre les universités d'Aix-Marseille et d'Oxford. Les scripts pour R ont été finalisés grâce à l'aide d'Adrien Méli.

Références

BALLIER, N. & MARTIN, P. (2013). Developing corpus interoperability for phonetic investigation of learner corpora, in DÍAZ-NEGRILLO, A., BALLIER, N. & THOMPSON, P. (eds.), *Automatic Treatment and Analysis of Learner Corpus Data,* Amsterdam: Benjamins, 33–64.

BIGI, B. (2012). SPPAS: a tool for the phonetic segmentation of speech, *Proceedings of the Language Resource and Evaluation Conference*, Istambul, Turkey, 1748-1755.

CHAN, D., FOURCIN, A., GIBBON, D., GRANDSTRÖM, B., HUCKVALE, M., KOKKINAKIS, G., KVALE, K., LAMEL, L., LINDERG, B., MORENO, A., MOUROPOULOS, J., SENIA, F., TRANSCOSO, I., VELT, C. & ZEILIGER, J. (1995). EUROM – A Spoken Language Resource for the EU, *Proceedings of Eurospeech* '95(1):867-880.

CRUTTENDEN, A. (1997). *Intonation*, Cambridge: Cambridge University Press, 2nd ed. (1st ed. 1986).

DELATTRE, P. (1966). Les Dix Intonations de base du français, *The French Review* 40 (1):1-14.

GRABE, E., POST, B. & NOLAN, F. (2001). Modelling intonational variation in English. The IViE system, in Puppel, S. & G. Demenko [eds.], *Proceedings of Prosody 2000*, Adam Mickiewicz University, Poznan, Pologne.

GUSSENHOVEN, C. (1984). *On the grammar and semantics of sentence accent*, Dordrecht : Foris.

HALLIDAY, M.A.K. (1967). *Intonation and Grammar in British English*, The Hague-Paris: Mouton.

HALLIDAY, M.A.K. & GREAVES, W.S. (2008). *Intonation in the Grammar of English*, London-Oakville : Equinox.

HERMENT, S., LOUKINA, A. & TORTEL, A. (2012). The AixOx corpus, SLDR. http://sldr.org/sldr000784/fr

HIRST, D. (2005). Form and function in the representation of speech prosody, in HIROSE, K., HIRST, D. & SAGISAKA, Y. (eds.) Quantitative prosody modeling for natural speech description and generation *Speech Communication* 46 (3-4):334-347.

MENNEN, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek, *Journal of Phonetics* 32:543–563.

MENNEN, I. (2007). Phonological and phonetic influences in non-native intonation, in J. Trouvain & U. Gut (eds.), *Non-native Prosody: Phonetic Descriptions and Teaching Practice*, 53–76, Mouton De Gruyter.

MENNEN, I. (à paraître). Beyond segments: towards an L2 intonation learning theory (LILT), in DELAIS-ROUSSARIE, E., AVANZI, M. & HERMENT, S. (eds.), *Prosody and languages in contact: L2 acquisition, attrition, languages in multilingual situations*, Springer Verlag, accepté.

RAMIREZ, D. & ROMERO, J. (2005). The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers, *Intercultural Pragmatics* 2 (2):151–168.

RASIER, L., & HILIGSMANN, P. (2007). Prosodic transfer from L1 to L2. Theoretical and methodological issues, *Nouveaux cahiers de linguistique française* 28:41–66.

SANTIAGO-VARGAS F. & DELAIS-ROUSSARIE, E. (2012). Acquiring phrasing and intonation in French as a second Language: the case of Yes-No questions produced by Mexican Spanish Learners, *Proceedings of Speech Prosody*, Shanghai, China.

WELLS, JC. (2006). *English Intonation, an Introduction*, Cambridge: Cambridge University Press.

XU, Y. (2013). ProsodyPro, A Tool for Large-scale Systematic Prosody Analysis, *Proceedings of the TRASP conference*, Aix-en-Provence.

Contribution à l'étude physiologique des possibilités articulatoires dans la cavité pharyngale

Chakir Zeroual[1,2], John H. Esling[3], Lise Crevier-Buchman[2,4]
(1) Faculté Polydisciplinaire de Taza, Maroc
(2) Laboratoire de Phonétique et Phonologie CNRS-UMR7018 et Sorbonne Nouvelle, Paris, France.
(3) University of Victoria, Victoria, Canada
(4) Unité d'exploration de la Voix, de la Parole, et de la Déglutition. Hôpital Européen Georges Pompidou (HEGP), Paris, France
chakirzeroual@yahoo.fr esling@uvic.ca lise.crevier@univ-paris3.fr

RESUME

Nous présentons des productions enregistrées par naso-endoscopie des consonnes postérieures de l'arabe marocain (AM) faites par un locuteur natif qui est également un phonéticien entraîné. Des constrictions pouvant correspondre à des linguo-pharyngales (ou upper-pharyngeals) fricatives et occlusives sont également présentées pour vérifier l'hypothèse de Maddieson (2009) selon laquelle ces dernières seraient possibles. D'après nos observations, l'AM a des consonnes épiglottales ayant une constriction épiglotto-pharyngale (ou lower-pharyngeal) et une autre aryépiglottale, le larynx étant élevé. Par contre, les articulations linguo-pharyngales occlusives et non-occlusives montrent une constriction linguo-pharyngale partielle, une compression du sphincter aryépiglottique et une position abaissée du larynx ; la glotte est fermée durant l'occlusive. Nos données semblent confirmer la prédiction d'Esling (1996) selon laquelle les articulations « pharyngales » possèdent une constriction aryépiglottique ; l'opposition entre les pharyngales et les épiglottales est contrôlée par la section du sphincter aryépiglottique et la position du larynx.

ABSTRACT

Contribution to the study of pharyngeal cavity articulations

In this study we present productions recorded by naso-endoscopy of Moroccan Arabic back consonants pronounced by a trained Moroccan Arabic phonetician. We also provide naso-endoscopic recordings of non-attested linguo-pharyngeal constrictions to test Maddieson's (2009) hypothesis according to which (upper)-pharyngeal fricative and plosive consonants are possible. We observe that Moroccan Arabic has epiglottal consonants produced with epiglotto-pharyngeal (lower-pharyngeal) and aryepiglottic constrictions, while the larynx is raised. The attempts to produce fricatives and plosive upper-pharyngeals show a linguo-pharyngeal partial constriction, aryepiglottic compression, and a lowered position of the larynx. The glottis is closed during the plosive. Our data seem to confirm Esling's prediction (1996) that the "pharyngeal" articulations are usually combined with aryepiglottic constriction, and that the pharyngeal versus epiglottal contrast is controlled by the posture of the aryepiglottic sphincter and the height setting of larynx.

MOTS-CLES : Articulations pharyngales, naso-endoscopie, modes phonatoires, arabe.
KEYWORDS : Pharyngeal articulations, naso-endoscopy, phonation types, Arabic.

1 Introduction

La caractérisation du lieu et du mode des consonnes pharyngales et épiglottales soulève encore plusieurs questions aussi bien phonétiques que phonologiques. Ce travail teste des prédictions faites par des modèles typologiques proposés pour ce type d'articulation (Catford, 1968, 1977 ; Ladefoged et Madieson, 1996 ; Esling, 1966, 1999). Pour ce faire, nous nous sommes basés sur des définitions phonétiques canoniques et d'autres spécialisées données à ces articulations, ainsi que sur l'analyse de nos données physiologiques collectées récemment pour les besoins de cette étude.

Dans les définitions canoniques, les consonnes pharyngales, notées /ħ ʕ/ par API, sont produites par le recul de la racine de langue vers la paroi arrière de la cavité oro-pharyngale. Cette définition est adoptée par Laver (1994) et Ladefoged et Maddieson (1996) ; ces derniers considèrent /ħ ʕ/ comme des « *upper pharyngeal* ». Catford (1977) qualifie ce type d'articulation de « linguo-pharyngal » et le distingue des « *faucal or transverse pharyngeals* » réalisées grâce uniquement à une compression latéral des « faucal pillars ». Ladefoged et Maddieson (1996) ajoutent que les articulations linguo-pharyngales sont souvent secondaires (exemples des voyelles et des consonnes pharyngalisées) ; les fricatives pharyngales sont très rares et les occlusives (upper) pharyngales sont impossibles. Pour Catford (1977), les pharyngales de l'arabe seraient des « *faucal or transverse pharyngeals* », mais cette affirmation a été infirmée par plusieurs études (voir plus bas).

Ces mêmes définitions canoniques considèrent que les consonnes épiglottales, notées /ʜ ʢ ʔ/ par IPA, sont réalisées dans la partie inférieure du pharynx d'où leur appellation de « *lower pharyngeals* » (Ladefoged et Maddieson, 1996) ou « *deep pharyngeals* » (Catford, 1977). Elles impliquent un recul de l'épiglotte qui aboutit à deux articulations, une épiglotto-pharyngale (entre le sommet de l'épiglotte et la paroi arrière de la cavité laryngo-pharyngale) et une autre aryépiglottale (entre le sommet des aryténoïdes, les cordes aryépiglottiques et la base de l'épiglotte). Des divergences apparaissent entre les auteurs quant à la question de savoir si la constriction principale est épiglotto-pharyngale ou ary-épiglottale. Pour Catford (1968, 1977), une articulation fricative, approximante ou vibrante épiglotto-pharyngale est possible ; par contre une occlusion totale dans ce même lieu d'articulation serait impossible. Ses analyses suggèrent que l'articulation principale des non-occlusives /ʜ ʢ/ serait épiglotto-pharyngale, et celle de l'occlusive /ʔ/ serait aryépiglottale.

Pour Esling (1996), toutes les consonnes /ħ ʕ ʜ ʢ ʔ/, qu'il qualifie de pharyngales, possèdent une constriction partielle aryépiglottale. Cette dernière est plus marquée durant /ʜ ʢ/ qui sont également accompagnées d'une vibration des aryténoïdes contre la base de l'épiglotte. /ʔ/ serait le résultat d'une constriction aryépiglottale totale. Le modèle d'Esling prédit également que les consonnes produites avec une constriction dans la cavité pharyngale seraient accompagnées d'une compression du sphincter aryépiglottique. Esling (1998) souligne que /ħ ʕ ʜ ʢ ʔ/ sont généralement produites avec une position plus élevée du larynx, mais leur production avec une position plus basse du larynx serait possible. Il n'exclue pas l'hypothèse que « phonetic distinctions which have been observed in the literature between "pharyngeals" and "epiglottals" are more likely due to the setting of the larynx height parameter »(Esling, 1998).

Dans ses travaux plus recents, Maddieson (2009) mentionne la possibilité de produire des constrictions au milieu du pharynx (linguo-pharyngale) avec des modes d'articulation différents : "it certainly seems possible to retract the tongue root sufficiently to close the pharynx in the middle, and pharyngealized uvular stops may well have an extended tongue contact reaching lower than the level of the uvula". Honjow et Isshiki (1971) ont montré que leur sujet, présentant des troubles d'articulation, produit [k] au voisinage de [a] et [u] avec une occlusion au niveau oro-pharyngale. Cependant, leurs observations ciné-radiographiques n'excluent pas la présence d'une autre occlusion au niveau aryépiglottale et/ou glottique comme cela est suggéré par les modèles d'Esling (1996, 1999) et de Catford (1968, 1977).

Des observations physiologiques semblent montrer qu'en arabe marocain, les consonnes généralement décrites comme des pharyngales et notées /ħ ʕ/ sont en fait des épiglottales ou plus exactement des aryépiglottales (Zeroual, 1999 ; Zeroual et al, 2011) que nous représentons ici par /ʜ ʢ/. En effet, des arguments aérodynamiques (Yeou et Maeda, 1995 ; Zeroual, 2003) ainsi qu'acoustiques (Yeou et Maeda, 1995) appuient l'hypothèse selon laquelle l'articulation principale de ces consonnes seraient située au niveau aryépiglottal.

D'autres données physiologiques montrent que les consonnes coronales de l'arabe marocain généralement appelées emphatiques et notées /S T D/ possèdent, en plus de leur articulation principale coronale, une autre secondaire produite par le recul de la langue au niveau oro-pharyngal. Cette articulation semble présentée des caractéristiques d'une articulation linguo-pharyngale, elle est généralement considérée comme une pharyngalisation. Cependant cette articulation n'est pas accompagnée d'une compression antérieure-postérieure de la cavité supraglottique (Zeroual, 1999 ; Zeroual et al, 2011) comme cela est prédit par Esling (1996).

Dans cette étude, nous présentons des données physiologiques enregistrées par fibroscopie sur des consonnes postérieures attestées en arabe marocain et produites par un locuteur natif qui est également un phonéticien entraîné. D'autres productions faites par ce même locuteur d'articulations qui correspondraient à des consonnes occlusives et non-occlusives *linguo-* ou *upper-pharyngeals* sont également décrites pour tester les prédictions de Maddieson (2009) selon lesquelles de telles articulations seraient possibles.

2 Méthode & matériel linguistique

Durant un enregistrement par naso-endoscopie à *l'Unité d'exploration de la Voix, de la Parole, et de la Déglutition. Hôpital Européen Georges Pompidou (HEGP)*, Paris, un nasofibroscope flexible (Kay-Pentax ® FNL10RP3, Japan ; doté d'une lumière froide Xenon 300 W) a été introduit par un médecin (laryngologiste), sans anesthésie, à travers les fausses nasales d'un locuteur adulte marocain (44 ans). Le bout externe de l'endoscope a été relié à un système d'enregistrement vidéo (DigitalStrobe ®, RLS91000, Kay Elemetrics, USA). Son bout interne a été placé dans deux positions différentes : (Position 1) au niveau oro-pharyngal derrière la luette pour observer les mouvements de la racine de la langue, des parois pharyngales arrières et latérales ainsi que de l'épiglotte ; (Position 2) au niveau laryngo-pharyngal derrière le sommet de l'épiglotte pour observer les

différentes postures de la cavité glottique et supra-glottique (figure 1).

Puisque les mesures endoscopiques sont non calibrées, nous avons réparti nos items en quatre listes. Chaque liste a été enregistrée durant une même session (sans pause). La liste 1 contenait les consonnes /s t k χ q h ʜ ʕ/ de l'arabe marocain prononcées de manière isolée, puis dans les contextes iCi (mots et non-mots, liste 2) et aCa (liste 3). La liste 4 regroupait /ʜ ʕ/ suivies par des productions qui pourraient correspondre à des linguo-pharyngales (ou upper-pharyngeals) occlusives et non-occlusives. Chaque liste a été répétée plusieurs fois en variant à chaque fois l'ordre de ses items prononcés sans phrase cadre et en plaçant le naso-endoscope d'abord au niveau oro-pharyngal, puis laryngo-pharyngal. Durant chaque session, la même intensité de la lumière a été gardée pour pouvoir comparer les items entre eux. En effet, le degré de luminosité d'une partie éclairée informe sur sa position verticale (Kagaya, 1971) : une luminosité plus intense implique une position plus proche de l'endoscope.

Le dispositif d'enregistrement nous a permis d'obtenir des fichiers séparés non compressés aux formats *.avi (images + son) ainsi que *.wav (audio). Les fichiers vidéos ont été segmentés et traités image par image grâce au logiciel AdobePremiere Pro CS5 et les fichiers audio au programme Praat.

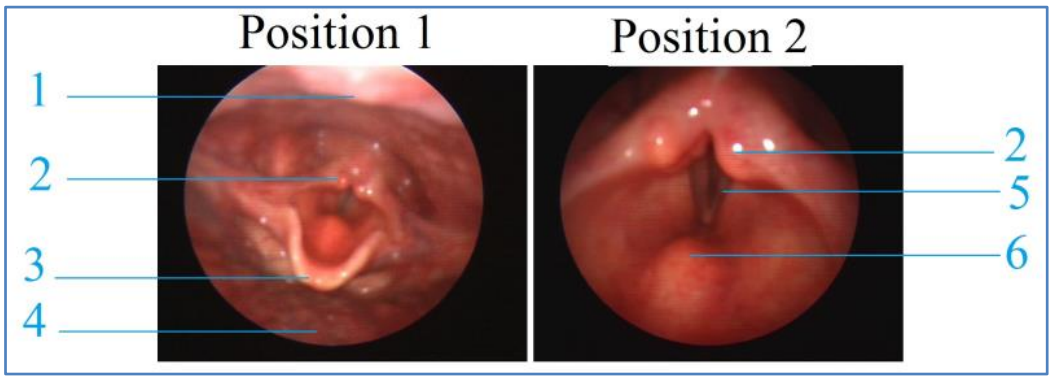


FIGURE 1 : Images issues des observations par nasofibroscopie. Position 1 : nasofibroscope derrière la luette. Position 2 : nasofibroscope derrière l'épiglotte. 1. Paroi pharyngale postérieure. 2. Aryténoïdes. 3. Sommet de l'épiglotte. 4. Partie postérieure de la langue. 5. Corde vocale. 6. Base de l'épiglotte.

3 Résultats et discussion

3.1 Les consonnes orales et glottales /s t h/

Durant la production des consonnes tenues /s t h/ (figure 2), l'épiglotte et la racine de la langue restent éloignées de la paroi pharyngale arrière. Dans /iCi/, la racine de la langue et notamment l'épiglotte apparaissent encore plus éloignées du pharynx due à la coarticulation avec les voyelles adjacentes /i/. Les observations au niveau de la cavité laryngée dans /iCi/ (figure 2) laissent apparaître une abduction totale (parties antérieure et postérieure) des cordes vocales durant /s t/ en accord avec la posture canonique des segments non-voisés. Par contre, dans /iCi/(figure 2), une abduction uniquement de la partie postérieure de la glotte est enregistrée durant /h/ due à sa réalisation, dans ce contexte, avec le mode phonatoire breathy-voice ; une posture glottique similaire a été également rapportée durant ce type de mode phonatoire (Ladefoged et Maddieson, 1996).

3.2 Les consonnes emphatiques et uvulaires /S T χ q/

Les emphatiques /S T/ se réalisent avec un recul de la partie postérieure de la langue qui se colle à l'épiglotte ; cette dernière reste assez écartée de la paroi pharyngale arrière (figure 2). Une rétraction relativement similaire est observée durant /χ/ et notamment /q/ où une partie postérieure plus élevée du corps de la langue est impliquée. La posture de la glotte durant /S/ semble très similaire à celle observée durant /s/ (figure 2).

Par contre, une absence d'abduction de la partie postérieure de la glotte combinée à une légère abduction de sa partie antérieure sont enregistrées durant /T/(figure 2). La section de la glotte apparaît plus étroite durant /T/ comparée à /t/. Cette différence articulatoire est nécessaire pour la production du VOT qui est plus long durant /t/ et très bref durant /T/ (78 msec et 28 msec respectivement devant /i/). Une posture de la glotte relativement similaire à celle de /T/ a été rapportée durant des occlusives voisées non aspirées de l'anglais (Lisker et al., 1969). Cette abduction antérieure très légère est généralement considérée comme passive et attribuée à l'augmentation de la pression intra-orale. Notons que des mesures aérodynamiques faites sur les consonnes de l'arabe marocain montrent que, durant /T/, la pression intra-orale augmente de manière très rapide et très importante pour atteindre une valeur maximale équivalente /t/ (Zeroual, 2002).

Nos observations montrent que /S T χ q/ n'induisent pas de compression du sphincter aryépiglottique bien qu'elles aient une constriction dans la cavité pharyngale. Ces observations semblent être en accord avec les études relativement récentes qui semblent montrer que les emphatiques seraient des uvularisées. En effet, pour Herzellah, (1990) et Bin-Muqbil (2005), les emphatiques seraient produites avec le recul du dos de la langue sans l'implication de sa racine. C'est ce qui permet d'expliquer le fait que les emphatiques induisent une baisse substantielle de F2 des voyelles adjacentes, qui est combinée à une élévation relativement moins importante de F1. Soulignons également que les emphatiques sont accompagnées d'un affaissement très important de la partie centrale de la langue (Zeroual, 2007) qui augmente le volume de la cavité antérieur et qui permet également l'abaissement de F2. Il n'est donc pas exclus que le recul du corps de la langue durant /S T/ ne soit que la conséquence de cet affaissement antérieur de la langue.

3.3 La consonne « épiglottale » non-occlusive

Durant /ʜ/ produite de manière tenue ou dans iCi (figure 2 et figure 3 : ligne 1), nous constatons une rétraction substantielle de l'épiglotte qui aboutit à une constriction épiglotto-pharyngale. Une partie très basse de la racine de la langue recule également, de même que les parois pharyngales latérales sont légèrement compressées. La section de la constriction épiglotto-pharyngale semble assez large pour générer la turbulence enregistrée durant cette consonne. Comparée à [s], /ʜ/ est accompagnée d'une compression antérieure-postérieure très importante de la cavité supra-glottique renforcée par un mouvement vers le haut et vers l'avant des cartilages aryténoïdes (figure 2 et figure 3 : ligne 1). Cette posture ressemble à celle que nous avons durant le chuchotement (Hirose, 1997 ; Zeroual, 2005) et s'accord avec le fait que /ʜ/ possède également les caractéristiques auditives de ce mode phonatoire. Ces observations semblent renforcer l'hypothèse selon laquelle le bruit de friction durant /ʜ/ est engendré au niveau aryépiglottale (voir introduction).

3.4 Les articulations simulées pouvant correspondant à des linguo-pharyngales non-occlusives et occlusives

Les essais en vue de produire une pharyngale non-occlusive montrent un recul très marqué de l'épiglotte et surtout d'une partie plus importante et plus élevée de la racine de la langue (figure 3, ligne 2 : position 1). Cette dernière finit par cacher le sommet de l'épiglotte pour s'approcher très étroitement de la paroi pharyngale arrière. La constriction maximale semble être linguo-pharyngale, elle est facilitée par la position plus basse de l'épiglotte ainsi que par la compression très importante des parois pharyngales latérales. La présence de la salive (secrétée très probablement de manière réflexive) montre une vibration aléatoire de sa partie située au niveau sagittale suggérant un écoulement central de l'air. Cette articulation linguo-pharyngale non-occlusive induit également une compression antérieure postérieure de la cavité laryngée combinée à une position plus basse du larynx et des cartilages aryténoïdes (figure 3, ligne 2 : position 2). Nos images ne permettent pas de dire si cette constriction linguo-pharyngale est plus étroite que la constriction aryépiglottale. La configuration de la glotte qui ressemble à celle enregistrée durant /ʜ/ et durant la parole chuchotée suggère que la turbulence est générée au niveau supra-glottique.

Les essaies de production d'une occlusive (upper) pharyngale montrent une rétraction très marquée de l'épiglotte et notamment d'une partie plus grande et plus haute de la langue, combinée à une compression substantielle des parois pharyngales latérales (figure 3, ligne 3 : position 1). Généralement, la racine de la langue et l'épiglotte restent légèrement éloignées de la paroi pharyngale arrière (figure 3, ligne 3 : position 1a) suggérant que l'occlusion serait glottique ou supra-glottique. Ceci est en accord avec nos observations au niveau de la cavité laryngée qui montrent que ce type d'articulation s'accompagne généralement d'une compression antérieure-postérieur de la cavité laryngée et d'une occlusion glottique totale (figure 3, ligne 3 : position 2a). Cette compression aryépiglottique n'est pas aussi importante que durant /ʜ/ due très probablement à la position plus basse des cartilages aryténoïdes et du larynx.

Dans quelques répétitions, une partie encore plus haute de la racine de la langue recule et cache l'épiglotte suggérant que la constriction totale serait linguo-pharyngale (figure 3, ligne 3 : position 1b). Cette déduction est renforcée par d'autres observations qui montrent que ce type d'articulation induit une adduction des cordes vocales qui n'aboutit pas, toutefois, à une constriction totale (figure 3, ligne 3 : position 2b). Cette légère abduction peut être passive due à la montée de la pression située derrière l'occlusion linguo-pharyngale.

4 Conclusion

L'arabe marocain possède une consonne sourde « épiglottale » /ʜ/ et plus exactement aryépiglottale combinant une rétraction d'une partie très basse de la racine de la langue et surtout de l'épiglotte, ainsi qu'une compression antérieure-postérieure du niveau aryépiglottique qui semble constituer son lieu d'articulation principal. Les articulations linguo-pharyngales très étroites sont possibles, mais également accompagnées d'une constriction aryépiglottique assez similaire à celle enregistrée durant /ʜ/.

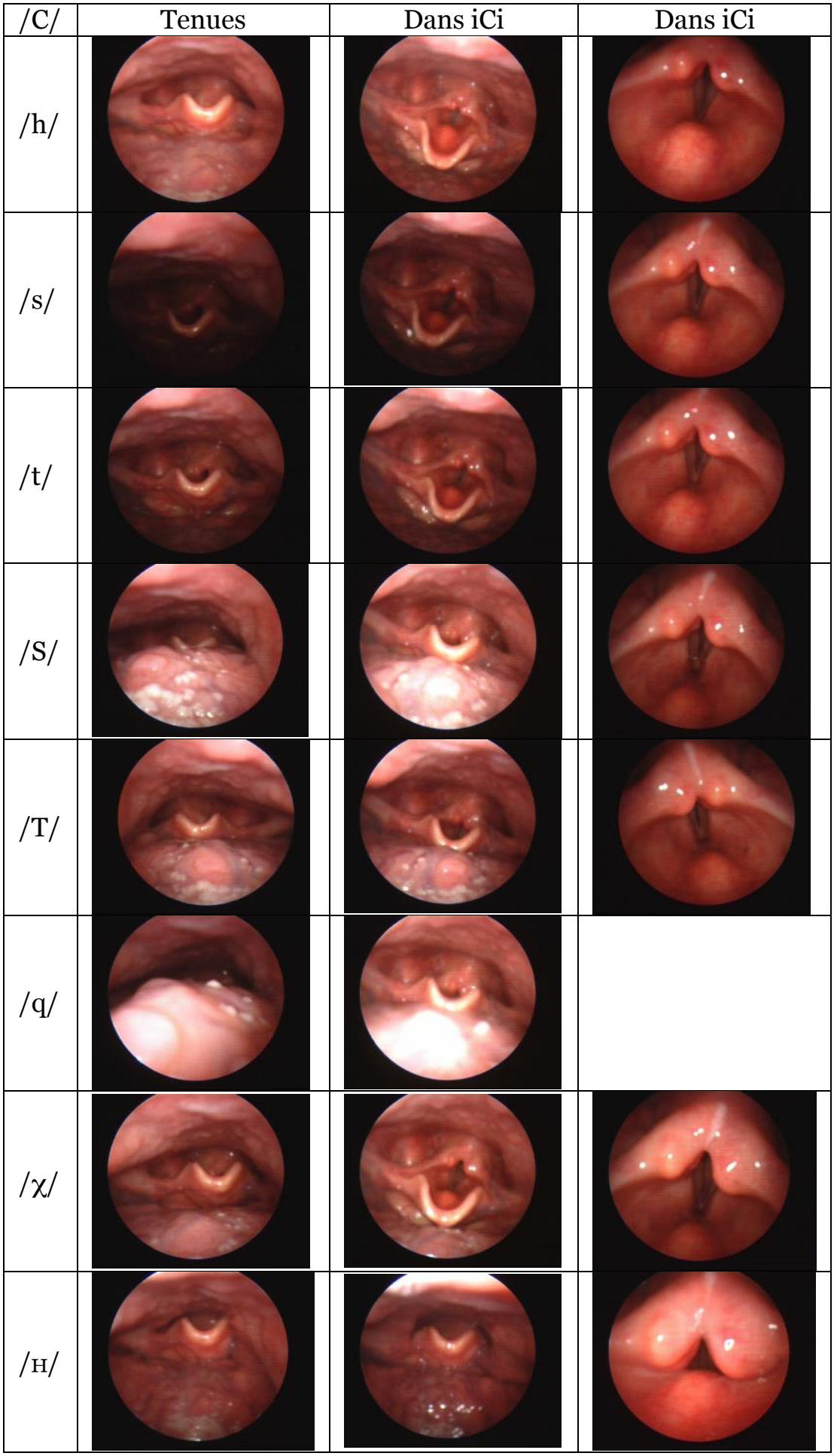


FIGURE 2 : Cavités pharyngale et laryngée durant /h s t S T q χ ʜ/ tenues et dans iCi

Les essaies de production d'une articulation linguo-pharyngale totale aboutissent généralement au recul très important d'une partie beaucoup plus haute de la racine de la langue et de l'épiglotte combiné à une constriction glottique totale et une compression antérieure-postérieure aryépiglottique. Elle peut également être produite avec une constriction totale qui semble être linguo-pharyngale combinée à une abduction partielle de la partie antérieure de la glotte et une adduction totale de sa partie postérieure.

D'autres investigations aérodynamiques sont nécessaires pour vérifier si la constriction linguo-pharyngale peut être totale (linguo-pharyngale occlusive) ou plus étroite (linguo-pharyngale non-occlusive) que la constriction aryépiglottique. Dans ces deux cas, nous devrons enregistrer une augmentation de la pression dans la cavité pharyngale située derrière la constriction linguo-pharyngale.

Toutes nos productions linguo-pharyngales sont généralement accompagnées d'une compression aryépiglottique et sont produites alors que le larynx reste dans une position abaissée. Ces observations s'accordent avec la déduction d'Esling (1996) selon laquelle les aryépiglottales qui sont généralement réalisées avec une montée du larynx peuvent également se réaliser avec une position plus basse du larynx.

Plusieurs arguments ont été présentés pour montrer que l'emphase serait un cas particulier des articulations produites dans la cavité pharyngale, puisqu'elle est produite avec une rétraction de la partie postérieure de la langue sans être accompagnée d'une constriction aryépiglottique.

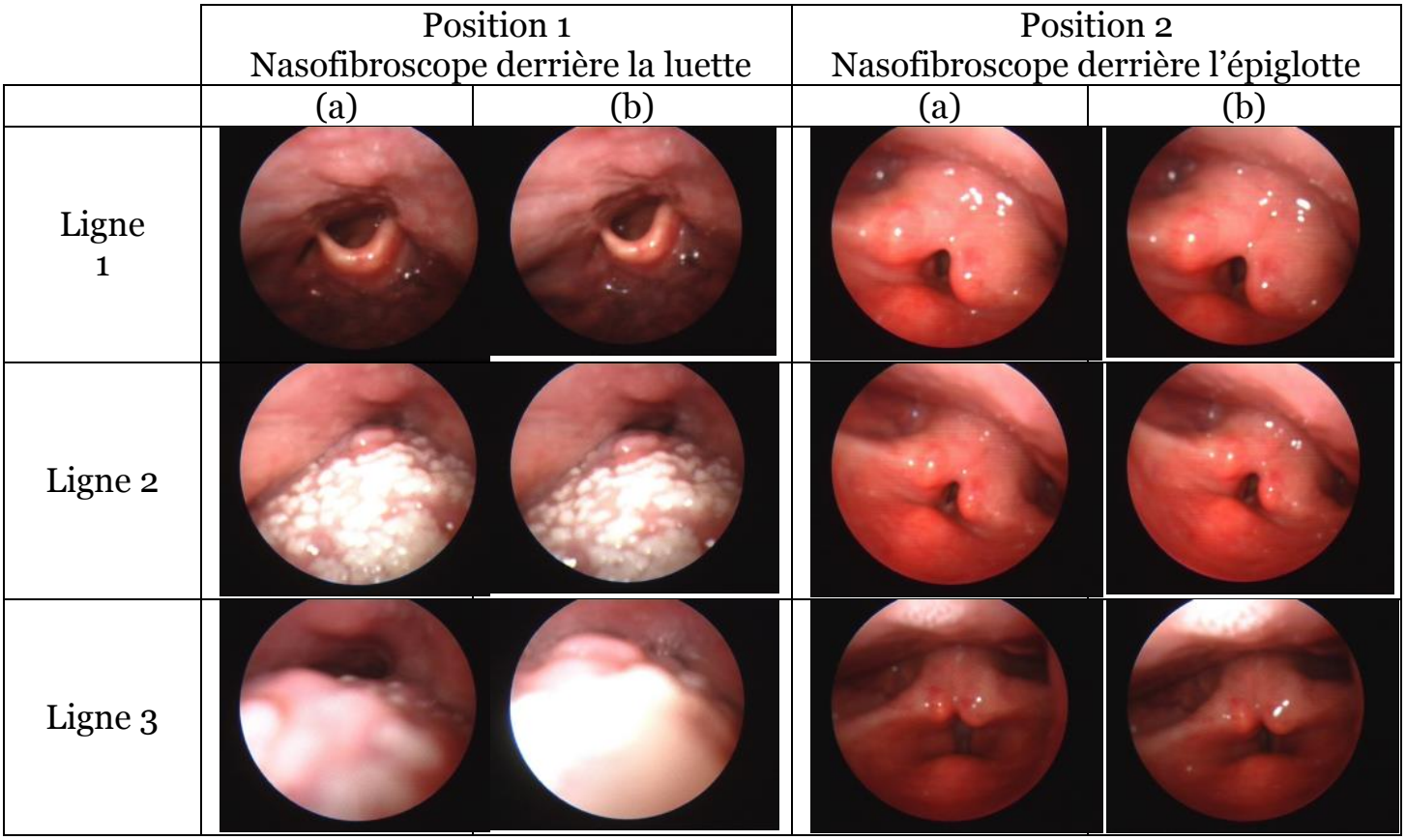


FIGURE 3 : Cavités pharyngale et laryngée durant /H/ tenues (ligne 1) ainsi que durant des linguo-pharyngales non-occlusives (ligne 2) et occlusives (ligne 3)

Références

BIN-MUQBIL, M. S. (2006). *Phonetic and phonological aspects of Arabic emphatic gutturals*. Ph.D. dissertation, University of Wisconsin-Madison.

CATFORD, J. (1968). The articulatory possibilities of man. in B. Malmberg, (ed.) *Manual of Phonetics*. North-Holland Publishing Compagny : Amsterdam, 45-75.

CATFORD, J. (1977). *Fundamental Problems in Phonetics*. Edinburgh University Press.

ESLING, J. H. (1996). Pharyngeal consonants and the aryepiglottic sphincter. *JIPA* 26.

ESLING, J.H. (1998). Laryngoscopic analysis of pharyngeal articulations and larynx-height voice quality settings. *Proceedings of the 5th ICSLP*, Sydney, 543-546.

HERZALLAH, R. (1990). *Aspects of Palestinian Arabic phonology: a Non-linear Approach*. Ph.D. dissertation, Cornell University.

HONJOW, I. and ISSHIKI, N. (1971). Pharyngeal stop in Cleft Palate speech. *Folia Phoniat.*, 23 : 347-354.

KAGAYA, R. (1971). Laryngeal gesture in Korean stop consonants. *Ann. Bull. RILP* 5.

LADEFOGED, P & MADDIESON, I. (1996). *The Sounds of the World's Languages*. Blackwell : Cambridge USA & Oxford UK.

LAVER, J. (1994). *Principle of Phonetics*. Cambridge University Press : Cambridge.

LISKER, L., ABRAMSON, A.S., COOPER, F.S. & SCHEVEY, M.H. (1969). Transillumination of the larynx in running speech. *JASA* 45: 1544-1546

MADDIESON, I. (2009). Typology and occurrence of pharyngeals and pharyngealization around the world. *International Workshop on Pharyngeals and Pharyngealisation*. Centre for Research in linguistics and Language Sciences. University of Newcastle, UK.

YEOU, M., & MAEDA, S. (1995). Uvular and pharyngeal consonants are approximants: An acoustic modeling and study. *Proceedings of the XIIIth ICPHS*, Stockholm, 586 589.

ZEROUAL, C. (2002). Etude aérodynamique et acoustique des occlusives emphatiques et non-emphatiques de l'arabe marocain. *XXIVèmes JEP*, Nancy, France, 365-368.

ZEROUAL, C. (2003). Aerodynamic study of Moroccan Arabic guttural consonants. *Proceedings of the XVth ICPHS*, Barcelona, 1859-1862.

ZEROUAL, C., ESLING J.H. & CREVIER-BUCHMAN, L. (2005) Physiological study of whispered speech in Moroccan Arabic. *INTERSPEECH*, Lisboa, Portugal, 1069-1072.

ZEROUAL, C., HOOLE, P., FUCHS, S., & Esling, J.H (2007). EMA study of the coronal emphatic and non-emphatic plosive consonants of Moroccan Arabic. *Proceedings of the XVIth ICPHS*, Saarbrücken, Germany, 397-400.

ZEROUAL, C., ESLING, J.H., & HOOLE, P. (2011). EMA, endoscopic, ultrasound and acoustic study of two secondary articulations in Moroccan Arabic: labial-velarization vs. emphasis. In Zeki, M.H. and B. Heselwood (eds.), *Instrumental Studies in Arabic Phonetics*. John Benjamins, 277-297.

Coordination/synchronisation gestes-voix dans la démence de la maladie d'Alzheimer en voix parlée et chantée : une étude de cas

Diane Caussade[1], Nathalie Vallée[1], Nathalie Henrich Bernardoni[1], Jean-Marc Colletta[2]
(1) GIPSA-Lab, DPC, UMR 5216 CNRS, BP25, 38040 Grenoble cedex, France
(2) LIDILEM, EA609, 1180 avenue Centrale, BP 25 38040 Grenoble, France
diane.caussade@gipsa-lab.grenoble-inp.fr, nathalie.vallee@gipsa-lab.grenoble-inp.fr, nathalie.henrich@gipsa-lab.grenoble-inp.fr, jean-marc.colletta@u-grenoble3.fr

RESUME

Dans la démence de la maladie d'Alzheimer, les études sur la production bimodale du langage n'abordent pas l'aspect concomitant parole-geste. Nous proposons un protocole expérimental original en voix parlée et chantée visant à décrire les troubles de la coordination/synchronisation entre parole et gestualité manuelle chez des patients Alzheimer. Le protocole a été testé lors d'une étude de cas sur une patiente et une participante contrôle. Les résultats obtenus avec cette étude pilote montrent que les capacités communicatives de la patiente sont affectées dans leur multimodalité aussi bien au niveau de la coordination/synchronisation gestes-voix qu'au niveau de la qualité de l'exécution des gestes, notamment s'ils sont effectués en voix parlée ou chantée. Des différences sont observées en fonction de la valeur (déictique, rythmique ou représentationnel) des gestes.

ABSTRACT

Coordination/synchronization between voice and gestures in speech and singing in the dementia of Alzheimer's disease: A case study
In Alzheimer-type dementia, studies on bimodal production of language do not deal with the concomitant aspect of speech-gesture (Rousseau, 2011). This paper is about an original protocol using speech, singing and manual gestures with the aim to describe coordination/synchronization troubles between verbal and non-verbal communication. The protocol was tested across a case study both on a patient and a control participant. Results show that the patient's communicative capacities tend to be affected in the multimodal dimension, in the voice-gestures coordination/ synchronization, as well as in the execution quality of gestures with differences depending whether they are executed during speech or singing. Differences are observed according to the gesture values.

MOTS-CLÉS : maladie d'Alzheimer, geste manuel, parole, chant, coordination/synchronisation
KEYWORDS: Alzheimer's disease, manual gesture, speech, singing, coordination/ synchronization

1 Introduction

La démence de la maladie d'Alzheimer (MA) est une affection neurodégénérative du système nerveux central provoquant la détérioration progressive et durable des fonctions cognitives du patient et affectant principalement ses capacités mnésiques mais aussi communicatives, avec des apraxies de la parole et du geste manuel (Rousseau, 2011), bien que les différences entre vieillissement normal et pathologique ne sont pas toujours évidentes (Albert, 1981). Au niveau linguistique, les troubles du langage associés à cette maladie se manifestent d'abord

au niveau de la production et ne semblent altérés que plus tardivement au niveau de la compréhension (Barkat-Defradas et al., 2008). Même si certaines procédures d'évaluation des capacités de communication de patients atteints de la MA prennent en compte l'aspect multimodal de la communication humaine, et en particulier les actes non verbaux (Rousseau, 2011), la coordination/synchronisation entre parole et gestes est absente des critères évaluant l'altération des capacités langagières des patients car peu ou pas étudiée à ce jour. Des travaux montrent pourtant que dans le cas de la MA, un désordre central sémantico-conceptuel sous-tend à la fois les détériorations communicatives linguistiques et gestuelles (Parakh et al., 2004 ; Rousseau, 2011).

Par ailleurs de nombreux auteurs ont démontré le caractère multimodal de la communication parlée (par exemple Kendon, 2004 ; Colletta et Guidetti, 2012). La production langagière n'est pas seulement vocale : elle s'accompagne de gestes manuels à visées communicationnelles. La communication visuo-gestuelle serait même, pour certains auteurs, l'origine du développement ontogénétique et phylogénétique du langage latéralisé. Pour l'aspect phylogénétique, plusieurs hypothèses existent : si pour certains auteurs, l'origine du langage serait purement vocale (MacNeilage, 2008), d'autres favorisent l'hypothèse d'une origine gestuelle du langage (entre autres McNeill, 2000 ; Corballis, 2003 ; Taglialatela et al., 2011), voire l'hypothèse d'une origine mixte (par exemple Masataka, 2008). Concernant l'aspect ontogénétique, des travaux récents montrent que dès 12 mois, l'enfant commence à utiliser des gestes manuels, et plus particulièrement des gestes déictiques. Ces gestes qui accompagnent le babillage de l'enfant constitueraient une première forme de communication intentionnelle (Goldin-Meadow et Butcher, 2000 ; Gonseth, 2013) précurseur de l'émergence de la syntaxe. La coordination entre le maintien (ou *stroke*) du geste et l'articulation des syllabes des mots affiliés aux gestes a été mise en évidence par Levelt (1989), suggérant un lien étroit entre la mise au point du programme moteur des gestes manuels et celle du plan articulatoire pour la parole. Une telle relation suppose qu'en cas de dégradation de l'articulation des mots, le timing du geste l'accompagnant serait influencé (Roustan, 2006). Dans le prolongement de cette idée, l'étude de la relation parole-geste pourrait sans doute aider au diagnostic de pathologies entraînant des troubles de la communication, telle que la démence de la MA. C'est dans ce cadre que s'inscrit l'étude présentée ici : elle a consisté à mettre en place et à tester un protocole expérimental en voix parlée et voix chantée afin d'établir un scénario de la dégradation de la coordination/synchronisation de la communication orale et gestuelle chez des patients atteints de la démence de la MA.

L'intégration de la modalité chantée dans un protocole d'étude de la relation temporelle geste-parole est originale. Elle s'appuie sur le fait que sur le plan phonatoire, voix chantée et voix parlée présentent des mécanismes communs tels qu'articulatoires, auditifs, cognitifs, proprioceptifs et physiques (Peretz et al., 2004). Certains chercheurs suggèrent que les liens fonctionnels et structurels entre parole et musique s'expliquent par le fait que ces deux systèmes auraient une genèse commune et universelle (Brown, 2000), et se seraient spécialisés par la suite (Brown, 2000 ; Mithen, 2005). Ainsi l'existence de capacités proto-musicales et proto-linguistiques chez les nouveau-nés humains pourrait être interprétée comme les racines d'un héritage biologique commun (Cross, 2003). D'un point de vue médical, il a été montré que la musique peut réactiver des zones de l'hémisphère gauche et bénéficier à des patients atteints de démences dégénératives comme celle de la MA (Brotons et Koger, 2000). En ce qui concerne notre étude, nous pensons que l'observation de la production des gestes en parole et en chant permettra de déterminer : 1/ quels types de gestes sont produits selon la modalité ; 2/ l'impact de la voix chantée sur le lien temporel entre voix et gestes chez des personnes atteintes de la démence de la MA. Nous faisons l'hypothèse que les troubles des capacités communicatives des patients Alzheimer affectent la communication dans sa multimodalité, orale et gestuelle, différemment en voix parlée et en

voix chantée. Les résultats préliminaires obtenus lors de cette étude de cas n'autorisent certes pas de généralisations, mais ils ont permis d'évaluer notre protocole expérimental et d'ouvrir les perspectives de recherche attendues.

2 Matériel et méthode

2.1 Participants

Notre protocole d'expérimentation est destiné à des sujets atteints de la démence de la MA recrutés sur les critères suivants :

- score au test MMSE (*Mini Mental State Examination*) entre 18 et 11 sur 30 (ce qui correspond à une atteinte cognitive modérée) ;
- droitier ;
- langue maternelle française ;
- non musicien professionnel.

Cette étude de cas vise à évaluer un protocole permettant d'étudier les capacités langagières d'une patiente dans des tâches de communication bimodale utilisant la voix parlée et la voix chantée et de comparer ses performances à celle d'une participante contrôle. Toutes les deux correspondent à nos critères de sélection et sont âgées respectivement de 68 et 69 ans. La patiente a une atteinte cognitive modérée (14/30 au test MMSE). Leur profil sociolinguistique a été évalué grâce au questionnaire ci-dessous (Table 1). Ce questionnaire a révélé que la participante atteinte de la démence de la MA a davantage d'expérience en communication non-verbale en raison de son expérience de loisir en théâtre, que la participante contrôle, qui elle a plus d'expérience en musique. Bien que nos participantes ne soient appariées ni au niveau sociolinguistique ni au niveau socioéducatif, cette étude préliminaire nous a permis de vérifier la faisabilité du protocole par une participante atteinte de démence de la MA et une participante contrôle ne souffrant ni de troubles cognitifs ni mnésiques (d'après son score au test MMSE).

1) Avez-vous fait de la musique ? (en cours ou de manière autodidacte)
2) Avez-vous fait du théâtre? (en cours ou de manière autodidacte)
3) Jouez-vous d'un instrument de musique ?
4) Avez-vous chanté dans le cadre de votre travail ?
5) Avez-vous chanté dans le cadre de vos loisirs ?
6) Avez-vous chanté des enfantines (berceuses, comptines, …) à vos enfants et/ou petits-enfants ?
7) Vos parents vous chantaient-ils des chansons ?
8) Pourriez-vous nous donner quelques exemples d'enfantines que vous connaissez ?
9) Connaissez-vous des comptines à gestes manuels ?
10) Avez-vous utilisé des comptines à gestes manuels avec des/vos enfants et/ou petits enfants ?
11) Avez-vous appris différentes langues étrangères (en cours ou de manière autodidacte) ?

TABLE 1 – Questionnaire sociolinguistique utilisé pour l'étude.

2.2 Corpus et protocole

Le protocole est composé de deux tâches :

- une tâche de narration ;
- une tâche de répétition.

Largement utilisé dans des recherches sur la communication multimodale (Colletta et al., 2009), nous avons fait le choix de faire la tâche de narration à partir d'un dessin animé de *Tom & Jerry* ®. Si nous avons choisi la tâche de répétition, c'est parce que ce procédé est utilisé pour évaluer, entre autres, les aphasies et apraxies. Il est à noter que ce protocole a été conçu pour des entretiens individuels.

La tâche de narration consiste à visualiser deux fois de suite un extrait de 2 minutes 47 d'un

dessin animé de *Tom & Jerry* ® et à en faire un résumé à la demande de l'expérimentatrice.
La tâche de répétition consiste à écouter l'expérimentatrice énoncer ou chanter le matériau une mesure à la fois (avec ou sans gestes) et de répéter chacune de ces mesures et de ces gestes, ainsi les participants atteints de la démence de la MA ne sont pas mis en difficulté par les troubles mnésiques dus à la maladie. Pour cette tâche, nous avons pris le parti de créer notre matériau afin de s'assurer qu'aucun des participants ne connaisse les supports utilisés dans ce protocole. Ce dernier est composé de cinq supports comme suit de 10 secondes chacune :

- un air en lallation sans gestes imposés ;
- une comptine en voix parlée sans gestes imposés ;
- une comptine en voix chantée sans gestes imposés ;
- une comptine en voix parlée avec gestes imposés ;
- une comptine en voix chantée avec gestes imposés (figure 1).

La comptine est choisie comme support en ce qu'il s'agit d'un genre de chanson, et plus particulièrement d'enfantine (berceuses, formulettes, rondes, comptines, …), en voix parlée et/ou voix chantée, aux textes courts et répétitifs, en vers, utilisant un lexique fréquent (Cornaz, à paraître).
Pour les gestes, nous nous sommes intéressés aux gestes déictiques, rythmiques et représentationnels. Les gestes déictiques et rythmiques font partie des premiers gestes coverbaux à apparaître dans l'ontogenèse du langage vers douze mois (Gonseth, 2013). Les habiletés verbales narratives et gestuelles représentationnelles se développent parallèlement chez l'enfant dès cinq ans (Colletta et al., 2012).
Les gestes imposés choisis sont les suivants :

- déictiques (gestes de pointage vers quelqu'un ou quelque chose de concret ou d'abstrait) ;
- rythmiques/de battement/*beats* (gestes biphasiques de la main, de la tête ou du pied ayant un rôle d'emphase) ;
- représentationnels/iconiques (gestes représentant une personne ou un objet concret ou abstrait et en lien avec l'imagerie mentale) (Roustan, 2006).

Chaque comptine avec gestes imposés comporte :

- 2 gestes déictiques ;
- 4 gestes représentationnels dont 2 sont associés à une action et 2 à une forme.

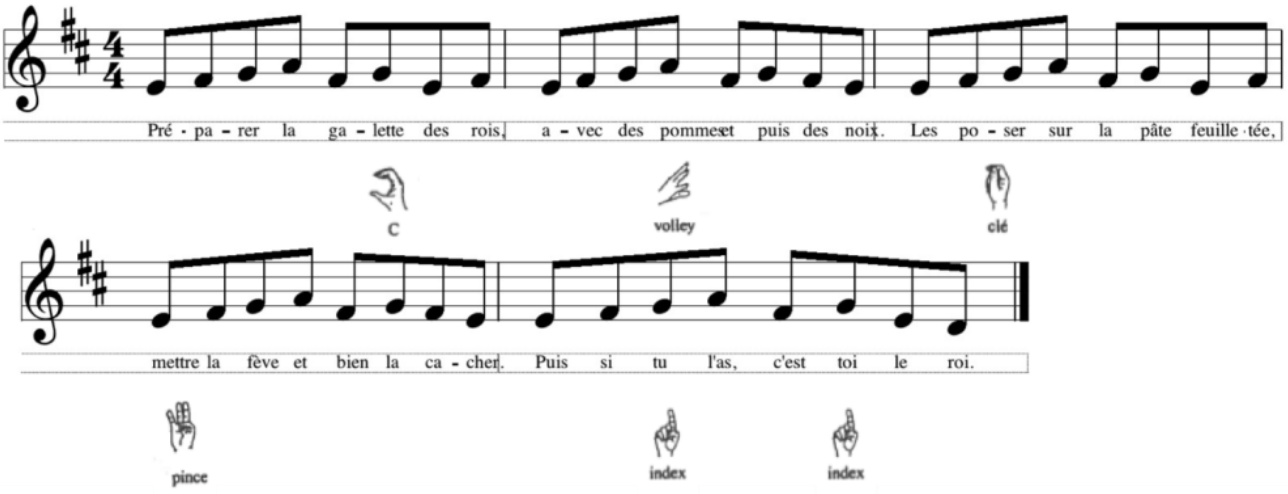


FIGURE 1 – Comptine en voix chantée avec gestes imposés.

2.3 Matériel et analyses

Les enregistrements audio-vidéo ont été réalisés au domicile des deux participantes à l'aide de deux caméscopes sur pied Sony Handycam HDR-XR500 (positionnés l'un de face, l'autre

de profil) et d'un micro-cravate Audio Technica ATR35S.
L'étiquetage des données a été fait sous ELAN® (EUDICO Linguistic Anotato). Ont été annotées les phases gestuelles, la valeur du geste en ne nous intéressant qu'aux gestes déictiques, représentationnels et rythmiques. Nous nous sommes également intéressés à la relation synchronique geste/parole. Pour les gestes imposés, nous avons retenu les critères suivants préconisés par le Manuel de codage de Colletta et al. (2009) :

- le mouvement ;
- l'emplacement ;
- la configuration ;

auxquels a été attribuée une valeur entre 0 et 2 :

- 0 quand le critère n'est pas respecté ;
- 1 quand il est produit mais pas à l'identique ;
- 2 lorsqu'il l'est de manière identique.

La qualité d'exécution de chacun des 3 critères est notée sur 2, donc celle de l'ensemble des critères sur 6. Les deux comptines avec gestes imposés, l'une en voix parlée et l'autre en voix chantée, nous ont permis d'analyser 12 gestes dont la qualité d'exécution a été évaluée à partir des critères ci-dessus.
Pour estimer le degré de coordination/synchronisation entre les deux modalités, nous avons calculé le laps de temps (ΔT) qui sépare le début de la phase de maintien du geste manuel du début de l'articulation du mot. Si $\Delta T = 0$, alors geste et parole sont synchrones. Si $\Delta T < 0$ ms, alors il y a anticipation du geste sur la parole. Si $\Delta T > 0$ ms, alors il y a retard du geste sur la parole. Nous nous intéressons également au délai d'anticipation du geste ($\Delta T < 0$ ms) entre le début de la phase de préparation du geste manuel et celui de la phase de maintien dans les modalités parlée et chantée.

3 Résultats

Dans la tâche de narration, la patiente n'a pu produire que 4 groupes de souffle en une phrase pour résumer le dessin animé présenté. Quant à la participante contrôle, elle a produit une narration résumant les actions du dessin animé en 37 groupes de souffle et une moyenne de 4,5 syllabes par groupe de souffle. Ainsi, la tâche de narration n'a pas été conclusive avec la participante atteinte de la démence de la MA car sa mémoire à court terme ou de travail est probablement affectée, ce qui ne nous a pas permis de recueillir des données analysables.
Dans la tâche de répétition, les participantes n'ont produit aucun geste manuel spontanément, mais tous les gestes manuels imposés, sauf un geste représentant un verbe pour la patiente.

3.1 Gestes produits et leur qualité d'exécution

En ce qui concerne la tâche de répétition qui intègre quatre comptines (dont deux en voix parlée et deux en voix chantée, avec ou sans gestes imposés), aucune des participantes n'a produit spontanément de gestes manuels, ni en voix chantée, ni en voix parlée. Cependant en voix chantée, la patiente a produit 2 gestes céphaliques à valeur rythmique et la participante contrôle 9.
Pour la condition avec gestes manuels imposés pour l'ensemble des deux modalités, la patiente a réussi à en répéter la majorité (10/12) et la participante contrôle la totalité. La qualité d'exécution de la participante contrôle était meilleure (62/72) que celle de la patiente (54/72). Pour cette dernière, la qualité d'exécution semble atteinte au niveau du mouvement (15/24). Les gestes déictiques étaient les mieux préservés, ils ne semblent pas affectés chez la patiente : tous étaient répétés avec une qualité d'exécution de 6/6.
La qualité d'exécution de la patiente pour les gestes représentant une forme est estimée à 4/6 pour 'pizza', 5/6 pour 'tomates', 5/6 pour 'galette des rois' et 6/6 pour 'pommes', et pour ceux

représentant une action : 4/6 pour 'déposer' et 5/6 pour 'poser'. La patiente n'a produit qu'un geste sur deux représentant une action en voix parlée comme chantée (table 2).
La qualité d'exécution de la patiente était légèrement meilleure en voix chantée (28/36) qu'en voix parlée (26/36). La même observation est faite pour la participante contrôle qui a produit une exécution de légèrement meilleure qualité en voix chantée (32/36) qu'en voix parlée (30/36). En voix chantée la qualité d'exécution des deux participantes était meilleure au niveau de la configuration et du mouvement pour des gestes représentant une forme ou une action.

Voix	Geste	Mot Cible	Patiente	Contrôle
parlée	déictique	'toi'	6	6
		'moi'	6	6
	action	'déposer'	4	4
		'parsemer'	0	4
	forme	'pizza'	4	5
		'tomates'	6	5
chantée	déictique	'tu'	6	6
		'toi'	6	6
	action	'poser'	5	5
		'mettre'	0	4
	forme	'galette des rois'	5	5
		'pommes'	6	6

TABLE 2 – Evaluation de la qualité d'exécution des gestes imposés en voix parlée et voix chantée en fonction du type de geste et le mot cible auquel chaque geste est rattaché selon les participants (cf. section 2.3. pour le calcul de l'estimation).

3.2 Timing parole-gestes

Dans les tables 3 et 4, les astérisques (*) correspondent à une erreur de la patiente au niveau du mot cible dû à un trouble de la décentration – dans la terminologie piagétienne, on appelle ainsi la capacité de la pensée à quitter un point de vue pour en envisager un autre, ce qui se traduit dans le langage par une décontextualisation (Brin et al., 2011) – propre à la démence de la MA (Delage et Lejeune, 2009). À noter que lorsque la patiente se trompait au niveau du mot cible, le geste déictique accompagnait tout de même sa parole.

	Partici-pante	Geste					
		déictique		action		forme	
Mot cible		'toi'	'moi'	'déposer'	'parsemer'	'pizza'	'tomates'
Mot prononcé	Patiente	'moi'*	'vous'*	'déposer'	'parsemer'	'pizza'	'tomates'
	Contrôle	'toi'	'moi'	'déposer'	'parsemer'	'pizza'	'tomates'
ΔT maintien-parole (ms)	Patiente	- 5	1	- 12		- 4	- 13
	Contrôle	- 1	0	- 7	- 16	0	- 3
ΔT préparation-maintien (ms)	Patiente	- 14	- 14	- 6		- 58	- 9
	Contrôle	- 8	- 10	- 6	- 8	- 4	- 5

* erreur au niveau du mot cible, le geste accompagne la parole

TABLE 3 - Relations temporelles gestes imposés et mots cibles en voix parlée selon les mots cibles et les participantes.

Les gestes déictiques semblent être répétés de manière davantage synchrone à la parole que la plupart des gestes représentationnels pour les deux participantes, aussi bien en voix parlée qu'en voix chantée. La modalité chantée semble influencer le temps de préparation du geste déictique : le temps de préparation est ici plus court en voix chantée. Notamment chez la patiente qui met trois fois moins de temps à préparer ses gestes en voix chantée qu'en voix parlée.
Pour les gestes représentationnels, la voix chantée ne semble pas améliorer ici la synchronie entre parole et gestes. Plus la charge cognitive était importante (modalité chantée), plus la patiente présentait le geste imposé avant de parler. Ce qui n'est pas le cas de la participante contrôle.

	Partici-pante	Geste					
		déictique		action		forme	
Mot cible		'tu'	'toi'	'poser'	'mettre'	'galette des rois'	'pommes'
Mot prononcé	Patiente	'je'*	'moi'*	'poser'	'mettre'	'galette des rois'	'pommes'
	Contrôle	'tu'	'toi'	'poser'	'mettre'	'galette des rois'	'pommes'
ΔT maintien-parole (ms)	Patiente	1	0	3		- 19	- 42
	Contrôle	- 1	- 1	- 2	0	- 8	- 5
ΔT préparation-maintien (ms)	Patiente	- 3	- 5	- 7		- 9	- 6
	Contrôle	- 6	- 7	- 3	- 2	- 2	- 7

* erreur au niveau du mot cible, le geste accompagne la parole

TABLE 4 - Relations temporelles gestes imposés et mots cibles en voix chantée selon les mots cibles et les participantes.

4 Discussion et perspective

Cette étude de cas nous a non seulement permis d'évaluer la faisabilité et la pertinence de notre protocole expérimental, mais aussi de recueillir et d'analyser des données qui semblent aller dans le sens de nos hypothèses de départ, à savoir que les troubles des capacités communicatives des patients Alzheimer affectent la communication dans sa multimodalité, orale et gestuelle de manière différente en voix parlée et en voix chantée. Ces résultats préliminaires vont dans le sens de ceux de Parakh et al. (2004) et Rousseau (2011). La dégradation que nous avons pu observer chez la patiente, bien que son niveau socioéducatif (NSE) soit plus élevé et que son expérience en communication non-verbale soit plus importante que ceux de la participante contrôle, montre l'intérêt d'observer finement la coordination/ synchronisation entre l'oralité et la gestualité. De plus, il semble, d'après nos résultats, que les gestes représentationnels sont affectés avant les gestes déictiques, aussi bien au niveau de leur qualité d'exécution que de leur relation temporelle avec la parole, ce qui nous amène à suggérer une atteinte graduelle en fonction du type de gestes. Au niveau de la qualité d'exécution des gestes représentationnels, ceux représentant une action semblent atteints avant ceux représentant une forme. Ce qui semble aller dans le sens des recherches de Tapir-Ladino (2003) ou encore de Lee, Hirsch, Gayraud et Barkat-Defradas (2011) qui montrent que chez les personnes atteintes de la MA, la distribution des pauses témoigne de difficultés à récupérer les verbes : la récupération de verbes pourrait demander une charge cognitive plus importante que celle de noms. Néanmoins cette étude de cas ne permet en rien de conclure à un traitement différencié des noms et des verbes, d'autant plus que cette question est au cœur des débats et qu'à l'heure actuelle il ne semble se dégager aucune ligne consensuelle à ce sujet. Nos résultats tendent aussi à montrer que la voix chantée améliorerait la qualité d'exécution des gestes manuels. Le fait que la patiente présente les gestes représentationnels avant la parole montre que le geste vient comme support/aide à celle-ci. Pour la diminution du temps de préparation des gestes en voix chantée, nous supposons un effet du rythme sur la production langagière.

Pour notre étude à plus grande échelle, nous allons recalibrer la tâche de narration en utilisant non plus seulement un support vidéo mais également des images statiques pour se libérer des difficultés mnésiques rencontrées dans la démence de la MA et afin de les contourner (Behrer et al., 2004). Nous allons également ajouter une tâche de répétition en voix chantée avec gestes manuels imposés à valeur rythmique afin d'intégrer ce type de gestes à l'analyse des gestes imposés et ainsi analyser leur coordination/synchronisation en voix parlée et voix chantée, et la comparer à celle des gestes spontanés. En ce qui concerne l'évaluation des participants, les résultats au test MMSE étant étroitement lié au NSE (Kalafat et al., 2003), nous allons évaluer le NSE grâce au test de Poitrenaud. Nous allons également utiliser la batterie brève d'évaluation des praxies gestuelles de Florence Mahieux-Laurent et al. (2009) afin de tester l'exécution sur commande verbale des gestes représentationnels et la répétition par imitation de gestes sans signification. Nous avons également composé un autre set de comptines en voix parlée et voix chantée avec ou sans gestes imposés afin de pouvoir recueillir davantage de données par participant.
Les résultats de cette étude pilote doivent maintenant être confirmés auprès d'un plus grand nombre de participants. Une étude longitudinale permettra de suivre l'évolution des capacités langagières et praxiques des personnes atteintes de la démence de la MA. L'étude des gestes communicatifs dans cette maladie est susceptible d'apporter des connaissances nouvelles quant au rôle de la communication non-verbale dans les activités langagières, et plus particulièrement dans le cas des aspects neurodégénératifs de cette maladie.

Remerciements

Cette étude pilote a pu être réalisée grâce au concours des Dr Olivier Moreaud et Lora Cohen du Pôle Psychiatrie et Neurologie du CHU de Grenoble, du CM2R Arc Alpin et du LPNC de Grenoble UMR CNRS 5105.
Nous remercions également les participantes.

Références

ALBERT, M. L. (1981). Changes in Language with Aging. *Seminars in Neurology*, 1(1): 43-46.

BARKAT-DEFRADAS, M., MARTIN, S., RICO DUARTE, L. & BROUILLET, D. (2008). Les troubles de la parole dans la maladie d'Alzheimer. *27[e] journée d'études sur la Parole*, Avignon.

BRIN, F., COURRIER, C., LEDERLE, E. & MASY, V. (2011). *Dictionnaire d'orthophonie*. Ortho Éditions: Isbergues.

BROTONS, M. & KOGER, S. M. (2000). The Impact of Music Therapy on Language Functioning in Dementia. *Journal of Music Therapy*, 37(3): 183-195.

BROWN, S. (2000). The 'musilanguage' model of music evolution. In N. L., Wallin, B., Merker & S., Brown (dir.). *The origins of music*: 271-300. MIT Press: Cambridge.

COLLETTA, J.-M., GUIDETTI, M. (éds.) (2012). *Gesture and Multimodal Development*. Benjamins Current Topics series: Amsterdam.

COLLETTA, J.-M., AUGUSTYN, M., CALBRIS, G., CAPIRCI, O., ECE DEMIR, O., GOLDIN-MEADOW, S., GRANDATY, M., GRAZIANO, M., GUIDARELLI, B., GUIDETTI, M., KENDON, A., KUNENE, R., LEVINE, S., MILADI, L., MILLET, A., MUGNIER, S., ÖZÇALISKAN, S., PELLEQ, C., ROUSSET, I. SIMON, J.-P. & VENOUIL, A. (2009). *Manuel de codage. Transcription et annotation de données multimodales sous ELAN*. Projet ANR Multimodalité 2005-2009.

CORBALLIS, M. C. (2003). From mouth to hand: Gesture, speech, and the evolution of right-handedness. *Behavioral and Brain Sciences,* 26: 199-260.

CORNAZ, S., CHOVELON, C. & JAUNEAU-CURY, N. (soumis). La vocalise et la comptine au service de la didactique des langues. *E-CRINI* : Rennes.

CROSS, I. (2003). Music and evolution: Causes and consequences. *Contemporary Music Review*, 22(3): 79-89.

DELAGE, M. & LEJEUNE, A. (2009). *La résilience de la personne âgée, un concept novateur pour prendre en soin la dépendance et la maladie d'Alzheimer*. De Boeck: Paris.

GOLDIN-MEADOW, S. & BUTCHER, C. (2000). Gesture and the transition from one- to two-word speech: When hand and mouth come together. In D., McNeill (ed.): *Language and gesture:* 235-257. Cambridge University Press, New York.

GONSETH, C. (2013). *Multimodalité de la communication langagière humaine : interaction geste/parole et encodage de la distance dans le pointage*. Thèse de doctorat en Sciences Cognitives, Université de Grenoble.

KENDON, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press: New York.

KALAFAT, M., HUGONOT-DIENER, L. & POITRENAUD, J. (2003). Standardisation et étalonnage français du "Mini Mental State" (MMS) version GRECO. *Revue de Neuropsychologie*, 13 (2): 209-236.

LEE, H., HIRSCH, F., GAYRAUD, F. BARKAT-DEFRADAS, M. (2011). Organisation temporelle de la parole dans la maladie d'Alzheimer : évolution des variables temporelles en fonction du stade de démence. *4[e] Journée de Phonétique clinique*, Strasbourg.

LEVELT, W. J. M. (1989). Speaking: From Intention to Articulation. MIT Press: Cambridge, MA.

MACNEILAGE, P.F. (2008). *The Origin of Speech*. Oxford University Press : Oxford.

MAHIEUX-LAURENT, F., FABRE, C., GALBRUN, E., DUBRULLE, A. & MORONI, C. (2009). Validation d'une batterie brève d'évaluation des praxies gestuelles pour consultation Mémoire. Evaluation chez 419 témoins, 127 patients atteints de troubles cognitifs légers et 320 patients atteints d'une démence. *Revue neurologique*, 65: 560-567.

MASATAKA, N. (2008). The gestural theory of and the vocal theory of language origins are not incompatible with one another. In N., Masataka (ed.). *The origins of language: Unraveling evolutionary forces:* 1-10. Springer: Heidelberg.

MCNEILL, D. (2000). Catchments and contexts: Non-modular factors in speech and gesture production. In D., McNeill (ed.): *Language and Gesture:* 313-328. Cambridge University Press: New York.

MITHEN, S. (2005). *The singing Neanderthals: The origins of music, language, mind and body*. Weidenfeld and Nicolson: London.

PARAKH, R., ROY, E., KOO, E. & BLACK, S. (2004). Pantomime and imitation of limb gestures in relation to the severity of Alzheimer's disease. *Brain and Cognition*, (55): 272-274.

PERETZ, I., GAGNON, L., HEBERT, S. & MACOIR, J. (2004). Singing in the brain: Insights from cognitive neuropsychology. *Music Perception*, 21(3): 373-390.

ROUSSEAU, T. (2011). *Maladie d'Alzheimer et troubles de la communication*. Elsevier-Masson : Issy-les-Moulineaux.

ROUSTAN, B. (2012). *Étude de la coordination gestes manuels/parole dans le cadre de la désignation*. Thèse de doctorat, Université de Grenoble.

TAGLIALATELA, J. P., RUSSELL, J. L., SCHAEFFER, J. A. & HOPKINS, W. D. (2011). Chimpanzee Vocal Signaling Points to a Multimodal Origin of Human Language. *PLoS One* 6(4) .

TAPIR-LADINO, M. (2003). Analisis de los enunciados producidos despues de una pausa en la conversacion de enfermos de Alzheimer y de senescentes. *Revista de Linguistica Teorica y Aplicada*, 41: 103-118.

Les voyelles fermées tendues, relâchées et allongées du français québécois : la contribution d'indices statiques/dynamiques et absolus/normalisés à la détermination de leur identité acoustique

Caroline Sigouin, Vincent Arnaud
Département des arts et lettres, Université du Québec à Chicoutimi, 555, boulevard de l'Université, Chicoutimi (Québec), G7H 2B1, CANADA
caroline.sigouin@uqac.ca, vincent.arnaud@uqac.ca

RÉSUMÉ

L'objectif de cette contribution est de déterminer les indices acoustiques discriminant le plus efficacement les variantes tendues, relâchées et allongées des voyelles fermées du français québécois. 1350 occurrences, extraites de la parole de 30 locuteurs des villes de Rouyn-Noranda, Saguenay et Québec, ont été analysées. Leur durée a été relevée, puis la fréquence fondamentale et la fréquence centrale des trois premiers formants (F_1, F_2, F_3) ont été estimées à 25-50-75 % de la durée. Les valeurs fréquentielles en Hertz et en scores-Z ont été considérées. Les résultats d'analyses discriminantes témoignent de la contribution de chaque indice acoustique, de la dynamique spectrale et de la normalisation à la répartition des occurrences au sein de leur classe vocalique d'appartenance. Alors que F_1 et F_2, en Hertz, relevés au centre des voyelles permettent d'obtenir une classification correcte à 58 %, d'autres combinaisons de prédicteurs parviennent à des taux s'élevant jusqu'à 97 %.

ABSTRACT

Quebec French tense, lax and lenghtened close vowels: The contribution of static/dynamic and raw/normalized cues to the determination of their identity

This study aims to estimate the contribution of various acoustic cues to the determination of tense, lax and lengthened variants of Quebec French close vowels identity. To do so, we analyzed the speech of 30 speakers from Rouyn-Noranda, Saguenay and Quebec cities who produced 1350 tokens of the variants under study. Their duration has been measured then the fundamental frequency and the central frequency of the first three formants (F_1, F_2, F_3) have been estimated at 25, 50 and 75% of this duration. Raw Hertz frequencies and Z-Scores transformed ones have been exploited. Discriminant analyses results show that each acoustic cue contributes in a manner to close vowels variants distinction. While mid-point Hertz values of F_1 and F_2 lead to an overall correct classification rate of 58%, other combinations of predictors may achieve as much as 97% of correct classification.

MOTS-CLÉS : voyelles fermées, français québécois, dynamique spectrale, normalisation.
KEYWORDS: close vowels, Quebec French, spectral change, normalization.

1 Introduction

1.1 Variantes phonétiques des voyelles fermées en français québécois

En français québécois (désormais FQ), les voyelles fermées /i/, /y/, /u/, dont les formes canoniques sont brèves et tendues, sont fréquemment soumises à différents phénomènes contextuels, nommément le relâchement, l'allongement, la diphtongaison, le dévoisement, la réduction et la syncope. Comme le remarque McLaughlin (1986), en observant la distribution de ces diverses réalisations, on constate qu'en syllabe non accentuée, différents phénomènes sont susceptibles de se produire (voire de se combiner) et que leur application est toujours facultative. Au contraire, les variantes présentes sous l'accent se distribuent catégoriquement dans trois contextes complémentaires : les variantes brèves tendues apparaissent en syllabe ouverte (abrégé V#), les variantes relâchées, en syllabe fermée par une consonne non allongeante (VK), et les variantes allongées (potentiellement diphtonguées), en syllabe fermée par une consonne allongeante (VR).[1] Par exemple, les mots *boue*, *boude* et *bouge* sont respectivement prononcés [bu], [bʊd] et [buːʒ] ou [bᵘuʒ] en FQ. C'est à ces trois derniers types de variantes (sous accent) que cette contribution est consacrée.

1.2 Caractéristiques acoustiques : un bref état de la question

Si par définition les variantes allongées présentent des durées plus importantes que les tendues et les relâchées, le rapport de durée entre ces deux derniers types de réalisations, lui, ne fait pas l'unanimité. En effet, dans certains travaux ayant abordé la question du relâchement des voyelles fermées en FQ (par exemple Dumas, 1974), aucune différence de durée n'a été relevée entre les variantes tendues et relâchées. Dans d'autres études, les variantes tendues sont rapportées comme plus longues que les relâchées (Gendron, 1966 ; Martin, 2002 ; Poliquin, 2006). Arnaud et coll. (2011) ont pourtant observé le rapport de durée inverse. Cela dit, cette discordance peut être imputée aux différences méthodologiques entre ces études. En effet, il paraît peu probable qu'une analyse impressionniste comme celle de Dumas (1974) soit à même de révéler de subtiles différences de durée entre voyelles brèves. Par ailleurs, les relâchées seraient plus longues que les tendues lorsque suivies d'une consonne voisée (comme c'est le cas dans Arnaud et coll. 2011), mais plus courtes devant une non-voisée – ce qui est systématiquement le cas dans Poliquin (2006) et majoritairement le cas dans Martin (2002). Les variantes tendues et relâchées ne se distingueraient donc pas intrinsèquement par leur durée (Sigouin, 2013).

En ce qui a trait aux caractéristiques spectrales des voyelles fermées du FQ, on sait depuis Gendron (1966) que les variantes relâchées ont un F_1 plus élevé que celui des tendues (sans pour autant atteindre celui des voyelles mi-fermées homorganiques) et que les relâchés apparaissent comme légèrement centralisés (F_2 plus bas pour /i/ et /y/, plus élevé pour /u/) par rapport à leurs contreparties tendues. Arnaud et coll. (2011) ont de plus remarqué que dans un espace F_1 / F_2, les variantes relâchées se centralisaient en cours d'émission alors que les tendues se déplaçaient en périphérie ; du moins lorsque ces segments étaient insérés dans des logatomes de forme /bV/ et /bVb/. Leur étude révèle également la pertinence de la prise en compte de la F_0, de F_1, de F_2, de F_3 et de leur évolution en cours d'émission dans la répartition des occurrences dans leur classe vocalique attendue.

[1] Nous ne tiendrons pas compte ici des mots, somme toute marginaux, faisant exception à cette distribution, comme *cool* [kuːl] ou *quiz* [kwɪz] (Côté, 2012 : 242).

En ce qui concerne les variantes allongées, leur timbre est sujet à débat et leurs caractéristiques spectrales ont été peu étudiées, généralement sous un angle ne permettant que difficilement la comparaison avec les variantes brèves. Sous l'accent, les voyelles fermées allongées sont-elles tendues (Gendron, 1966 ; Paradis, 1985), relâchées devant /ʁ/ mais tendues ailleurs (Côté, 2012), toujours diphtonguées (Dumas, 1974), parfois diphtonguées (Yaeger, 1979 ; Leblanc, 2012 ; Côté, 2012) ou ne présentent-elles qu'une variation de timbre (Dionne, 1971 ; Santerre et Millo, 1978) ? Au plan acoustique, certains chercheurs (Gendron, 1966 ; Santerre, 1971 ; Poliquin, 2006) n'ont considéré qu'un point de mesure par occurrence (ou une moyenne de plusieurs points), permettant difficilement de rendre compte d'une éventuelle diphtongaison. D'autres (Yaeger, 1979 ; Paradis, 1985) ont opté pour deux points de mesure uniquement lorsque les voyelles étaient jugées diphtonguées (et très peu de voyelles fermées – voire aucune – ont effectivement été jugées diphtonguées). Pour leur part, MacKenzie et Sankoff (2010) n'ont pris en compte qu'un point de mesure pour les voyelles brèves, mais deux pour les allongées, rendant hasardeuse toute comparaison entre ces deux classes vocaliques.

1.3 Objectifs

Les objectifs de cette étude sont a) de proposer une analyse acoustique unifiée des variantes tendues, relâchées et allongées des voyelles fermées du français québécois, b) d'estimer la contribution de la durée, de la F_0, de F_1, F_2, F_3, et de l'évolution temporelle de ces composantes spectrales à la classification d'occurrences de voyelles fermées dans leur classe phonétique attendue d'après leur distribution en syllabe accentuée, telle que proposée dans la littérature : (i#), (iK), (iR), (y#), (yK), (yR), (u#), (uK) ou (uR), et c) d'évaluer l'impact de l'application d'une procédure de normalisation aux valeurs fréquentielles sur cette même classification.

2 Méthodologie

2.1 Échantillon et corpus

La parole de 30 locuteurs du FQ, autant d'hommes que de femmes, étudiants ou diplômés universitaires, âgés de 20 à 31 ans et provenant à parts égales des villes de Saguenay, Québec et Rouyn-Noranda a été analysée. Les enregistrements de Saguenay et de Québec ont été réalisés en chambre anéchoïque ; ceux de Rouyn-Noranda, au domicile des locuteurs. Une partie de la tâche qui leur était demandée consistait à lire à haute voix, aussi naturellement que possible, une série de phrases porteuses pentasyllabiques et sémantiquement significatives contenant un mot cible en position finale. À la suite de la production de chaque phrase, le mot cible était répété de manière isolée après que l'enquêtrice eut simulé le fait qu'elle n'avait pas compris le mot final. Ce protocole permet 1) d'éviter que les locuteurs hésitent quant au sens des mots cibles, 2) de soustraire ces mots à l'influence de la phrase porteuse tout en prévenant un « effet de liste » et 3) de contrôler le nombre d'occurrences de chaque phonème et son contexte. Et comme les locuteurs naïfs du FQ ne semblent pas porter de jugement négatif à l'endroit du relâchement (Dumas, 1987 : 95) et que la diphtongaison des voyelles fermées passe le plus souvent inaperçue (Santerre et coll., 1985 : 51), il est peu probable que le degré de formalité induit par ces conditions expérimentales ait conduit les locuteurs à adopter une prononciation inhabituelle en ce qui concerne les voyelles fermées.

