Retour articulatoire visuel par échographie linguale augmentée : développements et application clinique

Diandra Fabre

▶ **To cite this version:**

Diandra Fabre. Retour articulatoire visuel par échographie linguale augmentée : développements et application clinique. Traitement du signal et de l'image [eess.SP]. Université Grenoble Alpes, 2016. Français. NNT : 2016GREAT076 . tel-01502322

HAL Id: tel-01502322
https://theses.hal.science/tel-01502322

Submitted on 5 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Signal, Image, Parole, Télécom (SIPT)**

Arrêté ministériel : 25 mai 2016

Présentée par

Diandra FABRE

Thèse dirigée par **Pierre BADIN** et
codirigée par **Mélanie CANAULT**

préparée au sein du **Département Parole et Cognition (DPC) du GIPSA-Lab, Grenoble** et du **laboratoire Dynamique du Langage, Lyon**
dans **l'École Doctorale Électronique, Électrotechnique, Automatique et Traitement du Signal (EEATS)**

Retour articulatoire visuel par échographie linguale augmentée : développements et application clinique

Thèse soutenue publiquement le **16 décembre 2016**,
devant le jury composé de :

M. Pierre BADIN
Directeur de Recherche, GIPSA-LAB, Grenoble (Directeur de thèse)
Mme Mélanie CANAULT
Maître de Conférences, Université Lyon 1 (Co-directrice de thèse)
M. Thomas HUEBER
Chargé de Recherche, GIPSA-LAB, Grenoble (Co-directeur de thèse)
Mme Nathalie BEDOIN
Maître de Conférences, Université Lyon 2 (Co-directrice de thèse)
M. Eric TRUY
Professeur des Universités – Praticien Hospitalier, Lyon (Président)
M. Olov ENGWALL
Professeur des Universités, KTH, Suède (Rapporteur)
M. Slim OUNI
Maître de Conférences, LORIA, Nancy (Rapporteur)
M. Thierry ARTIÈRES
Professeur des Universités, ECM, Marseille (Examinateur)

Sans la curiosité de l'esprit, que serions-nous ? Telle est bien la beauté et la noblesse de la science : désir sans fin de repousser les frontières du savoir, de traquer les secrets de la matière et de la vie sans idée préconçue des conséquences éventuelles.
Marie Curie

Le progrès, ce n'est pas l'acquisition de biens. C'est l'élévation de l'individu, son émancipation, sa compréhension du monde. Et pour ça il faut du temps pour lire, s'instruire, se consacrer aux autres.
Christiane Taubira

Hâtez-vous lentement, et, sans perdre courage,
Vingt fois sur le métier remettez votre ouvrage :
Polissez-le sans cesse et le repolissez ;
Ajoutez quelquefois, et souvent effacez.
Nicolas Boileau

Résumé

Dans le cadre de la rééducation orthophonique des troubles de la parole associés à un mauvais positionnement de la langue, il peut être utile au patient et à l'orthophoniste de visualiser la position et les mouvements de cet articulateur naturellement très peu visible. L'imagerie échographique peut pallier ce manque, comme en témoignent de nombreuses études de cas menées depuis plusieurs années dans les pays anglo-saxons. Appuyés par de nombreux travaux sur les liens entre production et perception de la parole, ces études font l'hypothèse que ce *retour articulatoire visuel* faciliterait la rééducation du patient. Lors des séances orthophoniques, le patient semble, en effet, mieux appréhender les déplacements de sa langue, malgré la difficulté d'interprétation sous-jacente de l'image échographique liée au bruit inhérent à l'image et à l'absence de vision des autres articulateurs. Nous développons dans cette thèse le concept d'*échographie linguale augmentée*. Nous proposons deux approches afin d'améliorer l'image échographique brute, et présentons une première application clinique de ce dispositif.

La première approche porte sur le suivi du contour de la langue sur des images échographiques. Nous proposons une méthode basée sur une modélisation par apprentissage supervisé des relations entre l'intensité de l'ensemble des pixels de l'image et les coordonnées du contour de langue. Une étape de réduction de la dimension des images et des contours par analyse en composantes principales est suivie d'une étape de modélisation par réseaux de neurones. Nous déclinons des implémentations mono-locuteur et multi-locuteur de cette approche dont les performances sont évaluées en fonction de la quantité de contours manuellement annotés (données d'apprentissage). Nous obtenons pour des modèles mono-locuteur une erreur de 1,29 mm avec seulement 80 images, performance meilleure que celle de la méthode de référence *EdgeTrak* utilisant les contours actifs.

La deuxième approche vise l'animation automatique, à partir des images échographiques, d'une *tête parlante articulatoire*, c'est-à-dire l'avatar d'un locuteur de référence qui révèle les structures externes comme internes de l'appareil vocal (palais, langue, pharynx, dents, *etc.*). Nous construisons tout d'abord un modèle d'association entre les images échographiques et les paramètres de contrôle de la langue acquis sur ce locuteur de référence. Nous adaptons ensuite ce modèle à de nouveaux locuteurs dits locuteurs *source*. Pour cette adaptation, nous évaluons la technique *Cascaded Gaussian Mixture Regression* (*C-GMR*), qui s'appuie sur une modélisation conjointe des données échographiques du locuteur de référence, des paramètres de contrôle de la tête parlante, et

des données échographiques d'adaptation du locuteur source. Nous comparons cette approche avec une régression directe par GMR entre données du locuteur source et paramètres de contrôle de la tête parlante. Nous montrons que l'approche par C-GMR réalise le meilleur compromis entre quantité de données d'adaptation d'une part, et qualité de la prédiction d'autre part. Enfin, nous évaluons la capacité de généralisation de l'approche C-GMR, et montrons que l'information *a priori* sur le locuteur de référence, exploitée par ce modèle, permet de généraliser à des configurations articulatoires du locuteur source non vues pendant la phase d'adaptation.

Enfin, nous présentons les premiers résultats d'une application clinique de l'échographie augmentée à une population de patients ayant subi une ablation du plancher de la bouche ou d'une partie de la langue. A l'aide de bilans orthophoniques classiques pratiqués entre chaque série de séances, nous évaluons l'usage du retour visuel en temps réel de la langue du patient et l'usage de séquences enregistrées préalablement sur un orthophoniste pour illustrer les articulations cibles. Les premiers résultats montrent une amélioration des performances des patients, notamment sur le placement de la langue.

Abstract

In the framework of speech therapy for articulatory disorders associated with tongue misplacement, providing a visual feedback might be very useful for both the therapist and the patient, as the tongue is not a naturally visible articulator. In the last years, ultrasound imaging has been successfully applied to speech therapy, as reported in several case studies. The assumption that *visual articulatory biofeedback* may facilitate the rehabilitation of the patient is supported on the links between speech production and perception. During speech therapy sessions, the patient seems to better grasp his/her tongue movements, despite the poor quality of the image due to inherent speckle noise and the lack of information about other speech articulators and structures (palate, pharyngeal wall, etc.). We develop in this thesis the concept of *augmented lingual ultrasound*. We propose two approaches to improve the raw ultrasound image, and describe a first clinical application.

The first approach focuses on automatic tongue tracking in ultrasound images. We propose a method based on supervised machine learning, where we model the relationship between the intensity of all the pixels of the image and the contour coordinates. The size of the images and of the contours is reduced using a principal component analysis, and a neural network models their relationship. We developed speaker-dependent and speaker-independent implementations and evaluated their performances as a function of the amount of manually annotated contours used as training data. We obtained an error of 1.29 mm for the speaker-dependent model with only 80 annotated images, which is better than a state-of-the-art baseline based on active contours.

The second approach intends to automatically animate an articulatory talking head from the ultrasound images. This talking head is an avatar of a so-called reference speaker. It reveals the external and internal structures of the vocal tract (palate, tongue, pharynx, teeth, etc.). First, we build a mapping model between ultrasound images and tongue control parameters acquired on the reference speaker. We then adapt this model to new speakers referred to as *source* speakers. This adaptation is performed using the Cascaded Gaussian Mixture Regression (C-GMR) technique, recently proposed in the context of cross-speaker acoustic-to-articulatory mapping. It is based on a joint model of the ultrasound data of the reference speaker, control parameters of the talking head, and adaptation ultrasound data of the source speaker. This approach is compared to a direct GMR regression between the source speaker data and the control parameters of the

talking head. We show that C-GMR approach achieves the best compromise between the amount of adaptation data and the prediction quality. We also evaluate the generalization capability of the C-GMR approach and show that prior information of the reference speaker helps the model generalize to articulatory configurations unseen during the adaptation phase.

Finally, we present preliminary results of a clinical application of augmented ultrasound imaging to a population of patients after partial glossectomy. We evaluate the use of visual feedback of the patient's tongue in real time and the use of sequences recorded with a speech therapist to illustrate the targeted articulation. Classical speech therapy probes are performed after each series of sessions. First results based on five patients show an improvement of the intelligibility, especially for phonemes involving accurate tongue placement.

Remerciements

Une thèse n'est pas une aventure dans laquelle on peut s'engager seule, et je tiens donc à remercier en premier lieu mes quatre encadrants pour leur présence et pour la confiance qu'ils m'ont accordée en me choisissant pour travailler sur ce projet. Pierre Badin a su me transmettre un peu de sa rigueur afin que je vérifie dix fois, cent fois mes scripts, m'apprenant ainsi à toujours remettre en question un résultat, toujours aller un peu plus loin dans son interprétation. Très disponible, sa porte est restée ouverte à chacune de mes nombreuses sollicitations. Thomas Hueber, moteur de ce projet, n'a jamais failli dans son encadrement. Dynamique et motivant, plein d'idées pour des améliorations, de nouvelles pistes, des lectures d'articles, il a su être très patient pour que je m'approprie peu à peu toutes les notions clés du sujet. Sur la partie orthophonie, j'ai eu la chance d'être encadrée par Mélanie Canault et de bénéficier de ses connaissances pour m'initier à ce vaste domaine. Ses conseils toujours pertinents de lecture et nos nombreux échanges m'ont permis de ne jamais m'égarer dans l'exploration de ce sujet. Nathalie Bedoin enfin a été présente à des moments plus ponctuels de cette thèse, mais d'une importance tout aussi grande, à chaque fois que nous devions travailler en collaboration entre les deux laboratoires, en apportant son expertise clinique, mais aussi à chaque phase de rédaction.

Aux rapporteurs et aux examinateurs qui ont accepté de faire partie du jury, j'adresse ces prochains mots. Merci à Olov Engwall et Slim Ouni pour leurs rapports complets, détaillés et précis qui m'ont permis de mieux préparer la soutenance orale. De même, je remercie Thierry Artières, examinateur, et Eric Truy, président du jury, qui ont apporté lors de leurs questions un regard expert sur des aspects que j'abordais avec des connaissances plus limitées.

Cette thèse a été financée par une Allocation Doctorale de Recherche de la Région Rhône-Alpes au sein de l'ARC6, Communauté de Recherche Académique « Technologies de l'Information et de la Communication et Usages Informatiques Innovants » dans le cadre du projet « *Voir sa langue pour mieux apprendre à parler ? Développement et évaluation de dispositifs de retour visuel articulatoire pour la rééducation orthophonique des troubles de la parole* ». Merci donc à la Région Rhône-Alpes d'avoir permis à ce projet d'exister.

Ce travail a été au cœur de nombreuses collaborations que je me dois de mentionner ici. Tout d'abord, merci à Philippe Revy qui a apporté son soutien lors de la constitution du projet de thèse en validant son intérêt. Je tiens à remercier chaleureusement Audrey Acher avec qui j'ai pu découvrir le milieu orthophonique en participant avec elle à une étude sur l'aphasie post-AVC. Mes remerciements vont ensuite à toutes les personnes impliquées dans le projet

ReviSon, soumis à un CPP avec l'aide bienvenue de Catherine Cereser des HCL. Merci à Eric Truy d'avoir accepté d'être promoteur de ce projet. Au Centre Médical Rocheplane, Marion Girod-Roux, qui a beaucoup contribué à l'élaboration du protocole de rééducation au cours de son stage au Gipsa-Lab, et Bérangère Gal nous ont permis de mener l'étude sur les cinq patients présentés dans ce travail. À l'institut d'Education Motrice d'Eybens, Claire Favier a été un contact privilégié pour nous permettre d'adapter le protocole aux enfants de l'institut. Dans un autre domaine, je remercie Lia Saki Bucar Shigemori de **l'**Institute of Phonetics de Munich, qui m'a généreusement fourni une grande quantité de données pour l'évaluation d'une des méthodes développées. Enfin, merci à Gina Yang avec qui j'ai travaillé à la fin de ma première année de thèse durant son stage.

En remontant un peu dans le temps, je souhaite remercier ceux qui ont encadré mes stages, à savoir Olivier Adam, puis Lori Lamel et tout Vocapia, et enfin Chantal Muller et David Rousseau. Avec leur confiance et leur expertise, j'ai pu explorer divers domaines, en musique, en parole et en images, qui m'ont donné des bases utiles pour cette thèse.

Venons-en maintenant tous ceux qui ont été importants au quotidien dans cette aventure. Merci d'abord à tous les chercheurs, ingénieurs, doctorants, post-docs, RH et service financier du Gipsa pour la bonne humeur et l'ambiance à l'intérieur et à l'extérieur du labo, les pauses et les (trop ?) nombreux goûters. Spéciale dédicace à mes cobureaux successifs, compagnons de galère le jour et souvent la nuit, sans qui il aurait été difficile de réaliser ce travail. Avec beaucoup de reconnaissance, j'adresse un sourire à tous ceux qui ont rendu ma route plus animée par leurs chants, leurs danses, leurs (fous) rires, nos histoires mais aussi nos peines, nos petits et nos grands changements de vie. A ces GGS, ces Timbawo, aux Zik'ets et bien sûr à la Ferme, aux Lionnes boiteuses, aux volontaires ICT Lyon de l'Euro 2016, au Club d'impro du G2Elab, aux merveilleux colocs des Taillées, à tous ceux pour lesquels je n'ai ni acronyme ni nom de code... À chacun d'entre vous j'adresse déjà souvent individuellement mes plus sincères sentiments, et à tous, je tiens à dire que ces trois ans auraient été tellement plus difficiles sans vous. Enfin, je veux m'adresser à ceux qui sont devenus thésards avec moi, répartis sur toute la France et le monde : nos nombreux échanges nous ont permis de nous rassurer régulièrement sur notre condition mentale au cours de cette aventure, sur nos capacités à finir… On y est arrivés !

Et enfin, la famille. Je remercie mes parents pour l'endurance dont ils ont fait preuve au quotidien tout au long de ma vie parfois un peu chaotique. Leur soutien pendant la fin de thèse était d'une importance capitale, les petits plats cuisinés pendant presque un mois de rédaction à domicile d'autant plus réconfortants. Mais en retour, quelle ardeur ont-ils mis à me rappeler à l'ordre lorsque je vagabondais un peu trop longtemps ! A mes sœurs, thésardes avant et après moi... Merci à l'une d'avoir rendu la route facile en la traçant de ses propres mains, et courage à l'autre, la fin est proche. Regarde, même moi, j'ai réussi.

Table des matières

Résumé 3
Abstract 7
Table des matières 9
Table des figures 15
Table des tableaux 19
Acronymes 21
Introduction 23
Chapitre 1. Visualiser les articulateurs de la parole : état de l'art 29
1.1. La langue dans la production de la parole 29
1.1.1. L'appareil phonatoire pour la production de la parole 29
1.1.2. Les troubles de l'articulation 31
1.2. L'illustration de l'articulation 33
1.2.1. Représenter l'articulation à partir de modèles 33
1.2.2. Représenter l'articulation à partir de données réelles 34
1.3. Le retour articulatoire visuel 37
1.3.1. L'électropalatographie 37
1.3.2. L'articulographie électromagnétique 38
1.3.3. L'échographie de la langue 40
1.3.4. L'inversion acoustico-articulatoire 41
1.4. Les relations entre production et perception : représentations internes de la parole 43
1.5. Conclusions de l'état de l'art 45
Chapitre 2. L'échographie linguale augmentée : suivi de la langue 47
2.1. Etat de l'art 48
2.1.1. Modèle de contours actifs 48
2.1.2. Modèles externes de la langue 49
2.1.3. Approche par réseau de neurones artificiels 50
2.2. Principe général de la méthode proposée 51
2.2.1. Approche EigenTongues pour le paramétrage des images ultrasonores 51
2.2.2. Approche EigenContours pour le paramétrage des contours de la langue 52
2.2.3. Modélisation de la relation EigenTongues-EigenContours par réseau de neurones artificiels 54
2.3. Protocole expérimental 57

2.3.1. Corpus de données ... 57
2.3.2. Scénarios mono- et multi-locuteur proposés ... 61
2.3.3. Choix du nombre de composantes pour le paramétrage des images et des contours ... 63
2.3.4. Apprentissage du réseau de neurones ... 65
2.3.5. Métrique de comparaison de contours ... 65
2.3.6. Choix des corpus d'apprentissage et de test ... 66
2.4. Résultats expérimentaux ... 67
2.4.1. Performances des modèles sur le Corpus MultiLoc ... 67
2.4.2. Comparaison avec la méthode état de l'art EdgeTrak ... 73
2.5. Conclusions et perspectives ... 75
Chapitre 3. L'échographie linguale augmentée : animation d'un modèle de langue ... 79
3.1. Etat de l'art ... 79
3.2. Méthode proposée ... 80
3.2.1. Principe général ... 80
3.3. Méthodologie ... 84
3.3.1. Gaussian Mixture Model (GMM) et Gaussian Mixture Regression (GMR) ... 84
3.3.2. Approche directe D-GMR ... 85
3.3.3. Approche par adaptation d'un modèle de conversion existant par C-GMR ... 86
3.4. Dispositif expérimental et évaluation des méthodes ... 92
3.4.1. Acquisition des données ... 92
3.4.2. Paramétrisation du signal audio et des images échographiques ... 92
3.4.3. Alignement des données ... 93
3.4.4. Choix du nombre de composantes des modèles D-GMR et C-GMR ... 95
3.4.5. Corpus d'apprentissage, de validation et de test ... 95
3.4.6. Métrique de comparaison des paramètres EMA ... 95
3.5. Résultats et interprétations ... 96
3.5.1. Performance du modèle de référence X-Y ... 96
3.5.2. Performance des approches D-GMR et C-GMR ... 97
3.5.3. Exemples illustratifs ... 107
3.6. Animation du modèle de lèvres et de la mâchoire de la tête parlante ... 107
3.7. Discussion et perspectives ... 110
Chapitre 4. Application clinique du retour visuel par échographie : études de cas sur des patients glossectomisés ... 113
4.1. Rééducation orthophonique par échographie : état de l'art ... 113
4.1.1. Troubles articulatoires isolés ... 114
4.1.2. Troubles de la parole : cas des personnes malentendantes de naissance ... 115
4.1.3. Résumé de la littérature ... 116

4.2. Rééducation orthophonique de patients glossectomisés 121
4.2.1. Définition de la glossectomie 121
4.2.2. Littérature sur la rééducation de patients glossectomisés 121
4.3. Protocoles de rééducation proposés 123
4.3.1. Dispositifs d'aide à la rééducation : sonde échographique et Ultraspeech-player 123
4.3.2. Protocoles de rééducation 125
4.3.3. Evaluation des progrès par des bilans orthophoniques 125
4.3.4. Déroulement de l'étude clinique 126
4.4. Résultats sur cinq patients ayant subi une chirurgie bucco-pharyngienne 127
4.4.1. Population de l'étude 127
4.4.2. Analyse descriptive des bilans 129
4.4.3. Discussion 139
4.5. Perspectives 140
Chapitre 5. Conclusion générale et perspectives 143
Contributions 147
Bibliographie 149
Annexe A : corpus Ultraspeech-player 159
Annexe B : bilans orthophoniques 161