Des 523 mots cibles répétés isolément qui ont été produits une fois par chaque locuteur, nous en avons retenu 45. Les autres mots, qui contenaient l'une ou l'autre des voyelles orales du FQ et qui

sont destinés à d'autres études, agissaient comme distracteurs. Un total de 1350 occurrences lexicales constitue donc notre corpus. Il s'agit, sauf exception, de noms et de formes verbales monosyllabiques dont le noyau est la voyelle /i/, /y/ ou /u/. Ils se répartissent également entre les classes vocaliques (i#), (iK), (iR), (y#), (yK), (yR), (u#), (uK) et (uR), telles que définies par le contexte. Les voyelles en contexte (VR) sont suivies de /ʁ/, /v/, /z/ ou /ʒ/ ; chacune de ces consonnes apparaissant au moins une fois à la suite de chaque voyelle. Notons enfin que seules 7 occurrences ont été rejetées (hésitations lors de leur production ou difficultés d'analyse acoustique). Au final, c'est donc 1343 occurrences vocaliques qui ont été analysées.

2.2 Analyse acoustique

Les voyelles ont été segmentées et étiquetées à l'aide du logiciel PRAAT. Leur durée a ensuite été relevée. Puis les paramètres de détection automatique de la F_0 (par autocorrélation) et de la fréquence centrale des pics formantiques (algorithme de Burg) ont été adaptés manuellement pour chaque occurrence après examen minutieux de la représentation spectrographique. Lorsque la transition entre la voyelle et la consonne était très graduelle (notamment lorsque la voyelle était suivie de /ʁ/), nous avons posé la frontière au milieu de la transition, là où, à l'audition, l'influence mutuelle des deux phones était minimale.

Étant donné la possible diphtongaison des segments en contexte (VR), nous avons choisi de tenir compte de l'évolution des composantes spectrales en cours d'émission. Ainsi, la F_0, le F_1, le F_2 et le F_3 de chaque occurrence ont été relevés à 25 %, à 50 % et à 75 % de leur durée. La localisation de ces trois points de mesure n'est pas guidée par des considérations théoriques impliquant les notions de noyau et d'appendice vocalique. Ce choix méthodologique est plutôt inspiré des nombreuses études consacrées à la théorie du *Vowel Inherent Spectral Change* (VISC) en anglais américain (pour un survol, voir Morrison et Assmann, 2013). La prise en compte d'un point de mesure dans la portion initiale de la voyelle et d'un autre dans la portion finale correspond notamment à l'hypothèse de la *dual target*, formulée par Nearey et Assmann (1986). La position exacte de ces deux points a été établie de façon à ce qu'ils soient « sufficiently extreme to include a substantial portion of observed formant frequency changes, while at the same time avoiding certain irregularities in source characteristics associated with absolute endpoints of the stimuli » (Nearey et Assmann, 1986 : 1298). Le point de mesure central permet quant à lui d'obtenir des résultats comparables aux études antérieures ne s'étant intéressées qu'à l'« état stable » des voyelles ou encore d'affiner la paramétrisation de la dynamique spectrale (lorsque les trois points de mesure sont pris en compte).

2.3 Normalisation

Comme notre échantillon comportait à la fois des hommes et des femmes, nous avons choisi de normaliser la F_0, le F_1, le F_2 et le F_3. Pour ce faire, nous avons adapté la formule des scores-Z proposée par Lobanov (1971), qui consiste à exprimer les valeurs d'une composante fréquentielle par leurs distances, mesurées en écarts-ypes, à la fréquence moyenne de cette composante pour chaque locuteur. Cette opération a pour effet de superposer et d'égaliser les espaces vocaliques des différents locuteurs (et ce, même si dans le cas présent, une partie seulement du système vocalique est représentée). La formule utilisée est la suivante :

$$(N_x)_i = \frac{(F_x)_i - (\overline{F_x})_i}{(s_x)_i} \qquad \text{où :}$$

- $(N_x)_i$ correspond à la fréquence normalisée (au score-Z) du x^e formant (ou de la F_0) d'une occurrence donnée d'un locuteur *i* ;
- $(F_x)_i$ correspond à la fréquence en Hz du x^e formant (ou de la F_0) de cette occurrence ;
- $(\overline{F_x})_i$ correspond à la moyenne du x^e formant (ou de la F_0) du locuteur *i* ;
- $(s_x)_i$ correspond à l'écart-type du x^e formant (ou de la F_0) du locuteur *i*.

2.4 Analyses statistiques

Nous avons procédé à une série d'analyses discriminantes quadratiques. Afin d'éviter une classification trop optimiste induite par un surajustement des fonctions discriminantes, nous avons utilisé une procédure de validation croisée. Celle-ci repose sur une partition de l'échantillon en un sous-échantillon d'entraînement et un sous-échantillon de test. Plus précisément, le discriminant était calculé sur les données provenant de la moitié des locuteurs de notre corpus (sexe et origine géographique le plus balancés possible) puis appliqué sur les données de l'autre moitié de notre échantillon. La variable dépendante était la classe vocalique. Diverses combinaisons de prédicteurs (F_0, F_1, F_2, F_3 – normalisés ou non, relevés en 1, 2 ou 3 points de mesure – avec ou sans durée) ont ainsi été soumises à des analyses discriminantes.

3 Résultats et analyse

Comme on peut l'observer sur la FIGURE 1, à 50 % de leur durée, les voyelles /i/, /y/ et /u/ se distinguent sans surprise par leur F_2 (plus élevé pour les antérieures) et leur F_3 (plus faible pour les arrondies). Les variantes relâchées se différencient pour leur part des tendues et des allongées par leur F_1 plus élevé. De façon moins marquée, les allongées diffèrent également des tendues par un F_1 plus élevé en moyenne (voir aussi la FIGURE 2).

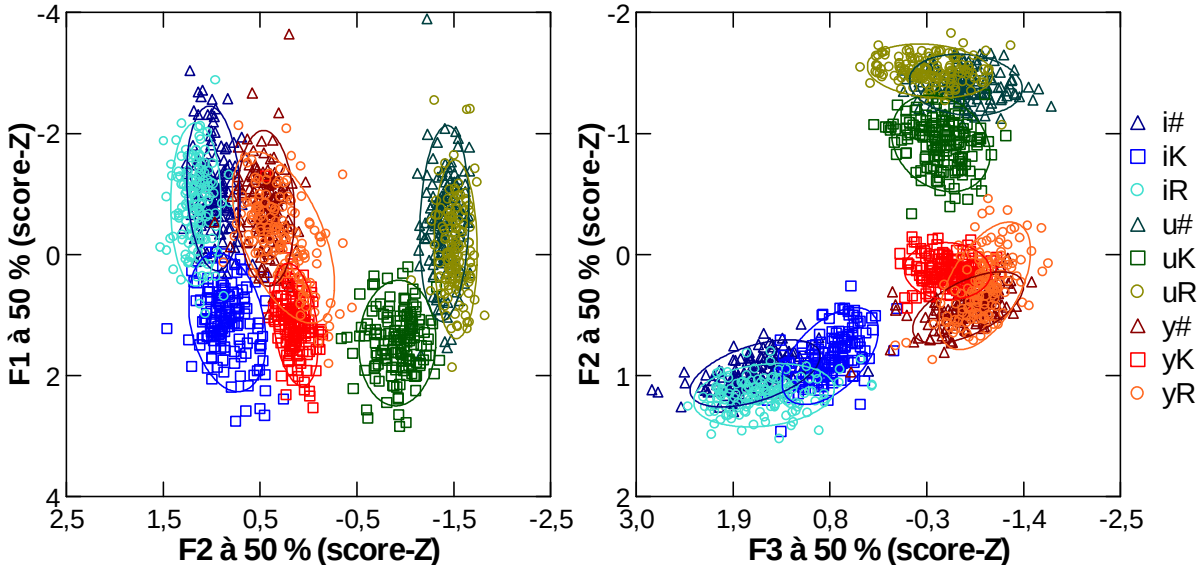


FIGURE 1 – Diagrammes biformantiques F_1 / F_2 et F_2 / F_3 normalisés présentant les 1343 occurrences analysées à 50 % de leur durée regroupées en fonction des classes vocaliques (les ellipses de dispersion englobent 86,5 % des occurrences d'une même classe vocalique, soit ± 2 *s*)

La FIGURE 2 présente quant à elle les trajectoires moyennes (de 25 % à 75 % de leur durée) des classes vocaliques dans un diagramme F_1 / F_2. On constate que même lorsqu'elles sont extraites d'environnements consonantiques variés, la centralisation des relâchées en cours d'émission notée par Arnaud et coll. (2011) se maintient, à tout le moins dans le cas de /i/ et de /u/. Les tendues et les allongées, au contraire, tendent vers la périphérie du trapèze vocalique. Les allongées présentent les trajectoires les plus importantes. D'après MacKenzie et Sankoff (2010), les trajectoires des voyelles fermées allongées sont toutefois moins importantes que celles des autres voyelles allongées généralement considérées comme diphtonguées. Des études ultérieures pourraient permettre de déterminer dans quelle mesure la trajectoire des voyelles fermées est corrélée à la perception de diphtongaison.

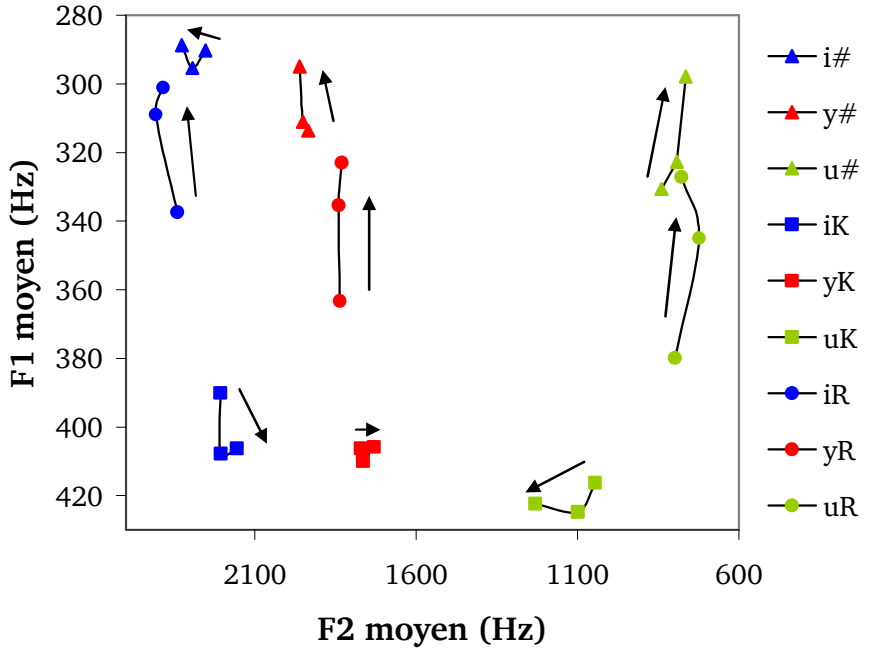


FIGURE 2 – Diagramme biformantique F_1 / F_2 présentant l'évolution moyenne des différentes classes vocaliques, tous locuteurs confondus

La TABLE 1 présente les taux de classification correcte obtenus lorsque différentes combinaisons de prédicteurs sont employées dans les analyses discriminantes.

		50 %		25, 75 %		25, 50, 75 %	
		sans durée	avec durée	sans durée	avec durée	sans durée	avec durée
F_1, F_2	Hz	58	82	68	87	73	85
	Score-Z	76	94	82	96	86	96
F_0, F_1, F_2	Hz	68	88	76	87	74	85
	Score-Z	81	95	87	89	89	95
F_1, F_2, F_3	Hz	66	90	76	93	79	92
	Score-Z	79	96	86	96	88	97
F_0, F_1, F_2, F_3	Hz	72	93	80	91	79	88
	Score-Z	84	96	89	96	92	96

TABLE 1 – Taux de classification correcte globaux (en %), résultats des analyses discriminantes visant à classifier les occurrences dans leur classe vocalique (i#), (iK), (iR), (y#), (yK), (yR), (u#), (uK), ou (uR) à partir de divers indices acoustiques relevés en différents points

On remarque tout d'abord que les données normalisées en scores-Z différencient plus efficacement les classes que les données brutes en Hertz, et ce, quels que soient les indices acoustiques et le nombre de points de mesure considérés. Conformément aux résultats d'Adank et coll. (2004), les valeurs relatives des composantes spectrales (ici exprimées en scores-Z) semblent donc un meilleur indicateur de l'identité des voyelles que leurs valeurs absolues en Hertz. Concernant la prise en compte de la dynamique spectrale, on peut constater que le relevé des indices acoustiques en deux points de mesure plutôt qu'en un seul permet d'améliorer la classification dans plusieurs cas. Cependant, lorsque la durée est également incluse dans les prédicteurs, l'impact positif de la dynamique sur la classification est moindre, voire disparaît selon les indices acoustiques considérés et selon que les données sont normalisées ou non. La prise en compte de trois points de mesure plutôt que deux améliore parfois la classification, mais jamais lorsque les données sont exprimées en Hertz et que la durée est incluse dans les prédicteurs, ni lorsque la F_0 est considérée.

La comparaison des matrices de confusion de chacune des analyses discriminantes dont le taux de classification correcte global est rapporté dans la TABLE 1 nous permet de déceler les classes vocaliques les mieux distinguées lorsqu'un indice acoustique donné est ajouté à F_1 et F_2 parmi les prédicteurs. Ainsi, même si les brèves (tendues et relâchées) se distinguent globalement des allongées sur le plan de la durée, l'inclusion ce prédicteur permet plus particulièrement de mieux différencier les tendues et les allongées ; les relâchées étant déjà relativement bien séparées des allongées par leur F_1. La prise en compte de F_3 réduit quant à elle la confusion entre les /i/ et les /y/. Étonnamment, bien que la F_0 moyenne des /i/ et celle des /y/ ne soient pas différentes (172 Hz tous contextes et locuteurs confondus, tant pour /i/ que pour /y/), la prise en compte de cet indice diminue également la confusion entre ces voyelles. En fait, comme on peut l'observer sur la FIGURE 3, la F_0 permet de séparer les /i/ produits par des hommes et les /y/ produits par des femmes, qui, autrement, présentent des valeurs de F_1 et de F_2 similaires.

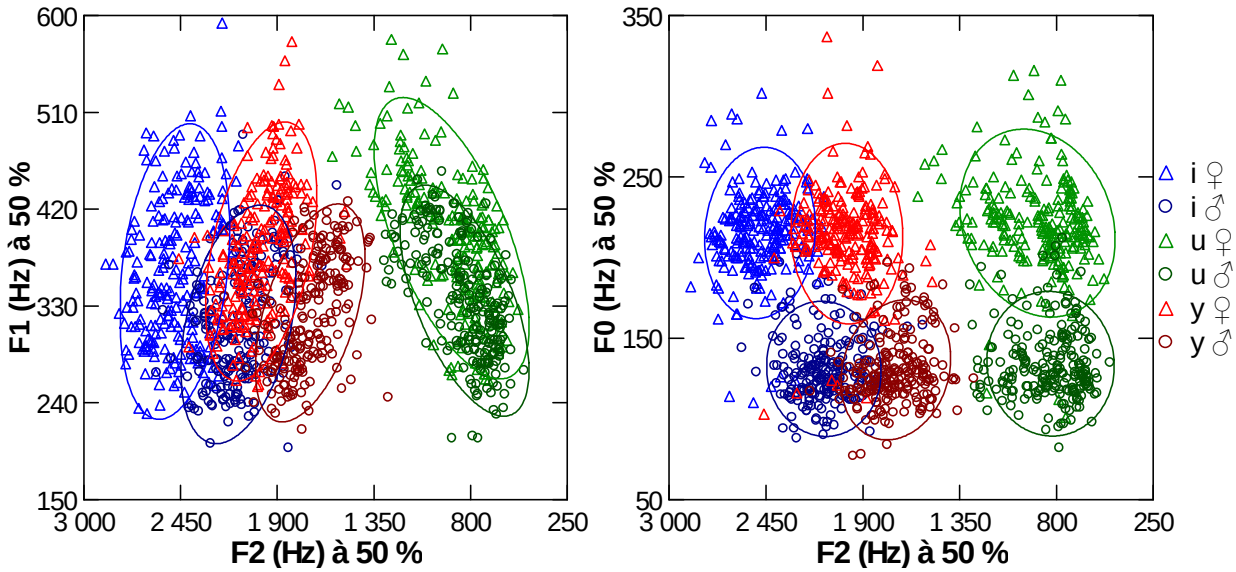


FIGURE 3 – Diagrammes de dispersion F_1 / F_2 et F_0 / F_2 présentant les 1343 occurrences de notre corpus à 50 % de leur durée regroupées d'après leur classe phonémique et le sexe du locuteur (les ellipses de dispersion englobent 86,5 % des occurrences d'un même groupe, soit ± 2 *s*)

En somme, les indices acoustiques pris en considération (F_0, F_1, F_2, F_3 et durée) participent tous à la différenciation des voyelles fermées tendues, relâchées et allongées du FQ. La normalisation des valeurs fréquentielles et la prise en compte de la dynamique spectrale permettent, dans bien des cas, une différenciation encore plus grande. Cela dit, diverses combinaisons de prédicteurs

permettent de distinguer plus efficacement que d'autres les classes vocaliques de notre corpus. Ainsi, avec un ensemble très économique de prédicteurs (F_1 et F_2 exprimés en scores-Z, à 50 % de la durée des occurrences, et durée vocalique) on obtient un taux de classification correcte de 94 %. Mais il est possible d'atteindre un taux similaire (93 %) sans recourir à la normalisation, en prenant également en compte la F_0 et F_3 ; ou encore en ajoutant seulement F_3 mais en prenant en compte deux points de mesure. Le taux de classification correcte le plus élevé (97 %) est quant à lui atteint lorsque la durée ainsi que F_1, F_2, et F_3 relevés en trois points de mesure et normalisés sont fournis comme prédicteurs. Des solutions plus économiques parviennent cependant à des résultats similaires (96 %). C'est notamment le cas lorsque les prédicteurs sont la durée ainsi que F_1 et F_2 relevés en deux points de mesure et normalisés.

4 Discussion et conclusion

Les résultats de la présente contribution doivent être considérés à la lumière des données sur lesquelles ils reposent et du cadre expérimental dans lequel cette étude s'inscrit. Au plan méthodologique, la catégorisation utilisée lors des analyses discriminantes repose sur la classe vocalique *attendue*, la prise en compte du caractère diphtongué / monophtongué des segments lors de cette tâche de classification automatique aurait-elle modifié le poids des différents prédicteurs considérés ? Par ailleurs, s'il a été question, dans l'interprétation de ces résultats, du *français québécois*, une généralisation des caractéristiques acoustiques identifiées à l'ensemble du FQ est sans nul doute prématurée. En outre, une analyse préliminaire de la variation régionale au sein de notre corpus laisse entrevoir certaines particularités à Saguenay. Les voyelles fermées allongées y seraient plus tendues et les relâchées devant consonne voisée, plus longues qu'à Rouyn-Noranda et Québec (sans pour autant que ces différences n'affectent les tendances globales observées dans la présente contribution). Rappelons également que notre corpus a été récolté à la fois dans des conditions très formelles et lors d'une tâche de lecture oralisée. Nos résultats devraient donc être mis à l'épreuve sur un échantillon plus vaste et diversifié, et dans une autre modalité de parole, telle que la parole spontanée. Soulignons par ailleurs que les indices acoustiques discriminants que nous avons mis au jour ne sont peut-être pas ceux sur lesquels repose effectivement la perception des voyelles à l'étude. Entre autres, il est sans doute peu probable que le système perceptif calcule des scores-Z.

Cela étant dit, nos résultats indiquent qu'au-delà de F_1 et F_2 au centre de l'état stable, la prise en compte de la dynamique spectrale constitue une plus-value, y compris pour la description acoustique de variantes brèves réputées ne pas présenter de variations de timbre. En outre, la dynamique spectrale qui distingue ces variantes leur semble propre ; plus précisément, elle ne semble pas strictement liée aux effets acoustiques de la coarticulation ; ce qui ne signifie pas que le contexte, entendu dans un sens plus large, n'a pas d'impact sur la dynamique spectrale. En effet, nous sommes ici en présence de variantes *contextuelles* qui diffèrent non seulement par leur timbre et leur durée, mais également par la structure syllabique dans laquelle elles apparaissent et la nature allongeante ou non de la consonne subséquente. Il est donc difficile d'attribuer à l'une ou l'autre ou à plusieurs de ces caractéristiques la dynamique spectrale observée.

Remerciements

Ce projet a reçu l'appui financier du *Fonds de recherche québécois – Société et culture* ainsi que du *Conseil de recherche en sciences humaines* du Canada.

Références

ADANK, P., R. SMITS et R. VAN HOUT (2004). A comparison of vowel normalization procedures for language variation research, ***JASA***, 116(5): 3099-3107.

ARNAUD, V., C. SIGOUIN et J.-P. ROY (2011). Acoustic description of Quebec French high vowels: First results. ***In Proceedings of the 17th ICPhS***, Hong Kong, 244-247.

CÔTÉ, M.-H. (2012). Laurentian French (Quebec). Extra vowels, missing schwas and surprising liaison consonants. *In* Gess, R., C. Lyche et T. Meisenburg (dir.). ***Phonological Variation in French. Illustrations from three continents***, John Benjamins Publishing Company, 235-274.

DUMAS, D. (1974). Durée vocalique et diphtongaison en français québécois. ***Cahier de linguistique***, (4):13-55.

— (1987), ***Nos façons de parler : les prononciations en français québécois***, PUQ.

GENDRON, J.-D. (1966). ***Tendances phonétiques du français parlé au Canada***, PUL.

LOBANOV, B. M. (1971). Classification of Russian vowels spoken by different speakers, ***JASA***, 49(2B):606-608.

MACKENZIE, L. et G. SANKOFF (2010). A quantitative analysis of diphthongization in Montreal French. ***UWPL: Selected Papers from NWAV 37***, 15(2):91-100.

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

MCLAUGHLIN, A. (1986). Une (autre) analyse de la distribution des variantes des voyelles hautes en français montréalais, ***Revue québécoise de linguistique théorique et appliquée***, 5(4): 21-60.

MORRISON, G. S. et P. F. ASSMANN (2013). ***Vowel Inherent Spectral Change***, Springer.

NEAREY, T. M. et P. F. ASSMANN (1986). Modeling the role of inherent spectral change in vowel identification. ***JASA***, 80(5):1297-1308.

PARADIS, C. (1985). ***An Acoustic Study of Variation and Change in the Vowel System of Chicoutimi and Jonquière (Québec),*** thèse de doctorat, University of Pennsylvania.

POLIQUIN, G. C. (2006). ***Canadian French Vowel Harmony***, thèse de doctorat, Harvard University.

SANTERRE, L. et J. MILLO (1978). Diphtongization in Montreal French. *In* Sankoff, D. (dir.). ***Linguistic Variation: Models and Methods***, Academic Press, 173-184.

SANTERRE, L., S.-P. DUFOUR et S. MCDUFF (1985). La perception de la diphtongaison : son importance dans les grands corpus, ***Revue québécoise de linguistique théorique et appliquée***, 4(4): 33-53.

SIGOUIN, C. (2013). ***Caractéristiques acoustiques des voyelles fermées tendues, relâchées et allongées en français québécois***, mémoire de maîtrise, Université du Québec à Chicoutimi.

YAEGER, M. (1979). ***Context-determined Variation in Montreal French Vowels***, thèse de doctorat, University of Pennsylvania.

Etude de la production des voyelles nasales du français chez des apprenants espagnols et colombiens

David Alejandro Bustamante, Angélique Amelot, Claire Pillot-Loiseau
Laboratoire de Phonétique et Phonologie (LPP) UMR7018, CNRS-Paris3/Sorbonne Nouvelle ; 19 rue des Bernardins, 75005 Paris
david.alejandro.bustamante@gmail.com

RESUME

L'objectif de cette étude est de décrire la production des voyelles nasales françaises dans des mots et des logatomes produits par deux apprenantes espagnoles, deux apprenantes colombiennes du Français Langue Etrangère (FLE) et deux françaises natives dans trois positions syllabiques : initiale absolue, initiale inter-consonantique dans une séquence CVCV, et position finale de mot. Nous avons comparé leurs productions à l'aide d'un accéléromètre piézoélectrique enregistrant les vibrations nasales. Les résultats montrent une importante variabilité dans les trois positions syllabiques concernant (1) la durée des voyelles nasales notamment chez les colombiennes et (2) le RMS (Root Mean Square) nasal. En position finale de mot, la différence de durée entre [a, ɛ, o] et [ɑ̃, ɛ̃, ɔ̃] n'est pas significative dans les deux groupes d'hispanophones. Le groupe colombien utilise plus d'énergie acoustique nasale par rapport au groupe espagnol.

ABSTRACT

Production of French nasal vowels by Spanish and Colombian learners of French

The aim of this article is to describe the production of the French nasal vowels in words and non-words produced by two Spanish female and two Colombian female learners of French as a Foreign Language (FFL) and two French native speakers in three different syllabic positions: word-initial, word-medial (interconsonantic position in a CVCV sequence), and final-word position. Their productions were compared using a piezoelectric accelerometer which recorded the nasal vibrations. The acoustic results show an important variability in the three syllabic positions regarding the duration of the nasal vowels and the nasal RMS (Root Mean Square) of the piezoelectric accelerometer signal. In final-word position, the difference of duration between [a, ɛ, o] and [ɑ̃, ɛ̃, ɔ̃] is not statistically significant in the two groups of Spanish speakers. The Colombian group uses more acoustic nasal energy than the Spanish group.

MOTS-CLES : voyelles nasales, accéléromètre piézoélectrique, FLE (Français Langue Etrangère), RMS nasal, RMS oral, nasalité vocalique, position syllabique, locuteurs hispanophones

KEY WORDS: nasal vowels, piezoelectric accelerometer, French as a Foreign Language, nasal RMS, oral RMS, vocalic nasality, syllabic position, Spanish speakers

Introduction

L'attribut de nasalité vocalique constitue un trait distinctif en français, comme dans d'autres langues indoeuropéennes comme par exemple le polonais et le portugais. Selon la base des données UPSID (UCLA Phonological Standard Inventory Data)[1], le trait de nasalité vocalique est présent dans 22,62% des langues du monde (soit 102 langues répertoriées dans la base) permettant le contraste phonologique entre voyelles orales et voyelles nasales. La nasalité consonantique, quant à elle, est présente dans 96,45% des langues de la base (soit 435 langues). En espagnol, les voyelles nasales n'existent pas, mais l'inventaire phonologique de cette langue comporte la nasalité dans le système consonantique. Ce n'est que sur le plan phonétique que les voyelles sont contextuellement nasalisées lorsqu'elles sont proches des consonnes nasales adjacentes. L'espagnol possède cinq voyelles orales ([i, e, a, o, u]) et cinq allophones nasals correspondants, ([ĩ, ẽ, ã, õ, ũ]), (Quilis, 2006), alors qu'il y a en français douze voyelles orales et trois voyelles nasales ([ɑ̃, ɛ̃, ɔ̃]), ou quatre selon une norme régionale ([œ̃]) (Detey & Le Gac, 2010b).

A notre connaissance, il existe très peu d'études concernant la production des voyelles nasales du Français langue Etrangère (FLE) par des apprenants hispanophones : Detey *et al.* (2010) ont travaillé avec des locuteurs japonais et espagnols sur l'identification de la voyelle nasale et l'évaluation du degré de « consonantisation » (présence/absence de la consonne nasale post-vocalique) de leurs réalisations ; Nawafleh (2013), avec des apprenants jordaniens, constate la difficulté de perception et de production que présentent ces locuteurs en confondant les voyelles nasales entre elles et avec les voyelles orales. Par ailleurs, l'accéléromètre piézoélectrique constitue un outil potentiellement didactique pour l'étude de la nasalité en FLE (Pillot-Loiseau *et al.*, 2011 ; Brkan *et al.*, 2012a, 2012b).

Les voyelles nasales représentent une source de difficulté d'apprentissage pour les apprenants hispanophones du FLE. Cette difficulté permet d'évoquer des travaux sur la perception des sons étrangers comme le *Speech Learning Model* de Flege (1995), qui suggère qu'un apprenant ne dispose que d'un inventaire phonologique pour la L1 et la L2, rendant difficile la discrimination du contraste dans la L2, et entre la L1 et la L2. Pareillement, les travaux de Best (*Perceptual Assimilation Model,* 1995) signalent que les sons d'une langue étrangère sont perçus en fonction des similarités ou différences phonétiques (articulatoires et acoustiques) entre langue native et langue non-native.

L'hypothèse principale de notre étude est que les apprenants hispanophones produisent les voyelles nasales françaises ([ɑ̃, ɛ̃, ɔ̃]) les plus proches de la réalisation des francophones natifs, quand celles-ci se situent en position finale de mot ; les hypothèses secondaires sont que ces productions sont différentes 1) entre les colombiennes et les espagnoles et 2) entre ces apprenantes et deux natives françaises également enregistrées. Une assimilation de la nasalisation dans certains dialectes d'Espagne et d'Amérique hispanique indique que les consonnes nasales en position finale de mot tendent parfois à disparaître tout en nasalisant la voyelle précédente (Zamora, 1967). Quoique moindre en

[1] La base de données UPSID est consultable à l'adresse suivante : http://web.phonetik.uni-frankfurt.de/cgi-bin/upsid_sounds.cgi [consulté le 28 avril 2012].

amplitude, nous considérons cette nasalisation dans la variation dialectale de l'espagnol comme un facteur intéressant pour la didactique de la prononciation des voyelles nasales du FLE.

1. Méthode

1.1 Matériel

Nous avons utilisé un accéléromètre piézoélectrique double (K&K Sound). Il s'agit d'une technique pour obtenir des informations sur la résonance nasale se produisant au niveau des cartilages latéraux du nez lors de la production des voyelles nasales. La méthodologie de l'enregistrement consiste à placer deux pastilles d'un diamètre de 0,5 cm, fixées avec un adhésif double face sur chaque côté du nez du locuteur (Lippman, 1981), afin de capter les vibrations provenant des cavités nasales. L'accéléromètre piézoélectrique est du type K&K Sound, branché à un préamplificateur (Presonus à 40 dB), et celui-ci à une carte d'acquisition (MOTU UltraLite mk3 hybride) et à un ordinateur. Calibrés avant l'enregistrement, l'accéléromètre piézoélectrique capture le signal acoustique nasal, et un microphone serre-tête (AKG C420L), enregistre le signal acoustique oral.

1.2 Participants

Nous avons enregistré deux locutrices françaises, deux locutrices colombiennes et deux espagnoles apprenant le français langue étrangère (FLE). Les locutrices françaises ont en moyenne 26 ans (E.T. 1,4), les locutrices colombiennes ont en moyenne 30,5 ans (E.T. 3,5) et les espagnoles ont en moyenne de 25,5 ans (E.T. 2,1) : elles sont toutes des étudiantes dans un établissement d'éducation supérieure. Le niveau de français des locutrices colombiennes est C2 pour l'une (Locutrice Col1), et C1 pour l'autre (Locutrice Col2), B2 pour une locutrice espagnole (Locutrice Esp2), et C1 pour l'autre (Locutrice Esp1)[2]. Les étudiantes colombiennes sont en France, depuis en moyenne de 15,2 mois, alors que les locutrices espagnoles ont un temps de résidence, en moyenne de 27,5 mois. Tous les participants hispanophones ont exprimé avoir des difficultés pour prononcer les voyelles nasales du français.

1.3 Corpus

Nous avons utilisé un corpus de 50 mots du français bisyllabiques où sont opposées les voyelles orales et les voyelles nasales dites « correspondantes » dans 25 phrases cadre (chaque phrase comporte donc un mot avec une voyelle orale et un autre mot avec la voyelle nasale « correspondante »), et un corpus de 36 logatomes bisyllabiques présentant le même type d'opposition entre voyelles orales et voyelle nasales : ayant les oppositions /ɛ - ɛ̃/, /a - ɑ̃/ et /o - ɔ̃/, ils se distinguent notamment par l'abaissement du voile du palais des voyelles nasales, ainsi que par une position spécifique de la langue et des lèvres. L'objectif est d'observer la production des voyelles nasales chez les apprenantes hispanophones dans trois positions syllabiques différentes : en position initiale absolue (#_CV) ; en position initiale inter-consonantique (C_CV) ; et en position

[2] Selon le Cadre européen commun de référence pour les langues (CECR), le niveau C2 est celui d'une « Maîtrise » de la langue, C1 est « Autonome », et B2 correspond à « Avancé ».

finale (CVC_#). Pour obtenir une production optimale des voyelles nasales, les deux corpus présentent un contexte phonétique défini par des consonnes occlusives et fricatives sourdes entourant les voyelles nasales ([p, t, k, f, ʃ]), afin d'éviter de possibles influences des consonnes voisées sur le signal de l'accéléromètre piézoélectrique avant et après la voyelle. Les phrases du corpus correspondant aux différentes positions syllabiques sont (Où V = [a, ɛ, o] ; C = [p, t, k, f, ʃ] ; et V_n = [ɑ̃, ɛ̃, ɔ̃]), exemple : Je répète frappa plus que frappant six fois.

Position	Phrases cadre	Nombre
#_CV	« Je répète **VCVC** plus que **VnCVC** six fois. »	367 phrases
C_CV	« Je répète **CVCV** plus que **CVnCV** six fois. »	390 phrases
CVC_#	« Je répète **CVCV** plus que **CVCVn** six fois. »	445 phrases

Tableau 1. Corpus de mots et de logatomes avec le type de position syllabique, le type de phrase cadre correspondant, et le nombre de phrases réalisées par toutes les locutrices.

1.4 Procédure

Les locuteurs ont été enregistrés en chambre sourde sur une tâche de lecture des corpus contenant les phrases cadre de mots et de logatomes (cinq répétitions) ; ils ont lu les phrases 5 fois chacune. Au total, nous avons enregistré 2580 occurrences, mais nous avons dû sélectionner 1202 occurrences en raison de réalisations inattendues des locutrices hispanophones, effectuant des liaisons d'ordre grammatical, lorsqu'une phrase présentait une voyelle nasale en position initiale absolue (#_CV), même si le corpus ne le proposait pas. Par exemple, « *Je répè**te a**ccord plus **qu'en**core six fois* », au lieu de « *Je répète accord plus **que encore** six fois* ».

1.5 Analyse des données

Les données ont été segmentées et étiquetées manuellement avec le logiciel PRAAT (Boersma et Weenink, 2012). En outre, puisque nos analyses portent sur la production des voyelles nasales chez quatre apprenantes hispanophones et deux locutrices françaises en fonction de l'influence que peut exercer la position de la syllabe, nous avons mesuré la durée des voyelles orales et nasales dans les trois positions syllabiques proposées, ainsi que l'amplitude du signal de l'accéléromètre piézoélectrique des voyelles nasales à travers des valeurs obtenues avec le Root Mean Square (RMS), de même que les voyelles orales produisant aussi de l'énergie acoustique dans les cavités nasales (Horii, 1980). Nous avons utilisé le ratio intra-locutrice du RMS du signal capté avec l'accéléromètre piézoélectrique (RMS nasal) au RMS du signal capté par le microphone (RMS oral) des voyelles pour comparer l'amplitude de nasalité des locutrices. Comme la calibration du signal de l'accéléromètre piézoélectrique n'était pas la même pour toutes les locutrices, nous avons comparé les ratios du RMS du signal nasal des voyelles orales et des voyelles nasales (RMS nasal et RMS oral), comme une mesure pour observer l'amplitude proportionnelle de l'acoustique nasale à l'acoustique orale de chacune des locutrices.

2. Résultats

2.1 Durée des voyelles orales et nasales

La durée des voyelles orales et les voyelles nasales des toutes les locutrices montre une différence attestée en français concernant une longueur plus importante des voyelles nasales (Delattre et Monnot, 1968, Pillot-Loiseau *et al.* 2011, Brkan *et al.* 2012). Cette distinction est plus visible dans les deux premières positions de mot et de logatome (figures 1, 2 et 3) : position initiale absolue (#_CV) et position initiale inter-consonantique (C_CV).

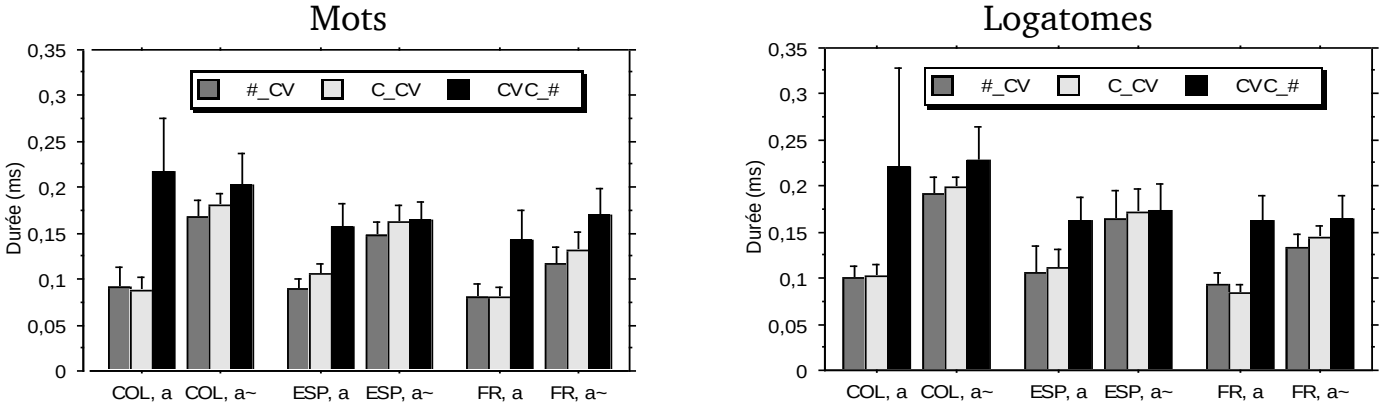


FIGURE 1 – Durée (ms) de [a] et [ã], réalisés par les locutrices colombiennes (COL), espagnoles (ESP) et françaises (FR) dans trois positions syllabiques de mot (gauche) et de logatome (droite).

Illustré dans la figure 1, le résultat d'une Anova à deux facteurs montre qu'il y a une différence significative de durée entre [a] et [ã], entre les groupes des locutrices dans les trois positions syllabiques : F*gr*(2,450)=88,7 ; p<0,0001. Le facteur position syllabique montre un effet significatif dans la réalisation de ces voyelles par les trois groupes : F*pos*(2,450)=267,2 ; p<0,0001. Un autre test Anova pour chaque position syllabique montre aussi qu'il y a une différence statistiquement importante de la durée entre [a] et [ã] dans les deux premières positions syllabique dans les trois groupes, mais ce n'est pas le cas en position finale de mot : l'effet du facteur groupe est très significatif (F*gr*(2,168)=42,1 ; p<0,0001), mais non pas pour le facteur voyelle (F*voy*(1,168)=1,6 ; p=209). Un test-PLSD de Post hoc Fisher confirme l'absence de différence significative entre les deux voyelles en position finale pour les trois groupes (p=0,2610), mais il indique que ce sont les locutrices colombiennes qui réalisent le [a] et [ã] significativement plus longs que les deux autres groupes en position finale (p<0,0001).

En outre, dans le corpus de logatomes, les effets des facteurs groupe et position syllabique sont très significatifs pour la réalisation de [a] et [ã] : F*gr*(2,352)=55 ; p<0,0001 ; F*pos*(2,352)=106,5 ; p<0,0001. On observe dans la figure 1 que dans les deux positions initiales la différence de [a] et [ã] est maintenue dans les trois groupes, et qu'il n'y a pas de différence significative de durée entre [a] et [ã] en position finale de logatome (F(1,122)=0,548 ; p=0,46), ainsi que de mot, chez toutes les locutrices.

Dans la figure 2, l'effet de la position syllabique sur [ɛ] et [ɛ̃], dans le corpus de mot et de logatome, est très significatif (p<0,0001). Quant à la durée entre ces voyelles, il y a une différence statistiquement importante dans les trois positions syllabiques de mot (F*voy*(1,410)=217,5 ; p<0,0001), ainsi que de logatome (F*voy*(1,330)=247,3 ; p<0,0001), sauf en position finale (F*voy*(1,122)=0,118 ; p=0,73).

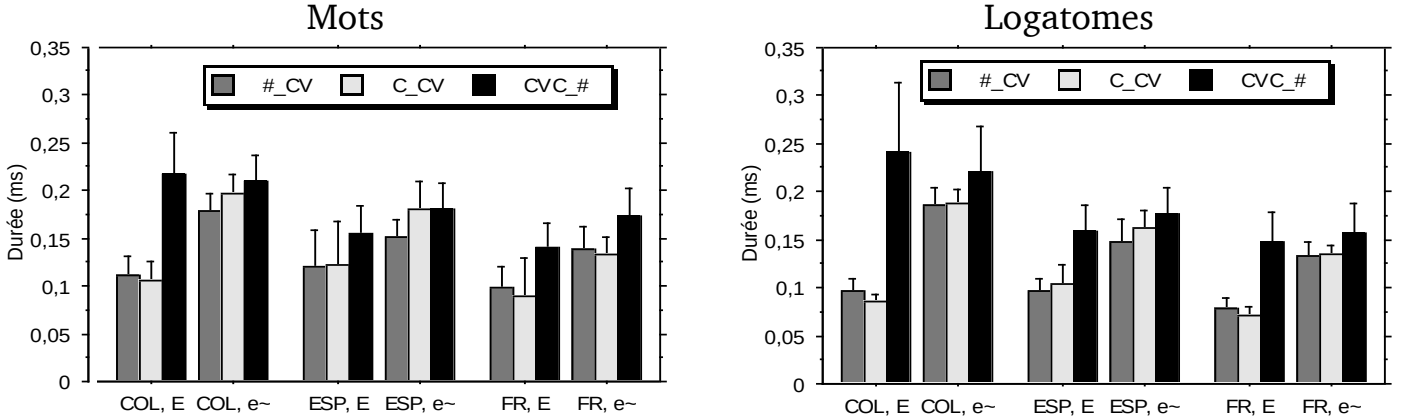


FIGURE 2 – Durée (ms) de [ɛ] et [ɛ̃], réalisés par les locutrices colombiennes (COL), espagnoles (ESP) et françaises (FR) dans trois positions syllabiques de mot (gauche) et de logatome (droite).

La distinction de durée entre [o] et [ɔ̃] (figure 3) est significative pour les trois groupes de locutrices dans les positions syllabiques observées (mot et logatome), sauf en position finale : F*voy*(1,410)=346,4 ; p<0,0001 ; F*voy*(1,348)=390,5 ; p<0,0001. De plus, comme pour les autres paires vocaliques, il y a un effet significatif de la position syllabique pour la durée entre [o] et [ɔ̃] de toutes les locutrices : dans les mots, F*pos*(2,410)=123,8 ; p<0,0001 ; et dans les logatomes, F*pos*(2,348)=107,6 ; p<0,0001.

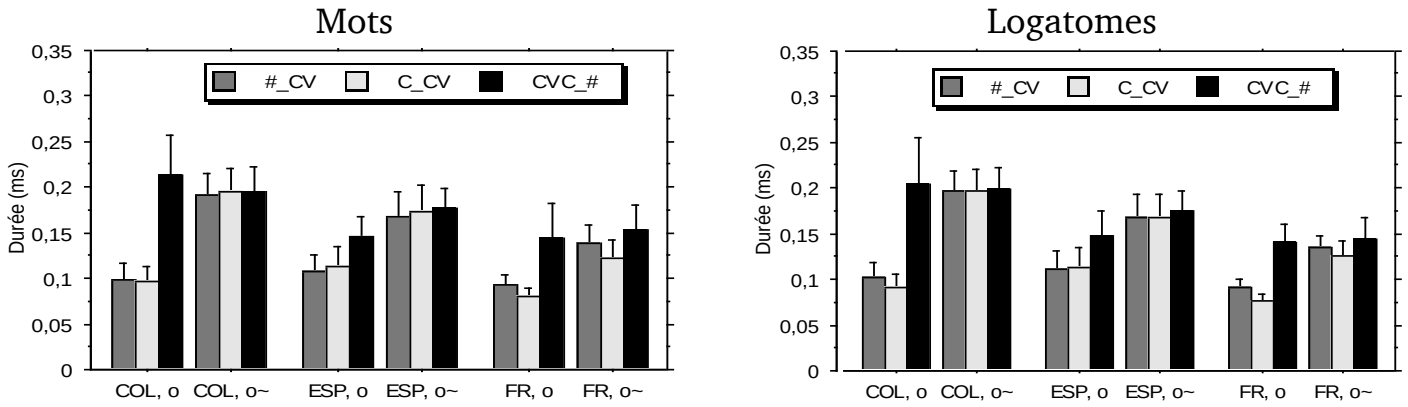


FIGURE 3. Durée (ms) de [o] et [ɔ̃], réalisés par les locutrices colombiennes (COL), espagnoles (ESP) et françaises (FR) dans trois positions syllabiques de mot (gauche) et de logatome (droite).

2.2 Ratio du RMS nasal au RMS oral des voyelles orales et nasales des locutrices colombiennes, espagnoles et françaises

Le résultat de la comparaison du ratio du RMS nasal sur le RMS oral des voyelles orales et des voyelles nasales permet de repérer un taux de nasalité chez toutes les locutrices. Dans la figure 4, la comparaison des données permet d'observer un taux de nasalité de [ɑ̃] plus important par rapport à [a] chez toutes les locutrices dans les trois positions syllabiques. Particulièrement, la locutrice Col1 présente un taux de nasalité beaucoup plus important de la voyelle nasale dans les trois positions syllabiques par rapport aux autres locutrices.

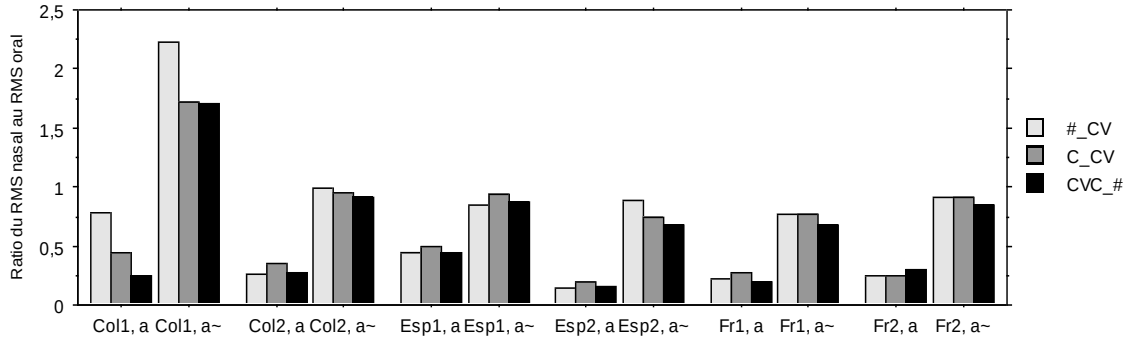


FIGURE 4. Ratio du RMS nasal au RMS oral de [a] et de [ã] des locutrices colombiennes, espagnoles et françaises dans trois positions syllabiques (mots et logatomes tous confondus).

En position initiale absolue (#_CV), la locutrice Col1 présente plus de nasalité dans [ã] que les autres locutrices. Les autres locutrices hispanophones semblent utiliser un même niveau de nasalité que les locutrices françaises. Par ailleurs, on observe que la locutrice Esp1 a un ratio de nasalité supérieur pour [a] dans les trois positions syllabiques par rapport aux autres locutrices, sauf pour Col1 en position initiale absolue.

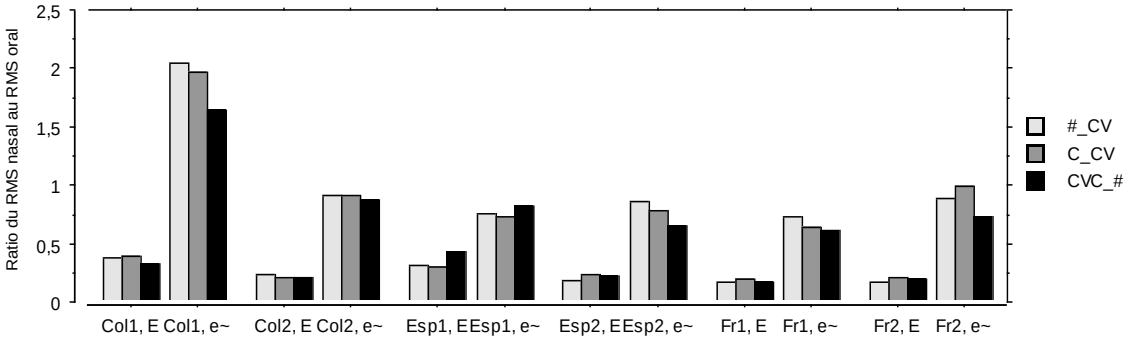


FIGURE 5. Ratio du RMS nasal au RMS oral de [ɛ] et de [ɛ̃] des locutrices colombiennes, espagnoles et françaises dans trois positions syllabiques (mots et logatomes tous confondus).

Dans la figure 5, la locutrice Col1 présente à nouveau un taux de nasalité plus important pour la voyelle nasale [ɛ̃] dans les trois positions syllabiques que les autres locutrices.

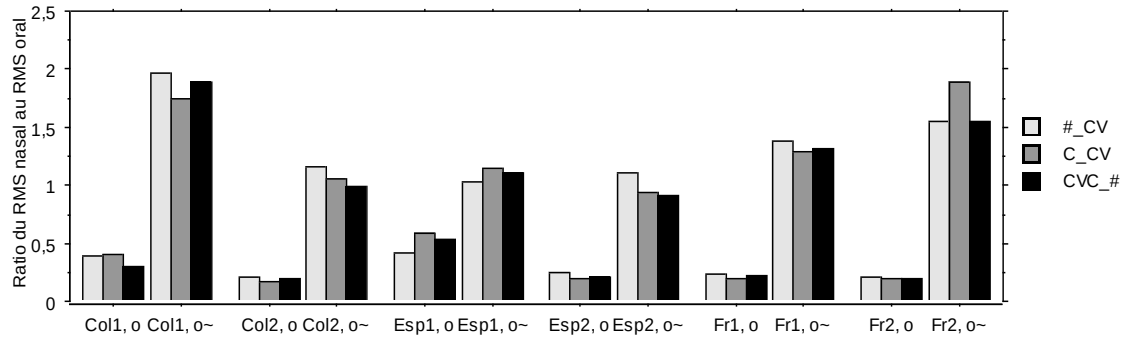


FIGURE 6. Ratio du RMS nasal au RMS oral de [o] et de [õ] des locutrices colombiennes, espagnoles et françaises dans trois positions syllabiques (mots et logatomes tous confondus).

Enfin, on observe (figure 6) que la voyelle [õ] exige plus de nasalité que les autres voyelles nasales, notamment pour les locutrices françaises. La locutrice Col2 et les deux espagnoles (Esp1, Esp2) présentent moins de nasalité pour [õ] dans les trois positions syllabiques que les locutrices françaises et la locutrice Col1.

3. Discussion et perspectives

Les résultats sur la durée des voyelles orales et nasales montrent dans cette étude que la position initiale absolue (#_CV) et la position initiale inter-consonantique (C_CV) sont des contextes syllabiques où la distinction de durée entre les voyelles nasales et les voyelles orales dites "correspondantes" est maintenue par toutes les locutrices. On observe qu'il y a une tendance chez les locutrices hispanophones, particulièrement une locutrice colombienne (Col1), à réaliser des voyelles orales et nasales plus longues que les locutrices françaises. Cependant, ces réalisations des apprenantes hispanophones peuvent être considérées comme appropriées, car la distinction de durée est maintenue dans les deux positions initiales, tout comme les françaises natives. Le ratio du RMS nasal au RMS oral a permis de visualiser un taux de nasalité des voyelles orales et nasales, où celui des voyelles nasales est plus important. La nasalité de [ɔ̃] chez toutes les locutrices est plus importante que pour les autres voyelles nasales, sauf pour la locutrice Col1 qui présente plus de nasalité dans [ɑ̃], spécialement en position initiale absolue. Malgré le fait que cette étude ne rend pas compte de l'articulation orale des voyelles nasales, ni de la position des lèvres, ni de la langue, il est possible d'affirmer que les locutrices ayant le niveau de français C1 (Col2 et Esp2) semblent maîtriser la production des voyelles nasales, ainsi que la locutrice Esp1 qui, ayant un niveau B2, présente aussi une réalisation distinctive des voyelles orales et nasales. La locutrice Col1 fait cette distinction encore plus marquée, bien que son niveau soit C2, il est possible qu'elle réalise une sur-articulation des voyelles nasales au moment d'être enregistrée. La variabilité observée ne permet de confirmer partiellement l'hypothèse sur la production des locutrices hispanophones des voyelles nasales en position finale : même si elles ne réalisent pas la distinction de durée dans cette position entre les voyelles orales et nasales comme les françaises natives, les locutrices hispanophones produisent une longueur des voyelles plus important ; et les locutrices colombiennes réalisent les voyelles nasales et orales plus longues en position finale que les locutrices espagnoles.

Vu qu'il y a peu de littérature sur la production des voyelles nasales du FLE chez des hispanophones, cette étude peut être un apport dans la recherche sur la production des sons du FLE, car les données ici présentées peuvent permettre la comparaison avec d'autres apprenants hispanophones du FLE. La présence de nasalité dans les voyelles orales représente également un facteur déterminant de référence pour la production des voyelles nasales du FLE. Un nombre assez faible de données, particulièrement pour les locutrices hispanophones dont nous avons dû éliminer certaines de leurs données en raison d'une lecture inattendue effectuant la liaison dans le corpus de mots (e.g. « *Je répète accord plus qu'encore…* »), même si le corpus ne proposait pas cette version, a représenté une limitation pour cette étude. Une reconsidération du corpus reste à faire pour de futures études. En outre, pour compléter cette étude, un nombre plus important de sujets hispanophones doit permettre de constituer des groupes afin de mieux aborder les tendances sur la production de la nasalité vocalique, de même que d'autres tâches de production sont à tenir en considération : répétition et parole spontanée. Enfin, la visualisation des données par rapport à un taux de nasalité des voyelles nasales pourrait se réaliser par une observation plus fine de l'endroit, dans le segment, où l'apprenant produit le plus de nasalité. Une étude future devrait intégrer la perception de ces voyelles afin d'étudier la variabilité observée dans la production.

Références

BEST, C. (1995). "A direct realist view of cross-language speech perception", In Strange, W. (éd.) *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, York, Timonium, MD: York Press, pp. 171–204.

BOERSMA, P. et WEENINK, D. (2012). Praat: doing phonetics by computer (Version 5.3.34), [en ligne] www.praat.org [consulté le 19 novembre 2012].

BRKAN, A., AMELOT, A., et PILLOT-LOISEAU, C. (2012a). Utilisation d'un accéléromètre piézoélectrique pour l'étude de la nasalité du Français Langue Etrangère. *In Conférence conjointe JEP-TALN-RECITAL 2012*, volume 1: JEP, Grenoble, 4 au 8 juin 2012, pages 689–696.

BRKAN, A., PILLOT-LOISEAU, C., AMELOT, A. (2012b). Etude de la nasalité du FLE à l'aide d'un accéléromètre piézoélectrique chez les locutrices natives et bosniaques: comparaison des données du corpus IPFC et d'un corpus spécifique. *Journées IPFC*, 10-11 décembre 2012, Paris.

DELATTRE P. et MONNOT M. (1968). The role of duration in the identification of French nasal vowels. *International Review of Applied Linguistics* 6, pages 267-288.

DETEY, S., RACINE, I., KAWAGUCHI, Y., ZAY, F., & BUEHLER, N. (2010a). Évaluation des voyelles nasales en français en L2 en production : de la nécessité d'un corpus multitâches. In: Neveu, F., Durand, J., Klingler, T., Prévost S., Muni-Toké V. (éds.). *In Actes de CMLF'10* [CD-ROM], ILF, 1289-1301.

DETEY, S., & LE GAC, D. (2010b). Le français de référence : quels locuteurs ? *In* (Detey, S., Durand, J., Laks, B., & Lyche, C., 2010) *Les variétés du français parlé dans l'espace francophone : Ressources pour l'enseignement.* Éditions OPHRYS, Paris, pages167-180.

FLEGE, J. (1995). "Second language speech learning: Theory, findings, and problems," In Strange, W., (éd) *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, York, Timonium, MD: York Press, pp. 233–277.

HORII, Y. (1980). An accelerometric Approach to Nasality Measurement: a preliminary report. *Cleft Palate Journal*, volume 17 (3), pages 254-261.

LIPPMAN, R.P. (1981). Detecting nasalization using a low cost miniature accelerometer. *Journal of Speech and Hearing Research*, volume 24, pages 314-317.

NAWAFLEH, A. (2013). *Difficultés de prononciation et de perception de voyelles du français par des apprenants arabophones: apprenants jordaniens.* Presses Académiques Francophones, 535 pages.

PILLOT-LOISEAU, C., AMELOT, A., et BRKAN, A. (2011). Utilisation d'un accéléromètre piézoélectrique pour l'étude de la nasalité du FLE : étude préliminaire. *Journées IPFC*, 6 décembre 2011, Paris.

QUILIS, A. (2006). *Tratado de fonología y fonética españolas*. GREDOS, Biblioteca Románica Hispánica. Madrid.

ZAMORA, V. (1967). *Dialectología española*. GREDOS, Madrid.

Portrait acoustique d'une variation régionale en français québécois contemporain : l'ouverture de la voyelle /ɛ/ en finale de mot

Josiane Riverin-Coutlée et Vincent Arnaud
Département des arts et lettres, Université du Québec à Chicoutimi
555, boul. de l'Université, Chicoutimi (Québec) Canada G7H 2B1
josiane.riverin-coutlee@uqac.ca, vincent.arnaud@uqac.ca

RESUME _____

Cette contribution est consacrée à l'étude acoustique d'un phénomène phonétique du français québécois qui à ce jour n'a reçu que peu d'attention : l'ouverture du /ɛ/ en finale absolue. Considéré en perte de vitesse au Québec, ce phénomène serait plus fréquemment produit par des locuteurs issus de milieux populaires en situation informelle, mais sans distribution géographique précise. L'analyse de la durée et de l'évolution temporelle des trois premiers formants de 473 occurrences produites en contexte formel (tâche de lecture en laboratoire) par 40 jeunes universitaires originaires des villes de Saguenay et de Québec (deux centres urbains de l'est du Québec) suggère néanmoins une vitalité du phénomène à Saguenay. Les résultats d'une série d'analyses discriminantes quadratiques indiquent que la prise en compte de F_3, de la durée et de la dynamique spectrale en plus de F_1 et F_2 permet d'affiner la classification des occurrences en fonction de l'origine géographique des locuteurs.

ABSTRACT_____

Acoustical Description of Word-Final /ɛ/ Lowering in Contemporary Quebec French: A Regional Variation

This acoustic study focuses on a Quebec French phonetic feature that has not been studied extensively: the lowering of /ɛ/ in absolute word-final position. This phenomenon, which is considered on the decline in the province of Quebec, has been said to happen more frequently with speakers from lower social classes, in informal situations, but without any precise geographical distribution. Our results contradictorily suggest that the lowering of /ɛ/ is still very much alive among speakers from Saguenay (a city in Eastern Quebec), based on the analysis of duration and temporal evolution of the first three formants from 473 tokens produced during a reading task (formal context) by 40 undergraduate students native of Saguenay and Quebec City. A series of quadratic discriminant analyses shows that taking into account F_3, duration and spectral changes, in addition to F_1 and F_2, improves dramatically the classification accuracy of tokens according to speakers' birthplaces.

MOTS-CLES : VISC, français québécois, variation régionale, voyelle /ɛ/, dynamique spectrale, analyse discriminante
KEYWORDS: VISC, Quebec French, regional variation, vowel /ɛ/, formant dynamics, discriminant analysis

1 Problématique et objectifs

Le français québécois (désormais FQ) est caractérisé par un faisceau de traits phonétiques qui, pour certains d'entre eux, ont été largement documentés, par exemple l'affrication des consonnes /t d/ ou la diphtongaison (voir entre autres Dumas, 1987). D'autres phénomènes, moins fréquemment mis de l'avant, font néanmoins partie intégrante du phonétisme du français en usage au Québec, parmi lesquels l'ouverture de la voyelle /ɛ/ en [æ] ou [a] en finale absolue (le mot *poulet* prononcé [pulæ]). Bien que considéré comme courant dès les premières descriptions du français en usage dans cette province (Squair, 1888), ce phénomène a reçu peu d'attention et n'a, au mieux, fait l'objet que de brèves remarques dans des ouvrages et études généralistes reposant sur l'analyse exclusivement auditive de données récoltées il y a plusieurs décennies. Gendron (1966 : 66) considère le phénomène comme caractéristique de locuteurs peu scolarisés, issus de milieux populaires. Dumas (1987 : 137) indique pour sa part qu'il s'agit d'une prononciation beaucoup plus fréquente en situation de communication informelle que formelle. Sur le site PHONO, qui constitue une synthèse d'une majeure partie des écrits sur le phonétisme du français en usage au Québec, Paradis et Dolbec (1998) constatent qu'il existe peu d'information en ce qui a trait à la répartition géographique du phénomène et que les données parcellaires récoltées à ce sujet laissent présumer qu'il serait commun à plusieurs régions du Québec. Plus récemment, Morin (2002 : 48) laisse entendre que le phénomène d'ouverture du /ɛ/ en finale absolue pourrait désormais être en situation de recul dans la province.

L'objectif premier de cette contribution est donc de faire le point sur l'éventuelle vitalité du phénomène en FQ. Pour ce faire, nous nous proposons de procéder à une analyse acoustique, ce qui n'a, à notre connaissance, jamais été effectué.

Les choix méthodologiques traditionnellement adoptés lors de l'analyse acoustique des voyelles sont fondés sur la prise en compte d'une seule estimation de la fréquence centrale des deux premiers formants (désormais F_1 et F_2) dans la portion médiane de chaque segment. Héritage des études pionnières des Laboratoires Haskins (Delattre et coll., 1952), ce paradigme d'analyse et de représentation des voyelles est encore d'actualité. Il peut néanmoins être enrichi en intégrant des indices acoustiques supplémentaires comme la durée, la fréquence fondamentale (désormais f_0), le troisième formant (F_3), dont l'intérêt pour la description acoustique des voyelles antérieures arrondies n'est plus à confirmer (Schwartz et coll., 1993), et le quatrième formant (F_4), mais aussi en prenant en compte la dynamique temporelle de ces divers indices (Arnaud, 2010).

En incluant la dynamique spectrale dans la paramétrisation acoustique des voyelles analysées, cette contribution se situe dans le prolongement de l'hypothèse du VISC (*Vowel Inherent Spectral Change*, Nearey et Assmann, 1986 : 1297). Peterson et Barney (1952 : 184) mettaient déjà en doute le fait que l'analyse acoustique d'une voyelle puisse être réduite à l'extraction d'indices chiffrés en une seule section temporelle. Cependant, Hillenbrand (2013 : 11, dans Morrison et Assmann, 2013) souligne qu'il faudra attendre les années 1980 pour que l'éventuel rôle des changements spectraux dans la description des timbres vocaliques soit mis à l'épreuve.

Les études s'appuyant sur l'hypothèse du VISC, dont la portée perceptive a été évaluée à plusieurs reprises en anglais américain, notamment en utilisant des *silent centre syllables*

(Jenkins et coll., 1983) et des stimuli synthétiques (Hillenbrand et Nearey, 1999), tendent à montrer que la dynamique spectrale des voyelles leur est inhérente et n'est pas strictement attribuable aux effets acoustiques de la coarticulation (Hillenbrand et coll., 2001).

Toutes les voyelles anglo-américaines, même celles traditionnellement considérées comme des monophtongues, présentent une trajectoire substantielle dans un espace de dispersion F1×F2 entre leur établissement (*onset*) et leur queue (*offset*). La trajectoire respective de ces voyelles demeure récurrente, qu'elles soient produites en contexte /hVd/ (dit neutre) ou dans divers contextes consonantiques (Hillenbrand et coll., 1995; Hillenbrand et coll., 2001).

La prise en compte de la dynamique spectrale favorise également la distinction entre les voyelles fermées tendues et leurs contreparties relâchées en anglais américain (Hillenbrand et coll., 1995) ou en FQ (Arnaud et coll., 2011; Sigouin, 2013). Les tendues et les relâchées présentent notamment des trajectoires opposées dans un espace de dispersion F_1×F_2, et ce, dans ces deux idiomes. Enfin, la prise en compte de la dynamique spectrale a également permis en anglais américain de décrire et de quantifier diverses variations régionales (Fox et Jacewicz, 2009).

Cette contribution propose donc de décrire la structure du /ɛ/ apparaissant en finale absolue en FQ et cela, en enrichissant le paradigme expérimental classique d'analyse acoustique des voyelles où seule la fréquence centrale de F_1 et de F_2 au « centre de l'état stable » des voyelles est considérée. Nous nous attacherons à montrer que la prise en compte de la durée et de l'évolution temporelle des trois premiers formants permet de : a) caractériser plus finement les propriétés de la voyelle /ɛ/ en finale absolue qu'une simple modélisation statique bidimensionnelle ; b) rendre compte, s'il y a lieu, d'éventuelles différences régionales concernant l'ouverture de cette voyelle.

2 Repères méthodologiques

L'échantillon de témoins dont les productions ont été analysées est constitué de 40 étudiants universitaires, autant d'hommes que de femmes, âgés de 20 à 29 ans, tous locuteurs natifs du FQ et originaires à parité des villes de Québec et de Saguenay. Ces deux centres urbains de l'est du Québec ne sont distants que de 200 km, mais se situent respectivement aux extrémités sud et nord de la Réserve faunique des Laurentides, territoire inhabité qui a longtemps constitué une barrière naturelle entre les deux villes (Paradis, 1985).

Chaque locuteur a été amené à lire à haute voix quelque 500 phrases porteuses pentasyllabiques sémantiquement significatives contenant, en position finale d'énoncé, des mots cibles[1]. Parmi ceux-ci, 12 se terminent par la voyelle /ɛ/ : *baie*, *épais*, *fait*, *gilet*, *guet*, *jet*, *lait*, *paix*, *poulet*, *prêt*, *raie* et *taie.* Les voyelles analysées sont extraites de la répétition isolée, à la suite d'une demande de reformulation de la part de l'enquêteur, de ces mots cibles. La tâche demandée aux témoins et les contraintes du lexique français ne nous permettaient pas de contrôler le contexte consonantique précédant la voyelle.

Les locuteurs ont été enregistrés en chambre anéchoïque au format numérique (44 100 Hz, 16 bits) à l'aide d'un enregistreur numérique TASCAM HD-P2 et d'un microphone serre-tête SHURE SM10A et le signal d'entrée a été amplifié par l'intermédiaire d'une console de mixage

[1] Il s'agit d'un vaste corpus de parole lue produit de 2010 à 2012.

BEHRINGER XENYX 802.

Pour des raisons techniques (productions indistinctes, bruits extérieurs, etc.), 7 occurrences ont été rejetées. Ce sont donc 473 voyelles qui ont tout d'abord été segmentées manuellement à l'aide du logiciel *Praat*. Nous avons ensuite relevé la durée puis, conformément au paradigme dans lequel s'inscrit notre analyse, nous avons estimé la fréquence centrale des trois premiers formants à 25 %, 50 % et 75 % de la durée. À la suite d'un examen attentif des représentations spectrographiques, les paramètres de l'algorithme de suivi formantique ont été ajustés manuellement pour chacune des occurrences.

Par la suite, une série d'analyses discriminantes quadratiques par *jackknife* ont été effectuées afin de mettre au jour dans quelles proportions une observation (ici une occurrence vocalique) peut être associée à une catégorie établie *a priori* (l'origine géographique des locuteurs) à partir de prédicteurs (différentes combinaisons d'indices acoustiques).

Les données récoltées n'ont pas été normalisées. Ce choix méthodologique repose, d'une part, sur le fait que l'utilisation de procédures de normalisation extrinsèques (par exemple les scores-Z) est difficilement envisageable lorsqu'une seule catégorie phonémique est considérée. D'autre part, l'application d'une procédure de transformation (mels, logHertz, etc.) ne semble pas avoir d'effet positif sur le taux de classification des occurrences lors d'analyses discriminantes, que ce soit en FQ (Arnaud et coll., 2011) ou en anglais américain (Hillenbrand et coll., 1995).

3 Résultats

La FIGURE 1 illustre la répartition des occurrences selon l'estimation de la fréquence centrale des trois premiers formants à 25 %, 50 % et 75 % de la durée vocalique, tandis que la TABLE 1 présente les moyennes et écarts-types de la durée et des fréquences formantiques calculés à partir des moyennes individuelles. Les données n'ayant pas été normalisées, les productions féminines et masculines sont présentées séparément.

				F_1 (Hz)			F_2 (Hz)			F_3 (Hz)		
Ville	Sexe	*n*	Durée (ms)	25 %	50 %	75 %	25 %	50 %	75 %	25 %	50 %	75 %
QC	H	119	96 (14)	468 (27)	490 (28)	493 (34)	1837 (94)	1881 (109)	1906 (99)	2541 (113)	2562 (115)	2580 (112)
	F	120	108 (25)	531 (23)	555 (29)	554 (38)	2279 (73)	2335 (73)	2371 (93)	2977 (92)	2998 (87)	3007 (110)
SG	H	118	106 (23)	526 (50)	565 (54)	597 (51)	1687 (117)	1722 (104)	1782 (123)	2568 (169)	2580 (164)	2609 (186)
	F	116	115 (33)	631 (57)	663 (64)	656 (71)	2048 (72)	2098 (69)	2162 (52)	2914 (130)	2913 (115)	2932 (115)

TABLE 1 – Moyenne et écart-type (entre parenthèses) de la durée et de la fréquence centrale des trois premiers formants à 25, 50 et 75 % de la durée des voyelles produites par les hommes (H) et les femmes (F) de Saguenay (SG) et de Québec (QC)

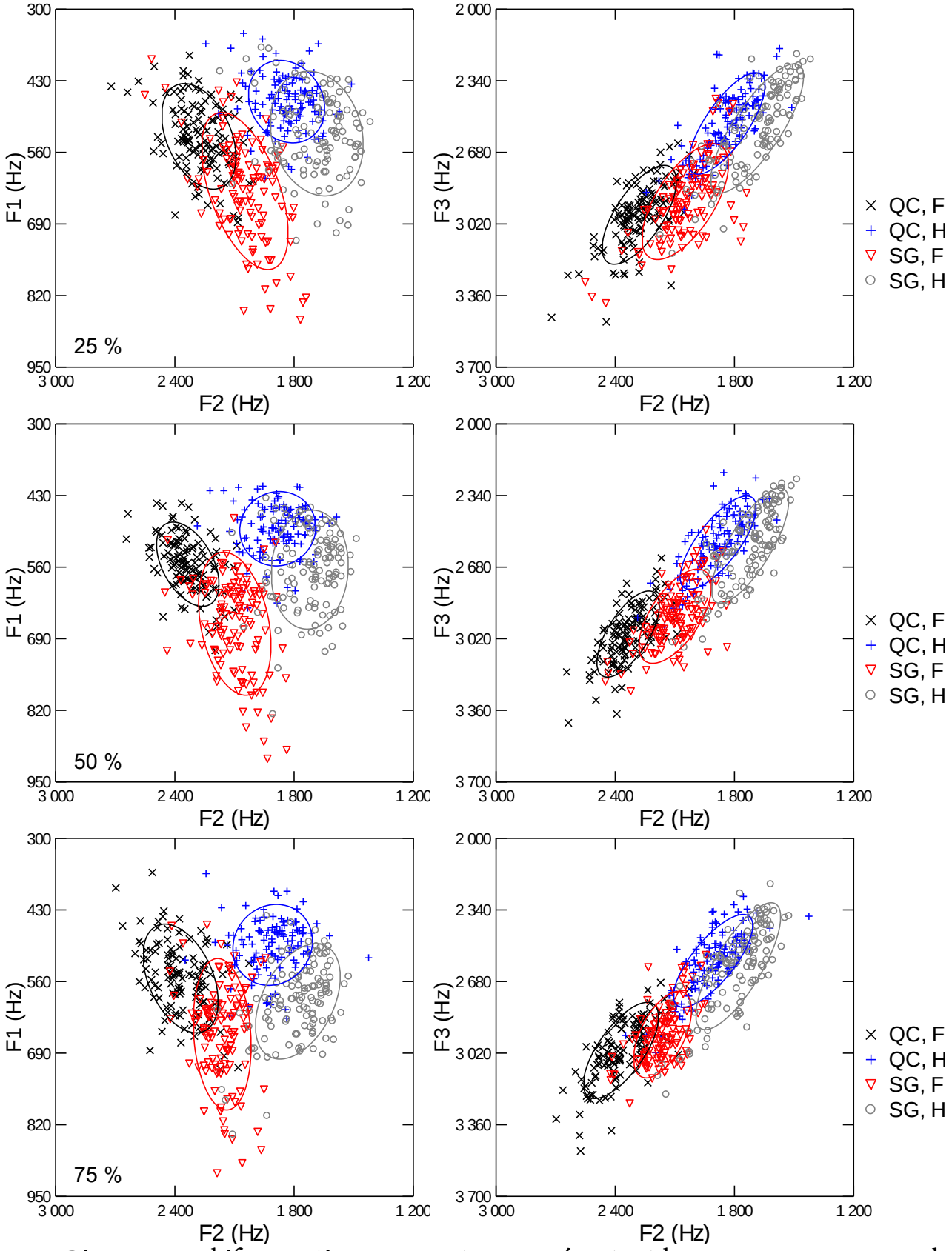


FIGURE 1 – Diagrammes biformantiques F_1×F_2 et F_2×F_3 présentant les 473 occurrences produites par les hommes (H) et les femmes (F) de Québec (QC) et de Saguenay (SG) à 25, 50 et 75 % de la durée vocalique. Les ellipses englobent 86,5 % des occurrences de chaque catégorie

D'après les données présentées dans la TABLE 1 et la FIGURE 1, les valeurs de F_1 et de F_2 semblent fortement influencées par l'origine géographique des locuteurs. En effet, les locuteurs de Saguenay tendent à produire des occurrences dont le premier formant est plus élevé et le

deuxième formant moins élevé que les locuteurs de Québec. Si, en première approximation, une augmentation de F_1 peut être interprétée comme une tendance à l'ouverture et une diminution de F_2, comme une tendance à la centralisation, dans notre corpus, les productions des locuteurs de Saguenay tendraient donc à être plus ouvertes et plus centralisées que celles des locuteurs de Québec. Ces tendances sont parfois si marquées qu'elles perturbent le rapport habituel entre fréquences formantiques féminines et masculines (voir Simpson, 2009 pour un survol). Contrairement à ce qui est attendu, la TABLE 1 et la FIGURE 1 révèlent que le F_1 des hommes de Saguenay est aussi (voire plus) élevé que le F_1 des femmes de Québec et que le F_2 des femmes de Saguenay est parfois aussi bas que le F_2 des hommes de Québec.

Quelle que soit l'origine des locuteurs, il appert également que les occurrences comprises dans les quatre catégories se chevauchent davantage à 25 % de la durée vocalique, puis se dissocient progressivement. Il n'est pas exclu que la plus grande variabilité des fréquences formantiques initiales soit un effet de la coarticulation (Hillenbrand et coll., 2001), le contexte consonantique précédant les voyelles analysées étant, comme nous l'avons mentionné, pour le moins varié.

La FIGURE 2 présente une synthèse de la trajectoire moyenne des occurrences sur les axes F_1, F_2 et F_3 de 25 % à 50 % et de 50 % à 75 % de la durée vocalique. Nous constatons dans l'espace F_1×F_2 que les occurrences tendent à se mouvoir vers la périphérie du trapèze vocalique, avec une augmentation de F_1 et de F_2 en cours d'émission. Bien que visuellement plus ténue, une augmentation de F_3 peut également être observée. La FIGURE 2 met en exergue l'importante augmentation du premier formant des occurrences produites par les hommes de Saguenay.

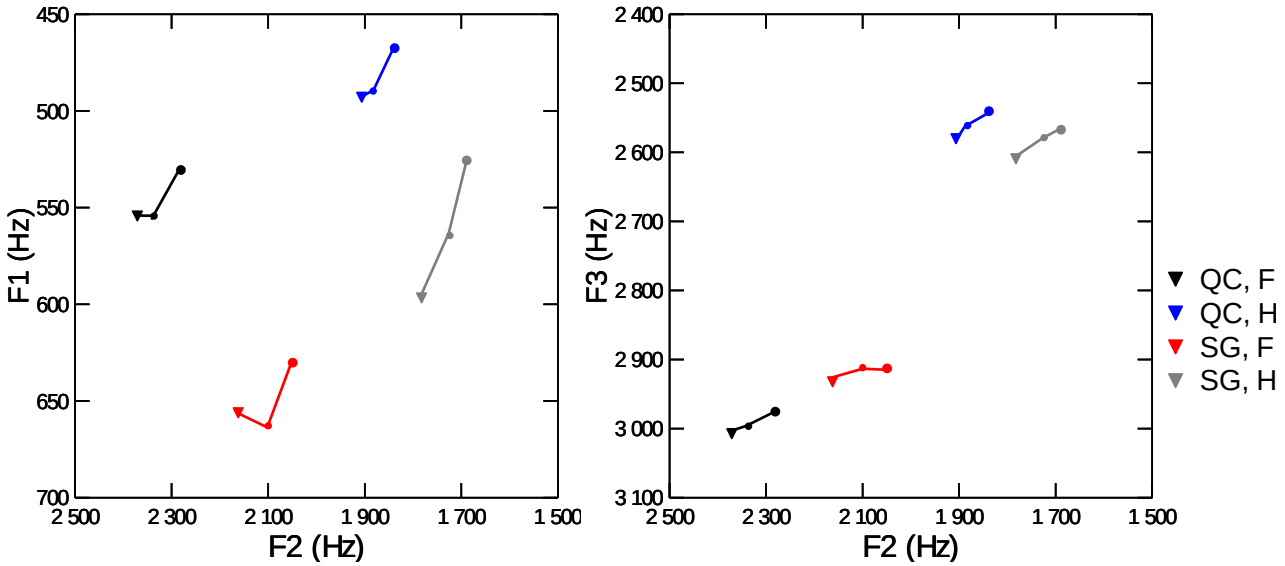


FIGURE 2 - Diagrammes biformantiques F_1×F_2 et F_2×F_3 présentant l'évolution temporelle moyenne des occurrences produites par les hommes (H) et les femmes (F) de Québec et de Saguenay (SG). Le triangle représente la valeur moyenne des mesures prises à 75 % de la durée

Au-delà de ces remarques qualitatives concernant l'apparente distinction entre les occurrences produites par les locuteurs de Saguenay et celles produites par les locuteurs de Québec, nous avons cherché à évaluer statistiquement dans quelle mesure les indices acoustiques considérés permettaient de séparer les occurrences en fonction de l'origine géographique des locuteurs les ayant produites. Pour ce faire, nous avons eu recours à une série d'analyses discriminantes, dont les résultats sont présentés dans la TABLE 2.

	50 %		25-75 %		25-50-75 %	
	~~Durée~~	Durée	~~Durée~~	Durée	~~Durée~~	Durée
F_1-F_2	84	85	85	86	86	87
F_1-F_2-F_3	86	88	88	**90**	87	89

TABLE 2 – Taux de classification correcte (en %) des occurrences dans les catégories *Saguenay* et *Québec* lors d'analyses discriminantes quadratiques où différents indices acoustiques relevés à différents points de mesure étaient utilisés comme prédicteurs

Le paradigme d'analyse acoustique traditionnel (F_1 et F_2 à un « état stable ») se voit enrichi par l'intégration, lors du calcul des fonctions discriminantes, d'indices supplémentaires tels F_3 et la durée, ainsi que par la prise en compte de l'évolution temporelle des caractéristiques spectrales des voyelles. L'analyse discriminante menée avec comme seuls prédicteurs F_1 et F_2 à 50% de la durée obtient le taux de classification correcte le plus faible (84 %), alors que le meilleur taux (90 %) est obtenu avec la prise en compte de la durée et des trois premiers formants à 25 % et à 75 % de la durée vocalique. Considérer un point de mesure central en plus de ceux à 25 % et à 75 % de la durée vocalique n'entraîne pas d'amélioration du taux de classification correcte. Ce résultat tend à suggérer que c'est bien l'évolution temporelle de la valeur des indices acoustiques considérés qui favorise une amélioration du taux de classification et non la simple adjonction de données supplémentaires. Quoi qu'il en soit, sachant qu'un seul phonème était à l'étude et que l'unique variable de regroupement des occurrences était la ville d'origine des locuteurs, un taux de classification correcte de 90 % s'avère remarquable.

4 Discussion

Le portrait proposé dans cette contribution s'avère plus complexe que celui auquel nous aurions abouti en ne considérant que la fréquence centrale des deux premiers formants estimée dans la section centrale de chaque occurrence. Le traditionnel diagramme F_1×F_2 demeure certes une représentation visuelle fondamentale et pédagogique en phonétique, mais les résultats proposés suggèrent que les faits de variation affectant le système vocalique pourraient être revisités en ne s'appuyant plus seulement sur une paramétrisation statique bidimensionnelle du signal sonore, mais sur une paramétrisation dynamique multidimensionnelle incluant F_3, la durée et la dynamique temporelle de chaque formant.

Même si notre étude est consacrée à la variabilité des réalisations d'un seul phonème dans une seule position prosodique et syllabique, nos observations sont congruentes avec celles proposées dans la littérature (Nearey et Assmann, 1986; Hillenbrand et coll., 1995, 2001; Arnaud et coll., 2011) et que Hillenbrand (2013 : 17, dans Morrison et Assmann, 2013) synthétise en ces termes : « static vowel targets are neither necessary nor sufficient for the recognition of vowel identity. »

L'analyse proposée réduit la dynamique temporelle des différents indices considérés à des trajectoires linéaires entre 25 % et 75 % de la durée. Cette modélisation correspond à l'hypothèse de la « double cible » (Nearey et Assmann, 1986) et semble plébiscitée dans la littérature, que ce soit en anglais (Morrison, 2013, dans Morrison et Assmann, 2013) ou en FQ (Arnaud et coll., 2011). Il aurait néanmoins pu être intéressant d'évaluer la pertinence des deux autres modélisations de base des trajectoires (*onset + slope* et *onset + direction*).

Notons également que la localisation temporelle des points de mesure appelés « établissement » (*onset*) et « queue » (*offset*) n'a pas été guidée par des considérations théoriques. Dans la lignée de Nearey et Assmann (1986) ou de Hillenbrand et coll. (1995), nous avons estimé que l'extraction des valeurs formantiques à 25 % et à 75 % de la durée vocalique offrait un intervalle temporel suffisant pour caractériser une éventuelle dynamique sans que les résultats de notre analyse ne soient bruités par les irrégularités caractérisant les cycles d'oscillation initiaux et finaux des segments (à ce propos, voir Nearey, 2013, dans Morrison et Assmann, 2013).

Le choix de recourir à l'analyse discriminante a l'avantage de permettre de considérer conjointement, dans un même test statistique, plusieurs indices acoustiques, afin d'évaluer leur poids relatif dans la différenciation des occurrences en fonction de l'origine géographique des locuteurs. Néanmoins, ce type d'analyses ne présente qu'une visée exploratoire puisque l'effet aléatoire de l'individu n'est pas pris en compte. L'utilisation de modèles mixtes d'ANOVAs à mesures répétées sera donc essentielle afin de tester l'effet du sexe et de l'origine géographique des locuteurs, mais aussi de la position du point de mesure (25 %, 50 % ou 75 %), sur la valeur des différents indices acoustiques considérés.

5 Conclusion

Les résultats obtenus lors de cette analyse convergent vers un même bilan : en ce qui a trait à la voyelle /ɛ/ en finale absolue, les usages des 40 locuteurs de Saguenay et de Québec constituant notre corpus sont bien distincts. Au plan acoustique, les principales différences résident dans la fréquence centrale des premier et deuxième formants, les occurrences produites par les Saguenéens présentant généralement un F_1 plus élevé et un F_2 moins élevé. Nous interprétons ces résultats comme une tendance à l'ouverture du /ɛ/ en finale absolue chez les locuteurs de Saguenay.

En dépit de ce que suggèrent les descriptions essentiellement impressionnistes fournies dans la littérature, le phénomène d'ouverture du /ɛ/ en finale absolue semble toujours vivant au Québec, y compris dans les productions très formelles de jeunes locuteurs universitaires. Il est également apparu que dans notre corpus, ce phénomène phonétique présentait un conditionnement géographique. Dès lors, la question de ses éventuels effets sur l'organisation du système vocalique des locuteurs étudiés mériterait d'être examinée plus avant. L'ouverture de la voyelle /ɛ/ en finale affecte-t-elle le timbre des voyelles adjacentes /e/ et /a/, sachant néanmoins que la présence d'un /a/ antérieur en finale absolue demeure rare en FQ (Dumas, 1987 : 137) ? Par ailleurs, nous nous sommes limités, dans le cadre de cette contribution, à une analyse acoustique du phénomène. Or, des auditeurs humains, naïfs ou expérimentés, seraient-ils en mesure de percevoir l'effet des patrons et des trajectoires différenciés mis en évidence par l'intermédiaire d'un processus de classification automatique ? La question demeure ouverte et actuellement au centre de nos investigations.

Remerciements

Ce projet a reçu l'appui financier du Fonds québécois de recherche sur la société et la culture (FQRSC) et du Conseil de recherches en sciences humaines du Canada (CRSH). Merci aux membres du laboratoire de phonétique expérimentale de l'UQAC, Jean Dolbec et Caroline Sigouin, pour leur relecture attentive.

Références

ARNAUD, V. (2010). Le rôle des changements spectraux dans la description des voyelles françaises : entre sociophonétique et acoustique. *In* Remysen, W. et D. Vincent (éd.). *Hétérogénéité et homogénéité dans les pratiques langagières*, PUL, 207-226.

ARNAUD, V., C. SIGOUIN et J.-P. ROY. (2011). Acoustic description of Quebec French high vowels: First results. *In Proceedings of the XVIIth ICPhS*, Hong Kong, 244-247.

DELATTRE, P., A. M. LIBERMAN, F. S. COOPER, et L. J. GERTSMAN. (1952). An experimental study of the acoustic determinants of vowel colour. *Word*, 8(3):195-210.

DUMAS, D. (1987). *Nos façons de parler. Les prononciations en français québécois*. PUQ.

FOX, R. A. et E. JACEWICZ. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *Journal of the Acoustical Society of America*, 126(5):2603-2618.

GENDRON, J.-D. (1966). *Tendances phonétiques du français parlé au Canada*. PUL.

HILLENBRAND, J. M., M. J. CLARK et T. M. NEAREY. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, 109(2):748-763.

HILLENBRAND, J. M., L. A. GETTY, M. J. CLARK et K. WHEELER. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5):3099-3111.

HILLENBRAND, J. M. et T. M. NEAREY. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America*, 105(6):3509-3523.

MORRISON, G. S. et P. F. ASSMANN. (2013). *Vowel Inherent Spectral Change*, Springer.

NEAREY, T. M. et P. F. ASSMANN. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80(5):1297-1308.

PARADIS, C. (1985). *An Acoustic Study of Variation and Change in the Vowel System of Chicoutimi and Jonquière (Québec)*, thèse de doctorat, University of Pennsylvania.

PARADIS, C. et J. DOLBEC. (1998). PHONO : caractéristiques phonétiques du français parlé au Québec, site Internet, http://phono.uqac.ca/

PETERSON, G. E. et H. L. BARNEY. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175-184.

SCHWARTZ, J.-L., D. BEAUTEMPS, C. ABRY et P. ESCUDIER. (1993). Inter-individual and cross-linguistic strategies for the production of [i] vs. [y] contrast. *Journal of Phonetics*, 21(4):411-425.

SIGOUIN, C. (2013). *Caractéristiques acoustiques des voyelles fermées tendues, relâchées et allongées en français québécois*, mémoire de maîtrise, Université du Québec à Chicoutimi.

SIMPSON, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2):621–640.

SQUAIR, J. (1888). A contribution to the study of Franco-Canadian dialect. *Proceedings of the Canadian Institute*, 6:161-168.

L'articulation des coronales occlusives en népalais

Rajesh KHATIWADA
Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS-Université Paris3/Sorbonne Nouvelle
19 rue des Bernardins, 75005 Paris
rajesh.khatiwada@gmail.com

RESUME

Le népalais, langue de la famille indo-aryenne parlée par plus de 35 millions de locuteurs, est la langue officielle du Népal. Elle fait également partie des langues officielles parlées en Inde, on la retrouve aussi au Bhoutan. La première grammaire de la langue remonte au début du 19è siècle et très peu d'études phonétiques et phonologiques expérimentales ont été réalisées depuis. La présente étude décrit l'articulation des coronales occlusives népalaises qui sont divisées en trois catégories phonologiques – les dentales simples, les affriquées et les rétroflexes. Chacune de ces catégories est à leur tour classifiée, en fonction de leur voisement et de leur aspiration, en 12 segments : /t, tʰ, d, dʱ/ (dentales), /ts, tsʰ, dz, dzʱ/ (affriquées) et /ʈ, ʈʰ, ɖ, ɖʱ/ (rétroflexes). Les méthodes palatographiques et linguographiques directes sont utilisées pour étudier les lieux d'articulation et les contacts linguaux chez 9 locuteurs.

ABSTRACT

Nepali, a language of the Indo-Aryan language family spoken by over 35 million speakers, is the official language of Nepal, and it is also spoken in India and in Bhutan. Despite the first grammar of the language which dates back to the early 19th century, very few phonetical and phonological experimental studies have been done on Nepali. The present study depicts the Nepalese coronal stops which are divided into three phonological categories - dentals, affricates and retroflex. Each of these categories is organized according to a quadruple opposition of voicing and aspiration providing the following 12 segments: /t, tʰ, d, dʱ/ *(dentals),* /ts, tsʰ, dz, dzʱ/ *(affricates) and* /ʈ, ʈʰ, ɖ, ɖʱ/ *(retroflex). Direct palatography and linguography are used to study the places of articulation and lingual contacts in the productions of 9 speakers.*

MOTS-CLES : coronales, rétroflexes, affriquées, népalais, consonnes, occlusives.
KEYWORDS: coronals, retroflex, affricates, Nepali, consonants, plosives.

1. Introduction

Le népalais fait partie des langues indo-aryennes. Il est parlé par plus de 35 millions de locuteurs dans le monde : au Népal où il est la langue officielle (népalais ou népali), mais aussi en Inde et au Bhoutan (Khatiwada 2009).

Si la première grammaire du népalais est conçue en 1820 par le révérend anglais Ayton, on a dû attendre les années 1980 pour voir les premières études expérimentales sur le népalais, avec notamment le travail de Poon et Mateer (1985) sur le VOT (Voice Onset Time) des occlusives népalaises et l'étude palatographique réalisée dans le cadre de la thèse de Pokharel (1989). Mais malgré ces deux importants travaux pionniers, les études expérimentales en phonétique sont restées à leurs balbutiements.

Le népalais, comme d'autres langues Indo-Aryenne, est intéressant de part la richesse de son inventaire consonantique et en particulier dans la zone coronale. Comme le hindi, le népalais possède 12 coronales occlusives : les dentales /t, tʰ, d, dʱ/, les affriquées /ts, tsʰ, dz, dzʱ/ et les rétroflexes /ʈ, ʈʰ, ɖ, ɖʱ/. Les grammaires traditionnelles du népalais, suivant la description du sanskrit, ont longtemps considéré les affriquées comme des palatales comme en sanskrit ou encore en hindi (Sigdyal 1951, Matthews 1984). Longtemps restées au stade d'analyses impressionnistes, grâce aux palatographies de Pokharel (1989), les études ont pu démontrer que les affriquées népalaises ne sont pas produites dans la région palatale mais dans la région alvéolaire. D'autre part, Pokharel met en doute la nature de la rétroflexion en népalais, en montrant que la série de rétroflexes est articulée dans la région alvéolaire à l'aide de la pointe de la langue sans aucun recourbement en arrière, et donc sont plutôt des apico-alvéolaires. Pour autant, les données de Pokharel sont basées sur les productions d'un seul locuteur. L'objectif du présent travail est donc de vérifier les résultats de Pokharel sur un nombre plus important de locuteurs afin de mieux documenter l'articulation de la série coronale en népalais et d'évaluer la variabilité articulatoire inter- et intra -locuteur de ces consonnes. La variété de népalais étudiée ici est une variété standard utilisée et considérée telle quelle dans les média, dans la capitale et les grandes villes népalaises.

2. La méthode

2.1 Matériel linguistique

Nous avons étudié, en premier lieu, les différences articulatoires entre les trois catégories de coronales occlusives (dentales, affriquées et rétroflexes) se distinguant en termes de voisement et d'aspiration (T, Tʰ, D, Dʱ)[1] et en deuxième lieu, nous avons observé l'effet de la coarticulation en fonction des trois contextes vocaliques (a, i, u) sur ces catégories. Les 2 corpora suivants ont été constitués pour ces deux objectifs en utilisant autant que possible des mots du népalais (seuls 3 logatomes ont été utilisés). Le contexte adjacent à la consonne test (*VC*^{2}*V)* a été contrôlé au maximum, et il faut noter que la consonne *C*^{1} n'a malheureusement pas pu être identique dans les quadruplets.

Le 1er corpus donné dans le tableau 1 vise à tester si les lieux d'articulation et les contacts linguaux des consonnes étudiées varient en fonction du voisement et de l'aspiration. Les consonnes tests sont toutes en contexte /aCa/.

	Dentales			Rétroflexes			Affriquées		
T	/**t**/	/mata/	'mère'	/ʈ/	/baʈa/	'bassine'	/**ts**/	/ba**ts**a/	'promesse'
Tʰ	/**tʱ**/	/matʰa/	'front'	/ɖ/	/gaɖa/	'véhicule'	/**tsʰ**/	/ba**tsʰ**a/	'veau'
D	/**d**/	/ada/	'moitié'	/ʈʰ/	/paʈʰa/	'agneau'	/**dz**/	/ba**dz**a/	'instrument de musique'
Dʱ	/**dʱ**/	/badʱa/	'empêchement'	/ɖʱ/	/gaɖʱa/	'foncé'	/**dzʱ**/	/ba**dzʱ**a/	'logatome'

TABLEAU 1 - Corpus 1 de la première expérience. T : sourde non-aspirée, Tʰ : sourde aspirée, D : sonore non-aspirée, Dʱ : sonore aspirée

Le deuxième corpus teste l'effet de la coarticulation sur les lieux d'articulation et sur la nature des contacts linguaux des catégories des coronales occlusives étudiées en fonction de trois contextes

[1] Les lettres capitales indiquent n'importe quelle occlusive coronale parmi ces trois catégories étudiées.

vocaliques /aCa/, /iCi, /uCu/ différents. Nous avons choisi le segment sourd et non-aspiré de chaque catégorie pour cette deuxième étude. Le tableau 2 suivant représente le corpus pour cette deuxième étude.

	Dental		Rétroflexe		Affriquée	
/aCa	/mata/	'mère'	/baʈa/	'bassine'	/batsa/	'promesse'
/iCi/	/miti/	'date'	/miʈi/	'logatome'	/mitsi/	'logatome'
/uCu	/kutu/	'logatome'	/muʈu/	'cœur'	/mutsu/	'logatome'

TABLEAU 2 - Corpus 2 de la deuxième expérience - aCa, iCi et uCu.

2.2 Participants

Cinq (3 hommes et 2 femmes) et quatre locuteurs (3 hommes et une femme) ont participé respectivement à la première et deuxième étude. Tous sont de langue maternelle népalaise et ne présentent aucun signe de troubles articulatoires. Hormis deux locuteurs (F1 et M1) vivant en France depuis plus de 20 ans (mais mariés à un népalais et parlant le népalais au quotidien) aucun de ces participants avaient passé plus d'un an en France ou de façon continue lors de la prise de données. Nous avons examiné deux productions pour chaque consonne étudiée.

2.3 Données articulatoires et analyses

L'analyse articulatoire des consonnes népalaises a été effectuée à l'aide de données de palatographie et de linguographie statiques (Marchal 1988, Ladefoged 2003). Bien que ces méthodes soient anciennes, la prise de donnée est simple, peu onéreuse (contrairement à l'électropalatographie) et peut être réalisé n'importe où.

La méthode palatographie directe comprend trois étapes successives : d'abord, nous badigeonnons la surface de la langue avec un mélange du charbon, du chocolat et de l'huile, ensuite, une fois que le locuteur a prononcé le mot contenant la consonne à l' étude, nous introduisons un petit miroir dans la bouche du locuteur. Enfin, nous prenons la photographie de l'image projetée sur le petit miroir. Ces images nous servent de base pour notre travail d'observation et de comparaison.

La linguographie est la méthode opposée à la méthode palatographie directe. Au lieu de badigeonner la langue et de prendre en photo la trace laissée sur le palais, ici, c'est le palais que nous badigeonnons. Une fois le palais enduit de la préparation, nous demandons aux locuteurs de prononcer le segment à l'étude, puis nous lui demandons de sortir leur langue et nous la photographions. Le linguogramme nous aide à observer la partie utilisée lors de la production de la consonne.

Définir les lieux d'articulation n'est pas une chose aisée. Nous trouvons différentes façons de délimiter les zones articulatoires. Notre travail s'appuie sur le plan de dentition proposé par Firth (1948). Mais contrairement à ce dernier, nous avons introduit une ligne mi-canine entre la ligne latérale incisive et la ligne canine. Ce plan de dentition qui sert aux analyses est présenté sur la gauche de la figure 1. Nous définissons donc une région alvéolaire plus réduite. Nous nous retrouvons ainsi avec une région alvéolaire plus proche de celle proposée par Ladefoged & Maddieson (1996), où la région « alvéolaire » dénote le contact sur ou dans la partie antérieure à l'alvéole. Le tableau 3 suivant donne les critères que nous avons utilisés pour définir les régions articulatoires et donc les lieux d' articulations.

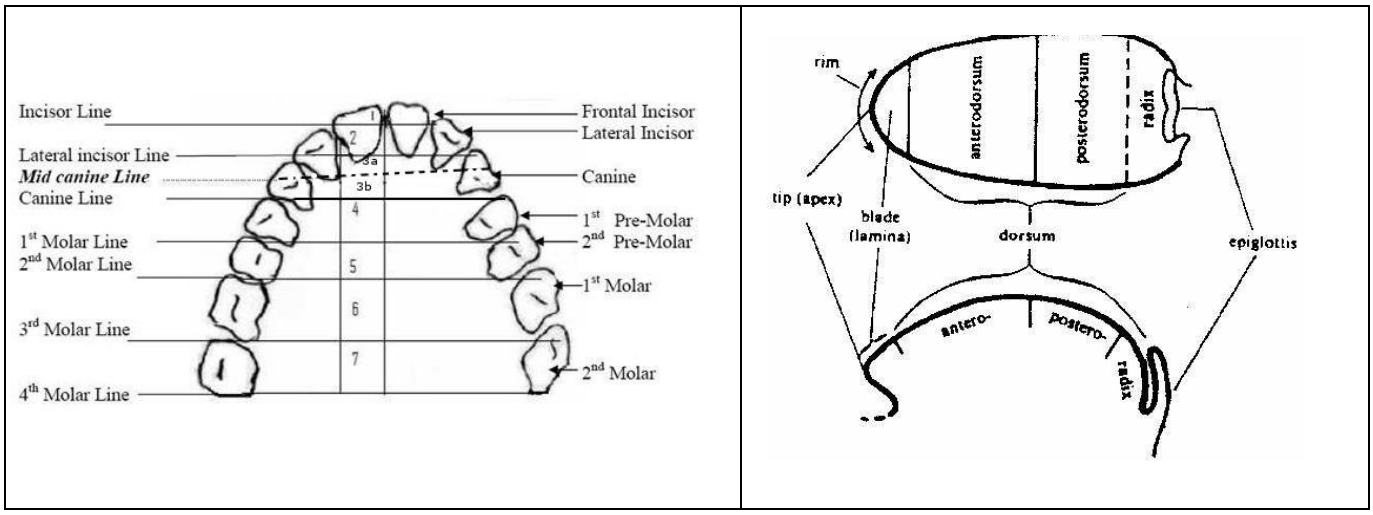


FIGURE 1 - Subdivision articulatoire basée sur le plan de dentition de Firth (1948 (à gauche) et la subdivision linguale d'après Catford (1977) à droit.

En ce qui concerne l'analyse des contacts linguaux nous nous inspirons de la division linguale proposée par Catford (1977) (figure 1 droite). Comme dans le cas de la délimitation des lieux d'articulation, il n'est pas non plus facile de définir la frontière entre la pointe de la langue (apex) et la lame de la langue, un contact large qui commence dans la région apicale jusqu'à la partie laminale sera considéré comme apico-laminal.

1	Ligne incisive	Dentale
3a	Ligne latérale incisive + mi-canine ligne	Alvéolaire
1,2 – 3a/ 3b	Incisive + latérale incisive + mi-canine	Dento-alvéolaire
4	Mid canine + première molaire	Post-alvéolaire
5	1re molaire + 2e molaire	Pré-palatale
6	2e molaire + 3e molaire	Palatale
7	Ligne 4e molaire	Vélaire

TABLEAU 3 - Critères de définition des lieux d'articulation.

Un contact lingual ne se comprend pas si nous n'éclaircissons pas la manière dont la langue est utilisée. Le contact de la langue avec le palais n'est pas une simple obstruction sans aucune considération de la configuration linguale comme laisse entendre souvent les descriptions phonétiques (Léon 1992, Ladefoged & Maddieson 1996).

En suivant Hála (1965), il faut noter que la langue est indivisible, et donc quand on parle de production de type « apicale » ou « laminale », on ne peut pas imaginer juste la partie concernée se soulever et créer une constriction sur le palais. L'auteur souligne qu'« *il vaudrait donc mieux entendre, par les termes en question, la prévalence de telle ou telle partie de la langue dans le cadre de l'articulatoire général. De cette façon, les consonnes apicales – […] seraient celles formées principalement par la pointe de la langue […]* » (Hála 1965). En se basant sur les ciné-radiographies des consonnes dites apicales, Hála avait noté que la pointe (de la langue) peut agir « *contre le palais de trois façon différentes : soit horizontalement, par sa surface de dessus ou par son dos, soit verticalement, par son extrémité de devant ou par son tranchant, soit enfin par sa*

surface de dessous (au dessous du tranchant) », comme illustré sur la partie droite de la figure 1.

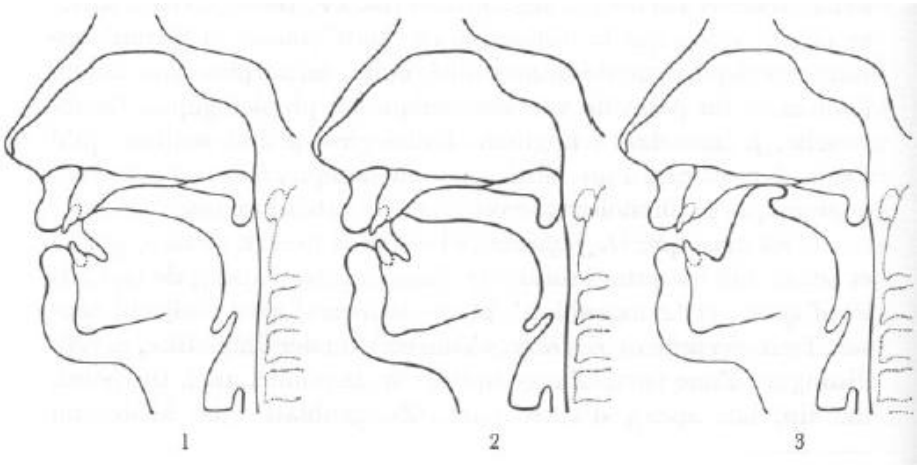


FIGURE 2 - Articulation de trois types « apical » d'après Hála, (1) le type « horizontal », le t tchèque ou français, (2) le type « vertical » comme t anglais et (3) le type « vertical » rétroflexe T en tamoul (Hála 1965).

Il définit donc [t] que nous trouvons par exemple, en français, tchèque ou allemand, illustré dans le schéma (1), comme « apicale de type horizontal », car cette consonne est formée horizontalement par les bords de la surface du dessus de la pointe. Lors de consonne [t] anglaise, schéma (2), l'articulation est « apicale de type verticale » formée par l'extrémité de la pointe appliquée verticalement contre le palais que l'auteur définit comme *cacuminale*. Enfin, la dernière consonne [ʈ] du schéma (3), articulée avec la pointe de la langue appliquée verticalement et recourbée en arrière est donc *rétroflexe*. En népalais, ces deux dernières, apicales de type vertical, à savoir cacuminal et rétroflexe semblent construire une seule et unique catégorie phonologique.

3. Les résultats et discussion

3.1 Expérience 1 : Dentales, affriquées et rétroflexes en fonction du voisement et de l'aspiration

Nous avons observé que lors de la production des dentales simples et des affriquées nos cinq locuteurs utilisent presque les mêmes lieux d'articulation et les mêmes parties de la langue.

(i) Dentales et affriquées:

Hormis un seul locuteur M1, qui articule la catégorie dite « dentale » systématiquement dans la zone alvéolaire et les affriquées dans la zone post-alvéolaire, tous les quatre autres produisent ces deux catégories des consonnes dans la région dento-alvéolaire. Excepté le M3 qui montre une variation de l'utilisation linguale, tous les quatre autres utilisent la lame de la langue pour produire ces deux catégories des consonnes.

FIGURE 3 – Les palatogrammes et linguogrammes de mata (F1) (à gauche) et de batsa (F2) (à droit)

Le tableau 4 suivant récapitule le résultat de notre première étude.

	/t/		/ts/		/tʰ/		/tsh/		/d/		/dz/		/dʰ/		/dzh/	
LOC	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**
F1	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L
F2	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L
M1	A	L	PA	L	A	L	PA	L	A	L	PA	L	A	L	PA	L
M2	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L	DA	L
M3	DA	AL	DA	L	DA	Ap	DA	L	D	Ap	DA	L	D	Ap	DA	L

TABLEAU 4 - Tableau récapitulatif des dentales et des affriquées népalaises. LA : lieu d'articulation, CL : contact lingual, D : Dental, DA : Denti-alvéolaire, A : Alvéolaire, L : Laminal, Ap : Apical, AL : Apico-laminal, PA : Post-alvéolaire.

(ii) Rétroflexes :

Les rétroflexes montrent une variation considérable non seulement dans les lieux d'articulation mais également en ce qui concerne les contacts linguaux. Nous constatons que même si la zone post-alvéolaire est celle qui est la plus utilisée, l'articulation peut s'étendre de la zone alvéolaire (comme chez Pokharel) jusqu'à la zone pré-palatale.

FIGURE 4 – Les palatogrammes et linguogrammes de [baʈa] (F1) type cacuminal (à gauche) et de [baʈa] (M 4) type rétroflexe (à droit)

Le tableau 5 suivant récapitule la stratégie articulatoire concernant l'utilisation de la langue lors de la production des rétroflexes. Dans cette première étude, les locuteurs semblent utiliser plutôt le type cacuminal (75% de nos participent articulent comme cacuminal) contre seulement 25% de type rétroflexe.

	/ʈ/		/ʈʰ/		/ɖ/		/ɖʰ/	
LOC	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**	**LA**	**CL**
F1	PA	C	PA	C	PA	C	PA	C
F2	PA	C	PA	C	PA	C	PA	C
M1	PA	C	PA	C	PA	C	PA	C
M2	PA	C	PA	C	PA	C	A	C
M3	PP	SA	PP	SA	PP	SA	PP	SA

TABLEAU 5 - Le tableau récapitulatif concernant la production des rétroflexes. LA : Lieu d'articulation, A : Alvéolaire, CL : contact lingual, DA : Dento-alvéolaire, L : Laminal, PP : Pré-palatal, C : Cacuminal, SA : Subapical, PA :Post-alvéolaire.

3.2 Deuxième étude : Influence du contexte:

Contrairement à la première expérience, nous avons trouvé un nombre moins important de productions de type alvéolaire comme réclamait Pokharel, et cette articulation est plutôt contextuelle (avec i). Hormis le locuteur M6, dans les contextes a et u, tous les autres ont produit les rétroflexes dans la région post-palatale et pré-palatale. Pour les consonnes rétroflexes les réalisations se partagent entre des productions de type « rétroflexe » et de type « cacuminal » articulées principalement dans la zone post-alvéolaire majoritairement, et mais également chez certaines locuteurs dans la zone encore postérieure (pré-palatal) (voir tableau 6).

Loc	**Contextes**	**[t]**		**/ʦ/**		**/ʧ/**	
		LA	**CL**	**LA**	**CL**	**LA**	**CL**
F3	a	DA	L	DA	L	PP	SA
	i	DA	L	DA	L	A	C
	u	DA	L	DA	L	PA	SA
M4	a	DA	L	DA	L	PA	SA
	i	DA	L	DA	L	A	C
	u	DA	L	DA	L	PP	SA
M5	a	DA	AL	DA	L !	PA	SA
	i	DA	AL	DA	L !	A	C
	u	DA	AL	DA	L	PA	SA
M6	a	DA	A	DA	L	A	C
	i	DA	A	DA	L	A	C
	u	DA	A	DA	L	PA	C

TABLEAU 6 - Le tableau récapitulatif concernant les productions des coronales occlusives sourdes non-aspirées de la deuxième expérience. LA : Lieu d'articulation, CL : contact lingual, DA : Dento-alvéolaire, A : Alvéolaire, L : Laminal, PP : Pré-palatal, C : Cacuminal, SA : Subapical, PA :Post-alvéolaire, AL : Apico-lamnial.

Les trois catégories de coronales occlusives présentent une légère variation au niveau de la cible articulatoire sur le palais (lieu d'articulation) mais aussi au niveau de l'utilisation de la langue. Nos données montrent que les contacts sont plus appuyés et nets pour les consonnes sourdes non-aspirées et également pour les sourdes aspirées (la première étant plus nette). Cependant, ces trois catégories laryngales utilisent la même cible articulatoire.

Les coronales simples, appelées fréquemment dans la littérature « dentales », sont produites avec la lame de la langue, majoritairement dans la région « dento-alvéolaire ». Cependant, nous avons également observé quelques productions peu fréquentes dans la région « dentale » et aussi dans la région alvéolaire. Hormis les laminales, nous avons également trouvé quelques articulations de type « apical » (avec un contact moins étendu), et également avec un contact plus étendu de type « apico-laminal » (très rare dans nos données) (voir le tableau 4).

La catégorie des affriquées montre une stabilité remarquable parmi ces trois classes de coronales

étudiées. Tous les participants ont utilisé la lame de la langue pour produire les affriquées. La région dento-alvéolaire est la préférée de la majorité de nos participants. Les affriquées népalaises ne sont pas des palatales comme nous pouvons encore le trouver dans un certain nombre d'écrits (Matthews 1984) mais plutôt de type lamino-dento-alvéolaire.

Les rétroflexes népalaises avec leurs variations, tant dans l'utilisation de la langue que dans le choix de la cible articulatoire, demeurent la catégorie la plus sujette à controverse. Nous avons observé (tableaux 4 et 5) que la région la plus utilisée dans nos expériences est la région post-alvéolaire. Nous avons également trouvé l'implication de la partie alvéolaire qui est très rare dans la première expérience faite uniquement dans un contexte /a/ et systématiquement présente dans la deuxième expérience dans le contexte /i/ et la partie pré-palatale (plus souvent qu'on attendait). Les résultats des deux expériences varient légèrement concernant l'utilisation de la langue dans l'articulation des rétroflexes. Nous avons observé que nos participants de la première expérience utilisent majoritairement la partie tranchante (*rim* en anglais) de la langue pour produire les consonnes de la catégorie dite « rétroflexe », tandis que dans la deuxième expérience, nous avons trouvé que la plupart des locuteurs recourbent systématiquement la pointe de la langue lors de l'articulation dans cette même catégorie des contextes /a/ et /u/.

Contrairement à Pokharel qui doute de l'existence de la rétroflexion proprement dite en népalais et pose l'apex sans le recourbement comme le seul articulateur actif, nos données illustrent une variation plus importante et une existence incontestable des rétroflexes surtout dans les contextes des voyelles /a/ et /u/.

Nos données confirment la variation existant au sein de la même langue lors de la production des rétroflexes (Ladefoged & Bhaskararao 1983, Ladefoged & Maddieson 1996). Cette variation entre le type rétroflexion et le type cacuminal se trouve chez un même locuteur, variant ou non en fonction des contextes vocaliques. Cela confirme que les deux types renvoient à la même classe phonologique. Aucune langue du monde semble n'avoir deux classes phonologiques différentes en fonction de l'utilisation du type rétroflexe et du type cacuminal ou apico-alvéolaire, au sens développé par Hála 1964 (Ladefoged & Maddieson 1996).

4. Conclusion

Nos présentes études articulatoires basées sur la méthode palatographique et linguographique ont essayé d'examiner les lieux d'articulation et les contactes linguaux lors de la production des coronales occlusives népalaises. Si la description des dentales et dans une certaine mesure les affriquées semblent s'accorder avec celles de Pokharel (1989), les rétroflexes montrent une différence fondamentale. Cette catégorie montre une variation très importante par rapport aux propositions de Pokharel. Ces variations (type cacuminal et type rétroflexe), se trouvent au niveau des interlocuteurs mais également en fonction de contextes vocaliques différents. Contrairement à Pokharel, à côté du type qu'il appelle « apico-alvéolaire », nous observons également le véritable type « rétroflexe » dans nos résultats articulés dans la région post-alvéolaire voir pré-palatale.

Remerciements

Je suis reconnaissant à Madame Annie Rialland pour son soutien et Isabelle Khatiwada et Cécile Fougeron pour la préparation de cette étude. Il va de soi que je suis le seul responsable pour les erreurs éventuelles.

Références

CATFORD, J.C. (1977). Fundamental problems in phonetics. Indiana University Press.

FIRTH J.R. (1948) Word-Palatograms and Articulation. In Papers in Linguistics 1931 – 1951. Oxford 148-55. (1957)

HÁLA, B. (1964). Apical, cacuminal, rétroflexe, coronal, dorsal. *Phonetica*. 11(3-4).

KHATIWADA, R. (2009). Nepali. *Journal of International Phonetic Association*. 39(3).

LADEFOGED, P. & MADDIESON, I. (1996). The Sounds of the World's Languages. Blackwell.

LADEFOGED, P. (2003). Phonetic data analysis: an introduction to fieldwork and instrumental techniques. Blackwell.

LEON, P. (1992). Phonétisme et Prononciations du français. Paris. Nathan.

MARCHAL, A. (1988). La palatographie. Edition du CNRS.

MATTHEWS, D. (1984). A Course in Nepali. Curzon.

POKHAREL, M.P.S. (1989). Experimental analysis of Nepali sound system. PhD dissertation, Deccan College, Pune.

POON, P.G. & MATEER, C.A. (1985). A Study of VOT in Nepali Stop Consonants. *Phonetica*. 42(1).

SIGDYAL, S. (1951). Madhyacandrika. Sajha Prakasha.

Boosting de bonzaïs pour la combinaison efficace de descripteurs : application à l'identification du rôle du locuteur

Antoine Laurent[1], Nathalie Camelin[2] Christian Raymond[3]
(1)LIMSI, Orsay, France
(2) LIUM, Le Mans, France
(3) IRISA-INSA de Rennes, France
antoine.laurent@limsi.fr,nathalie.camelin@lium.univ-lemans.fr
christian.raymond@irisa.fr

RÉSUMÉ

Dans ce travail, nous nous intéressons au problème de la détection du rôle du locuteur dans les émissions d'actualités radiotélévisées. Dans la littérature, les solutions proposées sont de combiner des indicateurs variés provenant de l'acoustique, de la transcription et/ou de son analyse par des méthodes d'apprentissage automatique. De nombreuses études font ressortir l'algorithme de boosting sur des règles de décision simples comme l'un des plus efficaces à combiner ces différents descripteurs. Nous proposons ici une modification de cet algorithme état-de-l'art en remplaçant ces règles de décision simples par des mini arbres de décision que nous appelons bonzaïs. Les expériences comparatives menées sur le corpus EPAC montrent que cette modification améliore largement les performances du système tout en réduisant le temps d'apprentissage de manière conséquente.

ABSTRACT

Boosting bonsai trees for efficient features combination : application to speaker role identification

In this article, we tackle the problem of speaker role detection from broadcast news shows. In the literature, many proposed solutions are based on the combination of various features coming from acoustic, lexical and semantic informations with a machine learning algorithm. Many previous studies mention the use of boosting over decision stumps to combine efficiently theses features. In this work, we propose a modification of this state-of-the-art machine learning algorithm changing the weak learner (decision stumps) by small decision trees, denoted bonzaï trees. Experiments show that using bonzaï trees as weak learners for the boosting algorithm largely improves both system error rate and learning time.

MOTS-CLÉS : identification rôle locuteur, boosting, arbre de décision.

KEYWORDS: Speaker role recognition, boosting, decision tree.

1 Introduction

Dans cet article nous nous intéressons à la détection du rôle des locuteurs dans les émissions radiotélévisées. Dans la littérature, ce problème est ramené à un problème de classification multi-classes où chaque locuteur d'une émission doit être associé avec un rôle. Suivant cette approche, les études précédentes ont attaqué le problème en utilisant des algorithmes de classification

supervisée utilisant comme descripteurs des indices lexicaux extraits de la transcription (Barzilay *et al.*, 2000; Damnati et Charlet, 2011; Liu, 2006), des indices acoustiques/prosodiques (Salamin *et al.*, 2009; Bigot *et al.*, 2010), ou les deux (Dufour *et al.*, 2012). Ces études ont mis en avant l'efficacité de l'algorithme de boosting sur des règles de décision à combiner ces différents descripteurs. Nous proposons ici une modification de cet algorithme de classification, en remplaçant les règles de décision simples par des mini arbres de décision que nous appelons bonzaïs afin d'améliorer l'efficacité de combinaison de ces différents descripteurs. Nous proposons alors un système d'identification de rôle de locuteur exploitant de multiples indices de descriptions comparable au système de (Dufour *et al.*, 2012) à l'exception notable de notre algorithme de classification. Nos expériences comparatives montrent que notre algorithme combine les indices de description de manière beaucoup plus performante que l'algorithme original tout en diminuant le temps nécessaire à l'apprentissage. Le taux de prédiction des rôles est ainsi amélioré relativement de 19,5% tandis que le temps d'apprentissage est diminué d'un facteur 3.

La suite du papier est organisée de la manière suivante : le chapitre 2 présente notre système de reconnaissance de rôle avec notamment les descripteurs utilisés, le chapitre 2.2 expose les modifications de l'algorithme de classification que nous proposons et les bénéfices que nous espérons en tirer, la section 3 présente les expériences comparatives effectuées sur le corpus EPAC.

2 Système de reconnaissance de rôle du locuteur

Le problème d'identification du rôle du locuteur est ici vu comme un problème de classification multi-classes avec un jeu de descripteurs variés. Les descripteurs utilisés puis l'algorithme de classification sont présentés dans les deux sous-sections suivantes.

2.1 Ensembles de descripteurs

2.1.1 Caractérisation de la parole spontanée

Une méthode permettant de détecter automatiquement la parole spontanée dans des documents audio a été proposée par (Dufour *et al.*, 2009). L'objectif de cet outil est d'associer à chaque segment de parole une des trois classes de spontanéité : parole *préparée*, *faiblement spontanée* et *fortement spontanée*. Des caractéristiques acoustiques (durées des voyelles, durées phonémiques, pitch...) et linguistiques (nombre de répétitions et de noms propres, taille du découpage syntaxique...) sont extraites pour chaque segment à partir d'un système de transcription automatique de la parole (Deléglise *et al.*, 2009). De plus amples informations sur les caractéristiques peuvent être trouvées dans (Dufour *et al.*, 2009).

2.1.2 Analyse des réseaux sociaux (SNA)

Nous avons également intégré des indices prenant en compte la position de chaque locuteur dans le dialogue, c'est à dire permettant de savoir comment celui-ci interagit avec les autres locuteurs. Des indices SNA sont exploités notamment dans (Garg *et al.*, 2008; Wang *et al.*, 2011). L'efficacité de ces indices, combinés à ceux de « **SPONTA** », a été montrée dans le système (Dufour *et al.*, 2012). L'objectif de cette méthode est d'être capable de déterminer la *centralité* de chaque locuteur par rapport aux autres dans une émission, en considérant que le locuteur i dialogue

avec le locuteur j si j intervient juste après i dans la transcription. En s'inspirant de (Vinciarelli, 2007), la centralité se calcule selon l'équation :

$$C_i = \frac{\sum_{j=1}^{nb} \chi D_{i,j}}{\sum_{j=1}^{nb} D_{i,j}} \quad (1)$$

Avec $\chi = 1$ si $D_{i,j} = 1$, et $\chi = 0$ sinon, C_i la centralité de i, nb le nombre de locuteurs et $D_{i,j}$ la distance entre i et j. Cette distance est exprimée en nombre de liens (orientés) à parcourir pour atteindre chaque nœud.

En plus de la centralité, la *couverture du locuteur*, correspondant au temps écoulé entre la première et la dernière intervention du locuteur dans l'émission, est calulée puis normalisée par la durée de l'émission.

Dans la suite, « **SNA** » fera référence aux indices *centralité* et *couverture du locuteur*.

2.1.3 *Ngrammes* multi-niveaux

Parmi les indices les plus pertinents, se trouvent ceux extraits de la transcription elle-même. La méthode usuelle est l'extraction de *Ngramme* de mots. Parfois d'autres niveaux d'informations sont ajoutés, tels que les lemmes, les catégories syntaxiques (POS) des mots ou tout autre niveau qui est pertinent pour la tâche visée. Usuellement, les *Ngrammes* sont extraits d'un unique niveau, indépendamment des autres (Wang *et al.*, 2011). Ainsi, les *Ngrammes* extraits des mots sont très spécifiques et ont peu de pouvoir de généralisation, tandis que les *Ngrammes* extraits de la séquence d'étiquettes syntaxiques correspondante généralisent souvent trop et ne sont pas assez précis.

Un classifieur peut tirer avantage des deux informations mais les considérer de manière séparée n'est pas optimal. Ici nous proposons d'extraire des *Ngrammes* multi-niveaux. Ces *Ngrammes* sont construits en générant la liste exhaustive des combinaisons possibles entre ces différents niveaux d'informations. L'idée est de capter les avantages des différents niveaux en composant des *Ngrammes* à partir de ces différents niveaux. Ainsi, on espère combiner la précision des *Ngrammes* de mots avec le pouvoir généralisant des *Ngrammes* de composants plus haut niveau. Dans ce travail, 3 niveaux d'informations associés à la transcription seront utilisés :

1. la transcription en mot
2. la séquence d'étiquette syntaxique correspondante (POS) prédite par TreeTagger (Schmid, 1994)
3. une combinaison de 2 éléments : l'entité nommée du mot si il en existe une, son lemme sinon. L'entité nommée est prédite selon le système (Raymond et Fayolle, 2010)

La présence de ces indices de descriptions dans notre système sera référencée par « **TEXT** ».

Sur la base de ces indices, 4 sortes de *Ngrammes* pourront être créés. Des *Ngrammes* uniquement composés de mots (W), des *Ngrammes* uniquement composés de POS (P), des *Ngrammes* uniquement extraits à partir du niveau EN/lemme (LNE) et également des *Ngrammes* multi-niveaux ($WPLNE$), *ie.* composés des 3 niveaux à la fois .

Lors de nos expériences, nous testerons le système avec différents jeux de descripteurs *Ngrammes* et notamment le jeu de descripteurs *Ngrammes* extrait à partir de tous les niveaux indépendamment, que nous noterons $W - P - LNE$, ou alors générés à partir de tous les niveaux (*Ngrammes* multi-niveaux) que nous noterons $WPLNE$.

Il est à noter que dans le cas d'utilisation d'*uni*-grammes, les *Ngrammes* multi-niveaux sont identiques à ceux extraits séparément.

2.2 Boosting de bonzaïs

De nombreuses études sur la reconnaissance de rôle du locuteur (Garg *et al.*, 2008; Barzilay *et al.*, 2000; Damnati et Charlet, 2011; Dufour *et al.*, 2009; Wang *et al.*, 2011) ont révélé l'efficacité de l'algorithme AdaBoost sur des règles de décision simples (*ie.* arbres de décision à deux feuilles) qui combinent différents descripteurs. Effectivement, les règles de décision sont induites directement à partir des différents indices tandis que l'algorithme de boosting permet de les combiner afin de produire un classifieur final très efficace.

En plus de cela, parmi les variantes d'algorithmes de boosting, certains d'entre eux sont multi-classes, comme AdaBoost.MH (Schapire et Singer, 2000). C'est notamment celui ci qui est appliqué dans de précédents travaux (Dufour *et al.*, 2012; Barzilay *et al.*, 2000). Bien que cet algorithme a montré de bons résultats, la combinaison linéaire des règles de décisions utilisée peut être insuffisante pour capturer certaines structures dans les données (*eg.* impossible de résoudre le problème du XOR). Afin de palier ce problème, l'idée appliquée ici consiste à implémenter l'algorithme de boosting sur des arbres de décision.

Nous avons décidé d'expérimenter le boosting sur de petits arbres de décision, que nous appelons *bonzaïs*, contraints par leur profondeur, suffisamment profond pour pouvoir capturer des structures dans les données mais pas trop pour conserver des règles générales et éviter le sur-apprentissage.

Dans les situations comme la notre exploitant des descripteurs de type *Ngramme*, le nombre de descripteurs s'élève très facilement à plusieurs millions même avec un N relativement faible. De ce fait, nous attendons 3 bénéfices de l'implémentation du boosting de bonzaïs :

1. un bonzaï étant plus complexe qu'une règle de décision, il est capable de modéliser et donc de capturer des informations plus complexes à partir des données d'apprentissage, nous espérons alors un gain de performance ;
2. pour les mêmes raisons, nous espérons que notre algorithme de boosting produise un classifieur aussi performant que l'original avec moins d'itérations que l'algorithme original n'en nécessite. La construction du bonzaï étant intrinsèquement parallélisable, alors que l'algorithme de boosting ne l'étant pas, la réduction du nombre d'itération permet d'espérer un gain de temps important sur l'apprentissage ;
3. avec l'utilisation de bonzaïs en tant que classifieur de base, nous nous demandons si le bonzaï n'est pas capable d'implicitement créer des *Ngrammes* à partir d'observations isolées telles que le sac de mots (*uni*-grammes). Si cela était le cas, il ne serait plus utile de générer explicitement les *Ngrammes* mais de se contenter de générer des *uni*-grammes, ce qui diminuerait drastiquement le nombre de descripteurs et par conséquent le temps d'apprentissage nécessaire à l'algorithme.

Nos bonzaïs sont induits suivant le critère du pseudo-loss décrit dans (Schapire et Singer, 2000)

qui est recommandé pour construire les règles de décision simples dans l'algorithme original. Nous répétons alors de manière récursive l'opération pour développer l'arbre. La récursion est stoppée en fonction de deux critères :

1. la profondeur : le bonzaï ne peux plus se développer lorsque on atteint une profondeur fixée
2. le gain selon le critère de pseudo-loss : si ce gain n'est pas positif lors de la tentative de subdivision d'un nœud, ce nœud devient une feuille

Notre implémentation est disponible en ligne (Raymond, 2010).

3 Expériences comparatives

3.1 Corpus

Le projet EPAC (Estève *et al.*, 2010) concerne le traitement de données audio non structurées. L'objectif principal de ce projet a été de proposer des méthodes d'extraction d'information et de structuration de documents spécifiques aux données audio. Les données audio traitées durant le projet EPAC proviennent d'émissions radiophoniques enregistrées entre 2003 et 2004 au sein de trois radios françaises : France Info, France Culture et RFI. Au cours de ce projet, 100 heures de données audio ont été manuellement annotées, avec principalement des émissions contenant une forte proportion de parole spontanée (interviews, débats, talk shows...). Le corpus EPAC inclut des informations supplémentaires au niveau des locuteurs et des émissions transcrites. Plus précisément, le rôle, la fonction et la profession de chaque locuteur ont été manuellement annotés selon la disponibilité des informations. Ainsi, un même locuteur a un seul rôle général (par exemple *Invité*, *Interviewé*,...) mais qui peut être affiné avec un maximum de deux autres étiquettes (par exemple, pour un *Invité* : *politicien* / *premier ministre*). L'intégralité du corpus a été manuellement annotée en 10 rôles par un expert linguiste.

3.2 Évaluation

Tous les résultats sont présentés en terme de Taux d'erreur en rôle (TER) et ont été obtenus sur la totalité du corpus EPAC via une validation croisée sur 20 sous-parties du corpus. Le TER est la somme des rôles erronés divisée par le nombre de rôles dans la référence.

Plusieurs paramétrages de notre système sont testés en faisant varier :

1. la taille du *Ngramme* : utilisation soit d'*uni*-grammes *1g*, soit de *bi*-grammes *2g*
2. la profondeur du bonzaï utilisé dans l'algorithme de boosting : de 1 à 3. Notons que la profondeur 1 (prof.1) est équivalent à une règle de décision simple et donc à l'algorithme original utilisé dans les travaux antérieurs de la littérature sur le sujet. La profondeur 3 (prof.3) correspond quant à elle à un arbre à 8 feuilles maximum.
3. les descripteurs : $SPONTA$, SNA, $TEXT$ indépendamment les uns des autres ou en combinaison
4. les différents niveaux du descripteur $TEXT$: uniquement les mots W, ou alors à la fois les mots, les pos et la combinaison entité nommée/lemme de manière indépendante $W-P-LNE$ ou multi-niveaux $WPLNE$ (*cf.* 2.1.3)

Notre système paramétré avec la configuration *TEXT(2g W)+SPONTA+SNA prof.1* est équivalent au système publié dans (Dufour *et al.*, 2012). Ce système noté $REFERENT$ nous permettra par la suite de faire des comparaisons avec les différents résultats de notre système.

3.2.1 Analyse sur l'apport des *Ngrammes* multi-niveaux

La table 1 montre les résultats obtenus par notre système utilisant l'ensemble des descripteurs décrits en 2.1 et faisant varier la profondeur du bonzaï et la façon dont les *Ngrammes* sont extraits à partir de $TEXT$. Les résultats sont donnés pour 2000 itérations, valeur où chaque système semble avoir atteint sa performance maximale.

TER pour 2000 itérations		**prof. 1**	**prof. 2**	**prof. 3**
1g	*W*	27.1	23.8	22.3
	W-P-LNE	26.1	24.5	22.3
	WPLNE	26.2	23.4	22.9
2g	*W*	26.1*	22.2	22.3
	W-P-LNE	25.7	**21.7**	21.5
	WPLNE	**24.7**	22.8	**21.0**

TABLE 1 – Résultats des système appris avec tous les descripteurs (TEXT+SNA+SPONTA) en fonction de la profondeur de l'arbre et des types de *Ngrammes* $TEXT$ utilisés. Le système $REFERENT$ est noté avec une étoile.

Les résultats présentés dans le tableau 1 montrent qu'avec l'utilisation de règles simples ($prof.1$) les *Ngrammes* multi-niveaux ($WPLNE$) améliorent les performances en comparaison avec les *Ngrammes* extraits séparément ($W-P-LNE$). En effet, la figure 1 montre que les *Ngrammes* multi-niveaux obtiennent rapidement de bien meilleurs résultats que les *Ngrammes* $W-P-LNE$, même si dans le cas de cette tâche, les *Ngrammes* de mots seuls montrent de bons résultats.

Il est à noter que la configuration $W-P-LNE$ et $W-PLNE$ en *uni*-grammes sont équivalentes et ne peuvent se distinguer que pour $N > 1$. La différence de 0.1% entre ces configurations en *1g* s'explique par le fait que parmi les millions de descripteurs *Ngrammes* beaucoup d'entre eux ont le même pouvoir discriminant et l'algorithme de construction décide arbitrairement de conserver le premier de la liste. Ainsi l'ordre des *Ngrammes* étant présentés différemment selon la configuration, le classifieur obtenu est donc légèrement différent et cela explique cette instabilité mineure des performances obtenues.

3.2.2 Analyse sur l'apport des bonzaïs

L'utilisation des bonzaïs améliore très nettement la performance de l'algorithme de boosting, le tableau 1 montre bien que chaque niveau supplémentaire dans la profondeur des arbres améliore le système de manière très significative. La figure 2 montre que la profondeur nécessaire à ce gain est petite : un bonzaï de profondeur 2 (4 feuilles max) améliore grandement les résultats obtenus avec une règle simple (équivalent à un bonzaï prof.1, deux feuilles). Le bonzaï prof.3 améliore sensiblement, ce qui n'est plus le cas pour des profondeurs supérieures.

La figure 2 montre que l'augmentation de la profondeur de l'arbre permet de converger avec moins d'itérations vers le taux d'erreur minimum. Ceci est particulièrement intéressant sur les tâches faisant intervenir des millions d'indices textuels (*Ngrammes*). En effet l'algorithme d'induction d'arbre fait des évaluations exhaustives de chaque descripteur tandis que l'algorithme de boosting est itératif et non parallélisable. En effet, la figure 2 révèle que le système configuré avec des *2g* et une profondeur 2 fait jeu égal avec le système configuré avec des *1g* et une

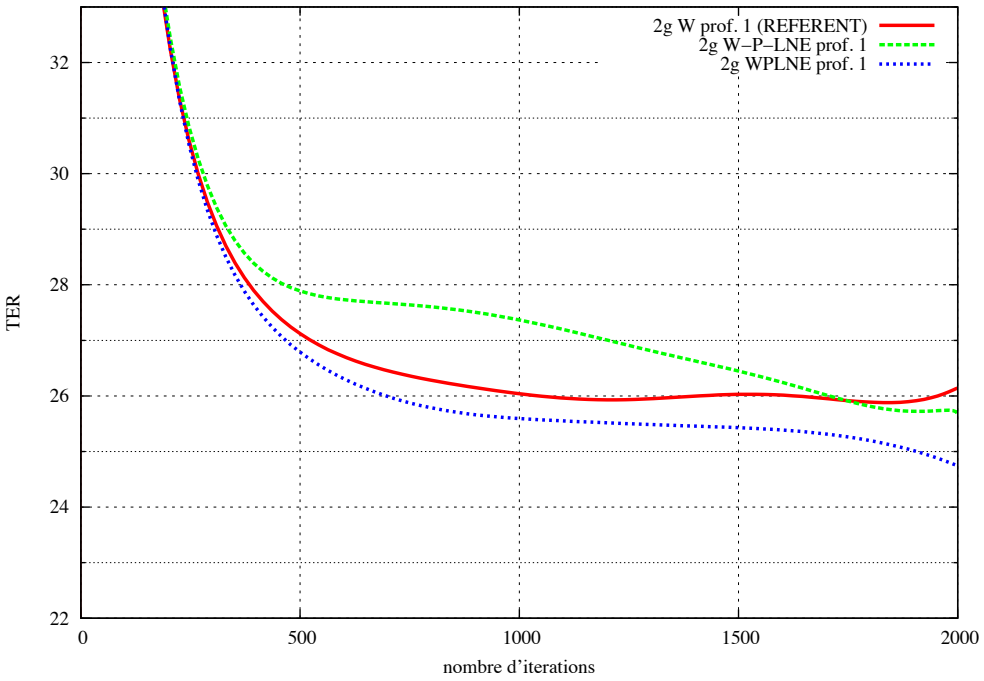


FIGURE 1 – Comparaison des perfomances des différents niveaux *Ngrammes TEXT* pour le système *TEXT(2g)+SNA+SPONTA*

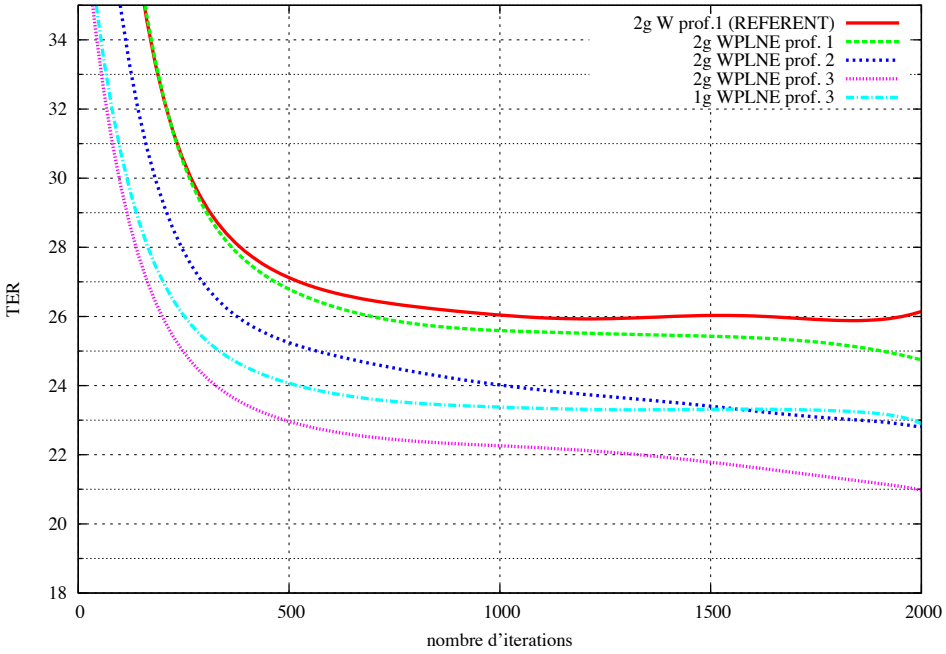


FIGURE 2 – Courbes de résultats pour différentes configurations de profondeur de bonzaïs et de descripteurs en fonction du nombre d'itération de boosting).

profondeur 3, la complexité générée par un arbre plus profond est largement compensée par l'économie faite par la réduction du nombre de descripteurs à évaluer : le gain de temps est d'un facteur 5. Ce même système comparé avec le système référent permet d'atteindre la même performance que celui ci avec 5 fois moins d'itérations ce qui lui permet un gain de temps d'un facteur 3.

Le tableau 2 présente la contribution de chaque catégorie d'indices isolée ou combinée avec les autres pour les deux systèmes : notre meilleure configuration *TEXT(2g WPLNE) prof.3* et la configuration référent (Dufour *et al.*, 2012) *TEXT(2g W) prof.1*. Il apparaît très clairement que l'utilisation des bonzaïs permet systématiquement d'améliorer les performances par rapport à l'utilisation de règles simples. Le système référent présente un TER de 26,1%, l'utilisation de bonzaïs permet avec exactement le même jeu de descripteurs de réduire le TER de 5,1% absolu.

Il est intéressant de noter que chacun des types de descripteurs apporte un gain. Notre meilleure configuration a choisi 65% de features de type $TEXT$, 27% de type $SPONTA$ et 8% de type SNA. Les bonzaïs permettent d'utiliser plusieurs types de features à chaque itération. Ainsi, quasiment 50% des itérations utilisent des features de types $TEXT$ et $SPONTA$ et près d'un tiers combinent les trois types de descripteurs.

indices/système	*proposé*	*référent*
TEXT	29.8	35.5
+ SPONTA	23.7	28.4
+ SNA	**21.0**	**26.1**
+ SNA	22.9	29.1
SPONTA	32.1	34.8
+ SNA	25.9	29.1
SNA	40.0	45.4

TABLE 2 – Résultats en terme de TER du système *proposé* et *référent* entraîné avec différentes combinaison d'indices

4 Conclusion

Nous avons proposé dans ce travail une modification de l'algorithme de boosting qui est très largement utilisé dans la tâche de la reconnaissance du rôle du locuteur. Nous proposons de remplacer les règles simples utilisées systématiquement dans la littérature en tant que classifieur faible dans l'algorithme de boosting par des mini arbres de décision appelés bonzaïs. Un système de reconnaissance de rôle à base de ce classifieur est évalué et comparé à un système état-de-l'art publié dans (Dufour *et al.*, 2012). Les expériences comparatives montrent qu'avec un jeu de descripteurs identique notre algorithme permet de mieux les exploiter en proposant une amélioration de 19,5% des performances tout en réduisant drastiquement le temps d'apprentissage du système.

Références

Barzilay, R., Collins, M., Hirschberg, J. et Whittaker, S. (2000). The rules behind roles : Identifying speaker role in radio broadcasts. *In AAAI*, pages 679–684.

Bigot, B., Ferrané, I., Pinquier, J. et André-Obrecht, R. (2010). Speaker role recognition to help spontaneous conversational speech detection. *In Searching Spontaneous Conversational Speech*, pages 5–10, Firenze, Italie.

Damnati, G. et Charlet, D. (2011). Robust speaker turn role labeling of tv broadcast news shows. *In ICASSP*, Prague, République Tchêque.

Deléglise, P., Estève, Y., Meignier, S. et Merlin, T. (2009). Improvements to the LIUM French ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? *In Interspeech*, pages 2123–2126, Brighton, Angleterre.

Dufour, R., Estève, Y., Deléglise, P. et Béchet, F. (2009). Local and global models for spontaneous speech segment detection and characterization. *In ASRU*, Merano, Italie.

Dufour, R., Laurent, A. et Estève, Y. (2012). Combinaison d'approches pour la reconnaissance du rôle des locuteurs. *In JEP*, Grenoble, France.

Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F. et Farinas, J. (2010). The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news. *In LREC*, pages 1686–1689, Valletta, Malte.

Garg, P. N., Favre, S., Salamin, H., Hakkani-Tür, D. et Vinciarelli, A. (2008). Role recognition for meeting participants : an approach based on lexical information and social network analysis. *In ACM Multimedia Conference (MM'08)*, pages 693–696, Vancouver, Canada.

Liu, Y. (2006). Initial study on automatic identification of speaker role in broadcast news speech. *In Human Language Technology Conference of the NAACL*, pages 81–84, New York, USA.

Raymond, C. (2010). bonzaiboost. http ://bonzaiboost.gforge.inria.fr/.

Raymond, C. et Fayolle, J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. *In Traitement Automatique des Langues Naturelles*, Montréal, Canada.

Salamin, H., Favre, S. et Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings : Using social affiliation networks for feature extraction. *In IEEE Transactions on Multimedia*, volume 11, pages 1373–1380.

Schapire, R. E. et Singer, Y. (2000). BoosTexter : A boosting-based system for text categorization. *Machine Learning*, 39:135–168.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Vinciarelli, A. (2007). Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transaction on Multimedia*, 9(6): 1215–1226.

Wang, W., Yaman, S., Precoda, K. et Richey, C. (2011). Automatic identification of speaker role and agreement/disagreement in broadcast conversation. *In ICASSP*, pages 5556–5559.

Le contraste quantitatif en saami de Lule : un système à trois niveaux

Nora Fangel-Gustavson[1], Rachid Ridouane[1]
(1) LLP (UMR 7018, CNRS-Sorbonne-Nouvelle)
nora.fangel-gustavson@univ-paris3.fr, rachid.ridouane@univ-paris3.fr

RESUME

Le saami de Lule présente un système phonologique avec trois niveaux de quantité consonantique, un phénomène typologiquement rare. Ce travail examine la manière dont ce contraste ternaire est acoustiquement implémenté en se basant sur les productions de 7 locuteurs natifs enregistrés à Tysfjord en Norvège. Les analyses, et le traitement statistique des données, indiquent que ce contraste phonologique se manifeste principalement par une augmentation systématique et substantielle de la durée consonantique d'une quantité à une autre. Un léger réajustement de la durée des voyelles adjacentes a aussi été observé, avec une réduction plus importante pour la voyelle qui suit. Au-delà de ces paramètres temporels, nous avons aussi pu dégager d'autres corrélats de la (super)gémination impliquant plus particulièrement l'intensité des voyelles adjacentes. Les résultats de l'analyse phonétique sont interprétés du point de vue phonologique en questionnant notamment la manière dans les géminées et supergéminées sont représentées dans le cadre de la théorie moraïque.

ABSTRACT

Quantity Contrast in Lule Saami: A Three-Way System
Lule Saami presents a phonological three-way consonantal quantity system, a typologically rare phenomenon. This study examines the acoustic implementation of this contrast based on acoustic data from 7 native speakers, recorded in Tysfjord in Norway. Results show that the three degrees of phonological quantity are primarily manifested through a substantial and systematic consonant duration increase form one quantity to another. There is furthermore a slight readjustment on adjacent vowel durations, particularly on the vowel following the target consonant. In addition to these temporal cues, other correlates are involved in the implementation of this contrast, including more specifically the intensity of the adjacent vowel. Implications of the phonetic results are discussed, with particular attention to the issue of (super)geminate representation within moraic theory.

MOTS-CLES : Langue saami, contraste ternaire de quantité, phonétique, phonologie
KEYWORDS: Saami language, three-way quantity contrast, phonetics, phonology

1 Introduction

Le saami de Lule est une des dix langues indigènes saami, de la branche finno-ougrienne des langues ouraliennes, parlées au nord de la Scandinavie et de la péninsule de Kola en Russie. Le saami de Lule, parlé par environ 600 locuteurs en Norvège et en Suède (Morén-Duolljá 2010), présente un phénomène typologiquement rare : son système phonologique distinguerait trois

degrés différents de longueur consonantique ; les consonnes peuvent être brèves, longues ou superlongues, que nous transcrivons ici [C], [C:], ou [C:C], respectivement. Le lexique (1a) et la morphologie (1b) se servent de ces trois degrés de quantité, appelés quantité 1 (Q1), quantité 2 (Q2) et quantité 3 (Q3), pour opposer des mots :

1)

a)	*Quantité 1*	*Quantité 2*	*Quantité 3*
	oare [ʔoɑrɪɛ]	oarre [ʔoɑr:ɪɛ]	oar're [ʔoɑ͡r:rɪɛ]
	« raisons »	« raison » OU « écureuils » (*mots homophones)*	« écureuil »

b)	*Quantité 1*	*Quantité 2*	*Quantité 3*
	(dån) basá [basɑ:]	(sån) bassá [bas:ɑ:]	(sån) bas'sá [bas:sɑ:]
	« (tu) laves »	« (il/elle) lave »	« (il/elle) commence à laver »

Les systèmes à trois degrés de quantité constituent un sujet de controverse aussi bien parmi les phonologues que les phonéticiens. D'un point de vue phonologique, plusieurs auteurs soutiennent que les oppositions de quantité sont maximalement binaires (e.g. Chomsky & Halle 1986, Odden 1997). McCarthy et Prince (1986), de leur côté, postulent que la syllabe est maximalement bimoraïque, excluant de fait l'existence de syllabe contenant des consonnes superlongues (voir plus bas). D'un point de vue phonétique, la question se pose quant à l'implémentation d'un contraste à trois degrés de quantité : s'agit-il réellement d'une opposition de durée entre les trois types de consonnes ou est-ce que d'autres corrélats (notamment qualitatifs) entrent en jeu ? Dans ce travail nous nous intéresserons plus particulièrement à cette dernière question à travers une analyse acoustique.

Des systèmes phonologiques à trois quantités consonantiques existent dans plusieurs langues de la famille finno-ougrienne. L'estonien, par exemple, a trois degrés de quantité vocalique et consonantique. La manifestation phonétique de cette triple opposition implique l'interaction entre plusieurs paramètres prosodiques : la durée, l'accent lexical, et la hauteur. Elle est aussi dépendante de la position de la syllabe dans le mot (Lehiste 1997, Ehala 2003, Spahr 2012). Le saami d'Inari a également trois niveaux de quantité consonantique. Mais il a été démontré que dans cette langue, et de manière systématique, la durée de la voyelle précédant la consonne a une corrélation négative avec la durée de la consonne, suggérant que le contraste de durée consonantique est fondamentalement binaire (Bye *et al*. 2009). En saami du Nord, autre langue considérée comme ayant une triple opposition de quantité, le contraste ternaire de surface serait le résultat d'un processus phonologique d'alignement entre les segments et les mores (Bals *et al*. 2012).

Le saami de Lule, l'objet de notre étude, est une langue peu documentée dans la littérature, et les recherches qui traitent de son système phonétique sont extrêmement limitées. A notre connaissance, une seule étude (Engstrand 1987) a examiné la quantité consonantique dans cette langue d'un point de vue expérimental. Les résultats obtenus dans ce travail suggèrent que la triple quantité phonologique est implémenté par une augmentation de durée de la consonne d'une quantité à l'autre, avec un ratio d'environ 1 : 2 : 3. Cependant, cette étude n'a traité que d'une

seule consonne ([n]) dans un corpus enregistré avec un seul locuteur. A l'évidence, une étude plus détaillée, examinant plus de consonnes et plus de locuteurs est nécessaire pour tenter de déterminer comment les trois degrés de quantité sont réalisées acoustiquement en saami de Lule. C'est l'objectif que nous nous fixons pour ce travail.

2 La quantité consonantique : phonétique et phonologie

Comme nous l'avons souligné plus haut, le contraste de durée à trois degrés a été la source de controverses sur les plans phonologique et phonétique. La Phonologie Autosegmentale (Goldsmith 1976) permet une représentation non-linéaire des segments et de leurs associations avec des positions sur une couche prosodique. Ces positions prosodiques sont soit des unités CV (McCarthy 1979, 1989 ; Halle & Vergnaud 1980 ; Clements & Keyser 1983) soit des positions moraïques (Hyman 1985 ; McCarthy & Prince 1986 ; Hayes 1989). Dans le cadre de la théorie moraïque, une consonne simple est représentée comme étant non-moraïque et une géminée comme étant mono-moraïque. De ce point de vue, une supergeminée serait bi-moraïque, donnant lieu à une syllabe tri-moraïque (si nous postulons que chaque syllabe doit comprendre un noyau vocalique comptant pour une more). Une telle syllabe trimoraïque serait problématique en vue du principe de « binarité maximale» (McCarthy & Prince 1986), qui exige qu'une syllabe ne puisse contenir plus de deux mores. Pour autant, des arguments existent suggérant qu'une syllabe peut contenir trois unités de poids. En hindi, par exemple, l'assignation de l'accent distingue entre les syllabes légères (CV), lourdes (CVC) et superlourdes (CVVC) (Broselow et al. 1997). De telles différences sont analysées dans le cadre de la théorie moraïque en postulant une différence de structure moraïque, entre syllabe mono-moraïque, bi- moraïque, et tri-moraïque. D'autres arguments en faveur de l'existence de syllabes tri-moraïques sont tirés du hongrois, de l'estonien, et du persan (Hayes 1989 ; Ham 1998). La question de la trimoraïcité est pertinente pour l'analyse de l'opposition de quantité en saami de Lule. Si l'on peut démontrer que le contraste de quantité dans cette langue est principalement implémenté par la durée, alors on peut légitimement poser une distinction phonologique à trois niveaux. Une représentation phonologique adéquate d'une telle opposition ternaire serait dès lors nécessaire, et la théorie moraïque peut fournir un moyen formel pour la développer.

Plusieurs travaux postulent que la durée phonétique est organisée autour de la more (Port et al. 1987 ; Hubbard 1995 ; Broselow et al. 1997 ; Ham 1998). Dans le travail très cité de Broselow et al. (1997), l'examen de 4 langues (hindi, malayalam, arabe levantin et arabe égyptien) a montré de manière particulièrement saisissante qu'il y a une corrélation étroite entre représentation moraïque et durée segmentale, suggérant que la structure moraïque est directement reflétée dans le timing phonétique de ces langues. Concernant la gémination, la littérature abonde en travaux montrant que le contraste entre consonnes simples et géminées se traduit principalement par des différences de durées acoustiques et/ou articulatoires. Cette différence de durées – notamment en position intervocalique – semble être un aspect universel, observé dans différentes langues du monde. Ridouane (2010) a ainsi montré que pour 24 langues opposant simples aux géminées, le seul paramètre acoustique partagé par les géminées est qu'elles sont significativement plus longues que leurs contreparties simples. Outre la durée, le contraste de gémination peut se signaler à travers des corrélats non-temporels, comme l'intensité. Ces corrélats sont souvent considérés comme étant secondaires, introduits pour renforcer d'avantage l'opposition entre simples et géminées (Keyser & Stevens 2006, Clements & Ridouane 2006). Concernant le contraste de longueur à trois degrés, la situation est moins claire. Comme nous l'avons précisé plus haut, la triple opposition en

estonien, par exemple, implique, outre la durée, des facteurs suprasegmentaux comme la manipulation de la F0. De même en saami d'Inari, la durée ne permet de contraster que deux degrés de quantité, des paramètres qualitatifs entre en jeu pour permettre de contraster les trois degrés. Dans ce travail, nous examinerons des paramètres temporels et non temporels pour déterminer si le contraste simples *vs*. géminées *vs*. supergéminées en saami de Lule est principalement un contraste temporel ou si d'autres corrélats entrent en jeu pour signaler cette opposition.

2.1 Méthode

Nous avons enregistré sept locuteurs natifs du saami de Lule, 4 hommes et 3 femmes, âgés de 26 à 81 ans (âge moyen = 49,9 ans), tous membres de la communauté saami de Lule à Tysfjord en Norvège. Tous les locuteurs sont bilingues et parlent le norvégien en plus du saami de Lule (comme c'est le cas pour l'écrasante majorité des locuteurs natifs de cette langue).

Groupe	Locuteur	Genre	Age
A (né après1985)	Loc. 1	Femme	26
A (né après1985)	Loc. 2	Homme	27
B (né 1960 - 1985)	Loc. 3	Homme	38
B (né 1960 - 1985)	Loc. 4	Femme	51
C (né avant 1960)	Loc. 5	Homme	63
C (né avant 1960)	Loc. 6	Femme	63
C (né avant 1960)	Loc. 7	Homme	81

TABLE 1 – Liste des locuteurs

Nous avons pris soin de choisir les sujets selon leur sexe et leur âge et nous avons obtenu une classification en 3 groupes, comme l'illustre la table 2. L'objectif ici est de déterminer si oui ou non il y'a un effet de génération sur la façon dont le contraste est implémenté, au vu de la pression de plus en plus forte exercée par le norvégien sur cette langue, en danger d'extinction.
Le corpus utilisé est constitué de six verbes dans trois formes grammaticales différentes (18 items en total), permettant d'obtenir des triplets minimaux, ayant une même structure syllabique (C)V**C**V, où C est simple, géminée ou supergéminée. Les consonnes testées sont [l], [m], [n], [r], [s] et [v] , chacune attestée dans un des 6 mots tests suivants ; *ballat*, *njammat*, *mannat*, *bårråt*, *bassat* et *tjavvat*. Le tableau 3 présente la liste des 18 items utilisés dans cette étude :

Forme infinitive du verbe	Q1 2p.SG	Q2 3p.SG	Q3 Incept. 3p.SG
« ballat » (« craindre »)	[palɑ:]	[pal:ɑ:]	[pal:lɑ:]
« ŋjammat » (« téter »)	[ŋamɑ:]	[ŋam:ɑ:]	[ŋam:mɑ:]
« mannat » (« partir / voyager »)	[manɑ:]	[man:ɑ:]	[man:nɑ:]
« bårråt » (« manger »	[pɔrɔ:]	[pɔr:ɔ:]	[pɔr:rɔ:]
« tjavvat » (« faire bouger la neige p.ex. par des rafales de vent »)	[tʃ͡avɑ:]	[tʃ͡av:ɑ:]	[tʃ͡av:vɑ:]
« bassat » (« laver »)	[pasɑ:]	[pas:ɑ:]	[pas:sɑ:]

TABLE 2 – Liste des items utilisés

Les voyelles précédant et suivant la consonne cible sont toujours les mêmes ([a] et [ɑɑ]), sauf pour la consonne [r] où aucun verbe avec ces voyelles dans les mêmes positions n'a pu être trouvé. A la place, le mot test contient les voyelles [ɔ] et [ɔɔ] (avant et après la consonne). Pour éliciter ces items tout en parant à toute influence éventuelle des formes écrites, nous avons présenté les verbes à la forme infinitive et demandé aux locuteurs de les conjuguer dans les trois formes grammaticales indiquées dans le tableau 2. Une phrase cadre a été utilisée – *Dån/Sån …, guoakta gålmmå* « Tu/il/elle …, deux trois » – dans le but d'éviter les effets de liste (intonation montante) sur le mot cible. Chaque phrase a été répétée par chaque locuteur 4 fois, donnant lieu à corpus composé de 504 items (18 items * 7 locuteurs * 4 répétitions). Les données acoustiques ont été segmentées, annotées et analysées avec Praat (Boersma & Weenick 2012). Nous avons analysé des variables acoustiques temporelles et non-temporelles, incluant la durée, la fréquence fondamentale (f0), la structure formantique F1, F2, F3, et l'intensité (dB). Nous avons extrait ces valeurs (valeurs moyennes des premiers et derniers 30 ms et valeur absolue du milieu) à partir de chaque segment de la séquence V<u>C</u>V suivant l'attaque initiale du mot.