Table des figures

Figure 1. Exemples d'images échographiques (a) surface de la langue bien visible : phonème /a/ (b) surface de la langue peu visible : phonème /i/ 25
Figure 2. Appareil phonatoire 30
Figure 3. La marionnette bavarde de Hoptoys 32
Figure 4. Illustrations : à gauche, Diadolab ; à droite, Canault 34
Figure 5. Ultraspeech-player 35
Figure 6. Exemples de représentations de têtes parlantes articulatoires. De gauche à droite : Fagel & Madany (2008) Bälter *et al.* (2005) et Badin *et al.* (2010) 37
Figure 7. Exemple de palais artificiel pour l'EPG (gauche) et de visualisation des contacts palato-linguaux (droite) 38
Figure 8. Dispositif EMA complet Carstens AG200 (gauche) et placement des bobines sur la langue (droite) 39
Figure 9. (a) exemple de positionnement d'une sonde échographique. (b) images ultrasonores de la langue dans le plan médio-sagittal (position de repos à gauche et lors d'un [k] à droite, extrait de Hueber (2009) 40
Figure 10. Extraction des vecteurs Eigentongues (b) à partir d'une base d'apprentissage redimensionnée à 64x64 pixels (a) et projection d'une image sur les K premiers vecteurs EigenTongues (c). 52
Figure 11. Grille semi-polaire placée sur les images échographiques pour l'annotation manuelle du contour. 53
Figure 12. Processus d'annotation des images échographiques avec P points de contour par image. 54
Figure 13. Fonctionnement général d'un neurone formel 55
Figure 14. Illustration d'un perceptron multicouche à deux couches cachées. 55
Figure 15. Schéma général de notre système de suivi de langue. 57
Figure 16 - Image du /a/ pour les 9 locuteurs avant et après recalage (locuteur de référence encadré en rouge). 61
Figure 17. Evolution de la variance en fonction du nombre de vecteurs propres Eigentongues. En rouge, les variations pour chaque locuteur de la base MultiLoc. En bleu, les variations pour un corpus incluant l'ensemble de la base MultiLoc (sans recalage en trait pointillé, avec recalage en trait continu). 63

Figure 18. Evolution de la variance en fonction du nombre de vecteurs propres EigenContours. En rouge, les variations pour chaque locuteur de la base MultiLoc. En bleu, les variations pour un corpus incluant l'ensemble de la base MultiLoc (sans recalage en trait pointillé, avec recalage en trait continu). 64
Figure 19. En rouge, le contour réel. En bleu, le contour estimé. Les flèches continues représentent la distance point à point, les flèches en pointillé la distance point à contour. 66
Figure 20. Moyenne des MSD en fonction de la quantité d'apprentissage pour les différents scénarios, avec leurs intervalles de confiance. En abscisse, le nombre d'images annotées du locuteur courant. En ordonnée, la RMSE de la MSD en mm. Les barres d'erreur représentent les intervalles de confiance (t-test avec $\alpha = 0{,}05$) 68
Figure 21. Erreur d'estimation des P points (ici P = 16) pour les différentes configurations du système multi-locuteur. Les points sont ordonnés de l'arrière à l'avant de la langue. 70
Figure 22. Exemples de contours estimés pour différentes configurations de la langue, différents locuteurs et différents scénarios 72
Figure 23. Résultats du suivi de contour par notre méthode et par EdgeTrak (noté eT), pour les locuteurs SK4, SK5, SK7, et SK8 de la base de données « Slovaque » et pour le locuteur MG. 74
Figure 24. Différence entre l'estimation de notre méthode et celle d'EdgeTrak pour les quatre phrases du locuteur MG. Plus la différence est importante et plus le contour fourni par notre méthode peut être jugé plus précis que celui fourni par EdgeTrak. Les barres verticales séparent les phrases. 75
Figure 25. Tête parlante articulatoire construite à partir des données IRM du locuteur de référence, développée au Gipsa-lab Badin *et al.* (2008) (a) Locuteur de référence et marqueurs utilisés pour les mouvements du visage (billes collées) ; (b) images IRM avec segmentation manuelle des articulateurs pour 46 différentes articulations tenues ; (c) tête parlante articulatoire 3D ; (d) mode de visualisation choisi pour le système de retour visuel proposé. 81
Figure 26. Bobines EMA sur la langue du locuteur de référence numérotées de 1 à 3 de l'arrière à la pointe de la langue (gauche), et visualisation sur la tête parlante articulatoire des paramètres de contrôle EMA associés (à droite). 82
Figure 27. Animation automatique du modèle de langue d'une tête parlante articulatoire à partir d'images échographiques 83
Figure 28. Représentation graphique du Direct-GMR (D-GMR), Split Cascaded GMR (SC-GMR) et Integrated Cascaded GMR (IC-GMR). Z et X sont respectivement les données échographiques des locuteurs source et référence. Y correspond aux

paramètres de contrôle EMA. m et k indicent les composantes des modèles de mélange. 89
Figure 29. Représentation schématique des variables X Y et Z utilisées pour les C-GMR. En pointillé, les données manquantes 91
Figure 30. Alignement par DTW d'une séquence d'images échographiques (paramétrées par l'approche EigenTongues) avec une séquence de paramètres EMA pour le locuteur de référence, en exploitant les signaux audio s_x et s_y associés respectivement aux données échographiques et aux données EMA. 93
Figure 31. Performance du mapping -vers- en RMSE (mm), avec un intervalle de confiance à 95 %, en fonction de la quantité de données d'enrôlement, pour le locuteur source F1 (en haut) et M1 (en bas) 99
Figure 32. Boxplots des RMSE obtenues pour la *baseline* et en généralisation sur le corpus des 11 phonèmes de test, pour M1 (en haut) et F1 (en bas), pour les modèles D-GMR, SC-GMR et IC-GMR 103
Figure 33. Moyenne des RMSE pour la généralisation sur les consonnes des trois modèles IC-GMR, SC-GMR, D-GMR, pour le locuteur M1 (en haut) et F1 (en bas) 105
Figure 34. Moyenne des RMSE pour la généralisation sur les voyelles des trois modèles IC-GMR, SC-GMR, D-GMR, pour le locuteur M1 (en haut) et F1 (en bas). 106
Figure 35. Exemples illustratifs d'animation du modèle de langue à partir d'images échographiques (après rotation). 107
Figure 36. Animation automatique du modèle de lèvres et de mâchoire d'une tête parlante articulatoire à partir de vidéos des lèvres. 108
Figure 37. Représentation des deux modes de visualisation disponibles (a) l'illustration proposée par Ultraspeech-player avec l'échographie de la langue de l'orthophoniste (b) un des contours proposés par Ultraspeech-biofeedback : la langue du patient est placée dans la zone transparente et normalisée par des transformations pour s'adapter à la forme du conduit vocal. Le même contour de palais est affiché dans les deux protocoles. 124
Figure 38. Scores obtenus pour trois bilans de T0 à T2 par IR001, RI002 et RI003 (IR = ILLUSTRATION puis RETOUR ; RI = RETOUR puis ILLUSTRATION). 132
Figure 39. Score moyen par bilan et par patient au MBLF pour la langue. Les intervalles de confiance sont indiqués pour chaque valeur (α=0,05). 133
Figure 40. Résultats pour le TPI par patient et par bilan. En ordonnée, le nombre d'erreurs pour chaque catégorie. 136

Table des tableaux

Tableau 1. Lieux d'articulation des consonnes françaises..... 31
Tableau 2. Informations sur la composition du corpus MultiLoc. Quantité totale et moyenne (par locuteur et par phrase) de phonèmes et d'images, et durée totale et moyenne des acquisitions (en secondes) pour l'ensemble des deux répétitions. 59
Tableau 3. Tableau récapitulatif des performances obtenues par les différents scénarios ($RMSE_{MSD}$ en mm). 67
Tableau 4. $RMSE_{MSD}$ (en mm) pour les quatre locuteurs du corpus Slovaque et le locuteur MG. 74
Tableau 5. Performance du GMR -vers-(mapping du locuteur de référence) : RMSE et intervalle de confiance à 95% (CI), pour chaque paramètre de contrôle EMA (tip, mid, back) 97
Tableau 6. Performances du modèle de référence en RMSE moyenne et intervalle de confiance (à 95%) dans trois situations : colonne de gauche, en haut, modèle construit entre les échographiques de la langue et les paramètres de contrôle de la langue ; colonne de gauche, en bas, modèle construit entre les images des lèvres et les paramètres de contrôle de la mâchoire et des lèvres ; colonne de droite, modèle conjoint construit entre les données jointes échographies + lèvres et l'ensemble des paramètres de contrôle de la tête parlante. TT = Tongue Tip ; TD = Tongue Dorsum ; TB = Tongue Back ; J = Jaw ; UL = Upper Lip ; LL = Lower Lip..... 109
Tableau 7. Etat de l'art sur la rééducation des troubles de la parole par échographie. N représente la taille de l'échantillon..... 118
Tableau 8. Déroulement des séances de rééducation. Le premier bilan est réalisé à T0 et marque le début de la prise en charge..... 127
Tableau 9. Critères d'inclusion et d'exclusion des patients. 128
Tableau 10. Informations générales sur les patients. Les scores indiqués sont obtenus à T0 avant le début de la rééducation. I = ILLUSTRATION ; R = RETOUR VISUEL. 129
Tableau 11. Score total par bilan et par patient. 130
Tableau 12. Score au test d'intelligibilité pour les cinq patients. 134
Tableau 13. Contenu des catégories d'erreurs A à M pour le bilan TPI. 134
Tableau 14. Erreurs réalisées au TPI par les patients à chaque bilan sur le lieu d'articulation et l'élévation de la langue 138

Acronymes

ACP	Analyse en Composantes Principales
GMM	Gaussian Mixture Model
D-GMR	Direct Gaussian Mixture Regression
SC-GMR	Split Cascaded Gaussian Mixture Regression
IC-GMR	Integrated Cascaded Gaussian Mixture Regression
ANN	Artificial Neural Network
RBM	Restricted Boltzmann machine
TPI	Test Phonétique d'Intelligibilité
MBLF	Motricité Bucco-Linguo-Faciale
BECD	Batterie d'Evaluation Clinique de la Dysarthrie
EMA	Articulographie Électromagnétique
EPG	Electropalatographie
MLP	Multi-Layer Perceptron
PCC	Percentage of Consonants Correct
PTCC	Percent Target Consonant Correct
PVM	Place-Voice-Manner
DEAP	Diagnostic Evaluation of Articulation and Phonology
PRC	Percent Rhotics Correct
AsIDS	Assessment of Intelligibility of Dysarthric Speech
CAS	Childhood Apraxia of Speech
KP	Knowledge of Performance
KR	Knowledge of Results

Introduction

Un son, puis plusieurs, et enfin, un premier mot. De la naissance à l'âge de cinq ans environ, un enfant acquiert tous les phonèmes qui lui donneront accès au langage oral. Ainsi, à l'âge de cinq ans, le langage de l'enfant est déjà bien structuré, mais de nombreux troubles peuvent entraver le développement de cette faculté. Plus tard, l'adulte peut lui aussi être touché par l'émergence de troubles impactant la sphère orale. Les origines et les manifestations de ces troubles sont diverses : anatomiques, neurologiques, cognitives, motrices. Dans tous les cas, leur prise en charge est effectuée par un orthophoniste à la suite de bilans précis.

Dans le cadre de cette thèse, nous nous sommes orientée plus particulièrement vers deux troubles : les troubles articulatoires et les troubles phonologiques. Un trouble articulatoire est lié à l'incapacité de prononcer un ou plusieurs phonèmes de sa langue maternelle, indépendamment du contexte. Le trouble phonologique consiste quant à lui en la substitution systématique d'un phonème par un autre, mais dans un contexte donné. Ces troubles sont souvent liés à un mauvais positionnement de la langue, un des articulateurs majeurs de la parole. Or, durant les séances de rééducation, l'orthophoniste n'a qu'assez peu d'informations sur la position de la langue, qui reste la plupart du temps cachée dans la cavité buccale. Il la déduit de son interprétation de la production orale réalisée par son patient. Il l'amène alors à corriger la position et l'articulation de sa langue en fonction de ce retour. Le patient possède quant à lui trois informations : une information tactile (contact de la langue avec les autres articulateurs), une information auditive (résultat de l'articulation) et une information proprioceptive (capacité du cerveau à connaître la position de tout élément du corps même lorsqu'il nous est invisible). Ajouter une information visuelle sur cette articulation permettrait donc au patient comme à l'orthophoniste de comprendre plus précisément la nature de l'erreur et de la corriger.
Ce paradigme s'inscrit donc dans le cadre général du biofeedback, ou *retour visuel*, développé dans différentes applications médicales. Ainsi, l'utilisation d'un électrocardiogramme informe un patient sur le comportement de son cœur. Il peut aussi être utilisé pour renseigner un sportif sur ses performances physiques, en transformant le signal en informations visuelles claires et exploitables par l'utilisateur. Nous nous intéressons dans le cadre de cette thèse au retour visuel pour la rééducation des troubles de la parole et de l'articulation.
Au cours des dernières années, de nombreuses études de perception de la parole ont appuyé l'hypothèse de l'intérêt du retour visuel pour la rééducation orthophonique. De

plus, certaines études semblent révéler une capacité chez certains sujets non spécialistes, c'est-à-dire sans entrainement particulier en phonétique, d'exploiter dans des conditions défavorables, comme le bruit, une information visuelle sur le mouvement de la langue pour identifier certains phonèmes.

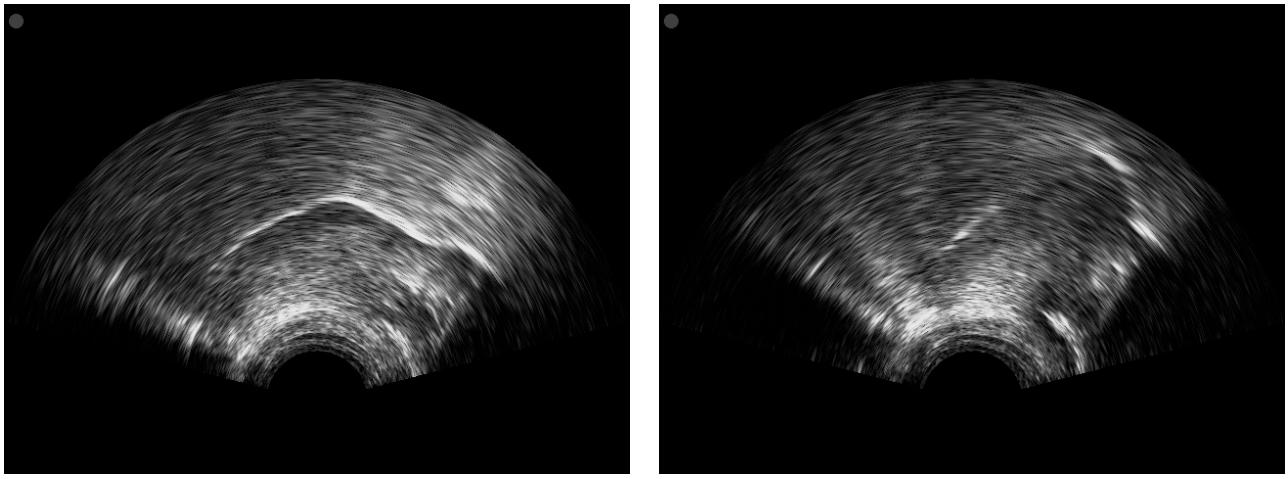
De ce fait, il semble pertinent de proposer un mode de visualisation de la langue dans le cadre d'une séance de rééducation orthophonique. Dans cette thèse, nous nous intéresserons à deux types de représentations, que nous appellerons (1) l'illustration visuelle linguale, et (2), le retour visuel lingual. Par souci de concision, nous utiliserons dans la suite *illustration* et *retour visuel*.
Dans le cadre de l'*illustration*, nous cherchons à montrer au patient un geste cible qui est représenté dans un espace articulatoire dit de référence (*i.e.* non-pathologique), différent de celui du patient. Nous verrons plus en détail dans le Chapitre 1 qu'il peut s'agir d'une schématisation d'un conduit vocal générique, ne correspondant donc à aucun locuteur spécifique, de données d'imagerie médicale acquises sur un locuteur dit de *référence*, ou bien d'une tête parlante articulatoire, c'est à dire un clone logiciel de la sphère orofaciale de ce locuteur de référence. Ce mode de visualisation a pour objectif de fournir au patient un cadre attractif (et donc motivant) pour mieux comprendre les notions de base de l'articulation, ce qui lui permettra notamment de mieux interpréter les consignes du praticien.

Dans le cadre du *retour visuel articulatoire*, nous cherchons à faire visualiser au patient sa propre articulation. Cette visualisation s'effectuera idéalement en temps-réel, c'est-à-dire que l'information visuelle accompagnera le geste articulatoire avec un délai constant et très faible. Le retour visuel vient ici se combiner aux retours auditif, tactile et proprioceptif. Dans le cadre de la rééducation orthophonique, nous faisons l'hypothèse que ce nouveau canal d'information favorisera une calibration plus efficace des relations sensori-motrices (modèles internes) liées à la production et à la perception de la parole. Dans cette thèse, nous testerons cette hypothèse dans le cadre de la prise en charge post-opératoire de patients ayant subi une ablation partielle de la langue ou du plancher de la bouche au cours du traitement d'un cancer.

Plusieurs technologies, détaillées dans le Chapitre 1, peuvent être envisagées pour capturer le mouvement de la langue pendant la production de la parole. Nous pouvons mentionner en particulier l'électropalatographie (EPG), qui consiste à mesurer les points de contact de la langue avec le palais à l'aide d'une matrice d'électrodes disposées sur un palais artificiel. Cette technique a fait l'objet de nombreuses études cliniques et a prouvé son efficacité pour certaines pathologies. Dans le cadre de cette thèse, nous porterons notre intérêt sur

l'échographie (ou imagerie ultrasonore). Placée sous la mâchoire d'un locuteur, une sonde échographique permet de visualiser en temps-réel les mouvements de la langue dans les plans médio-sagittal et coronal. Cette technique d'acquisition est simple d'utilisation, ne nécessite pas d'équipement encombrant, est très peu invasive et est jugée sans risque.
Au cours des dix dernières années, plusieurs travaux ont été réalisés afin de valider l'utilisation du retour visuel articulatoire (notamment par échographie) dans le cadre de la rééducation orthophonique. Ces études portent sur différentes pathologies et se présentent principalement sous la forme d'études de cas, menées pour la plupart dans les pays anglo-saxons. Elles semblent montrer un impact positif sur les progrès réalisés par le patient. Une validation à plus grande échelle (sur une population de patients) reste à ce jour à mener. Il s'agit d'un des objectifs de cette thèse dans laquelle nous proposons un protocole complet ainsi que des premiers résultats obtenus pour cinq patients.

Si l'utilisation de l'échographie dans le cadre d'une rééducation orthophonique par retour visuel semble prometteuse, elle pose néanmoins plusieurs problèmes. Dans ce travail de thèse, nous considérons que le premier d'entre eux est la lisibilité de l'image échographique par le patient. En effet, l'image échographique est très bruitée en raison de la présence du bruit caractéristique dit de *speckle*, visible en Figure 1.



(a) (b)

Figure 1. Exemples d'images échographiques (a) surface de la langue bien visible : phonème /a/ (b) surface de la langue peu visible : phonème /i/

Par ailleurs, elle ne permet qu'une représentation partielle de la surface supérieure de la langue (information sur l'apex et le dos de la langue parfois manquante), et ne fournit aucune information sur la localisation du palais, des dents ou d'autres éléments de la cavité buccale. Cette mauvaise qualité relative couplée à l'absence de repères la rend souvent difficile à interpréter pour un non-expert**.** Afin de pallier ces manques, nous

proposons de développer dans cette thèse le concept d'*échographie linguale augmentée*, par analogie à la réalité augmentée. Il s'agit d'ajouter en temps-réel à une échographie linguale, certains éléments initialement non-visibles par cette technique, afin d'en améliorer la compréhension. Ce travail s'articule donc autour de deux objectifs :
1) Développer différents systèmes d'illustration et de retour visuel par échographie linguale augmentée.
2) Évaluer ces systèmes dans le cadre d'une étude clinique sur la prise en charge d'un trouble de l'articulation.