3 Résultats

3.1 Paramètres temporels

Le triple contraste de quantité en saami de Lule est principalement implémenté par des différences significatives (p <.0001) de durée entre les consonnes simples, géminées et supergéminées. Tous mots et locuteurs confondus, la durée moyenne était de 90 ms pour la consonne simple (Q1), 215 ms pour la géminée (Q2), et 323 ms pour la supergéminée (Q3). Le plus grand contraste de durée entre une consonne en Q2 et une en Q3 a été trouvé pour [v] dans le mot *tjavvat*; où la durée moyenne était de 204 ms en Q2 et de 319 ms en Q3, soit une augmentation de 56%. Le contraste le plus petit a été constaté pour [s] dans le mot *bassat* ; l'augmentation de durée était de 253 ms en Q2 à 353 ms en Q3, soit une augmentation de 40%. La figure 1 illustre les résultats et présente les moyennes de durée pour V1, <u>C</u> et V2 pour les trois contrastes phonologiques.

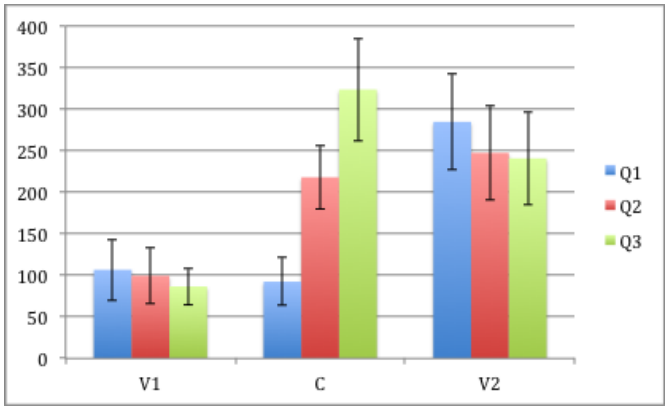


FIGURE 1 – Durées moyennes des segments V1, C, V2 pour les trois types de quantité (Q1, Q2, Q3) pour les 7 locuteurs. Les barres d'erreur représentent les écarts types.

Comme le montre la figure 1, en plus de la durée de la consonne cible, nous avons constaté un léger réajustement temporel des voyelles adjacentes. L'effet le plus important était sur la voyelle suivant la consonne cible (V2), avec une durée moyenne allant de 286 ms en Q1 à 250 ms en Q2, et à 242 ms en Q3. De même, la durée moyenne de toute la séquence VCV (qui suit l'attaque initiale) était systématiquement plus importante de Q1 à Q2, et de Q2 à Q3 ; allant de 484 ms en

Q1, à 565 ms en Q2 et à 651 ms en Q3. La figure 2 illustre la relation entre les segments dans la séquence V1CV2 pour les trois contrastes de quantité.

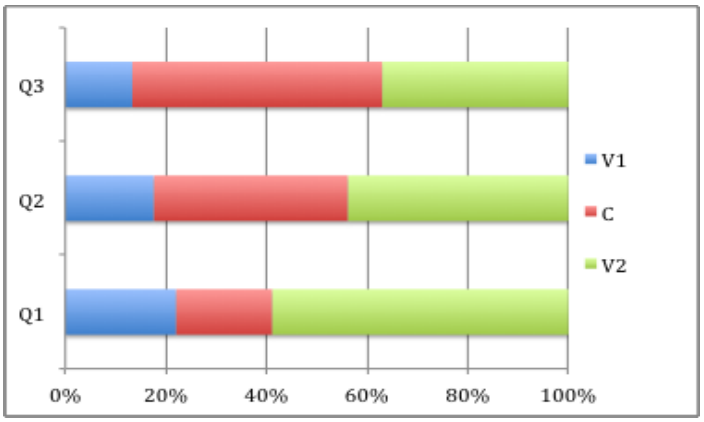


FIGURE 2 – Durées moyennes relatives des segments post-attaque (V1, C, V2) dans les trois contextes (Q1, Q2, Q3) pour les 7 locuteurs.

Une question subsidiaire soulevée au moment de la présentation de nos données a été de savoir si la façon dont le contraste de quantité est réalisé en saami de Lule varie d'un point de vue générationnel. Nous avons évalué l'influence du groupe d'âge sur la durée consonantique et n'avons trouvé aucune différence significative entre les trois groupes pour les quantités 1 et 2. Cependant, nous avons trouvé que les locuteurs du groupe B et C produisent la quantité 3 avec une durée significativement plus longue comparés aux locuteurs plus jeunes du groupe A (p <.0001). Il est prématuré de tirer des conclusions définitives de ces résultats. L'examen d'un groupe de locuteurs plus large est nécessaire pour déterminer si la tendance observée ici traduit un réel changement générationnel dans la façon dont le contraste ternaire de quantité est phonétiquement implémenté.

3.2 Paramètres non-temporels

3.2.1 Fréquence fondamentale

Plusieurs locuteurs ont produit un pattern d'intonation haut-bas dans la séquence dissyllabique VCV ; le pic de f0 coïncidant avec la frontière gauche de la consonne cible dans les trois quantités. Pour certains locuteurs, le pic était légèrement plus élevé dans Q2 et Q3 que dans Q1, mais ces différences de F0 pour les 3 quantités ne sont pas significatives (p>.05). La figure 3 illustre les variations de F0 pour la séquence VCV pour les 3 quantités, pour la forme « tjavvat » :

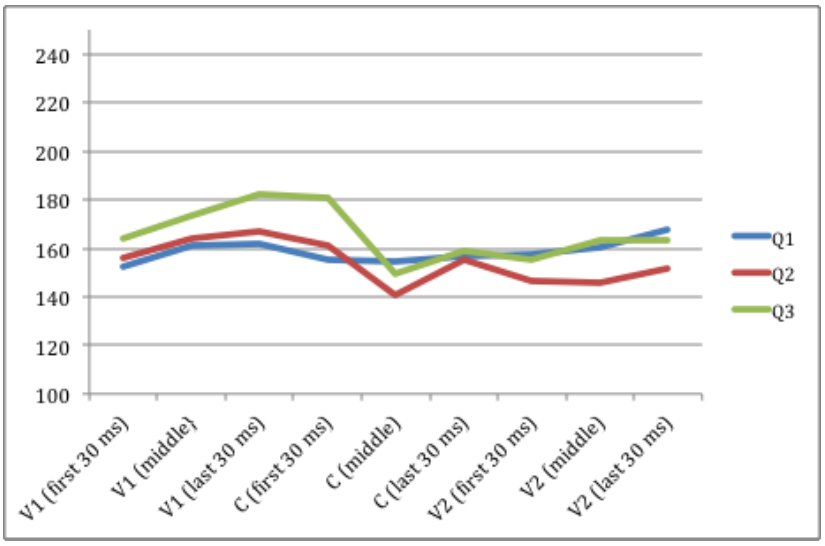


FIGURE 3 – Valeurs moyennes de la fréquence fondamentale (f0) des segments post-attaque pour la suite /avɑ:/ dans les trois contextes (Q1, Q2, Q3) pour les 4 locuteurs masculins.

Compte tenu des différences considérables parmi les locuteurs, nous n'avons pas pu dégager de stratégies tonales spécifiques permettant de distinguer les trois quantités.

3.2.2 Intensité

La quantité a un effet significatif sur l'intensité des voyelles adjacentes (V1 : p = .0124, V2 : p = .02) mais pas sur celle de la consonne cible (p = 0.18). Nous avons constaté que l'intensité de V1 augmentait d'une quantité à l'autre, alors que l'intensité de V2 baissait.

3.2.3 Qualité des voyelles adjacents

Les structures formantiques des voyelles adjacentes n'ont pas changé significativement d'une quantité à l'autre. Pour autant, nous avons observé une différence qualitative entre V1 et V2, mais liée aux différences de quantité entre ces voyelles. V1, qui est brève, est plus ouverte et plus antérieure que sa contrepartie longue V2, plus fermée et plus postérieure. Mais ces qualités observées ne changent pas selon la quantité de la consonne cible.

4 Conclusion et discussion

Les résultats de notre étude acoustique montrent que le contraste ternaire de quantité en saami de Lule se manifeste principalement par une augmentation systématique et substantielle de la durée consonantique d'une quantité à une autre, allant de 90 ms en moyenne pour la quantité 1, à 215 ms pour la quantité 2, et jusqu'á 323 ms pour la supergéminée. La quantité a un effet sur la durée des voyelles adjacentes, avec une réduction plus importante pour les voyelles qui suivent. Au-delà de ces paramètres temporels, nous avons aussi pu dégager d'autres corrélats de la (super)gémination impliquant plus particulièrement l'intensité des voyelles adjacentes. Ces corrélats semblent secondaires et jouent probablement un rôle de renforcement permettant de maximaliser la distance entre les trois catégories consonantiques.
Partant du postulat qu'il y a une relation étroite entre représentation phonologique et représentation phonétique (Pierrehumbert 1990, Broselow et al. 1997), et que cette relation doit se refléter dans la théorie linguistique, les caractéristiques phonétiques des (super)géminées en saami de Lule peuvent être rendues en représentant ces segments comme monomoraïques (pour les géminées) et

bi-moraïques (pour les supergéminées). La figure 4 illustre de telles représentations.

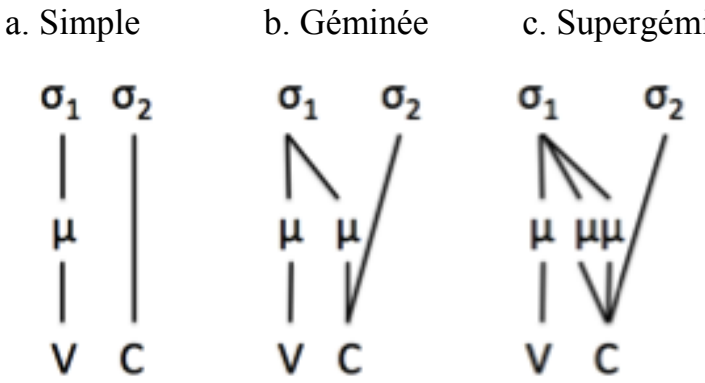


FIGURE 4 – Représentation moraïque des 3 quantités de consonnes en saami de Lule.

Une telle représentation permet d'établir un lien étroit entre les processus morphologiques qui manipulent la quantité en saami de Lule (passage de Q1 à Q2 et de Q2 à Q3) et la durée acoustique de ces segments (Q1 < Q2 < Q3). Passer d'une quantité à une autre correspondrait structurellement à l'ajout d'une more et se traduirait phonétiquement par l'augmentation de la durée. En hongrois, Ham (1998) a observé une corrélation entre la durée de la syllabe et le nombre de mores contenus dans cette syllabe: plus le nombre de mores augmente, plus la durée de la syllabe augmente. Cela correspond aux résultats que nous avons obtenus pour le saami de Lule. Les effets de réajustement observés dans notre étude, qui se manifestent par une légère diminution des durées des voyelles adjacents, n'affectent pas l'augmentation de la durée totale la séquence, montrant qu'il y a effectivement une augmentation du nombre de mores d'une quantité à l'autre.

Références

BALS, B.A., D. ODDEN & C. RICE. (2007). *The phonology of gradation in North Saami*. http://www.ling.ohio- state.edu/~odden/SaamiGradation2007.pdf. [Consulté le 27/01/2014]

BALS BAAL, B. A., ODDEN, D., & RICE, C. (2012). An analysis of North Saami gradation. *Phonology*, 29(02), 165-212.

BROSELOW, E., S.-I. CHEN & M. HUFFMAN. (1997) Syllable Weight: Convergence of Phonology and Phonetics. *Phonology* 14, 47-83.

BYE, P., SAGULIN, E., & TOIVONEN, I. (2010). "Phonetic duration, phonological quantity and prosodic structure in Inari Saami". *Phonetica*, 66(4), 199-221.

CHOMSKY, N. & M. HALLE. (1968) *The Sound Pattern of English*. Harper & Row, Publishers, New York.

CLEMENTS, G. N. & S. J. KEYSER. (1983). *CV Phonology. A Generative Theory of the syllable*. Cambridge, Mass. : MIT Press

CLEMENTS, G. N., & RIDOUANE, R. (2006). Part I: Quantal Phonetics and Distinctive Features: a Review. Part II: Distinctive Feature Enhancement: a Review. Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 17-24 and 28-30.

EHALA, M. (2003). Estonian quantity: Implications for moraic theory. *Generative Approaches to Finnic and Saami Linguistics*. *Stanford: CSLI*, *5*, 79.

ENGSTRAND, O. (1987). Durational Patterns of Lule Sami Phonology. *Phonetica 44: 11*. 117-128

GOLDSMITH, J. A. (1976). *Autosegmental phonology* (Vol. 159). Bloomington: Indiana University Linguistics Club.

HALLE, M., & VERGNAUD, J. R. (1980). Three dimensional phonology. *Journal of linguistic research*, *1*(1), 83-105.

HAM, W. (1998). *Phonetic and Phonological Aspects of Geminate Timing*. Ph.D. dissertation, Cornell University.

HAYES, B. (1989). Compensatory lengthening in Moraic Phonology. *Linguistic Inquiry* (The Massachusetts Institute of Technology) 20 (2): 253–306

HUBBARD, K. (1995). Toward a theory of phonological and phonetic timing: Evidence from Bantu. *Phonological and Phonetic Evidence: Papers in Laboratory Phonology IV*. Cambridge: CUP, 168-187.

HYMAN, L. M. (1985). *A theory of phonological weight*. Dordrecht: Foris.

KEYSER, S. & STEVENS, K. (2006). *Enhancement and overlap in the speech chain*. Ms. MIT

LEHISTE, I. *(*1997). Search for phonetic correlates in Estonian Prosody. In I. Lehiste, & J. Ross (Eds.), *Estonian prosody: Papers from a symposium* (pp. 11–33). Tallinn: Institute of Estonian Language.

MORÉN-DUOLLJÁ, B. (2010.) *De Samiske språkene: vakre, unike og uerstattelige*. Bårjås 2010 Kultuvrra ja luonndoárbbe. Àrran: Drag

MCCARTHY, J. and A. PRINCE. (1986). *Prosodic morphology*. Manuscript

MCCARTHY, J. J. (1979). On stress and syllabification. *Linguistic Inquiry*, 10, 443-465.

MCCARTHY, J. J. (1989). Linear order in phonological representation. *Linguistic inquiry*, *20*, 71-99.

SPAHR, C. (2012). Rethinking the morphophonology of Estonian quantity. Nordic Prosody XI

ODDEN, D. (1997). Some theoretical issues in Estonian prosody. In *Estonian prosody: Papers from a symposium* (pp. 165-195).

PORT, R. F., DALBY, J., & O'DELL, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America*, *81*, 1574.

PIERREHUMBERT, J. (1990). Phonological and phonetic representation. *Journal of Phonetics*, *18*(3), 375-394.

RIDOUANE, R. (2010). Geminates at the junction of phonetics and phonology. *Papers in laboratory phonology X*, 61-90.

iPhocomp : calcul automatique de l'indice de complexité phonétique de Jakielski

Hyeran Lee[1] Philippe Gambette[2] Melissa Barkat-Defradas[1, 3]
(1) Praxiling UMR5267 CNRS & Université Montpellier 3, Route de Mende, 34199 Montpellier
(2) LIGM, 5 boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée
(3) Institut des Sciences de l'Evolution UMR5554 CNRS & Université Montpellier 2, Place Eugène Bataillon – C.C. 065 – 34095 Montpellier cedex 05
hyeranleejean@gmail.com, philippe.gambette@u-pem.fr, melissa.barkat@univ-montp3.fr

RESUME

L'indice de complexité phonétique a été proposé par Jakielski afin d'estimer la difficulté à prononcer certains mots en anglais. Des études comparatives sur plusieurs types de pathologies du langage, dans des langues diverses, ont permis d'observer des comportements différents parmi les sujets étudiés, face à des mots d'indices de complexité phonétique variés. Nous proposons une interface web basée sur un dictionnaire phonétique afin d'estimer automatiquement l'indice de complexité phonétique d'un ensemble de mots en français : iPhocomp. Nous comparons l'étiquetage humain à l'étiquetage automatique fourni par iPhocomp afin d'estimer la performance de l'interface.

ABSTRACT

iPhocomp, automatically computing Jakielski's index of phonetic complexity

The index of phonetic complexity was introduced by Jakielski in order to estimate the difficulty to pronounce some words in English. Comparative studies about several language pathologies, in several languages, have observed distinct behavior among the studied subjects facing words with various phonetic complexity indices. We provide a web interface based on a phonetic dictionary to automatically estimate the phonetic complexity index of a set of words in French: iPhocomp. We compare the human computation of this index with the automatic computation provided by iPhocomp in order to estimate the performance of the interface.

MOTS-CLES : indice de complexité phonétique, dictionnaire phonétique, interface web.
KEYWORDS: index of phonetic complexity, phonetic dictionary, web interface.

1 Introduction

Les recherches antérieures ont montré que la complexité phonétique des mots a une influence sur leur planification et le maintien d'une production fluide de la parole, l'augmentation de la complexité nécessitant un plus grand effort pour l'activation et l'organisation des structures phonologiques et la planification motrice (Coalson et *al.*, 2012, Howell, 2004). Contrôler cette variable est ainsi important dans les évaluations langagières ou de la mémoire utilisant des stimuli verbaux et elle peut apporter des éléments explicatifs à l'étude de la performance langagière des locuteurs.

Au cours de ces dernières années, plusieurs mesures de complexité phonétique ont été proposées : *index of phonetic complexity* (Jakielski, 2000), *phonological mean length of utterance* (Ingram et al., 2001), *word complexity measure* (Stoel-Gammon, 2010), etc. Dans cette étude, nous nous intéresserons tout particulièrement à l'*index of phonetic complexity* (IPC) de Jakielski. En effet, l'IPC est une mesure complète qui permet de quantifier la complexité phonétique à travers plusieurs caractéristiques de la production langagière. Ainsi, il nous semblait intéressant d'examiner un large éventail de facteurs afin d'obtenir de plus amples informations. De plus, l'IPC a été utilisé dans divers domaines d'étude comme l'acquisition du langage (Charlier Bererd et *al.*, 2007), les pathologies du langage telle que la dyslexie (Bose et *al.*, 2011), le bégaiement (Howell et *al.*, 2006, Lasalle et *al.*, 2011), l'autisme (Ferré et *al.*, 2013), la maladie d'Alzheimer (Lee, 2012). Cet indice initialement conçu pour l'anglais a été utilisé dans différentes langues, par exemple en allemand (Dworzynski et *al.*, 2004) et en espagnol (Howell et *al.*, 2007).

L'IPC de Jakielski est basé sur la théorie de « cadre puis contenu » de MacNeilage et Davis (1990). Selon cette théorie, le développement vocal précoce est conditionné par la maturation physiologique et la maîtrise des habiletés motrices de l'appareil phonatoire, et ce de façon universelle. En effet, ces auteurs ont observé des caractéristiques régulières dans le babillage des enfants, par exemple l'apparition précoce des consonnes labiales ou coronales par rapport aux consonnes dorsales, ces dernières nécessitant un contrôle plus précis de l'appareil articulatoire. De même, avec la maîtrise de la variation de la forme du conduit vocal, les consonnes fricatives émergent dans le répertoire phonétique des enfants, après les consonnes occlusives. Enfin, avec l'accroissement de la capacité articulatoire, les enfants élargissent leur éventail de production des consonnes isolées aux groupes consonantiques. Ainsi, la théorie de « cadre puis contenu » possède un caractère prédictible.

Jakielski (1998) a observé 8 caractéristiques du babillage des enfants :

1. Les jeunes enfants utilisent moins fréquemment les consonnes dorsales que les enfants plus âgés.
2. Les consonnes occlusives, nasales et semi-consonnes sont récurrentes.
3. Les voyelles rhotiques sont rares.
4. Les syllabes produites ont tendance à être les syllabes ouvertes.
5. Les mots produits sont courts.
6. Les consonnes isolées sont réduplicuées dans le groupe des jeunes enfants, alors que les productions des enfants plus âgés comportent des consonnes de lieu d'articulation varié.
7. Avec l'âge, l'utilisation de groupes consonantiques augmente.

8. Les groupes consonantiques se développent de groupe consonantique homo-organique au groupe consonantique hétéro-organique.

À partir de l'observation de l'apparition précoce ou tardive de ces caractéristiques dans le babillage, Jakielski (2000) a déterminé des éléments simples ou complexes à produire et a identifié 8 paramètres afin de calculer la complexité phonétique.

Paramètres		**Points non attribués**	**Points attribués**
1	Lieu d'articulation des consonnes	Labiales, coronales, glottales	Dorsales
2	Mode d'articulation des consonnes	Occlusives, nasales, semi-consonnes	Fricatives, affriquées, liquides
3	Classe de voyelle	Monophtongues, diphtongues	Rhotiques
4	Final du mot	Se termine par une voyelle	Se termine par une consonne
5	Longueur de mot	Monosyllabique, bisyllabique	≥ 3 syllabes
6	Variation de lieu d'articulation des consonnes isolées	Réduplicquées	Variées
7	Groupe consonantique	Absence	Présence
8	Type de groupe consonantique	Homo-organique	Hétéro-organique

TABLE 1 – 8 paramètres pour calculer la complexité phonétique (Jakielski, 2000)

Dans l'IPC de Jakielski, chaque partie d'un mot composé de l'un des 8 paramètres reçoit un point de complexité. La somme de ces points permet de déterminer la valeur de la complexité d'un mot.

La valeur de complexité phonétique d'un corpus français peut être calculée à partir des règles suivantes (Eldridge, 2006) :

1. Le lieu d'articulation de la consonne : 1 point est attribué à chaque présence d'une consonne dorsale [k], [g], [ʁ], [ŋ].
   Exemple : « agrafeuse » [a**gʁ**aføz] = 2 points.

2. Le mode d'articulation de la consonne : 1 point est attribué pour chaque consonne fricative ou liquide [f], [v], [s], [z], [ʃ], [ʒ], [l], [ʁ].
   Exemple : « calculatrice » [ka**l**ky**l**at**ʁ**i**s**] = 4 points.
3. La classe de voyelle : ce paramètre (i.e. présence de voyelle rhotique) n'étant pas pertinent pour le français, il a été écarté dans cette étude.
4. La fin du mot : la présence d'une consonne en position finale de mot vaut 1 point.
   Exemples : « archet » [aʁ**ʃ**ɛ] = 0 point / « ark » [aʁ**k**] = 1 point.
5. La longueur du mot : un nombre de syllabes supérieur ou égal à 3 correspond à 1 point.
   Exemples : « anneaux » [a-no] = 0 point / « bibliothèque » [bi-bli-jɔ-tɛk] = 1 point.
6. La variation de lieu d'articulation des consonnes isolées : lorsque les consonnes isolées du mot appartiennent à des lieux d'articulation différents (labiales : [p], [b], [m], [f], [v] ; coronales : [t], [d], [n], [l], [s], [z], [ʃ], [ʒ], [ɲ] ; dorsales : [k], [g], [ʁ], [ŋ]), 1 point est octroyé.
   Exemple : « baguette » [**b**a**g**ɛ**t**] = 1 point. Ici, on observe la variation de lieu d'articulation des consonnes isolées de labiale [b] à dorsale [g], puis de dorsale [g] à coronale [t].
   En cas de présence de groupes consonantiques et d'une seule consonne isolée, ce paramètre n'est pas pris en compte.
   Exemple : « empreinte » [ɑ̃**p**ʁɛ̃**t**] = 0 point.
   En cas de présence de groupes consonantiques et de deux consonnes isolées ou plus, la variation de lieu d'articulation est examinée entre les consonnes isolées.
   Exemple : « chronomètre » [**k**ʁɔ**n**ɔ**m**ɛ**t**ʁ] = 1 point. Ici, on observe la variation de lieu d'articulation de coronale [n] à labiale [m] des consonnes isolées.
7. Le groupe consonantique : 1 point est attribué à chaque présence d'un groupe consonantique.
   Les consonnes produites de façon consécutive sont considérées comme un groupe consonantique. Par exemple, dans « volcan » [vɔlkɑ̃], [lk] est considéré comme un groupe consonantique, même si [l] et [k] appartiennent à des syllabes différentes.
   Exemple : « rétroprojecteur » [ʁe**t**ʁo**p**ʁɔʒɛ**k**tœʁ] = 3 points.
8. Enfin, 1 point est attribué à chaque groupe consonantique hétéro-organique, c'est-à-dire impliquant des consonnes articulées en des lieux distincts.
   Exemples : « agrafeuse » [a**g**ʁaføz] = 0 point. Le groupe consonantique [gʁ] est composé de deux consonnes dorsales / « brique » [**b**ʁik] = 1 point. Le lieu d'articulation des consonnes du groupe consonantique [bʁ] varie de labiale [b] à dorsale [ʁ].

La quantification de la valeur de complexité phonétique, obtenue grâce à l'IPC de Jakielski, peut être utilisé dans divers domaines d'étude ou d'application comme l'étude de la maturation du système de la parole chez les enfants, le développement d'une méthode d'enseignement efficace dans le domaine de l'enseignement des langues étrangères en identifiant des éléments faciles ou difficiles à maîtriser, la planification d'une prise en charge orthophonique adaptée dans le domaine des pathologies du langage en discernant les éléments altérés précocement ou préservés en fonction de la complexité phonétique.

Malgré le grand intérêt qu'a suscité cette grille, il n'existait pas, jusqu'à présent, d'outil de traitement automatique de la complexité phonétique. Pourtant, automatiser le processus d'analyse réduirait le temps d'évaluation et limiterait le biais lié à la subjectivité des évaluateurs. Ainsi,

notre objectif est de développer un outil de calcul automatique de la complexité phonétique qui fournit un résultat rapide et fiable, permettant de faciliter la recherche et la mise en pratique de cette mesure.

2 Méthode

Notre méthode iPhocomp de calcul automatique de la complexité phonétique de Jakielski est décrite en Figure 1. Pour chaque mot, une transcription phonétique est obtenue à partir de la base de données Lexique 3 (New et *al.*, 2001). Les règles décrites plus haut sont alors appliquées sur cette transcription pour évaluer 8 paramètres quantitatifs. Ces paramètres sont additionnés pour obtenir la valeur de l'indice de complexité phonétique. Pour les mots dont la transcription phonétique n'a pas été trouvée dans la base de données, l'indice de complexité phonétique n'est pas évalué, mais une proposition de transcription phonétique est donnée. L'utilisateur pourra alors la corriger, et l'ajouter à un fichier personnel de transcriptions qu'il pourra utiliser en relançant le calcul.

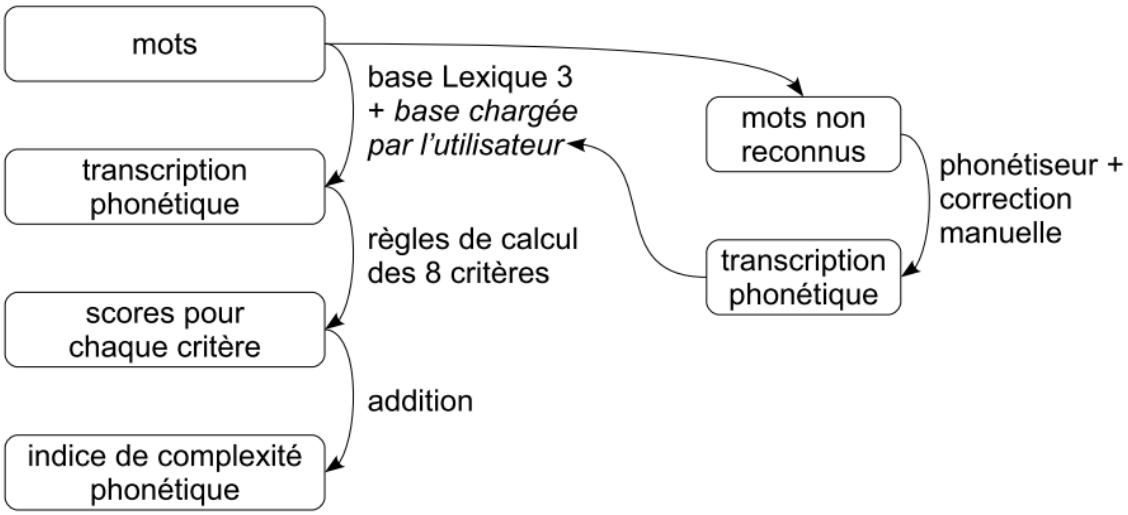


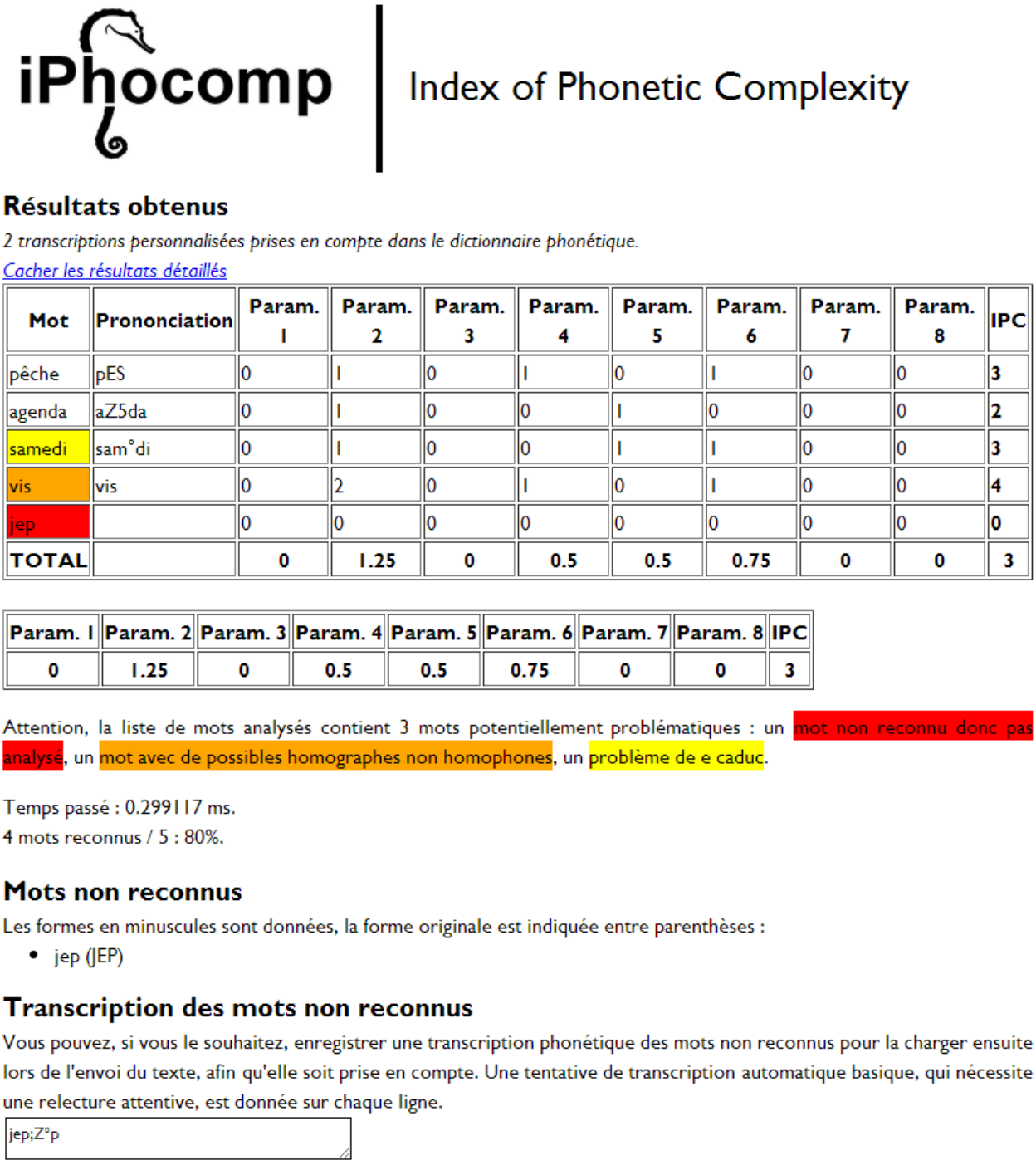
FIGURE 1 – Schéma descriptif de l'approche automatique implémentée dans iPhocomp

Cette méthode a été programmée sous forme d'une interface web à l'aide du langage PHP. Elle est disponible à l'adresse http://igm.univ-mlv.fr/~gambette/iPhocomp/. Le code source du logiciel est également téléchargeable à la même adresse, et réutilisable selon la licence GPLv3 afin de favoriser l'adaptation de cet outil à d'autres langues.

L'ensemble des mots à analyser est à fournir sous forme d'une simple liste de mots (un mot par ligne) dans un cadre de texte, sur la page web d'iPhocomp. L'utilisateur peut également charger une liste personnalisée de transcriptions phonétiques, qui seront prioritaires par rapport aux 125 653 transcriptions prédéfinies dans la base de données du logiciel. Ceci permet d'inclure des mots absents de notre base de données, en particulier pour des études sur les pseudo-mots, ainsi que sur les éventuelles variations de prononciation dues à des facteurs géographiques, stylistiques ou pathologiques.

L'ensemble des règles d'évaluation des 8 paramètres qui composent l'indice de complexité phonétique d'un mot, détaillées dans la Table 1, peut se déduire de sa transcription phonétique. Ainsi, nous avons choisi d'utiliser une ressource librement disponible afin de disposer de transcriptions phonétiques : la base de données Lexique 3 (New et *al.*, 2001). Lexique 3 contient

plusieurs homographes, homophones ou non. Les homographes homophones (comme « abats », qui désigne la deuxième personne du singulier du verbe « abattre » à l'indicatif, mais aussi le nom « abat » au pluriel) ont été traités comme des doublons et retirés. La base de données contient également 80 homographes non homophones, incluant par exemple les noms « actions » et « vis », qui se prononcent différemment des verbes conjugués « actions » et « vis ». Pour ces mots, seule la transcription phonétique la plus fréquente, selon la fréquence d'apparition à l'oral, également fournie par la base Lexique 3 (i.e. fréquence élaborée à partir de la transcription de dialogues de films), a été conservée. Elle a été marquée afin de pouvoir signaler ces mots comme potentiellement mal transcrits lors de l'affichage du résultat final, en couleur orange (voir « vis » en Figure 2).



Mot	Prononciation	Param. 1	Param. 2	Param. 3	Param. 4	Param. 5	Param. 6	Param. 7	Param. 8	IPC
pêche	pES	0	1	0	1	0	1	0	0	3
agenda	aZ5da	0	1	0	0	1	0	0	0	2
samedi	sam°di	0	1	0	0	1	1	0	0	3
vis	vis	0	2	0	1	0	1	0	0	4
jep		0	0	0	0	0	0	0	0	0
TOTAL		0	1.25	0	0.5	0.5	0.75	0	0	3

Param. 1	Param. 2	Param. 3	Param. 4	Param. 5	Param. 6	Param. 7	Param. 8	IPC
0	1.25	0	0.5	0.5	0.75	0	0	3

FIGURE 2 – Interface d'iPhocomp

La base Lexique 3 propose également un traitement des « e » caducs distinct d'autres bases de données comme le TLFI (e.g. « batterie » est transcrit en [batʀi] selon http://www.cnrtl.fr/definition/batterie), en suggérant qu'ils sont prononcés, à l'aide du symbole « ˚ ». Les mots contenant des « e » caducs sont identifiés par iPhocomp à l'aide d'une règle simple (présence d'un « ˚ » dans la transcription phonétique, précédé et suivi par une seule consonne) et signalés par une couleur jaune (voir « samedi » en Figure 2).

Enfin, les mots pour lesquels il n'existe pas de transcription phonétique dans la base Lexique 3 ne sont pas analysés. Ils sont colorés en rouge et une transcription phonétique automatique basique est fournie (voir « JEP » en Figure 2). L'utilisateur est ainsi invité à vérifier soigneusement les transcriptions phonétiques des mots colorés, et éventuellement à relancer l'analyse en fournissant des transcriptions phonétiques personnalisées des mots mal transcrits.

iPhocomp fournit la valeur de complexité phonétique des mots, mais aussi les résultats détaillés pour chacun de ces 8 paramètres (Figure 2). Cela ouvre des perspectives d'analyses plus qualitatives et permet d'adapter l'IPC en fonction des corpus traités, par exemple en attribuant des poids différents aux 8 paramètres.
En plus des valeurs pour chaque mot, iPhocomp peut également calculer la valeur moyenne de la complexité phonétique d'une liste de mots. Cela permet de mener des analyses comparatives en évaluant ces divers paramètres sur deux listes de mots distinctes.

3 Résultats

Afin d'estimer la validité d'iPhocomp, nous avons procédé à une étude comparative de l'analyse manuelle et de l'analyse automatique.

Un des auteurs de cet article a effectué l'analyse manuelle d'IPC sur la base de données de Bonin et *al.* (2003). Cette base comporte un ensemble de 299 images associées à une dénomination de référence en français, ainsi que divers critères quantitatifs comme les valeurs d'accord nom-image, de familiarité conceptuelle, d'âge d'acquisition, de fréquence, de neutralité culturelle, etc. (Lee et *al.*, 2013).

Les résultats obtenus par l'analyse automatique et l'analyse manuelle ont alors été comparés. Cela a permis de corriger des erreurs initiales de conception du logiciel, par exemple sur les mots composés. Nous avons également pu repérer des erreurs dans l'analyse manuelle : sur le paramètre 2 pour « cafetière » et « scaphandrier », le paramètre 5 pour « oreiller », le paramètre 6 pour « chronomètre » et « portefeuille », et le paramètre 8 pour « flocon », « jeep » et « moustique ». Comme il n'y a pas d'ambiguïté de codage, toutes les erreurs manuelles relevaient d'erreurs d'inattention. Le calcul manuel est donc sujet à des erreurs, même s'il s'agit d'un taux très réduit (8 valeurs paramètres sur l'ensemble des 2093 valeurs de paramètres évaluées, soit moins de 0.4 % d'erreur).

Un second test a été effectué afin d'évaluer le temps d'analyse gagné sur un corpus correspondant à un cas d'usage : 118 mots extraits du discours de 13 patients atteints de la maladie d'Alzheimer, ainsi que le même nombre de mots extraits du discours d'individus sains. En utilisant la transcription phonétique fournie par Lexique 3, nous avons constaté que le temps de calcul humain des 8 paramètres se montait à environ 2 heures et demie pour ces 236 mots,

alors que le calcul automatique avec iPhocomp sur le même corpus prenait 0.54 ms. Il est à noter que si une intervention manuelle est nécessaire, en raison des types de corpus ou des choix de transcription décrits dans la partie 2, cela peut engendrer un temps de correction supplémentaire. On peut prévoir 10 secondes supplémentaires par mot à corriger puisqu'il ne s'agit pas de faire une transcription complète mais une simple modification de celle-ci.

4 Conclusion et perspectives

iPhocomp permet de réduire significativement le temps total d'évaluation de l'indice de complexité phonétique, par rapport à l'évaluation manuelle. Il permet aussi d'éviter l'erreur humaine due à l'aspect répétitif de la tâche ou à une mauvaise application des règles de calcul des paramètres, optimisant ainsi la comparabilité des résultats de différentes études.

Il nous semble important d'approfondir l'étude de la complexité phonétique de la langue française (e.g. paramètre 3). Nous pensons que l'intégration des aspects phonologiques (Pellegrino, 2009) et l'étude de divers phénomènes observables (e.g. l'ordre d'émergence des segments, la maîtrise de certaines unités, les difficultés de production, etc.) dans l'acquisition du langage, dans l'apprentissage de langues étrangères et dans les pathologies du langage en français, permettraient de dégager des éléments pertinents pour enrichir nos connaissances et notre compréhension de la complexité phonétique, permettant de mieux adapter l'indice de complexité phonétique de Jakielski aux spécificités de la langue française.

iPhocomp peut être utilisé pour toute langue disposant des ressources linguistiques adaptées. Ainsi, il nous semble intéressant d'étendre cette interface aux autres langues afin de faciliter les études translinguistiques.

Le logiciel iPhocomp est librement mis à la disposition de la communauté scientifique. Nous espérons que ce programme donnera un nouvel élan d'une part aux recherches fondamentales pour mieux appréhender les propriétés physiques et structurelles des unités sonores du langage et d'autre part aux études appliquées telles que celles conduites dans le champ de la phonétique clinique.

Références

BONIN, P., PEERELMAN, R., MALARDIER, N., MEOT, A., & CHALARD, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, 35(1):158–167.

BOSE, A., COLANGELO, A., & BUCHANAN, L. (2011). Effect of phonetic complexity on word reading and repetition in deep dyslexia. *Journal of Neurolinguistics*, 24:435–444.

CHARLIER BERERD, M., & JUHEM, A. (2007). *Evolution de la complexité phonético-phonologique et sélection lexicale chez des enfants français entre 12 et 27 mois*. Mémoire d'orthophonie, Université Lyon 1.

COALSON, G., BYRD, C., & DAVIS, B. (2012). The influence of phonetic complexity on stuttered speech. *Clinical linguistics & phonetics*, 26(7):646–659.

DWORZYNSKI, K., & HOWELL, P. (2004). Predicting stuttering from phonetic complexity in German. *Journal of fluency disorders*, 29:149–173.

ELDRIDGE, K. (2006). *Phonological complexity and speech disfluency in young children*. Thèse de doctorat, University of Pittsburgh.

FERRE, S., DOS SANTOS, C., & BONNET-BRILHAULT, F. (2013). Phonological delay or phonological impairment in autism: an intergroup comparison. *International meeting for autism research*, San Sebastian, Espagne.

HOWELL, P. (2004). Assessment of some contemporary theories of stuttering that apply to spontaneous speech. *Contemporary issues in communication science and disorders*, 31:123–140.

HOWELL, P., & AU-YEUNG, J. (2007). Phonetic complexity and stuttering in Spanish. *Clinical linguistics & phonetics*, 21(2):111–127.

HOWELL, P., AU-YEUNG, J., SCOTT YARUSS, J., & ELDRIDGE, K. (2006). Phonetic difficulty and stuttering in English. *Clinical linguistics & phonetics*, 20(9):703–716.

INGRAM, D., & INGRAM, K. (2001). A whole-word approach to phonological analysis and intervention. *Language, speech, and hearing services in schools*, 32:271–283.

JAKIELSKI, K. (1998). Motor organization in the acquisition of consonant clusters. Thèse de doctorat, University of Texas at Austin.

JAKIELSKI, K. (2000). Quantifying phonetic complexity in words: An experimental index. *Child phonology conference*, Cedar Fallas, IA.

LASALLE, L., & WOLK, L. (2011). Stuttering, cluttering, and phonological complexity: case studies. *Journal of fluency disorders*, 36:285–289.

LEE, H. (2012). *Langage et Maladie d'Alzheimer : Analyse multidimensionnelle d'un discours pathologique*. Thèse de doctorat, Université Montpellier 3.

LEE, H., GAMBETTE, P., GAYRAUD, F., & BARKAT-DEFRADAS, M. (2013). Élaboration d'un outil d'évaluation des performances en dénomination pour les patients bilingues atteints de la maladie d'Alzheimer. *Rééducation orthophonique*, 253:143–152.

MACNEILAGE, P., & DAVIS, B. (1990). Acquisition of speech production: Frames, then content. In M. Jeannerod (Ed.), *Attention and performance XIII: Motor representation and control*. Hillsdale: Lawrence Erlbaum.

NEW, B., PALLIER, C., FERRAND, L., & MATOS, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101:447–462. www.lexique.org

PELLEGRINO, F. (2009). *De l'identification des langues à la complexité phonologique*. Habilitation à diriger des recherches, Université Lumière Lyon 2.

STOEL-GAMMON, C. (2010). The word complexity measure: description and application to developmental phonologic and disorders. *Clinical linguistics & phonetics*, 24:271–282.

Influence du développement phonético-phonologique sur les premières productions lexicales d'enfants arabophones : une étude longitudinale

Hacène Bellemmouche[1] Frédérique Gayraud[2]
Rim Hamdi-Sultan[3] Melissa Barkat-Defradas[1, 4]

(1) Praxiling, UMR5267 CNRS &Université Montpellier 3
(2) Dynamique du Langage UMR 5596 CNRS & Université Lyon 2
(3) Savoirs, Textes, Langage UMR 8163 CNRS &Université Lille 3
(4) Institut des Sciences de l'Évolution UMR 5554 CNRS & Université Montpellier 2
hacene.bellemmouche@univ-montp3.fr, frederique.gayraud@univ lyon2.fr, rim.hamdi@univ-lille3.fr, melissa.barkat@univ-montp3.fr

RÉSUMÉ

Ce travail examine l'évolution de la complexité phonétique des premiers mots d'enfants arabophones en utilisant l'index de complexité phonétique de Jakielski (2000). Dans cette étude, 4 enfants (8 à 24 mois) ont été enregistrés en situation de communication naturelle. Les résultats indiquent que la complexité des productions enfantines augmente avec l'âge et la taille du vocabulaire. Nous observons qu'au cours du développement, les enfants produisent des mots de plus en plus complexes, mais dont la complexité phonétique est toujours inférieure à celle des mots cibles. Les premiers mots procèdent ainsi d'une forme de « sélection lexicale » opérée en fonction de la maturité phonatoire. De plus, les mères produisent des mots dont la forme sonore est simplifiée afin de s'adapter aux capacités motrices de leur enfant.

ABSTRACT

Influence of phonetic-phonological development on first lexical productions of Arabic-speaking children

The aim of this longitudinal study is to measure the evolution of the Index of Phonetic Complexity (Jakielski, 2000) in native arabic infants' first lexical productions. To answer this question, 4 native young speakers of Tunisian Dialectal Arabic have been recorded at home every other week during one hour while speaking with their mother. Our preliminary results show that the IPC increases significantly with vocabulary size. Moreover, we showed that although children's IPC increases with time, its value is always inferior to that of target words. We propose that childrens' realizations are the result of somekind of lexical selectivity which is determined by their phonatory maturity (i.e. words composed of already attested segments in their phonetic inventory). In addition our results reveal that mothers adapt their language to their childrens' abilities.

MOTS-CLÉS : Acquisition, Complexité phonétique, Développement lexical, Arabe.
KEYWORDS: Acquisition, Phonetic complexity, Lexical development, Arabic.

1 Introduction

Le lexique productif précoce de l'enfant est souvent décrit comme quantitativement et qualitativement très différent du lexique de l'adulte (Fenson et *al.,* 1993). Entre 1 et 3 ans, celui-ci est relativement limité et les premières réalisations produites par les enfants sont souvent très éloignées du mot cible. Plusieurs facteurs peuvent être invoqués pour expliquer ces différences parmi lesquels, l'évolution de la maturité phonatoire (Bornstein et *al.*, 2004). Bien que l'étude du développement du lexique précoce ait donné lieu à de nombreux travaux dans les langues de grande diffusion telles que l'anglais ou le français (Jakielski, 2000 ; Matyasse, 2005 ; Pellegrino et *al.*, 2009), les études en acquisition portant sur l'arabe dialectal –langue maternelle des locuteurs arabophones– sont encore peu fréquentes (Badri-Zalami, 2007) et, quand elles existent, elles concernent soit l'apprentissage tardif et formel de l'arabe standard ; soit l'acquisition de variétés d'arabe dialectal largement documentées dans la littérature comme l'arabe dialectal égyptien (Omar, 1973) ou jordanien (Ameyrah, 2003).

Nous proposons dans ce travail de contribuer au développement des études acquisitionnelles en arabe dialectal en nous intéressant plus particulièrement au cas de l'arabe maghrébin et notamment, de l'arabe dialectal tunisien (i.e. parler de Tunis). Cette étude longitudinale[1] a pour objectif l'examen des premières productions enfantines considérées du point de vue de leur complexité phonétique. Nous avons appliqué aux données verbales l'indice de complexité phonétique (ICP), échelle initialement développée pour l'anglais par Jakielski (2000). Nous entendons ainsi inférer la maturité phonatoire de nos jeunes locuteurs à travers la mesure objective de la complexité phonétique de leurs premières productions.

1.1 Système phonologique du parler de Tunis

L'arabe dialectal tunisien est, par nature, une langue « mixte » résultant du contact de diverses variétés linguistiques. En effet, tout comme les dialectes algérien et marocain, il s'établit sur un substrat berbère, certes fortement influencé par l'arabe, mais aussi par les langues des peuples ayant vécu ou administré cette région au cours de l'histoire, dont le turc, l'italien, l'espagnol et le français. Du point de vue descriptif, le tunisois (i.e. *tounsi*) est un dialecte urbain au même titre que celui de Kairouan, de Bizerte, de Nabeul ou de Sousse (Versteegh, 2001). On y retrouve certaines des réalisations typiques des parlers de sédentaires (i.e. maintien des fricatives interdentales et prononciation sourde de l'occlusive uvulaire). Le système consonantique du tunisois est présenté dans le tableau 1, où les symboles entre parenthèses correspondent aux réalisations pour lesquelles il n'y a pas de consensus quant à leur statut de phonèmes ou d'allophone.

[1]Ce travail s'inscrit dans le cadre du projet ANR « *Influence du développement phonologique et de l'input sur les premiers mots* » (AA-PPPP-000/2012-15, coordinatrice : Sophie KERN DDL-UMR5596 Lyon).

labial	coronal			dorsal			
bilabial	dental	alvéolaire	palato-alvéolaire	vélaire	uvulaire	pharyngal	glottal
(p)		t tˤ		k	q		ʔ
b (bˤ)		d		g			
m (mˤ)		n (nˤ)					
f	θ	s sˤ	ʃ		x	ħ	h
(v)	ð	z (zˤ)	ʒ		ɣ	ʕ	
	ðˤ						
		l (lˤ)					
w		r (rˤ)	j				

TABLE 1 – Système consonantique du tunisois moderne (d'après Gibson, 2009)

Le système vocalique de l'arabe classique se présente, en tunisois, dans un état de « relative conservation » (Marçais, 1977:13). On y retrouve en effet la triade [i] [a] [u] de l'arabe standard, chacun de ces timbres de base étant doublé d'une corrélation de quantité (on n'atteste cependant pas d'opposition de durée en position finale absolue, sauf dans les monosyllabiques). On constate par ailleurs l'absence des diphongues classiques [aj]et [aw] respectivement réalisées [iː] et [uː] (e.g. [lajl] > [liːl] « nuit ; [mawt] > [muːt] « mort »). Enfin, considérant l'effet de la pharyngalisation sur le timbre des voyelles, on constate qu'en présence d'une consonne pharyngalisée, les voyelles cardinales subissent des variations qualitatives importantes tant au niveau de l'aperture que du trait de postériorisation (Barkat-Defradas, 2009a). Cet état de fait se manifeste également en tunisois où, dans l'environnement d'une consonne emphatique : [i] > [ɛ] ; [u] > [ɔ] et [a] > [ɑ]. Enfin, notons que le phénomène d'imāla (i.e. hausse de la voyelle centrale vers [æ] voire [e] y est particulièrement marqué en contexte neutre, notamment dans les monosyllabes à voyelle finale (e.g. [sməːʒ] « ciel » ; [ʒəːʒ] « il est venu »).

1.2 L'acquisition de l'arabe langue maternelle : aspects phonético-phonologiques

Les quelques études consacrées à l'étude de l'acquisition phonologique des consonnes de l'arabe révèlent cinq stades du développement, tel qu'illustré dans le tableau 2. Il est par ailleurs intéressant de remarquer l'émergence précoce de l'occlusive glottale et de la liquide /l/ laquelle est attribuée à la fois à la fréquence élevée et à la charge fonctionnelle importante de ces sons en langue (Amrayeh, 2003 ; Amrayeh & Dyson, 2000). On constate également que l'acquisition des pharyngalisées est particulièrement tardive, ceci peut s'expliquer par la complexité articulatoire dont elles procèdent. Une interprétation alternative pose que la pharyngalisation serait traitée comme un trait suprasegmental. L'argument majeur en faveur de cette hypothèse repose sur le fait que les consonnes neutres sont acquises les unes après les autres de façon séquentielle alors que les différents segments pharyngalisés apparaissent simultanément au cours du développement, comme si l'enfant surimposait aux articulations principales antérieurement acquises et maîtrisées le trait de pharyngalisation (Eldine, 1989).

	babillage	14 - 24 mois	2 – 3 ; 1 ans	4 – 6; 4 ans	6 ; 5 – 8 ans
occlusives	b	d, t, ʔ	k, q		tˁ, dˁ
fricatives	h	ʃ, ʕ, ħ	f	s,x,ð,ɣ,θ,sˁ	ðˁ
nasales/liquides	m	n, l		r	
approximantes	w, j				

TABLE 2 – Stades d'acquisition des phonèmes consonantiques de l'arabe d'après Omar (1973) et Amayreh & Dyson (2000)

L'objectif de ce travail est d'examiner l'évolution de la complexité phonétique des premières productions lexicales d'enfants arabophones d'origine tunisienne en utilisant l'index de complexité phonétique de Jakielski (2000) afin d'évaluer l'influence de la maturité phonatoire – mesurée à travers l'indice de complexité phonétique – sur les premières productions lexicales enfantines. Nous chercherons également à voir dans quelle mesure la complexité phonétique du langage adressé à l'enfant contribue au processus d'acquisition.

2 Méthode

2.1 Corpus

Notre corpus est composé d'interactions spontanées entre mères et enfants (n = 4), âgés de 8 à 24 mois. Les données ont été acquises en situation de communication écologique, les enregistrements audio vidéos ayant été réalisés au domicile des familles tous les 15 jours pendant une heure. Les enregistrements ont démarré à l'émergence du premier mot (entre 8 et 15 mois) et se sont poursuivis jusqu'à la production, par l'enfant, de 50 mots différents– mots produits soit spontanément, soit par imitation par rapport au modèle maternel, i.e. mots-cibles – (n = 50*4). Nous avons également transcrit 50 mots produits par les mères s'adressant à leur enfant afin de comparer les productions enfantines à celles de l'adulte (n = 50*4). Ceci nous a permis d'évaluer l'influence de la maturité phonatoire sur les premiers mots. L'ensemble des données a été transcrit manuellement en API par un locuteur natif d'arabe tunisien[2] à l'aide du programme Phon 1.6.2[3].

2.2 Mesure de complexité phonétique

L'indice de complexité phonétique (désormais ICP) est une échelle de mesure qui permet de quantifier la complexité phonétique à travers différents paramètres articulatoires. Initialement développée pour étudier l'acquisition phonologique en anglais (Jakielski, 2000), elle a également été utilisée pour différentes langues comme l'allemand (Dworzynski et *al.*, 2004) ou l'espagnol (Howell et *al.*, 2007). L'ICP est basé sur la théorie du « cadre puis contenu » (MacNeilage et Davis, 1990) selon laquelle le

[2] Les auteurs tiennent à remercier Inès Ghazoueni pour sa contribution au traitement des données.

[3] http://childes.psy.cmu.edu/phon/

développement phonatoire précoce est conditionné par la maturation physiologique et la maîtrise de l'appareil phonatoire. C'est ainsi qu'il est possible d'expliquer que, de façon quasi-universelle, les consonnes labiales et coronales émergent plus tôt que les consonnes dorsales (ces dernières nécessitant un contrôle plus précis de l'appareil phonatoire), et que les fricatives (dont l'articulation nécessite la maîtrise de la variation du conduit vocal) apparaissent après les occlusives. A l'origine, l'ICP a été construit sur la base de 8 paramètres lesquels ont été déterminés à partir du caractère « simple » vs « complexe » à acquérir, étant entendu que les éléments simples à acquérir seront acquis plus précocement que les éléments complexes. Le tableau 3 reprend chacun de ces paramètres en indiquant – pour chaque catégorie– le nombre de points à attribuer.

Paramètres		0 point	1 point / observation
1	Lieu d'articulation des consonnes	Labiales, coronales, glottales	Dorsales
2	Mode d'articulation des consonnes	Occlusives, nasales, semi-consonnes	Fricatives, affriquées, liquides
3	Classe de voyelle	Monophtongues, diphtongues	Rhotiques
4	Finale du mot	Se termine par une voyelle	Se termine par une consonne
5	Longueur de mot	Monosyllabique, bisyllabique	≥ 3 syllabes
6	Variation de lieu d'articulation des consonnes isolées	Réduplicquées	Variées
7	Groupe consonantique	Absence	Présence
8	Type de groupe consonantique	Homo-organique	Hétéro-organique

TABLE 3 – Index de complexité phonétique de Jakielski (2000)

Dans le cadre de ce travail, nous avons été amenés à adapter l'ICP aux spécificités de l'arabe. C'est pourquoi seuls 7 des 8 paramètres d'origine ont été retenus, le critère relatif aux voyelles rhotiques ayant été écarté dans la mesure où ce phénomène n'est pas attesté en arabe. L'indice de complexité phonétique a ensuite été appliqué à nos deux ensembles de données, soit (i) aux mots-cibles extraits du langage adressé à l'enfant et (ii) aux productions enfantines (spontanées et/ou produites par imitation de la cible). Nous avons ensuite calculé l'indice de complexité phonétique moyen en divisant le nombre total de points obtenus dans chaque ensemble de données par le nombre total de mots analysés. Par ailleurs, afin d'étudier l'évolution de la complexité phonatoire en fonction de l'âge, nous avons déterminé *a priori* 3 stades de développement : (i) 8-15 mois; (ii) 16-19 mois; (iii) 20-24 mois. Ce découpage s'inspire des différentes périodes de développement langagier proposées par Boysson-Bardies (1996).

3 Résultats et discussions

3.1 Taille du vocabulaire et complexité phonétique

La figure 1 montre que la taille du vocabulaire et la complexité phonétique moyennes augmentent de façon significative avec l'âge ($F_{(2,1)=}$48,46; p=.0002) et et $F_{(3,1)}$=14,65 ; p=0,004 respectivement). En outre, il apparaît que la taille du vocabulaire et la complexité phonétique sont positivement corrélées (r_s=0,79 ; p=0.002).

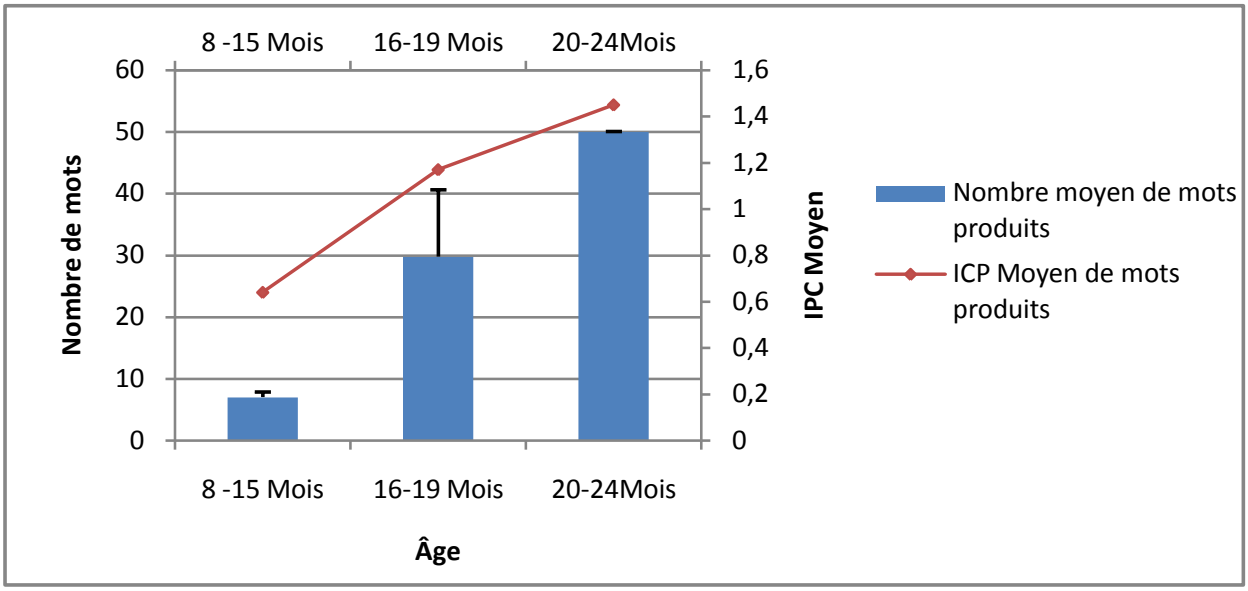


FIGURE 1 – Évolution de la taille du vocabulaire et de l'ICP en fonction du stade de développement

(Les barres de dispersion correspondent aux différences individuelles observées sur chaque groupe d'âge)

D'autre part, nous observons qu'au fil du temps, les enfants produisent des mots certes, de plus en plus complexes, mais dont la complexité phonétique est toujours inférieure à celle des mots-cibles (figure 2).

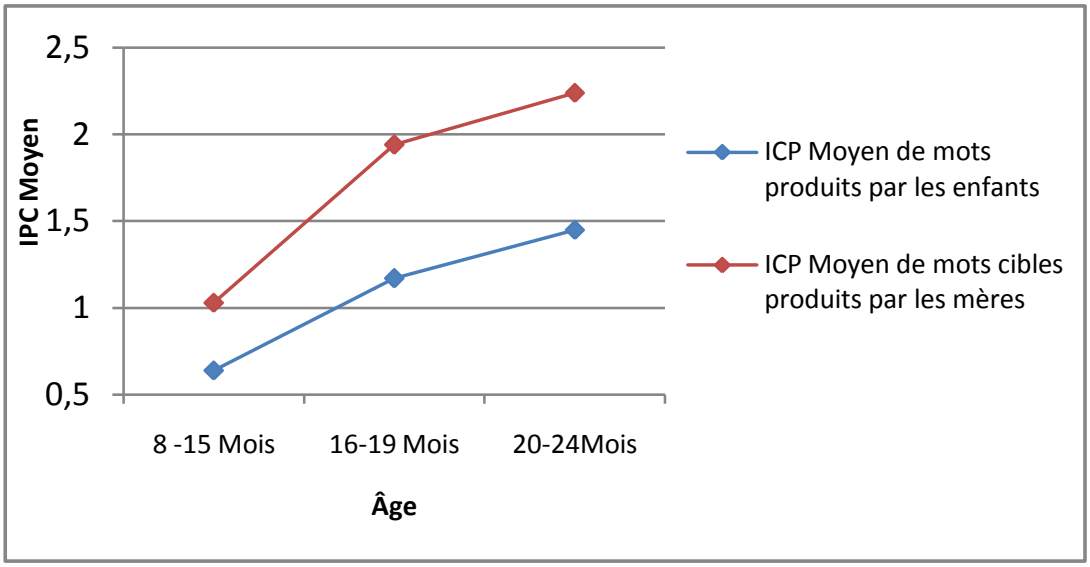


FIGURE 2 – Évolution de l'IPC des mots produits vs des mots-cibles en fonction de l'âge

Une explication possible à cette observation est que les réalisations enfantines procèdent d'une forme de « sélection lexicale » que les enfants semblent opérer en fonction de leur maturité phonatoire (i.e. productions de mots articulatoirement « simples » composés de sons déjà présents dans leurs répertoire phonétique à la période *T*) et, d'autre part, que les mères s'adaptent aux capacités de leur enfant en produisant des mots de plus en plus complexes au cours du développement. En effet la figure 3 montre que la proportion de mots articulatoirement très simples (i.e. indice de complexité = 0) diminue au cours du temps au profit de l'introduction dans le langage des mères adressé à l'enfant, de formes phonétiquement plus complexes.

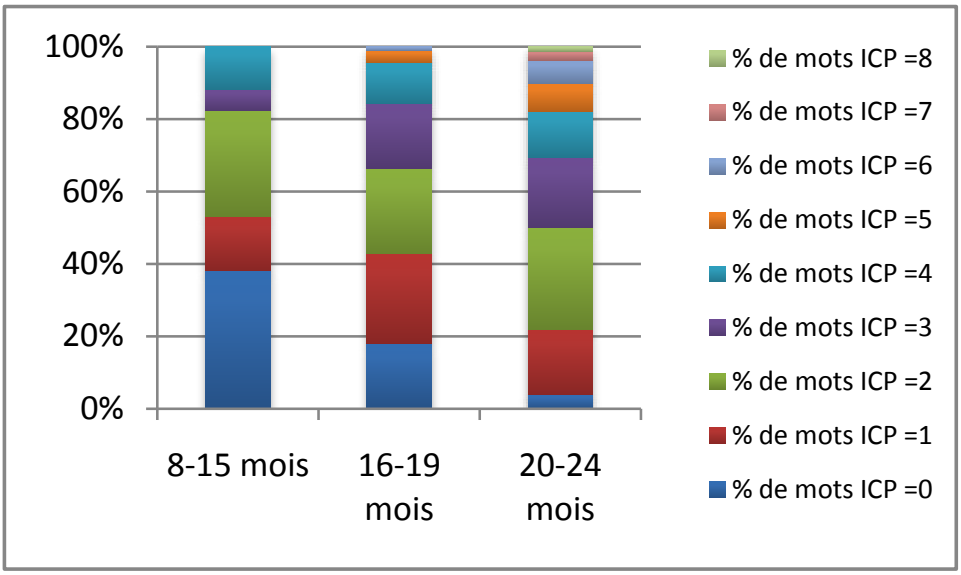


FIGURE 3 – Proportion de mots +/– complexes produits par les mères en fonction du stade de développement

Un examen plus approfondi des différents paramètres de l'ICP chez l'enfant et chez l'adulte (cf. Figure 4) révèle que certains critères montrent des écarts particulièrement importants entre les productions maternelles et les productions enfantines ($F_{(1,2)} = 6.08$; $p = 0.04$). On observe également un effet du paramètre (notamment du mode d'articulation, et dans une moindre mesure, de la variation consonantique et des clusters) ($F_{(6,1)} = 87{,}35$; $p < 0.0001$).

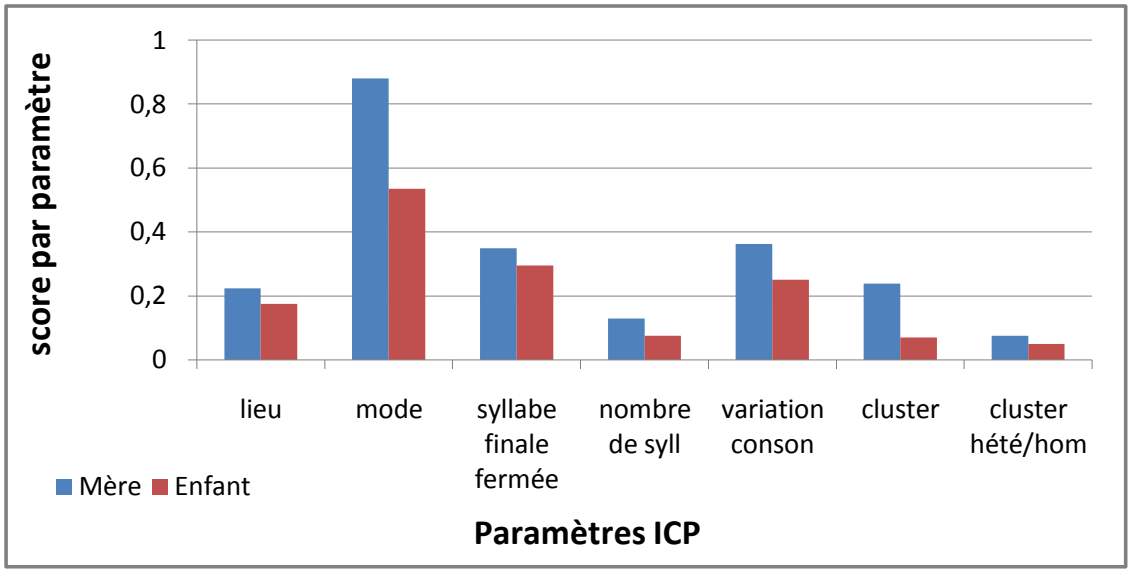


FIGURE 4 – Scores par paramètre d'IPC chez l'enfant et chez l'adulte

Le tableau 4 présente l'ordre d'acquisition des consonnes de l'arabe dialectal tunisien chez nos sujets.

	8 - 15 mois	16 - 19 mois	20 - 24 ans	consonnes non acquises
occlusives	b, t, d	ʔ, k, q		tˤ
fricatives/affriquées	h	ʃ, ʕ, ħ, s	z, f, χ	θ, ð, ðˤ, ʒ, ɣ, sˤ
nasales/liquides	m	n, l		r
approximant	w, j			

TABLEAU 4 – Ordre d'acquisition des consonnes en arabe tunisien

Dans la première période, on observe la production d'occlusives, d'approximantes et de la nasale bilabiale ; les fricatives et les liquides apparaissant plus tardivement au cours du deuxième et troisième stade. On observe qu'à 24 mois, plusieurs sons consonantiques – pour la plupart fricatifs – restent à acquérir. On note également qu'aucune consonne pharyngalisée n'est attestée dans les productions précoces. La réalisation des consonnes à double articulation linguale (i.e. pharyngalisées) requiert en effet, outre une articulation principale – apicale ou laminale –, une articulation secondaire nécessitant le recul de la racine de la langue vers la paroi postérieure du pharynx avec, par voie de conséquence, une dépression de son dos en plateau (Barkat-Defradas et Embarki, 2009). Cette articulation nécessite un contrôle moteur particulièrement fin, ce qui explique qu'elles émergent très tardivement. D'une façon générale, ces résultats s'avèrent être conformes à ceux de la littérature.

4 Conclusion

L'étude de l'Indice de Complexité Phonétique que nous avons conduite a montré que cet indice augmente au cours des trois périodes examinées, et ce aussi bien dans les productions enfantines ce qui traduit, entre autres, l'évolution de leurs capacités motrices. Nos résultats révèlent par ailleurs, que l'ICP des adultes augmente de façon significative au cours du temps, ce qui semble témoigner d'une adaptation des mères aux capacités de l'enfant. Par ailleurs, en dépit des spécificités de l'arabe tunisien, les productions lexicales précoces des quatre enfants de notre corpus semblent fondamentalement façonnées par les contraintes biomécaniques universelles décrites dans la théorie du cadre puis du contenu. En effet, comme on a pu l'observer dans de nombreuses langues, les consonnes qui nécessitent un contrôle moteur fin apparaissent plus tardivement que les autres. Afin de mieux identifier l'influence des spécificités de la langue maternelle sur l'ordre d'acquisition, nous envisageons à l'avenir de conduire des comparaisons translinguistiques, seules à même de nous informer dans quelle mesure l'acquisition est conditionnée par des contraintes universelles ou par des spécificités de la langue en cours d'acquisition.

Références

AMAYREH, M. et DYSON, A. (2000). Phonetic inventories of young Arabic-speaking children. *Clinical linguistics & phonetics,* 14 (3): 193–215.

AMAYREH, M. (2003). Completion of the Consonant Inventory of Arabic. *Journal of Speech, Language and Hearing Research* 46 (3): 517-529.

BARKAT-DEFRADAS, M. (2009a). "Vowel backing & vowel lowering in Arabic vernaculars". *Encyclopaedia of Arabic Language & Linguistics*, Vol. 4, K. Versteegh Ed., Brill, Leiden, The Netherlands, p. 669-674

BARKAT-DEFRADAS, M. et EMBARKI, M., (2009). « Mécanismes articulatoires et corrélats acoustiques de l'emphase en arabe et en berbère : un état de l'art », in Etudes de phonétique et linguistique berbères Hommage à Naïma Louali (1961-2005), S. Chaker & A. Mettouchi & G. Philipson Eds. Ussun amaziɣ - Editions Peeters, Paris/Louvain. pp. 21-42.

BORNSTEIN, M. et *al.* (2004). Cross-linguistic Analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development,* 75 (4): 1115–1139.

DAVIS, B. et MacNeilage, P.F. (1990). The acquisition of vowels: a case study. Journal of Speech and Hearing Research ,33, 16-27

DE BOYSSON BARDIES, B. (1996). *Comment la parole vient aux enfants*. Odile JACOB.

FENSON, L. et *al.* (1993). MacArthur Communicative Development Inventories: user's guide and technical manual. San Diego, CA: Singular Press. 188 p.

GIBSON, M.L., (2009). Tunis Arabic, in *Encyclopaedia of Arabic Language & Linguistics*, Vol. 4, K. Versteegh Ed., Brill, Leiden, The Netherlands, p. 563-570.

JAKIELSKI, K.J. (2000). Quantifying phonetic complexity in words: An experimental index. *Paper presented at the Annual child Phonology Conference*, Cedar Falls, IA.Jescheniak & Levelt, 1994.

MARÇAIS, Ph. (1977). Esquisse grammaticale de l'arabe maghrébin, Paris: Librairie Adrien-Maisonneuve.

MATYASSE, R. (2005). Analyzing the acquisition of phonetic complexity in one-year-old children. Senior Thesis, Augustana College, Rock Island, IL.

OMAR, M.K (1973). The acquisition of Egyptian Arabic as a native language. GUP/Washington, DC.

PELLEGRINO, F. et *al.* (2009). Approaches to phonological complexity. Phnonology & phonetics, 16. Mouton de Gruyter : Berlin. New York.

VERSTEEGH, C. (2001). The arabic language. Edinburgh University Press.

ZALAMI, F.B. (2007). First language acquisition. in *Encyclopedia of Arabic Language and Linguistics*, Boston, Brill:Ed, (2) :96–10.

L'effet des contextes segmentaux et prosodiques sur la glottalisation de la voyelle initiale de mot en persan

ASSADI Suzanne

Laboratoire de Phonétique et Phonologie (UMR 7018) CNRS / Sorbonne Nouvelle, Paris- France

suassadi@yahoo.fr

Résumé

Cette communication étudie l'effet des contextes segmentaux et prosodiques sur la glottalisation de la voyelle initiale de mot en persan. L'enregistrement du corpus est basé sur 20 minutes de conversation avec trois sujets ainsi que sur 18 mots isolés.

Les résultats montrent que la pause et la voyelle sont des contextes favorables pour l'occurrence du coup de glotte et de la glottalisation respectivement alors que la consonne (occlusive et fricative) les favorise le moins. La comparaison des voyelles initiales des mots accentués (mise en relief) avec celles des mots non-accentués révèle l'importance de l'accent d'insistance sur l'occurrence du coup de glotte et/ou de la glottalisation. Au contraire, l'accent lexical n'a pas d'effet sur cette occurrence et il n'y a pas de différence significative entre la durée de la partie glottalisée des voyelles accentuées et non accentuées.

MOTS CLÉS: Persan, glottalisation, voyelle, accent.

Influence of segmental and prosodic contexts on the glottalization of the word initial vowel in Persian

Abstract

This paper examines the occurrence of glottal stop and glottalization of the word-initial vowels in Persian according to segmental and prosodic contexts. Data analysis was conducted using 20 minutes of conversation with three native Persian speakers and 18 isolated words.

Pause and vowel are determining factors for the occurrence of glottal stop and glottalization respectively while with consonant (plosive and fricative) this effect is less important. The comparison of initial vowels of focally accented vs. unaccented words reveals the importance of focal accent on the occurrence of glottal stop and /or glottalization. On the contrary, the lexical stress has no effect on this occurrence and there is no significant difference between the duration of the glottalized part of the vowel in accented and unaccented vowels.

KEYWORDS: Persian, glottalization, vowel, stress.

1. Introduction

Les phénomènes de glottalisation recouvrent divers termes dans les langues et peuvent avoir des fonctions différentes dont les plus importantes consistent à : i) marquer la frontière de mot ou de morphème commençant par une voyelle, ii) modifier la qualité de la voix à la fin d'une phrase assertive iii) servir en contraste phonémique et iv) être un

allophone des plosives sourdes dans certaines positions. Cette recherche aborde, la première fonction mentionnée ci-dessus et les contextes qui influencent l'occurrence du coup de glotte et/ou de la glottalisation de la voyelle initiale de mot en persan.

Divers facteurs peuvent favoriser l'occurrence de la glottalisation dans les langues : l'accent d'insistance (Carton et al. 1977 en français, Umeda 1978 et Pierrehumbert 1995 en anglais, Kohler 1994 en allemand), la structure prosodique (Dilley & al. 1996 ainsi que Pierrehumbert 1995 en anglais, Fougeron 1998 en français), le contexte segmental (Kohler, 1994 en allemand et Dilley, 1996 en anglais), le style (Dilley et Shattuck-Hufnagel 1995 en anglais) et le débit (Assadi 2003a et 2003b en persan).

Selon Umeda (1978), en anglais, l'occlusive glottale apparaît au début de la voyelle accentuée initiale, surtout après une pause, ou à la frontière de deux mots (V-V). Pierrehumbert et Talkin (1992), et plus tard Dilley et al. (1996) ont insisté sur l'importance du contexte prosodique sur la glottalisation. Allen (1957), au contraire, donne moins d'importance aux facteurs prosodiques. Selon lui, l'usage de l'occlusive glottale est conditionné par la position de la transition V-V à la frontière du mot.

Ficher-Jørgensen (1989) et Kohler (1994) ont, à leur tour, souligné le rapport entre la glottalisation et la prosodie. Ficher-Jørgensen (1989) a effectué une analyse détaillée du «stød» danois. Du point de vue phonétique, le «stød» est une sorte de phonation reliée à la voix craquée. D'après elle, historiquement, il est relié aux tons du mot. Du point de vue phonologique, le «stød» est une marque prosodique définissant la syllabe dans certains types de mots. Kohler (1994) a étudié la fréquence de différentes réalisations de l'occlusive glottale en allemand. Selon lui, il y a un rapport significatif entre l'occurrence du coup de glotte avec la voyelle accentuée suivante et l'élision de celui-ci avec la voyelle inaccentuée suivante.

Les recherches phonétiques sur le persan sont peu nombreuses et souvent basées sur l'impression perceptive des chercheurs. Nos études expérimentales antérieures montrent que le débit influence la glottalisation de la voyelle initiale de mot. L'effet du débit rapide se révèle significativement par une version affaiblie du coup de glotte (la glottalisation). Par ailleurs, l'enchaînement des mots ainsi que des cas non glottalisés sont deux fois plus fréquents en débit rapide qu'en débit lent.

A présent, l'hypothèse posée est que d'autres facteurs comme les contextes prosodiques et segmentaux influencent l'occurrence du coup de glotte et/ou de la glottalisation de la voyelle initiale de mot en persan. Les paramètres prosodiques examinés dans cette étude seront donc l'effet de l'accent lexical et de l'accent d'insistance sur la glottalisation. L'accent lexical est en général, situé sur la dernière syllabe des noms et des adjectifs. En revanche, les formes verbales personnelles ont un accent récessif. Par ailleurs, les paramètres acoustiques de l'accentuation du persan sont la durée, la f0 et l'intensité. Le but de cette recherche est d'examiner l'occurrence de différentes réalisations de l'occlusive glottale en fonction des contextes segmentaux et prosodiques. Le sujet de cette étude est le persan parlé de Téhéran (Iran). C'est la langue de toutes les classes de la société dans leurs relations courantes. Le renouveau littéraire a aujourd'hui fait entrer le persan parlé dans la littérature. Le persan littéral (style « soutenu ») des journaux et de

l'administration s'en écarte. Chaque apprenant doit donc pouvoir maîtriser ces deux niveaux de langue.

2. Étude expérimentale

2.1. Méthode

2.1.1. Locuteurs : trois sujets issus du milieu universitaire, originaires de Téhéran et âgés de 32 à 42 ans ont été enregistrés au laboratoire de phonétique à l'université paris III.

2.1.2. Corpus

Le corpus est composé de deux parties :

- 20 minutes de conversation (six à sept minutes d'entretien individuel avec trois sujets). Nous avons demandé à chacun de nous raconter le meilleur film qu'il a vu et ensuite de nous expliquer la recette d'un gâteau. Le corpus a été analysé et 189 mots avec une voyelle en position initiale ont été relevés.

-18 mots isolés, inscrits sur des cartes et se terminant par un point.

Ces mots ont été présentés à deux des trois locuteurs ci-dessus avec un intervalle de quelques secondes. Les locuteurs ont lu les mots. Les mots isolés constituent des phrases, composées d'un seul mot. Nous n'avons pas inséré ces mots dans des phrases cadres car dans ce cas, le mot cible est précédé d'une voyelle, un contexte qui favorise la glottalisation.

2.2. Analyse et résultats

Nous étudions les questions suivantes :

1. Dans quelle mesure le contexte précédent influence-t-il le coup de glotte et / ou la glottalisation de la voyelle initiale de mot ?

2. L'accent d'insistance et l'accent lexical sont-ils des facteurs importants pour l'occurrence du coup de glotte et /ou de la glottalisation de la voyelle initiale de mot ?

3. La durée de la glottalisation augmente-t-elle en fonction de l'accent lexical ?

Critères de classement du coup de glotte et de la glottalisation: Les chercheurs (entre autres, Kohler 1994, Dilley et Shattuck-Hufnagel 1995) utilisent souvent une typologie qui permet de classer les différentes réalisations des locuteurs et facilite donc leur interprétation. Nous nous sommes inspirée en partie de l'étude de Kohler (1994) en allemand sur les différentes réalisations de «l'occlusive glottale ». Nous avons ainsi attribué les termes de coup de glotte à un silence suivi d'un relâchement et glottalisation aux striations verticales irrégulières sur le spectrogramme correspondant aux vibrations irrégulières des plis vocaux. Cette recherche fait partie d'une analyse détaillée des phénomènes de glottalisation en persan. Le choix du classement a été fait en fonction de l'examen des premières périodes au début de la voyelle initiale de mot, l'examen spectrographique, la courbe de F0 et l'aspect perceptif. Nous avons représenté les différentes réalisations acoustiques de « l'occlusive glottale» ainsi :

Type 1 : coup de glotte + glottalisation (figure 1).

Type 2 : coup de glotte sans glottalisation (figure 2).

Type 3 : glottalisation (figure 3).

Autres : abaissement de l'amplitude entre deux voyelles, absence de glottalisation due à l'enchaînement avec la syllabe du mot précédent, cas ambigus.

2.2.1. Effet du contexte précédent sur la voyelle initiale de mot

Question : Dans quelle mesure le contexte précédent influence-t-il le coup de glotte et / ou la glottalisation de la voyelle initiale de mot ?

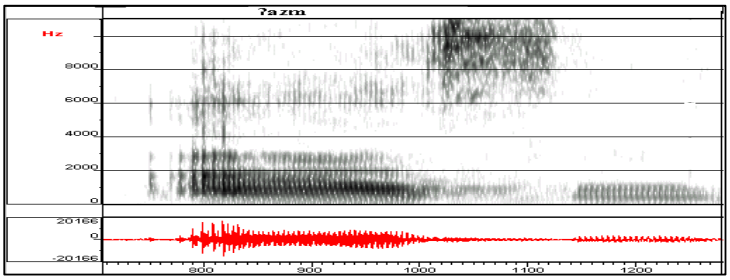


FIGURE 1: Spectrogramme du mot [ʔazm] (intention), type 1: coup de glotte + glottalisation.

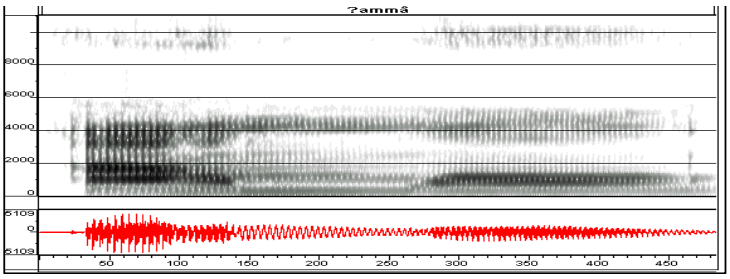


FIGURE 2 : Spectrogramme du mot [ʔamma] (mais), type 2 : coup de glotte sans glottalisation.

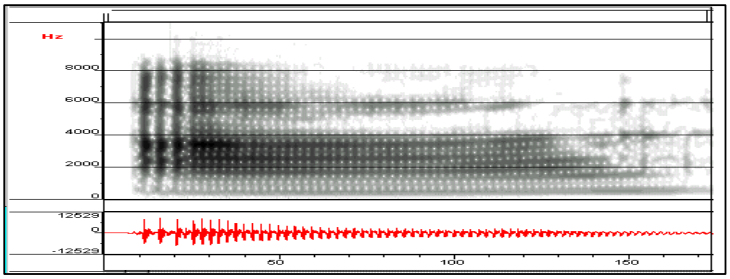


FIGURE 3 : Spectrogramme du mot [ʔaz] (de), type 3: glottalisation.

Le tableau 1 montre l'occurrence de coup de glotte et/ou de la glottalisation de la voyelle initiale de mot pour l'ensemble des locuteurs, en tenant compte des contextes précédents : pause, voyelles, obstruents "occlusives et fricatives" et sonantes [l, m, n, r]).

Critères de classement selon la pause : Dilley et al. (1996) analysent la glottalisation en anglais en tenant compte de pauses d'une durée minimum de 50 ms avant le segment cible qui indique juste une interruption sur le signal. Duez (1982) définit la durée des pauses « silencieuse » perceptibles entre 180 à 250 ms. Concernant notre recherche, nous avons choisi les pauses *perceptibles* dont le seuil est de 180 ms, mentionné par Duez

(1982). Les pauses considérées étaient également *perceptibles* par une mini-expérience auprès de trois auditeurs persans.

Distinction de pause et de coup de glotte : Umeda (1978) considère un silence de moins de 100 ms sur le signal comme un coup de glotte. Selon Umeda, ce dernier et la pause ont la même fonction. Comme Kohler (1994), nous distinguons ici un vrai silence du discours (pause), et un silence qui fait partie d'une articulation, c'est-à-dire une plosive sourde ou occlusion silencieuse qui est le résultat d'une articulation et qui se présente dans le contexte d'une consonne et non d'une pause. Donc, dans cette étude, la présence ou l'absence d'une articulation consonantique (un silence suivi d'un relâchement) distinguera entre les deux cas.

Contexte précédent	Nombre de voyelle avec coup de glotte et / ou glottalisation	% de voyelle avec coup de glotte et / ou glottalisation
Voyelle	46	26 %
Pause	29	16 %
Sonante	20	11 %
Cons occ + fric	8	5 %
Total	103	58 %

TABLEAU 1: Fréquence du coup de glotte et / ou de la glottalisation de la voyelle initiale de mot (types 1, 2 et 3), selon le contexte précédent (pour l'ensemble des locuteurs).

Il est évident que la présence d'une pause implique que la frontière considérée est celle d'un constituant prosodique supérieur au mot. Mais, ici, compte tenue du manque d'études détaillées sur le système prosodique du persan, je ne considérerai que l'effet d'une présence ou absence de pause sans faire d'interprétation sur le niveau de frontière.

Afin d'examiner lesquels des quatre contextes étudiés favorisent le plus le coup de glotte et / ou la glottalisation, nous avons utilisé un test chi 2 simple (analyse univariée). Il permet de conclure que la voyelle et, dans une moindre mesure, la pause sont des contextes favorables à la réalisation de l'ensemble du coup de glotte et / ou de la glottalisation, et que la consonne (occlusive et fricative) les favorise le moins : 103 / 4 = 25,75 et chi 2 = 29,85 avec $p < 0{,}0001$. Ce qui montre que la distribution du coup de glotte et / ou la glottalisation en fonction des différents contextes n'est pas homogène (tableau 1).

Le tableau de la page suivante montre l'occurrence de différentes réalisations de l'occlusive glottale selon le contexte précédent pour l'ensemble des locuteurs. Parmi les 189 voyelles initiales du corpus, 103 présentent un coup de glotte et /ou une glottalisation, 35 comportent un abaissement de l'amplitude (une version affaiblie de la glottalisation) et 37 présentent un enchaînement avec la syllabe précédente. Il y a 14 cas ambigus. En ce qui concerne les différentes réalisations acoustiques selon le contexte, on constate que la voyelle précédente favorise la glottalisation (type 3, c'est-à-dire 28 cas sur 46 qui correspondent à 61% d'occurrences). La pause est un contexte favorable pour

l'occurrence du coup de glotte avec ou sans glottalisation (les types 1 et 2, c'est-à dire 27 cas sur 29 correspondant à 93% d'occurrences). Il y a donc une différence significative entre la voyelle et la pause en ce qui concerne l'occurrence du coup de glotte : chi $2 = 21{,}55$ avec $p < 0{,}0001$.

Les sonantes favorisent la glottalisation (type 3) mais de façon moindre que les voyelles. Par contre, après les obstruents (occlusive et fricative) la glottalisation est rare et l'enchaînement avec la syllabe du mot suivant est fréquent.

Classement	Type 1: coup de glotte + glottalisation	Type 2: coup de glotte	Type 3: glottalisation	Total	Abaissement de l'amplitude	Enchaînement
Pause	16	11	2	29 (16%)	-	-
Obstruents	3	2	3	8 (5%)	2 (1%)	19 (11%)
Voyelle	16	2	28	46 (26%)	27(15%)	4 (2%)
Sonante	5	-	15	20 (11%)	6 (3%)	14 (8%)

TABLEAU 2: Classement de différentes réalisations de l'occlusive glottale selon le contexte précédent pour l'ensemble des locuteurs.

La prochaine étape consiste à étudier l'effet de l'accent d'insistance.

2.2.2 Effet de l'accent d'insistance

Question : L'accent d'insistance est-il un facteur important pour l'occurrence du coup de glotte et /ou de la glottalisation de la voyelle initiale de mot en persan ?

L'accent d'insistance signifie ici les mots mis en relief. Pour les trois sujets les mots monosyllabiques mis en relief commençant par une voyelle ont été pris en compte et nous les avons comparés avec leur correspondant non accentué. En choisissant les mots monosyllabiques, l'accent tombe sur les voyelles cibles. Le contexte précédent reste constant et se termine également par une voyelle. La comparaison s'est donc fait ainsi :

voyelle # voyelle dans les mots mono-syllabiques (20 mots mis en relief)

voyelle # voyelle dans les mots mono-syllabiques (57 mots non-mis en relief).

L'analyse des données montre que toutes les voyelles cibles accentuées comportent un coup de glotte /et/ ou une glottalisation, alors que seulement 45% des mots non-accentués comportent un coup de glotte /et/ ou une glottalisation (tableau 3). Ceci confirme l'étude de Pierrehumbert (1995), sur l'ampleur des gestes sous accent. Fisher-Jørgensen (1989) et Kohler (1996) décrivent la glottalisation comme une version affaiblie du coup de glotte. Ces deux versions (coup de glotte et glottalisation), correspondraient, selon Fisher-Jørgensen, à différents degrés d'activité des muscles du larynx. Dans le reste des cas, on observe un faible abaissement de l'amplitude indiquant la transition d'une voyelle à l'autre ou l'absence de la glottalisation due à la contraction des deux voyelles adjacentes.

Dans la suite de notre étude, nous avons voulu savoir si l'accent lexical jouait également un rôle dans l'occurrence du coup de glotte et /ou de la glottalisation.

Différentes réalisations acoustiques	Mots mis en relief (20)	Mots non-mis en relief (57)
coup de glotte et glottalisation	10	6
coup de glotte	2	0
glottalisation	8	20
Abaissement de l'amplitude	-	27
Enchaînement	-	4
Total coup de glotte et /ou glottalisation	20 (100%)	(26) 45%

TABLEAU 3 : Différentes réalisations de la voyelle initiale de mot dans les mots mis en relief et non-accentués (ensemble des sujets).

2.2.3. Effet de l'accent lexical

Question : l'accent lexical est- il un facteur important pour l'occurrence du coup de glotte et /ou de la glottalisation de la voyelle initiale de mot ?

Étant donné que l'accent lexical tombe sur la dernière syllabe des mots étudiés en persan, deux types de mots commençant par la [*a*] ont été étudiés. Le premier groupe est monosyllabique, tel que le mot "âb" (eau), alors que le second est dissyllabique dans "âbi" (bleu). Les premiers comportent un accent sur la voyelle initiale cible, mais dans le second groupe, la voyelle initiale n'est pas accentuée. Si nous avons choisi des mots commençant par la voyelle [*a*], c'est à cause de sa grande fréquence dans les mots mono et dissyllabiques.

L'un des avantages que présente l'étude des voyelles en position initiale dans les mots prononcés isolément, est qu'il n'y a pas d'enchaînements ni de coarticulation avec les mots précédents, comme ce serait le cas dans une phrase. Cependant, un mot présenté en isolation est une phrase, le début du mot isolé est un début de phrase et de syntagme. On ne peut pas non plus distinguer la variation due à l'accent lexical de celle de l'accent de phrase. Comme tous les mots ont été lus de la même façon, et dans les mêmes conditions, ils peuvent faire l'objet d'une comparaison.

L'analyse des données montre que dans la majorité des cas il n'y a pas de différence entre les réalisations des locuteurs dans les deux types de mots. Cela pourrait être dû au fait que l'accent lexical est assez faible en persan (Lazard, 2006, P.28).

On pourrait s'attendre à ce qu'un mot accentué commençant par voyelle comporte plus d'irrégularités qu'un mot non accentué. Afin de savoir si la durée de la portion glottalisée de la voyelle augmente en fonction de l'accent, nous avons mesuré la durée totale de chaque voyelle accentuée et non accentuée ainsi que celle de la partie glottalisée. Ces

durées sont légèrement plus grandes dans les syllabes accentuées que non accentuées (figure 4). Le test Anova à mesures répétées indique toutefois que la différence de la portion glottalisée n'est pas significative entre les deux types de voyelles : F (1, 17) = 1,56 ; P = 0,2276.

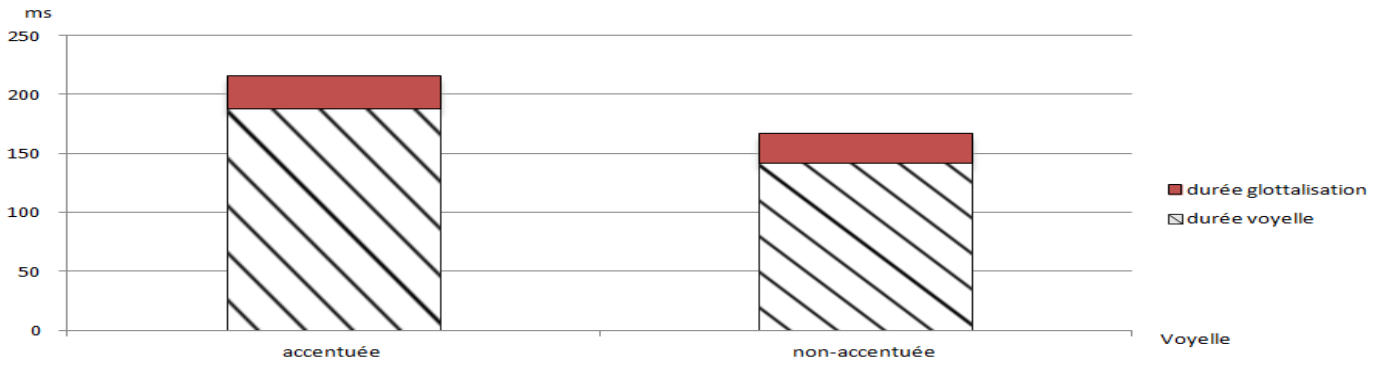


FIGURE 4 : Durée de voyelle ainsi que celle de la partie glottalisée pour les voyelles accentuées et non-accentuées.

3. Conclusion et discussion

Cette étude montre qu'en persan, comme en anglais et en allemand, l'accent d'insistance et la pause sont des facteurs favorisant le coup de glotte au début de la voyelle initiale de mot. Mais à la différence de l'anglais, l'accent lexical en persan ne semble pas y jouer un rôle. On peut donc dire qu'il existe une occlusive glottale avant la voyelle initiale de mot mais, elle n'apparait que lorsque l'ampleur et la durée des gestes sont suffisantes (Pierrehumbert, 1995) puisque le mécanisme des vibrations des plis vocaux n'est pas linéaire.

Un coup de glotte représente une rupture (Vaissière, 2003). Il indique également un trait de renforcement par rapport à la glottalisation correspondant aux différents degrés d'activité des muscles du larynx (Ficher-Jørgensen, 1989).

Contrairement à la consonne (occlusive et fricative), la voyelle est un contexte favorable pour la glottalisation de la voyelle initiale de mot. Dans le cas où deux voyelles sont en hiatus à la frontière de mot, malgré les variations inter locuteurs, pour chaque sujet, nous avons trouvé les cas suivants : la glottalisation, l'abaissement de l'amplitude, la contraction des deux voyelles adjacentes ou bien encore l'insertion des semi-voyelles. Par ailleurs, un fort abaissement de l'amplitude accompagné d'une baisse de la F0 donnait la perception d'une voix craquée. Nous avons également démontré (Assadi 2003a et b) que pour ce type de voix, le signal laryngographique indique soit un dédoublement de la période vibratoire soit des cycles d'amplitude basse alternant avec des cycles d'amplitude élevée. Ces deux sortes de signaux sont définies comme voix craquée «creaky voice» par Fourcin (1981).

Cette recherche contribue à montrer certains aspects universels des phénomènes de glottalisation basés sur les gestes articulatoires et est une base pour les études ultérieures sur l'occurrence de la glottalisation selon les différents constituants prosodiques.

Références

ALLEN J. (1970). The glottal stop as a junctural correlate in English ». *Journal of the Acoustical Society of America* 47: 57 (résumé).

ASSADI, Sh. S. (2003a). Laryngographic, Acoustic and Perceptive Study of Glottalization Phenomena in Persian. *Proceedings of the XVth International Congress of Phonetic Sciences.* Barcelona. 3117-3120.

ASSADI, Sh. S. (2003b). *Les phénomènes de glottalisation en persan (langue standard / langue parlée*). Thèse de doctorat. Université de la Sorbonne Nouvelle. Paris.

CARTON, F. HIRST, D. MARCHAL, A. & SEGUINOT, A. (1977). *L'accent d'insistance / Emphatic stress*. Didier. Montréal. Paris. Bruxelles.

DILLEY, L. and SHATTUCK-HUFNAGEL, S. (1995). Variability in glottalization of word onset vowels in American English . *Proceedings of the XIIIth International Congress of Phonetic Sciences Stockholm*. 586-589.

DILLEY, L. SHATTUCK-HUFNAGEL, S. & OSTENDORF, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24: 423-444.

DUEZ, D. (1982). Silent and non silent pauses in three speech-styles. *Language and Speech*: 21-32.

FICHER-JøRGENSEN, E. (1989). "Phonetic analysis of the stød in Standard Danish". *Phonetica* 46: 1 –59.

FOUGERON, C. (1998). *Variations articulatoires en début de constituants prosodiques de différents niveaux en français.* Thèse de doctorat. Université paris III.

FOURCIN, A. J. (1981). Laryngographic assessment of phonatory function. *Proceedings of the Conference on the Assessment of Vocal Pathology*. Report 11: 116-127.

KOHLER, K.J. (1994). Glottal stops and glottalization in German. Data and theory of connected speech processes. *Phonetica* 51: 38-51.

KOHLER, K.J. (1996). Glottalization across languages. *Arbeitsberichte* des *Institute für Phonetik und digitale Sprachverarbeitung (AIPUK)* 31: 207-216. Kiel.

LAZARD, G. (2006). *Grammaire du persan contemporain*. Eds. Farhang Moaser. Téhéran.

PIERREHUMBERT, J. (1995). Prosodic Effects on Glottal Allophones. In *Vocal Fold Physiology*: 39-60. Fujimura 0. and Hirano M. Eds. Singular Publishing Group, Inc. San Diego. California.

UMEDA, N. (1978). Occurrence of glottal stop in fluent speech. *Journal of the Acoustical Society of America.* 64 (1) : 88-94.

VAISSIÈRE, J. (2001). Changements de sons et changements prosodiques : du latin au français. *Revue Parole* 15/16: 53-88.

Opposition /y/-/u/ chez sept apprenantes japonophones avec et sans aide de l'ultrason lingual: 2. identification perceptive des voyelles isolées par 16 francophones natifs

Takeki Kamiyama[1,2], Claire Pillot-Loiseau[2], Tanja Kocjančič Antolík[2]
(1) Linguistique Anglaise Psycholinguistique (LAPS), EA 1569, Université Paris 8 : 2 rue de la Liberté, 93526 Saint-Denis Cedex
(2) Laboratoire de Phonétique et Phonologie (LPP) UMR 7018 CNRS / Université Paris 3 Sorbonne Nouvelle : 19 rue des Bernardins, 75005 PARIS, FRANCE
takeki.kamiyama@univ-paris8.fr, claire.pillot@univ-paris3.fr, tkocjancic@googlemail.com

RESUME

Quatre apprenantes japonophones (EXP) ont bénéficié d'un entraînement de prononciation avec rétrocontrôle par ultrason pour produire /u/-/y/ français. Elles ont été enregistrées avant (ENR1), une semaine après (ENR2), et deux mois après l'entraînement (ENR3). Trois autres apprenantes (CTR) ont aussi été enregistrées deux fois. Seize francophones natifs ont écouté des stimuli (4 occurrences de /y/ et /u/ isolés) prononcés par chaque locutrice à ENR1, 2, et 3, et ont identifié la voyelle avec une note. En ENR1, une EXP a produit le /u/ avec une note moyenne supérieure à 2/3 ; en ENR2 et ENR3, le /u/ de 2 EXP (ENR2) et celui de 3 EXP (ENR3) atteint une note supérieure à 2. Cellle des CTR n'a pas dépassé 1,3. Le retour ultrason est donc utile pour /u/, dont la position de la langue est invisible.

ABSTRACT

/y/-/u/ contrast in French by seven Japanese-speaking learners with and without the help of Ultrasound: 2. perceptual identification of isolated vowels by 16 native French listeners

Four Japanese-speaking learners (EXP) received pronunciation training with ultrasound feedback in producing French /y/-/u/. They were recorded before (REC1), a week after (REC2), and two months after the training (REC3). Three other learners (CTR) were recorded twice at similar intervals. Sixteen French native subjects listened to stimuli (4 tokens of /y/ and /u/ in isolation) pronounced by each learner in REC1, 2, and 3. They identified the vowel, providing a goodness rating. In REC1, one EXP produced /u/ that obtained a mean goodness score above 2/3 for /u/; in REC2 and REC3, the /u/ productions of 2 and 3 EXP's respectively reached a score above 2. Those produced by the CTR's did not obtain a score above 1.3. So ultrasound feedback may be useful especially for /u/, for which the tongue position is invisible.

MOTS-CLES : apprenantes japonaises, entraînement phonétique, ultrason, position de la langue, évaluation perceptive, auditeurs francophones, voyelles isolées
KEYWORDS: Japanese-speaking learners, pronunciation training, ultrasound, tongue position, perceptual evaluation, French-speaking listeners, vowels in isolation.

1 Introduction et but de l'étude

Il a été largement montré que le système vocalique d'une langue maternelle a un impact significatif sur l'acquisition des voyelles de langues étrangères. Les voyelles arrondies fermées antérieures, comme /y/, et postérieures, comme /u/, sont connues pour poser des difficultés aux locuteurs de langues qui ne possèdent pas ces deux types de phonèmes (Rochet 1995, Levy et Law 2010, parmi d'autres ; Flege 1987, Lauret 1998, Kamiyama and Vaissière 2009).

La voyelle antérieure arrondie fermée /y/ et la voyelle postérieure arrondie fermée /u/ sont phonologiquement en opposition dans le français parisien contemporain et parvenir à réaliser cette opposition est d'une grande importance pour les apprenants du français langue étrangère (FLE). Des locuteurs japonophones natifs de Tokyo ont tendance à produire le /u/ français avec un deuxième formant (F2) plus élevé que les locuteurs francophones natifs, ce qui résulte en une voyelle identifiée plutôt comme /ø/ par des auditeurs francophones (Kamiyama et Vaissière 2009). Cela est directement lié à la nature de la voyelle fermée non antérieure /u/ en japonais de Tokyo, communément transcrite [ɯ]. D'un point de vue acoustique, cette voyelle montre un F2 plus élevé (supérieur à 1000Hz) que celui du /u/ français ; d'un point de vue articulatoire, la langue est située moins loin en arrière et les lèvres sont moins arrondies (Bothorel *et al.* 1986; Uemura & Takada 1990).

Il est donc nécessaire, pour les japonophones apprenant le français, de positionner leur langue aussi loin en arrière que le font les francophones natifs quand ils produisent la voyelle /u/. Toutefois, parce que la position linguale en arrière ne peut pas être observée durant la parole continue, la correction de cette position se pratique couramment uniquement sur la base d'une évaluation perceptive. D'un autre côté, l'imagerie ultrason de la langue, qui est la technique la moins invasive, la plus fiable et la plus économique, utilisée pour observer la langue durant la parole, permet le contrôle direct et en temps réel de la position linguale. Cette visualisation a montré son efficacité pour traiter des troubles de la parole (Bernhardt *et al.* 2008), mais il n'existe pas à notre connaissance de recherches prouvant l'efficacité de cet équipement pour améliorer la prononciation du FLE.

Le but de notre recherche est d'évaluer la production de l'opposition entre les voyelles françaises /y/ et /u/ par quatre apprenantes japonaises de niveau intermédiaire ayant suivi des leçons spécifiques à l'aide de l'ultrason lingual pour la production de ces voyelles, et de trois autres apprenantes de même origine n'ayant pas bénéficié de cette aide spécifique. Cet article présente une étude d'évaluation auditive des productions acoustiques des apprenantes par des auditeurs francophones natifs.

2 Méthode

Les apprenantes du groupe expérimental (EXP), EXP1 (33 ans, niveau B2), EXP2 (31 ans, B1), EXP3 (28 ans, B1) et EXP4 (29 ans, B2), de niveau de français intermédiaire, ont, en plus de ces trois leçons spécifiques avec l'ultrason, bénéficié de 12 semaines de formation traditionnelle à la phonétique du FLE, comme pour

celles du groupe contrôle (CTR), CTR1 (29 ans, niveau B1), CTR2 (33 ans) et CTR3 (33 ans, niveau B1), qui n'ont pas bénéficié de ces leçons. Toutes sont natives de Tokyo ou des environs, sauf CTR2 de Fukuoka et EXP4 de Kobe. Elles ont commencé à apprendre le français adulte et vivent en France depuis au moins deux ans, sauf CTR1 (5 mois). Une locutrice native française de 42 ans (NAT1) a également été enregistrée.

Pour mettre en évidence les différences dans l'articulation de [y] et [u], les données ultrason et acoustiques ont été enregistrées en même temps avec une synchronisation de ces deux signaux, en utilisant le logiciel *Articulate Assistant Advanced* (AAA, Articulate Instruments Ltd, 2008). La TABLE 1 résume l'ensemble des enregistrements réalisés par sujet et leur chronologie.

Chronologie	**1[er] enregistrement**	**2[ème] enregistrement**	**3[ème] enregistrement**	**Groupe**
Enregistre-ments et sujets	**ENR 1** (1 semaine avant 1[ère] leçon et formation 12 semaines)	**ENR 2** (1 semaine après 3[ème] leçon et formation 12 semaines)	**ENR 3** (2 mois après le ENR 2)	
NAT1	X			Référence
EXP1	X	X	X	Expéri-mental
EXP 2	X	X	X	
EXP 3	X	X	X	
EXP 4	X	X	X	
CTR 1	X (sans leçon ultrason)	X (sans leçon ultrason)		Contrôle
CTR 2	X (sans leçon ultrason)	X (sans leçon ultrason)		
CTR 3	X (sans leçon ultrason)	X (sans leçon ultrason)		

TABLE 1 – Enregistrements réalisés pour chaque sujet (19 au total).

Les quatre EXP bénéficié de **trois sessions d'entraînement** de 45 minutes dans lesquelles l'ultrason lingual a été utilisé pour réaliser et contrôler la position de la langue pour les voyelles cibles /y/ et /u/. L'entraînement a débuté avec la répétition avec puis sans image ultrason, puis la lecture de : 1) voyelles /y/ et /u/ isolées puis alternées, 2) syllabes avec puis sans allongement de la voyelle cible, où /y/ et /u/ apparaissaient en position initiale, finale puis médiane. Les voyelles cibles de ces syllabes apparaissaient dans des contextes phonétiques facilitants, neutres puis difficiles tels qu'ils ont été définis pour le français par Callamand (1981). 3) Ces leçons se sont poursuivies avec la répétition puis la lecture de mots respectant ces contextes, puis de phrases et textes contenant ces mots. Le protocole exact de cet apprentissage était ajusté aux capacités et aux préférences de chaque participant (par exemple : position finale des voyelles cibles parfois plus favorisante que la position initiale dans les syllabes, donc abordée avant ; la réalisation du [u] est favorisée par le contexte [uk] chez certaines, [uʁ] chez d'autres).

Le corpus était composé de dix répétitions (par deux séries de cinq) de [y] et [u] isolés, ainsi que les voyelles isolées [a], [i] et, pour les sept japonophones, [ɯ]. Des logatomes, mots et phrases avec [y] et [u] ont aussi été enregistrés mais dans l'étude

d'évaluation auditive présentée dans cet article, seules les voyelles isolées ont été utilisées comme stimuli.

Les données acoustiques, enregistrées à 22050Hz, 16bits, ont permis l'analyse des quatre premiers formants mesurés à 25%, 50% et 75% de la durée de chaque voyelle sous Praat (Boersma & Weenink 2013).

Parmi les 10 répétitions de chaque locutrice dans chaque séance d'enregistrement, 4 (les 3[e], 4[e], 8[e] et 9[e]) ont été extraites, afin d'éviter que le test soit trop long (ce qui a permis de limiter le temps de passation à 25 minutes environ). Ainsi ont été sélectionnés 228 stimuli composés de 3 voyelles cibles (/u/, /y/ ainsi que /i/) x 4 répétitions x 19 séances d'enregistrement (= 4 EXP x 3 + 3 EXP x 2 + 1 EXP x 1). À cela ont été ajoutées 19 occurrences du /a/ produites par les 8 locutrices, ce qui fait 247 stimuli au total. Ces stimuli ont été placés dans un ordre semi-aléatoire pour que la même voyelle cible ne se succède pas 3 fois ou plus, et que la même locutrice ne se suive pas. Deux listes ont été préparées avec l'ordre inverse des stimuli : l'une a été présentée à la moitié des auditeurs, l'autre aux autres auditeurs.

Un total de 16 auditeurs francophones natifs originaire de la France métropolitaine, âgés de 23 à 63 ans, ont participé à l'expérience. Ils ont été informés qu'ils allaient entendre des voyelles du français prononcées par des étudiantes étrangères, et qu'ils devraient indiquer quelle voyelle ils auraient entendue et si la voyelle aurait été un plus ou moins bon exemple de la catégorie vocalique choisie (mauvais / moyen / bon). Les stimuli ont été présentés sur un ordinateur à travers un casque audio, en utilisant Praat (Boersma & Weenink 2013), et les auditeurs ont répondu en cliquant sur les cases correspondantes sur l'écran. Chaque bouton de réponse contenait le symbole phonémique de la voyelle et un mot exemple (ex. « /i/ (lit) »). Chaque stimulus était précédé par un bip sonore (440 Hz, 50 ms) et un silence d'une seconde environ. Avant de passer le test, les auditeurs ont 1) lu les consignes écrites présentées sur un écran ; 2) passé une phase d'habituation où ils pouvaient entendre les voyelles isolées d'une locutrice native et voir les symboles phonétiques correspondants ainsi que des mots exemples ; et 3) passé un entraînement. Le temps de passation était de 30 minutes environ.

3 Résultats

3.1 Analyse des réponses

Les réponses enregistrées ont été traitées de manière suivante (cf. Guion et al. 2000). Premièrement, le nombre et les types de réponses (voyelles) obtenus pour chacune des deux voyelles cibles (/u/ ou /y/) prononcées par chaque locutrice dans chaque séance d'enregistrement ont été calculés (64 réponses = 16 auditeurs x 4 productions). Deuxièmement, le score a été calculé pour chaque auditeur en multipliant la note d'évaluation par la fréquence de réponses. Ainsi, la voyelle cible /u/ prononcée par une locutrice dans une séance d'enregistrement a été identifiée 3 fois comme /u/ et 1 fois comme une autre voyelle (/ø/, par exemple) par un auditeur donné, avec des notes de 2, 2, 1, et 1, respectivement, obtient des scores de 1,25 (= (2+2+1)/4) pour la réponse /u/ (correspondance entre la cible et la

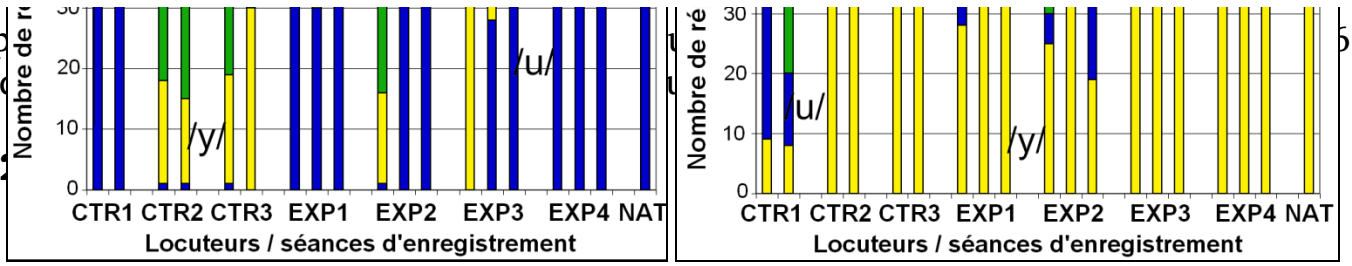


FIGURE 1 montre le nombre et les types de réponses obtenues pour les cibles /u/ et /y/.

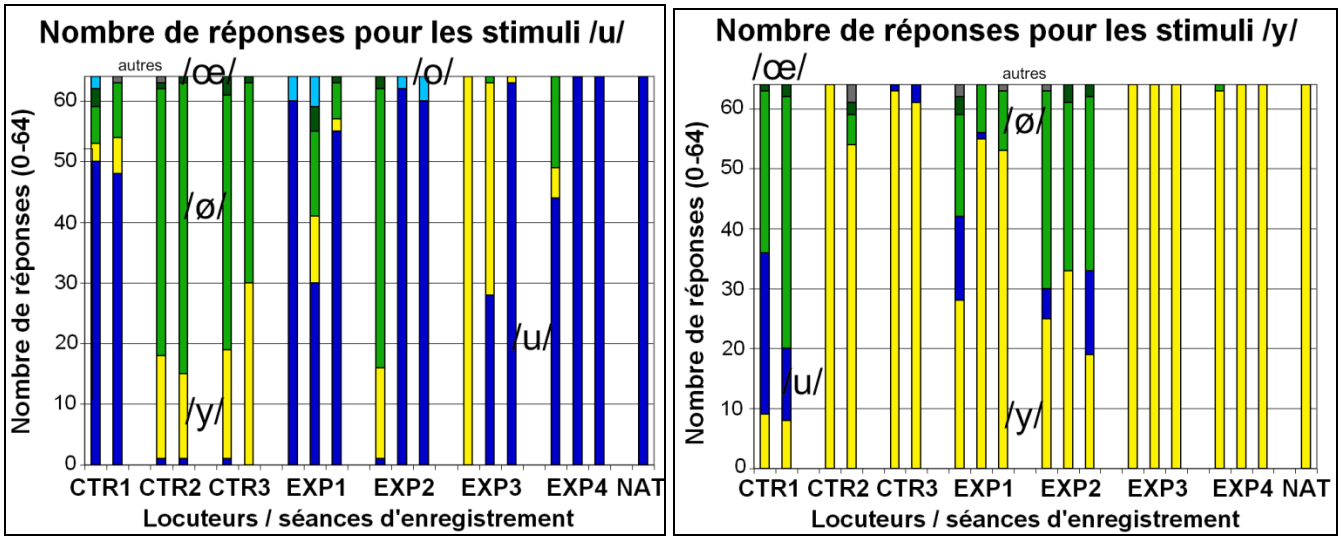


FIGURE 1 – Nombre et types de réponses d'identification aux stimuli des voyelles cibles /u/ (gauche) et /y/ (droite). De gauche à droite : ENR1, ENR2, ENR3 pour chaque locutrice.

3.2.1 Voyelle cible /u/

Dans le groupe CTR, seule CTR1 a obtenu la majorité (50 en ENR1, 48 en ENR2) de réponses « attendues » (correspondance entre la voyelle cible et la voyelle identifiée par les auditeurs). CTR2 et CTR3 ont produit des voyelles identifiées comme /ø/ ou /y/ dans les deux séances d'enregistrement. Chez les apprenantes EXP, nous pouvons observer une amélioration, sauf chez EXP1. Le nombre de réponses pour /u/ augmente progressivement (mise à part la diminution de ENR2 à ENR3 chez EXP2 (de 62 à 60 sur 64)), et atteint presque le plafond (entre 60 et 64) en ENR2 (EXP2 et EXP4) et en ENR3 (EXP2, EXP3, EXP4). Notons cependant que 3 apprenantes (CTR1, EXP1 et EXP4) ont obtenu la majorité de réponses « attendues » dès ENR1, mais le score moyen est inférieur à 1,5/3 sauf EXP1 (voir section 3.3).

3.2.2 Voyelle cible /y/

Parmi les CTR, CTR2 et CTR3 ont obtenu une grande majorité de réponses attendues malgré une légère baisse de ENR1 à ENR2 (de 64 à 54 chez CTR2, de 63 à 61 chez CTR3), alors que les productions de CTR1 ont été identifiées plutôt comme /ø/ (27 en ENR1 et 42 en ENR2) ou /u/ (27 en ENR1 et 12 en ENR2). Quant aux EXP, EXP3 et EXP4 ont obtenu la quasi-totalité de réponses attendues. Chez EXP1, il y a une amélioration entre ENR1 (28) et ENR2 et 3 (55 et 53, respectivement), alors que EXP2

montre des valeurs moins élevées (25, 33 et 19), avec des réponses pour /ø/ aussi nombreuses (33, 28 et 29). Notons cependant que 4 apprenantes (CTR2, CTR3, EXP3 et EXP4) ont obtenu la majorité de réponses « attendues » dès ENR1, avec des scores moyens supérieurs à 2,5/3, sauf CTR3 (voir section 3.3).

3.3 Scores de réponses attendues (bonne identification)

Nous nous concentrons maintenant sur les scores obtenus pour les réponses attendues pour les voyelles cibles /u/ et /y/, afin de savoir s'il y a une amélioration de qualité de production (FIGURE 2). Tout d'abord, il n'y a pas d'effet significatif de l'auditeur sur les résultats ($F_{(15, 592)} = 1{,}27$, $p = 0{,}21$). En ENR1, une seule EXP a produit le /u/ obtenant une note moyenne supérieure à 2 (sur 3) ; en ENR2, le /u/ de 2 EXP (EXP 2 et 4), et en ENR3, celui de 3 EXP (EXP 2, 3, 4) atteint une note supérieure à 2. Celui des CTR ne dépasse pas 1,3. Quant à /y/, 2 EXP (EXP 3 et 4) ont obtenu un score supérieur à 2,5 dans toutes les séances d'enregistrement, mais une seule CTR (CTR2) atteint à ce niveau seulement en ENR1.

Entre les séances d'enregistrement, nous n'observons aucune amélioration chez les CTR, même s'il y a des baisses de score en ENR2 par rapport à ENR1. En revanche, les EXP montrent une amélioration progressive pour le /u/ sauf EXP1, et pour le /y/, soit une amélioration progressive (EXP1), soit un effet plafond (EXP3 et 4), sauf EXP2.

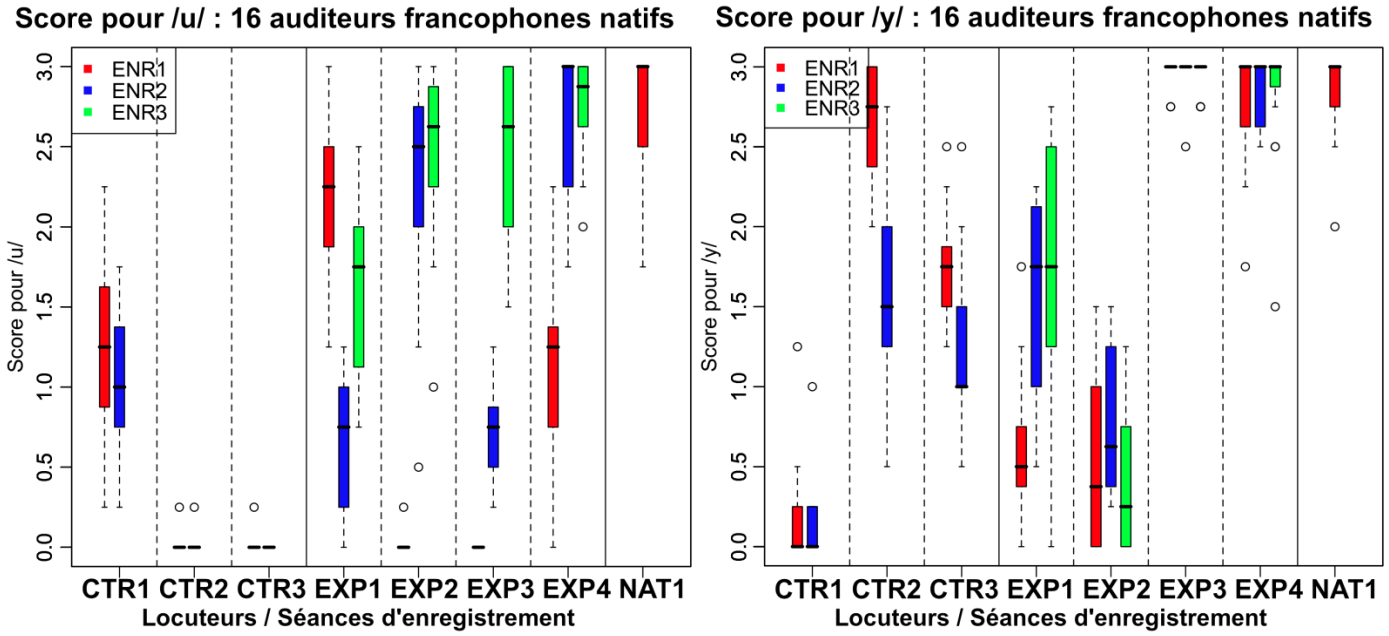


FIGURE 2 – Score moyen (0-3) pour la réponse /u/ obtenue par les stimuli de /u/ (gauche) et pour la réponse /y/ obtenu pour /y/ (droite). De gauche à droite : ENR1, ENR2, ENR3 pour chaque locutrice.

L'ANOVA mesure réptétée montre un effet significatif du groupe de locuteur (CTR/EXP/NAT : $F_{(2, 598)} = 94{,}61$, $p < 0{,}0001$), de la voyelle cible ($F_{(1, 598)} = 40{,}73$, $p < 0{,}0001$) et de la séance d'enregistrement ($F_{(1, 598)} = 33{,}43$, $p < 0{,}0001$) sur les scores obtenus.

3.3.1 Groupe EXP : comparaison entre les séances d'enregistrements

3.3.1.1 Voyelle cible /u/

Les scores moyens de chaque séance d'enregistrement sur les 4 apprenantes (ENR1 : 0,85 ; ENR2 : 1,60 ; ENR3 : 2,32) étaient significativement différents (ENR1- ENR2 : test t de Student : $t_{(63)} = 3,91$, $p = 0,0001$; ENR2- ENR3 : $t_{(63)} = 6,93$, $p < 0.0001$).

3.3.1.2 Voyelle cible /y/

Le score moyen de ENR2 sur les 4 apprenantes (2,04) était significativement différent de celui de ENR1 (1,74 : $t_{(63)} = 4,58$, $p < 0,0001$, selon le test t apparié de Student), mais celui de ENR3 (1,97) ne l'était pas de celui de ENR2 ($t_{(63)} = 1,24$, $p = 0,22$). Celui de ENR3 était significativement supérieur à celui de ENR1 ($t_{(63)} = 2,83$, $p = 0,006$).

3.3.2 Comparaison entre les EXP et CTR

Les différences de scores entre ENR1 et ENR2 (« amélioration » = ENR2- ENR1) ont été comparées entre les EXP et CTR.

L'amélioration moyenne du /u/ sur toutes les apprenantes de chaque groupe était de -0,10 pour les CTR et de 0,75 pour les EXP, ce qui correspond à une différence significative (test t non-apparié de Student : $t_{(110)} = 3,75$, $p = 0,0003$). Celle du /y/ était de -0,54 pour les CTR et de 0,30 pour les EXP. La différence est également significative ($t_{(110)} = 7,83$, $p < 0,0001$).

3.4 Corrélation avec les données acoustiques

La FIGURE 3 montre la distribution des scores de réponses attendues pour les voyelles cibles /u/ et /y/ en fonction des données acoustiques.

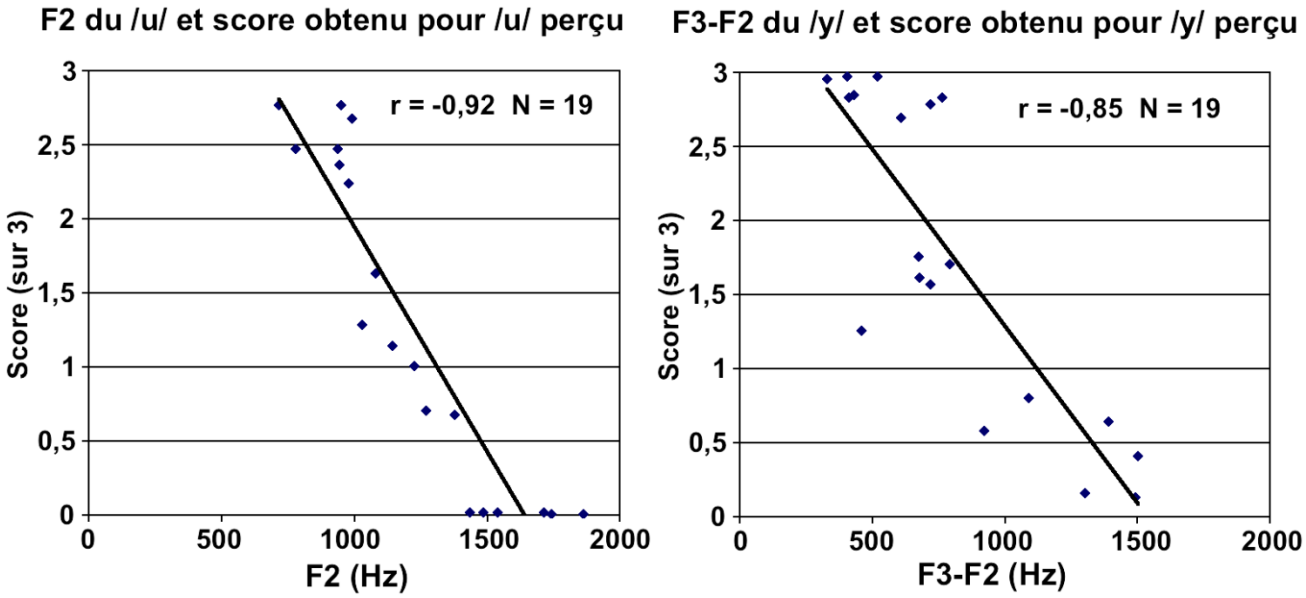


FIGURE 3 – Corrélation entre : gauche : la réponse /u/ obtenue par les stimuli de la voyelle cible /u/ (ordonnée) et F2 moyen des stimuli correspondants (abscisse) (gauche) ; droite : la réponse /y/ que ceux de la voyelle cible /y/ ont obtenu (en ordonnée) et la distance moyenne entre F2 et F3 des stimuli correspondants (en abscisse) (droite).

Les scores pour /u/ que les stimuli du /u/ ont obtenus montrent une corrélation hautement significative avec la moyenne du F2 des stimuli correspondants (coefficient de corrélation de Pearson : $r = -0,92$, $N = 19$, $p < 0.001$). Une corrélation également significative est observée entre les scores pour /y/ que les stimuli du /y/ ont obtenus et la distance entre F2 et F3 des stimuli correspondants ($r = -0,85$, $N = 19$, $p < 0.001$).

Ces résultats corroborent les caractéristiques acoustiques des voyelles isolées du français chez les locuteurs natifs décrites dans la littérature (F1 et F2 proches et bas pour /u/ ; F2 et F3 proches pour /y/ : Liénard 1977, Vaissière 2007, entre autres ; notons qu'une petite distance F1/F2 pourrait également être atteinte par des F1 et F2 plus élevés, dans lequel cas les auditeurs francophones percevraient [o], [ɔ], ou encore [ɑ]).

4 Discussion et conclusion

L'expérience d'évaluation auditive présentée ci-dessus montre que les apprenantes EXP avaient tendance à obtenir un meilleur score après l'entraînement par ultrason (ENR 2 et 3), alors que la même tendance n'a pas été observée chez les CTR en ENR2. Ces données suggèrent que le retour ultrason est utile pour améliorer la production de /u/-/y/ du français, mais tout particulièrement pour la voyelle postérieure /u/, dont la position de la langue est invisible.

EXP1 présente des données intéressantes : le score du /u/ diminue de ENR1 (2,23 de moyenne) à ENR2 (0,7), alors que celui du /y/ augmente entre ces deux séances (de 0,64 à 1,56) ; en ENR3 elle obtient 1,63 pour /u/ et 1,7 pour /y/. Cela suggère qu'à la fin, cette apprenante a amélioré la production des deux voyelles en les distinguant mieux, ce qui est confirmé par les données articulatoires d'ultrason qui montrent que les positions de langue pour les deux voyelles sont séparées de plus en plus nettement d'une séance d'enregistrement à l'autre (locutrice CS dans Pillot-Loiseau *et al.* 2013).

Il sera donc nécessaire de réexaminer plus en détail les liens entre les données articulatoires ultrason, les données acoustiques et les résultats de l'évaluation auditive.

Le présent article traite uniquement des voyelles isolées, mais des analyses acoustiques des logatomes sont présentées dans la première partie de cette étude, faisant l'objet d'un autre article de ces actes (Pillot-Loiseau et al.).

Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-10-LABX-0083.

Références

BERNHARDT, B. M., BACSFALVI, P., ADLER-BOCK, M., SHIMIZU, R., CHENEY, A., GIESBRECHT, N., O'CONNELL, M., SIRIANNI, J., RADANOV, B. (2008). Ultrasound as visual feedback in speech habilitation: Exploring consultative use in rural British Columbia, Canada. *Clinical Linguistics & Phonetics*, 22(2): 149–162.

BOERSMA, P., WEENINK, D. (2013). Praat: doing phonetics by computer [logiciel]. Version 5.3.45, retirée en avril 2013 de http://www.praat.org/

BOTHOREL, A., SIMON, P., WIOLAND, F., ZERLING, J.P. (1986). *Cinéradiographie des voyelles et consonnes du français*. Strasbourg, Publications de l'Institut de Phonétique de Strasbourg.

GUION, S. G., FLEGE, J. E., AKAHANE-YAMADA, R., PRUITT, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America*, 107(5): 2711–2724.

KAMIYAMA, T., VAISSIÈRE, J. (2009). Perception and production of French close and close-mid rounded vowels by Japanese-speaking learners. *Acquisition et interaction en langue étrangère Aile... Lia*, 2: 9-41.

LIÉNARD, J.-S. (1977). *Les processus de la communication parlée : Introduction à l'analyse et la synthèse de la parole*. Masson.

PILLOT-LOISEAU, C., KOCJANČIČ ANTOLÍK, T. & KAMIYAMA, T. (2013). Contribution of ultrasound visualisation to improving the production of the French /y/-/u/ contrast by two Japanese learners. Poster présenté au Workshop *Phonetics, phonology and languages in contact : Contact varieties, multilingualism, and second language learning*, Paris, France.

UEMURA, Y., TAKADA, M. (Kokuritsu Kokugo Kenkyûjo [National Institute for the Japanese Language]) (1990). *Nihongo no boin, shiin, onsetsu: chouon undou no jikken-onseigakuteki kenkyuu* [Vowels, consonants and syllables in Japanese: experimental phonetics study in articulatory movements]. Tokyo, Shûei shuppan.

VAISSIÈRE, J. (2007). Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. In SOLÉ, M. J., BEDDOR, P. S., OHALA, M. (eds), *Experimental Approaches to Phonology*, p. 54-71, Oxford University Press.

Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français

Fethi Bougares[1] Anthony Rousseau[1] Paul Deléglise[1]
Yannick Estève[1] Loïc Barrault[1] Holger Schwenk[1]
Sylvie Brunessaux[2] Khaled Khelif[2] Mathieu Manta[3]
Laboratoire d'Informatique de l'Université du Maine[1] prénom.nom@lium.univ-lemans.fr
Airbus Defence and Space[2] prénom.nom@cassidian.com
Direction Générale de l'Armement[3] mathieu.manta@intradef.gouv.fr

RÉSUMÉ

Le pashto fait partie des langues peu dotées caractérisées par un manque remarquable d'outils de traitement automatique. Ce travail présente une première expérience dans le LIUM avec la langue pashto. Dans cet article, nous présentons les résultats de nos premières expériences de traduction de la parole en pashto/français. La traduction de la parole pashto est réalisée en deux étapes : dans un premier temps, un système de transcription de la parole est utilisé pour transcrire de la parole. La sortie de la transcription est par la suite traduite avec un système de traduction statistique. Nous décrivons dans cet article les développements des différents modèles de chaque système. Ces travaux ont été réalisés en collaboration avec Airbus Defence and Space (anciennement Cassidian) dans le cadre du projet TRAD financé par la DGA.

ABSTRACT

Development and Evaluation of a Pashto/French Spoken Language Translation System

This paper presents an overview of the development of a Pashto spoken language translation system. We present a number of challenges encounter the development of this system highlighting the lack of Pashto resources and tools of all sorts. Based on open source tools and in collaboration with Airbus Defence and Space (formerly Cassidian) under the TRAD project funded by the French Ministry of Defence (DGA), a speech recognition and machine translation systems are developed, described and evaluated in this work.

MOTS-CLÉS : Traduction automatique de la Parole, Langue peu dotée, Pashto.

KEYWORDS: Spoken Language Translation, Under-resourced Language, Pashto.

1 Introduction

L'article 40 de la déclaration universelle des droits linguistiques signé à Barcelone en juin 1996 par l'UNESCO et plusieurs organismes gouvernementaux affirme depuis 1996 que «*Toute communauté linguistique a le droit de disposer, dans le domaine de l'informatique, d'équipements adaptés à son système linguistique et d'outils de production dans sa langue, afin de profiter pleinement du potentiel qu'offrent ces technologies pour l'auto-expression, l' éducation, la communication, l'édition, la traduction, et en général le traitement de l'information et de la diffusion culturelle.*»

Malgré cette affirmation, et malgré les nombreuses initiatives organisées pour informatiser des langues peu dotées, toutes ne sont pas égales devant le processus d'informatisation et beaucoup de langues restent aujourd'hui mal dotées ou avec peu de ressources linguistiques informatisées. Le manque de telles ressources et les caractéristiques méconnues de ces langues rendent le développement de systèmes de traitement automatique difficile.

Dans cet article, nous présentons le développement et l'analyse d'un système de traduction automatique de la parole pour une langue particulièrement peu dotée : le pashto. Le système consiste à transcrire, dans un premier temps, automatiquement des enregistrements audio des conversations en Pashto et de les traduire, ensuite vers le français avec un système de traduction statistique. Dans la section 2, nous décrivons les particularités de la langue pashto ainsi que les difficultés liées à son traitement automatique. La section 3 est réservée à la description du développement de système de transcription et son développement. Avant de conclure, la section 4 présente le système de traduction statistique et les résultats obtenus.

2 Particularités et traitement automatique du pashto

Le pashto est une langue parlée par environ quarante-cinq à cinquante millions de personnes dans le monde. Elle constitue l'une des deux langues officielles de l'Afghanistan, particulièrement parlée par le peuple *pachtoune* dans le sud, l'est et le sud-ouest du pays, ainsi que dans le nord et le nord-ouest du Pakistan mais également en Iran, au Tadjikistan et en Inde (Uttar Pradesh et Cachemire).

2.1 Particularités linguistiques

Le pashto appartient à la famille des langues indo-iraniennes, basée sur les lettres de l'alphabet arabe, augmenté par des lettres supplémentaires empruntées aux alphabets perse et urdu. Son alphabet est composé de quarante-quatre lettres, auxquelles s'ajoutent quatre marques diacritiques. Le pashto s'écrit de droite à gauche et un mot peut se composer de plusieurs éléments (proclitique, préfixe, base, suffixe et enclitique). De plus, cette langue a la particularité de se décliner en quatre dialectes, selon la région d'origine du locuteur, ajoutant une difficulté supplémentaire à son traitement automatique. Cependant, la majeure difficulté provient du fait que le pashto est une langue très peu dotée en ressources linguistiques : par exemple, il n'existe pas de corpus pashto dans les catalogues de LDC [1].

2.2 Particularités de traitement

Pour l'écriture du pashto, les lettres changent de glyphe selon leur position dans le mot. Chaque lettre a donc une représentation graphique différente en fonction de sa position. Dans le cadre d'encodage électronique de texte, comme pour l'arabe, chaque lettre doit être idéalement encodée

1. www.ldc.upenn.edu

par la valeur principale Unicode correspondante, indépendante de la position de la lettre et donc du glyphe affiché. Par exemple, le bloc de valeurs principales Unicode pour l'écriture de l'arabe est compris entre U+FB50 et U+FDFF et entre U+FE70 et U+FEFF. Dans ces conditions, c'est à la charge du moteur de rendu graphique de gérer l'affichage du bon glyphe correspondant au contexte d'un caractère.

Ce point est très important car la pluralité des sources de texte rend ces conditions difficiles à respecter. Ainsi, il est très fréquent d'observer dans les données rencontrées des lettres encodées non pas par leur valeur Unicode principale, indépendante du glyphe, mais directement par la valeur du glyphe affiché. Ainsi, un même mot d'un même corpus peut être représenté électroniquement par plusieurs séquences de valeurs différentes. Il s'agit là d'un problème crucial car fréquent dans les données, entraînant du bruit dans les modèles de langage, perturbant le processus de phonétisation et biaisant les évaluations.

Une difficulté se présente également, similaire à celle que l'on rencontre dans le traitement de l'arabe : un mot n'est pas transcrit avec ses voyelles et, pour une même représentation orthographique, peut se prononcer différemment en fonction du contexte lexical dans lequel apparaît la suite de consonnes le représentant. Cette représentation orthographique peut aussi faire référence à des mots ayant des sens différents.

3 Reconnaissance automatique de la parole en pashto

3.1 Ressources linguistiques

Les données accessibles au LIUM pour l'apprentissage des différents modèles du système de reconnaissance automatique de la parole (SRAP) en pashto se composent de corpus produits dans le cadre du projet TRAD par ELDA (Mostefa *et al.*, 2012) et par Appen Butler Hill. Dans l'objectif de représenter au mieux la variabilité dialectale entre locuteurs, le corpus utilisé pour le SRAP est composé des enregistrements de locuteurs (hommes et femmes) provenant de trois différentes régions (Afghan du sud, Afghan du nord et pakistanais).

Ces données sont de plusieurs types et sources :

- des données textuelles issues de pages Web pour l'apprentissage des modèles de langage (ensemble web produit par ELDA) ;
- des enregistrements audio de journaux télévisés et leur transcriptions manuelles (ensemble H4 produit par ELDA) ;
- des enregistrements audio de la parole conversationnelle et leur transcriptions manuelles (ensemble H5 produit par Appen Butler Hill).

L'objectif étant de créer un système de traduction de la parole conversationnelle, nous avons extrait un corpus de développement (dev-H5) du corpus H5. Le tableau 1 présente les caractéristiques de ces données.

En plus de ces données d'apprentissage et de développement, afin de construire notre dictionnaire, nous disposons de la phonétisation manuelle de 9000 des mots les plus fréquents en pashto. Cette fréquence a été mesurée sur les transcriptions de parole à notre disposition. Néanmoins, ce dictionnaire ne contient que la prononciation canonique d'un mot, représenté sous la forme

Corpus	Durée de parole (h)	Nombre de segments/phrases	Nombre de mots	Durée moyenne d'un segment (s)
H4	86,06	42 095	967,3k	7,36
H5	31,67	31 038	407,5k	3,67
web	N/A	1 202 839	108,97M	N/A
dev-H5	0,97	877	N/A	3,98

TABLE 1 – Caractéristiques des corpus utilisés pour l'apprentissage et l'optimisation des SRAP en pashto.

d'une séquence de consonnes.

3.2 Choix du système principal de reconnaissance automatique de la parole

Le LIUM a choisi de travailler principalement sur la base du décodeur Kaldi [2] pour développer ses SRAP. Deux raisons nous ont amenés à faire ce choix :

— Kaldi est un système très performant et à l'état de l'art quant à la modélisation acoustique. Le LIUM l'a déjà utilisé avec succès, combiné avec Sphinx, dans le cadre de la campagne d'évaluation interne au défi ANR REPERE [3] en janvier 2013 ;
— dans le cadre de nos expériences, il était nécessaire de développer un SRAP fonctionnant en conservant une latence faible. Le système Kaldi a ceci de particulier puisqu'il est basé sur une approche par transducteurs qui implique que l'espace de recherche est calculé avant le décodage : cela lui permet d'être plus rapide que la plupart des SRAP d'une autre nature. Le prix de cette vitesse est une consommation mémoire importante, qui peut toutefois être maitrisée par un vocabulaire de taille limitée. La vitesse de traitement de Kaldi a été un critère déterminant ;

3.3 Phonétisation du Pashto

Afin de disposer d'un dictionnaire contenant un nombre suffisant d'entrées pour l'apprentissage d'un SRAP, nous avons utilisé les 9000 mots phonétisés manuellement comme corpus d'apprentissage pour un système de phonétisation automatique, basé sur une approche probabiliste. Le système utilise l'outil Moses pour apprendre un système de traduction statistique avec comme corpus d'apprentissage bilingue les mots avec leur phonétisation. L'idée est d'utiliser le paradigme de la traduction automatique pour traduire un mot vers sa phonétisation, la méthode est détaillée dans (Laurent *et al.*, 2009). Dans nos expérimentations, sur un sous-ensemble des 9000 mots de départ, nous avons relevé que le taux de mots correctement phonétisés était de 83.29% (qui représente le score BLEU de la traduction mots-phonèmes).

À l'aide cet outil, le dictionnaire que nous avons constitué est composé des 40700 mots les plus fréquents rencontrés dans nos transcriptions de parole : nous avons donc phonétisé automatiquement

2. http ://kaldi.sourceforge.net/about.html
3. http ://www.defi-repere.fr

près de 32000 mots.

Pour pallier la difficulté induite par le manque de voyelles dans les transcriptions, nous avons relâché les contraintes utilisées lors d'un alignement forcé des corpus d'apprentissage en autorisant plusieurs voyelles à être alignées dans certains contextes. Puis, en analysant le résultat du processus d'alignement automatique, nous avons injecté des variantes de prononciation observées dans le corpus d'apprentissage dans notre dictionnaire.

3.4 Modélisation du langage

Les modèles de langage ont été estimés sur les données H4 et H5 à notre disposition, en plus des données textuelles extraites depuis le Web.

3.4.1 Modèles de langage H4

Pour le type de données H4, le modèle de langage principal est un modèle trigramme. Nous avons expérimenté l'utilisation d'un modèle quadrigramme, mais il dégradait les performances mesurées par un calcul de perplexité sur le corpus de développement. Nous pensons que cela provient du manque de données proches des données H4 dans les données provenant d'Internet.

Pour l'estimation de ce modèle, nous avons d'abord calculé un modèle de langage par source (H4, H5 et web) avec le vocabulaire de 40700 mots. Ensuite, une interpolation linéaire a été réalisée à l'aide de coefficients calculés de manière à minimiser la valeur de perplexité du modèle interpolé sur le corpus de développement H4. La perplexité du modèle final calculée sur ce corpus de développement est de 176 (185 pour le modèle quadrigramme). Un modèle de langage bigramme, pour le décodage principal, a également été estimé de manière similaire. Il obtient une valeur de perplexité égale à 236.

3.4.2 Modèles de langage H5

La même procédure a été suivie pour le modèle de langage H5 qui est lui aussi un modèle trigramme. Le vocabulaire utilisé contient 13244 mots : nous avons restreint la taille du vocabulaire en raison de la faible taille du corpus d'apprentissage et de l'application beaucoup plus contrainte d'un point de vue sémantique. Ce vocabulaire est composé de tous les mots présents dans le corpus d'apprentissage et le corpus de développement H5, plus les 5000 mots les plus fréquents du corpus d'apprentissage H4 absents du corpus d'apprentissage H5. Les trois sources (H4, H5 et web) ont été utilisées pour le modèle trigramme. Sur le corpus de développement H5, la valeur de perplexité obtenue est de 128. Un modèle de langage bigramme a également été estimé qui obtient une valeur de perplexité égale à 134.

3.5 Modélisation acoustique et décodage

Au cœur des SRAP utilisés pour la reconnaissance du pashto, les modèles acoustiques se présentent sous la forme classique de modèles de Markov cachés (HMM), dont chaque état est associé à une

mixture de gaussiennes. Chaque HMM modélise un phonème en contexte (contexte phonétique gauche, contexte phonétique droit, position dans le mot).

Les modèles acoustiques ont été estimés sur les données audio en pashto avec transcriptions manuelles disponibles. Pour leur apprentissage, nous avons appliqué plusieurs approches successives : d'abord une estimation LDA (Linear Discriminant Analysis) (Haeb-Umbach et Ney, 1992) avec transformation MLLT (Maximum Likelihood Linear Transform), puis une approche SAT (Speaker Adaptative Training) (McDonough *et al.*, 2002) avec transformation fMLLR (Maximum Likelihood Linear Regression) et enfin une estimation fMMI (Minimum Mutual Information).

Le SRAP développé est basé sur un noyau issu du système Kaldi (Povey *et al.*, 2011) pour le décodage principal qui génère un graphe de mots. Ce décodage principal s'effectue en deux étapes : un premier décodage, indépendant du locuteur, qui permet d'obtenir une première transcription. Cette transcription est utilisée pour calculer des matrices de transformation fMLLR qui permettent d'adapter le décodage acoustique à la voix du locuteur et aux conditions acoustiques. Le second décodage exploite ces matrices et génère des graphes de mots. Ces deux décodages utilisent un modèle bigramme plutôt qu'un modèle trigramme afin d'accélérer le traitement en réduisant la taille de l'espace de recherche. Chaque graphe de mots généré par le noyau Kaldi est ensuite traité pour réévaluer les scores linguistiques à l'aide d'un modèle trigramme.

4 Traduction automatique de la parole pashto

La présence d'une grande quantité de données d'apprentissage adéquates ainsi que l'utilisation de logiciels libres permettent la construction rapide de systèmes de traduction automatique produisant de bons résultats. Toutefois, la traduction vers ou à partir de langues peu dotées doit faire face au manque de données linguistiques disponibles. De plus, la majorité des langues peu dotées ne disposent pas d'outils de traitement linguistique dédiés comme par exemple les analyseurs morphologiques qui se montrent utiles dans le cadre de la traduction de langues morphologiquement riches. Dans l'objectif de surmonter ce manque de ressources linguistiques, plusieurs travaux s'orientent vers la création de données de manière non supervisée en exploitant par exemple des données comparables. Dans cette section, nous présentons notre premier système de traduction pour la paire de langues pashto-français.

4.1 Données d'apprentissage

La construction manuelle d'un corpus bilingue est généralement réalisée par des locuteurs natifs de la langue cible (la langue française dans notre cas). Dans notre contexte, il est très difficile de trouver un locuteur français natif maitrisant la langue pashto. C'est pour cette raison que les corpus d'apprentissage ont été traduits du pashto par des locuteurs natifs (langue source) et révisés ensuite par des locuteurs natifs français (langue cible) (Mostefa *et al.*, 2012).

Le système de traduction a été entraîné à partir de données distribuées par ELDA et Appen Butler Hill. Deux corpus bilingues ont été utilisés. Ces données représentent la traduction des corpus pashto H4 et H5 transcrits par des locuteurs natifs et présentés dans la section 3.1.

Dans l'objectif d'augmenter la couverture des données d'apprentissage, nous avons ajouté deux corpus bilingues. Le premier compte 5737 traductions d'entités nommées (nom de personnes

et de villes) et le second contient 737 conversations usuelles en langue pashto ainsi que leurs traductions françaises. Ces deux corpus sont téléchargeables gratuitement sur le site du LIUM [4].

La Table 2 présente la taille, en nombre de phrases, de l'ensemble des données bilingues utilisées pour l'apprentissage et le développement du système de traduction.

Corpus	Taille (#phrases)
H4	33k
H5	29K
names	5737
Web-unsup	737

TABLE 2 – Corpus bilingues utilisés pour construire et développer le système de traduction pashto-français

La Table 3 présente les données monolingues utilisées pour apprendre le modèle de langage en langue cible (la langue française en l'occurrence). En plus de la partie cible des corpus bilingues, nous avons utilisé deux corpus libres et téléchargeables sur Internet : le corpus OpenSubtitles2012 [5] et le corpus de Wikipédia [6].

Corpus	(#phrases)
H4-fr	33k
H5-fr	29K
wiki-extract	13M
OpenSubtitles-2012	4,3M

TABLE 3 – Corpus monolingues utilisés pour le système de traduction pashto-français

Les données d'apprentissage (bilingues et monolingues) ont fait l'objet d'un pré-traitement pour s'adapter au style de la sortie du SRAP. Ce pré-traitement comprend, entre autres, la suppression de la ponctuation et de la casse, ainsi que la conversion des chiffres en lettres.

4.2 Système de traduction

Le système de traduction développé est un système à base de segments (*phrase-based*). La table de traduction et le modèle de réordonnement sont estimés à l'aide des outils issus du décodeur Moses [7] (Koehn *et al.*, 2007). Le modèle de langage est un modèle quadrigramme de repli Kneyser-Ney modifié obtenu après une interpolation des différents modèles présentés dans la table 3 à l'aide des outils SRILM (Stolcke, 2002). La perplexité de chaque modèle sur le corpus de développement ainsi que les cœfficients d'interpolation sont présentés dans la Table 4. La tokénisation des corpus est réalisée via l'outil inclus dans le décodeur Moses et les textes parallèles ont été alignés au niveau des mots avec l'outil mGiza. Enfin, les paramètres du système (14 paramètres du modèle log-linéaire) ont été optimisés avec l'algorithme MERT (Och, 2003) sur le corpus de développement.

4. http://www-lium.univ-lemans.fr/~bougares/personnel.php
5. http://www.opensubtitles.org/fr
6. http://dumps.wikimedia.org
7. http://www.statmt.org/moses/

Corpus	Perplexité	Cœfficient
H5-fr	61,88	0,66
H4-fr	130,23	0,09
wiki-extract	102,65	0,19
OpenSubtitles2012	149,37	0,06
Interpolation	49,57	n/a

TABLE 4 – Perplexité sur le corpus de développement et interpolation des modèles de langage estimés sur chacun des corpus monolingues.

4.3 Performances du système

La table 5 présente le taux d'erreur mots de système de transcription et les scores BLEU obtenus par le système de traduction selon le type de l'entrée de traduction.

Entrée de traduction	WER	BLEU
Transcription de référence	n/a	17,78
Transcription automatique+audio-dev	25,9	n/a
Transcription automatique	36,0	13,22

TABLE 5 – Score BLEU de la traduction de différent type de texte

Le système a été utilisé dans un premier temps pour traduire la transcription manuelle du corpus de DEV. Cette première expérience montre le score maximum qu'on pourrait avoir avec une transcription parfaite. La traduction de la transcription de référence donne une sortie avec un score BLEU de 17.78 points. Le système a été utilisé par la suite pour traduire la sortie de système de transcription. Le tableau montre bien la corrélation entre la qualité de la transcription et celle de la traduction. En effet, la traduction d'une transcription automatique de 36% de WER réduit le score BLEU de 4.56 points absolus. Dans l'objectif de mesurer l'effet de manque de données sur la qualité de la transcription, nous avons augmenté le corpus d'apprentissage acoustique en ajoutant le corpus de DEV (le modèle de langage n'a pas été modifié). L'intégration de ces données permet une réduction absolue de WER 10,1 point ce qui montre le manque évident de données d'apprentissage de modèle acoustiques.

Conclusion

Dans cet article nous avons présenté un premier investissement du LIUM dans le domaine de traitement automatique des langues peu dotées. Nous avons développé un système de traduction automatique de la parole pour le pashto dans le cadre du projet TRAD en collaboration avec Airbus Defence and Space pour le compte de la DGA. Ce travail présente les premiers résultats qui montrent que l'application des technologies de traitement des langues à l'état de l'art aux langues peu dotées ne permet pas d'avoir des performances équivalentes à leur utilisation avec les langues ayant beaucoup de ressources numériques.

Références

HAEB-UMBACH, R. et NEY, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. *In Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 13–16, Washington, DC, USA. IEEE Computer Society.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. *In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

LAURENT, A., DELÉGLISE, P. et MEIGNIER, S. (2009). Grapheme to phoneme conversion using an SMT system. *In INTERSPEECH*, pages 708–711.

MCDONOUGH, J. W., SCHAAF, T. et WAIBEL, A. (2002). On Maximum Mutual Information speaker-adapted training. *In ICASSP*, pages 601–604. IEEE.

MOSTEFA, D., CHOUKRI, K., BRUNESSAUX, S. et BOUDAHMANE, K. (2012). New language resources for the pashto language. *In LREC 2012*.

OCH, F. J. (July 2003). Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., GLEMBEK, O., GOEL, N., HANNEMANN, M., MOTLICEK, P., QIAN, Y., SCHWARZ, P., SILOVSKY, J., STEMMER, G. et VESELY, K. (2011). The kaldi speech recognition toolkit. *In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

STOLCKE, A. (2002). SRILM — an extensible language modeling toolkit. *In in Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA,.

Détection et Caractérisation d'erreurs dans des transcriptions automatiques pour des systèmes de traduction parole-parole

Frédéric Béchet, Benoit Favre
LIF/CNRS, Aix-Marseille Université, 163 avenue de Luminy, Marseille, France
prenom.nom@lif.univ-mrs.fr

Résumé

Quelle que soit la qualité des modules de transcription de la parole, les erreurs de reconnaissance sont inévitables étant donné d'une part les ambiguïtés intrinsèque à toute langue naturelle et d'autre part aux limites technologiques des systèmes actuels (vocabulaire fermé, séquentialité des traitements). L'étude et la caractérisation des erreurs de transcription représente un champ d'étude à part entière avec comme finalité d'atténuer l'impact de ces erreurs sur tout module exploitant des transcriptions automatiques. Dans le cadre d'un système de traduction parole-parole cette étude présente un module de détection et de caractérisation d'erreurs basé sur un étiqueteur de séquence utilisant des indices acoustiques, lexicaux et syntaxiques.

Abstract

Even though small ASR errors might not impact downstream processes that make use of the transcript, larger error segments like those generated by OOVs can have a considerable impact on applications such as speech-to-speech translation and can eventually lead to communication failure between users of the system. This work focuses on error detection in ASR output targeted towards significant error segments that can be recovered using a dialog system. We propose a CRF system trained to recognize error segments with ASR confidence-based, lexical and syntactic features.

Mots-clés : Traduction parole-parole, mesures de confiance, détection d'erreurs, analyse linguistique.

Keywords: Speech-to-speech translation, confidence measures, error detection, linguistic analysis.

1 Introduction

Les performances actuelles des systèmes de Reconnaissance Automatique de la Parole ont permis le développement de systèmes de communication innovants entre humains et machines tels que les assistants personnels commandés par la voix ou bien les traducteurs vocaux permettant à deux interlocuteurs de mener une conversation orale, chacun dans une langue différente.

Ces systèmes comportent tous un module de transcription automatique de la parole permettant de traduire le signal vocal en texte pouvant servir d'entrée au module d'interprétation sémantique, pour les systèmes de communication humain-machine, ou bien au module de traduction automatique pour les systèmes de traduction parole-parole. Cependant, quelle que soit la qualité du module de transcription, les erreurs de transcriptions sont inévitables étant donné d'une part les ambiguïtés intrinsèque à toute langue naturelle et d'autre part les limites technologiques des systèmes actuels (vocabulaire fermé, séquentialité des traitements).