Dans le cadre de l'échographie linguale augmentée, deux pistes ont été étudiées, qui vont être détaillées dans les chapitres qui suivent. Dans le Chapitre 1 de cette thèse, nous dresserons tout d'abord un état de l'art des techniques permettant de visualiser les articulateurs de la parole. Nous confronterons l'utilisation de l'illustration à celle d'un retour visuel et justifierons le choix de l'échographie pour la rééducation orthophonique.

Dans le chapitre 2, nous présenterons une première méthode d'échographie augmentée, s'intéressant à la problématique du suivi automatique du contour de langue. En dépit de l'existence de nombreux travaux, il s'agit d'un problème encore non résolu, et aucune technique ne semble adaptée à une utilisation pratique (*i.e.* robuste, temps-réel, multi-locuteur, ne nécessitant qu'une phase limitée de calibration) notamment dans le cadre d'une rééducation orthophonique. Nous proposons ici une nouvelle méthode qui cherche à atteindre cet objectif. Cette dernière est basée sur la modélisation des relations entre forme et texture à l'aide de réseaux de neurones par apprentissage supervisé[1].

Dans le chapitre 3, nous présenterons le second système développé, où nous chercherons à faire apparaître des structures manquantes dans l'image échographique (palais, dents, pharynx, etc.). Pour cela, nous proposons d'animer automatiquement et en temps réel, la tête parlante articulatoire développée au GIPSA-lab Badin, Elisei *et al.* (2008). Nous présentons ainsi plusieurs techniques basées sur l'apprentissage artificiel (ou *machine learning*) supervisé, et notamment la modélisation par mélange de gaussiennes (*GMM*). Toujours dans l'optique de développer des systèmes transférables à une utilisation clinique, nous avons cherché à maintenir une performance acceptable tout en minimisant la quantité de données d'apprentissage, et en permettant au système de traiter une parole pathologique, présentant un inventaire phonétique incomplet. Notre système s'appuie notamment sur la technique *Cascaded-Gaussian Mixture Regression*, récemment proposée par Hueber, Girin *et al.* (2015).

[1] Cette méthode a fait l'objet d'une publication à la conférence Interspeech 2015

Outre le développement de technologies, nous avons souhaité évaluer l'impact du retour visuel par échographie sur des populations rarement étudiées. En France, où cette technologie émerge tout juste, le spectre des possibilités est large. Nous avons donc mis en place une étude permettant d'évaluer l'impact du retour visuel sur des adultes ayant subi une chirurgie bucco-pharyngée. Le Chapitre 4 nous amènera donc à étudier une application clinique du retour visuel par échographie. Nous exposerons le protocole mis en place pour la rééducation orthophonique qui nous a permis de comparer l'utilisation de l'*illustration* et celle du *retour visuel*. Nous présenterons les premiers résultats sous forme d'études de cas.

Chapitre 1. Visualiser les articulateurs de la parole : état de l'art

Dans ce chapitre, nous détaillerons l'état de l'art de la visualisation des articulateurs de la parole, et en particulier de la langue, pour la rééducation orthophonique ou l'apprentissage des langues secondes. Nous commencerons par une partie sur la production de la parole et en particulier l'implication de la langue. Nous mentionnerons la rééducation orthophonique dans le cadre de parole pathologique. Nous détaillerons les techniques d'illustration et de retour visuel, et argumenterons notre choix de l'échographie comme retour visuel. Enfin, nous appuierons les études cliniques présentées en soulignant les relations entre la production de la parole et la perception, en particulier dans le cas de la langue, un articulateur normalement invisible.

1.1. La langue dans la production de la parole

Dans cette section, nous présentons le fonctionnement de l'appareil phonatoire pour la production de la parole. Nous détaillons l'implication de la langue dans les sons du français. Nous définissons ensuite les troubles de l'articulation et les conditions classiques de rééducation des orthophonistes pour traiter ces troubles.

1.1.1. L'appareil phonatoire pour la production de la parole

La production de la parole repose sur l'enchaînement de trois mécanismes : la respiration, la phonation, l'articulation. En premier lieu, pendant la respiration, les poumons insufflent à l'ensemble du système l'énergie nécessaire, transmise par l'air expulsé (pour la parole en français, d'autres langues ayant aussi des sons produits à partir d'air inspiré). Cet air arrive au niveau des plis vocaux, communément appelés les cordes vocales. Lorsqu'un son fait l'objet d'une vibration des plis vocaux, il est dit *voisé,* sinon, il est *non voisé*. Ce flux d'air vient ensuite résonner dans les cavités supraglottiques (cavités orales et cavité nasale) dont la géométrie, et donc les caractéristiques acoustiques, est modelée par la position des différents articulateurs de la parole. Dans le cadre de nos travaux sur le *retour visuel*, nous diviserons l'ensemble de ces articulateurs en deux catégories : les articulateurs visibles comme les lèvres et la mâchoire, ainsi que la pointe de la langue lorsque la bouche

est ouverte ; les articulateurs invisibles : la langue et du velum. L'ensemble des structures et organes impliqués dans la production de la parole constitue l'*appareil phonatoire*, présenté en Figure 2.

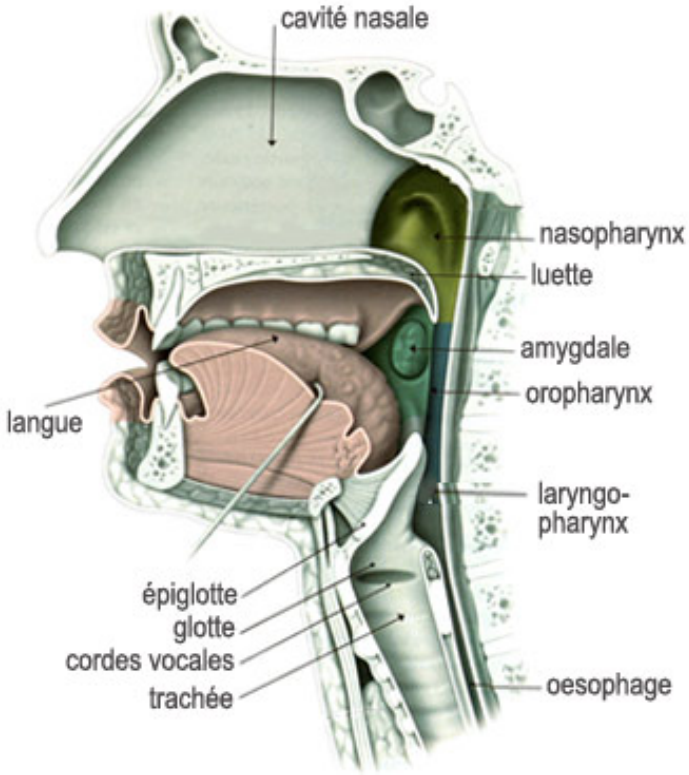


Figure 2. Appareil phonatoire[2]

Parmi les articulateurs de la parole, la langue joue un rôle de première importance. Sa position est déterminante pour une grande majorité des consonnes, comme le montre la zone encadrée en rouge dans le Tableau 1, et pour toutes les voyelles. La distinction de certains sons, comme le /t/ et le /k/, repose exclusivement sur cette position (ici dentale versus vélaire). Un trouble de l'articulation linguale est donc susceptible d'être à l'origine d'une altération forte de l'intelligibilité de la parole produite, et donc de la communication parlée en général.

[2] http://lecerveau.mcgill.ca/flash/capsules/outil_bleu21.html

Tableau 1. Lieux d'articulation des consonnes françaises

MODE D'ARTICULATION			MODE D'ARTICULATION										
			Bilabiale	*Labio-dentale*	*Dentale*	*Alvéolaire*	*Prépalatale*	*Palatale*	*Vélaire*	*Uvulaire*			
MODE D'ARTICULATION	Occlusive	Médiane	p		t				k		Non-voisé	Orale	MODE D'ARTICULATION
			b		d				g		Voisé		
			m		n			ɲ				Nasale	
	Constrictive			f		s	ʃ				Non-voisé	Orale	
				v		z	ʒ	j		ʁ	Voisé		
		Latérale			l								
		Médiane	ɥ, w					ɥ	w				

1.1.2. Les troubles de l'articulation

Selon la Fédération Nationale des Orthophonistes[3] : « *L'orthophonie s'attache aux dimensions plurielles du concept de langage, comme moyen d'expression, d'interaction et d'accès à la symbolisation dans toutes ses dimensions […] linguistiques […], cognitives […], psycho-affectives […], sociales […]. L'orthophonie s'intéresse également à toutes les altérations de la sphère oro-faciale sur les plans moteur, sensitif et physiologique* ». Aussi, les troubles de la parole et du langage sont multiples. Dans ce travail de thèse, nous nous intéressons aux altérations de la sphère oro-faciale sur le plan moteur, et, plus précisément, aux troubles de l'articulation associés à un mauvais placement de la langue. Nous distinguons pour ce faire ici deux types de troubles :

- les troubles articulatoires : le patient est dans l'incapacité de prononcer un ou plusieurs phonèmes de sa langue maternelle (exemple : sigmatisme)
- les troubles phonologiques : le patient est capable de prononcer tous les phonèmes de la langue, mais substitue systématiquement un phonème par un autre dans un contexte donné (exemple : /tʁ/ vs. /kʁ/)

[3] http://www.fno.fr/lorthophonie/lorthophonie-et-les-orthophonistes/quest-ce-que-lorthophonie-2/ : définition issue du référentiel d'activités du Certificat de Capacité en Orthophonie, publié au Bulletin Officiel N°32 du Ministère de l'Enseignement supérieur et de la Recherche du 5 septembre 2013

Ces troubles peuvent trouver leur source dans des origines diverses : il peut s'agir d'une malformation, d'un trouble neurologique, d'un changement structurel lié à une intervention chirurgicale, ou encore, surtout chez l'enfant, d'un retard de langage ou d'un mauvais apprentissage lors de l'acquisition du langage.
Lors des séances de rééducation de ces troubles, les orthophonistes cherchent à corriger la position de la langue du patient par l'usage de nombreux exercices. Dans la suite du manuscrit, lorsque nous parlerons de séances de rééducation orthophonique, nous considérerons seulement celles traitant de ces troubles.

L'orthophoniste dispose de diverses méthodes pour la prise en charge de ses patients. Certaines de ces méthodes visent à fournir une représentation visuelle de la *cible articulatoire* à atteindre, c'est-à-dire le mouvement à effectuer pour atteindre un but acoustique. Exemple : faire claquer une partie du corps de la langue contre le palais pour produire une occlusive comme /g/. Une telle représentation est particulièrement utile pour la langue dont un sujet naïf n'a qu'une conscience très partielle de la position. L'orthophoniste utilise des outils diversifiés pour fournir cette représentation visuelle. Il peut réaliser des exercices de praxies, en utilisant des supports variés comme l'utilisation d'un miroir, pour que le patient puisse voir ses lèvres et sa mâchoire, de dessins, ou de marionnettes comme en Figure 3.

Figure 3. La marionnette bavarde de Hoptoys

Au cours des dernières années, avec l'avènement de l'informatique, plusieurs logiciels permettant une *illustration visuelle* de l'articulation ont fait leur apparition. Ces derniers seront détaillés à la section 1.2.
Plus récemment, un nouveau paradigme de prise en charge des troubles de l'articulation a émergé. Il s'agit du *retour visuel* (ou plus communément *visual biofeedback*). Le retour articulatoire visuel permet à un locuteur de visualiser sa propre articulation au cours de la production. Il faut donc le distinguer du paradigme d'illustration visuelle mentionné

précédemment, qui tend à montrer au patient le geste cible et non pathologique d'une tierce personne, ou d'un locuteur virtuel, comme nous le verrons plus loin.

Dans le cadre de ce travail, l'information fournie au patient est une représentation anatomique des mouvements de sa langue, à la manière d'un miroir qui permettrait de visualiser les structures cachées de l'appareil vocal. Nous n'aborderons pas les approches basées sur la visualisation par le patient d'un sonagramme, tel que proposé par Byun & Hitchcock (2012), Shuster, Ruscello *et al.* (1992) et Shuster, Ruscello *et al.* (1995).

Nous dressons dans les sections qui suivent un état de l'art des technologies d'illustration et de retour visuel. Nous appuyons ces deux procédés par des études et discutons de leur importance dans la rééducation orthophonique.

1.2. L'illustration de l'articulation

Au détour de collaborations multiples entre ingénieurs, chercheurs et orthophonistes, plusieurs logiciels ont émergé pour la rééducation des troubles de la parole. Dans cette partie, nous parcourons la littérature afin de les répertorier. Certains s'appuient sur des représentations schématiques du conduit vocal, d'autres s'appuient sur des données articulatoires réelles, obtenues sur un ou plusieurs locuteurs.

1.2.1. Représenter l'articulation à partir de modèles

Le logiciel Diadolab, développé par Menin-Sicard & Sicard (2012), propose des animations de contours d'articulateurs d'enfant simplifiés et inspirés de données articulatoires (illustration en Figure 4, gauche). Une application de ce logiciel à la rééducation orthophonique est proposée dans le mémoire d'orthophonie de Bezard (2015). Cette étude de cas est menée sur deux enfants déficients auditifs de 9 et 13 ans, à travers une séance hebdomadaire, pendant dix semaines. Un moulage dentaire et des cartes en coupe sagittale représentant différentes consonnes françaises complètent l'utilisation de ce logiciel lors des séances. Cette étude, comme le soulignent les auteurs, doit être étendue à une population plus large afin de valider les premiers effets positifs observés lors de l'évaluation (par jury d'écoute et scores). Dans le même registre, Canault propose un outil en ligne intitulé « Le conduit vocal en action »[4] qui modélise les mouvements de l'ensemble des articulateurs impliqués dans la parole (voir une illustration à la Figure 4, droite). Il s'agit initialement

[4] Ce logiciel peut être testé sur la page http://anatomie3d.univ-lyon1.fr/webapp/website/website.html?id=3346735&pageId=223201

d'un outil pédagogique mis à disposition des étudiants en orthophonie au cours de leurs études.

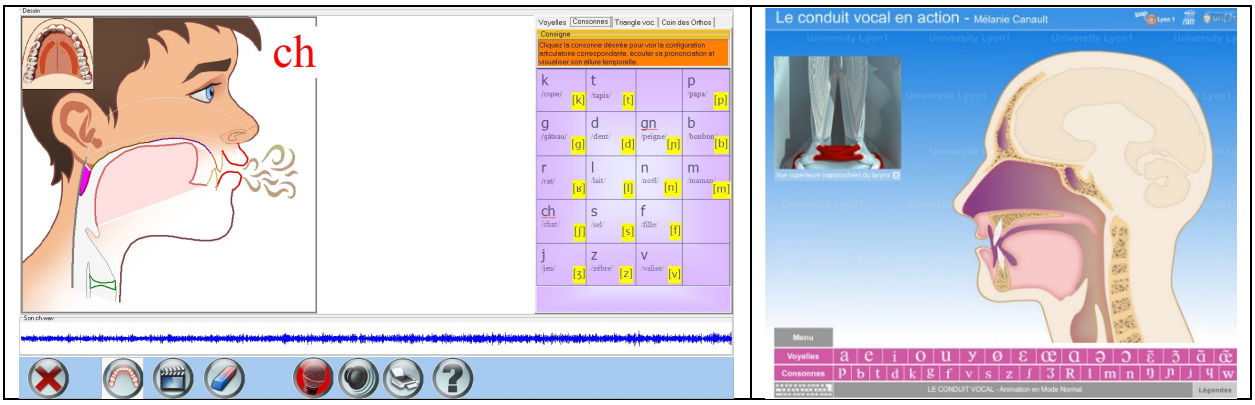


Figure 4. Illustrations : à gauche, Diadolab ; à droite, Canault

1.2.2. Représenter l'articulation à partir de données réelles

Une autre façon d'illustrer les mouvements articulatoires est d'utiliser des données réelles, enregistrées sur des locuteurs qualifiés dans la suite de ce manuscrit de *locuteurs de référence*.

Hueber (2013) présente le logiciel Ultraspeech-player illustré en Figure 5. Ce logiciel[5] s'appuie sur une grande base de films échographiques à haute vitesse de la langue, enregistrés de façon synchrone avec le signal de parole et une vidéo haute vitesse des lèvres, à l'aide du logiciel Ultraspeech Hueber, Chollet *et al.* (2008). L'acquisition d'images échographiques est détaillée dans la suite de cet état de l'art. Cette base couvre l'ensemble des voyelles et des consonnes du français, et présente également des logatomes de type VCV ou CVC, des clusters de consonnes, des mots et des phrases simples, des déglutitions, pour plusieurs locuteurs (dont une orthophoniste). Ultraspeech-player embarque un mécanisme de traitement en temps-réel des flux audio et vidéo, permettant à l'utilisateur de contrôler la vitesse du geste articulatoire présenté, et celle du signal sonore associé. L'objectif est de ralentir le geste articulatoire pour mieux l'observer. L'utilisation de données articulatoires acquises sur un véritable locuteur permet de restituer la dynamique réelle des mouvements, notamment au niveau des patrons de coarticulation. Ce logiciel a été mis en application lors d'une étude menée dans le cadre d'un mémoire d'orthophonie, et détaillée dans Fabre, Hueber *et al.* (2016) et Bach & Lambourion (2014). Cette étude porte sur la rééducation chez l'enfant d'un trouble phonologique avec substitution de [tʁ]

[5] Ce logiciel est téléchargeable sur la page www.ultraspeech.com

par [kʁ]. Les résultats de cette étude soulignent l'apport efficace de ce logiciel à la prise en charge orthophonique d'un trouble phonologique chez de jeunes enfants âgés de 5 à 7 ans.

Figure 5. Ultraspeech-player

Une autre façon de montrer des mouvements articulatoires cibles au patient est d'utiliser une *tête parlante articulatoire*, dont la particularité est de permettre la visualisation, sous n'importe quel angle, des articulateurs normalement cachés comme la langue et le voile du palais. Chen, Johnson *et al.* (2016) proposent une revue de la littérature sur l'apport de ces techniques d'illustration. Une dizaine d'études mettant en jeu une tête parlante y est présentée. Divers modèles de têtes parlantes articulatoires 3D, basés sur des schémas simplifiés ou des modèles plus complexes, sont proposés pour des applications à des populations variées : personnes malentendantes dotées d'un implant cochléaire, enfants présentant un sigmatisme ou encore personnes aphasiques. Certaines têtes parlantes offrent une visualisation de la langue dans le plan médio-sagittal 2D, à l'aide d'un mécanisme de peau semi-transparente. L'illustration proposée dans chaque cas semble accélérer l'apprentissage et être facilement assimilée par le patient, qui semble ainsi parvenir à associer ce qu'il voit à l'écran, dans l'espace articulatoire d'un autre locuteur, avec sa propre articulation. Eriksson, Bälter *et al.* (2005) recueillent des informations et des conseils auprès des orthophonistes et de neuf patients âgés de 8 à 15 ans. Les personnes interrogées s'accordent généralement à dire que si l'orthophoniste est irremplaçable, il est intéressant de disposer d'un outil simple et motivant, utilisable même hors des séances encadrées, et facilement modulable en fonction des besoins des orthophonistes et des troubles de l'enfant.