Pouvoir détecter ces erreurs est crucial pour les traitements successifs des systèmes de transcription afin d'éviter de les propager au niveau de l'interprétation ou de la traduction. Cependant toutes les erreurs n'ont pas le même impact sur les performances globales d'un système. Ainsi une erreur d'accord, une confusion entre plusieurs formes conjuguées d'un même verbe ou bien l'ajout ou l'omission d'un mot grammatical tel que déterminant ou préposition, peuvent avoir un impact sur la structure syntaxique d'un énoncé, sans pour autant affecter de manière notable les performances. En effet les modules linguistiques utilisant les transcriptions automatiques sont entraînés à pouvoir traiter des énoncés bruités de ce type. En revanche, faire une confusion sur un syntagme prédicatif ou sur un groupe nominal peut avoir un impact très fort sur le système global.

Les mesures de confiance données par les systèmes de transcription automatique, telles que les probabilités a posteriori, ne prennent pas en compte cette notion d'impact. L'étude menée dans cet article consiste à proposer des indices supplémentaires permettant de mieux détecter et caractériser ces erreurs de transcription, dans le cadre d'un système de traduction parole-parole.

Le contexte expérimental de cette étude est celui de la tâche de traduction parole/parole du projet DARPA BOLT (Ayan et al., 2013). Cette tâche a pour but d'améliorer les systèmes de traduction parole/parole en permettant à la machine, en cas d'ambiguïté détectée lors de la transcription ou de la traduction, d'engager un dialogue de clarification avec l'utilisateur.

Cet article est organisé de la manière suivante : le paragraphe 2 présente des travaux connexes à cette étude dans le domaine de la détection d'erreurs et de mots hors-vocabulaire ; le paragraphe 3 présente la tâche visée dans le cadre du projet BOLT ; le paragraphe 4 détaille le modèle de détection d'erreurs développé qui est évalué dans le paragraphe 5.

2 Travaux connexes

Estimer la confiance d'une hypothèse de transcription pose plusieurs problèmes : choisir la granularité de l'erreur à détecter (Hazen et al., 2002) (mot, segment conceptuel, syntagme ou phrase entière), définir l'ensemble de paramètres qui seront intégrés dans le calcul de la

confiance (paramètres directement issus du processus de transcription, indices syntaxiques, indices contextuels), combiner les différents paramètres et mettre au point une stratégie de décision prenant en compte l'ensemble d'entre eux (Sarikaya et al., 2005). Quel que soit le modèle développé, la majorité des approches partage tout du moins deux étapes de base : générer autant de paramètres que possible pour caractériser chaque hypothèse produite ; estimer des scores de confiance à partir de ces paramètres ou indices.

Un problème connexe à celui de la détection d'erreurs de transcription est celui de la détection de segments de mots hors-vocabulaire. En effet, étant donné que tout système de transcription se base sur un vocabulaire fermé, tout mot inconnu de ce vocabulaire va générer nécessairement un segment d'erreur. Les systèmes actuels de détection de ces segments d'erreurs correspondant à des mots hors-vocabulaire sont basés sur des classifieurs prenant en compte un très grand nombre de paramètres provenant de plusieurs étapes de décodage (avec des unités lexicales et sous-lexicales) (Rastrow et al., 2009) ainsi que des indices prosodiques ou syntaxiques (Hirschberg et al., 2004). Ce problème est envisagé sous l'angle d'un problème de classification binaire ou chaque mot (ou segment d'un réseau de confusion) est classé comme mot hors-vocabulaire (Out Of Vocabulary - OOV) ou pas.

Plus récemment des études telles que (Parada et al., 2010) ont proposé de considérer le problème de la détection de segments hors-vocabulaire comme un problème d'étiquetage de séquence. L'intuition derrière cette idée est basée sur le fait que les mots hors vocabulaire ont tendance à générer plusieurs erreurs, d'une part en remplaçant chaque mot inconnu par une séquence de mots connus plus petits, et d'autre part à cause de l'effet "boule de neige" provoqué par ces même mots dans leur contexte immédiat. Une approche à base de Conditional Random Fields (CRF) a été proposée, intégrant tous les indices présentés auparavant, et visant à prédire la meilleure séquence d'étiquettes `B_OOV` (début de segment OOV), `I_OOV` (à l'intérieur d'un segment OOV) et `O_OOV` (en dehors d'un segment OOV) à partir d'une séquence de mots hypothèses. Cette approche a obtenu des gains significatifs en performance comparée à l'approche basée uniquement sur des décisions locales sur les mots.

Nous proposons dans cette étude de généraliser cette méthode à tous les différents types d'erreurs de reconnaissance. Cet article est une extension de (Béchet et Favre, 2013). En se basant sur une approche à base de CRF, nous générons un graphe d'hypothèses de segments d'erreurs permettant d'améliorer la détection de séquences d'erreurs ayant un impact important sur les performances des traitements suivants la phase de transcription de la parole.

3 Erreurs de reconnaissance et traduction parole-parole

Cette étude a été menée dans le cadre de la tâche de traduction parole-parole du projet DARPA BOLT. Le cadre applicatif est la communication entre deux personnes ne parlant pas la même langue, sans support visuel (uniquement à la voix), pour des tâches de collecte d'information. L'originalité de BOLT, par rapport à des projets similaires tels que TRANSTAC, consiste à doter les modules de transcription et de traduction d'une capacité à générer un dialogue de clarification avec l'utilisateur lorsqu'une erreur est détectée (Ayan et al., 2013).

La difficulté principale de ce type de système est d'engager des dialogues de clarification uniquement lorsque les erreurs détectées sont susceptibles de bloquer le dialogue entre les

locuteurs.

Un exemple est donné dans la figure 1.

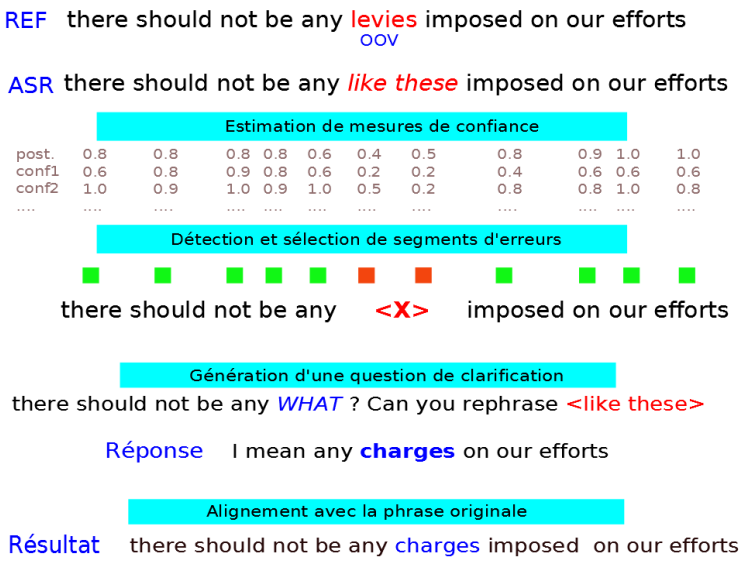


Figure 1 – Exemple de déroulement d'un dialogue de clarification dans BOLT

Cette différence d'importance entre erreurs de reconnaissance est illustrée par la table 1 présentant deux exemples provenant du corpus BOLT avec les hypothèses de transcription (ASR 1-best), les mesures de confiance du système de reconnaissance (ASR posteriors) et la séquence de mots effectivement prononcée par le locuteur (Réference).

Le premier exemple contient un mot hors-vocabulaire "acetaminophen" qui est remplacé par le système de reconnaissance par la séquence de 6 mots : "I see them in a fan". Dans ce cas, il est très important de détecter cette erreur et engager un dialogue de clarification avant la phase de traduction. Il est intéressant de constater qu'au niveau des mesures de confiance, seul le début et la fin du segment erroné ont des valeurs basses, les erreurs de milieu de segment ont toutes des valeurs acceptables. Ce phénomène justifie l'approche à base d'étiquetage de séquence utilisée dans cette étude.

exemple 1													
ASR 1-best	is	the	town	low	on	supplies	of	I	see	them	in	a	fan
ASR posteriors	1	1	1	0.9	1	1	1	0.1	0.9	0.9	0.9	0.8	0.6
Réference	is	the	town	low	on	supplies	of	acetaminophen					

exemple 2									
ASR 1-best	they	travel	to	cross	to	desert	on	camera	back
ASR posteriors	0.7	0.8	0.8	0.8	0.8	0.9	0.9	0.7	0.8
Réference	they	traveled		across	the	desert	on	camelback	

Table 1 – Exemple de transcriptions automatiques avec des segments d'erreurs

Le second exemple contient deux types d'erreurs : des confusions sur des mots proches ("travel/traveled", "cross/across"), une confusion sur un mot monosyllabique ("to/the"), enfin de nouveau une erreur due à un mot hors-vocabulaire ("camera back/camelback").

Sur ces erreurs, deux clarifications seraient nécessaires dans cet exemple avant d'effectuer la traduction : d'une part l'ambiguïté entre "to desert" (déserter) et "the desert" ; d'autre part la clarification sur le mot hors-vocabulaire "camelback".

Il est clair sur ces exemples que les seules mesures de confiance de type probabilités a posteriori ne permettent ni de détecter facilement ces erreurs, ni de les distinguer les unes des autres. C'est pour cela que nous proposons un traitement intégrant à la fois des indices linguistiques, mais aussi une méthode à base d'étiquetage de séquences.

4 Détection d'erreurs par étiquetage de séquence

La méthode de détection de segments d'erreurs présentée dans cette étude est basée sur un étiqueteur de séquence de type Conditional Random Field (CRF). Cet étiqueteur va attribuer un label binaire à chaque mot : erreur ou correct.

Réference

word	we	need	judicious	men
POS	PRP	VBP	JJ	NN
synt. dep.	SBJ(need)	ROOT()	NMOD(men)	OBJ(need)

ASR 1-best

word	we	need	your	dishes	that
POS	PRP	VBP	PRP	NN	WDT
synt. dep.	SBJ(need)	ROOT()	NMOD(dishes)	OBJ(need)	NONE()

Table 2 – Exemple d'analyse linguistique de transcription de référence et d'hypothèses

Cet étiqueteur est appris sur un corpus de transcriptions automatiques sur lesquelles les étiquettes erreur et correct sont obtenues par alignement dynamique avec les transcriptions de référence. Trois niveaux de paramètres sont utilisés pour décrire les exemples du corpus d'apprentissage :

1. Paramètres issus du système de transcription de parole : nous utilisons les probabilités a posteriori générées durant le décodage. Ces probabilités sont discrétisés par la méthode décrite dans (Fayyad et Irani, 1993) avec l'implémentation de (Raymond, 2012). Nous utilisons les valeurs du mot courant, du mot précédent et du mot courant.
2. Paramètres lexicaux : le mot courant et son contexte droit et gauche font partie des paramètres. Ils sont associés à un paramètre relatif à la taille du mot (en nombre de caractères) ainsi qu'à un paramètre indiquant si les 3 trigrammes contenant le mot courant et son contexte immédiat ont été vus dans le corpus d'apprentissage du modèle de langage.
3. Paramètres syntaxiques : les transcriptions sont analysées avec la boite à outils MACAON (Nasr et al., 2011). Nous utilisons un étiqueteur morphosyntaxique ainsi qu'un analyseur en dépendance. Les séquences d'étiquettes syntaxiques sont utilisées comme paramètres, ainsi que les liens de dépendance entre le mot courant et son gouverneur.

Un exemple d'analyse linguistique est donné dans la table 2. Tous les paramètres syntaxiques sont produits à partir de ces analyses.

5 Expérimentation et résultats

Nous utilisons dans cette étude des corpus en langue anglaise provenant du projet DARPA BOLT et transcrit automatiquement par le système de SRI Dynaspeak (Franco et al., 2002). Ce corpus contient deux types de phrases :

– Corpus 1 : cette partie du corpus contient des dialogues enregistrés durant le projet TRANSTAC, contenant des conversations centrées sur la collecte d'information, proche des corpus utilisés pour entraîner le système de reconnaissance. Il y a peu de mots hors-vocabulaire et le taux d'erreurs mots est assez faible ($< 10\%$).
– Corpus 2 : ce corpus a été défini durant le projet BOLT pour contenir spécifiquement des phrases difficiles avec des ambiguïtés provenant soit d'un mot hors-vocabulaire, soit d'une ambiguïté de type homophone, soit d'un mot volontairement mal prononcé. Le taux d'erreur de ce corpus est bien plus élevé (35%).

Ces corpus correspondent ainsi à deux situations différentes : le premier corpus représente les erreurs classiques d'un système de reconnaissance, qui peuvent évidemment perturber le système global, mais qui ne nécessitent pas nécessairement un dialogue de clarification. Le deuxième corpus contient les erreurs pouvant avoir un impact important sur la suite des traitements qui doivent être clarifiées avant de pouvoir lancer la phase de traduction.

Corpus	#mots	#phrases	WER	taille err.	fertilité
Corpus 1	84405	6527	8.4	1.5	1.2
Corpus 2	4919	570	35.8	2.6	4.8

Table 3 – Description des corpus avec leurs caractéristiques

Ces deux corpus sont décrits dans la table 3. Nous remarquons que la taille moyenne des erreurs est bien supérieure pour le corpus 2. La fertilité d'une erreur correspond aux nombres de mots erronés générés par le système de transcription pour un mot mal reconnu dans la transcription de référence. La fertilité moyenne donnée dans le tableau indique ainsi le nombre moyen d'insertion observé dans les transcriptions automatique pour un mot de la référence non reconnu. Nous remarquons que pour le corpus 2, la fertilité est très forte : presque 5 insertions pour chaque erreur. Ceci s'explique par le fait que les mots hors-vocabulaire ont tendance à être remplacés par des séquences de mots courts, comme illustré dans la table 1. Afin de ne pas faire d'a priori sur le type d'erreurs que l'on va rencontrer, les résultats présentés ici ont été obtenus avec une approche de type 10-fold, où les corpus 1 et 2 ont été mélangés, puis découpés en 10 partitions. L'étiqueteur CRF (Okazaki, 2007) est appris sur les 9/10 de chaque partition, puis évalué sur la dixième. Il y a plusieurs méthodes pour évaluer les performances d'un détecteur d'erreurs. La méthode la plus simple consiste à évaluer la prédiction erreur/correct pour chaque mot. Cette métrique est présentée dans la table 4 dans la colonne correct. Etant donné que la plupart des mots sont corrects, cette mesure n'est pas très informative. Il est plus intéressant de se placer dans un contexte de détection d'évènements où les seules évènements à détecter sont les erreurs de reconnaissance et non pas les mots corrects. Nous donnons ainsi les résultats en termes de précision (P), rappel

(R), Fausse ALarme (FA), et détection manquée (Miss) pour les erreurs de reconnaissance uniquement (un système prédisant que tous les mots sont corrects aurait 0% de précision et rappel). La table 4 présente ces résultats séparément pour les deux corpus, bien qu'ils aient été réunis durant la phase d'apprentissage et de test. Plusieurs conditions sont comparées : en utilisant dans le modèle CRF uniquement les probabilités a posteriori, puis en rajoutant les paramètres lexicaux, et enfin les paramètres syntaxiques.

Corpus 1					
Paramètres	correct	P	R	FA	Miss
ASR post.	94.8	51.1	22.0	48.9	78.0
+ lexical	95.1	55.1	32.1	44.9	67.9
+ syntaxique	95.1	55.4	36.8	44.6	63.2
Corpus 2					
Paramètres	correct	P	R	FA	Miss
ASR post.	78.7	76.6	41.5	23.5	58.5
+ lexical	81.3	74.0	57.8	26.0	42.2
+ syntaxique	82.4	74.0	63.7	26.0	36.3

Table 4 – Résultats en terme de classification erreur/non erreur (correct) ; puis de détection des erreurs uniquement sur les deux corpus C1 et C2

Comme on peut le voir les prédictions d'erreurs sont meilleures sur Corpus 2. C'est d'une part attendu, car les erreurs sont plus saillantes, mais d'autre part c'est aussi espéré car il s'agit des erreurs pour lesquelles il faut avoir une stratégie de correction. Les courbes (Miss/FA) sont présentées dans la figure 2. Elles sont obtenues en faisant varier un seuil sur le score de détection d'une erreur donné par l'étiqueteur CRF.

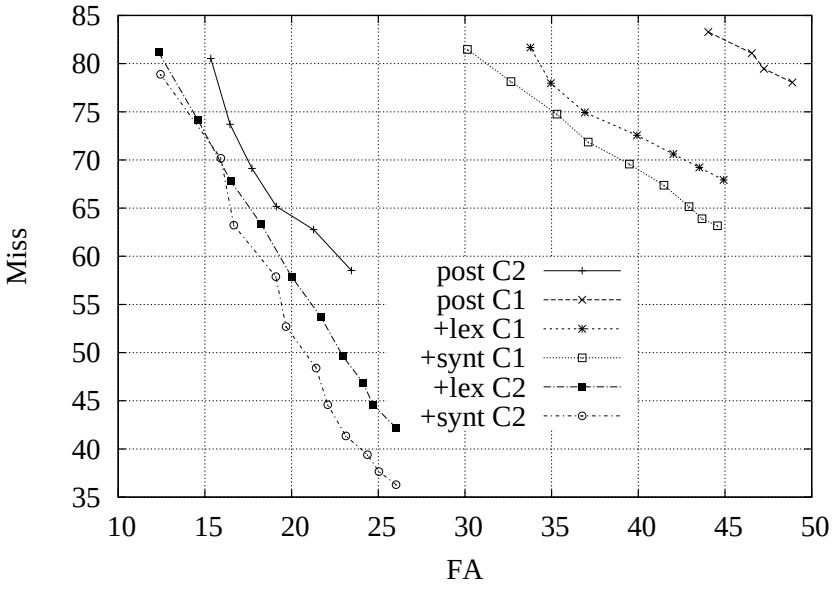


Figure 2 – Courbe Miss/FA avec différents jeu de paramètres sur les corpus C1 et C2

Toutes les mesures précédentes sont données au niveau des mots. Cependant, dans cette étude, on s'intéresse principalement aux segments d'erreurs, considérant la fertilité forte des erreurs impactant les performances finales du système de traduction. Nous présentons ainsi une mesure plus proche de la tache visée : la mesure du taux d'erreurs mots après avoir agrégé toutes les détections successives d'erreurs au sein d'un même symbole `X`. Cette méthode ne permet pas de corriger des erreurs de transcription, elle permet cependant de diminuer le nombre d'insertions erronées. Notons aussi que c'est une mesure assez dure dans la mesure où aucun gain n'est donné pour la détection correcte d'une erreur constituée d'un seul mot : un mot erroné est remplacé par `X`, qui est toujours considéré comme une confusion par le programme d'alignement avec la référence. Par exemple, sur une phrase comme "I saw that man at Izamni", si la transcription automatique est "I saw that man at is on me" on aura un taux d'erreurs de 50%. Si on détecte correctement le segment d'erreurs et qu'il est remplacé dans la transcription automatique par le symbole `X`, nous obtenons la transcription : "I saw that man at X" qui a un taux d'erreurs de 16.6%. Cette mesure alternative de la qualité d'un prédicteur d'erreurs de reconnaissance est particulièrement intéressante car elle permet de juger de l'utilité objective d'un tel prédicteur (baisse du taux d'erreurs sur les mots) sans être liée à un cadre applicatif précis.

Corpus/WER	ASR	Oracle	P>0.5	P>0.8
Corpus 1	8.4	7.5	9.2	8.3
Corpus 2	35.8	19.9	31.1	29.6

Table 5 – Taux d'erreur mots obtenus en agrégeant tous les segments d'erreurs détectés automatiquement dans les transcriptions automatiques par un seul symbole `X` puis en comparant à la référence

On peut voir dans la figure 5, grâce au taux Oracle, l'importance de bien détecter les segments d'erreurs. On peut faire chuter très significativement le taux d'erreurs mots, même sans la moindre correction, si on arrive à détecter correctement les segments erronés. Les taux d'erreurs sont donnés selon deux conditions : P>0.5 (respectivement P>0.8) signifie qu'un mot est considéré comme erreur si son score de confiance est supérieur à 0.5 (respectivement 0.8). On voit qu'une baisse sensible du taux d'erreurs mots sur les corpus 2 est constaté (gain de 6% absolu). Ce résultat est encourageant, il montre que notre approche basée sur un étiqueteur de séquence et des indices linguistiques parvient bien à détecter les segments d'erreurs importants, sans pour autant impacter les performances pour le corpus 1.

6 Conclusion

Dans le cadre d'un système de traduction parole-parole cette étude a présenté un module de détection et de caractérisation d'erreurs basé sur un étiqueteur de séquence utilisant des indices acoustiques, lexicaux et syntaxiques. La partie expérimentale, effectuée sur le corpus BOLT, a montré la validité de l'approche pour la détection des erreurs "importantes" correspondant par exemple aux mots hors vocabulaire, qui ont un impact important sur les traitements successifs à la phase de reconnaissance.

Remerciements

Cette étude a été financée en partie par le contrat BOLT DARPA HR0011-12-C-0016 en tant que sous-traitant de SRI International.

Références

Ayan, N. F., Mandal, A., Frandsen, M., Zheng, J., Blasco, P., Kathol, A., Béchet, F., Favre, B., Marin, A., Kwiatkowski, T. et al. (2013). "can you give me another word for hyperbaric ?" : Improving speech translation using targeted clarification questions. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8391–8395. IEEE.

Béchet, F. et Favre, B. (2013). Asr error segment localization for spoken recovery strategy. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 6837–6841. IEEE.

Fayyad, U. et Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning.

Franco, H., Zheng, J., Butzberger, J., Cesari, F., Frandsen, M., Arnold, J., Gadde, V., Stolcke, A. et Abrash, V. (2002). Dynaspeak : Sri's scalable speech recognizer for embedded and mobile systems. In Proceedings of the second international conference on Human Language Technology Research, pages 25–30. Morgan Kaufmann Publishers Inc.

Hazen, T., Seneff, S. et Polifroni, J. (2002). Recognition confidence scoring and its use in speech understanding systems. Computer Speech & Language, 16(1):49–67.

Hirschberg, J., Litman, D. et Swerts, M. (2004). Prosodic and other cues to speech recognition failures. Speech Communication, 43(1):155–175.

Nasr, A., Béchet, F., Rey, J., Favre, B. et Le Roux, J. (2011). Macaon : An nlp tool suite for processing word lattices. Proceedings of the ACL 2011 System Demonstration, pages 86–91.

Okazaki, N. (2007). Crfsuite : a fast implementation of conditional random fields (crfs) `http://www.chokkan.org/software/crfsuite/`.

Parada, C., Dredze, M., Filimonov, D. et Jelinek, F. (2010). Contextual information improves oov detection in speech. In North American Chapter of the Association for Computational Linguistics (NAACL).

Rastrow, A., Sethy, A. et Ramabhadran, B. (2009). A new method for oov detection using hybrid word/fragment system. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pages 3953–3956. IEEE.

Raymond, C. (2012). Discretized for CRF. `http://www.irisa.fr/texmex/people/raymond/Tools/tools.html`.

Sarikaya, R., Gao, Y., Picheny, M. et Erdogan, H. (2005). Semantic confidence measurement for spoken dialog systems. Speech and Audio Processing, IEEE Transactions on, 13(4):534–545.

L'identification vocale : pour une quantification des effets de la familiarité

Julien Plante-Hébert[1, 2], Victor J. Boucher[1, 3]
(1)Laboratoire de sciences phonétiques, Université de Montréal,
Montréal, Québec, Canada, H3T 1N8
julien.plante-hebert@umontreal.ca[2], victor.boucher@umontreal.ca[3]

RESUME

La présente étude porte sur les effets de la familiarité dans l'identification de locuteurs dans une situation de parade vocale. Cette technique, inspirée d'une procédure d'identification visuelle d'individus, consiste en la présentation de plusieurs voix avec des aspects acoustiques similaires définis selon des critères reconnus. L'objectif principal était de déterminer si un contrôle de la familiarité d'une voix dans une parade vocale permet d'obtenir un haut taux d'identification vocale (> 99 %). Notre étude est la première à quantifier le critère de familiarité selon le degré de contact entre un témoin et un individu associé à une « voix cible» dans une parade vocale. Les résultats indiquent qu'un très haut taux d'identification vocale s'obtient par l'application d'un index de familiarité en contrôlant pour la longueur des stimuli vocaux.

ABSTRACT

Voice identification: for a quantitative approach to the effects of familiarity

The present study aims to determine the effect of familiarity on speaker identification in voice line-ups. The voice line-up technique is analogous to visual line-ups consisting in the presentation of a number of voices sharing similar acoustics properties that are well recognized in the literature. The main objective was to investigate if, by controlling voice familiarity, one can obtain high rates of correct voice identification (>99 %). This study is the first to quantify the familiarity criterion in terms of the degree of contact between a witness and an individual associated with a voice in a line-up. The results indicate that a very high rate of voice identification can be obtained by applying a criterion index of familiarity and controlling for the length of the voice stimuli.

MOTS-CLES : Familiarité, parade vocale, identification.

KEYWORDS: Familiarity, voice line-up, identification.

1 Introduction

L'identification d'un locuteur par le biais d'enregistrements audio ou de signaux téléphoniques possède plusieurs applications. Dans un contexte légal ou criminel, l'identification vocale peut être critique surtout lorsqu'il n'y a aucun document visuel pour appuyer une défense ou mise en accusation. Dans ces situations, certains organismes tels Interpol (2001) recommandent l'utilisation de la technique des parades vocales. Cette technique est analogue à la technique des parades visuelles, qui est bien connue des enquêteurs (Jessen 2008). Les cas légaux ayant recours à la parade vocale sont, dans une grande majorité, des cas où le témoignage d'un individu ou d'un groupe d'individus est critique dans l'établissement d'un jugement. . Le recours à la parade vocale, dans ce contexte, invite à une grande prudence et à l'application de techniques qui sont de plus en plus standardisées. Par exemple, il existe actuellement des recommandations formelles régissant le choix des voix dans l'élaboration de parades vocales : celles-ci doivent avoir une fréquence fondamentale similaire, on doit utiliser les mêmes énoncés, présenter des conditions d'écoute identiques pour les témoins (etc., Yarmey 1995, Nolan 2003, Jessen 2008). En considérant les facteurs qui peuvent faire varier l'exactitude des identifications vocales des témoins, une des variables évidentes est le niveau de familiarité des voix entendues.

Bien qu'un certain nombre d'études (p. ex. Foulkes 2000 ; Yarmey 2001) ont examiné l'effet de la familiarité des voix dans le contexte des parades vocales, aucune recherche n'a délimité de manière opérationnelle de catégories objectives de familiarité vocale, même s'il est reconnu que l'être humain possède de grandes capacités à identifier les voix familières. En fait, il y a une tendance dans la communauté scientifique spécialisée à référer à des techniques automatisées alors que l'identification par le biais humain et les parades vocales offre une possibilité de taux d'identification supérieurs aux techniques informatiques et une plus grande efficacité. Le tableau 1 résume certaines comparaisons faites par Jessen (2008) et Campbell (2009) qui mettent en lumière les avantages respectifs de la parade vocale et de l'identification automatisée.

Paramètres	**Parades vocales**	**Techniques automatisées**
Aucun besoin d'entrainement	☑	☐
Meilleurs résultats	☑	☐
Considération du contenu para et extra linguistique	☑	☐
Coût matériel	☑	☐
Données interprétables par spécialistes	☑	☐
Rapidité des procédures	☐	☑
Accessibilité	☐	☑
Fiabilité des données	☐	☑

TABLEAU 1 – Comparaison entre les avantages et les inconvénients de la parade vocale et des techniques d'identification par la voix automatisées.

Par rapport aux points de comparaisons du tableau ci-dessus, notons simplement que l'utilisation de témoins humains dans une parade vocale peut offrir des taux d'identification supérieurs (selon les conditions utilisées dans la présente étude) aux techniques d'identification automatisées. Notons aussi que les participants à une parade ne nécessitent pas de période d'entraînement,

puisqu'ils sont déjà en mesure de catégoriser des voix familières, contrairement aux techniques automatisées qui peuvent requérir un nombre indéterminé d'expositions à des exemplaires afin d'établir les attributs acoustiques pertinents à l'identification. L'efficacité de la perception humaine dans l'identification vocale peut se saisir en pensant à la capacité que nous avons à distinguer la voix de proches. Par exemple, nous pouvons facilement identifier, parmi plusieurs voix, celle de notre mère ou celle d'un frère peu importe ce qui est dit, alors qu'un système automatique nécessiterait un vaste corpus pour en arriver à des taux qui peuvent fluctuer selon les énoncés.

En dehors des applications légales, l'étude de l'identification par la voix suscite aussi un intérêt pour la linguistique fondamentale. Ainsi, l'accroissement du taux d'identifications correctes en fonction du niveau de familiarité suggère un raffinement des «théories d'exemplaires» selon lesquelles l'emprunte mnémonique de la parole se renforce par la fréquence d'exposition (Johnson, 1997). En fait, les effets de la familiarité sur l'identification de la voix suggèrent que les «exemplaires» qui se développent ne sont pas seulement des «traits» phonologiques mais aussi des éléments d'informations associés aux voix d'individus. Les exemplaires se raffinent à force d'exposition au point où des voix familières pourraient être identifiées avec un très faible taux d'erreur. Dans cette optique, le présent travail vise à démontrer que le contrôle du niveau de familiarité des voix dans une technique de parade vocale permet d'obtenir de hauts taux d'identification qui dépassent les possibilités actuelles de techniques informatiques.

2 Méthode

2.1 Participants et questionnaire sur la familiarité

Les participants étaient 18 hommes et 26 femmes (n=44), âgés entre 18 et 65 ans et ayant le français comme langue maternelle. Ces individus n'avaient aucun trouble d'audition apparent ou diagnostiqué. Les participants étaient avisés, lors du recrutement, que l'expérience impliquait l'identification de voix dans trois parades vocales et que certaines des voix (dites «voix-cibles») pouvaient être d'individus qui leur étaient familiers. Les participants à l'étude ont été recrutés selon leur degré de familiarité *a priori* avec la voix cible d'une parade vocale à laquelle ils seraient exposés et le degré de familiarité a été établi au moyen d'un questionnaire (voir section 2.2.1 ci-dessous). Ainsi, les voix cibles pouvaient être des voix de parents proches, de collègues de travail, d'amis, de connaissances ou des voix de personnes inconnues.

2.1.1 Critères de familiarité de la «voix cible»

Lors de l'expérimentation, chaque participant devait remplir un questionnaire visant à établir son niveau de familiarité avec la voix cible dans une parade avant la tâche expérimentale. Le questionnaire a permis l'élaboration d'un «index de familiarité » basé essentiellement sur trois critères:

1. La fréquence des communications: il est ici question de la fréquence moyenne des communications orales entre le participant et la voix cible depuis le premier événement.
2. La récence des communications: le participant devait mentionner à quand remonte la dernière communication orale avec la voix cible.
3. La durée des communications: la durée moyenne des communications entre le

participant et la voix cible depuis le premier événement.

Les participants ont été répartis en nombres égaux dans chacun des niveaux d'index de familiarité utilisés. Ceux regroupés dans l'index de familiarité 1 étaient des individus n'ayant eu aucun contact avec la voix cible préalablement à l'expérimentation. Ces derniers ont entendu, environ 30 minutes avant le début de la tâche expérimentale, un enregistrement de la voix qu'ils devaient identifier.

2.2 Stimuli

2.2.1 Les voix de la parade

Trois ensembles de parades vocales ont été élaborés, chaque ensemble contenait quatre longueurs d'énoncés (voir section 2.3.2) pour un total de 24 parades vocales. Chacune des parades contenait 10 voix d'hommes âgés entre 18 et 50 ans dont une voix cible. Toutes les voix dans les parades avaient une fréquence fondamentale moyenne en parole (F0mp) similaire, c'est-à-dire que l'écart entre les F0mp des voix présentées était moins d'un semi ton (Hudson, 2007), respectant les standards en vigueur dans l'élaboration de parades vocales. Aussi, tous les volontaires à l'élaboration des stimuli étaient des locuteurs natifs du français québécois et aucun ne possédait d'accent régional marqué ou de particularités articulatoires (idiosyncrasies). Enfin, aucun des locuteurs ne possédait de trouble d'articulation apparent ou diagnostiqué.

2.2.2 Longueurs d'énoncés et enregistrements audio

Les énoncés utilisés dans les parades vocales étaient des phrases familières produites avec une intonation neutre (p. ex. «merci beaucoup», «comment vas-tu ?») et variaient en longueur syllabique (1, 4, 10 et 18 syllabes). Chacune de ces longueurs comportait deux énoncés dont l'un était composé principalement de sons oraux et l'autre comportait plusieurs sons nasaux (l'analyse des effets associés à la nasalité ne fait pas l'objet du présent rapport). Tous ces enregistrements ont été effectués dans une salle à l'épreuve du son au laboratoire de phonétique de l'Université de Montréal à l'aide d'un microphone *Electro-Voice* (modèle A635), d'une carte sonore interne d'ordinateur (16-bit, *Sony Vaio NW Series*) régler à un taux d'échantillonnage de 44,1 kHz..

2.2.3 Filtrage des stimuli pour répliquer les conditions d'écoute par téléphone

Il est utile de noter que l'identification vocale dans des situations paralégales implique souvent des voix transmises par le biais de téléphones sans fil (Fernandez Gallaro, 2012). Afin de représenter ces situations, tous les stimuli enregistrés ont été filtré (de 300Hz à 3500Hz) au moyen d'un filtre blackman reproduisant les courbes de téléphone cellulaire. De plus, un bruit de fond audible (largeur de bande de 0.25kHz à 6.5 kHz et d'une amplitude maximale de 24 dB) mais qui n'altérait pas la perception de la parole, a été ajouté aux enregistrements.

2.3 Procédure expérimentale

La procédure impliquait une tâche à deux étapes. Dans la première étape, les participants avaient pour directives d'écouter d'abord chacun des 10 énoncés dans une parade, un énoncé à la fois,

sans répétition et dans un ordre prédéterminé. Ils devaient, lors de l'écoute, noter, sur un formulaire, les voix pouvant représenter un individu qu'ils connaissaient. Immédiatement après cette première écoute, il devait indiquer la voix la plus susceptible d'appartenir à un individu connu. Dans une deuxième étape, le participant pouvait réécouter autant de fois qu'il voulait les voix dans une parade afin de valider ou modifier sa réponse. Les résultats ci-dessous, bien que très similaires à ceux obtenus à la première étape, portent sur l'identification vocale faite à la deuxième étape. Ces résultats regroupent 8 essais par participants (n = 44) pour un total de 352 essais.

Pour l'écoute des stimuli, tous les participants ont utilisé un casque d'écoute (*Beyerdynamic*, modèle DT250) et les signaux étaient rejoués à partir d'un ordinateur portable (même système que celui utilisé pour l'enregistrement) à une amplitude maximale d'environ 68 dBa (un niveau confortable d'écoute) tel que mesurée avec un sonomètre et un adaptateur (*Digital Recordings*, modèle DR-1).

3 Résultats

Les résultats présentés dans la Figure 1 démontrent une hausse du pourcentage d'identifications correctes en fonction de l'index de familiarité. Ces résultats incluent les identifications obtenues avec les énoncés de 4, 10 et 18 syllabes en éliminant les valeurs associées aux énoncés monosyllabiques (qui n'ont pas mené à des taux acceptables d'identifications). La figure 1 montre qu'un très haut taux de réussite à la tâche d'identification (100%) est obtenu avec un index de familiarité de niveau 4. Notons que les identifications à partir d'énoncés monosyllabiques n'ont démontré aucune tendance particulière.

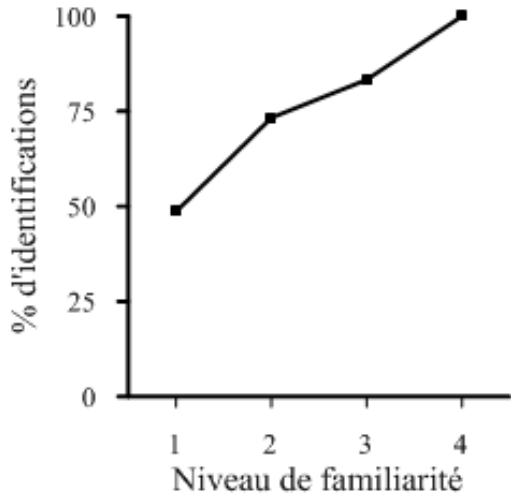


FIGURE 1 – Pourcentage global d'identifications correctes en fonction du niveau de familiarité. Les énoncés monosyllabiques ont été exclus.

Quant aux variations associées à la longueur des énoncés, la figure 2 démontre que les taux d'identification diffèrent très peu pour les longs énoncés (10 et 18 syllabes). Par contre, il s'avère que l'identification de voix, même très familière, n'est pas possible avec de courts énoncés de moins de quatre syllabes.

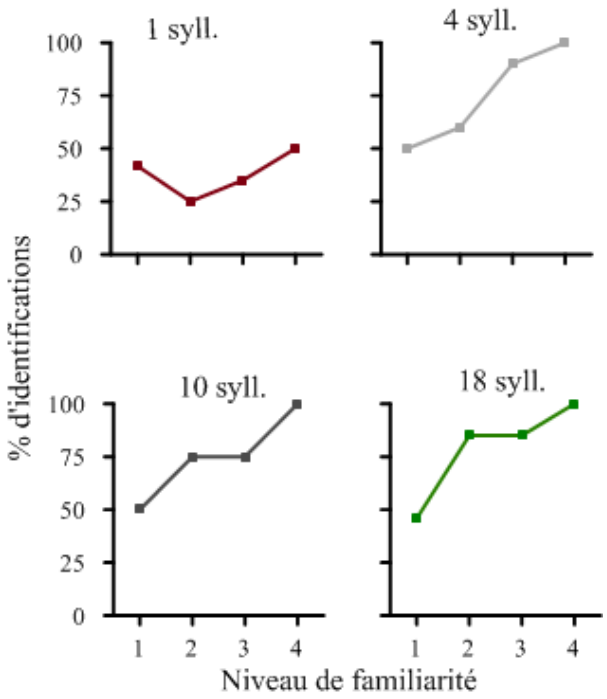


FIGURE 2 – Pourcentage d'identifications correctes en fonction du niveau de familiarité pour chaque longueur d'énoncé.

Notons par rapport à la Figure 1 que, la probabilité que tous les participants ayant un niveau 4 de familiarité (n=12) identifient la voix cible par hasard est de $p < 0{,}000001$.

Le tableau I illustre les coefficients de corrélation entre la réussite à la tâche expérimentale et le pointage obtenu pour chaque paramètre évalué par le questionnaire de familiarité [1]. Ces coefficients sont statistiquement significatifs pour les paramètres de fréquence, durée et période.

Paramètres	r_s	p
Récence	-0.052	< 0.743
Fréquence	0.542	< 0.001
Durée	0.494	< 0.01
Période	0.686	< 0.001

TABLEAU I – Coefficients de corrélation entre le pointage obtenu pour les paramètres de la familiarité et le pourcentage moyen d'identifications.

4 Discussion et conclusion

Les résultats ci-dessus permettent d'affirmer que de hauts taux d'identification approchant 100% peuvent être obtenus dans le cadre d'une parade vocale élaborée selon les recommandations en vigueur si on contrôle le niveau de familiarité des voix. Notons que les taux rapportés ont été obtenus pour des enregistrements reflétant les conditions d'écoute de téléphones cellulaires, ce qui permet d'espérer que les résultats soient généralisables à des situations d'écoutes difficiles. Par ailleurs, l'analyse indique que le facteur de la longueur des énoncés a peu d'influence sur les

[1] Pour des raisons de confidentialité et de propriété intellectuelle, veuillez contacter l'auteur pour plus de détails concernant le barème de pointage utilisé

taux d'identification : c'est-à-dire, avec un index de familiarité de 4, on obtient des taux d' identification s'approchant 100% avec des énoncés de quatre syllabes ou plus. Par conséquent, contrairement à certaines recherches (Eriksson, 2007), il ne semble pas y avoir d'effet majeur de longueur de stimuli lorsque les voix entendues sont familières. En conclusion, cette étude indique que le critère de familiarité joue un rôle majeur dans l'identification d'individu par la voix en contexte de parade vocale. La quantification de la familiarité serait donc primordiale surtout dans un contexte d'application légal. Par contre, un problème qui nuit à l'application du critère de familiarité est le fait que les plus hauts taux d'identification sont obtenus avec des personnes qui connaissent bien les individus associés aux voix cibles. Rappelons qu'un index 4 de familiarité est un individu avec qui il y a un contact très fréquent (plus d'une fois semaine) et un contact de longue durée (un ami de longue date, un membre de la famille, etc.) et un contact récent (dans le dernier mois). Or, dans un contexte d'application légal où un individu serait amené à identifier une voix d'un proche, il serait probable que l'individu évite de répondre afin d'éviter une possible incrimination. Nos travaux ultérieurs visent à répondre à cette difficulté en exploitant des techniques d'identification axées sur les potentiels évoqués.

Références

CAMBELL, J. 2009. Forensic speaker identification: A need for caution. *IEEE Signal Processing Magazine.*03: 95-103

ERIKSSON, E. 2007. That voice sounds familiar. Factors in speaker recognition. *Umeä Studies in Cognitive Science 1.*

FERNANDEZ GALLARNO, L. 2012. Comparison of Human Speaker Identification of Known Voices Transmitted through Narrowband and Wideband Communication Systems. *ITG Fachbericht 236 : Sprachkommunikation.*

FOULKES, P. 2000. Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*. 7(2) : 180-198.

HUDSON, T. 2007. F0 statistic for 100 young male speaker of standard southern brithish english. *XVI International Conference od Phonetic Sciences (ICPhS 2007).* Saarbrücken, Allemagne.

INTERPOL. 2001. Forensic speech and audio analysis forensic linguistics. *Forensic Science Symposium (2001)*. Lyon, France.

JESSEN, M. 2008. Forensic Phonetics. *Language and Linguisitcs Compass.* 2(4) : 671-711.

JOHNSON, K. 1997. Speech Perception without Speaker Normalization : An Exemplar Model. *Talker variability in speech processing.* Acamdemic Press.

NOLAN, F. 2003. A recent Voice Parade. *Speech, Language and the Law.*10(2) : 277-291.

YARMEY, D. 1995. Earwitness Speaker Identification. *Psychology, Public policy and Law*, 1(4) : 792-816.

YARMEY, D. 2001. *Commonsens Beliefs and the Identification of Familiar Voices.* Applied Cognitive Psychology. 15 : 283-299.

A la recherche des traces dialectales dans l'arabe standard : production des voyelles et des fricatives inter-dentales par des locuteurs tunisiens et marocains.

Zeineb Ammar, Cécile Fougeron, Rachid Ridouane
1 Laboratoire de Phonétique et Phonologie, UMR 7018 CNRS/Université Paris 3 Sorbonne Nouvelle, Labex EFL
zeineb.ammar@etud.sorbonne-nouvelle.fr ; {cecile.fougeron ; rachid.ridouane}@univ-paris3.fr

RESUME

Ce travail s'intègre dans un projet plus large visant à examiner les transferts phonologiques des dialectes arabes dans les productions des locuteurs arabophones natifs lorsqu'ils s'expriment en arabe standard. Nous étudions ici la production des voyelles et fricatives inter-dentales produites par 11 locuteurs du dialecte tunisien sur un corpus lu en arabe standard. Ces productions sont comparées à celles de 5 locuteurs marocains servant de groupe contrôle. Nos résultats montrent que plusieurs indices segmentaux permettent de distinguer les productions en arabe standard des locuteurs d'origines dialectales différentes : les timbres vocaliques, la production du contraste de durée entre les voyelles longues et brèves, et la réalisation des fricatives inter-dentales. Ces indices sont interprétés comme des traces de leur dialecte d'origine.

ABSTRACT

Looking for dialectal cues in Standard Arabic: vowel and interdental fricative production of Tunisian and Moroccan speakers

As part of a wider project aiming at investigating phonological transfers of Arabic dialects in Standard Arabic, this work focuses on the production of Tunisian speakers when they speak Standard Arabic. We study the production of vowels and interdental fricative of 11 Tunisian speakers in a read corpus of Standard Arabic. These productions are compared with those of 5 Moroccan speakers used as a control group. Several segmental cues are found to distinguish the Standard Arabic productions of the two groups of speakers: they concerned the acoustic vowel formants, the realization of the quantity contrast between long and short vowels, and the realization of interdental fricatives. These cues are interpreted as surface manifestations of the phonological transfer from speakers' native dialects.

MOTS-CLES : Transfert, variantes dialectales, arabe standard, dialecte tunisien et marocain, voyelles courtes et longues, fricatives inter-dentales

KEYWORDS: Transfer, dialectal variants, Standard Arabic, Tunisian and Moroccan dialects, short and long vowels, interdental fricatives.

1. Introduction

Les dialectes arabes sont généralement subdivisés en deux grandes zones dialectales : la zone orientale qui regroupe les pays du Moyen-Orient et la zone occidentale ou maghrébine (Cohen 1973 ; Barkat 2000 ; Embarki 2008). L'intercompréhension entre les différents dialectes varie en fonction de l'éloignement des zones dialectales, et les locuteurs (lettrés généralement) de pays différents peuvent recourir à l'arabe standard pour communiquer entre eux. L'arabe standard est considéré comme le style formel de la langue, un registre « soutenu ». C'est une langue qu'apprennent les enfants à l'école comme une langue étrangère dans sa structure et souvent aussi dans son vocabulaire. Il résulte de cela que les productions en arabe standard varient en fonction de l'origine dialectale des locuteurs, et il n'est pas rare que les auditeurs natifs affirment pouvoir identifier le dialecte ou du moins la zone dialectale de l'origine du locuteur parlant en arabe standard.

Plusieurs études se sont penchées sur la question des marqueurs phonétiques permettant l'identification des dialectes arabes, soit en comparant des dialectes entre eux (voir notamment Barkat (2000) et les références qui y sont citées) soit en cherchant des traces de l'origine dialectale d'un locuteur lorsqu'il parle une autre langue de référence. Barkat (1999, 2000, 2001) montre que l'information la plus pertinente pour l'identification de l'origine dialectale d'un locuteur arabophone est liée à l'organisation qualitative et quantitative de son système vocalique. A partir d'un corpus d'échantillons de parole spontanée de 6 variétés dialectales (marocaine, algérienne, tunisienne, syrienne, libanaise, jordanienne), l'auteur appuie ses résultats sur l'analyse acoustique des indices identifiés par les locuteurs eux mêmes. Les auditeurs, arabophones natifs de parlers arabes différents, sont capables de distinguer ces différents parlers et catégorisent correctement les différents stimuli selon les deux zones géographiques principales mentionnées plus haut. Outre ces indices segmentaux, des indices prosodiques, comme les variations de f0, d'énergie et de durée permettent de différencier les dialectes arabes maghrébins des dialectes arabes moyen-orientaux (Barkat et al. 1999). Boula de Mareuil et al. (2004) ont également étudié le rôle des indices segmentaux et suprasegmentaux dans l'identification de l'origine dialectale de locuteurs maghrébins lorsqu'ils parlent français. Leurs tests d'identification sur des échantillons de parole lue en français standard par 6 locuteurs arabes de dialectes différents (2 marocains, 2 tunisiens et 2 algériens), 6 locuteurs kabyles et 6 locuteurs français, étaient faits par deux groupes d'auditeurs : un groupe d'auditeurs français qui devaient attribuer un degré d'accent au stimuli et un groupe d'auditeurs algériens qui devaient identifier l'origine dialectale du locuteur. Ils concluent que les indices segmentaux sont plus pertinents que les indices supragsegmentaux pour l'identification de l'origine dialectale de locuteurs maghrébins parlant le français, surtout si le locuteur est jugé ayant un fort accent. Lorsque l'on compare les scores d'identification des différents locuteurs, il ressort que ce sont les locuteurs d'origine tunisienne qui sont les moins bien identifiés (30.5%), ce qui contraste avec l'identification des locuteurs de dialecte proche, comme les marocains qui sont identifiés correctement à 56.2%. Rjaibi-Sahbi (1993) a également étudié la discrimination de l'origine dialectale de locuteurs arabophones dans une langue de référence, mais cette fois-ci en utilisant des productions en arabe standard. L'auteur s'intéresse essentiellement aux trois fricatives inter-dentales de l'arabe standard /θ, ð, ðˁ/ qui connaissent une grande variabilité inter-dialectale. L'auteur s'appuie sur une étude phonétique et phonologique de 8 parlers arabes (Yémen, Iraq, Liban, Syrie, Egypte, Soudan, Tunisie, Algérie). Sur le plan phonétique, des tests perceptifs, basés sur des passages radiophoniques en arabe standard extraits de 6 radios arabes (Tunisie, Maroc, Egypte, Emirats

Arabes Unis, Algérie et Syrie), ont été effectués auprès de 6 auditeurs arabophones naïfs dont la tâche consistait à reconnaître l'origine géographique des locuteurs. Le taux de reconnaissance obtenu par l'auteur est particulièrement élevé, de l'ordre de 90%. Ce travail constitue une des rares études à avoir traité de l'identification de l'origine dialectale des locuteurs arabophones à partir de l'arabe standard (voir aussi Abu-Haidar 1991). Il nous est donc paru intéressant de compléter cette investigation qui peut avoir des applications importantes, notamment dans le domaine de l'identification automatique des langues.

Nous nous intéressons donc ici aux caractéristiques phonétiques de l'arabe standard tel qu'il est produit par des locuteurs natifs d'un dialecte maghrébin qui semble à priori plus difficilement identifiable que les autres : le tunisien. Cette difficulté est probablement liée au statut particulier du dialecte tunisien, situé géographiquement entre le pôle occidental et le pôle oriental. Il est en ce sens un dialecte intermédiaire aux points de vue des caractéristiques phonétiques (Hamdi et al. 2004 ; Boula de Mareuil et al. 2004). Nous examinerons dans ces productions des traces des particularités connues du dialecte tunisien (Jomaa, 1992). Ces réalisations seront comparées aux productions d'un groupe contrôle d'origine dialectale géographiquement proche, les locuteurs marocains. Le dialecte arabe marocain est choisi à dessein, car il constitue la variante arabe la plus occidentale, et dont le phonétisme est particulièrement influencé par celui du berbère (Cohen 1973 ; Embarki 2004). L'arabe tunisien et l'arabe marocain possèdent des systèmes phonologiques qui diffèrent principalement par leurs timbres et quantités vocaliques et par la production des fricatives inter-dentales. Ce sont justement ces caractéristiques vocaliques et consonantiques qui feront l'objet de l'analyse proposée dans cette étude.

Les analyses faites par Barkat sur la distribution des segments vocaliques dans les parlers maghrébins montrent que la dispersion des segments vocaliques en dialecte tunisien est plutôt périphérique, alors qu'elle est plutôt centrale en dialecte marocain. Pour autant, le système vocalique du dialecte tunisien se distingue par rapport à celui de l'arabe standard par une plus grande variation allophonique des voyelles brèves /a, i, u/ et de certaines correspondantes longues. Par exemple, les voyelles brèves /i/ /a/ /u/ disparaissent en syllabe ouverte inaccentuée et /i/ et /u/ sont réalisées [I] et [ʊ] en syllabe fermée, respectivement (Jomaa 1992). Metoui (1989) propose un quatrième « timbre vocalique » [æ] avec sa correspondante longue [æ:] qui résulterait d'une prononciation plus antérieure et moins ouverte de la voyelle /a/ correspondante en arabe standard. Dans le parler du sahel cette voyelle peut être plus fermée avec une réalisation [ɛ], voire [e] (Jomaa 1992 ; Baccouche 1969). Toutefois, la littérature reste assez partagée quant à la considération de [æ] comme un phonème ou comme une variante allophonique de /a/ (Barkat 2000 ; Jomaa 1992).

En ce qui concerne la réalisation de l'opposition de durée vocalique, l'ensemble des études traitant des dialectes arabes et notamment du dialecte tunisien (Cantineau 1960 ; Ghazeli 1979 ; Barkat 2000, etc.), semble être d'accord sur le fait que le dialecte tunisien conserve l'opposition de quantité vocalique (voyelles brèves *vs.* voyelles longues), à l'instar de l'arabe standard. Par contre, les avis divergent en ce qui concerne le dialecte marocain. Cantineau (1960), par exemple, avance l'hypothèse selon laquelle la seule opposition binaire de quantité vocalique qui subsiste encore serait entre /u/ et /u:/. Embarki (2012), dans son étude dédiée à ce sujet, observe à partir d'une analyse acoustique de corpus de parole spontanée, que l'opposition entre voyelles longues et voyelles brèves n'est plus faite en dialecte marocain au niveau phonologique et que la perception de cette opposition par certains auditeurs est due essentiellement à la nature de la syllabe ou à la morphologie classique du mot (mot emprunté à l'arabe classique). Barkat (2000)

observe de son côté « qu'en parole spontanée, c'est-à-dire lorsque tous les paramètres de durée co-intrinsèque entrent en interaction, l'opposition de durée vocalique en dialecte marocain est toujours pertinente ».

Concernant la production des fricatives inter-dentales /θ, ð, ðˤ/ de l'arabe standard, la situation semble plus claire. La plupart des études menées sur le dialecte marocain (Cantineau 1960 ; Barkat 2000 ; Al-Tamimi 2007), sont d'accord sur l'absence de fricatives inter-dentales [θ, ð, ðˤ] dans ce parler. Elles sont généralement remplacées par les occlusives dentales /t d dˤ/ ou les sifflantes /s z/. Par ailleurs, Cantineau (1960), observe également la présence d'une fricative pharyngale /zˤ/ qui peut être la substitution de la fricative inter-dentale /ðˤ/. En arabe tunisien, par contre, ces fricatives inter-dentales sont réalisées comme telles.

En comparant la réalisation des voyelles, du contraste de durée vocalique et des fricatives inter-dentales de locuteurs des deux dialectes différents parlant en arabe standard, ce travail cherche à tester si les systèmes phonologiques de ces deux dialectes transparaissent au niveau de leurs productions sous la forme de transferts phonologiques.

2. Méthode

Le système vocalique de l'arabe standard comprend 3 voyelles courtes /a i u/ et leurs correspondantes longues /a: i: u:/. Dix paires minimales (ou quasi-minimales) mettant en jeux soit l'opposition de timbre, soit l'opposition de durée ont été utilisées. Les mots sont de structure /(C)V1(:)CV2(:)C/ ou /CV1(:)CVCV2(:)/ et les voyelles étudiées sont /a a:/ et /u u:/ dans la position V1 et /i i:/ dans la position V2 (ex. /qatala/ *vs.* /qa:tala/ ; /kutiba/ *vs.* /ku:tiba/ ; /qa:dim/ *vs.* /qadi:m/). Les contextes consonantiques ont été contrôlés en fonction des paires disponibles en arabe standard de façon à équilibrer les contextes entre les 3 timbres vocaliques. Les consonnes précédentes et suivantes ont été choisies de façon à exclure l'adjacence avec des pharyngales ou des glottales afin d'éviter des problèmes éventuels de segmentation, liés notamment à la délimitation des voyelles cibles.

Concernant l'examen des fricatives inter-dentales /θ, ð, ðˤ/, nous nous sommes basés sur un corpus contenant 30 mots (dix mots pour chaque fricative). Nous avons contrôlé les contextes d'occurrence des fricatives de sorte à ce que ces trois types de consonnes apparaissent dans des positions semblables (initiales, intervocaliques, finales, et médiane précédant une consonne, ex. en position médiane de mot (/ jaθqifu / ; / jaðkuru / ; / taðˤjuru /).

Les 40 mots cibles ont étés insérés dans des phrases courtes et ont été lus par 11 locuteurs tunisiens et 5 locuteurs marocains dans un ordre aléatoire avec deux répétitions par phrase. Les locuteurs sont tous de sexe masculin, et ont entre 25 et 30 ans. Ils ont tous un niveau universitaire et sont en contact quotidien avec l'arabe standard (via les médias audio-visuels ou des lectures personnelles). Les locuteurs tunisiens sont tous natifs de la même région (le sahel tunisien) et résident tous à la ville de Sousse. Les locuteurs marocains résident tous en France depuis plus de 3 ans et sont tous originaires de la ville de Casablanca.

Les données acoustiques ont été segmentées, annotées et analysées avec Praat (Boersma & Weenink 2012). La segmentation des voyelles a été faite manuellement en se basant sur le signal acoustique et le spectrogramme. La durée vocalique a été définie à partir de l'onset et l'offset du F2 de la voyelle. Les voyelles sont annotées comme longues ou brèves selon que la voyelle est étymologiquement longue ou brève en arabe standard. Les valeurs des formants F1 et F2 ont été

prises au milieu des voyelles. Les mesures temporelles et spectrales par item ont été moyennées sur les deux répétitions. La segmentation des fricatives inter-dentales a été déterminée par inspection visuelle du signal acoustique et du spectrogramme et par vérification auditive. Les réalisations ont été annotées comme des fricatives lorsqu'elles présentaient un bruit de friction au delà de 2500Hz ou comme des occlusives lorsqu'elles présentaient un silence et un bruit de relâchement sur le même intervalle. La réalisation fricative est annotée comme une sifflante sourde [s] ou voisée [z] si elle présentait un bruit de friction au-delà de 4 KHz.

3. Résultats :

3.1. Effet dialectal sur l'opposition de durée vocalique

Les durées moyennes des voyelles brèves (VB) et longues (VL) de l'arabe standard produites par les 11 locuteurs tunisiens sont présentées dans la figure 1 pour chaque paire de voyelles. Pour tous les locuteurs, le contraste entre voyelles longues et brèves est réalisé (/a/vs /a :/ $F(1,108)$ $p<0.001$; /i/ vs /i:/ $F(1,64)$ $p<0.01$ et /u/ vs /u:/ $F(1,42)$ $p=0.008$) et le rapport VL/VB moyen chez les locuteurs tunisiens est de 1.63 (tous timbres confondus).

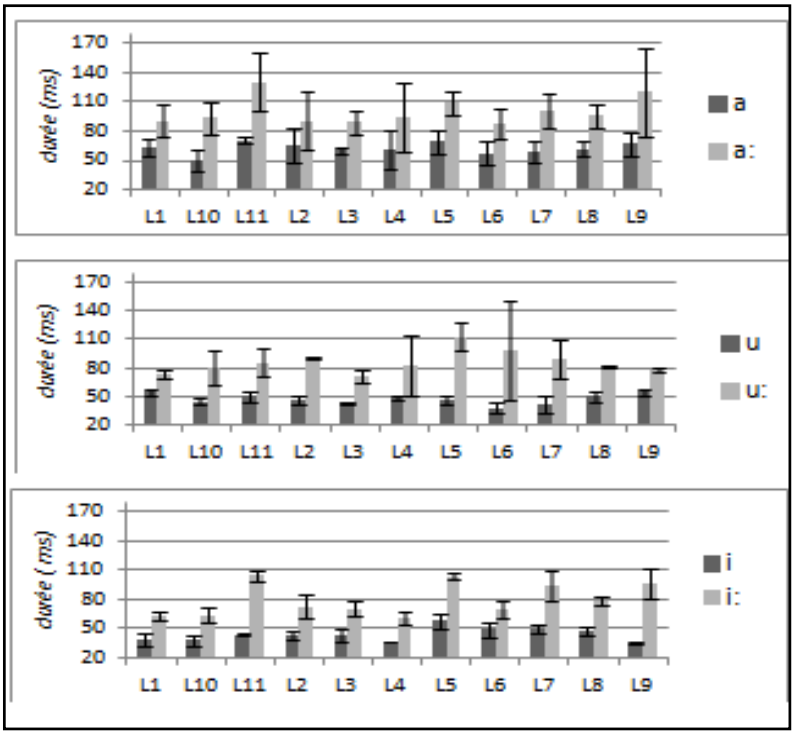


FIGURE 1 : Durées moyennes (et écart type) des voyelles /a/ vs. /a:/ ; /i/ vs. /i:/ ; et /u/ vs. /u:/ pour les 11 locuteurs tunisiens (L1 à L11)

A l'instar des locuteurs tunisiens, les 5 locuteurs marocains produisent aussi l'opposition de durée vocalique en arabe standard, avec des VL plus longues que leurs correspondantes brèves. La comparaison entre locuteurs tunisiens et marocains sur la production de cette opposition de quantité est illustrée Figure 2. Le contraste phonétique de durée produit par les 5 locuteurs marocains est plus important (avec une rapport VL/VB de 1.9) du fait de la production de voyelles longues plus longues que celles des tunisiens (/a/ vs /a:/ $F(1,58)$ $p<0.001$; /i/ vs /i:/ $F(1,38)$ $p<0.01$; /u/ vs /u:/ $F(1,28)$ $p=0.0001$).

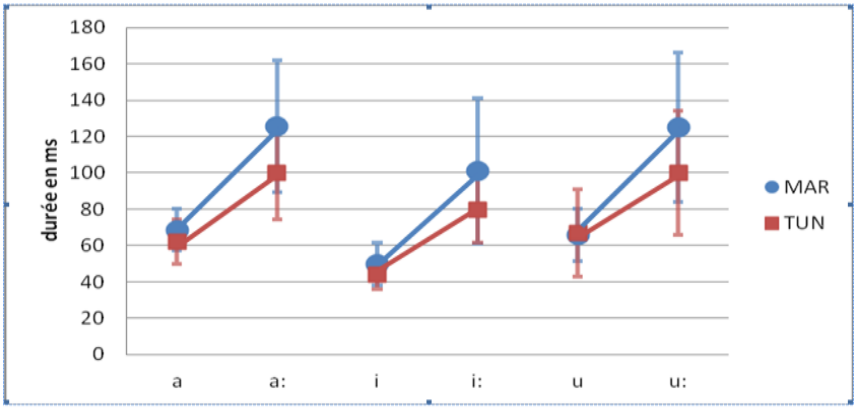


FIGURE 2 : Durées moyennes (et écart type) des voyelles brèves et longues produites par les locuteurs tunisiens (en rouge) et marocains (en bleu).

3.2 Effet dialectal sur le timbre des voyelles

La distribution des voyelles dans l'espace acoustique F1/F2 est illustrée sur la Figure 3 pour les locuteurs tunisiens et marocains. Chez les tunisiens, on observe une réalisation plus périphérique du /i:/ par rapport à /i/. Le timbre de cette voyelle courte dans la production de nos locuteurs tunisiens en arabe standard se rapproche donc d'une réalisation [ɪ] voire [e] décrite dans la littérature pour le dialecte tunisien et elle se distingue bien du /i/ produit par les marocains (autant sur F1 que sur F2). Cette variation de timbre en fonction de la durée ne s'observe pas sur les voyelles /a/ *vs.* /a:/ et /u/ *vs.* /u:/ chez les tunisiens. Ceci contraste avec les productions des locuteurs marocains pour qui l'allongement des voyelles s'accompagne d'un élargissement de l'espace vocalique avec des voyelles longues toutes plus périphériques.

Concernant la réalisation des voyelles ouvertes /a/ et /a:/, on retrouve également dans les productions en arabe standard des traces du dialecte tunisien avec des réalisations plutôt mi-fermées et plus antérieures de type [æ, æ:], voire [ɛ, ɛ:] pour certains locuteurs. Cette dernière réalisation est typique de la région du sahel à laquelle appartiennent nos locuteurs. Pour autant, la comparaison avec les locuteurs marocains montre que cette réalisation de la voyelle ouverte n'est spécifique aux tunisiens que pour la voyelle longue. En effet, le timbre du /a/ court est similaire pour les deux dialectes (Pour les locuteurs tunisiens, $F(1,108)$ $p>0.05$ sur F1 et F2 et pour les locuteurs marocains, $F(1,58)$ $p>0.05$ sur F1 et F2) alors que le F1 de /a:/ est significativement plus élevé chez les marocains, suggérant une articulation plus ouverte.

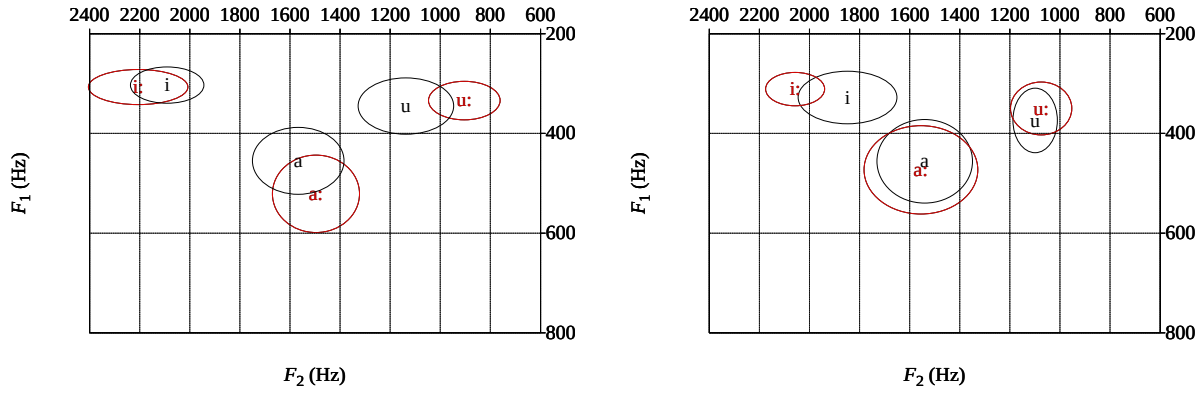


FIGURE 3. Dispersions sur F1/F2 des voyelles brèves (en noir) et longues (en rouge) pour les locuteurs tunisiens (à gauche) et les locuteurs marocains (à droite).

3.3. Production des fricatives inter-dentales /θ, ð, ðˁ/ en arabe standard

Les analyses des productions des locuteurs tunisiens en arabe standard révèlent clairement et sans exception que les fricatives inter-dentales sont maintenues chez ce groupe de locuteurs et ce quel que soit le contexte. Ceci contraste avec les productions des locuteurs marocains dont la réalisation de ces consonnes varie. Comme indiqué dans le tableau 1, les consonnes inter-dentales cibles de l'arabe standard ont été soit réalisées correctement (majorité des /θ/ et moitié des /ð/ et /ðˁ/), soit remplacées par des occlusives dentales ou des fricatives alvéolaires. Cette variabilité dans la réalisation des fricatives inter-dentales de l'arabe standard s'observe chez les 5 locuteurs marocains enregistrés et semble être fonction du voisement de la consonne cible et de sa position dans le mot. Une piste qu'il sera intéressant d'étudier dans de futures recherches.

Cible	Réalisations	% des réalisations
/θ/	s	16%
	t	4%
	θ	80%
/ð/	d	44%
	ð	56%
/ðˁ/	dˁ	50%
	ðˁ	45%
	zˁ	5%

TABLEAU 1 : Réalisations des fricatives inter-dentales chez les 5 locuteurs marocains

Les figures 4 et 5 ci-dessous illustrent deux réalisations différentes de la fricative /ð/ produites par deux locuteurs marocains (L3 et L2) dans le mot /jaðu:qa/. Dans ces exemples, la fricative sonore /ð/ a été soit réalisée correctement (figure 4) soit remplacée par l'occlusive /d/ (figure 5).

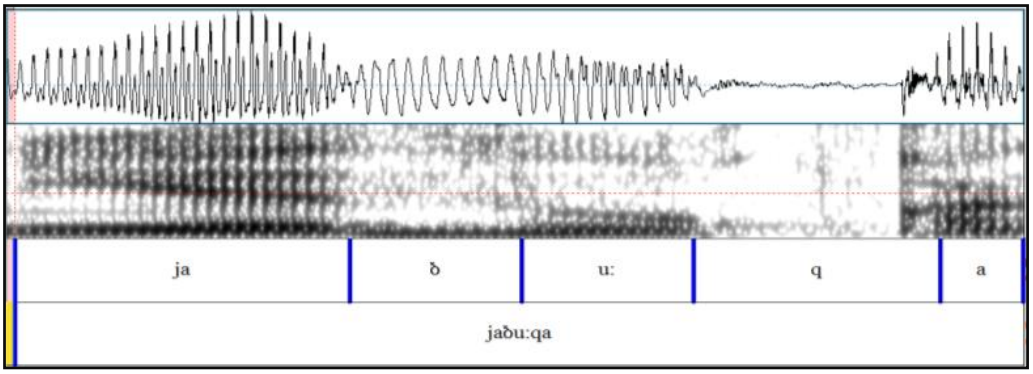


FIGURE 4: Réalisation correcte de /ð/ dans le mot /jaðu:qa/ par le locuteur marocain L3

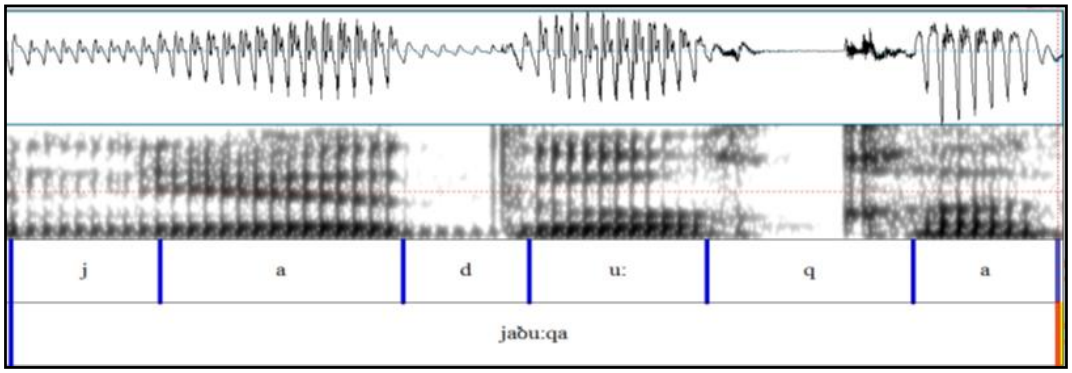


FIGURE 5 : Substitution de / ð / par [d] dans le mot /jaðu:qa/ par le locuteur marocain L2

4. Discussion et conclusion

L'objectif de cette étude a été de tester si l'on pouvait identifier dans la production de locuteurs tunisiens s'exprimant en arabe standard des caractéristiques de leur origine dialectale. Trois aspects segmentaux ont été examinés : le timbre des voyelles, la réalisation du contraste de durée vocalique et la réalisation des fricatives inter-dentales. Sur ces trois aspects il nous est apparu pertinent de comparer les réalisations des locuteurs tunisiens à celles de locuteurs d'un dialecte proche géographiquement mais phonologiquement distinct, notamment en ce qui concerne l'opposition de durée vocalique et la production des consonnes inter-dentales.

Nos résultats soutiennent ceux de Barkat (2000, 2001) concernant l'importance de l'organisation qualitative et quantitative du système vocalique pour l'identification de l'origine dialectale. En effet, il apparaît que les productions des locuteurs tunisiens en arabe standard comportent des traces dialectales en ce qui concerne le timbre des voyelles, notamment sur le /i/ et les /a/-/a:/ qui sont peu périphériques. Les productions des tunisiens se distinguent également de celles des marocains par la façon avec laquelle le contraste de durée entre voyelles brèves et voyelles longues de l'arabe standard est produit. Chez les tunisiens, qui connaissent cette opposition dans leur dialecte d'origine, le contraste a été produit principalement par un allongement. Chez les marocains, pour qui ce contraste semble instable dans leur dialecte, la distinction entre voyelles longues et brèves a été réalisée par une différence de durée mais aussi de timbre (les voyelles courtes étant plus centralisées).

Enfin, outre ces différences entre dialectes sur les voyelles, nos données montrent aussi que la réalisation des fricatives inter-dentales peut s'avérer importante pour l'identification de l'origine dialectale de nos deux groupes de locuteurs, comme l'a montré Rjaibi-Sahbi (1993). En effet, la réalisation correcte des fricatives inter-dentales de l'arabe standard peut être considérée comme une particularité des locuteurs tunisiens qui connaissent ces consonnes dans leur dialecte d'origine. Cette réalisation inter-dentale et fricative systématique des consonnes /θ, ð, ðˁ/ leur permet de ce distinguer des locuteurs marocains chez qui nous avons observé des substitutions par des variantes occlusives ou sifflantes.

Les différences observées dans la production des voyelles et des fricatives inter-dentales de l'arabe standard entre nos locuteurs d'origines dialectales différentes confirment donc l'existence d'un transfert phonologique du dialecte tunisien en arabe standard. Ces résultats en production nous permettent d'envisager une poursuite de ce travail, en l'élargissant à d'autres dialectes et en incluant des expériences en perception dont l'objectif sera de tester le taux d'identification de l'origine dialectale à partir de l'arabe standard.

Références

ABU-HAIDAR, F. (1991). *Variabilité et invariance du système vocalique de l'arabe standard.* Thèse de Doctorat. Université de Besançon.

BACCOUCHE, T., ATTIA, A. (1969). *Travaux de phonologie. Parlers de Djemmal, Gabès et Mahdia*. Cahiers du CERES, Tunis.

BARKAT, M. (1999). Identification of Arabic dialects and experimental determination of discriminative acoustic cues . *The International Congress of Phonetic Sciences*, pp.901-903.

BARKAT, M., OHALA, J., PELLEGRINO, F. (1999). Prosody as a Distinctive Feature for the Recognition of Arabic Dialects. *In Eurospeech'99*, pp. 395-398, Budapest.

BARKAT, M. (2000). *Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes*. Thèse de Doctorat. Université Lumière Lyon 2.

BARKAT, M. (2001). Vers l'identification automatique des parlers arabes. *Revue Langue, Revue Internationale de Linguistique*. pp. 47-73, Maroc.

BARKAT, M., MEUNIER, F., SYNUKYUV, L., MAGRIN-CHANGNOLLEAU, I. (2002). *Intelligibility of Time-Reversed Speech in French.* TIPS (Avril 2002), Aix-en-Provence.

BOERSMA, P., WEENINK, D. (2012). *Praat, a system for doing phonetics by computer*. Glot International 5:9/10, pp. 341-345.

BOULA De MAREUIL, P., BRAHIMI, B., GENDROT, C. (2004). Role of segmental and suprasegmental cues in the perception of Maghrebian-accented French. *8th International Conference on Spoken Language Processing,* pp. 341–344.

CANTINEAU, J. (1960). *Etude de linguistique arabe.* Klincksieck, Paris.

COHEN, D. (1970). *Koiné, langue commune et dialectes arabes*. Janua Linguarum, Série Practica , Mouton.

COHEN, D. (1973). Pour un atlas linguistique de l'arabe. *In Actes du 1er Congrès d'étude des cultures méditerranéennes d'influence arabo-berbère*, pp. 63-69, Alger, Algérie.

EMBARKI, M. (2004).Quantité vocalique et nature du lexique en arabe Marocain. *JEP XXVe (JEP 2004)*, Fès, Maroc.

EMBARKI, M., GUILLEMINOT, C., YEOU, M. (2007). Le voisement en arabe moderne, un indice de classement dialectal. *Rencontre internationale Typologie des parlers arabes*, Paris.

EMBARKI, M. (2008). Les dialectes arabes modernes : état et nouvelles perspectives pour la classification géo-sociologique. *Arabica 55*, pp. 583-604.

EMBARKI, M. (2012). Clarté de la parole et effets coarticulatoires en arabe standard et dialectal. *Actes de la conférence conjointe JEP-TALN-RECITAL, volume 1: JEP,* pp. 209–216, Grenoble.

GHAZALI, S. (1979). *Du statut des voyelles en arabe. Analyses et Théorie*, pp. 199-217.

GHAZALI, S., BRAHAM, A. (1992). Voyelles longues et voyelles brèves en arabe standard : organisation temporelle. *JEP XIXe*, pp. 19-24, Bruxelles.

GHAZALI, S., HAMDI, R., BARKAT, M. (2002). Rhythm variations in Arabic Dialects. *Speech Prosody*, pp. 331-34, Aix-en-Provence.

HAMDI, R., BARKAT, M., PELLEGRINO, F. (2004). De la caractérisation linguistique à l'identification automatique des dialectes arabes. *Actes de Workshop MIDL (Carré des Sciences 2004)*, Paris.

JOHNSON, K. (2000). Adaptative dispersion in vowel perception. *Phonetica, 57,* pp181-188.

JOMAA, M. (1992). *Organisation temporelle acoustique et articulatoire de la quantité en arabe tunisien : comparaisons avec l'arabe standard et l'arabe koweitien*. Thèse de Doctorat. Université Stendhal, Grenoble III.

RJIABI-SABHI, S. (1993). *Approches historique, phonologique et acoustique de la variabilité dialectale arabe : caractérisation de l'origine géographique en arabe standard,* Thèse de Doctorat. Université de Besançon.

Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit

Lionel Fontan Cynthia Magnen Julien Tardieu Pascal Gaillard
U.R.I. Octogone (EA 4156), Université de Toulouse II
lionel.fontan@univ-tlse2.fr, cynthia.magnen@univ-tlse2.fr,
julien.tardieu@univ-tlse2.fr, pascal.gaillard@univ-tlse2.fr

RÉSUMÉ

L'objectif de cette étude est de fournir des données de référence sur l'intelligibilité et la compréhension de la parole en français dans les cas de presbyacousie. Les principaux effets de la presbyacousie ont été simulés afin d'obtenir un jeu de stimuli reproduisant les caractéristiques de la parole perçue par des patients atteints à divers degrés de sévérité, que ce soit dans des conditions de bruit ambiant (brouhaha) ou de silence. Soixante participants ont écouté ces stimuli et exécuté trois tâches situées sur le continuum entre l'intelligibilité et la compréhension de la parole : la répétition de mots (T1), la répétition de phrases (T2) et la réponse à des consignes verbales (T3). Tout en offrant des perspectives différentes et complémentaires sur les performances des auditeurs, les trois tests révèlent des effets significatifs de la simulation de la presbyacousie et du bruit.

ABSTRACT

Simulation of the effects of presbycusis on the intelligibility and comprehension of speech in noisy and quiet environments

This study deals with the effects of presbycusis on speech intelligibility and comprehension in French, both in quiet and noisy conditions. Speech stimuli representative of hearing losses at various ages have been created by simulating the main effects of presbycusis on audio signals. Sixty participants listened to these stimuli and were asked to perform three tasks situated along the continuum between speech intelligibility and comprehension : repeating words (T1), repeating sentences (T2) and responding to oral commands (T3). While offering different and complementary perspectives on listeners' performances, the three tests reveal significant effects of presbycusis simulation and noise.

MOTS-CLÉS : Presbyacousie, intelligibilité, compréhension, parole dans le bruit.