Certaines de ces têtes parlantes sont construites à partir de données réelles acquises sur les locuteurs de référence. Nous pouvons ainsi citer Massaro & Light (2004), Bälter, Engwall *et al.* (2005) ou Badin, Tarabalka *et al.* (2010), illustrées à la Figure 6. Elles ont été utilisées dans le cadre de la rééducation de la parole ou de l'apprentissage d'une deuxième langue. Nous nous contentons ici de décrire celle de Bälter *et al.* (2005). Celle de Badin *et al.* (2010), au cœur d'un des systèmes développés dans le cadre de cette thèse, sera décrite plus en détail dans le Chapitre 3 de ce manuscrit. Bälter *et al.* (2005) développent ARTUR (ARticulation TUtoR). Cette tête est conçue à partir des informations statistiques extraites des images IRM d'un locuteur, et des données EMA (détail de la technique en section 1.3.2) acquises sur le même locuteur. Une étude est menée avec cette tête parlante, par une procédure dite de Magicien d'Oz, sur des enfants et des adultes suivant une prise en charge orthophonique. Un expert phonéticien, agissant comme Magicien, anime depuis une autre pièce la tête parlante en fonction de ce qui est prononcé par un participant à travers un microphone. Il récupère pour cela dans une banque de données l'animation articulatoire qu'il considère être la plus proche de celle réalisée par le participant pour produire le mot. Le participant peut de son côté rejouer plusieurs fois l'animation, la ralentir ou visualiser l'articulation correcte, à l'aide de boutons. A la fin de l'investigation, les avis des participants s'avèrent généralement positifs. L'affichage proposé est en effet perçu comme clair et facile à interpréter. Des progrès sont observés sur la parole des enfants. Cependant, n'avoir que dix choix à disposition peut sembler limité lorsqu'on souhaite proposer à un patient le retour le plus proche possible de ce qu'il produit. Accéder directement aux mouvements des articulateurs cachés pourrait ainsi aider à proposer une animation plus automatique et plus pertinente de la tête parlante.

Fagel & Madany (2008) présentent « Vivian », une tête parlante animée semi automatiquement à partir des informations fournies par des images IRM de langue, du velum et d'une partie du mur pharyngé dans le plan médio-sagittal. Elle est contrôlée par huit paramètres, dont la hauteur de trois points de la langue et l'ouverture des lèvres. Vivian est mis en application dans le cadre de la rééducation de huit enfants âgés de 4 à 8 ans atteints d'un trouble articulatoire sur le groupe /s,z/. Les enfants ont semblé comprendre rapidement l'image fournie par la tête parlante et des progrès ont été mesurés, sans qu'il puisse être possible d'affirmer qu'ils sont dus exclusivement à la tête parlante.

Figure 6. Exemples de représentations de têtes parlantes articulatoires. De gauche à droite : Fagel & Madany (2008) Bälter *et al.* (2005) et Badin *et al.* (2010)

1.3. Le retour articulatoire visuel

A la différence de l'illustration, le paradigme de retour visuel articulatoire tend à fournir au patient une information visuelle sur sa propre articulation. Nous détaillons dans cette section quatre technologies permettant cette visualisation en temps-réel : l'électropalatographie, l'articulographie électromagnétique, l'échographie ultrasonore et l'inversion acoustico-articulatoire. Nous comparons ces quatre procédés afin d'essayer de déterminer celui qui pourrait être le plus pertinent pour la rééducation orthophonique.

1.3.1. L'électropalatographie

L'*électropalatographie* (EPG) est une technique relativement ancienne, avec des travaux répertoriés dès 1957, comme en témoigne la revue de la littérature réalisée par Gibbon (2011). Cette technologie consiste en un palais artificiel sur lequel est réparti un ensemble d'électrodes (voir Figure 7, gauche). Un dispositif recueille les signaux de ces électrodes et permet de visualiser, en temps-réel, les contacts de la langue avec le palais. Cette technologie est assez largement utilisée de nos jours dans les pays anglophones pour le traitement de troubles variés. Gibbon, Hardcastle *et al.* (2013) s'intéressent, par exemple, à des enfants présentant une fente palatine. Gibbon & Lee (2015) présentent quant à eux une méthode d'utilisation de l'EPG applicable aux adolescents et aux adultes ayant un trouble articulatoire, et propose une étude de cas avec une adolescente de 12 ans.

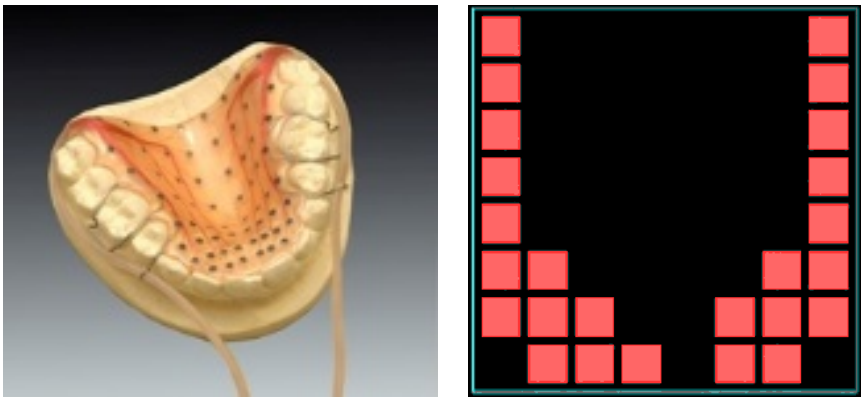


Figure 7. Exemple de palais artificiel pour l'EPG (gauche) et de visualisation des contacts palato-linguaux (droite)

Cependant, l'EPG ne mesurant que les contacts langue-palais, elle n'est pas adaptée pour le travail des voyelles et des consonnes n'impliquant pas ce type de contact comme le /ʁ/. De plus, l'EPG nécessite de construire un palais artificiel pour chaque patient. Pour un enfant suivi durant plusieurs années de rééducation, l'évolution de la forme de son palais exigera donc un changement régulier de matériel.

1.3.2. L'articulographie électromagnétique

L'*articulographie électromagnétique* (Electromagnetic Articulography ou EMA) est une autre technique de capture du mouvement des articulateurs. Elle est basée sur de petites bobines électromagnétiques réceptrices collées sur les articulateurs. La position de ces bobines est déterminée en mesurant les courants induits par des bobines émettrices, fixées soit sur un casque (comme dans le système AG200 illustrée en Figure 8), soit sur un boîtier placé à proximité du locuteur (comme par exemple dans le système NDI Wave). Ainsi, il est possible de mesurer avec précision et d'inférer en temps-réel les mouvements des articulateurs correspondants.

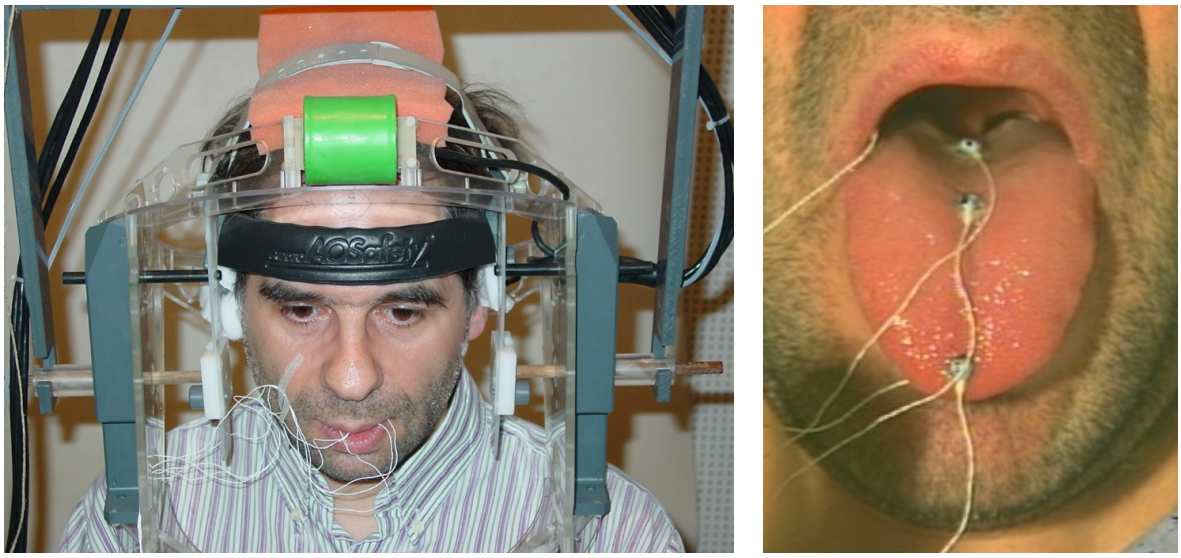


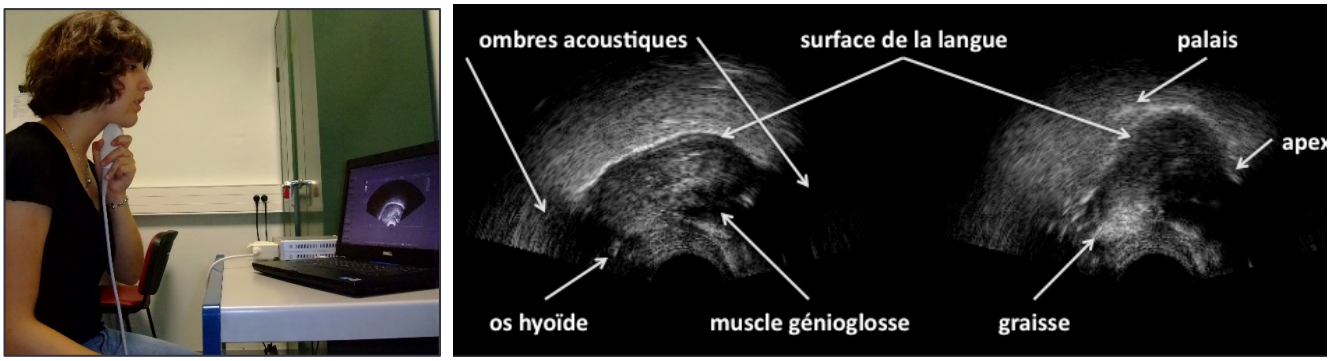
Figure 8. Dispositif EMA complet Carstens AG200 (gauche) et placement des bobines sur la langue (droite)

Les applications cliniques de cette technique dans le domaine orthophonique sont rares. (Katz & McNeil, 2010) proposent une preuve de concept pour des cas d'apraxie pour plusieurs patients. Cette étude nécessite un approfondissement quant à son efficacité réelle en comparaison avec des séances traditionnelles, en augmentant le nombre de patients. Une autre étude de Katz & Mehta (2015), menée sur cinq étudiants, s'intéresse non pas à la rééducation mais à l'apprentissage d'un nouveau phonème /ɖ/. Cinq bobines sont placées sur la langue des participants, permettant d'animer un modèle de langue 3D (OptiSpeech). La cible à atteindre pour la pointe de la langue est matérialisée sur l'image par une sphère. L'expérience se déroule en trois étapes : une étape d'entraînement sans retour visuel, mais avec des retours donnés par l'investigateur ; une étape avec retour visuel, où les réussites sont illustrées par un changement de couleur de la sphère affichée à l'écran ; une dernière étape sans retour visuel. Tous les mouvements linguaux sont enregistrés et permettent d'évaluer la progression. Les participants semblent améliorer le positionnement de la langue par cette méthode.

Si cette technique permet une visualisation précise de la langue pendant la production, mais elle nécessite un équipement coûteux et surtout très invasif, peu adaptable à l'enfant. Elle est donc difficilement applicable à notre contexte. Dans la section suivante, nous détaillons l'échographie (ou imagerie ultrasonore), une technique non invasive permettant, de même, la visualisation de la langue.

1.3.3. L'échographie de la langue

L'échographie, ou imagerie ultrasonore, est une technique d'imagerie biomédicale permettant observer les mouvements de langue pendant la production de parole. Elle est jugée inoffensive et est assez peu invasive pour le locuteur Epstein (2005). Une sonde échographique médicale est placée sous la mâchoire du locuteur (voir Figure 9a). Des ondes ultrasonores se propagent dans les tissus de la cavité buccale et sont généralement réfléchies lorsqu'elles atteignent la surface supérieure de la langue. L'analyse de ces réflexions, ou échos, permet de constituer une image de la langue. En fonction de la position de la sonde, une image dans le plan médio-sagittal (plan le plus utilisé, voir Figure 9b) ou coronal (perpendiculaire au plan médio-sagittal) est obtenue. Cette technique d'imagerie présente de bonnes résolutions temporelles (de l'ordre de 80 images par seconde dans une configuration standard) et spatiales (résolution inférieure à 1 mm), et ne nécessite pas d'équipement volumineux. Parmi les travaux pionniers sur l'utilisation de l'échographie pour l'étude de la production de la parole, nous pouvons citer Stone & Shawker (1986) ou Stone & Davis (1995). Plus récemment, Hueber (2009) a décrit en détail les différents aspects de l'étude du conduit vocal par échographie (concepts physiques sous-jacents, fonctionnement d'un système d'échographie, protocole expérimental pour l'acquisition de données articulatoires, interprétation et traitement des images ultrasonores).



(a) (b)

Figure 9. (a) exemple de positionnement d'une sonde échographique. (b) images ultrasonores de la langue dans le plan médio-sagittal (position de repos à gauche et lors d'un [k] à droite, extrait de Hueber (2009)

L'échographie semble aujourd'hui être une technique privilégiée dans le cadre de l'utilisation du retour visuel pour la rééducation de certains troubles de l'articulation. Les travaux existants sont variés, sur la forme comme sur le fond. Dès 1985, Shawker & Sonies (1985) proposent d'utiliser des images échographiques de la langue pour la rééducation du

/ɹ/ anglais, avec une patiente âgée de 9 ans. Un enregistrement de l'articulation correcte est joué en boucle conjointement avec l'audio associé, pendant que l'enfant répète et visualise sa propre articulation sur un deuxième écran. Ils émettent alors l'hypothèse que ce retour pourrait bénéficier aux populations malentendantes, à condition que le patient soit capable d'interpréter ce qu'il voit et que suffisamment de séances soient consacrées à cette technique. Depuis cette première étude de cas, de nombreuses autres études ont été réalisées, majoritairement dans les dix dernières années. Un état de l'art détaillé de l'application clinique de cette technique est proposé dans le Chapitre 4. Il souligne la diversité des études proposées, que ce soit dans le cas de l'apprentissage pour des personnes sourdes implantées, résumé dans Gallagher (2013), ou pour des enfants ayant des difficultés à acquérir certains phonèmes Jonathan L. Preston, Maas *et al.* (2016). Des applications existent aussi dans le cadre de l'apprentissage d'une langue seconde (L2) incluant des phonèmes différents de la langue maternelle de l'apprenant, comme dans les études de Wu, Gendrot *et al.* (2015) et Pillot-Loiseau, Kamiyama *et al.* (2015). Enfin, une étude réalisée par Cleland, McCron *et al.* (2013) compare l'efficacité de l'échographie et de l'EPG auprès de vingt adultes non pathologiques. Ce travail ne permet pas de mettre en avant une des deux techniques, mais insiste sur le fait que la faculté de lecture linguale n'est pas présente chez tous les participants. L'échographie semble cependant gagner la confiance des orthophonistes par les résultats qu'elle propose et par sa simplicité de mise en œuvre. Alors que les premières études réalisées sont limitées à des études de cas sur un patient, des travaux menés actuellement à plus large échelle tentent de confirmer l'apport positif de cette technologie.

1.3.4. L'inversion acoustico-articulatoire

Les techniques d'inversion acoustico-articulatoire, visant à estimer les mouvements des articulateurs directement à partir du signal de parole capté par un microphone, peuvent être mises en œuvre dans le cadre de systèmes de retour visuel. Les limites de l'inversion acoustico-articulatoire reposent sur un problème récurrent du traitement automatique de la parole, donnant lieu encore aujourd'hui à une littérature abondante. Il s'agit d'un problème *mal posé* au sens mathématique du terme : deux configurations articulatoires différentes peuvent correspondre à des réalisations acoustiques proches, et réciproquement (Atal, Chang *et al.* (1978)). Ce problème est classiquement abordé aujourd'hui à l'aide des techniques dites de modélisation par apprentissage statistique (ou *machine learning*, approche qui sera notamment détaillée dans le Chapitre 2). Différentes techniques ont été proposées pour modéliser les relations complexes entre le contenu spectral d'une part, et la position des articulateurs d'autre part, parmi lesquelles les modèles de Markov Cachés Zen, Nankaku *et al.* (2011), les modèles de mélanges de gaussiennes (GMM) pour Toda,

Black *et al.* (2008) ou les réseaux de neurones Richmond (2002). Cette liste est non exhaustive. Nous citerons enfin les approches par dictionnaire (codebook) tel que proposées dans Ouni & Laprie (2005). Hueber *et al.* (2015) proposent par ailleurs une technique d'inversion visant à estimer les mouvements articulatoires d'un locuteur de référence à partir du signal de parole d'un autre locuteur. Au cœur d'un des systèmes de retour visuel développé dans cette thèse, cette méthode sera détaillée dans le Chapitre 3.

La thèse de Ben Youssef (2011) présente un premier système complet de retour visuel basé sur une tête parlante articulatoire, animée automatiquement à partir du signal de parole de n'importe quel locuteur, par inversion acoustico-articulatoire (approche basée sur l'utilisation de HMM). Une implémentation temps-réel à partir de GMM a également été proposée dans Hueber, Youssef *et al.* (2012). Ce système serait très intéressant pour une utilisation en orthophonie car il ne requiert aucun capteur externe coûteux, mais uniquement un microphone et un ordinateur pour l'affichage. L'utilisation de ce système dans ce contexte se heurte cependant aux problèmes suivants :

- les performances d'inversion sur certaines consonnes, dont les plosives, sont relativement faibles. Ceci peut s'expliquer par la présence dans le signal audio d'une zone de silence avant la plosion (*burst*) qui correspond à la phase d'occlusion et pendant laquelle les relations acoustico-articulatoires sont difficilement modélisables.
- il est fréquent qu'un orthophoniste fasse travailler le patient sans vocalisation. C'est notamment le cas dans le travail de certaines praxies, ou bien lorsque le patient a une capacité d'articulation intacte mais une phonation altérée. Cela peut se produire dans le cas d'une aphasie dite globale, où les patients sont souvent mutiques en période aiguë de l'AVC. Dans ces cas, ce système ne peut être utilisé.

Aussi, si l'inversion acoustico-articulatoire reste une piste de recherche très prometteuse, les performances ne semblent pas encore assez satisfaisantes pour une utilisation en rééducation orthophonique. Dans ce travail, nous privilégions une approche visant à capturer directement le mouvement de la langue à l'aide de l'échographie.

La plupart des études actuelles sur le retour visuel, par échographie ou par d'autres méthodes, s'effectue dans une démarche clinique, qui cherche à quantifier les progrès que permet cette technique. Il peut cependant être intéressant de discuter ce paradigme d'un point de vue plus fondamental, en recherchant quels sont les mécanismes physiologiques et cognitifs qui sous-tendent ces progrès. Le paradigme du retour visuel questionne en effet les relations existant entre la production de la parole d'une part et la perception visuelle,

acoustique et proprioceptive d'autre part. Ces relations sont brièvement discutées dans la section suivante, au regard de notre contexte applicatif.

1.4. Les relations entre production et perception : représentations internes de la parole

Dès les premiers jours et tout au long de la vie, nous apprenons en imitant ceux qui nous entourent. La très célèbre étude de Meltzoff & Moore (1997) montre que dès l'âge de six semaines un nourrisson peut imiter une protrusion de la langue réalisée par des adultes. Le bébé est ainsi capable de reproduire les gestes d'autres personnes sans avoir aucune conscience visuelle de ses propres gestes, ni aucune conscience de son propre corps. Cette faculté serait imputable au système miroir. Jusqu'alors observée chez les singes, l'existence de *neurones miroirs* chez l'homme est étudiée dès les travaux de Rizzolatti, Fadiga *et al.* (1996) à travers une observation par TEP (Tomographie par Emission de Positron). Elle semble finalement démontrée en 2010 par les travaux de Mukamel, Ekstrom *et al.* (2010). Les auteurs remarquent que certains neurones s'activent à la fois lors de l'observation et lors de l'imitation d'expressions faciales et de saisie de la main, suggérant l'existence de ces neurones. L'apprentissage par imitation nécessite de pouvoir percevoir les autres. Dans le cas de l'apprentissage du langage, la perception des autres et de soi-même est indispensable, comme le notent J.S. Perkell (2012) ou Turgeon, Prémont *et al.* (2015) pour l'audition, ou encore Mills (1987) pour la vision. Mills (1987) rapporte les difficultés des enfants non-voyants à apprendre les contrastes entre /m/ et /n/, faciles d'un point de vue visuel, mais difficiles d'un point de vue acoustique. Cowie, Douglas-Cowie *et al.* (1982) mettent en évidence une tendance des malentendants post-linguaux à dégrader la parole qu'ils produisent.

Afin de formaliser les liens entre production et perception, différentes études proposent des modèles internes de représentation de la parole. Ainsi, J.S. Perkell, Guenther *et al.* (2000) développent une théorie de la composante segmentale du contrôle moteur. Cette théorie considère que la programmation des mouvements articulatoires pour atteindre des cibles auditives (phonèmes, segments) est basée sur un modèle interne. Ce modèle, aussi appelé *copie d'efférence*, prédit les conséquences acoustiques à partir des configurations articulatoires. Le retour auditif permettrait en premier lieu à l'enfant d'acquérir ce modèle interne lors de l'apprentissage du langage, et ensuite de le maintenir tout en l'adaptant aux changements de morphologie liés à la croissance. La *copie d'efférence* permet d'assurer la correction des commandes motrices à partir des différences entre les cibles acoustiques et les cibles prédites par le modèle en l'absence de retour effectif (J. Perkell, Matthies *et*

al. (1997)). Le cerveau se base sur les informations qu'il reçoit de ses différents organes, comme la tension des tendons, la longueur des muscles ou les contacts pour situer dans l'espace l'ensemble des parties du corps, même lorsqu'elles sont invisibles pour l'œil : c'est la proprioception. Les configurations articulatoires sont liées à ces informations (J. Perkell *et al.* (1997)).

L'acquisition de la parole passe donc par l'exploitation du retour auditif et orosensoriel, et par la visualisation de l'autre. De nombreuses études ont mis en évidence et quantifié l'apport des articulateurs visibles pour la perception de la parole (*e.g.* Erber (1975), Sumby & Pollack (1954), Benoît & Le Goff (1998)).

Nous avons donc établi des liens existant entre la production et la perception de la parole. Cependant, la seule vision des lèvres et du visage, bien qu'utile, ne fournit qu'une information phonétique très incomplète, en particulier sur les articulateurs invisibles. Montgomery (1981) a montré que l'humain possède néanmoins une certaine conscience articulatoire, c'est-à-dire une connaissance de la place de ses articulateurs. Il semble donc possible d'évaluer la capacité d'un locuteur à exploiter une information visuelle sur un articulateur pourtant invisible, dont nous ne possédons qu'un retour proprioceptif. Badin *et al.* (2010) ont donc testé l'hypothèse selon laquelle les humains seraient capables d'utiliser la vision de la langue pour la reconnaissance des phonèmes, comme ils le font en lecture labiale. Cette étude montre que les sujets sont en effet dotés d'une certaine capacité de *lecture linguale*, c'est-à-dire qu'ils sont capables d'exploiter une représentation visuelle des mouvements de la langue pour améliorer leur perception (notamment des consonnes) lorsque le signal audio est fortement dégradé ou absent.

La visualisation du mouvement d'un articulateur invisible semblerait donc bénéfique pour l'apprentissage de celui-ci. Cependant, il n'est pas prouvé que, pour un mouvement donné, il soit plus pertinent de visualiser son propre mouvement plutôt qu'un mouvement cible à atteindre. Dans l'étude de Badin *et al.* (2010) les participants sont en effet confrontés à la représentation visuelle d'une articulation qui n'est pas la leur. Ouni (2014) s'intéresse à cette problématique à travers une expérience d'acquisition de nouveaux mouvements linguaux. Les participants sont divisés en deux groupes. Pour les deux groupes, il est demandé de réaliser une série de mouvements de la langue. Ces mouvements sont enregistrés grâce à une sonde échographique, dont le fonctionnement est détaillé en section 1.3.3. La sonde donne une information sur le mouvement de leur langue, mais les participants n'ont pas accès à cette image. Pour le deuxième groupe, une session d'entraînement est ajoutée, avec un retour visuel en temps-réel par échographie. Il en ressort que le deuxième groupe, ayant une connaissance des mouvements de la langue,

présente de meilleures performances dans l'acquisition de nouveaux mouvements. Fournir un retour visuel même succinct semblerait donc faciliter l'acquisition de nouveaux mouvements. L'auteur précise que la façon d'intégrer le retour visuel efficacement dans le processus d'apprentissage reste néanmoins à déterminer. Il se pourrait que ce soit l'ajout d'une étape, et donc l'allongement de la durée d'apprentissage, qui facilite l'acquisition du mouvement, plus que la nature même du travail effectué dans cette étape. Dans les deux cas, le retour visuel permet cependant aux participants de prendre conscience de leur propre articulation, et de se concentrer ainsi sur les conséquences d'un mouvement dont ils ont l'information visuelle, plutôt qu'en s'aidant simplement du retour auditif et éventuellement perceptif habituel.

Ces deux études soulignent les capacités naïves des participants à exploiter une information visuelle sur un articulateur invisible. Nous aurions donc conscience de nos articulateurs invisibles, et en particulier des mouvements de notre propre langue. Reste en suspens la question suivante : **visualiser la langue améliore-t-il réellement l'apprentissage d'une nouvelle articulation, ou suffit-il de visualiser un modèle de référence ?**

1.5. Conclusions de l'état de l'art

Cet état de l'art fournit plusieurs informations importantes sur les liens entre production et perception de la parole. Au cours de l'apprentissage, l'enfant a conscience de l'articulation de sa langue et exploite cette connaissance pour réajuster ses productions. Les travaux de Badin *et al.* (2010) et Ouni (2014) nous confortent sur l'intérêt d'utiliser un retour articulatoire visuel ou une illustration pour la rééducation orthophonique. Pour ce faire, nous avons fait le choix de la sonde échographique pour le retour visuel. En effet, malgré son coût qui peut être un investissement important pour un orthophoniste, cette technique ne nécessite aucune installation particulière. Elle est totalement indépendante de l'utilisateur, et n'altère en rien sa production puisque rien n'est ajouté sur la langue ou dans la cavité vocale. L'échographie fournit une image en temps-réel de la véritable articulation du patient, donnant ainsi des informations supplémentaires et pertinentes à l'orthophoniste comme au patient. Elle présente donc un intérêt important pour l'orthophonie par rapport aux méthodes traditionnelles.

Nous avons vu à travers cette revue de la littérature que l'utilisation de l'échographie comme retour articulatoire visuel semble avoir un impact positif sur la rééducation orthophonique. Ces constats se basent cependant sur des études incluant peu de

participants, souvent des études de cas, comme nous le détaillerons au Chapitre 4. Une étude à plus large échelle pourrait permettre de valider ces observations.

Cependant, comme illustré dans l'introduction en Figure 1, un des problèmes posé par le retour visuel par échographie est la difficulté, pour le patient, d'interpréter l'image qu'il voit. L'image échographique ne fournit qu'une information dans un plan 2D sur le contour supérieur, parfois incomplet, de la langue. Elle est de plus dégradée par la présence d'un bruit dit de *speckle* et ne donne aucune information sur les limites de la cavité orale comme le palais ou les dents.

Afin de pallier ces manques, nous proposons donc le concept d'*échographie linguale augmentée* (par analogie à *réalité augmentée*). Il s'agit d'ajouter en temps-réel à une échographie linguale certains éléments initialement non-visibles afin d'en améliorer la compréhension. Nous proposons deux approches :

- Mettre en surbrillance dans l'image échographique le contour correspondant à la surface supérieure de la langue. Il s'agit d'un problème de segmentation d'image que nous abordons à l'aide d'une technique d'apprentissage statistique (*machine learning*). Notons que dans ce mode de visualisation, le retour visuel se situe dans l'espace articulatoire de l'utilisateur.
- Animer automatiquement le modèle de langue d'une tête parlante, qui permet de visualiser l'ensemble des structures qui composent l'appareil vocal (palais, pharynx, etc.). Le retour visuel se situe ici dans un espace articulatoire *a priori* différent de celui de l'utilisateur. Nous verrons qu'il s'agit donc d'un problème de régression entre deux modalités et deux locuteurs que nous abordons également à l'aide d'une approche par apprentissage statistique.

Enfin, nous proposons une évaluation clinique d'une version simplifiée de ces systèmes d'échographie augmentée dans le cadre de la rééducation de personnes glossectomisées. Nous présentons ici cinq études de cas d'un essai clinique encore en cours au moment de la rédaction de ce document.

Ces trois pistes d'exploration constituent les trois prochains chapitres de ce manuscrit.

Chapitre 2. L'échographie linguale augmentée : suivi de la langue

Nous avons vu dans le chapitre 1 que proposer un retour visuel par échographie aux patients suivant une prise en charge orthophonique était susceptible de les aider à mieux comprendre leur trouble en visualisant leurs propres mouvements de langue. Cependant, un des problèmes posé par le retour visuel est la difficulté, pour le patient, d'interpréter l'image qu'il voit. Notre hypothèse est que cette difficulté peut s'expliquer en partie par le caractère bruité de l'image échographique, qui ne fournit qu'une information sur le contour supérieur de la langue, dans le plan 2D choisi (médio-sagittal ou coronal), et n'informe en aucun cas sur les limites de la cavité orale offertes par le palais ou les dents par exemple. De plus, certaines parties de ce contour peuvent être mal imagées lorsque le contour de la langue prend une position quasiment parallèle au faisceau ultrasonore incident. Cette limitation est classique pour les systèmes d'échographie.

Dans la suite de ce manuscrit, nous proposerons deux méthodes d'échographie augmentée afin de pallier ces carences. Ces deux méthodes ont été développées en exploitant des algorithmes d'apprentissage automatique ou *machine learning*. Le *machine learning* regroupe un ensemble d'algorithmes permettant d'apprendre un modèle de façon automatique à partir d'une base de données d'exemples ou base d'apprentissage. Dans le cadre de cette thèse, nous nous plaçons dans le contexte de l'apprentissage *supervisé*, pour lequel la base de données d'apprentissage met en regard des *observations d'entrée* avec des *observations de sortie*. L'objectif est ici de modéliser les relations entrées-sorties et d'être capable de prédire une observation de sortie à partir d'une observation d'entrée non vue pendant l'apprentissage : on parle alors de *généralisation*. Lorsque les observations de sorties sont des vecteurs de variables continues, il s'agit alors d'un problème de régression ; dans le cas discret, il s'agit d'un problème de classification.

Dans ce chapitre, nous proposerons une méthode d'extraction automatique du contour de la langue dans les images échographiques. Notre objectif est de rendre l'image plus lisible et donc interprétable plus intuitivement, pour un patient. Comme nous le verrons en section 2.1, le suivi de la langue dans les images échographiques est un problème qui a déjà fait l'objet de plusieurs travaux. Dans notre contexte d'application en rééducation

orthophonique, nous aborderons cette question de traitement d'image avec les contraintes suivantes :

- disposer d'une méthode aussi automatique que possible, et nécessitant un minimum d'intervention humaine ;
- disposer d'une méthode plus robuste que celles exposées dans l'état de l'art, notamment la méthode EdgeTrak, basée sur la technique des contours actifs, décrite en section 2.1.1.

La méthode proposée s'appuie sur l'idée selon laquelle il existe un lien statistique entre la distribution de l'intensité des pixels de l'ensemble de l'image et la position du contour de la langue dans cette même image. Autrement dit, **notre hypothèse est qu'une partie manquante du contour peut être estimée non seulement à partir de la connaissance des autres parties de ce contour, mais également sur la base des autres structures présentes dans l'image** : les tissus et muscles visibles sous la surface de la langue, mais aussi la distribution du *speckle* dans l'image, que nous supposons légèrement corrélée à la position de la langue dans la cavité buccale. D'autres travaux en traitement d'images échographiques s'appuient sur cette même idée d'exploiter le bruit de *speckle* pour suivre des objets déformables (voir Yeung, Levinson *et al.* (1998) pour une revue de la littérature).

Ce chapitre est organisé de la façon suivante. Nous commencerons par un état de l'art des méthodes d'extraction automatique du contour de la langue dans les images échographiques. Nous décrirons ensuite les principes théoriques de la méthode proposée. Puis, nous présentons le protocole expérimental utilisé pour son évaluation. Enfin, nous présenterons les résultats expérimentaux obtenus et nous les discuterons.

2.1. Etat de l'art

D'après les définitions disponibles dans le domaine du traitement d'images, la langue, dans un plan 2D, est un objet déformable. Contrairement aux structures osseuses de son environnement, comme la mâchoire par exemple, sa forme varie au cours du temps. De plus, l'absence de points saillants dans l'image échographique, existant dans les images naturelles, rend le problème difficile.

2.1.1. Modèle de contours actifs

Une des techniques les plus utilisées pour traiter cette question du suivi de la langue est une adaptation des modèles de contours actifs (ou *snakes*) à la problématique du suivi de

la langue. L'objectif des modèles de contour actif est de minimiser l'énergie totale associée à l'objet d'intérêt. Cette énergie est définie par l'équation suivante :

$$E_{total} = \alpha E_{int} + \beta E_{ext}$$

où E_{int} correspond à l'énergie interne (propriétés géométriques et élastiques de l'objet déformable) et E_{ext} représente l'énergie externe du modèle (propriétés globales de l'image). Li, Kambhamettu *et al.* (2005) présentent une implémentation de ce modèle appliqué au contour de langue dans le logiciel EdgeTrak. La première image est segmentée manuellement par l'utilisateur, et les images qui suivent dans la séquence sont segmentées automatiquement. Cette approche donne de bons résultats, tant que le contour de la langue est clairement visible. Cependant, la qualité de la segmentation décroît fortement dès qu'une partie de la langue vient à disparaitre Roussos, Katsamanis *et al.* (2009). Les derniers travaux autour d'EgdeTrak présentés par Xu, Yang *et al.* (2016) semblent améliorer les performances de cet algorithme, avec un fonctionnement sur de longues séquences intégrant une réinitialisation automatique du contour. Cependant, le corpus d'évaluation de la méthode n'est pas clairement détaillé dans l'article.

Tang, Bressmann *et al.* (2012) présentent une autre méthode de suivi de contour de langue, basée sur des champs aléatoires de Markov, qui sont des modèles graphiques non-orientés. Dans cette méthode, appelée TongueTrack, une fonction d'énergie similaire à celle d'EdgeTrak est calculée. L'équation générale s'écrit comme suit :

$$E_{total} = \alpha E_{data} + \beta E_{temporal} + \gamma E_{spatial} + \delta E_{length}$$

Cette fonction d'énergie dépend non seulement des données d'apprentissage du modèle, mais aussi des informations temporelles (trames précédentes) et d'informations spatiales de longueur. TongueTrack peut annoter jusqu'à 500 trames consécutives pour une erreur moyenne de 3 mm, ce qui est, d'après les auteurs, meilleur que les performances d'EdgeTrak sur le même corpus.

2.1.2. Modèles externes de la langue

D'autres études mettent en avant l'utilisation d'un modèle externe de la langue pour régulariser le processus de segmentation. Roussos *et al.* (2009) proposent ainsi une approche basée sur les Modèles Actifs d'Apparence (*Active Appearance Model* ou AAM), inspirée des travaux de Cootes, Edwards *et al.* (1998). Le modèle est ici construit sur une autre modalité : une base de 700 images ciné-radiographiques (X-ray), annotées

manuellement. Un avantage de cette technique est sa capacité à extrapoler le contour de la langue à l'avant comme à l'arrière de la cavité buccale. Toujours dans cette même perspective de régularisation, Loosvelt, Villard *et al.* (2014) proposent d'utiliser un modèle biomécanique de la langue pour contraindre le déplacement des points de contours au cours du temps. Les deux approches semblent surpasser les techniques de l'état de l'art basées sur des *snakes*. Cependant, la première méthode de Roussos et al. ne semble pas être très adaptée à la rééducation orthophonique en raison de la nécessité d'acquérir des images ciné-radiographiques du patient. Pour la seconde méthode, il n'est pas précisé dans l'article si l'évaluation a été menée auprès de plusieurs locuteurs.

Fasel & Berry (2010) proposent de modéliser les relations entre l'intensité de l'ensemble des pixels d'une image et la position du contour de la langue. Les auteurs proposent d'apprendre cette relation à l'aide des techniques récentes d'apprentissage dit profond (ou *deep learning*). La méthode proposée dans ce chapitre s'inscrivant dans cette même démarche de modélisation des relations entre pixels et contours, nous la présentons ici plus en détail.

2.1.3. Approche par réseau de neurones artificiels

Fasel & Berry (2010) propose une méthode basée sur un réseau de neurones artificiel dit *profond* présentant une architecture particulière nommée *translational Deep Beliefs Networks* ou *tDBN*. Ces réseaux sont entraînés pour modéliser directement la relation entre les pixels bruts de l'image échographique et les coordonnées des points qui décrivent le contour de la langue. Ces réseaux, non détaillés ici, sont d'abord entraînés de façon non supervisée à répliquer les observations d'entrées par un apprentissage génératif basé sur l'utilisation de Machines de Boltzmann restreintes (RBM). Dans la lignée des travaux récents sur le *Deep Learning*, il s'agit ici de construire de façon automatique des représentations des observations dites de haut-niveau (également appelées abstractions), à partir des données brutes, sans passer par une phase d'extraction de descripteurs. Cette modélisation brute mène aujourd'hui à des avancées significatives dans certains domaines comme la reconnaissance d'images ou de parole. Elle nécessite cependant la mise en place de réseaux très complexes, c'est-à-dire avec un très grand nombre de paramètres ajustables, et entraînés sur de très grandes bases de données. C'est le cas du réseau proposé par Fasel & Berry (2010), qui contient 5514 neurones répartis sur trois couches cachées, avec des observations d'entrée/sortie de dimensions 646 (*i.e.* l'image originale réduite à 18x34 pixels). Le nombre de paramètres ajustables est donc très important. L'entraînement se fait sur une base de données de 8000 images, annotées manuellement à partir d'une base de sept locuteurs. Dans Fasel & Berry (2010), l'évaluation a été effectuée

sur un corpus de mots incluant le son /l/ dans des phrases porteuses. D'après les auteurs, les résultats semblent égaler la méthode *EdgeTrak*, avec une erreur moyenne de 0,75 mm. Le contenu des corpus évalués étant différent, il est cependant impossible de l'affirmer.
Cette méthode nous apparaît à ce jour comme étant la plus précise. Cependant, nous considérons que la nécessité de disposer d'une base de données de plusieurs milliers d'images est un facteur limitant pour une utilisation en orthophonie. De plus, il n'a pas été démontré expérimentalement que ce type de modèle généralise à un nouveau locuteur, ni à de nouvelles conditions d'acquisition, induites par la position de la sonde ou par un nouvel échographe.

2.2. Principe général de la méthode proposée

La méthode que nous proposons s'appuie :

1) sur le paramétrage de l'ensemble des pixels d'une région d'intérêt par ACP (on parlera de l'approche par EigenTongues),

2) sur le paramétrage des contours de langue également par ACP (par analogie, on parlera de l'approche EigenContours)

3) une mise en correspondance de ces deux espaces de représentation à l'aide d'un réseau de neurones artificiel. La méthode proposée est évaluée sur deux corpus.