KEYWORDS: Presbycusis, speech intelligibility, speech comprehension, speech in noise.

1 Introduction

La *presbyacousie* est une atteinte cochléaire bilatérale et relativement symétrique liée au vieillissement naturel du système auditif. Le phénomène survient chez des sujets âgés de plus de 50 ans et se caractérise par une diminution progressive et irrémédiable de l'acuité auditive.

Cette surdité de perception se caractérise d'un point de vue fonctionnel par une perte de cellules sensorielles qui débute au départ de la cochlée et lèse en premier les cellules ciliées externes (CCE). Dès lors, la sensibilité et la sélectivité fréquentielles disparaissent, de même que la capacité à séparer le signal du bruit ambiant (Bouccara *et al.*, 2005). Du point de vue audiométrique, l'atteinte se traduit par une chute sélective dans les fréquences aiguës (les seuils étant relativement conservés jusqu'à 1 kHz), ce qui provoque des difficultés de compréhension de la parole, en particulier en milieu bruyant. Par ailleurs, le rôle de compression de l'intensité sonore que jouent les CCE est également largement atténué. En résulte une distorsion de la sensation d'intensité : les sujets deviennent progressivement intolérants aux sons forts produits dans leur entourage.

Ces pertes auditives pouvant s'accompagner de dysfonctionnements cognitifs également causés par le vieillissement, les conséquences de la presbyacousie sur le traitement de la parole sont difficiles à mettre en lumière. Une solution est alors de simuler les effets de la pathologie sur des signaux de parole, que l'on fait ensuite écouter à des sujets sains.

1.1 Simulation des effets de la presbyacousie

Il existe deux grandes approches pour simuler les pertes auditives liées à l'âge (Arriöz, 2012). La première consiste à masquer la parole dans du bruit pour simplement réduire l'audibilité du signal. La seconde approche implique d'effectuer des traitements sur les signaux de parole pour simuler les effets produits par le système auditif déficient. Sont alors pris en compte les trois effets qui accompagnent la plupart des pertes auditives (Moore, 2007) :

1. l'augmentation du seuil d'audibilité. L'intensité du signal de parole est réduite de manière plus ou moins importante en fonction des bandes de fréquences concernées ;
2. la réduction de la sélectivité fréquentielle. La procédure habituellement utilisée, décrite dans Baer et Moore (1993), consiste à effectuer un lissage du spectre fréquentiel ;
3. le recrutement de sonie. L'enveloppe du signal est élevée à la puissance 2 ou 3 pour accroître les variations d'intensité (Moore et Glasberg, 1993).

Ces trois traitements sont appliqués à divers degrés en fonction de la sévérité du cas de presbyacousie dont on cherche à simuler les effets.

1.2 Effets de la presbyacousie sur l'intelligibilité de la parole

De manière générale, plus la presbyacousie est importante, plus elle dégrade l'intelligibilité de la parole. Plusieurs auteurs ont montré que pris ensemble, les différents facteurs de dégradation décrits ci-dessus n'ont pas les mêmes conséquences sur l'intelligibilité de la parole, et que les effets sont aggravés par la présence d'un bruit de fond. Par exemple, Baer et Moore (1993) ont montré que l'effet du lissage du spectre fréquentiel sur l'intelligibilité de la parole est beaucoup plus important lorsque les signaux de parole sont diffusés dans du bruit. Pour que la simulation des effets de l'âge sur l'intelligibilité de la parole soit réaliste, il faut donc combiner les différents facteurs de dégradation associés à la presbyacousie ainsi que la présence ou non d'un bruit de fond (Moore, 2007; Nejime et Moore, 1997).

1.3 Intelligibilité et comprehension de la parole

Contrairement aux tests d'intelligibilité qui se concentrent sur la perception d'unités linguistiques, les tests de compréhension visent à vérifier la capacité de l'auditeur à interpréter des énoncés entendus, en contexte ou non. Les tests de compréhension portent donc sur des énoncés complets (phrases) et engagent l'auditeur dans différentes tâches dont la réalisation implique leur traitement sémantique. Les auditeurs peuvent ainsi être amenés à répondre à des questions sur le contenu sémantique d'un texte entendu, à apparier des images avec les énoncés qu'elles illustrent, ou encore à réagir à des phrases impératives en exécutant les actions commandées sur des objets.

Les données recueillies dans plusieurs études ont suggéré que les scores d'intelligibilité de la parole n'étaient pas ou très peu corrélés avec les scores obtenus dans des tâches tournées vers la compréhension de l'auditeur (Fontan, 2012; Hustad, 2008). Ces études suggèrent en effet que les tests d'intelligibilité et de compréhension de la parole offrent des perspectives différentes sur le traitement d'énoncés linguistiques par des auditeurs. Les tests d'intelligibilité, laissant moins de part à la mobilisation de connaissances extralinguistiques par l'auditeur, ont une validité interne accrue : les mesures sont plus sensibles aux variations dans la qualité du signal de parole. D'un autre côté les tests de compréhension de la parole peuvent être envisagés comme ayant une meilleure validité externe (une meilleure représentativité), car les tâches qu'ils impliquent s'approchent davantage de situations de communication réelles, dans lesquelles les auditeurs peuvent exploiter leurs connaissances encyclopédiques ou bien liées au contexte de communication pour interpréter les énoncés (Fontan, 2012). En conséquence, pour capturer l'image la plus exhaustive des effets de la presbyacousie sur le traitement d'énoncés linguistiques perçus dans le silence et dans le bruit, il est nécessaire de convoquer différents tests situés le long du continuum entre l'intelligibilité et la compréhension de la parole.

1.4 Objectif de l'étude

L'objectif de l'étude est de mesurer les effets de la presbyacousie simulée sur l'intelligibilité et la compréhension de la parole en français, dans le silence et dans le bruit.

En accord avec les études précitées, nous nous attendons à ce que la dégradation du signal par la simulation des effets de la presbyacousie, ainsi que par l'ajout de bruit, provoque une baisse générale des scores d'intelligibilité et de compréhension de la parole. Nous émettons également l'hypothèse que ces effets seront d'autant plus importants que la tâche dans laquelle est engagée l'auditeur ne lui permet pas de mobiliser des connaissances externes au signal de parole. Dit autrement, les scores d'intelligibilité devraient faire preuve d'une plus grande sensibilité aux dégradations du signal de parole que les scores de compréhension.

2 Méthode

2.1 Dessein experimental général

Afin de couvrir le continuum partant de l'intelligibilité vers la compréhension présenté ci-dessus, trois tests ont été développés.

T1 – Répétition de mots. Ce type de test, très utilisé par les audioprothésistes, consiste pour l'auditeur à répéter des mots dissyllabiques. Le matériel sonore du test a été sélectionné à partir d'un sous ensemble des listes de Fournier (Fournier, 1951). Le corpus est constitué de 10 mots d'entraînement et 60 mots de tests.

T2 – Répétition de phrases. Les phrases utilisées dans ce second test sont issues de la version française du test HINT (pour *Hearing in Noise Test*, développé initialement en anglais par Nilsson *et al.*, 1994; Vaillancourt *et al.*, 2005, pour la version française). Ce test est également très utilisé par les audioprothésistes pour évaluer l'intelligibilité de la parole en milieu bruité. Le corpus est constitué de 10 phrases d'entraînement et de 60 phrases de test.

T3 - Exécution de commandes verbales. Ce test de compréhension (Fontan *et al.*, 2014; Fontan, 2012) consiste à demander aux auditeurs de répondre à des consignes verbales leur demandant de déplacer des images sur un écran d'ordinateur (ex. : « Mettez - objet n°1 - position - objet n°2 »). Les phrases du test ont été élaborées afin que la proportion de déplacements à gauche/droite et au-dessus/au-dessous soit la plus équilibrée possible, et que chaque objet cible n'apparaisse qu'une fois dans le test. Au total, 10 phrases d'entraînement (ex : « Mettez l'hélicoptère au dessus de la girafe ») et 30 phrases de test ont été définies. Pour chaque phrase, les deux images cibles sont accompagnées de quatre autres images pouvant induire l'auditeur en erreur.

2.2 Participants

Soixante participants (nombre de femmes = 34, âge moyen = 21,3 écart type = 2,2) ont été sélectionnés et rémunérés pour cette expérience. Les critères d'inclusion étaient les suivants : francophones natifs (langue maternelle française, ayant toujours vécu en France), âgés de 18 à 30 ans inclus, étudiant dans des disciplines autres que musique, sciences du langage, langues étrangères ou psychologie, sans problème de vue non corrigé par des lentilles ou des lunettes. Enfin, le niveau d'audition de chaque participant a été vérifié par un audiogramme tonal (critère d'inclusion : moyenne des pertes entre 2kHz et 8kHz < 15 dB).

2.3 Stimuli

Afin que nos enregistrements soient représentatifs d'un panel de voix variées, deux locuteurs adultes (un homme et une femme, 46 et 47 ans) et une enfant de 12 ans ont été enregistrés dans la cabine audiométrique PETRA (http ://petra.univ-tlse2.fr), avec un microphone omnidirectionnel Sennheiser MD46, une console de mixage TASCAM DM-3200 et un ordinateur MacPro équipé du logiciel Reaper (http ://www.reaper.fm).

Pour simuler la présence d'un bruit ambiant, nous avons choisi de créer un bruit de type « brouhaha » (*babble noise*). Pour cela nous avons mixé les phrases enregistrées pour T2, en prenant soin d'alterner entre les locuteurs ainsi qu'en introduisant un décalage temporel aléatoire entre les phrases (cette procédure correspond à celle suivie dans Fontan, 2012). Ce fichier de bruit a ensuite été mixé aux phrases et aux mots avec un rapport signal sur bruit de 5 dB A.

Les stimuli de parole avec et sans bruit ont ensuite été traités pour simuler les effets de la presbyacousie. La simulation a été effectuée dans MATLAB à partir des algorithmes initialement développés par Nejime et Moore (1997). À partir des données issues de la littérature (étude

épidémiologique sur 3753 sujets de Cruickshanks *et al.*, 1998), et afin de couvrir une large plage de presbyacousie, dix niveaux de dégradation ont été définis : un premier niveau correspondant au signal non dégradé ; puis 9 niveaux distribués linéairement de l'âge théorique de 60 ans à 110 ans (dégradation maximale).

2.4 Procédure

Chaque participant était invité à passer deux des trois tests, selon la répartition suivante : (1) Groupe 1 (30 participants) : T1 et T3, (2) Groupe 2 (30 participants) : T2 et T3 [1]. Dans chaque groupe, les deux tests étaient passés en alternant l'ordre de présentation afin d'éviter un éventuel effet d'entrainement à la voix des locuteurs.

Les trois tests ont été effectués avec le même équipement que celui qui a permis l'enregistrement des stimuli (*cf.* section 2.3). Les participants étaient assis à un mètre des haut-parleurs (Tannoy Precision 6D), avec un niveau de diffusion des phrases non dégradées (condition 1) de 60 dB A. L'interface dédiée aux tests T1 et T2 et créée avec le logiciel Max/MSP permettait la présentation audio des stimuli et l'enregistrement sonore des réponses données par les participants. Les réponses ont ensuite été transcrites par les expérimentateurs afin de calculer les scores d'intelligibilité. Concernant T3, les images, l'interface et le logiciel permettant de récupérer les réponses des auditeurs proviennent d'une étude consistant à évaluer la compréhension de la parole pathologique (Fontan, 2012; Fontan *et al.*, 2014).

3 Résultats

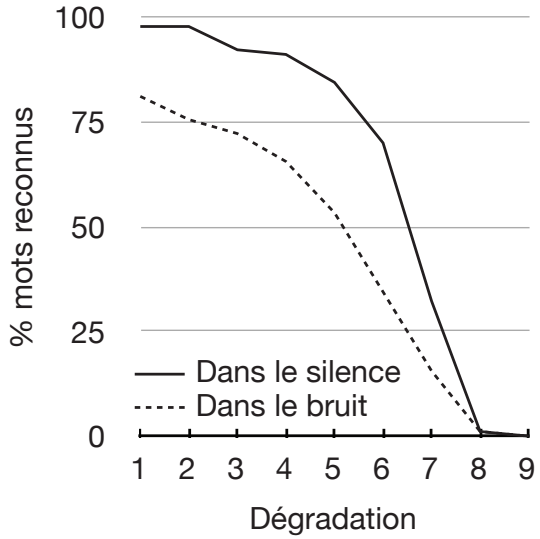
3.1 T1 : intelligibilité de listes de mots

Pour chaque mot, le score est égal à 100 % si le mot est totalement reconnu (l'article précédent le mot n'est pas pris en compte) ou 0 % si il n'est que partiellement reconnu ou non reconnu. Les résultats sont présentés sur la Figure 1 et montrent que plus la dégradation du signal de parole augmente, plus les scores de reconnaissance diminuent, aussi bien dans le silence que dans le bruit. On note aussi que dans le bruit, les scores sont toujours plus faibles que dans le silence (différence de 18,6 % en moyenne). Les données ont été analysées en utilisant un modèle linéaire mixte généralisé, et ont révélé un effet significatif de la dégradation ($\chi^2(9) = 205,3; p < 0,001$) et du bruit ($\chi^2(1) = 38,9; p < 0,001$).

3.2 T2 : intelligibilité de listes de phrases

Pour ce deuxième test, tous les mots des phrases sont pris en compte dans le calcul des scores de reconnaissance, le score est égal au nombre de mots reconnus divisé par nombre de mots total dans la phrase. Les résultats de ce test sont présentés sur la Figure 2. On observe les mêmes tendances que pour le test 1 avec cette fois une différence moyenne entre les scores dans le

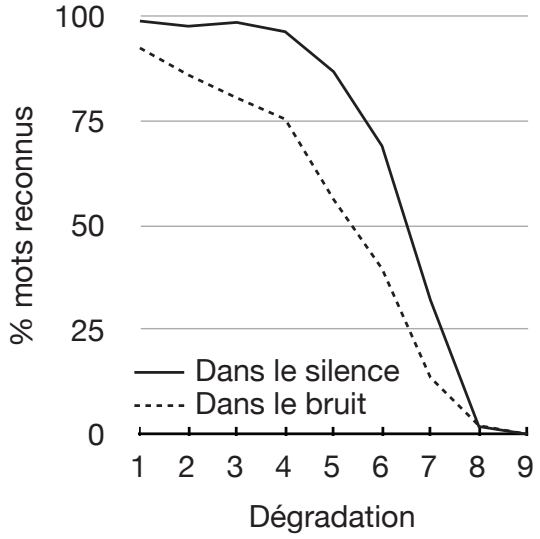
1. Nous avons volontairement doublé le nombre de sujets pour T3 car nous envisageons dans une étape ultérieure de faire une étude portant précisément sur les effets de contexte impliqués dans ce test.



	% mots reconnus			
Dég.	Silence		Bruit	
1	97,8	(14,8)	81,1	(39,3)
2	97,8	(14,8)	75,6	(43,2)
3	92,2	(26,9)	72,2	(45,0)
4	91,1	(28,6)	65,6	(47,7)
5	84,4	(36,4)	53,3	(50,1)
6	70,0	(46,0)	34,4	(47,7)
7	32,2	(46,9)	15,6	(36,4)
8	1,1	(10,5)	1,1	(10,5)
9	0	(0)	0	(0)

FIGURE 1 – T1 : pourcentages de mots reconnus en fonction du niveau de dégradation, dans le silence et dans le bruit. Tableau : pourcentages moyens et écarts-types entre parenthèses

silence et le bruit de 15 %. Les données ont été analysées en utilisant un modèle linéaire mixte et ont révélé un effet significatif du niveau de dégradation ($F(9; 1711) = 572,8; p < 0,001$) et du bruit ($F(1; 1711) = 188,3; p < 0,001$).



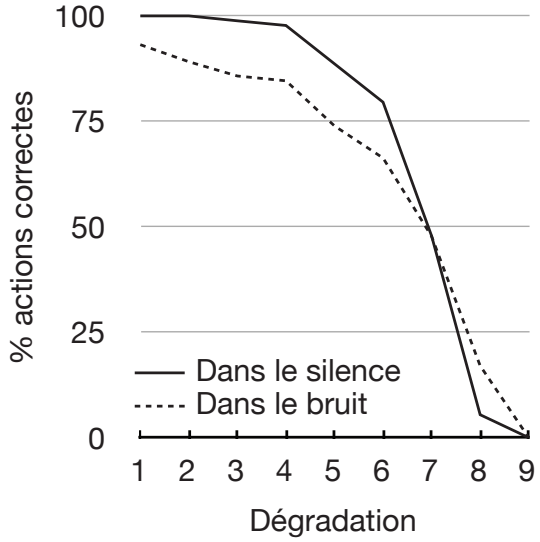
	% mots reconnus			
Dég.	Silence		Bruit	
1	98,9	(8,7)	92,4	(15,4)
2	97,7	(8,3)	86,0	(22,0)
3	98,6	(7,5)	80,5	(28,7)
4	96,3	(9,1)	75,4	(33,3)
5	86,8	(23,7)	56,2	(38,0)
6	69,0	(37,0)	39,6	(37,7)
7	32,1	(37,3)	13,6	(24,8)
8	1,9	(6,9)	2,1	(8,2)
9	0	(0)	0	(0)

FIGURE 2 – T2 : pourcentages de mots reconnus dans les phrases en fonction du niveau de dégradation, dans le silence et dans le bruit. Tableau : pourcentages moyens et écarts-types entre parenthèses

3.3 T3 : compréhension de listes de phrases

Pour ce test, le score est égal à 100% si l'action est correctement réalisée ou 0% si elle ne l'est que partiellement ou pas du tout. Les scores sont présentés sur la Figure 3. Ici aussi les

scores décroissent avec la dégradation et sont plus faibles dans le bruit (différence moyenne de 6,7 %). Les données ont été analysées en utilisant un modèle linéaire mixte, et ont révélé un effet significatif de la dégradation ($F(9;551) = 75,7; p < 0,001$) et du bruit ($F(1;551) = 4,6; p < 0,05$).



	% actions correctes			
Dég.	Silence		Bruit	
1	100	(0)	93,2	(25,4)
2	100	(0)	89,1	(31,3)
3	98,8	(10,7)	85,8	(35,2)
4	97,7	(15)	84,6	(36,3)
5	88,6	(31,9)	73,9	(44,2)
6	79,5	(40,6)	66,3	(47,6)
7	47,8	(50,2)	47,7	(50,2)
8	5,4	(22,8)	17,0	(37,8)
9	0	(0)	0	(0)

FIGURE 3 – T3 : pourcentages d'actions correctes en fonction du niveau de dégradation, dans le silence et dans le bruit. Tableau : pourcentages moyens et écarts-types entre parenthèses

4 Discussion

L'étude présentée dans cet article a permis de mesurer les effets de la simulation de la presbyacousie – pour des âges théoriques compris entre 60 et 110 ans – sur les scores obtenus par 60 participants à deux tests d'intelligibilité (T1 : répétition de mots et T2 : répétition de phrases) et à un test de compréhension (T3 : exécution de commandes verbales).

Dans un premier temps et comme attendu, les résultats observés rejoignent ceux des travaux précédents sur la simulation de la presbyacousie (Moore, 2007) et montrent que les effets produits sur le signal de parole conduisent à une perte d'intelligibilité significative et importante pour l'auditeur. Dans cette étude la même observation a été faite concernant cette fois la compréhension de la parole, testée à travers la réaction des auditeurs à des commandes verbales.

La seconde hypothèse testée était que selon la nature du test utilisé, les scores ne démontreraient pas la même sensibilité aux deux dégradations étudiées. Nos données semblent confirmer cette hypothèse. Tout d'abord, les courbes des scores observées montrent que les valeurs plafond et plancher sont atteintes plus rapidement dans T3 que dans T2, et plus rapidement dans T2 que dans T1. C'est en effet dans ce dernier test que les scores suivent l'évolution la plus linéaire, et en particulier pour la condition bruitée.

Cette observation va dans le sens d'une plus grande sensibilité des tests d'intelligibilité, pour lesquels de faibles variations dans la qualité du signal se traduisent immédiatement par des variations dans les scores des participants. Les différences de résistance au bruit observées

dans les trois tests viennent également appuyer cette seconde hypothèse : la chute des scores provoquée par l'ajout de bruit est moins importante pour le test de compréhension que pour les tests d'intelligibilité ; et parmi les deux tests d'intelligibilité convoqués, la baisse des scores est moins importante pour le test de répétition de phrases que pour le test de répétition de mots.

Ces résultats renforcent les conclusions qui ont découlé d'études sur les différences entre l'intelligibilité et la compréhension de la parole pathologique (Fontan, 2012; Hustad, 2008), et selon lesquelles les scores d'intelligibilité présentent de meilleures caractéristiques de validité interne (meilleure sensibilité), alors que les scores de compréhension offrent davantage de validité externe (meilleure représentativité de la dimension que l'on cherche à mesurer). Dans le cas présent les auditeurs ont pu mobiliser davantage de connaissances de haut niveau dans le test T3 (grâce à la présence du contexte imagé) et le test T2 (grâce à la présence du contexte phrastique), ce que l'on peut concevoir comme une situation plus proche d'une communication réelle que la tâche de répétition de mots isolés.

5 Conclusions et perspectives

En conclusion, cette étude offre différentes perspectives sur les performances d'auditeurs atteints de presbyacousie, dans le silence et dans le bruit. Le corpus de stimuli constitué et les scores associés dans les tests d'intelligibilité peuvent servir de référence pour des études reposant sur l'observation d'évolutions subtiles dans les performances de l'auditeur (par ex. en vue d'optimiser de fins réglages dans les prothèses auditives) et les scores de compréhension de la parole peuvent donner une indication plus globale de la capacité de l'auditeur à communiquer dans la vie quotidienne.

Ce travail de constitution de données de référence pourra être poursuivi en affinant les mesures relevées dans cette étude. Les scores d'intelligibilité pourront par exemple être nuancés en prenant en compte le nombre de phonèmes reconnus par les 60 auditeurs pour chaque stimulus. Concernant le test de compréhension, les effets de la distance phonologique entre les noms correspondant aux images cibles et ceux correspondant aux autres images présentes dans le contexte pourront être étudiés ; ceci permettra de mieux contrôler les effets de contexte et ainsi d'accroître la validité et la précision des mesures de compréhension.

À terme, les données quantitatives et qualitatives relevées serviront à l'étude de la prédictibilité des performances d'auditeurs atteints de presbyacousie. Dans cette optique, la voie que nous privilégions est le rapprochement des données subjectives d'intelligibilité et de compréhension de la parole avec les performances objectives de systèmes informatiques de reconnaissance automatique de la parole.

Remerciements

Cette étude a été réalisée dans le cadre d'un projet financé par la Région Midi-Pyrénées.

Références

ARRIÖZ, U. (2012). *Developping subject-specific hearing loss simulation to apply different frequency lowering algorithms for the enhancement of sensorineural hearing losses*. Middle-East Technical University.

BAER, T. et MOORE, B. C. J. (1993). Effects of spectral smearing on the intelligibility of sentences in the presence of noise. *Journal of the Acoustical Society of America*, 94(4):1229–1241.

BOUCCARA, D., FERRARY, E., MOSNIER, I., BOZORG GRAYELI, A. et STERKERS, O. (2005). Presbyacousie. *EMC – Oto-Rhino-Laryngologie*, 2(4):329–342.

CRUICKSHANKS, K. J., WILEY, T. L., TWEED, T. S., KLEIN, B. E., KLEIN, R., MARES-PERLMAN, J. A. et NONDAHL, D. M. (1998). Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin. The Epidemiology of Hearing Loss Study. *American Journal of Epidemiology*, 148:879–886.

FONTAN, L. (2012). *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication*. Université de Toulouse.

FONTAN, L., GAILLARD, P. et WOISARD, V. (2014). Comprendre et agir : les tests pragmatiques de compréhension de la parole et EloKanz. *In* SOCK, R., VAXELAIRE, B. et FAUTH, C., éditeurs : *Travaux en phonétique clinique*, pages 131–144, Mons (Belgique). CIPA. (sous presse).

FOURNIER, J.-E. (1951). *Audiométrie vocale : les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités*. Maloine.

HUSTAD, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language and Hearing Research*, 51(3):562–573.

MOORE, B. C. J. (2007). *Cochlear hearing loss : Physiological, psychological and technical Issues*. Wiley.

MOORE, B. C. J. et GLASBERG, B. R. (1993). Simulation of the effects of loudness recruitment and treshold elevation on the intelligibility of speech in quiet and in background of speech. *Journal of the Acoustical Society of America*, 94(4):2050–2062.

NEJIME, Y. et MOORE, B. C. J. (1997). Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise. *Journal of the Acoustical Society of America*, 102:603–615.

NILSSON, M., SOLI, S. et SULLIVAN, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, 95(2):1085–1099.

VAILLANCOURT, V., LAROCHE, C., MAYER, C., BASQUE, C., NALI, M., ERIKS-BROPHY, A., SOLI, S. D. et GIGUÈRE, C. (2005). Adaptation of the HINT (Hearing in Noise Test) for adult Canadian Francophone populations. *International Journal of Audiology*, 44:358–369.

Combinaison de mots et de syllabes pour transcrire la parole

Luiza Orosanu[1, 2, 3] Denis Jouvet [1, 2, 3]
Équipe PAROLE, LORIA
(1) Inria, Villers-lès-Nancy, F-54600, France
(2) Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
(3) CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{luiza.orosanu, denis.jouvet}@loria.fr

RÉSUMÉ

Cet article analyse l'intérêt de modèles de langage hybrides pour transcrire de la parole. L'objectif est d'utiliser une telle solution pour aider à la communication avec des personnes sourdes, et de la mettre en œuvre sur un terminal portable, ce qui introduit des contraintes sur la taille du modèle. Les unités linguistiques considérées pour cette tâche sont les mots et les syllabes. Des lexiques de différentes tailles sont obtenus en variant le seuil de sélection associé aux fréquences d'occurrence des mots dans les données d'apprentissage, les mots les moins fréquents sont alors décomposés en syllabes. Ce type de modèle de langage peut reconnaître entre 69% et 96% des mots (le reste étant représenté par des syllabes). En ajustant le seuil sur les mesures de confiance associées aux mots reconnus, les hypothèses de mots les plus fiables peuvent être identifiées (à un taux de bonne reconnaissance variant entre 70% et 92%).

ABSTRACT

Combining words and syllables for speech transcription

This paper analyzes the use of hybrid language models for automatic speech transcription. The goal is to later use such an approach as a support for helping communication with deaf people, and to run it on an embedded decoder on a portable device, which introduces constraints on the model size. The main linguistic units considered for this task are the words and the syllables. Various lexicon sizes are studied by setting thresholds on the word occurrence frequencies in the training data, the less frequent words being therefore syllabified. Using this kind of language model, the recognizer can output between 69% and 96% of the words (whereas the other words, will be represented by syllables). By setting different thresholds on the confidence measures associated to the recognized words, the most reliable word hypotheses can be identified, and they have correct recognition rates between 70% and 92%.

MOTS-CLÉS : modèle de langage hybride, mots, syllabes, mots hors vocabulaire, mesure de confiance, surdité.

KEYWORDS: hybrid language model, words, syllables, out-of-vocabulary words, confidence measure, deaf people.

1 Introduction

Le projet RAPSODIE porte sur l'extraction d'informations pertinentes à partir du signal de parole, afin d'aider à la communication avec des personnes sourdes ou malentendantes. Compte tenu des contraintes imposées par les ressources disponibles (capacité mémoire et puissance de calcul) avec un système embarqué, la solution optimale doit reposer sur le meilleur compromis entre coût de calcul et performance pour le modèle de reconnaissance, ainsi que sur la meilleure manière de présenter l'information reconnue (en mots, syllabes, phonèmes ou en les combinant).

Au cours des dernières décennies, les scientifiques ont essayé d'offrir une meilleure compréhension de la parole pour la communauté des sourds, en affichant des phonèmes pour aider la lecture labiale (Sokol, 1996), en affichant des informations en langage des signes grâce à un avatar (Cox *et al.*, 2002) et, bien entendu, en affichant des sous-titres générés d'une manière semi-automatique ou entièrement automatique. Les aspects ergonomiques et les conditions d'utilisation de la reconnaissance vocale pour aider les personnes sourdes ont été analysés dans (Woodcock, 1997).

L'un des principaux inconvénients des systèmes de reconnaissance de la parole est leur incapacité à reconnaître les mots qui n'appartiennent pas à leur vocabulaire. Compte tenu de la quantité limitée des données d'apprentissage, et aussi de la capacité mémoire et de la puissance de calcul limitées des systèmes de traitement, en particulier, pour des systèmes intégrés dans un dispositif portable, il est impossible de concevoir un système de reconnaissance qui couvre tous les mots, et encore moins tous les noms propres et les abréviations. Par conséquent, un système de reconnaissance grand vocabulaire (basé sur les mots) ne peut pas être une solution idéale.

IBM a ainsi testé le sous-titrage en phonétique du discours d'un orateur, avec le système appelé LIPCOM (Coursant-Moreau et Destombes, 2002). Le système de reconnaissance était monolocuteur, et a été testé dans une école pour enfants sourds. L'application était basée sur un décodage phonétique (sans aucun vocabulaire préalablement défini) et le résultat était affiché en phonèmes codés sur une ou deux lettres.

Toutefois, étant donné que notre objectif est de trouver une solution adaptée aux personnes sourdes et malentendantes, les priver de transcriptions basées sur des mots serait une erreur. Par conséquent, l'utilisation en complément de mots, d'unités de sous-mots, pourrait résoudre ce problème. Cela éviterait d'afficher systématiquement des informations erronées quand des mots hors vocabulaire sont prononcés ; les parties hors vocabulaire seraient affichées en sous-mots. Le résultat serait alors un texte lisible et compréhensible, et la taille raisonnable du lexique limiterait les besoins en ressources de traitement.

Des études ont été menées pour étendre les lexiques de mots avec des fragments de mots, dans le but de diminuer les erreurs liées aux mots hors vocabulaire. La méthode proposé dans (Yazgan et Saraclar, 2004) utilise un modèle de langage hybride qui combine des mots avec des unités de sous-mots tels que les phonèmes ou les syllabes. Une étude sur les vocabulaires ouverts a été faite aussi dans (Bisani et Ney, 2005) où les mots et des fragments de mots ont été mélangés conjointement dans un modèle de langage hybride ; les fragments de mots résultent des séquences de lettres obtenues avec les multigrammes conjoints (sequences de lettres et séquences de phonèmes). Dans (Rastrow *et al.*, 2009), les unités de sous-mots correspondent à des séquences de phonèmes de longueurs variables déterminées automatiquement à partir de données ; une telle extension est censée fournir des meilleures correspondances acoustiques

sur les parties hors-vocabulaire du signal de parole, ce qui réduit le taux d'erreur phonétique. Des études plus récentes (Shaik *et al.*, 2011) ont examiné la combinaison de plus de deux types d'unités lexicales dans le même vocabulaire et modèle de langage : les mots les plus fréquents sont conservés, les autres étant remplacés par des morphèmes graphémiques ou des syllabes.

L'apport des mesures de confiance (Jiang, 2005) dans l'emploi de transcriptions automatiques de parole pour des personnes sourdes a été analysé dans (Razik *et al.*, 2008). Des tests subjectifs ont même montré une préférence pour l'affichage sous forme phonétique pour les mots avec une faible mesure de confiance. Cependant, le phonème présente de nombreuses irrégularités dans ses réalisations acoustiques et une unité de reconnaissance plus grande devrait être considérée pour capter les variations comme celles introduites par la coarticulation.

La syllabe a été étudiée dans le passé en tant qu'unité acoustique (Cerf-Danon *et al.*, 1989; Wu *et al.*, 1998; Zhang et Edmondson, 2002; Tachbelie *et al.*, 2011), pour la reconnaissance de la parole continue avec un grand vocabulaire, généralement en combinaison avec des phonèmes dépendants du contexte (Ganapathiraju *et al.*, 2001; Hämäaläainen *et al.*, 2005) ou pour le décodage phonétique uniquement (Le Blouch et Collen, 2006). Dans (Wu *et al.*, 1998), la syllabe a été décrite comme une unité attrayante pour la reconnaissance grâce à sa plus grande stabilité, à son lien naturel entre l'acoustique et l'accès lexical et à sa capacité d'intégrer les informations prosodiques dans la reconnaissance. Dans (Le Blouch et Collen, 2006), la coarticulation a été modélisée entre les phonèmes à l'intérieur de la syllabe, mais aucune modélisation dépendante du contexte n'a été prise en compte entre les syllabes, de plus le modèle de langage appliqué au niveau de la syllabe était un bigramme.

Dans nos travaux précédents, nous avons considéré la syllabe comme une unité linguistique pour la création de modèles de langage pour un décodage phonétique (Orosanu et Jouvet, 2013). Même si les résultats ont montré que les meilleures performances sont toujours obtenues avec un système de reconnaissance grand vocabulaire, le choix d'utiliser un modèle de langage syllabique ne doit pas être ignoré. Il offre des performances de décodage phonétique qui sont seulement 4% inférieures à celles fournies par le système de reconnaissance grand vocabulaire, mais nettement supérieures à celles obtenues avec un modèle phonétique et un trigram sur les phonèmes. Cela nous a conduit à considérer un modèle de langage hybride des mots et des syllabes.

Pour se positionner par rapport aux travaux précédents, il faut préciser que nous voulons non seulement réduire les erreurs liées aux portions correspondant aux mots hors vocabulaire, mais aussi maximiser la compréhension de la transcription résultante pour la communauté de personnes sourdes. Pour ce faire, nous avons cherché à modéliser les suites de sons, plutôt que les suites de lettres. Les particularités de la langue française nous ont aussi poussé à chercher une solution qui prenne en compte la prononciation ou non du e-muet ou des phonèmes de liaison. Ce qui explique notre choix final de fabriquer des modèles de langage hybrides basés sur la transcription de prononciations de parole (au niveau mots, et au niveau phonèmes en passant par un alignement forcé). Par conséquent, l'objectif des travaux présentés dans cet article est de déterminer si la combinaison de mots et de syllabes (dans le lexique et le modèle de langage) conduit néanmoins à un bon taux de reconnaissance de mots. L'autre objectif est l'étude de l'apport et de la pertinence des mesures de confiance pour identifier les mots correctement reconnus.

Le papier est organisé comme suit : la section 2 fournit une description du modèle de langage hybride, la section 3 est consacrée à la description des corpus de données et des outils utilisés dans nos expériences, et la section 4 présente et analyse les résultats.

2 Contexte expérimental

2.1 Données

Les corpus de parole utilisés dans nos expériences proviennent des campagnes d'évaluation ESTER2 (Galliano *et al.*, 2009) et ETAPE (Gravier *et al.*, 2012), et du projet EPAC (Estève *et al.*, 2010). Les données d'ESTER2 et d'EPAC sont des bulletins d'informations français recueillis auprès de diverses stations de radio, ils contiennent de la parole "préparée" et de la parole spontanée (par exemple les interviews). Une grande partie des données de parole sont de qualité studio, et certaines parties sont de qualité téléphonique. À l'inverse, les données d'ETAPE correspondent à des débats, et ont été collectées auprès de diverses chaînes de radio et de télévision. Il s'agit essentiellement de parole spontanée.

Les données de parole de l'ensemble d'apprentissage d'ESTER2 et d'ETAPE, ainsi que les données transcrites du corpus EPAC, ont été utilisées pour apprendre les modèles acoustiques. Les données d'apprentissage s'élèvent à environ 300 heures de parole et 4 millions de mots. Le modèle de langage hybride de mots et de syllabes a été appris sur les corpus des transcriptions d'ESTER2, d'ETAPE et d'EPAC (après alignement forcé).

Les variantes de prononciation ont été extraites du lexique BDLEX (de Calmés et Pérennou, 1998) et des lexiques internes, quand disponibles. Pour les mots manquants, les variantes de prononciation ont été obtenues automatiquement en utilisant des convertisseurs graphème-phonème JMM et CRF (Illina *et al.*, 2011).

2.2 Configuration

Les modèles de langage statistiques ont été fabriqués avec les outils SRILM (Stolcke, 2002). Les outils de Sphinx3 (Placeway *et al.*, 1996) ont été utilisés pour apprendre les modèles acoustiques des phonèmes. Le décodage des signaux audio, ainsi que le calcul des mesures de confiance (probabilité a posteriori des mots), a été effectué avec Pocketsphinx (Huggins-daines *et al.*, 2006). L'analyse acoustique MFCC (Mel Frequency Cepstral Coefficients) donne 12 paramètres MFCC et le logarithme de l'énergie par trame (fenêtre de 32 ms, décalage de 10 ms). Les modèles acoustiques HMM des phonèmes utilisent des mélanges de 64 gaussiennes, et sont adaptés aux données hommes et aux données femmes. L'unité acoustique utilisée est toujours le phonème, avec une modélisation dépendante du contexte.

3 Création du modèle de langage hybride

La création d'un modèle de langage hybride, combinant mots et syllabes, implique de constituer un corpus d'apprentissage qui repose sur ces deux unités lexicales. Ensuite, le vocabulaire est défini en sélectionnant que les mots les plus fréquents, i.e. les mots qui apparaissent au moins θ fois dans le corpus d'apprentissage, et les mots hors vocabulaire sont décomposés en syllabes ; les syllabes correspondantes sont alors ajoutées au lexique.

Un point de départ consiste à appliquer un alignement forcé sur les données d'apprentissage,

afin de déterminer quelle variante de prononciation a été effectivement utilisée pour chaque mot prononcé (fréquent ou pas). À noter que les vocabulaires utilisés dans la reconnaissance de la parole peuvent avoir plusieurs variantes de prononciation, et que un ou plusieurs phonèmes peuvent manquer dans certaines d'entre elles. Ces variantes de prononciation sont nécessaires afin de tenir compte des événements de type liaison ou réduction. Ensuite, les mots peu fréquents (non sélectionnés par rapport au seuil choisi) sont remplacés par leur variante de prononciation telle que résultant de l'alignement forcé. Enfin, les séquences de phonèmes ainsi obtenues, correspondant aux segments de parole entre les mots sélectionnés, sont traitées par l'outil de syllabation, qui est basé sur les règles décrites dans (Bigi *et al.*, 2010). Deux principes sont appliqués : une syllabe contient une seule voyelle et une pause désigne la frontière d'une syllabe.

Par exemple, dans la phrase "une femme a été blessée " :
- mots fréquents = {une, femme, a, été}
- mot non-fréquent = {blessée}
- remplacement du mot "blessée" par sa prononciation "b l eh s e" (variante déterminée par alignement forcé) ⇒ "une femme a été b l eh s e"
- structuration de la séquence de phonèmes "b l eh s e" en deux syllabes "_b_l_eh" et "_s_e"

  ⇒ **mots et syllabes :** "une femme a été _b_l_eh _s_e".

L'utilisation d'un modèle de langage hybride vise à assurer une reconnaissance correcte des mots les plus fréquents et à proposer des suites de syllabes pour les segments de parole correspondant à des mots hors vocabulaire. Les syllabes choisies correspondant à des suites de phonèmes, donc ayant une prononciation unique, devraient faciliter la compréhension des segments de parole hors vocabulaire, du moins plus facilement que de devoir interpréter une suite de petits mots dont l'une des variantes de prononciation correspond au segment de parole hors vocabulaire (cas fréquent lorsque le lexique et le modèle de langage n'utilisent pas d'unités sous-lexicales).

Différents seuils minimaux sur la fréquence d'occurrence des mots ont été considérés : $\theta \in \{3, 5, 10, 25, 50, 100, 300\}$. Chaque choix de seuil entraîne la création d'une transcription différente du corpus d'apprentissage (avec un nombre différent de mots et de syllabes), ce qui conduit à un lexique et un modèle de langage correspondant.

Condition	**Données d'apprentissage**					**Modèles de langage**		
	# Mots			**# Syllabes**		**# Lex**	**# 3g**	**Taille** [MO]
	unique	total	couv.	unique	total			
min3occ	31298	3.57 M	98.9%	2287	0.11 M	33585	1.85 M	14.2
min5occ	23052	3.54 M	98.1%	3209	0.19 M	26261	1.86 M	13.9
min10occ	15052	3.49 M	96.7%	4050	0.33 M	19102	1.87 M	13.5
min25occ	8179	3.38 M	93.8%	4739	0.62 M	12918	1.86 M	12.8
min50occ	4895	3.27 M	90.6%	5126	0.91 M	10021	1.83 M	12.3
min100occ	2868	3.13 M	86.7%	5391	1.27 M	8259	1.79 M	11.6
min300occ	1066	2.83 M	78.3%	5730	1.99 M	6796	1.69 M	10.5

TABLE 1: Description des modèles de langage hybrides

Les différentes transcriptions et les modèles de langage associés sont décrits dans le tableau 1. En ce qui concerne les lexiques, les nombres d'entrées uniques (mots d'une part, et syllabes d'autre part) sont mentionnés dans le tableau. Chaque mot dans le lexique correspond à environ deux

variantes de prononciation (essentiellement du fait des liaisons possibles et de la présence ou non de e-muet). Étant donné que les syllabes correspondent à une seule séquence de phonèmes, elles ne peuvent avoir qu'une seule prononciation possible (Orosanu et Jouvet, 2013). Le tableau indique aussi le nombre total d'occurrences des mots fréquents et des syllabes fréquentes dans le corpus d'apprentissage, ainsi que le taux de couverture avec les mots fréquents.

Les modèles de langage utilisées dans notre étude sont des modèles statistiques trigrammes, donc pour chaque séquence de trois unités lexicales, la probabilité de la dernière unité lexicale dépend de l'identité des deux unités qui la précèdent.

Pour compléter la description des modèles, seules les syllabes observées au moins trois fois dans le corpus d'apprentissage sont conservées dans les lexiques et dans les modèles de langage.

4 Résultats et discussions

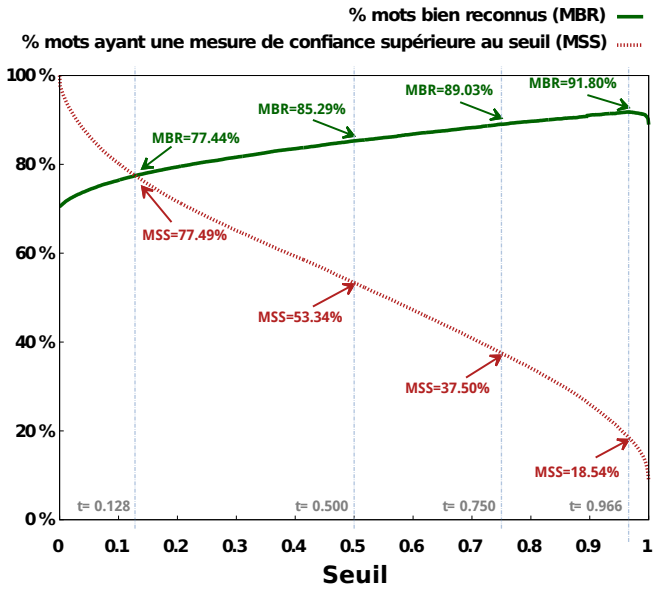
Les ensembles de développement d'ESTER2 (excepté les radios africaines, soit environ 42,000 mots) et d'ETAPE (environ 82,000 mots) sont utilisés dans les expériences rapportées ci-dessous.

Étant donné que les modèles de langage mélangent des mots et des syllabes, et que nous sommes intéressés à récupérer le message porté par la parole, nous avons particulièrement analysé combien de mots ont été produits par le décodeur (dans la séquence de mots et de syllabes), et parmi ces mots, combien d'entre eux ont été effectivement bien reconnus.

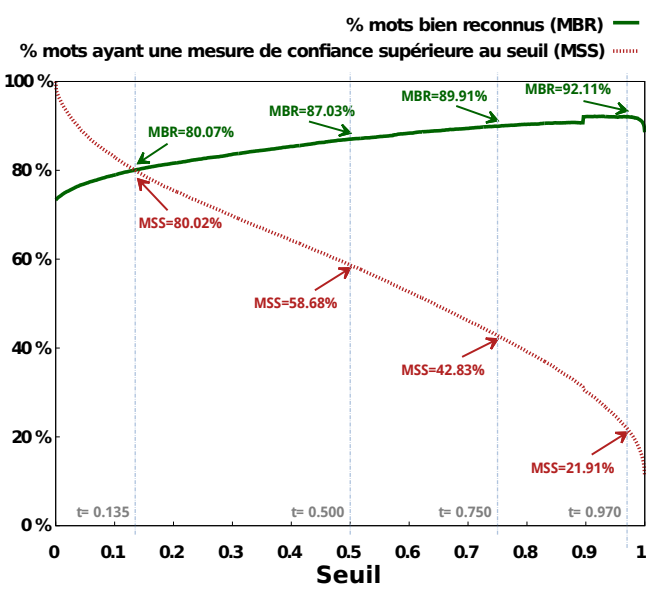
Les résultats pour le corpus ETAPE sont affichés dans le tableau 2 : le modèle de langage qui combine les mots vus au moins 3 fois avec les syllabes vues au moins 3 fois ("min3occ") génère environ 67K mots (couvrant 96,40% d'unités générées par le décodeur), dont 47K ont été bien reconnus (70,03%). Avec le modèle de langage qui combine les mots vus au moins 300 fois et les syllabes vues au moins trois fois ("min300occ"), 71,11% des hypothèses mots sont correctes, mais le nombre de mots reconnus est plus faible (53K, couvrant 68,58% d'unités générées par le décodeur). Pour ESTER2, les pourcentages de mots générés par le decodeur sont similaires aux valeurs obtenues pour ETAPE (une moyenne d'environ 83%). Cependant, les pourcentages de mots bien reconnus sont un peu plus élevés (73%) .

ML	Unités en sortie du décodeur			Mots corrects	
	# mots	# syllabes	% mots	nombre	pourcentage
min3occ	66568	2485	96.40	46616	70.03
min5occ	66085	3787	94.58	46397	70.21
min10occ	65293	5841	91.79	46111	70.62
min25occ	62325	9710	86.52	44195	70.91
min50occ	61627	14405	81.05	43795	70.92
min100occ	59322	19507	75.25	42134	71.03
min300occ	53542	31605	68.58	38075	71.11

TABLE 2: Statistiques sur les séquences d'unités en sortie du décodeur pour le corpus ETAPE : nombre de mots et de syllabes, taux d'unités mots, et statistiques sur mots bien reconnus



(a) Corpus = ETAPE , ML = min3occ



(b) Corpus = ESTER , ML = min3occ

FIGURE 1: Analyse de l'impact du seuil pour l'acceptation de mots, basée sur leurs mesures de confiance : les courbes rouges (MSS) indiquent le pourcentage de mots ayant une mesure de confiance supérieure au seuil, et les courbes vertes (MBR), le pourcentage de ceux bien reconnus

Il est intéressant de filtrer les mots reconnus (fournis pas le décodeur) selon leurs mesures de confiance, afin de maximiser le pourcentage de mots qui sont corrects, i.e. bien reconnus par le système (MBR, "mots bien reconnus"). Le rapport MBR est calculé comme suit :

$$MBR(seuil) = \frac{\#\text{ mots : bien reconnus \& MC} \geq \text{seuil}}{\#\text{ mots : MC} \geq \text{seuil}}$$

La figure 1 montre les performances obtenues par la mise en place de divers seuils sur les mesures de confiance, en utilisant le modèle de langage "min3occ" sur le corpus ETAPE (a) et sur le corpus ESTER (b) (les résultats obtenus avec les autres modèles de langage sont similaires). Un seuil supérieur à 0,96 sur les mesures de confiance conduit à un ratio de mots bien reconnus supérieur à 90% pour les deux corpus. Cependant, un grand seuil rejette de nombreuses hypothèses de mots (par exemple, 80% des hypothèses de mots ont une mesure de confiance inférieure au seuil de 0.96). L'augmentation du seuil réduit le nombre de mots conservés, mais augmente la pertinence des hypothèses conservées. En pratique, un compromis doit être fait entre le nombre de mots qui sont rejetés, et la performance qui peut être obtenue sur les mots non rejetés.

Les mesures de confiance sur les syllabes semblent moins pertinentes ; avec le modèle de langage "min3occ", environ 50% des syllabes ont une mesure de confiance inferieure à 0,07.

5 Conclusions

Cet article a analysé l'intérêt de modèles de langage hybrides pour transcrire de la parole, dans le but d'utiliser une telle solution pour aider à la communication avec des personnes sourdes, et de la mettre en œuvre sur un terminal portable. L'objectif était de déterminer si la combinaison de mots et de syllabes (dans le lexique et le modèle de langage) conduit néanmoins à un bon

taux de reconnaissance de mots. L'autre objectif est l'étude de l'apport et de la pertinence des mesures de confiance pour identifier les mots correctement reconnus.

Les tests ont été effectués sur deux corpus de parole français : ETAPE et ESTER2. Les résultats ont montré que parmi les 69 à 96% de mots qui sont reconnus par le système (les autres étant représentés par des syllabes), environ 70% d'entre eux sont effectivement correctement reconnus. En ajustant le seuil sur les mesures de confiance associées aux mots reconnus, nous avons constaté, comme attendu, que plus nous augmentons le seuil, plus le pourcentage de mots ayant une mesure de confiance supérieure à ce seuil est réduit et plus le pourcentage de mots correctement reconnus parmi les mots conservés est élevé (entre 70% et 92%). Par conséquent, un compromis doit être fait entre le nombre de mots qui sont rejetés, et la performance qui peut être obtenue sur les mots conservés.

L'étude se poursuit avec l'analyse des mesures de confiance sur les syllabes, et des travaux parallèles portant sur la recherche de l'affichage le plus pertinent pour optimiser la communication. Les travaux futurs se concentreront sur la détection d'informations complémentaires, comme la détection des questions, afin d'indiquer aux utilisateurs qu'une question leur a été adressée et qu'ils doivent répondre ou demander une clarification (si nécessaire).

Remerciements

Le travail présenté dans cet article fait partie du projet Rapsodie, et a reçu le soutien du Conseil Régional de Lorraine et de la Région Lorraine (FEDER) (http ://erocca.com/rapsodie).

Références

Bigi, B., Meunier, C., Bertrand, R. et Nesterenko, I. (2010). Annotation automatique en syllabes d'un dialogue oral spontané. *In Journées d'Étude de la Parole*, pages 1–4.

Bisani, M. et Ney, H. (2005). Open vocabulary speech recognition with flat hybrid models. *In INTERSPEECH*, pages 725–728.

Cerf-Danon, H., Derouault, A., Elbeze, M. et Mérialdo, B. (1989). Speech recognition in french with a very large dictionary. *In Eurospeech*, pages 2150–2153.

Coursant-Moreau, A. et Destombes, F. (2002). Lipcom, prototype d'aide automatique à la réception de la parole par les personnes sourdes. *Glossa*, 53(68):36–40.

Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M. et Abbott, S. (2002). Tessa, a system to aid communication with deaf people. *In ACM conference on Assistive technologies*, pages 205–212. ACM.

de Calmés, M. et Pérennou, G. (1998). Bdlex : a lexicon for spoken and written french. *In Language Resources and Evaluation (LREC1998)*, pages 1129–1136.

Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F. et Farinas, J. (2010). The epac corpus : Manual and automatic annotations of conversational speech in french broadcast news. *In Language Resources and Evaluation (LREC'10)*.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ester 2 evaluation campaign for rich transcription of french broadcasts. *In INTERSPEECH*.

GANAPATHIRAJU, A., HAMAKER, J., PICONE, J., ORDOWSKI, M. et DODDINGTON, G. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(4):358–366.

GRAVIER, G., ADDA, G., PAULSON, N., CARRÉ, M., GIRAUDEL, A. et GALIBERT, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. *In Language Resources and Evaluation (LREC'12)*.

HUGGINS-DAINES, D., KUMAR, M., CHAN, A., BLACK, A.-W., RAVISHANKAR, M. et RUDNICKY, A. (2006). Pocketsphinx : A free, real-time continuous speech recognition system for hand-held devices. *In ICASSP*.

HÄMÄÄLÄAINEN, A., BOVES, L. et de VETH, J. (2005). Syllable-length acoustic units in large-vocabulary continuous speech recognition. *In SPECOM*, pages 499–502.

ILLINA, I., FOHR, D. et JOUVET, D. (2011). Grapheme-to-phoneme conversion using conditional random fields. *In INTERSPEECH*, pages 2313–2316.

JIANG, H. (2005). Confidence measures for speech recognition : A survey. *Speech Communication*, 45(4):455–470.

LE BLOUCH, O. et COLLEN, P. (2006). Reconnaissance automatique de phonèmes guidée par les syllabes. *In Journées d'Etude de la Parole*.

OROSANU, L. et JOUVET, D. (2013). Comparison of approaches for an efficient phonetic decoding. *In INTERSPEECH*.

PLACEWAY, P., CHEN, S., ESKENAZI, M., JAIN, U., PARIKH, V., RAJ, B., RAVISHANKAR, M., ROSENFELD, R., SEYMORE, K., SIEGLER, M., STERN, R. et THAYER, E. (1996). The 1996 hub-4 sphinx-3 system. *In DARPA Speech Recognition Workshop*.

RASTROW, A., SETHY, A., RAMABHADRAN, B. et JELINEK, F. (2009). Towards using hybrid word and fragment units for vocabulary independent lvcsr systems. *In INTERSPEECH*, pages 1931–1934.

RAZIK, J., MELLA, O., FOHR, D. et HATON, J.-P. (2008). Transcription automatique pour mal-entendants : amélioration à l'aide de mesures de confiance locales. *In Journées d'Étude de la Parole*.

SHAIK, M. A. B., MOUSA, A. E.-D., SCHLÜTER, R. et NEY, H. (2011). Hybrid language models using mixed types of sub-lexical units for open vocabulary german lvcsr. *In INTERSPEECH*.

SOKOL, R. (1996). *Réseaux neuro-flous et reconnaissance de traits phonétiques pour l'aide à la lecture labiale*. Ph.d. thesis in computer sciences, Université de Rennes.

STOLCKE, A. (2002). Srilm an extensible language modeling toolkit. *In Spoken Language Processing*.

TACHBELIE, M., BESACIER, L. et ROSSATO, S. (2011). Comparison of syllable and triphone based speech recognition for amharic. *In Language Technology Conference (LTC)*, pages 207–211.

WOODCOCK, K. (1997). Ergonomics and automatic speech recognition applications for deaf and hard-of-hearing users. *In Technology and Disability*, pages 147–164.

WU, S.-L., KINGSBURY, B., MORGAN, N. et GREENBERG, S. (1998). Incorporating information from syllable-length time scales into automatic speech recognition. *In ICASSP*, pages 721–724.

YAZGAN, A. et SARACLAR, M. (2004). Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. *In ICASSP*, pages I–745–8.

ZHANG, L. et EDMONDSON, W.-H. (2002). Speech recognition using syllable patterns. *In Spoken Language Processing*.

La réorganisation gestuelle interagit avec la phonotactique spécifique à chaque langue

Ioana Chitoran[1] Mark Tiede[2]
(1) Université Paris Diderot, CLILLAC-ARP, Case 7003, 5 rue Thomas Mann, Paris 75250 Cedex 13 France
(2) Haskins Labs, 300 George St., New Haven, CT 06511, USA
ioana.chitoran@univ-paris-diderot.fr, tiede@haskins.yale.edu

RESUME

L'étude de la réorganisation articulatoire obtenue dans la parole accélérée permet d'étudier le planning dans la parole. Les travaux précédents ont montré l'émergence de structures coordinatives stables, où des patrons VC se réorganisent en CV, VCC en CVC, et l'ordre coronal-labial en labial-coronal. Ces modes stables sont interprétés comme résultant de contraintes biomécaniques imposées par le système articulatoire/auditif. Nous adoptons une perspective inter-langues, pour déterminer si des modes stables peuvent aussi provenir de patrons linguistiques appris, propres à chaque langue. On étudie le géorgien et le français, différents dans leur :

- prosodie (proéminence initiale vs. finale en CVCV)
- phonotactique (le géorgien permet des attaques complexes ignorant la sonorité)

Nous analysons des données acoustiques préliminaires des disyllabes [pata], [tapa] (géorgien, français) et [paki] (géorgien) répétées au rythme d'un métronome accéléré. Les résultats suggèrent que la structure linguistique spécifique à chaque langue compte aussi comme médiateur de la réorganisation gestuelle.

ABSTRACT

Gestural reorganization interacts with language-specific phonotactics

Studies of articulatory reorganization occurring under rate-driven production pressure can provide a window into speech planning. Previous work shows evidence for stable coordinative structures in speech in which VC patterns reorganize to CV, VCC to CVC, and coronal-labial to labial-coronal order. Such stable modes are argued to result from general physical-biological constraints imposed by the articulatory/auditory system. Here we adopt a cross-linguistic approach, to examine whether stable modes can also arise from linguistic patterns learned on a language-specific basis. The case studies are Georgian and French. Georgian differs drastically from French in: (i) prosody (initial vs. final prominence in CVCV), and (ii) phonotactics (it licenses complex onsets disregarding sonority). We analyze preliminary acoustic data from Georgian and French native speakers repeating [pata], [tapa], and Georgian speakers repeating [paki] as they track an accelerating metronome. The results suggest that language-specific structure also mediates gestural reorganization, along with the established biomechanical constraints.

MOTS-CLES : phonotactique, réorganisation gestuelle, géorgien
KEYWORDS: phonotactics, gestural reorganization, Georgian

1 Introduction

L'étude de la réorganisation articulatoire obtenue dans la parole accélérée a montré que les gestes articulatoires subissent une réorganisation. Ces études ont montré l'émergence de structures coordinatives stables, où des patrons VC se réorganisent en CV, VCC en CVC (Stetson 1951, Tuller & Kelso 1990, Gleason 1999), et l'ordre coronal-labial en labial-coronal (Rochet-Capellan & Schwartz 2007, Lancia & Fuchs 2011). Ces modes stables sont interprétés comme résultant de contraintes biomécaniques imposées par le système articulatoire/auditif. La parole, comme d'autres aspects du contrôle moteur, se caractérise par une réorganisation menée par la fréquence. Par exemple, un ratio de fréquence de 1:1 est préféré dans les syllabes (C et V alternants, un cycle de mouvement de la mâchoire par syllabe).

L'objectif de notre étude est de déterminer si des modes stables peuvent aussi surgir de patrons linguistiques appris, propres à chaque langue. Selon notre hypothèse, la réorganisation provoquée par un débit accéléré ne reflète pas uniquement les contraintes générales, physiques et biologiques imposées par le système articulatoire et auditif, mais, sous certaines conditions, les contraintes imposées par les routines articulatoires développées et apprises par rapport à chaque langue comptent aussi comme médiateurs de la réorganisation gestuelle. Nous proposons spécifiquement que les routines articulatoires qui jouent un rôle linguistique et qui interviennent dans l'émergence de modes de coordination stables intrinsèques sont la phonotactique et la prosodie.

Nous comparons des données de deux langues – le géorgien et le français – différentes par rapport à leur structure phonotactique et prosodique. Le géorgien permet des attaques complexes ignorant la sonorité et suit à la place un pattern de phasage ou le degré de chevauchement des gestes articulatoires varie par rapport à l'ordre des constrictions : les séquences consonantiques antérieur-postérieur comme /pt/ sont plus chevauchées que les séquences postérieur-antérieur comme /tp/ (Chitoran et al., 2002). Dans cette étude nous analysons des données acoustiques préliminaires provenant de locuteurs natifs géorgiens et français. Les locuteurs ont prononcé les mots [pata], [tapa], [paki] au rythme d'un métronome accéléré. Les résultats préliminaires montrent que des patrons différents de réorganisation sont obtenus dans les deux langues, ce qui correspond à notre hypothèse, suggérant que la structure spécifique à chaque langue interagit avec les contraintes biomécaniques dans la médiation de la réorganisation gestuelle. La comparaison de tels cas d'effets de structure linguistique sur les modes de coordination stable par défaut permet de mieux comprendre l'interaction des facteurs biomécaniques et cognitifs dans le planning de la parole.

Dans la section 1 nous présentons la méthodologie de notre étude, suivie par la discussion des résultats dans la section 2. Nous présentons l'analyse qualitative des patrons de réorganisation observes pour les formes : [pata] vs. [tapa] en géorgien (section 2.1), [pata] vs. [tapa] en français (section 2.2), [paki] en géorgien (section 2.3). La section 3 contient les conclusions.

2 Méthodologie

Quinze locuteurs natifs géorgiens ont été enregistrés à Tbilissi, en Géorgie, dans une salle de classe isolée de l'Université d'Etat de Tbilissi. Six locuteurs natifs français ont été enregistrés à Lyon, dans des conditions similaires. Tous les participants ont produit chaque mot synchronisé avec un métronome auditif. On leur a demandé de ne pas s'arrêter pour respirer, ni pour se corriger. Le métronome a débuté à un rythme constant de 600 ms entre les battements, ensuite s'est accéléré progressivement, atteignant un deuxième plateau de 100 ms entre les battements. La durée de la portion accélérée entre les plateaux du début et de la fin est de 8 secondes. Les enregistrements sont duplex, avec le bruit du métronome sur la plage de gauche et la voix du locuteur sur la plage de droite. Ils ont été effectués avec un enregistreur numérique Roland Edirol R-05 et un microphone externe Audio-Technica ATM63HE.

Six formes CVCV ont été enregistrées, contenant toutes les combinaisons de plosives sourdes [p], [t], [k]. Pour chaque mot, chaque série de répétitions suivant le métronome a été enregistrée trois fois. Nous allons discuter ici des patrons observés pour trois de ces formes : [pata], [tapa], [paki]. Pour les locuteurs géorgiens elles représentent des vrais mots, pendant que pour les locuteurs français ce sont des non-mots. Nous reconnaissons que cette différence peut mener à des patrons de réorganisation différents dans les deux langues (Frank 2011), car les modes de coordination employés dans des vrais mots peuvent être plus résistants à la réorganisation que dans les non-mots.

3 Résultats

Nous proposons que la structure phonotactique et la structure prosodique spécifiques à chaque langue peuvent interagir dans l'émergence de modes de coordination stables intrinsèques dans la réorganisation due à un débit accéléré. Les prédictions faites pour les deux langues sont contraires. Elles sont présentées dans le Tableau 1.

Langue	**CVCV > CCV plus probable**	**CVCV > CCV moins probable**
géorgien	attaques complexes sans contraintes	proéminence initiale
français	proéminence finale	attaques complexes modérées

TABLEAU 1 – Patrons de réorganisation prédits sur la base de la phonotactique et de la prosodie spécifiques à chaque langue

1. La prédiction basée sur les différences phonotactiques est que, suite à l'accélération, la réorganisation gestuelle de CVCV à CCV est plus probable en géorgien, où les attaques complexes sont fréquemment permises. En français, où il y a plus de restrictions sur les attaques complexes, ce type de réorganisation devrait être moins probable.

2. La prédiction basée sur les différences prosodiques est que, suite à l'accélération, les disyllabes à proéminence initiale (géorgien) seront plus résistantes à la réorganisation

de CVCV à CCV que les disyllabes à proéminence finale (français).

Si la réorganisation était déterminée uniquement par des contraintes biomécaniques sur l'articulation on s'attendrait à trouver des patrons de réorganisation similaires pour toutes les formes dans les deux langues. Des études récentes basées sur le paradigme de l'accélération ont montré que la proéminence initiale peut bloquer la transition de deux à un cycle de mâchoire par syllabe (Lancia & Fuchs 2011 pour l'allemand ; Carrissimo-Bertola 2011 pour l'italien). Par le choix de deux langues qui diffèrent aussi dans leur structure phonotactique, l'étude présente peut éclaircir l'importance relative des deux patrons linguistiques dans l'établissement de routines articulatoires.

Vues les différences individuelles de débit de parole, nous allons présenter et comparer ici les données de locuteurs individuels.

3.1 Ordre antérieur-postérieur (pata) vs. postérieur-antérieur (tapa) en géorgien

Le mot géorgien [pʰatʰa] 'stalemate' a été répété trois fois par chaque locuteur, suivant un métronome. Le géorgien possède deux séries de plosives sourdes : aspirées et éjectives. Tous les locuteurs ont systématiquement perdu l'aspiration et la glottalisation très tôt durant chaque répétition.

L'utilisation du débit accéléré sous forme sigmoïde a eu comme résultat deux types de réorganisation : une pendant la phase d'accélération avant d'arriver au plateau final, une autre pendant le plateau final. Nous avons observé d'abord une transition de la proéminence initiale à une proéminence finale (pAta > patA), suivie par une transition à [pta]. La figure 1 montre un exemple représentatif des deux types de réorganisation provenant du même locuteur géorgien (pAta > patA > ptA).

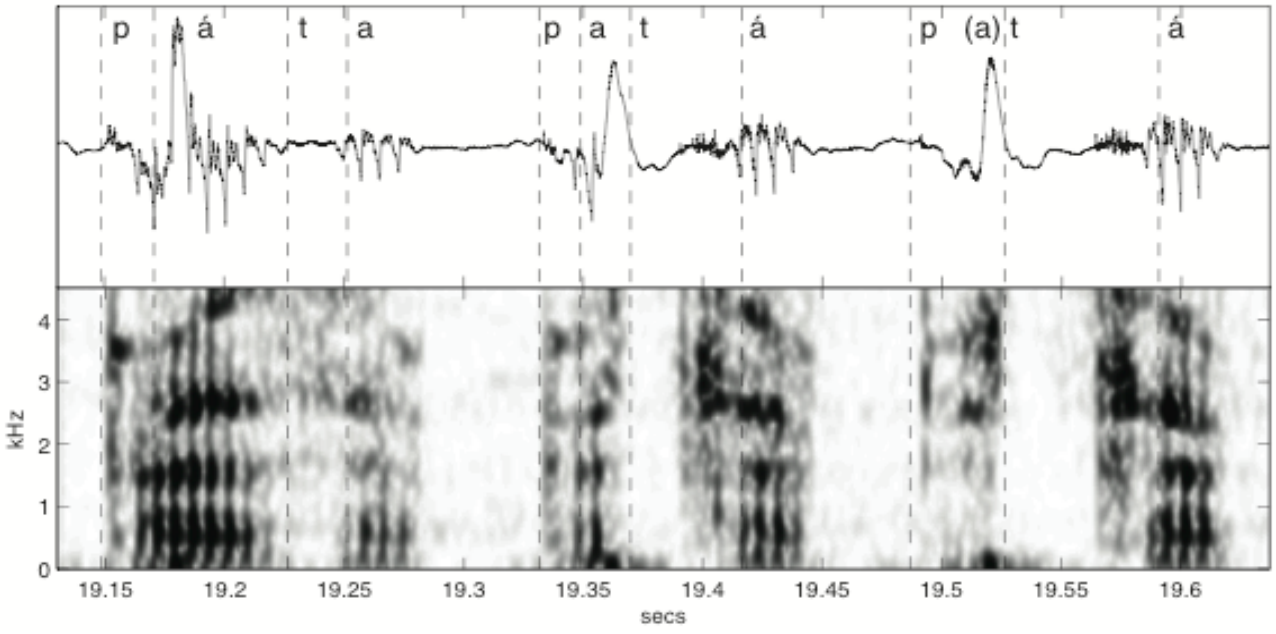


FIGURE 1 – Géorgien. Exemple de transition de proéminence initiale à proéminence finale dans [pʰatʰa] (la durée de V2 augmente par rapport à V1), suivie par élision de V1 (*pata → pta*). Typiquement les locuteurs persistent dans le patron CCV suivant la

réorganisation.

Les données de neuf des quinze locuteurs géorgiens ont été analysées. Tous les locuteurs ont fait la transition de proéminence initiale à finale *(pAta → patA),* et sept d'entre eux ont fait la transition à *pta*. L'élision de V1 ne s'oppose pas aux regles phonotactiques du géorgien.

On remarque relativement moins de réorganisation pour *tʰapʰa* 'small hill', conformément à notre prédiction. Il n'y a pas de réorganisation à CCV. Tous les neuf locuteurs réduisent la voyelle finale *(tapa → tap)*. La transition d'accent initial à accent final *(tApa → tapA)* se produit pour un seul locuteur, dans deux des trois répétitions, mais n'est pas suivie par l'élision de V1. L'absence d'élision de V1 dans *tapa* n'est pas prédite par les contraintes biomécaniques, mais elle est en accord avec l'asymétrie de l'ordre de constriction du géorgien, où les séquences d'ordre postérieur-antérieur *(t-p)* sont moins étroitement coordonnées que les séquences antérieur-postérieur *(p-t)*.

3.2 Ordre antérieur-postérieur (pata) vs. postérieur-antérieur (tapa) en français

Nous avons examiné les données de deux des six locuteurs français. Pour un des locuteurs le seul changement observé a été la transition de la proéminence finale à une proéminence initiale dans les trois répétitions des deux formes : *patA → pAta* et *tapA → tApa*. Pour le second locuteur la transition de proéminence a été observée seulement pour [tapa], avec réduction de la voyelle finale non-accentuée : *tapA → tApa → tAp*. Aucun changement n'a été observé pour [pata]. Nous remarquons que l'absence de vrais mots n'a pas encouragé la réorganisation par rapport au cas du géorgien, ou des vrais mots ont été utilisés.

L'absence d'élision de V1 est en accord avec la phonotactique du français, mais Rochet-Capellan & Schwartz (2007) ont observé cette élision pour les deux formes : *pata → pta* et *tapa → pata → pta*. Nos résultats préliminaires sont donc inattendus, mais nous devons analyser aussi les autres locuteurs.

3.3 [paki] en géorgien

Une forme CVCV avec des voyelles différentes – [p'ak'i] – teste la prédiction que la réorganisation de deux cycles de mâchoire par énoncé (fréquence 2:1) à un seul (fréquence 1:1) peut se réaliser par coalescence des voyelles. Nous avons observé chez deux des locuteurs géorgiens la coalescence des deux voyelles dans un patron CVVC *(paik)* durant la phase d'accélération (Figure 2). A mesure que le débit augmente, on observe que le phasage de /k/ change, découvrant la transition /a:i/. Il n'y a pas d'indice de réduction vocalique, mais la constriction du dos de la langue est appliquée plus tard, ce qui indique une réorganisation.

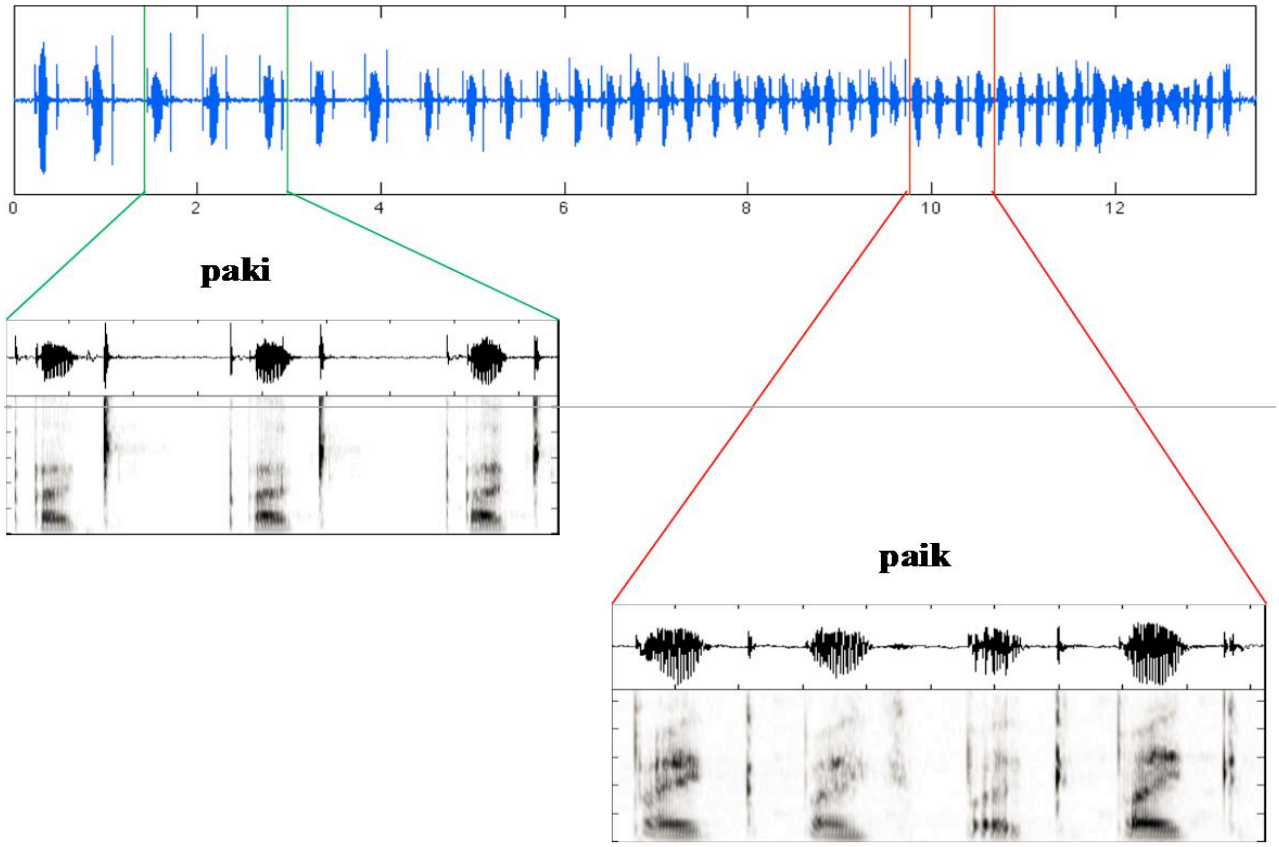


FIGURE 2 – *paki* → *paik* en géorgien. Re-phasage de /k/ à débit accéléré.

On pourrait attribuer le changement *paki* → *paik* simplement à une coarticulation V-V renforcée et à la réduction de la voyelle finale non-accentuée, les deux étant attendues à un débit de parole accéléré. Si c'était le cas, la portion vocalique coarticulée située entre les deux plosives serait aussi de plus en plus courte. Mais si le changement est dû à la réorganisation gestuelle plutôt qu'à la réduction, la portion vocalique coarticulée ne devrait pas diminuer avec l'accélération. Au contraire, dans la transition de [paki] à [paik], la portion vocalique entre les deux plosives devrait représenter un pourcentage plus élevé de la durée de l'intervalle entre les deux relâchements (inter-burst interval – IBI), entre le début du relâchement de [p] et le début du relâchement de [k]. Si le geste de [k] avance vers la droite, il découvrira une plus grande partie de la voyelle [i]. L'hypothèse de la réorganisation gestuelle prédit donc que la durée de la partie vocalique va progressivement augmenter avec le débit accéléré et que la proportion représentée par la partie vocalique par rapport à l'intervalle [p]-[k] va augmenter aussi. La figure 3 montre les données obtenues pour les trois répétitions de [paki] pour le même locuteur. Les panneaux supérieurs montrent la proportion entre la durée de la voyelle et l'intervalle IBI. On observe que la proportion augmente avec le débit accéléré. Les panneaux inférieurs montrent la durée de l'intervalle IBI en bleu avec la durée de la partie vocalique superposée en rouge. On observe que l'intervalle IBI se réduit à mesure que le débit accélère.

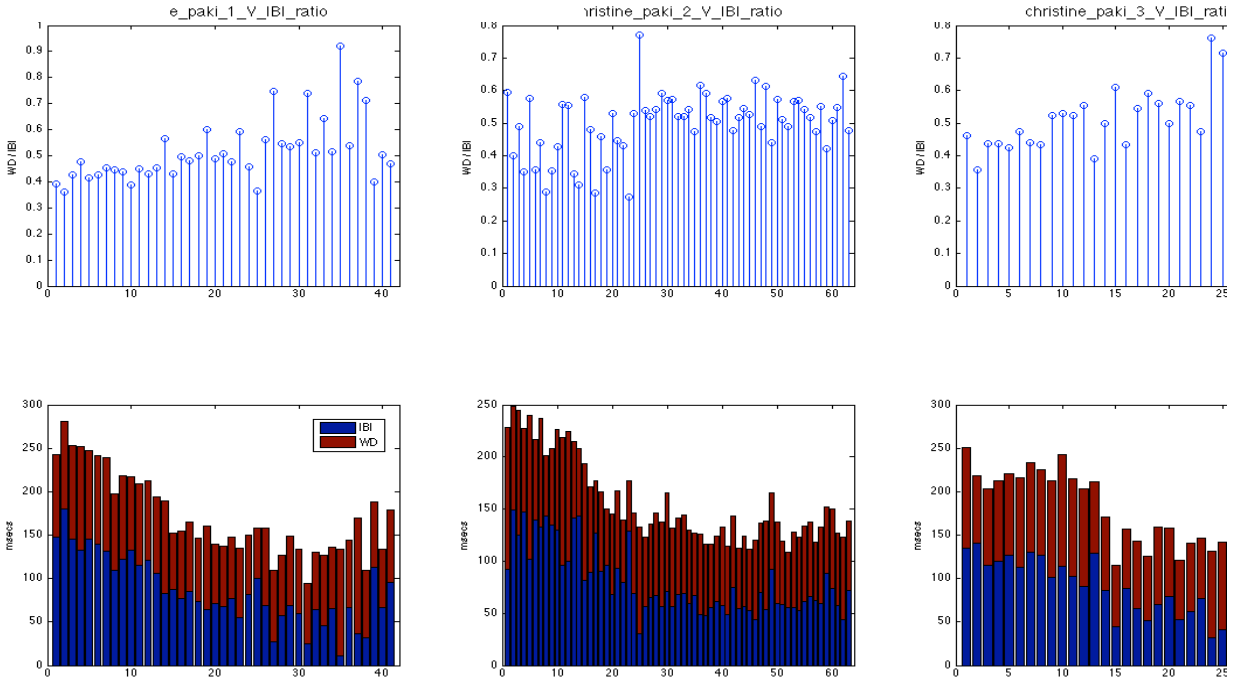


FIGURE 3 – *paki* → *paik* en géorgien. 3 répétitions. En haut: proportion partie vocalique : intervalle IBI. En bas: durée de la partie vocalique en rouge (ms) superposée sur la durée de l'intervalle IBI en bleu (ms).

Ce résultat confirme que le paradigme de l'accélération entraine une réorganisation des gestes impliqués plutôt que leur réduction.