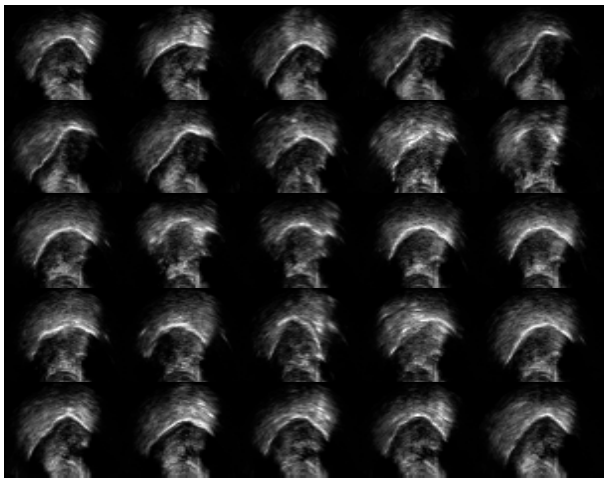
2.2.1. Approche EigenTongues pour le paramétrage des images ultrasonores

Dans la méthode proposée, les images ultrasonores sont paramétrées à l'aide de la technique EigenTongues, proposée par Hueber, Aversano *et al.* (2007). Cette méthode est une adaptation directe aux images échographiques linguales de la méthode des EigenFaces proposée par Turk & Pentland (1991) dans le cadre de la reconnaissance faciale et peut se résumer comme suit. Considérons un ensemble de M images d'apprentissage. Redimensionnées à une taille NxN, ces images sont ensuite vectorisées (ligne par ligne ou colonne par colonne) après avoir soustrait la moyenne afin de former une matrice A de taille N^2xM. L'analyse en composantes principales de A permet d'obtenir un nouvel espace de représentation des images défini par :

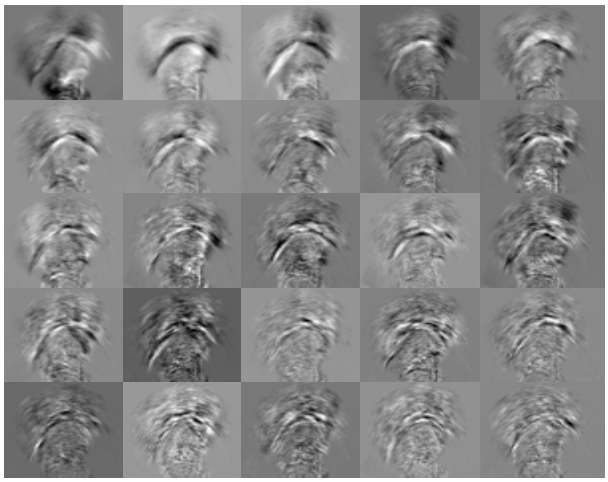
$$C = \frac{1}{M}(AA^\top) \; et \; R^\top CR = \Lambda$$

où C est la matrice de covariance de A, R la matrice des vecteurs propres et Λ la matrice des valeurs propres associées. Les vecteurs propres sont de taille N^2 et peuvent donc être représentés sous la forme d'images de taille NxN. Ils sont appelés EigenTongues. Chaque

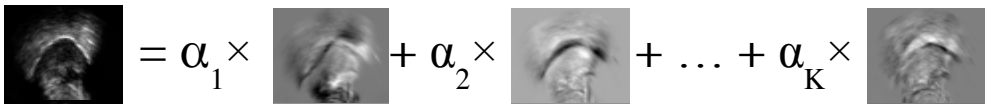
image échographique peut donc être représentée comme une combinaison linéaire de EigenTongues, comme illustré à la Figure 10. Une nouvelle image (non comprise dans la base d'apprentissage) peut donc être paramétrée par ses K premières coordonnées dans la base des EigenTongues. Comme détaillé par la suite, nous montrons que cette technique permet d'extraire un vecteur de descripteurs visuels de taille K avec $K \ll N^2$.



(a) Base d'apprentissage



(b) Vecteurs EigenTongues



(c) Projection d'une image sur les K premiers EigenTongues

Figure 10. Extraction des vecteurs Eigentongues (b) à partir d'une base d'apprentissage redimensionnée à 64x64 pixels (a) et projection d'une image sur les K premiers vecteurs EigenTongues (c).

2.2.2. Approche EigenContours pour le paramétrage des contours de la langue

La méthode proposée s'appuie sur une base d'images d'apprentissage segmentées manuellement. Une grille semi-polaire est placée sur l'image échographique, comme illustré en Figure 11. Le centre de la grille correspond au centre de la sonde, dont la position sur l'image est connue. L'angle d'ouverture de la grille est adapté à la morphologie du locuteur : les lignes externes de la grille sont choisies pour être alignées en moyenne sur les ombres créées par l'os hyoïde et l'os de la mâchoire (Figure 9), délimitant ainsi l'espace de visualisation de la langue. Ce choix d'une géométrie variable adaptée au locuteur vise à normaliser les différences morphologiques entre les locuteurs. L'annotation manuelle est

facilitée par l'utilisation de courbes de Bézier : l'utilisateur manipule l'ensemble de la courbe grâce à quelques points de contrôle, au lieu de cliquer sur chaque point d'intersection entre la langue et les axes de la grille. Cependant, à l'issue de l'annotation manuelle d'une image, le contour final de la langue est représenté par les coordonnées x/y des points d'intersection entre le contour et les axes.

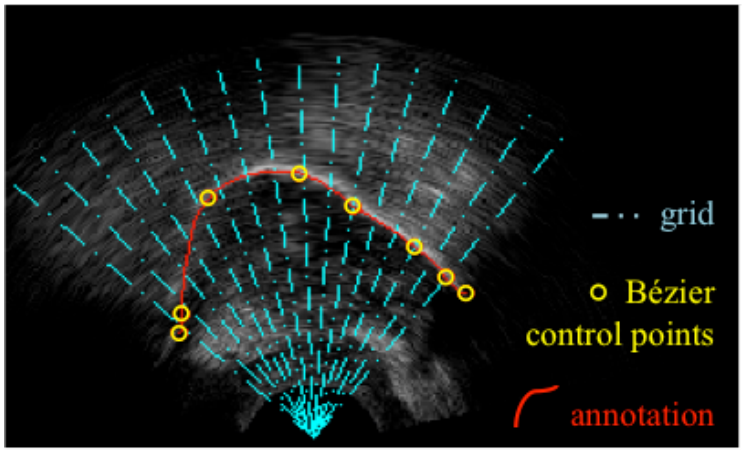


Figure 11. Grille semi-polaire placée sur les images échographiques pour l'annotation manuelle du contour.

Lorsque le contour de la langue ne coupe pas certains axes, un code spécifique (*ex.* NaN) est affecté aux coordonnées correspondantes, signifiant ainsi que la valeur est manquante. Ces points sont généralement placés aux extrémités de la langue.

La Figure 12 présente le processus d'annotation, qui se décrit comme suit : d'une image brute, on extrait P points de coordonnées (x,y) qui sont placés en alternance dans un vecteur. A chaque image annotée correspond ainsi un vecteur unique de dimension 2P.
Dans l'ensemble des expériences de ce chapitre, nous avons choisi P = 16.

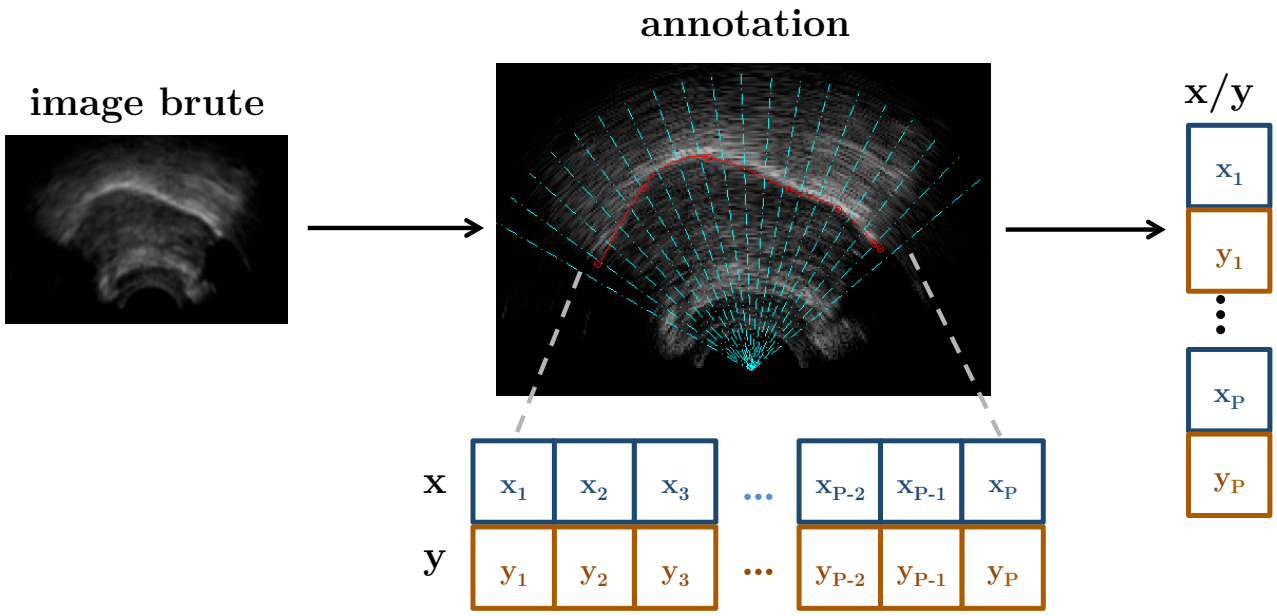


Figure 12. Processus d'annotation des images échographiques avec P points de contour par image.

De la même manière que pour les images, des vecteurs *EigenContours* sont extraits à partir des contours annotés manuellement sur les images. Cependant, certains contours présentent des données manquantes. Nous avons appliqué une ACP probabiliste (p-PCA) pour estimer ces points absents à partir des autres données du corpus. L'algorithme est proposé par Porta, Verbeek *et al.* (2005). La p-PCA repose sur l'utilisation de l'algorithme itératif EM (Expectation-Maximization), où les valeurs manquantes sont estimées à chaque itération de l'algorithme, tout en minimisant l'erreur entre les données réelles et ré-estimées.

2.2.3. Modélisation de la relation EigenTongues-EigenContours par réseau de neurones artificiels

Nous commençons par rappeler brièvement le fonctionnement d'un réseau de neurones artificiel. Ce rappel se veut concis ; la littérature sur les réseaux de neurones et en particulier les réseaux profonds étant en pleine explosion, une description exhaustive des architectures possibles (C.M. Bishop (1995)) dépasserait le cadre de ce travail.
Un neurone formel est une fonction mathématique réalisant une transformation non-linéaire g de la combinaison linéaire des n entrées $\mathbf{x_i}$ avec les paramètres de pondération (ou *poids*) $\mathbf{w_i}$ avec i=1:n, tel que : $s = g(\sum_{i=0}^{n} w_i x_i)$ où s est la sortie du neurone formel.
Parmi les non-linéarités, ou fonction d'activation, les plus communément utilisées, nous

citerons la fonction sigmoïde, ainsi que la ReLU (*Rectified Linear Unit*), très utilisée dans les réseaux profonds. Le fonctionnement général d'un neurone formel est illustré par la Figure 13

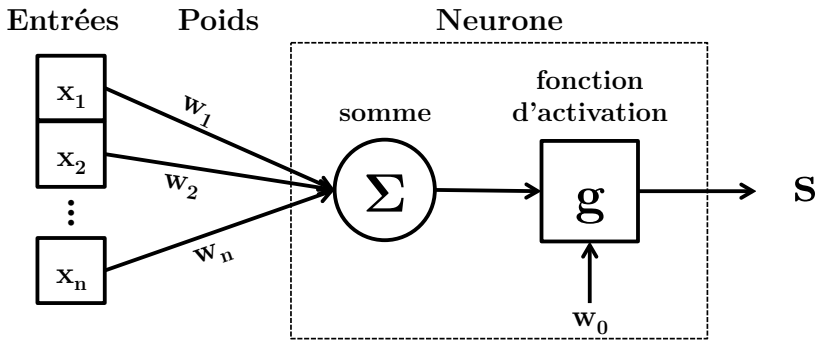


Figure 13. Fonctionnement général d'un neurone formel

Un réseau de neurones est une structure composée de plusieurs neurones formels, organisés en *couches* et connectés les uns aux autres selon une architecture particulière. L'architecture la plus classique est le perceptron multicouche (*MultiLayer Perceptron* ou *MLP*), qui est composé : d'une couche d'entrée ; d'une ou plusieurs couches dites cachées ; d'une couche de sortie pour laquelle la fonction d'activation est souvent linéaire dans le cas d'un problème de régression, comme le nôtre. La Figure 14 illustre schématiquement l'organisation d'un perceptron multicouche.

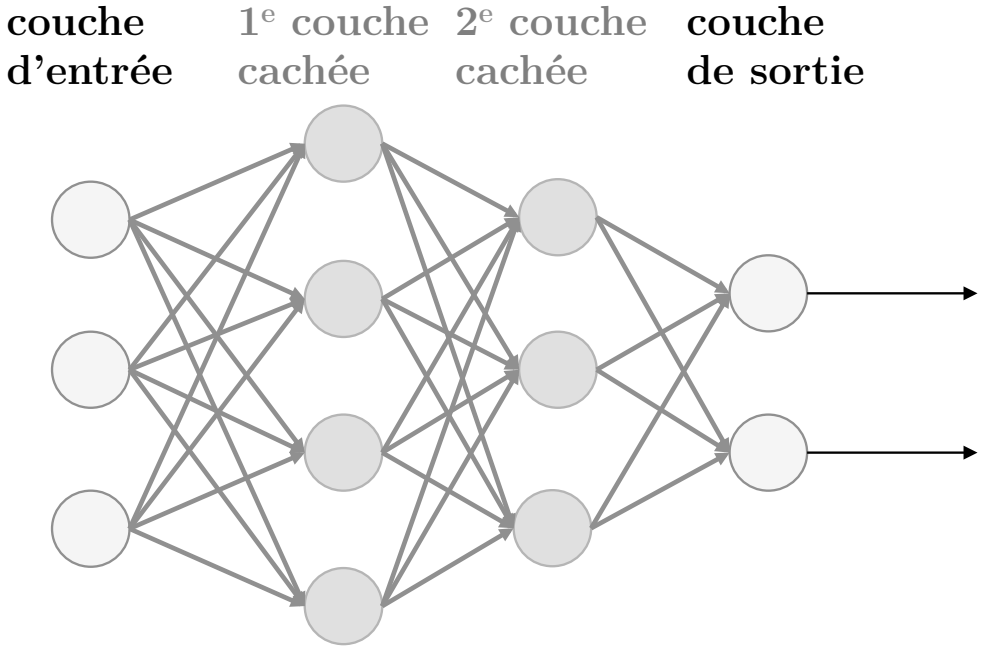


Figure 14. Illustration d'un perceptron multicouche à deux couches cachées.

Un MLP calcule la valeur d'une observation de sortie à partir de la valeur d'une observation d'entrée, qui lui est présentée sur sa première couche. Il s'agit d'un réseau *feed-forward* : l'information ne circule que dans un sens, depuis la couche d'entrée vers la couche de sortie. Bien d'autres architectures existent, comme les réseaux récurrents pour lequel une ou plusieurs boucles de rétroaction sont présentes dans le réseau (Pineda (1987)). Un réseau de neurones artificiel peut être utilisé pour résoudre des problèmes de régression ou de classification. Il peut s'entraîner de façon supervisée ou non supervisée, ou par une succession d'apprentissages non-supervisés puis supervisés : c'est notamment le cas pour l'initialisation des réseaux profonds, la partie supervisée étant alors dite de *fine-tuning*. L'estimation des poids du réseau à partir d'une base de données d'apprentissage est un problème d'optimisation : les poids des neurones, initialisés à des valeurs aléatoires, sont ensuite réajustés en fonction de l'erreur commise par le réseau dans l'estimation de la sortie. Il s'agit donc de minimiser l'erreur entre les prédictions du réseau et les valeurs originales de la base d'apprentissage jusqu'à convergence. La procédure générale d'optimisation est connue sous le nom de rétro-propagation du gradient (*backpropagation*). De nombreuses variantes de cette technique ont été proposées, comme l'algorithme Levenberg-Maquardt proposé par Marquardt (1963) bien adapté aux réseaux de petite taille (*i.e.* une seule couche cachée), ou l'algorithme SGD (*stochastic gradient descent*), très utilisé pour l'apprentissage des réseaux profonds, notamment en cas d'entraînement par mini-batch.

La méthode de détection de la langue proposée dans notre étude s'inscrit dans la continuité des travaux de Fasel & Berry. Dans la méthode proposée, nous modélisons également la relation entre l'intensité des pixels de l'ensemble de l'image échographique d'une part, et le contour de la langue d'autre part, à l'aide d'un réseau de neurones artificiel (ANN). Mais, contrairement à Fasel & Berry, nous cherchons à minimiser la taille de la base d'apprentissage tout en maintenant une bonne capacité de généralisation. Nous rejoignons ainsi d'une approche plus classique en *machine learning*. Avec une étape préalable d'extraction de descripteurs visant à représenter l'information pertinente à l'aide des approches EigenTongues et EigenContours, nous réduisons la dimension des observations d'entrée et de sortie. Rappelons que Fasel & Berry utilisent un paradigme de *deep learning* dans lequel les observations d'entrées du réseau sont les pixels bruts, les premières couches du réseau étant dédiées à l'extraction de descripteurs discriminants. Nous pouvons alors mettre en place des réseaux de neurones avec moins de paramètres ajustables, et donc *a priori* entraînables sur des bases de données plus restreintes qu'un réseau profond.

Le fonctionnement général de la méthode proposée est illustré en Figure 15.

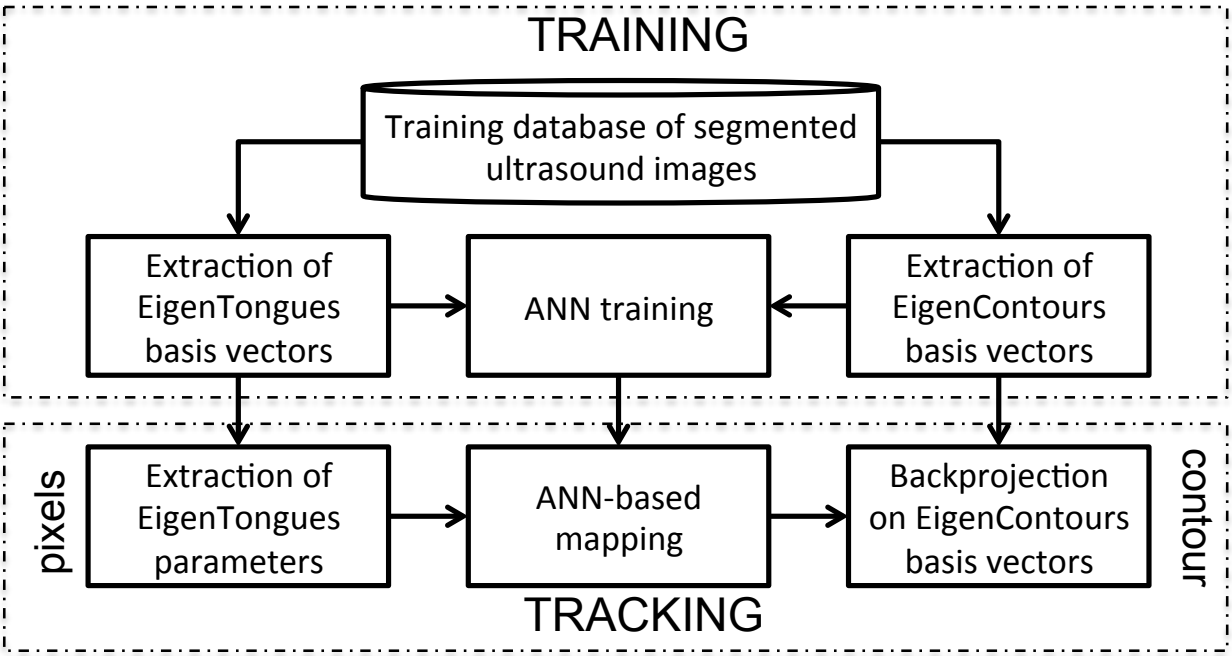


Figure 15. Schéma général de notre système de suivi de langue.

Durant la phase d'apprentissage (*training*), une base d'images échographiques annotées permet la construction d'un espace de projection EigenTongues et EigenContours. Les images et les contours sont projetés sur ces bases. Un réseau de neurones est ensuite entraîné pour modéliser la relation entre descripteurs EigenTongues et EigenContours, sur la base de ces images d'apprentissage. En phase de suivi (*tracking)*, un vecteur de coefficients EigenContours est estimé à l'aide du réseau à partir d'un vecteur de coefficients EigenTongues non vu pendant l'apprentissage. Le contour est ensuite reconstruit par rétro-projection sur la base des EigenContours. Notons ici que l'estimation du contour est réalisée trame à trame. Elle ne nécessite donc pas d'avoir des trames consécutives, et peut en outre tourner en temps-réel.