4 Conclusions

Les données que nous avons présentées soutiennent l'hypothèse que la réorganisation entrainée par la manipulation d'un paramètre de contrôle (le débit de parole) n'est pas affectée uniquement par des contraintes biomécaniques. Elle montre aussi des effets de timing spécifiques à une langue. En géorgien, la répétition accélérée transforme la forme [p^{h}at^{h}a] en une attaque complexe antérieur-postérieur [pta], ce qui ne se produit pas dans la répétition accélérée de [t^{h}ap^{h}a]. Ce résultat correspond à l'asymétrie de l'ordre de constriction observe en géorgien (Chitoran et al. 2002), qui prédit une réorganisation gestuelle différente par rapport à l'ordre des constrictions. Il est moins en accord avec l'effet labial-coronal (Locke 1983, MacNeilage & Davis 2000, Rochet-Capellan & Schwartz 2007) qui prédit la réorganisation dans [pta] pour les deux formes. En français, l'absence de réorganisation CVCV à CCV correspond à la phonotactique de la langue.

La réorganisation de [paki] à [paik] est aussi en accord avec la phonotactique du géorgien. Même si [pk] est une attaque permise en géorgien, il est possible que la fréquence 1:1 obtenue par la coalescence des voyelles (CVVC) est préférable a CCV

lorsque la combinaison vocalique résultante est attestée dans la langue. Cette question fera le sujet de prochaines études et d'une comparaison avec [paki] en français.

L'étude de la réorganisation articulatoire est importante, car elle permet d'observer le planning de la parole en observant l'optimisation de la synergie gestuelle sous l'accélération et la réorganisation des proportions de fréquences intra-syllabiques. L'optimisation consiste dans la simplification du cycle du mouvement de la mâchoire. Les expériences préliminaires présentées ici confirment que l'étude de la réorganisation dans une perspective inter- langues est essentielle si on veut comprendre l'apport des forces biomécaniques et cognitives qui gèrent la parole humaine.

Remerciements

Ce travail a été en partie financé par une bourse Fulbright-Hays.

Références

CARRISSIMO-BERTOLA, M. (2011). *Tendances et universaux des structures syllabiques : l'effet Labial-Coronal en italien*. (Mémoire Master2). Université Stendhal, Grenoble.

CHITORAN, I., L. GOLDSTEIN, D. BYRD (2002) Gestural overlap and recoverability: Articulatory evidence from Georgian. In Gussenhoven, C., T. Rietveld, N. Warner (Eds.) *Laboratory Phonology 7*. Berlin: Mouton de Gruyter, 419-447.

FRANK, A. (2011) *Integrating linguistic, motor, and perceptual information in language production*. PhD dissertation, University of Rochester.

GLEASON, P.L. (1999) *Syllable structure and stress effects on articulatory coordination*. PhD dissertation, Florida Atlantic University.

LANCIA, L. AND S. FUCHS. (2011) The labial coronal effect revised. In Ostry, D., S.R. Baum, L. Menard, V. Gracco (Eds.) *Proceedings of ISSP 9*, Montreal, Canada, June 20-23, 2011, 187-194.

LOCKE, J.L. (1983) *Phonological acquisition and change*. New York: Academic Press.

MACNEILAGE, P. AND B. DAVIS. (2000). On the origins of internal structure of word forms. *Science 288*, 527-531.

ROCHET-CAPELLAN, A. AND J-L. SCHWARTZ (2007). An articulatory basis for the labial-coronal effect: /pata/ seems a more stable articulatory pattern than /tapa/. *JASA 121* (6), 3740-3754.

STETSON, R.H. (1951). Motor phonetics: a study of speech movements in action. Amsterdam: North Holland.

TULLER, B. AND J.A.S. KELSO. (1990). Phase transitions in speech production and their perceptual consequences. In Jeannerod, M. (Ed.) Attention and Performance XIII. Motor Representation and Control. Hillsdale, NJ, Hove, and London: Lawrence Erlbaum Associates, 429-452.

Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques

Benjamin Bigot[†‡], Grégory Senay[†∓], Georges Linarès[†], Corinne Fredouille[†], Richard Dufour[†] *

[†]Laboratoire Informatique d'Avignon (LIA), Université d'Avignon, France
[‡]School of Computer Engineering, Nanyang Technological University, Singapour
[∓]Panasonic R&D Company of America, Cupertino, États-Unis
prénom.nom@univ-avignon.fr

RÉSUMÉ

La détection de noms de personnes dans des flux audios est un problème critique en reconnaissance et compréhension de la parole, en particulier parce que ces noms portent une information essentielle sur le contenu sémantique du document. La difficulté de la reconnaissance vient de leur nombre, de leur diversité et du caractère imprévisible de leur contexte d'apparition : ils sont souvent liés à un fait d'actualité ponctuel. Ces caractéristiques rendent difficile la conception de lexiques de reconnaissance suffisamment couvrants. Nous traitons ce problème en proposant une méthode de modélisation des contextes lexicaux liés aux noms de personnes ; ces modèles de contexte doivent permettre d'évaluer la probabilité d'un nom par l'analyse du contexte lexical de son occurrence. Nos premières expériences sont réalisées dans le cadre du défi ANR REPERE. Il valident à la fois l'intérêt de l'approche et la pertinence des modèles proposés.

ABSTRACT

Continuous modeling of contexts for person name spotting)

The detection of proper names utterances in speech streams is an important cue for speech recognition. We propose a method aiming at modeling the context of proper names. This method allows us to detect names even is there were not recognized by the automatic speech recognition system. A new context modeling paradigm is proposed in order to capture lexical and structural information surrounding a spoken name. The fundamental hypothesis of this study has been validated on broadcast TV documents available in the context of the REPERE challenge.

MOTS-CLÉS : Reconnaissance de la parole, indexation audio, Détection de mots clefs..

KEYWORDS: Speech recognition, word spotting, audio indexing.

1 Introduction

La détection de noms de personnes dans les flux audiovisuels est un problème critique pour l'analyse des contenus parlés, ces noms portant une information essentielle à l'intelligibilité du document. Dans un contexte d'indexation audio ou de recherche d'information dans des

∗. Ce travail a été réalisé dans le cadre du projet ContNomina financé par l'Agence Nationale de la Recherche (ANR) sous le contrat ANR-12-BS02-0009.

documents parlés, les noms propres sont souvent directement liés à l'attente de l'utilisateur, soit qu'il s'agisse d'éléments de la requête soit qu'ils y soient indirectement liés.

La communauté du traitement automatique du langage naturel a consacré de nombreuses études à la détection des entités nommées, souvent sur des données relativement bien structurées. La détection des noms propres dans des flux de parole pose des problèmes particuliers liés d'une part à la nature des données et, d'autre part, aux composants impliqués dans les systèmes de détection. L'approche standard consiste à s'appuyer sur un système de reconnaissance automatique de la parole (SRAP) dont le lexique intègre l'ensemble des noms qui doivent être détectés. Dans un contexte réel, cette technique rencontre une difficulté majeure liée au nombre, à la diversité des noms propres et au fait qu'ils sont souvent attachés à des événements ou des lieux particuliers, ce qui limite la fréquence d'un nom et augmente la difficulté de sa modélisation statistique. De fait, la construction de lexiques exhaustifs de noms propres est impossible.

La plupart des méthodes proposées en reconnaissance de la parole se sont concentrées soit sur des mises à jour automatiques (ou semi-automatiques) des lexiques, soit sur des méthodes de recherche acoustique dans les graphes de reconnaissance, méthode qui autorise la recherche dans des bases de données préalablement décodées avec un lexique incomplet. Cette dernière voie repose sur l'idée que, bien qu'on ne dispose pas de données *a priori* permettant l'intégration de tous les noms possibles dans les dictionnaires et dans les modèles de langage, la recherche au niveau phonétique reste possible et souvent suffisante.

Notre travail se place dans ce contexte de recherche de noms dans des grandes bases transcrites automatiquement par un SRAP dont le lexique ne couvre pas l'ensemble des noms de personnes contenues dans le flux audio. L'idée initiale est que les noms manquants sont généralement rares et souvent attachés à des contextes particuliers (actualité, thème peu fréquent, etc.). Cette caractéristique implique une dépendance forte entre un nom et le contexte global dans lequel il est prononcé. C'est cette dépendance que nous voulons modéliser puis exploiter pour permettre la détection de noms qui n'auraient pas été reconnus par un SRAP.

La plupart des méthodes proposées dans le passé pour la représentation des contextes reposent soit sur des statistiques à court terme, typiquement des modèle n-grammes, soit sur une approche sac-de-mots dans laquelle la structure temporelle des contextes est perdue. Ici, nous présentons une méthode de représentation des contextes qui intègre à la fois cette structure temporelle et qui permet de capturer, dans un même formalisme relativement simple, les dépendances à court et à long termes. Cette méthode repose sur la projection des contextes lexicaux dans un espace continu de faible dimension.

La suite de l'article présente un survol assez général des approches proposées dans la littérature pour la détection de personnes (i.e. section 2). La section 3 décrit l'approche originale que nous proposons pour une modélisation continue des contexte lexicaux. La section 4 présente les expériences menées et analyse les premiers résultats obtenus. La dernière section conclut et présente quelques perspectives ouvertes par cette étude.

2 Modéliser les contextes : état de l'art

La plupart des systèmes de détection des entités nommées reposent sur l'extraction d'informations de natures assez diverses, notamment relatives aux contextes lexicaux, structurels, syntaxique,

sémantiques... d'occurrence d'une entité. En particulier, les modèles n-grammes ont été largement utilisés pour représenter la séquentialité de *N* mots précédant un nom donné. Dans (Béchet *et al.*, 2000), les entités nommées sont extraites automatiquement pour apprendre un classifieur statistique. Dans ce type d'approches, la séquentialité des mots est capturée, mais seulement sur des fenêtres de tailles assez réduites ; pourtant, les dépendances lexicales à plus long terme portent une information relative au sujet ou au contexte global dans lequel le nom apparaît... Une alternative couramment utilisée consiste à considérer les fréquences des mots du contexte sans prendre en compte leur séquentialité ; cette approche à base de représentation vectorielle d'un sac-de-mots est un stanadrd en traitement automatique des langues (Pedersen *et al.*, 2005) et en recherche d'information (Pedersen *et al.*, 2006; Kozareva et Ravi, 2011). Elle a l'avantage d'une certaine robustesse et d'une facilité d'estimation, mais la structure temporelle du contexte (ou du document) est perdue. Par ailleurs, cette approche peut permettre l'estimation de modèles distributionnels des contextes (Miller et Charles, 1991; Harris, 1968).

3 Contexte temporel et lexical

L'approche proposée dans ce papier repose sur l'hypothèse que le nom d'une personne est lié au contexte lexical dans lequel ce nom apparaît. Afin de modéliser ces dépendances *nom vers contexte*, nous proposons de considérer un nom propre comme une variable latente survenant dans divers contextes observables. Au regard de cette hypothèse, un nom de personne doit être découvert dans les transcriptions automatiques en tenant compte de son contexte, même si le nom n'a pas été proposé par le système de transcription. Ceci est particulièrement vrai pour le cas des noms propres, puisque nombre d'entre-eux ne sont pas dans le vocabulaire du système de transcription, et ne pourront donc jamais être reconnus.

Notre approche consiste à capturer les dépendances *nom vers contexte* à l'aide de modèles statistiques. Les techniques classiques reposent sur la modélisation du contexte proche, ou sur les co-occurrences éloignées, représentées dans un modèle vectoriel (SVM). En premier lieu, les modèles de contexte proche sont généralement basés sur des statistiques n-grammes, où *n* est inférieur à 6 pour des raisons de complexité. Ces modèles manquent alors les dépendances longues portées alors qu'elles véhiculent l'information sémantique et pragmatique. D'autre part, les modèles de co-occurrences représentent un document comme un sac-de-mots, sans aucune information relative à sa structure temporelle. Par conséquent, l'estimation des statistiques sur ces co-occurrences est possible sur des corpus relativement limités, mais l'information critique, telle que la séquentialité des mots notamment, est définitivement perdue.

Ici, l'accent est mis sur l'extraction de l'information discriminante détenue par les mots autour de noms propres. En effet, nous considérons la temporalité du contexte des mots comme un élément important, tout comme dans les n-grammes et les méthodes à base de règles. De plus, nous supposons que les dépendances proches et éloignées des mots pourraient apporter des informations pertinentes pour le nom propre latent. Ceci est particulièrement vrai pour les émissions télévisées, où de nombreux noms hors-vocabulaire apparaissent exclusivement dans un contexte spécifique.

Pour ces raisons, nous proposons une représentation vectorielle qui détient des informations relatives aux positions des mots du contexte, couplée avec une fenêtre de contexte plus large comparable à celles utilisées dans les approches fondées sur la représentation par sac-de-mots.

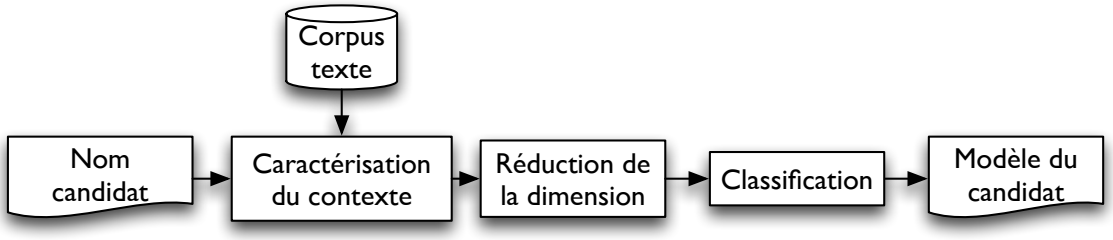


FIGURE 1 – Cadre général pour l'extraction et la modélisation des contextes lexicaux de noms de personnes

En effet, cette fenêtre permet de prendre en compte le cas où l'information discriminante se situe dans des phrases éloignées de plusieurs dizaines de mots du nom de personne. Cette approche originale, illustrée sur la figure 1, se déroule en trois étapes, décrites ci-dessous, pour la modélisation des contextes d'un nom de personne à partir d'exemples trouvés dans une base de textes.

3.1 Caractérisation du contexte

Considérons un nom de personne W, et supposons qu'il dispose d'un corpus de documents textuels sans erreur. La première étape consiste à estimer, à partir de la collection de contextes du nom, le modèle qui sera capable de retrouver le nom de personne manquant W dans des transcriptions automatiques.

De manière formelle, un contexte est une séquence de $2N + 1$ mots centrée sur un nom de personne. Le système extrait tout d'abord, à partir des documents de la base de données, tous les contextes C_W associés à W. Cet ensemble de contextes est noté S_{C_W}. Les documents du corpus ont été pré-traités en enlevant la ponctuation et en mettant l'ensemble du texte en minuscule, tout comme les sorties du système de transcription.

Dans un second temps, un étiqueteur morpho-syntaxique (Schmid, 1994) est appliqué sur les contextes. Les mots étiquetés comme *Nom*, *Verbe* et *Adjectif* sont conservés, mais remplacés par leur forme lémmatisée. Les mots ne possédant pas ces étiquettes sont substitués par l'étiquette *Null*. Enfin, ce processus supprime les mots fonctions, et réduit le lexique de S_{C_W} en considérant les formes canoniques des mots. À la fin de cette seconde étape, un lexique global est construit avec l'ensemble des mots lemmatisés apparaissant au moins une fois dans les contextes de S_{C_W}. Le lexique L_W contient l'information lexical de notre représentation contextuelle.

Par définition, un contexte est une séquence de mots $w^{(i)}$ centrée autour W, où i est la position relative de w_i par rapport à W :

$$w^{-N} \quad \ldots \quad w^{(-1)} \quad W \quad w^{(+1)} \quad \ldots \quad w^{(+N)}$$

Les mots sont pondérés en fonction de leurs positions relatives dans le contexte. Les poids permettent de prendre en compte le fait que les mots proches de W caractérisent mieux le nom de personne que les mots plus éloignées dans le contexte. Pour un mot $w^{(i)}$ situé à la position i,

le poids $P_{w^{(i)}}$ est calculé selon :

$$P_{w^{(i)}} = \frac{1+\log 10}{1+\log|i|}, i \neq 0$$

La fonction n'est pas définie en 0 puisque le mot latent W est supposé ne pas être observé.

En combinant le lexique L_W et les mots pondérés de chaque contexte C_W de S_{C_W}, une matrice dénotée M_W est construite, dans laquelle une ligne représente un contexte, une colonne représente un mot de L_W, et les cellules sont remplies selon la position pondérée P_{w^i} du mot w^i. Lorsque qu'un mot apparaît plusieurs fois dans un contexte, le mot ne sera considéré qu'une seule fois avec son occurrence la plus proche de W. Si un mot de L_W n'apparaît pas dans un contexte, sa cellule correspondante dans M_W est mise à la valeur 0.

3.2 Réduction de la dimension

L'étape précédente a produit une matrice M_W avec les dimensions $C \times L$, où C est le nombre de contextes extraits à partir d'un ensemble de documents d'entrainement et L est la taille du lexique global. Puisque la matrice contient une proportion importante de valeurs égales à 0, une stratégie de réduction de la dimension est nécessaire. Ici, une décomposition en valeurs singulières (ou SVD, *Singular Value Decomposition*) est appliquée. Elle permet de réduire la rareté des données, tout en réduisant la dimension du lexique.

La réduction de la dimension est réalisée en séparant la matrice M_W en trois différentes matrices :

$$M_W = U \times \Sigma \times V^T$$

où Σ représente une matrice diagonale, et U et V des matrices orthogonales. Les vecteurs composant Σ sont ré-ordonnés en fonction des valeurs singulières les plus hautes. La réduction de la dimension est, ensuite, achevée en supprimant les dimensions représentant le moins de variation. La matrice de transformation est obtenue grâce au produit de M_W par V. La SVD a été réalisée au moyen de l'outil *SVDLIBC*.

3.3 Détection des noms

L'espace réduit obtenu à partir de la SVD permet de projeter chaque séquence de mots dans un vecteur de faible dimension. Le processus de détection sera exploité à ce niveau, en cherchant une correspondance dans le flux de vecteurs de caractéristiques obtenu à partir des mots issus des transcriptions automatiques.

Le modèle contextuel de noms est entrainé pour détecter qu'un contexte lexical inconnu correspond ou non à un nom dans le flux de vecteurs. Nous avons choisi d'utiliser un classifieur linéaire "Support Vector Machine" (SVM) avec deux classes. Notre stratégie de classification s'appuie sur la concurrence entre deux matrices de contextes contradictoires.

La première matrice de contextes M_{C_W}, contenant les contextes centrés sur le nom de personne W, est appelée *matrice d'acceptation* de W. La seconde est appelée *matrice de rejet* de W. Elle a été construite au moyen des contextes où le nom W n'est pas présent. La classification SVM a été réalisée avec l'outil *libSVM* (Chang et Lin, 2011).

4 Expériences

4.1 Contexte expérimental

Les expériences menées dans ce papier ont été réalisées avec les données issues de la campagne d'évaluation REPERE (Kahn *et al.*, 2012). Les contextes utilisés pour construire les matrices d'acceptation et de rejet sont issus de trois différentes types de documents. Le premier ensemble est composé d'environ 135 000 dépêches AFP (Agence France Presse) de 2009 à 2011. Le deuxième ensemble provient de Wikipedia (2012). Il contient environ 1,2 millions d'articles. Le dernier ensemble contient les transcriptions manuelles de 280 heures d'émissions radios et télévisées françaises issues de précédentes campagnes d'évaluation françaises (ESTER1&2, EPAC, ETAPE et REPERE).

Dans ces expériences, les données de test sont issues d'un corpus de développement fourni par les organisateurs de REPERE. Il contient 135 fichiers audio pour une total de 24 heures de parole annotées incluant 4 172 annotations de noms de personnes prononcées. Nous avons réduit cet ensemble à 4 130 occurrences de noms, les personnes dont l'identité n'ayant pu être clairement définie étant enlevées. Au final, cet ensemble de noms correspond à 845 identités différentes parmi lesquelles nous trouvons principalement les journalistes des émissions, des politiciens français et étrangers, des sportifs, des personnes connues. . .Dans l'ensemble de test, 40 % des personnes identifiées correspondent à 85 % des noms de personnes annotées. Les noms des derniers 60 % ne sont pas prononcés plus de deux fois.

Les transcriptions automatiques ont été produites au moyen du système de transcription du LIA (Nocera *et al.*, 2002). Le taux d'erreur-mot sur l'ensemble de test est de 29,4 %. Nous notons que 66 % des identités contenues dans notre corpus de test se trouvent dans le dictionnaire du système de transcription.

Dans nos expériences, la longueur des fenêtres d'estimation utilisée pour l'extraction des contextes a été fixée à 201 mots (*i.e.* une fenêtre centrée de $2N + 1$ mots sur un nom de personne avec $N = 100$). À la fin du processus d'extraction du contexte, nous avons observé qu'aucun contexte n'a été trouvé pour 83 identités. De plus, si nous considérons les 845 noms de personnes présents dans les annotations manuelles, seulement 323 noms (38,2 % des noms de personnes) possèdent plus de 100 contextes. Cet ensemble de 323 noms couvre 63,3 % des occurrences des noms de personnes prononcés dans l'ensemble de test (2 615 occurrences sur un total de 4 130). Cette observation nous a conduit à nous limiter à ces 323 noms pour nos expériences. Pour chacun d'entre eux, les contextes sont alors utilisés pour l'apprentissage du modèle. Une fois construit, chaque matrice de contextes est réduite à 100 dimensions au moyen de la SVD. À l'étape de classification, afin d'entraîner un classifieur SVM plus discriminant, pour un nom W donné, nous avons construit la matrice de rejet correspondante au moyen d'un échantillonnage aléatoire des contextes des noms de personnes différentes de W.

5 Résultats

Nous proposons de réaliser la reconnaissance et la localisation des noms de personnes dans les transcriptions automatiques au niveau des segments de parole. Notre système de base consiste à

regarder directement les noms de personnes dans la meilleure hypothèse en sortie du système de transcription au moyen d'expressions régulières. Puisque que notre approche suppose qu'un nom de personne peut être trouvé dans les résultats des systèmes de transcription même si le nom prononcé n'est pas présent dans les transcriptions, nous cherchons à évaluer si notre représentation temporelle et lexicale du contexte a capturé assez d'information contextuelle pour apprendre un modèle discriminant. D'un autre côté, nous voulons également évaluer la complémentarité du système de base avec notre proposition.

L'évaluation est réalisée en considérant seulement 323 noms de personnes des références manuelles pour lesquelles nous avons été capable d'apprendre un modèle du contexte. L'ensemble de test contient alors 2 615 occurrences de noms. L'évaluation indique que 21 % (549 occurrences) des noms de personnes ne peuvent pas être trouvés, que ce soit dans les transcriptions automatiques ou le modèle des contextes. Nous avons également évalué que 40 % des noms de personnes peuvent être trouvés directement dans les transcriptions automatiques ou en utilisant le modèle des contextes. Notre méthode s'appuyant sur les contextes est incapable de trouver 14 % des occurrences, alors que les noms prononcés ont pu être retrouvés avec des expressions régulières. Cependant, un résultat très intéressant concerne le fait que **25 % des noms prononcés peuvent être correctement extraits** au moyen de notre approche, alors que ces résultats ne peuvent être trouvés dans les transcriptions automatiques. Ce dernier résultat confirme la validité de notre hypothèse de départ. Enfin, en considérant la combinaison de ces deux approches (base + contexte), nous observons que 79 % des noms de personnes considérés peut être correctement retrouvés.

6 Conclusion et perspectives

Dans ce papier, nous avons présenté une méthode pour la modélisation du contexte afin de retrouver des noms de personnes dans la meilleure hypothèse d'un système de transcription dans le cadre d'émissions télévisées. Notre approche suppose que l'occurrence d'un nom de personne est dépendant du contexte lexical dans lequel il apparaît, et la modélisation des dépendances du contexte d'un nom peut aider la recherche de noms de personnes dans les sorties de systèmes automatiques de transcription. Partant de cette idée, nous avons proposé un paradigme permettant la modélisation de contextes au moyen de vecteurs de caractéristiques intégrant la structure temporelle de contextes lexicaux très éloignés. Ces vecteurs sont réduits avec un processus SVD classique, et les modèles SVM sont entrainés afin d'identifier le nom de personne recherchés dans un espace vectoriel de faible dimension.

Les expériences ont été menées sur les données issues de la campagne d'évaluation REPERE. Les résultats valident notre hypothèse de départ : notre méthode améliore de 54 % à 79 % le taux de détection des noms de personnes.

Des travaux futurs s'attarderont à augmenter le nombre de noms de personnes considérés, puisque nos expériences ne concernent pour l'instant que 323 noms de personnes. Nous étudierons également l'utilisation des mesures de confiance fournies par les systèmes de transcription afin d'augmenter la précision de notre système. De plus, une fois que les noms de personnes ont été détectés, nous pouvons essayer d'étudier la façon dont nous pouvons répondre aux questions *Qui parle ?* et *De qui l'on parle ?* en appliquant ces méthodes telles que (Tranter, 2006; Canseco *et al.*, 2005).

Références

BÉCHET, F., NASR, A. et GENET, F. (2000). Tagging unknown proper names using decision trees. *In Proc. of the 38th Annual Meeting on ACL*, pages 77–84.

CANSECO, L., LAMEL, L. et GAUVAIN, J.-L. (2005). A comparative study using manual and automatic transcriptions for diarization. *In Proc. of IEEE ASRU*, pages 415–419.

CHANG, C. et LIN, C. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27 :1–27 :27.

HARRIS, Z. (1968). *Mathematical Structures of Language*. Wiley.

KAHN, J., GALIBERT, O., QUINTARD, L., CARRE, M., GIRAUDEL, A. et JOLY, P. (2012). A presentation of the repere challenge. *In CBMI*, pages 1–6.

KOZAREVA, Z. et RAVI, S. (2011). Unsupervised name ambiguity resolution using a generative model. *In Proc. of the 1rst Workshop on Unsupervised Learning in NLP*, pages 105–112. ACL.

MILLER, G. et CHARLES, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

NOCERA, P., LINARÈS, G., MASSONIÉ, D. et LEFORT, L. (2002). Phoneme lattice based a* search algorithm for speech recognition. *In Proc. of the 5th Int. Conf. on Text, Speech and Dialogue*, pages 301–308.

PEDERSEN, T., KULKARNI, A., ANGHELUTA, R., KOZAREVA, Z. et SOLORIO, T. (2006). An unsupervised language independent method of name discrimination using second order co-occurrence features. *Computational Linguistics and Intelligent Text Processing*, pages 208–222.

PEDERSEN, T., PURANDARE, A. et KULKARNI, A. (2005). Name discrimination by clustering similar contexts. *Computational Linguistics and Intelligent Text Processing*, pages 226–237.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proc. of Int. Conf. on New Methods in Language Processing*.

TRANTER, S. (2006). Who really spoke when ? finding speaker turns and identities in broadcast news audio. *In in Proc. of ICASSP*. IEEE.

Catégorisation multi-thématique de dialogues téléphoniques

Xavier Bost[1] Marc El-Bèze[1] Renato De Mori[1,2]
(1) LIA, Université d'Avignon, 339 chemin des Meinajariès, Avignon, France
(2) McGill University, School of Computer Science, Montréal, Québec, Canada
{xavier.bost, marc.elbeze, renato.demori}@univ-avignon.fr

RÉSUMÉ

Ce papier porte sur l'analyse automatique de dialogues téléphoniques entre un client et un agent d'un centre d'appel d'un service clientèle. Le but de l'analyse est d'identifier, parmi un ensemble thématique prédéfini, les thèmes des problèmes évoqués dans le dialogue. Un dialogue peut contenir des thèmes multiples mentionnés dans des segments entremêlés difficiles à délimiter. Deux méthodes sont proposées pour conjecturer les thèmes multiples. La première repose sur une mesure de similarité *cosine* appliquée au sac de mots extrait de la totalité du dialogue. La seconde méthode introduit le concept de densité thématique distribuée autour de positions spécifiques du dialogue. En plus des unigrammes, sont également pris en compte les bigrammes, avec d'éventuels trous entre les deux termes. Les résultats expérimentaux obtenus par les méthodes proposées apparaissent supérieurs à ceux obtenus par des machines à support de vecteurs appliquées aux mêmes données.

ABSTRACT

Multi-theme categorization of telephone dialogues

This paper deals with the automatic analysis of dialogues between a customer and an agent in a call centre of a customer care service. The purpose of the analysis is to identify, among a predefined thematic set, themes about problems discussed in the dialogue. A dialogue may contain multiple themes mentioned in interleaved segments that cannot be well defined. Two methods are proposed for multiple theme hypothesization. One of them is based on a cosine similarity measure using a bag of words extracted from the entire dialogue. The other method introduces the concept of thematic density distributed around specific word positions in a dialogue. In addition to unigrams, word bigrams with possible gaps between the two words are also considered. The experimental results obtained with the proposed methods turn out to outperform the results obtained with support vector machines on the same data.

MOTS-CLÉS : classification multi-thématique de documents audio, analyse de dialogues humain/humain, *speech analytics*, bigrammes à distance.

KEYWORDS: multi-topic audio document classification, human/human dialogue analysis, speech analytics, distance bigrams.

1 Introduction

Ces dernières années, l'intérêt porté aux possibilités offertes par les technologies de la parole pour le suivi des services téléphoniques est allé croissant. Dans un souci d'efficacité et de satisfaction du client, il existe actuellement un consensus sur l'importance de l'amélioration des systèmes d'analyse de dialogues entre humains afin d'obtenir des rapports sur les problèmes rencontrés par les clients et la façon dont les conseillers ont pu les résoudre. A partir de tels rapports, diverses statistiques peuvent être établies sur la typologie des problèmes, l'efficacité des solutions qui leur sont apportées, le comportement et le taux de satisfaction du client.

Le contexte applicatif pris en considération dans ce papier concerne l'analyse de dialogues entre un agent d'un centre d'appel et un client dont le comportement reste imprévisible. Le client est supposé demander des informations et/ou évoquer des difficultés en lien avec le système de transport parisien (RATP) et les services afférents. La documentation du centre d'appel décrit succinctement un nombre fini de motifs d'appels possibles, assimilables aux thèmes du dialogue.

Dans ce papier, nous nous concentrons sur la détection des thèmes, éventuellement multiples, présents dans le dialogue. Différents thèmes peuvent être mentionnés dans des segments disjoints du discours. Dans certains cas, les références aux différents thèmes peuvent coexister dans de courts segments ou même dans une seule phrase. Même quand les différents thèmes sont mentionnés dans des segments distincts, les frontières des segments peuvent être difficiles à tracer, tant à cause des erreurs commises par le système de transcription automatique (ASR) que de la difficulté à appréhender convenablement les structures linguistiques mises en œuvre dans une situation réelle par des usagers occasionnels.

Ce papier est structuré comme suit. La section 2 évoque les travaux antérieurs. La section 3 introduit le cadre applicatif et les descripteurs utilisés pour la détection de thèmes. La première des deux approches utilisées pour la tâche de classification multi-thématique, fondée sur une mesure de similarité *cosine*, est introduite dans la section 4. La seconde approche, décrite dans la section 5, introduit le concept de densité thématique, propre à permettre une détection locale des segments thématiques constitutifs du dialogue. Les résultats expérimentaux sont présentés dans la section 6.

2 Travaux antérieurs

Les méthodes d'analyse de dialogues entre humains ont été récemment passées en revue dans (Tur et Hakkani-Tür, 2011). Des méthodes d'identification thématique dans les documents audio sont examinées dans (Hazen, 2011). Des solutions pour la détection de segments de dialogues porteurs de différents thèmes ont été proposées dans de nombreuses publications, récemment passées en revue dans (Purver, 2011). Des solutions intéressantes ont été proposées pour des modèles linéaires de segmentation non hiérarchique. Certaines approches proposent des méthodes d'inférence pour extraire des points de segmentation localisés aux *maxima* de fonctions de cohésion. Les fonctions utilisent des paramètres extraits de chacune des phrases du dialogue ou d'une fenêtre qui couvre quelques phrases. Certaines méthodes de recherche détectent la cohésion d'un dialogue à partir de modèles de langage et d'autres prennent en considération des thèmes cachés partagés à travers les documents.

On peut trouver une revue critique de la notion de cohésion lexicale dans (Eisenstein et Barzilay, 2008), qui propose une approche non supervisée pour conjecturer des points de segmentation en utilisant des répliques automatiquement extraites de données non annotées. On trouve une évaluation de la segmentation *à grain grossier* dans (Niekrasz et Moore, 2010).

La catégorisation multi-thématique est abordée dans (Tsoumakas et Katakis, 2007) et (de Carvalho et Freitas, 2009), qui introduit une méthode dite de *création* qui consiste dans le cas de vastes corpus à créer de nouveaux labels composites par association des thèmes multiples attribués à une instance.

Ce papier généralise certains concepts de la littérature en introduisant une version des bigrammes à distance : dans une séquence de trois unigrammes, l'unigramme intermédiaire est ignoré et le bigramme est formé par concaténation des deux termes extrêmes. Ces termes sont utilisés dans le cadre de stratégies de décision fondées sur une mesure de similarité *cosine* et sur un nouveau concept de densité thématique locale.

3 Contexte applicatif et descripteurs

Notre papier propose une nouvelle approche pour annoter automatiquement les dialogues entre un agent et un client à l'aide d'un ou plusieurs des labels thématiques définis par le cadre applicatif, éléments de l'ensemble $\mathbb{C}$ défini comme suit :

$\mathbb{C}$:= {**itinéraire**, **objets perdus**, **horaires**, **tarifs**, **carte de transport**, **état du trafic**, **pv**, **appels spécialisés**, **autres**}

Chacune des deux méthodes de catégorisation multi-thématique présentées peut être définie comme une application $\gamma_i : \mathbb{X} \rightarrow \mathscr{P}(\mathbb{C})$ $(i = 1, 2)$ qui à tout dialogue d du corpus $\mathbb{X}$ associe une partie de $\mathbb{C}$. Les méthodes sont apprises à partir d'un sous-ensemble $\mathbb{T} \subset (\mathbb{X} \times \mathscr{P}(\mathbb{C}))$ de dialogues dont les thèmes ont été manuellement annotés.

A chaque couple (d, c) formé d'un dialogue d et d'une classe thématique c, on fait correspondre un couple de vecteurs $(\mathbf{v_d}, \mathbf{v_c})$.

Les composantes $w_c(t)$ du vecteur $\mathbf{v_c}$ sont données, pour tout terme t, par le produit $w_c(t) = df_c(t).idf(t).G(t)$, où $df_c(t)$ désigne le nombre de dialogues de la classe c contenant t, $idf(t)$ la fréquence inverse de document de t et $G(t)$ la pureté de t définie selon le critère de Gini par :

$$G(t) = \sum_{c \in \mathbb{C}} \mathbb{P}^2(c|t) = \sum_{c \in \mathbb{C}} \left(\frac{df_c(t)}{df_{\mathbb{T}}(t)} \right)^2 \quad (1)$$

avec $df_{\mathbb{T}}(t)$ correspondant au nombre de dialogues du corpus d'apprentissage qui contiennent au moins une fois t.

Les composantes $w_d(t)$ du vecteur $\mathbf{v_d}$ sont quant à elles définies, pour tout terme t, par $w_d(t) = tf_d(t).idf(t).G(t)$, où $tf_d(t)$ désigne le nombre total d'occurrences du terme t dans le dialogue d.

Afin d'enrichir les descripteurs, des bigrammes ont été ajoutés au vocabulaire initial de 7 217 mots en conjonction avec des bigrammes à distance formés par concaténation de mots distants

au plus de deux termes dans le cours du dialogue. Le nombre de descripteurs s'élève alors à 160 433 termes.

Afin d'éviter les effets induits par une telle dispersion des données, un ensemble réduit de descripteurs a été obtenu par sélection des termes sur la base de leur pureté $G(t)$, développée dans la formule (1), et de leur couverture.

4 Mesure globale de similarité *cosine*

La première méthode de classification γ_1 repose sur le calcul de l'indice de similarité *cosine* $\mathrm{sc}(d,c)$ entre les deux vecteurs $\mathbf{v_d}$ ($d \in \mathbb{X}$) et $\mathbf{v_c}$ ($c \in \mathbb{C}$), dont la formule est donnée par :

$$\mathrm{sc}(d,c) = \cos(\widehat{\mathbf{v_d}, \mathbf{v_c}}) = \frac{\sum_{t \in d \cap c} w_d(t).w_c(t)}{\sqrt{\sum_t w_d(t)^2 . \sum_t w_c(t)^2}}$$

La classe, notée $\hat{c}$, la plus similaire au dialogue traité lui est alors attribuée. D'éventuels thèmes secondaires lui sont en outre assignés si les scores qu'ils obtiennent excèdent un seuil fixé empiriquement et défini en proportion du score $\mathrm{sc}(d,\hat{c})$ obtenu par la classe la plus proche du dialogue. Toutefois, si le score maximal obtenu par la classe $\hat{c}$ reste en deçà d'un second seuil, fixé empiriquement en valeur absolue, aucun thème secondaire n'est attribué au dialogue et seul le thème dominant est retenu.

5 Mesure locale de densité thématique

Dans le cadre de la seconde approche, le dialogue n'est plus considéré comme un sac de mots dont l'ordre est indifférent mais linéairement dans l'ordre de son déroulement temporel. Un dialogue de N mots est alors vu, en incluant les espaces intermédiaires, comme une suite finie $(p_1, ..., p_n)$ de n positions (avec $n = 2N - 1$). Le $k-$ième unigramme du dialogue se situe alors à la position $2k-1$; le $k-$ième bigramme à la position $2k$ et le $k-$ième bigramme à un trou à la position $2k+1$.

Densité thématique : définition

La contribution $w_c(p_i)$ de la $i-$ème position du dialogue à la classe thématique $c \in \mathbb{C}$ est donnée par la somme des contributions normalisées à ce thème des termes (unigrammes et bigrammes) qui y sont localisés :

$$w_c(p_i) = \frac{1}{\| \mathbf{v_c} \|} \sum_{t \in T_{p_i}} w_c(t) \quad (i = 1, ..., n)$$

où T_{p_i} désigne l'ensemble des termes (unigrammes et bigrammes) situés à la $i-$ème position du dialogue.

A chaque position p_i du dialogue, on peut alors associer une mesure de densité thématique $d_c(p_i)$ pour chaque classe thématique c, évaluée selon la formule suivante :

$$d_c(p_i) = \frac{\sum_{j=1}^{n} \frac{w_c(p_j)}{\lambda^{d_j}}}{\sum_{j=1}^{n} \frac{1}{\lambda^{d_j}}} \; (i = 1, ..., n)$$

où $\lambda \geqslant 1$ est un paramètre de sensibilité au contexte local dont la valeur est estimée empiriquement à partir du corpus de développement ; et où d_j, la distance entre la position p_j et la position de référence p_i, est donnée par $d_j := |i - j|$.

Profil thématique d'un dialogue

La mesure de densité thématique en une position du dialogue permet d'en construire le profil thématique.

La FIGURE 1 montre le profil de la transcription automatique d'un dialogue pour $\lambda = 1.05$ (FIG. 1-a), et $\lambda = 3$ (FIG. 1-b). La densité thématique est tracée en fonction de la position dans le dialogue. Le dialogue porte sur un service de desserte aéroportuaire par bus : une mère s'enquiert pour le compte de sa propre fille, qui doit prendre dans quelques jours un avion, des horaires (notés HORR), ainsi que du tarif (noté TARF) de la desserte. Trois fonctions sont tracées : deux pour ces thèmes et une troisième pour le thème *itinéraire* (noté ITNR).

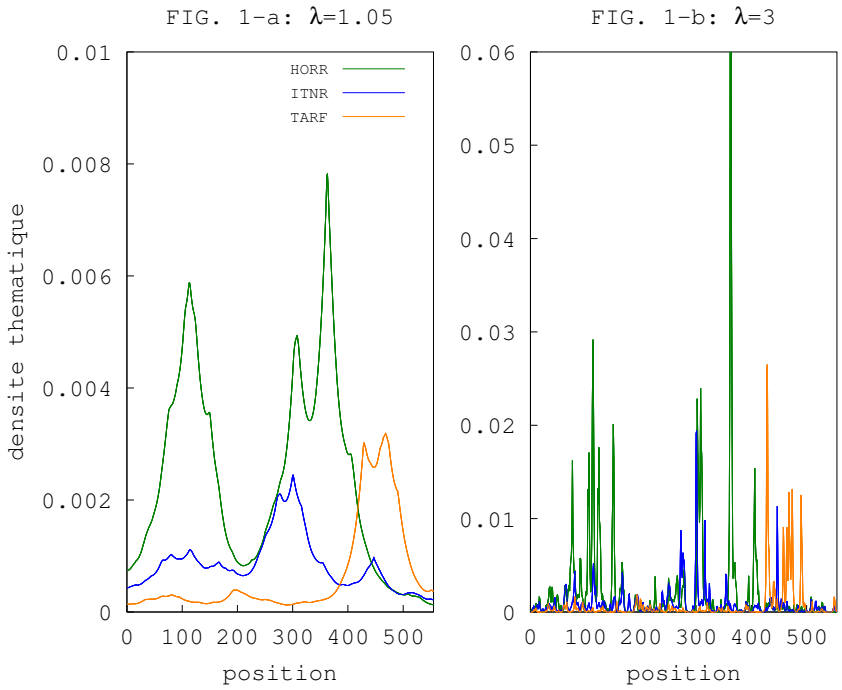


FIGURE 1 – Densités thématiques en fonction de la position dans le dialogue. Les densités sont tracées pour $\lambda = 1.05$ (FIG. 1-a) et $\lambda = 3$ (FIG. 1-b) et trois thèmes : horaires (noté HORR, tracé vert), itinéraire (noté ITNR, tracé bleu), et tarif (noté TARF, tracé orange).

On trouvera ci-après un extrait, manuellement transcrit pour plus de lisibilité, de ce dialogue. Les positions correspondantes dans la transcription automatique telles qu'elles apparaissent sur la figure 1 sont signalées entre crochets :

– *Client* [82–100] : je voudrais savoir si les (...) bus commencent tôt le matin.
– *Agent* [101–118] : cinq heures quarante-cinq le premier départ à l'opéra.

– (...)
– *Agent* [304–314] : à (...) huit heures quarante elle veut être **là-bas** ou neuf heures ?
– (...)
– *Agent* [442–470] : ah non tarification spécifique desserte aéroportuaire je vais vous donner le tarif (...) alors le tarif neuf euros dix
– *Client* [471–496] : neuf euros dix on vous donne (...) neuf euros dix en espèces ?

Pour $\lambda = 3$ (FIG. 1-b), le contexte thématique tend à être négligé. La décision d'affectation thématique peut alors indûment reposer sur des termes isolés fortement connotés isolément alors que le contexte thématique est tout autre. Par exemple, l'adverbe *là-bas* dans le tour de parole [304–314] (souligné dans l'extrait ci-dessus) tend à orienter la polarité thématique dans le sens d'une demande d'itinéraire avec un pic de densité pour ce thème autour de la position 300, alors qu'il est en fait utilisé dans le contexte d'une demande d'horaires.

Le contexte local est davantage pris en compte pour $\lambda = 1.05$ (FIG. 1-a) : la contribution de l'adverbe *là-bas* au thème *itinéraire* n'est pas soutenue par l'environnement thématique immédiat et voit sa valeur atténuée.

Convenablement estimé, le paramètre λ de sensibilité au contexte permet ainsi de prendre une décision fondée sur la cohérence thématique de segments entiers du dialogue en lissant les contributions thématiques locales par leur contexte.

Règles de décision

Deux règles de décision sont appliquées conjointement : un thème c est attribué à un dialogue d s'il est de densité dominante en une position quelconque du dialogue et si la somme de ses densités dans les positions où il est dominant excède un seuil fixé empiriquement.

6 Expériences et résultats

Cadre expérimental

Les expériences ont été menées en utilisant le système d'ASR décrit dans (Linarès *et al.*, 2007), à base de modèles de Markov cachés (HMM) acoustiques triphones avec des mélanges gaussiens appartenant à un ensemble de 230 000 distributions. Les paramètres du modèle ont été estimés selon une probabilité maximum *a posteriori* par adaptation des 150 heures de dialogues téléphoniques formées par les données du corpus d'apprentissage. Un corpus de 1 658 dialogues téléphoniques a été collecté au centre d'appel du service public des transports parisiens (RATP), puis décomposé en trois sous-ensembles : apprentissage (884 dialogues), développement (196) et test (578). Un modèle de langage trigramme a été obtenu en adaptant aux transcriptions de l'ensemble d'apprentissage un modèle de langage de base. Un ensemble initial d'expériences a été mené avec ce système avec un taux global d'erreurs mots (WER) sur le corpus de test de 57%. Ces taux d'erreurs élevés s'expliquent pour l'essentiel par les disfluences du discours et par des environnements acoustiques bruités pour certains dialogues quand, par exemple, des clients sont en train d'appeler depuis des gares ou des rues bruyantes avec des téléphones mobiles. Par ailleurs, le signal de certaines phrases est saturé ou de faible intensité à cause de la distance entre les locuteurs et les téléphones. Les corpus de développement et de test ont été annotés *a posteriori* par maximisation de l'accord entre trois annotateurs et le corpus d'apprentissage à la

volée par les agents eux-mêmes avec parfois pour consigne de n'indiquer que le thème principal.

Métriques d'évaluation

Les approches proposées ont été évaluées selon des procédures spécifiées dans (Tsoumakas et Katakis, 2007) en utilisant des métriques issues du domaine de la recherche d'informations, mais adaptées au cas particulier de la classification multi-thèmes.

Pour un corpus $\mathbb{X}$, un ensemble de classes $\mathbb{C}$ et une méthode de classification multi-thématique $\gamma : \mathbb{X} \to \mathscr{P}(\mathbb{C})$, ces métriques d'évaluation, en notant $L(d)$ l'ensemble des thèmes manuellement attribués au dialogue d, sont données par :

$$\textbf{Rappel :}\ R(\gamma,\mathbb{X}) = \frac{1}{|\mathbb{X}|}\sum_{d\in\mathbb{X}}\frac{|\gamma(d)\cap L(d)|}{|L(d)|} \quad - \quad \textbf{Précision :}\ P(\gamma,\mathbb{X}) = \frac{1}{|\mathbb{X}|}\sum_{d\in\mathbb{X}}\frac{|\gamma(d)\cap L(d)|}{|\gamma(d)|}$$

$$\textbf{F-score :}\ F(\gamma,\mathbb{X}) = \frac{2P(\gamma,\mathbb{X})R(\gamma,\mathbb{X})}{P(\gamma,\mathbb{X})+R(\gamma,\mathbb{X})} \quad - \quad \textbf{Accuracy :}\ A(\gamma,\mathbb{X}) = \frac{1}{|\mathbb{X}|}\sum_{d\in\mathbb{X}}\frac{|\gamma(d)\cap L(d)|}{|\gamma(d)\cup L(d)|}$$

Résultats

Aux deux méthodes de catégorisation γ_1 et γ_2 respectivement introduites dans les sections 4 et 5, un troisième classifieur γ_3 fondé sur des SVM a été adjoint à titre de référence. Pour l'entraînement et l'application des SVM, un noyau linéaire a été utilisé avec le même lexique de référence que les méthodes γ_i $(i = 1, 2)$. Pour chaque classe $c_j \in \mathbb{C}$, un classifieur binaire $\gamma_j : \mathbb{X} \to \{c_j, \overline{c_j}\}$ est défini. A tout couple $(d, c_j) \in \mathbb{X} \times \mathbb{C}$, on peut ainsi associer un score issu du classifieur binaire correspondant. Les thèmes conjecturés pour le dialogue courant d sont alors ceux dont le score représente une certaine proportion du score maximal. On impose de plus, comme condition d'attribution d'un thème secondaire, que le score de la classe dominante excède un certain seuil empiriquement fixé.

Les résultats obtenus par les SVM, la mesure de similarité *cosine* et la mesure de densité thématique sont reportés dans le tableau 1 pour le corpus de développement et dans le tableau 2 pour le corpus de test.

DEV	MAN			ASR		
	SVM	**cos.**	**dens.**	**SVM**	**cos.**	**dens.**
Accuracy	0.76	0.86	0.85	0.74	0.82	0.80
Précision	0.84	0.95	0.94	0.82	0.92	0.92
Rappel	0.87	0.89	0.88	0.88	0.84	0.85
F-score	0.86	**0.92**	0.91	0.85	**0.88**	**0.88**

TABLE 1 – Résultats obtenus par les SVM, la mesure de similarité *cosine* et la densité thématique sur le corpus de développement.

Analyse des résultats

Le corpus de développement a été collecté pendant la même période que le corpus d'apprentissage (automne) alors qu'une partie du corpus de test a été collectée pendant l'été. Les différences

TEST	MAN			ASR		
	SVM	**cos.**	**dens.**	**SVM**	**cos.**	**dens.**
Accuracy	0.73	0.81	0.79	0.69	0.75	0.75
Précision	0.81	0.90	0.88	0.76	0.84	0.84
Rappel	0.86	0.85	0.84	0.85	0.79	0.79
F-score	0.84	**0.88**	0.86	0.80	**0.81**	**0.81**

TABLE 2 – Résultats obtenus par les SVM, la mesure de similarité *cosine* et la densité thématique sur le corpus de test. Pour la méthode fondée sur la similarité *cosine*, l'intervalle de confiance estimé est de ± 2.74% pour les transcriptions manuelles et de ± 3% pour les sorties de l'ASR.

entre les résultats obtenus sur les ensembles de test et de développement peuvent donc en partie s'expliquer par l'hétérogénéité des motifs d'appel entre ces deux périodes (par exemple, les appels relatifs aux grèves, plus fréquents à l'automne, deviennent plus rares l'été et inversement, les appels relatifs aux travaux de maintenance sont caractéristiques de la période estivale).

Par ailleurs, les différentes stratégies développées pour conjecturer un ou plusieurs thèmes secondaires en plus du thème dominant permettent toutes d'améliorer les performances par rapport à une stratégie qui affecterait au dialogue le seul thème dominant. Étendue à la recherche d'éventuels thèmes secondaires dans les dialogues du corpus de développement, la méthode de densité thématique permet par exemple d'améliorer le F-score de 0.89 à 0.91 pour les transcriptions manuelles et de 0.84 à 0.88 pour les transcriptions automatiques. Les améliorations obtenues sur le corpus de test sont du même ordre.

Enfin, en utilisant le corpus de développement pour inférer une règle de rejet d'un dialogue si les scores obtenus selon les deux classes dominantes sont trop proches, le F-score obtenu sur le corpus de test se trouve rehaussé à 0.83 si l'on applique un taux de rejet de près de 10%, soit le taux de désaccord entre des annotateurs humains.

7 Conclusion et perspectives

Pour conjecturer un ou plusieurs des thèmes évoqués dans un dialogue entre un agent de centre d'appel et un client, on a proposé d'utiliser un lexique d'unigrammes, de bigrammes et de bigrammes à distance automatiquement extraits du domaine de l'application.

Deux approches ont été introduites pour conjecturer des thèmes multiples. La première repose sur une mesure globale de similarité *cosine* appliquée aux termes extraits du dialogue ; la seconde se fonde sur une nouvelle notion de densité thématique sensible au profil thématique du dialogue. Chacune de ces deux approches s'est montrée plus performante qu'un classifieur à base de SVM fondé sur le même vocabulaire et appliqué aux mêmes données.

Les recherches à venir se concentreront sur la question des mesures de confiance appropriées à la tâche de catégorisation multi-thématique et sur l'extraction automatique de brefs rapports des dialogues à partir de leur profil thématique. Sur la base de tels rapports, des statistiques utiles à l'amélioration du service rendu à l'usager pourront être déduites, par exemple sur la fréquence des problèmes rencontrés par les clients et l'efficacité des solutions qui leur sont apportées par les conseillers.

Références

de Carvalho, A. C. et Freitas, A. A. (2009). A tutorial on multi-label classification techniques. *In Found. of Computational Intelligence Volume 5*, pages 177–195. Springer.

Eisenstein, J. et Barzilay, R. (2008). Bayesian unsupervised topic segmentation. *In Proceedings of the 2008 Conference on Emp. Methods in Nat. Lang. Processing*, pages 334–343. Association for Computational Linguistics.

Hazen, T. J. (2011). Topic identification. *In Spok. Lang. Underst.*, pages 319–356. John Wiley & Sons, Ltd.

Linarès, G., Nocéra, P., Massonie, D. et Matrouf, D. (2007). The lia speech recognition system : from 10xrt to 1xrt. *In Text, Speech and Dialogue*, pages 302–308. Springer.

Niekrasz, J. et Moore, J. D. (2010). Unbiased discourse segmentation evaluation. *In Spok. Lang. Technology Workshop (SLT)*, pages 43–48. IEEE.

Purver, M. (2011). Topic segmentation. *Spok. Lang. Underst. : Syst. for Extracting Semantic Inf. from Speech*, pages 291–317.

Tsoumakas, G. et Katakis, I. (2007). Multi-label classification : An overview. *Int. J. of Data Warehous. and Min. (IJDWM)*, 3(3):1–13.

Tur, G. et De Mori, R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.

Tur, G. et Hakkani-Tür, D. (2011). Human/human conversation understanding. *Spok. Lang. Underst. : Syst. for Extracting Semantic Inf. from Speech*, pages 225–255.

Analyse factorielle pour une représentation vectorielle des états des Modèles de Markov Cachés

Mohamed Bouallègue, Driss Matrouf et Georges Linarès
Université d'Avignon, LIA, France
{mohamed.bouallegue, driss.matrouf, georges.linares}@univ-avignon.fr

RÉSUMÉ

Dans cet article nous proposons une représentation vectorielle des états des modèles de Markov cachés (HMM, Hidden MArkov Models). L'idée de base nous a été inspirée par l'approche SGMM (Subspace Gaussian Mixture Models paradigm). Modéliser les états des HMMs par de simples vecteurs (au lieu de GMMs) permet de faciliter un grand nombre de tâches dans le cadre du traitement automatique de la parole, comme la classification automatique des états ou des phonèmes. La représentation vectorielle des états permet de voir les états comme un nuage de points dans un espace multi-dimensionnel, ce qui permet d'étudier ses caractéristiques en utilisant des techniques d'analyse de données. Dans cet article, nous expliquerons comment obtenir cette représentation et comment l'utiliser pour réaliser la procédure de partage d'états pour estimer des modèles contextuels avec des états partagés. Nous comparons notre approche avec celle fondée sur l'utilisation des arbres de décision.

ABSTRACT

Subspace Gaussian Mixture Models for Vectorial HMM-states Representation

In this paper we present a vectorial representation of the HMM states that is inspired by the Subspace Gaussian Mixture Models paradigm (SGMM). This vectorial representation of states will make possible a large number of applications, such as HMM-states clustering and graphical visualization. Thanks to this representation, the Hidden Markov Model (HMM) states can be seen as sets of points in multi-dimensional space and then can be studied using statistical data analysis techniques. In this paper, we show how this representation can be obtained and used for tying states of an HHM-based automatic speech recognition system without any use of linguistic or phonetic knowledge.

MOTS-CLÉS : Représentation vectorielle des états des HMM, partage d'états, classification des états d'un HMM.

KEYWORDS: HMM-state vector representation, Speech recognition, Acoustic Modelling, HMM states clustering..

1 Introduction

Depuis les premières applications dans le domaine du traitement de la parole, la problématique du calcul de la similitude entre phonèmes (ou allophones) a été posée comme sujet de recherche par la communauté scientifique. En reconnaissance de la parole, cette mesure est utilisée principalement dans la procédure de "partage d'états". Cette approche proposée par Young (Young, 1992)

consiste à associer la même fonction de densité de probabilité (probability density function : pdf) aux états des phonèmes contextuels acoustiquement proches. Le "partage d'états" est une procédure incontournable dans le processus de la modélisation acoustique afin d'avoir un équilibre entre la quantité de données d'apprentissage (pour les phonèmes contextuels) et le nombre de paramètres à estimer dans les modèles acoustiques. Elle est basée sur la définition d'une distance entre les états des HMM (mesure de similitude). En supposant que chaque phonème est modélisé par une seule gaussienne, Young (Young et Woodland, 1993) défnit la similitude entre phonèmes en utilisant une distance entre deux gaussiennes, appelée la divergence de Kullback-Leibler [1]. Pour calculer la distance entre deux mélanges de gaussiennes, Mak a proposé une expression alternative de la similitude basée sur la distance de Bhattacharyya [2] (Mak et Barnard, 1996). La distance de Bhattacharyya permet de mesurer la distance théorique entre deux distributions gaussiennes. Mak a proposé dans son travail de l'étendre au calcul de la distance entre deux mélanges de gaussiennes.

Dans ce travail, nous proposons une nouvelle vision de la classification des phonèmes. Au contraire des travaux mentionnés auparavant, les phonèmes sont caractérisés par des vecteurs estimés par l'approche d'analyse factorielle (Kenny *et al.*, 2005). Un intérêt indéniable de cette représentation vectorielle est la possibilité de traiter les états par des techniques d'analyse de données, telle que l'analyse factorielle des correspondances ou l'analyse en composantes principales. Le fait de représenter les états par des vecteurs permet aussi de mesurer efficacement la similarité entre eux, en utilisant des distance adaptées (distance euclidienne ou distance de Mahlanobis). Il est important de noter que cette tâche (mesure de similarité) était très complexe et assez approximative lorsque les états ne sont représentés que par des GMMs. Nous proposons ici l'utilisation de cette représentation vectorielle dans la modélisation acoustique pour la reconnaissance de la parole. Nous nous basons sur les vecteurs pour réaliser la procédure de regroupement d'états des MMCs. Nous appelons les vecteurs représentatifs des états des MMCs *facteurs d'états*.

Dans la section 2.1 nous exposons la méthode d'estimation des *facteur d'états*. Nous détaillons dans la section 2.2 les différentes étapes de la procédure de regroupement des états. Nous présentons ensuite les résultats obtenus sur la langue française. Dans la section 4 nous montrons l'intérêt de l'application de cette méthode dans la modélisation acoustique pour les langues peu dotées, telles que le vietnamien.

2 Représentation vectorielle des états

Dans cette section nous exposons le modèle utilisé ainsi que la procédure de regroupement d'états pour une mutualisation des paramètres dans le cadre de la modélisation acoustique.

1. www.aiaccess.net/french/Glossaires/GlosMod/f_gm_kullbakh̀tm
2. http ://www.cse.yorku.ca

2.1 Analyse factorielle pour une représentation vectorielle des états de MMC

Dans la modélisation acoustique basée sur l'analyse factorielle, tous les GMMs associés aux états des MMC sont dérivés d'un seul modèle générique appelé "modèle du monde" (GMM-UBM). C'est un mélange de gaussiennes modélisant tout l'espace acoustique de la parole. Les moyennes et les poids des états sont obtenus à partir de l'UBM, alors que les variances restent inchangées par rapport à l'UBM. Soit $\mathbf{m}$ le super-vecteur du modèle du monde résultant de la concaténation des moyennes de ses gaussiennes. Le super-vecteur $\mathbf{m}_e$ de l'état e est obtenu par l'équation suivante :

$$\mathbf{m}_e = \mathbf{m} + \mathbf{U}\mathbf{x}_e \tag{1}$$

Dans ce modèle, la matrice $\mathbf{U}$, de faible dimension R, représente le sous-espace de variabilité inter-états. Elle est estimée sur toutes les données de tous les états des MMC. $\mathbf{x}_e$ est un vecteur de dimension R estimé sur les données propres à l'état e. C'est un vecteur caractéristique de l'état que nous appelons *facteur d'état*. L'algorithme d'estimation de $\mathbf{U}$ et des $\mathbf{x}_e$ est décrit en détail dans (Matrouf *et al.*, 2007).

Nous proposons d'adopter les *facteurs d'état* comme une représentation vectorielle des états des MMC. Les *facteurs d'états* simplifient le calcul des distances entre les états des MMC. En effet, nous remplaçons les mélanges des gaussiennes par un simple vecteur de faible dimension (maximum 100 paramètres). Les distances utilisées généralement dans la classification phonétique sont remplacées par un simple calcul de distance entre vecteurs. Avec la représentation vectorielle, la classification phonétique devient un problème de classification classique dans un espace R^d. Par ailleurs, cette représentation vectorielle permettra d'exploiter les résultats scientifiques obtenus au cours de plusieurs années de recherche dans le domaine de l'analyse de données. Ces résultats peuvent servir dans l'analyse de la variabilité acoustique et de la variation de l'articulation phonétique. Enfin, les *facteurs d'états* peuvent être utilisés dans d'autres applications comme la phonétique clinique, la détection d'un dialecte ou l'identification automatique de la langue.

2.2 Procédure de regroupement d'états basée sur des facteurs d'états

Dans la littérature, le regroupement est souvent réalisé par un algorithme de classification utilisant les arbres de décision (Reichl et Chou, 1998). Plusieurs inconvénients sont inhérents à l'utilisation de cette méthode : elle nécessite notamment des connaissances linguistiques (pour construire le jeu de questions), qui peuvent ne pas être disponibles pour certaines langues. De plus, le temps de calcul est long à cause de l'évaluation des vraisemblances pour chaque question à chaque noeud de la structure de l'arbre. Comme alternative, nous proposons d'utiliser les *facteurs d'états* dans le regroupement des états des MMC. Avec les *facteurs d'états* le calcul de la probabilité est remplacé par un simple calcul de distance entre vecteurs et le regroupement des états peut être formulé comme un problème de classification classique dans R^d. Cette méthode ne nécessite plus de connaissances phonétiques ou linguistiques, ce qui nous permet de contourner le problème d'insuffisance ou l'absence de ce type d'informations pour les langues peu dotées. Dans la suite nous exposons les différentes étapes de la réalisation du partage d'états basé sur les *facteurs d'états*.

Nous présentons dans cette section les étapes nécessaires pour réaliser le regroupement d'états en utilisant les facteurs d'états.

– Etape-1 : utilisation d'un système de reconnaissance de la parole existant pour faire la segmentation par rapport aux états, par un classique alignement forcé ; nous utilisons l'algorithme de Viterbi pour cette segmentation. Chaque phonème indépendant du contexte est représenté par trois états émetteurs.

– Etape-2 : contextualisation des phonèmes dans la segmentation obtenue dans l'étape-1. En effet, nous associons à chaque phonème le phonème précédent et le phonème suivant. L'objectif de cette étape est de trouver tous les phonèmes contextuels dans notre corpus d'apprentissage.

– Etape-3 : estimation des *facteurs d'états* $\mathbf{x}_s$ pour chaque état avec l'équation 1. Une bonne estimation des $\mathbf{x}_s$ nécessite un nombre minimum de trames pour chaque état. Pour cela nous ignorons les états qui sont rarement observés dans le corpus d'apprentissage et nous ne calculons les *facteurs d'états* que pour les états qui ont suffisamment de trames (dans nos expériences nous avons traité les états qui ont plus de 50 trames).

– Etape-4 : utilisation des *facteurs d'états* obtenus précédemment pour classifier les états en utilisant l'algorithme de classification non-supervisée k-means [3]. Cette classification permet de regrouper les états dépendants du contexte qui sont acoustiquement proches. Nous appelons les classes obtenues, *classe-états*.

– Etape-5 : modélisation des états appartenant à la même classe par un seul GMM appelé *GMMs-classe-états*. Ces GMMs sont dérivés du GMM-UBM en utilisant une adaptation par maximum a posteriori (MAP) au moyen des données appartenant à chaque *classe-état* (obtenu dans l'étape-4).

– Etape-6 : affectation des états non classifiés dans l'étape-4 (qui ont moins de 50 trames) à la classe la plus proche. Pour ce faire, nous calculons la vraisemblance des trames de chaque état non classifiées sur les *GMMs-classe-état* obtenus dans l'étape-5. Après, nous assemblons ces trames avec les trames de classe la plus vraisemblable.

– Etape-7 : adaptation de chaque *GMMs-classe-état* obtenu dans l'étape-5 sur les données de la *classe-états* correspondante incluant les données des états nouvellement classifiés.

– Etape-8 : construction des MMCs en utilisant les *GMMs-classe-état* obtenus dans l'étape précédente.

– Etape-9 : Pour améliorer les performances de notre modèle, nous appliquons la procédure récursive standard du ré-alignement/ré-estimation des paramètres.

3. l'algorithme k-means permet une classification non-hiérarchique en minimisant la variance intra-classe basée sur la distance Euclidienne ; cette procédure est un moyen simple pour classifier un ensemble de données dans un certain nombre de classes (noté k) préalablement fixé

3 Résultats expérimentaux

Dans ces expériences, nous avons utilisé le système SPEERAL avec un processus de transcription en deux passes. Les nouveaux modèles acoustiques sont estimés et évalués sur le corpus d'évaluation ESTER. Les modèles acoustiques de base sont des MMC gauche-droit avec 13316 phonèmes contextuels. Pour modéliser tous ces phonèmes contextuels indépendamment, nous avons besoin de 39.948 états dans les MMCs. Ce nombre est réduit à 5050 états par l'application du regroupement d'états basé sur des arbres de décision descendantes. Cette approche utilise des questions linguistiques pour construire les arbres de décision.

Dans cette expérience, nous avons fixé le nombre de paramètres des *facteurs d'états* à 60. Nous rappelons que les GMMs d'états des MMCs sont dérivés d'un seul modèle du monde avec une adaptation MAP. Pour améliorer les performances, nous proposons de modéliser indépendamment chaque état en utilisant l'algorithme EM (Espérance-Maximisation). Dans notre procédure de modélisation, une fois la classification terminée, nous estimons un GMM pour chaque classe d'états avec l'algorithme EM. Les résultats sont exposés dans le tableau 1.

	modèle de base	Modèle MAP	modèle EM
F.Inter	30.88	30.10	29.38
RFI	16.62	16.09	15.84
TVME	25.74	24.90	24.77
AFRICA	29.42	28.89	27.47
Moyenne	27.50	26.61	25.66
gain absolu	-	0.89	1.84

TABLE 1 – *Les performances du modèle MAP et modèle EM comparées avec le modèle de base.*

4 Modèles acoustiques pour les langues peu dotées : le vietnamien

Parmi les 6 912 langues parlées dans le monde, seul un tout petit nombre d'entre elles possède les ressources nécessaires pour implémenter des technologies issues du traitement du langage naturel. Au niveau de la modélisation acoustique, plusieurs travaux sont proposés dans la littérature pour contourner les difficultés liées à l'absence partielle ou totale de ressources linguistiques et informatiques. La solution la plus utilisée aujourd'hui est le *bootstraping* (Osterholtz *et al.*, 1992). Cette solution consiste à obtenir un tableau de correspondances phonétiques (*phone mapping*) entre une ou plusieurs langues sources et la langue cible (la langue peu dotée). Nous distinguons deux classes de méthodes pour réaliser cette solution : la première classe comprend les méthodes manuelles à base de connaissances linguistiques et phonétiques (Le et Besacier, 2009). Elles consistent à chercher les couples de phonèmes source/cible les plus proches dans le tableau d'Alphabet Phonétique International(API). Dans une deuxième classe, nous trouvons les méthodes automatiques (Anderson *et al.*, 1994). Ces méthodes utilisent des modèles acoustiques de la langue source et un corpus vocal transcrit de la langue cible pour calculer la matrice de confusion entre les phonèmes et trouver le tableau de correspondances phonétiques. Ces méthodes sont utilisées pour construire des modèles acoustiques indépendants du contexte. Pour

construire des modèles acoustiques contextuels, des travaux comme (Beulen et Ney, 1998) (Singh *et al.*, 1999) proposent d'utiliser la procédure classique basée sur les arbres de décision. Les questions nécessaires à la construction des arbres sont générées automatiquement. D'autres travaux proposent de combiner les modèles acoustiques (au niveau des phonèmes contextuels) de différentes langues sources afin d'obtenir des modèles contextuels de la langue cible (Beyerlein, 1998) (Schultz et Waibel, 2001).

La plupart des travaux concernant la modélisation acoustique contextuelle, cités auparavant, sont basés sur l'idée d'association entre une ou plusieurs langues sources et la langue cible. Les différentes applications montrent que l'association à base de connaissances linguistiques donnent de meilleures résultats que les méthodes automatiques basées sur le calcul de distance entre modèles acoustiques. Mais ces solutions (à base de connaissances linguistiques) se retrouvent toujours face au problème de la couverture phonétique. De plus, elles nécessitent des connaissances linguistiques et phonétiques des deux langues source et cible, ce qui n'est pas toujours disponible (en quantité et en qualité) pour la langue cible. Afin de contourner ces difficultés, nous proposons d'appliquer la méthode de modélisation acoustique contextuelle proposée dans la section 2.2. Dans la suite nous montrons les résultats obtenus sur la langue vietnamienne.

La langue vietnamienne appartient au groupe Viet-Muong, qui est un membre de la branche mon-khmer qui fait partie de la branche austro-asiatique [4]. Elle est parlée par environ 82 millions de personnes, principalement au Vietnam. Le vietnamien est une langue tonale qui possède six tons avec des caractères accentués pour les tons. Le vietnamien possède 41 phonèmes, dont 23 consonnes, 13 voyelles simples, 3 diphtongues et 2 semi-voyelles, représentées par 29 lettres dans l'alphabet (N. Thi, 2006).

Corpus de la parole : dans ces expériences nous avons utilisé une partie du corpus VNSPEECH-CORPUS (Tran, 2003). C'est un corpus de 39 heures de parole enregistrée au centre MICA [5]. En 2005, il contenait 39 locuteurs, 19 femmes et 20 hommes, venant des régions nord, centre et sud du Vietnam. Nous utilisons uniquement les enregistrements de locuteurs d'une langue standard (nord du Vietnam). Au total, environ 9 heures de parole sont enregistrées, ce qui correspond à 18 locuteurs : 10 hommes et 8 femmes. Nous avons utilisé 7 heures pour l'apprentissage des modèles (8 hommes et 6 femmes) et 2 heures pour le corpus de test (2 hommes et 2 femmes).

Corpus de texte : le corpus de texte vietnamien, utilisé pour estimer le modèle de langage, est collecté exclusivement à partir du web et des journaux numériques. Ce corpus est constitué de 2,7 millions de phrases avec 45 millions de syllabes.

Modèles contextuels obtenus en utilisant les arbres de décision

Le modèle de phonèmes indépendants du contexte est obtenu avec le *bootstraping* (Osterholtz *et al.*, 1992). Un tableau de correspondances phonétiques est construit entre les phonèmes vietnamiens et les phonèmes du français en utilisant l'Alphabet Phonétique International. Les modèles des phonèmes contextuels sont obtenus par un arbre de décision basé sur des questions générées automatiquement (Singh *et al.*, 1999). Nous avons construit plusieurs modèles acoustiques de différentes tailles afin de trouver le nombre optimal de paramètres. Nous avons fait varier les nombres d'états des MMCs et le nombre de gaussiennes par état. Dans le tableau 2, nous exposons les résultats en termes de taux d'erreur mot des différents modèles. Chaque colonne contient les nombres d'états dans les MMCs. Les lignes correspondent aux nombres de gaussiennes par état.

4. http ://www.ethnologue.com
5. http ://www.mica.edu.vn

Les résultats montrent que le modèle qui est composé de 600 états avec 64 gaussiennes obtient les meilleures performances (32,70% WER). Ce modèle sera notre modèle de base.

	200s	400s	600s	800s	1200s
32g	33.11	33.51	34.70	34.90	36.10
64g	32.81	32.73	**32.70**	33.00	34.90
128g	33.83	33.80	34.40	35.30	36.40
256g	34.00	33.97	36.10	35.40	38.40

TABLE 2 – *Performances des modèles acoustiques contextuels de différentes tailles où le regroupement des états contextuels est basé sur des arbres de décision.*

***modèle contextuels du vietnamien basé sur les* facteurs d'états**

Dans ces expériences, nous appliquons la méthode de modélisation contextuelle proposée sur le vietnamien. Dans une première expérience, nous avons estimé quatre groupes de *facteurs d'états* qui ont respectivement 20, 40, 80 et 120 coefficients. Par la suite, nous utilisons chaque groupe de vecteurs pour regrouper les états dans 1000, 1200, 1400, 1600, 1800, 2000 et 2200 classes respectivement. Le nombre de gaussiennes par état est fixé à 128. Dans le tableau 3, nous exposons le WER obtenu par les différents modèles. Chaque ligne du tableau correspond au nombre d'états des modèles. En colonne, nous avons les tailles des *facteurs d'états* qui ont été utilisées dans la phase de regroupement.

	20	40	60	80	120
1000	28.80	29.50	36.15	38.50	57.10
1200	28.00	28.80	35.10	38.40	60.70
1400	28.10	28.20	33.80	35.70	52.30
1600	27.20	27.80	33.20	35.10	49.90
1800	26.80	27.20	30.70	34.20	50.50
2000	**26.60**	27.50	30.90	33.80	49.00
2200	26.90	27.50	31.10	33.55	49.24

TABLE 3 – *Performances des modèles de différents nombres d'états MMC, avec 128 gaussiennes par états.*

Le meilleur modèle donne un gain absolu de 6,10% par rapport au modèle de base. Les résultats obtenus montrent que les modèles ayant le plus grand nombre d'états obtiennent de meilleurs résultats. En outre, nous observons qu'à partir de 1800 états, le gain se stabilise. Ces résultats montrent que, à cause de la quantité limitée de données d'apprentissage disponible pour chaque état, les meilleurs *facteurs d'états* sont ceux de plus faible dimension (20 coefficients).

Dans une deuxième expérience, nous avons essayé de trouver le nombre optimal de gaussiennes par états. Le nombre de paramètres des *facteurs d'états* est fixé à 20. Nous évaluons le nombre d'états de 1600, 1800 et 2000. Pour chaque modèle, le nombre des gaussiennes varie de 64, 128 et 256 gaussiennes.

Le tableau 4 montre que le meilleur modèle est celui avec 2000 états et 256 gaussiennes par état, avec un gain de 7,8% absolu par rapport au modèle de base. Pour améliorer les performances du

	64g	128g	256g
1600	30.60	27.80	26.00
1800	29.80	27.20	25.50
2000	29.30	26.60	**24.90**

TABLE 4 – *Performances des modèles en fonction du nombres d'états et du nombre de gaussiennes.*

système, nous avons testé une approche de regroupement guidé. Nous rappelons que chaque phonème contextuel *ph* est modélisé par trois états. Dans une première étape, nous regroupons ensemble les *facteurs d'états* qui représentent les différents contextes du même phoneme. Nous obtenons ainsi 38 classes. Dans une deuxième étape, nous utilisons les *facteurs d'états* pour classifier les états à l'intérieur de chaque classe de phonèmes *ph*. Le nombre de gaussiennes par état est de 256. Les résultats présentés dans le tableau 5 montrent que le regroupement guidé nous permet un gain de 0,7% absolu par rapport au regroupement global. Nous obtenons donc un gain absolu 8,50% par rapport au modèle de base.

	1600s	1800s	2000s	2200s	2400s
WER	25.00	24.80	24.90	24.30	**24.20**

TABLE 5 – *Results de regroupement guidé.*

5 Conclusion

Dans cet article, nous avons proposé une nouvelle représentation vectorielle des états des MMCs. Cette représentation est obtenue avec le paradigme d'analyse factorielle. Dans la modélisation acoustique contextuelle, nous avons exploité cette représentation vectorielle dans la réalisation de la procédure de regroupement d'états des MMC. La classification des états dans cette procédure est basée seulement sur l'information portée par les vecteurs *facteurs d'états*. L'application de cette méthode sur la langue française a monté une amélioration significative des performances. Nous avons évaluer cette méthode basées sur les *facteurs d'états* dans la modélisation acoustique pour les langues peu dotées. La modélisation acoustique pour cette classe de langues souffrent du manque de ressources informatiques et linguistiques. Avec les *facteurs d'états* nous sommes arrivés à contourner ces contraintes pour construire des modèles acoustiques dépendants du contexte ayant de bonnes performances, avec un gain relatif de 23% pour la langue vietnamienne par rapport au système de référence obtenu par la technique de l'arbre de décision standard.

Références

ANDERSON, O., DALSGAARD, P. et BARRY, W. (1994). On the use of data-driven clustering techniques for language identification of poly and mono-phonemes for four european languages. pages 121–124. ICASSP, Adelaide.

Beulen, K. et Ney, H. (1998). Automatic question generation for decision tree based state tying. *In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, pages 805–809, Seattle, WA, USA.

Beyerlein, P. (1998). Discriminative model combination. *In ICASSP*, volume 1. Institute of Electrical Engineers (IEE).

Kenny, P., Boulianne, G., Ouellet, P. et Dumouchel, P. (2005). Factor Analysis Simplified. *In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005)*, volume 1, pages 637–640.

Le, V. B. et Besacier, L. (2009). Automatic speech recognition for under-resourced languages : Application to vietnamese language. *IEEE Transactions on Audio, Speech and Language Processing*, 17(8):1471–1482.

Mak, B. et Barnard, E. (1996). Phone clustering using the bhattacharyya distance. *In ICSLP*. ISCA.

Matrouf, D., Scheffer, N., Fauve, B. G. B. et Bonastre, J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. *In INTERSPEECH*, pages 1242–1245.

N. Thi, H. (2006). *Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens*. Thèse de doctora, Université Henri Poincaré, Nancy, France.

Nonyme, A. (2013a). Un article de revue. *Une grande revue*, 2(1):20–40.

Nonyme, A. (2013b). *Un livre*. Éditeur du livre.

Nonyme, A. (2014). Un magnifique article. *In XXX[e] Journées d'Études sur la Parole (JEP 2014)*, Le Mans, France.

Osterholtz, L., Augustine, C., McNair, A., Rogina, I., Saito, H., Sloboda, T., Tebelskis, J. et Waibel, A. (1992). Testing generality in janus : A multi-lingual speech translation system.

Reichl, W. et Chou, W. (1998). Decision tree state tying based on segmental clustering for acoustic modeling. *In icassp*, volume 2, pages 801–804.

Schultz, T. et Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1-2):31–51.

Singh, R., Raj, B. et M.Stern, R. (1999). Automatic clustering and generation of contextual questions for tied states in hidden markov models. volume vol. 1, pages 117–120. International Conference on Spoken Language Processing (ICSLP).

Tran, D.-D. (2003). Building a large vietnamese speech database. Rapport de master tic, Vietnam.

Young, S. J. (1992). The general use of tying in phoneme-based hmm speech recognisers. *In ICASSP*, pages 569–572.

Young, S. J. et Woodland, P. C. (1993). The use of state tying in continuous speech recognition. *In EUROSPEECH*. ISCA.

Remerciements à :