2.3. Protocole expérimental

Dans cette section, nous détaillons les corpus de données utilisés pour évaluer la méthode proposée ainsi que son implémentation pratique.

2.3.1. Corpus de données

2.3.1.1 Corpus MultiLoc

Les données échographiques et audio de ce corpus ont été enregistrées grâce au système Ultraspeech Hueber *et al.* (2008) (www.ultraspeech.com). Ultraspeech est un logiciel

s'interfaçant directement avec le *hardware* d'un échographe afin d'acquérir des séquences d'images à leurs résolutions spatiale et temporelle maximales, contrairement aux systèmes utilisant la sortie vidéo analogique des échographes qui brident l'acquisition à 30fps (*frames per second*) et introduisent des artéfacts dans les images. Ultraspeech permet l'acquisition synchrone des images avec un ou plusieurs signaux audio et un flux vidéo. Pour nos acquisitions, nous utilisons Ultraspeech avec l'échographe Terason T3000, une sonde micro-convexe (128 éléments répartis sur un angle de 140°). La bande de fréquences des ultrasons a été fixée à 3-5 MHz, et la profondeur maximale d'exploration à 7 cm : il s'agit d'un paramétrage classique pour imager la langue, également utilisé dans Hueber (2009).

Le corpus MultiLoc est constitué de données recueillies chez neuf locuteurs : quatre femmes et cinq hommes adultes de langue maternelle française. Les images ont été enregistrées à 60fps avec une résolution de 320×240 pixels (soit une résolution spatiale de 0,5 mm/pixel). Nous avons enregistré, pour chaque locuteur, les dix phrases de la première liste du corpus de Combescure (1981), détaillée dans la Table 1. Cette liste contient 234 phonèmes et comme les autres listes du corpus, elle est phonétiquement équilibrée : la fréquence d'apparition des phonèmes est similaire à celle observée dans la langue française. Le choix d'une de ces listes pour constituer notre corpus de données est donc lié à la variété phonétique et donc articulatoire qu'elle propose.

Table 1. Première liste de Combescure.

1. Il se garantira du froid avec ce bon capuchon.
2. Annie s'ennuie loin de mes parents.
3. Les deux camions se sont heurtés de face.
4. Un loup s'est jeté immédiatement sur la petite chèvre.
5. Dès que le tambour bat, les gens accourent.
6. Mon père m'a donné l'autorisation.
7. Vous poussez des cris de colère ?
8. Ce petit canard apprend à nager.
9. La voiture s'est arrêtée au feu rouge.
10. La vaisselle propre est mise sur l'évier.

Les phrases sont présentées sous forme orthographique à l'écran aux locuteurs par le logiciel Ultraspeech. Deux répétitions de la liste, soit vingt phrases, permettent d'obtenir environ 5000 images par locuteur. Des informations sur la composition du corpus MultiLoc sont présentées dans le Tableau 2. Nous y reportons des informations moyennes sur le

nombre d'images par locuteur et par phrase, ainsi que la durée moyenne par locuteur et par phrase.

Tableau 2. Informations sur la composition du corpus MultiLoc. Quantité totale et moyenne (par locuteur et par phrase) de phonèmes et d'images, et durée totale et moyenne des acquisitions (en secondes) pour l'ensemble des deux répétitions.

	Total	**Moyenne par locuteur**	**Moyenne par phrase**
Phonèmes	468	468	23,4
Images	44491	4943	247
Durée (s)	687,7	76,4	4,1

Afin de se placer dans les conditions d'une séance d'orthophonie, **le locuteur tenait lui-même la sonde pendant l'acquisition**. Nous n'utilisons donc aucun système de stabilisation de la sonde par rapport à la tête du locuteur, contrairement aux protocoles classiques de phonétique expérimentale (*e.g.* Stone & Davis (1995), Hueber *et al.* (2008)). La sonde n'étant pas fixée sous la mâchoire au cours de l'acquisition, elle pouvait donc se déplacer et faire varier l'angle d'inclinaison.

2.3.1.2 Corpus Slovaque

Ce corpus a été acquis à **l'***Institute of Phonetics and Speech Processing* (Munich, Allemagne) dans le cadre de la thèse de Lia Saki Bucar Shigemori, qui a généreusement accepté de partager de nombreuses séquences d'images annotées. Une partie est utilisée dans Shigemori, Pouplier *et al.* (2015). Le corpus Slovaque est constitué des productions de quatre locuteurs slovaques. Nous n'avons en notre possession que les images échographiques. Il est constitué de syllabes extraites de mots inclus dans des phrases porteuses. Ces phrases sont *« pozri aj ron mi* ***pNp****ap dal »* ou *« pozri aj ron mi pi****pLep*** *dal »* avec N = {e,l,r}, et L = {r,l}. Seule la syllabe marquée ici en gras est conservée et annotée manuellement dans ce corpus. Bien plus limité que le corpus MultiLoc en termes de variabilité, il présente cependant l'avantage d'avoir des images successives annotées (12800 en tout). Ce corpus est utilisé ici uniquement pour l'évaluation de notre méthode et non pour l'apprentissage du modèle neuronal utilisé pour le suivi.

2.3.1.3 Pré-traitement (recalage)

Afin de limiter la variabilité liée à la position de la sonde et entre les locuteurs, nous avons appliqué une technique dite de recalage d'images (ou *image registration*) sur le corpus MultiLoc. Nous avons choisi arbitrairement un des neuf locuteurs du corpus MultiLoc comme locuteur de référence. Nous avons sélectionné trois images par locuteur représentant les articulations vocaliques les plus éloignées : les phonèmes /a/, /i/ et /u/. Nous avons pris à chaque fois la première image correspondant au début de la phrase pour trois phrases de la Table 1 : la phrase 1 fournit le /i/ dans « il » ; la phrase 2 fournit le /a/ dans « Annie » ; la phrase 7 fournit le /u/ dans « vous », le phonème /v/ n'impliquant pas de mouvement de la langue.

Pour chacune de ces images, nous avons sélectionné cinq points d'intérêt, en essayant de retrouver, d'un locuteur à l'autre, les mêmes positions sur la langue compte tenu de sa forme. Il est évident que les caractéristiques de l'image échographique de la langue ne permettent pas de placer des points de repères, parfaitement équivalents d'un locuteur à l'autre. Une matrice de transformation affine, calculée à partir de ces points, a alors été appliquée aux images et aux contours annotés manuellement. Cette matrice est construite comme suit :

$$\begin{bmatrix} k cos\alpha & -k sin\alpha & 0 \\ k sin\alpha & k cos\alpha & 0 \\ Tx & Ty & 1 \end{bmatrix}$$

avec α = angle de la rotation, k = facteur de l'homothétie et Tx et Ty = coordonnées associées à la translation. Un résultat de cette transformation est illustré en Figure 16.

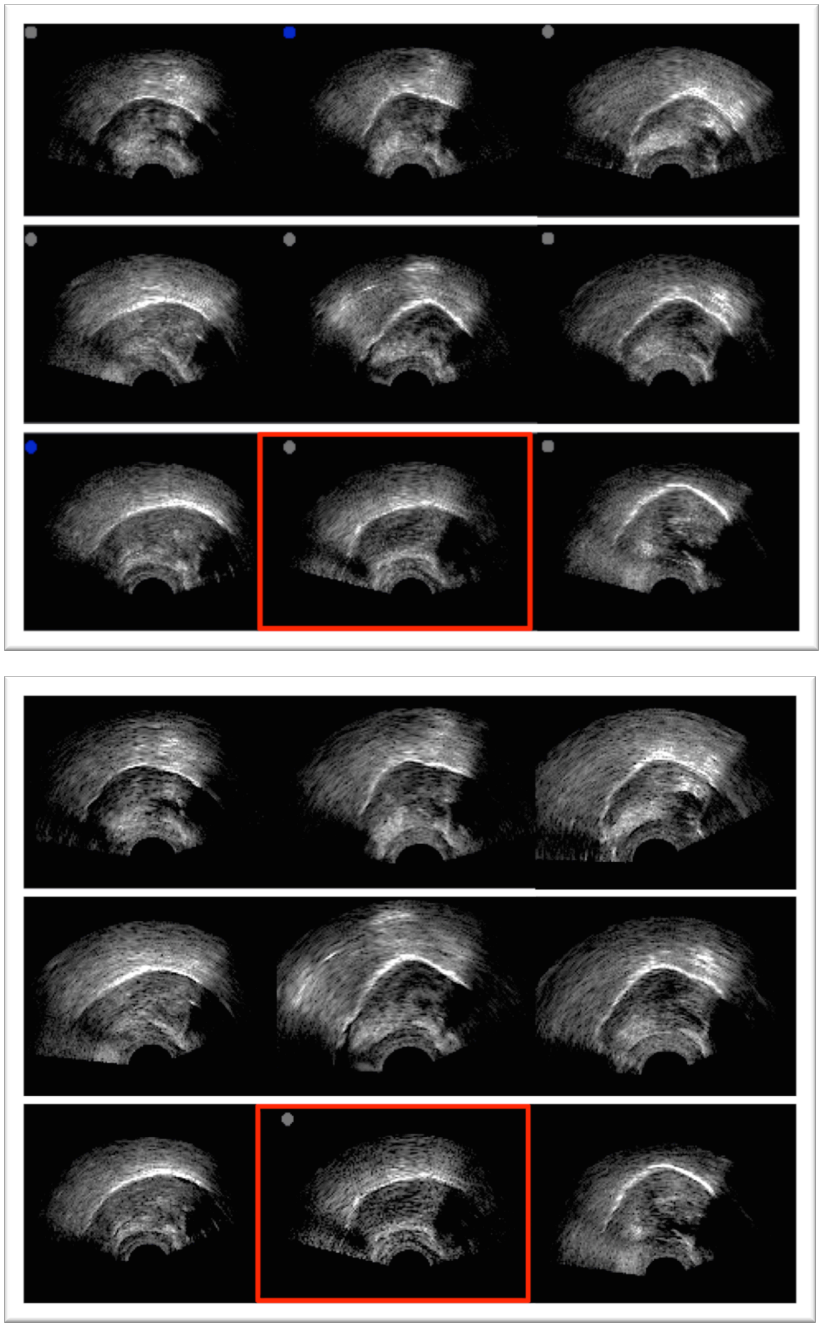


Figure 16 - Image du /a/ pour les 9 locuteurs avant et après recalage (locuteur de référence encadré en rouge).

2.3.2. Scénarios mono- et multi-locuteur proposés

Trois scénarios ont été mis en place afin d'évaluer la performance de la méthode proposée. A travers ces scénarios, nous proposons deux approches principales : une approche *mono-locuteur* pour laquelle nous construisons un modèle de suivi propre à un locuteur, et une approche *multi-locuteur*, pour laquelle ce modèle est estimé sur l'ensemble des locuteurs

disponibles dans la base d'apprentissage. Nous observons notamment la capacité de ce modèle multi-locuteur à généraliser à un nouveau locuteur. Pour chacun des scénarios, nous évaluons les performances en fonction de la quantité de données d'apprentissage, c'est-à-dire le nombre d'images annotées manuellement. Cela nous renseigne directement sur le niveau d'intervention humaine que requiert notre méthode.

2.3.2.1 Approche mono-locuteur

Dans ce scénario, les paramètres EigenTongues et EigenContours (base de projection) sont estimés indépendamment pour chacun des locuteurs de la base MultiLoc. La base EigenContours est construite à partir d'un sous-ensemble de N_0 images annotées manuellement (avec $N_0 \leq N$). Cette sélection s'effectue à l'aide de la procédure décrite à la section 2.3.6 : sélection des N_0 images les plus différentes parmi les N disponibles à l'aide de l'algorithme K-means (K=20). Ensuite, la base EigenTongues est construite à partir de l'ensemble des N images disponibles pour chacun des locuteurs. Cette étape ne nécessite aucune annotation, la sonde échographique permettant d'acquérir facilement des images sur des locuteurs.

2.3.2.2 Approche mono-locuteur avec locuteur de référence

Le locuteur de référence choisi pour le recalage des images peut être utilisé pour compléter les données acquises sur le locuteur courant. En effet, notre hypothèse est que le recalage effectué (section 2.3.1.3) est suffisamment précis pour que les données du locuteur de référence puissent améliorer le modèle construit sur le locuteur courant.
Nous construisons donc l'espace EigenTongue sur N images du locuteur de référence et N images du locuteur courant. La performance est évaluée en fonction de la quantité de données d'apprentissage, avec cette fois un nombre N_0 de contours du locuteur courant + N contours du locuteur de référence, utilisés pour construire l'espace EigenContours et entraîner le modèle neuronal. Par souci de simplification, nous nommerons par la suite cette approche mono-locuteur+réf.

2.3.2.3 Approches multi-locuteur

Dans ce scénario, nous exploitons un modèle pré-entraîné sur des données d'autres locuteurs que le locuteur que nous voulons traiter (nous appellerons ce dernier le locuteur courant). Ce modèle peut soit être appliqué directement pour segmenter des images du locuteur courant, soit être adapté à ce dernier. Le processus d'adaptation prend simplement la forme d'un réapprentissage des paramètres du modèle neuronal sur une base

de données contenant un ensemble limité de N_0 images du locuteur courant, ainsi qu'un plus grand nombre de N images d'autres locuteurs (avec typiquement N> N_0).

2.3.3. Choix du nombre de composantes pour le paramétrage des images et des contours

En Figure 17, nous pouvons observer, pour le corpus Multiloc, l'évolution de la variance cumulée en fonction du nombre de vecteurs EigenTongues utilisés. Les courbes en rouge représentent les valeurs obtenues pour chacun des locuteurs, et les courbes en bleu pour un corpus incluant les neufs locuteurs de la base MultiLoc (sans recalage en trait pointillé, avec recalage en trait continu). Avec 40 vecteurs, il est possible de représenter 90 % de la variance des données pour chaque locuteur, contre plus de 90 vecteurs sur l'ensemble de la base MultiLoc.

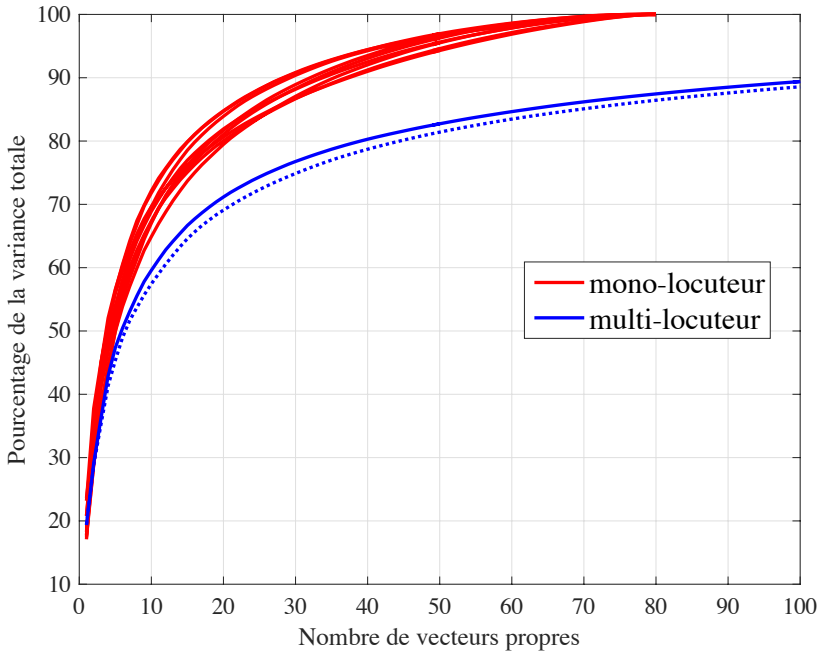


Figure 17. Evolution de la variance en fonction du nombre de vecteurs propres Eigentongues. En rouge, les variations pour chaque locuteur de la base MultiLoc. En bleu, les variations pour un corpus incluant l'ensemble de la base MultiLoc (sans recalage en trait pointillé, avec recalage en trait continu).

En Figure 18, nous pouvons observer l'évolution de la variance cumulée en fonction du nombre de vecteurs EigenContours utilisés. Le nombre de vecteurs nécessaires pour

atteindre 90% de la variance est très inférieur à celui observé pour les EigenTongues. Les quatre premiers vecteurs contiennent plus de 90% de la variance en mono-locuteur comme en multi-locuteur.

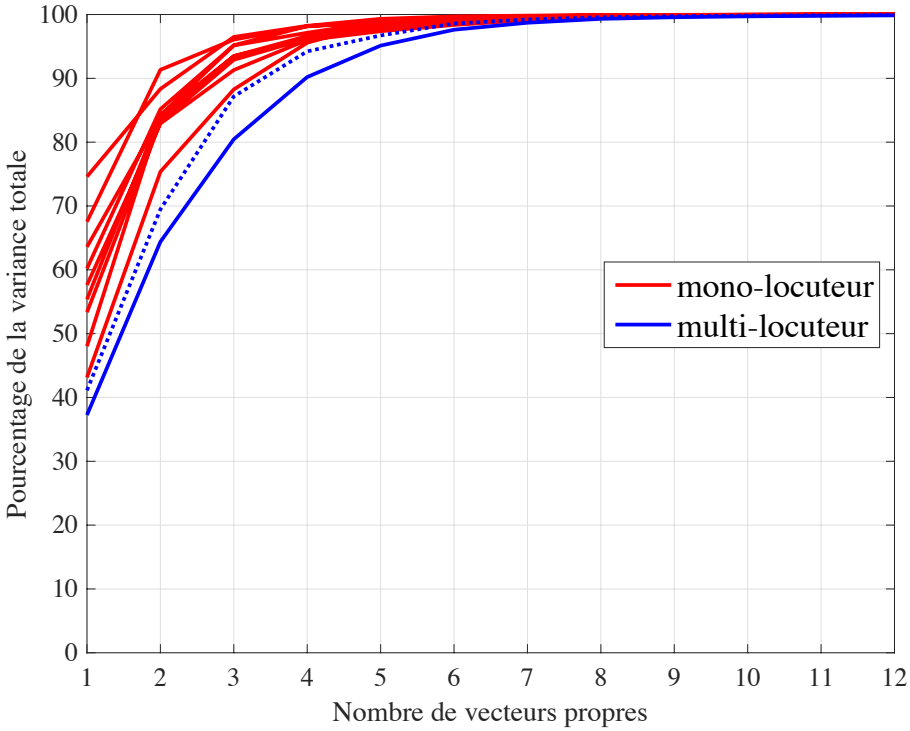


Figure 18. Evolution de la variance en fonction du nombre de vecteurs propres EigenContours. En rouge, les variations pour chaque locuteur de la base MultiLoc. En bleu, les variations pour un corpus incluant l'ensemble de la base MultiLoc (sans recalage en trait pointillé, avec recalage en trait continu).

Nous avons évalué la capacité de reconstruction des vecteurs EigenContours. Nous avons constitué une base de vecteurs par locuteur, sur l'ensemble des contours disponibles. Nous avons ensuite évalué l'erreur de reconstruction des contours, tout d'abord en transformant les contours en paramètres EigenContours, puis en réalisant l'opération inverse. L'erreur de reconstruction a été mesurée entre le contour ainsi reconstruit et le contour annoté manuellement. Cette erreur, calculée sur la moyenne des performances des neuf locuteurs, s'élève à 0,28 mm et correspond donc à la meilleure performance possible de nos modèles.

Pour le reste des expériences, nous décrirons les images ultrasonores par 40 coefficients EigenTongues et 8 coefficients EigenContours pour les scénarios mono-locuteur et mono-

locuteur+réf. Pour le scénario multi-locuteur, nous utiliserons 80 coefficients EigenTongues et 12 coefficients EigenContours. Dans les 3 cas, nous utilisons une représentation qui explique au moins 90% de la variance observée dans la base de données MultiLoc.

2.3.4. Apprentissage du réseau de neurones

La mise en correspondance de ces observations est effectuée par un MLP (voir Figure 14), à une seule couche cachée. Le nombre de neurones cachés est déterminé sur un corpus de validation, c'est-à-dire un corpus indépendant du corpus d'apprentissage (utilisé pour l'estimation des poids) et du corpus de test. Lors de l'apprentissage, les poids sont initialisés aléatoirement par une distribution gaussienne avec un écart-type de 0,0001. Le critère d'erreur est défini comme l'erreur quadratique moyenne entre les valeurs prédites et réelles. La minimisation de cette erreur est réalisée avec la méthode du gradient conjugué sur des lots successifs (mini-batch) : à chaque itération, les échantillons de données d'apprentissage sont mélangés aléatoirement. La fonction d'activation est la fonction sigmoïde. Les entrées et les sorties sont centrées réduites (*z-score*) avant d'être fournies au réseau. La régularisation du réseau de neurones durant l'apprentissage est réalisée d'une part avec une stratégie de *early-stopping*, arrêtant l'apprentissage lorsque l'erreur sur le corpus de validation se maintient ou augmente à nouveau, et d'autre part en ajoutant une pénalité aux poids du réseau lors de la propagation de l'erreur.

Le nombre de neurones de la couche cachée est fixé par validation croisée, sur 20% des données d'apprentissage (non utilisées pour l'estimation des poids). **Pour les scénarios mono-locuteur et mono-locuteur+réf, cette valeur est fixée à 35 neurones, et s'élève à 50 pour le scénario multi-locuteur.**

2.3.5. Métrique de comparaison de contours

Pour évaluer la qualité du contour estimé, nous utilisons la mesure MSD (Mean Sum of Distance) définie dans Li *et al.* (2005) comme suit :

$$MSD_k = \frac{1}{2N}\sum_{i=1}^{N}(min_j(v_i \rightarrow u_j) + min_j(u_i \rightarrow v_j))$$

où $u_i \rightarrow v_j$ indique la distance euclidienne entre le $i^{ème}$ point u_i du contour estimé et le $j^{ème}$ point v_j du contour manuellement annoté (N est le nombre de lignes dans la grille d'annotation). Cette mesure correspond donc, sur l'ensemble des points d'un contour, à la

moyenne de la distance de chaque point estimé au contour réel, cumulée à la distance de chaque point réel au contour estimé.
Cette mesure permet de calculer une distance entre deux contours pour lesquels nous ne pouvons pas mettre en correspondance les coordonnées des points qui le définissent, mais uniquement la forme générale. Cette métrique est donc bien adaptée à l'image ultrasonore de la langue, pour laquelle nous ne disposons pas de points caractéristiques (ou points de chair). La Figure 19 illustre un cas où l'on souhaite mesurer la distance entre le contour estimé (en bleu) et le contour réel. Ainsi, à l'extrémité, lorsque la distance est mesurée point à point, elle est bien plus importante que la distance point à contour, alors même que le point concerné respecte bien la forme générale de la langue.

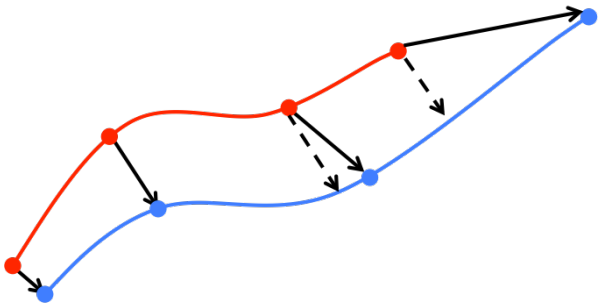


Figure 19. En rouge, le contour réel. En bleu, le contour estimé. Les flèches continues représentent la distance point à point, les flèches en pointillé la distance point à contour.

L'erreur moyenne en millimètres pour K images est définie par

$$MSD_{RMS} = R\sqrt{\frac{1}{K}\sum_{k=1}^{K} MSD_k^2}$$

où R est la résolution de l'image fournie par le système ultrasonore (R=0.5mm/pixel dans ce cas).

2.3.6. Choix des corpus d'apprentissage et de test

La base MultiLoc contient environ 5000 images pour chacun des neuf locuteurs de la base. Afin d'éviter la segmentation manuelle de ces 45000 images, nous ne travaillons que sur un sous-ensemble d'images. Ce sous-ensemble est obtenu en sélectionnant, pour chacun des 9 locuteurs, les 100 images les plus différentes. Nous annotons ainsi 900 images au lieu de 45000. Ces images sont extraites automatiquement à l'aide de l'algorithme K-means. Nous avons déterminé un nombre adapté de clusters (*i.e.* K) en nous basant sur les

caractéristiques phonétiques des différentes articulations et coarticulations présentes dans la langue française. Ce nombre a été fixé empiriquement à 20 clusters pour l'ensemble des expériences réalisées.

Nous étudions la performance du modèle en fonction de la taille du corpus d'apprentissage, c'est-à-dire du niveau d'intervention manuelle nécessaire. Rappelons que nous cherchons la méthode la plus automatique possible pour limiter l'intervention du praticien. Pour chaque locuteur, nous faisons donc varier la taille du corpus d'apprentissage que nous notons $\mathbf{N_0}$ pour $\mathbf{N_0}$ compris entre 0 et N=80. Le choix de ces $\mathbf{N_0}$ images les plus éloignées parmi les N images s'effectue également à l'aide de l'algorithme K-means avec K=20 comme défini précédemment. Toutes les expériences ont été systématiquement réalisées en utilisant une stratégie classique de validation croisée avec cinq partitions de N/5 images.

2.4. Résultats expérimentaux

2.4.1. Performances des modèles sur le Corpus MultiLoc

Les résultats obtenus sur le Corpus MultiLoc sont reportés en Figure 20. Les trois scénarios sont représentés. Chaque point représente la MSD moyenne sur l'ensemble des images. Nous avons choisi d'afficher les moyennes afin d'ajouter l'intervalle de confiance associé. Le Tableau 3 répertorie quant à lui les valeurs de RMSE des MSD obtenues pour chacune des expériences.

Tableau 3. Tableau récapitulatif des performances obtenues par les différents scénarios (RMSE$_{MSD}$ en mm).

RMSE$_{MSD}$	0	20	40	60	80
mono-locuteur		2,92	2,15	1,68	1,28
mono-locuteur+réf	4,07	2,08	1,70	1,54	2,61
multi-locuteur (sans recalage)	4,09	2,35	2,13	1,89	1,95
multi-locuteur	3,11	2,29	2,04	1,89	1,95

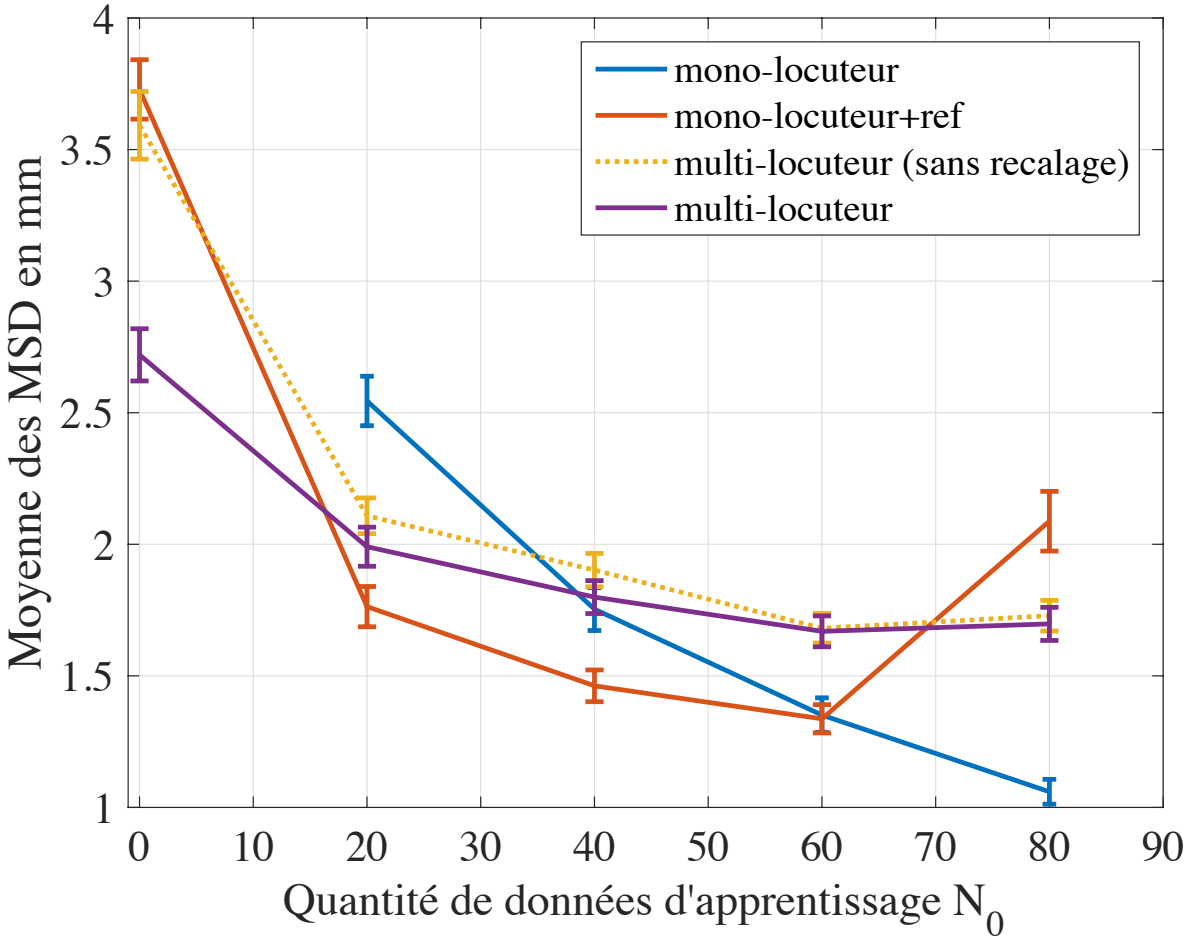


Figure 20. Moyenne des MSD en fonction de la quantité d'apprentissage pour les différents scénarios, avec leurs intervalles de confiance. En abscisse, le nombre d'images annotées du locuteur courant. En ordonnée, la RMSE de la MSD en mm. Les barres d'erreur représentent les intervalles de confiance (t-test avec $\alpha = 0{,}05$).

La tendance globale est la même quel que soit le scénario réalisé. Plus la quantité de données connues pour le locuteur courant augmente et meilleure est la performance du modèle. Cependant, nous pouvons observer des variations importantes entre les divers scénarios.

Commençons par le scénario **mono-locuteur**. De façon attendue, la meilleure performance est en effet obtenue lorsqu'on utilise le plus gros corpus d'apprentissage possible. Avec 80 images, nous obtenons une erreur de 1,28 mm, contre 2,92 mm avec seulement 20 images. La performance du modèle avec 60 images (1,68 mm) peut sembler

un bon compromis, qui sera cependant laissé à l'appréciation du praticien en fonction de la précision et du temps d'intervention manuelle disponible.

Intéressons-nous maintenant au scénario **mono-locuteur+réf.** A la différence du scénario **mono-locuteur**, il est possible de réaliser une estimation du contour même lorsqu'aucune image du locuteur courant n'est annotée. Cependant, cette performance est supérieure à 4 mm d'erreur, ce que nous considérons comme une précision trop faible pour l'orthophoniste : cette différence représente l'écart entre deux articulations linguales comme le /s/ et le /l/. Cependant, jusqu'à 60 images manuellement annotées, ce modèle s'avère plus performant que le modèle **mono-locuteur**, avec dans ce cas 1,54 mm d'erreur contre 1,68 pour le scénario **mono-locuteur**. L'erreur d'estimation diminue progressivement jusqu'à ce seuil, pour ensuite augmenter à nouveau. Les données du locuteur de référence semblent donc aider à l'estimation d'un contour cohérent pour le locuteur courant dont on possède peu de données. Cependant, elles semblent dégrader les performances de cette estimation lorsqu'on a assez de données sur le locuteur courant, ce qui peut être dû à la variabilité anatomique inter-locuteur.

Orientons maintenant nos observations vers le scénario multi-locuteur. L'influence positive du recalage d'images sur l'estimation du contour est perceptible au premier coup d'œil. Sans aucune annotation du locuteur courant, un modèle construit sur huit locuteurs, recalés sur un locuteur de référence, présente une erreur d'estimation de 3,11 mm, contre 4,09 mm sans recalage. Par contre, à partir de 60 images, les performances du modèle sans recalage rejoignent celles du modèle avec recalage pour les égaler. Comme pour le modèle mono-locuteur, le modèle multi-locuteur présente les meilleures performances avec le plus grand nombre de données disponibles (80 images par locuteur), même si aucune différence significative n'est observée en utilisant seulement 60 images (1,95 mm contre 1,89 mm). Cependant, avec ou sans recalage, les modèles multi-locuteur sont surpassés par le modèle mono-locuteur autour de 40 images annotées.

Il est intéressant d'observer que des informations dérivant des autres locuteurs semblent dégrader l'estimation au lieu de l'améliorer, lorsqu'on a suffisamment de données du locuteur courant. Ce constat expliquerait la brusque diminution de performance du modèle **mono-locuteur+réf** au-dessus de 60 images. L'effet opposé est observé avec moins de 40 images par locuteur, où le modèle multi-locuteur surpasse significativement le modèle mono-locuteur. Dans ce cas, le modèle semble extrapoler les informations des données des autres locuteurs. Pour le scénario **mono-locuteur+réf**, cet effet est observé dès 20 images où les estimations du modèle sont supérieures à celles du modèle **multi-locuteur**.

Considérons maintenant l'erreur obtenue pour chaque point des contours. Un contour est constitué de P points, ordonnés de l'arrière à l'avant de la langue. La Figure 21 illustre les résultats obtenus pour le modèle multi-locuteur, qui sont comparables à ceux observés pour les systèmes mono-locuteurs. Comme nous pouvons le constater, les erreurs les plus importantes se trouvent aux extrémités du contour de la langue, alors que toute la partie centrale présente des meilleurs résultats. Rappelons que ces dernières sont souvent mal-imagées en raison des ombres acoustiques de la mâchoire et de l'os hyoïde, qui rendent la segmentation difficile.

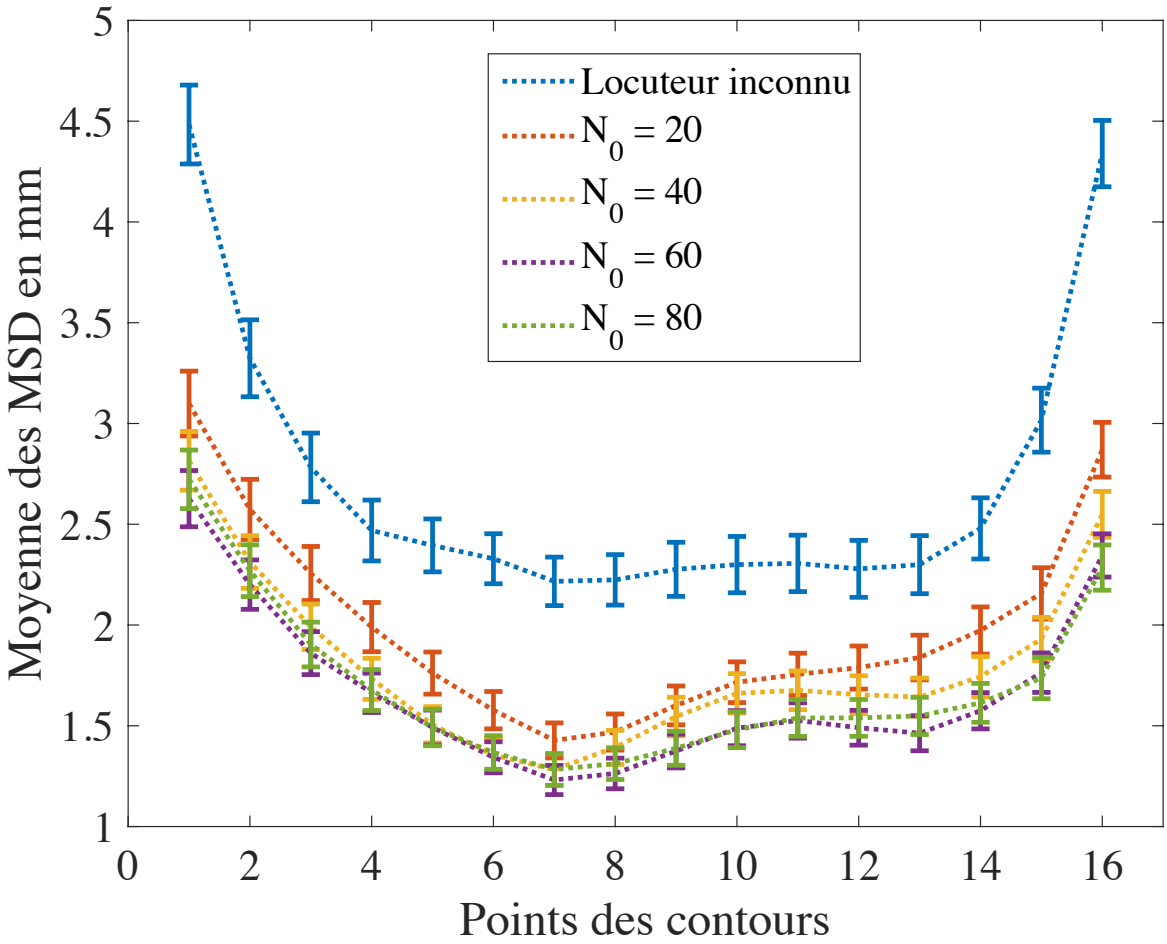


Figure 21. Erreur d'estimation des P points (ici P = 16) pour les différentes configurations du système multi-locuteur. Les points sont ordonnés de l'arrière à l'avant de la langue.

Nous avons donc pu comparer les différents modèles entre eux à partir des erreurs d'estimation obtenues. Mais gardons à l'esprit notre hypothèse initiale : utiliser l'image complète lors de la recherche du contour de langue devrait améliorer son suivi, en

exploitant les informations fournies entre autres par le *speckle*. Il nous faut maintenant observer les résultats obtenus sur des images, en particulier lorsqu'elles sont de très mauvaise qualité, ou lorsqu'une partie du contour est manquante. La Figure 22 illustre les résultats obtenus pour chacun des modèles, pour différents locuteurs et différentes articulations, dans leurs conditions optimales : 80 images pour le système **mono-locuteur,** 40 images pour le système **mono-locuteur+réf** et aucune image pour le système **multi-locuteur**.

Chaque colonne correspond, respectivement, à l'image brute, l'annotation manuelle, puis l'estimation par le modèle mono-locuteur, mono-locuteur+réf et enfin multi-locuteur. Notons que pour les colonnes mono-locuteur+réf et multi-locuteur, les images ont été recalées sur le locuteur de référence, contrairement aux trois autres images d'une même ligne. L'exemple I9 (Figure 13) est une image du locuteur de référence, pour lequel le modèle mono-locuteur+réf n'a évidemment pas été construit.

L'exemple I1 est une illustration d'estimation lorsque le contour est parfaitement visible sur l'image. Dans ce cas, les performances des modèles sont globalement très bonnes, avec de très légères variations observables par endroit. Les autres lignes illustrent des positions extrêmes de la langue, ou des exemples où une partie de la langue est totalement absente. Nous pouvons constater que les performances du modèle multi-locuteur sont très variables en fonction de l'image observée. Ainsi, pour l'exemple I3 comme pour l'exemple I7, l'estimation proposée n'est absolument pas pertinente, contrairement à l'exemple I1. Cependant, l'obstacle rencontré par ce modèle ne semble pas être le manque d'information visible, mais plutôt les différences anatomiques entre le locuteur courant et les locuteurs connus du modèle. Le modèle mono-locuteur est très proche de l'annotation manuelle et semble suffisamment pertinent dans le cadre de la rééducation orthophonique. Enfin, le modèle mono-locuteur+réf, bien que moins précis, semble offrir une performance suffisante pour compléter le contour de la langue.

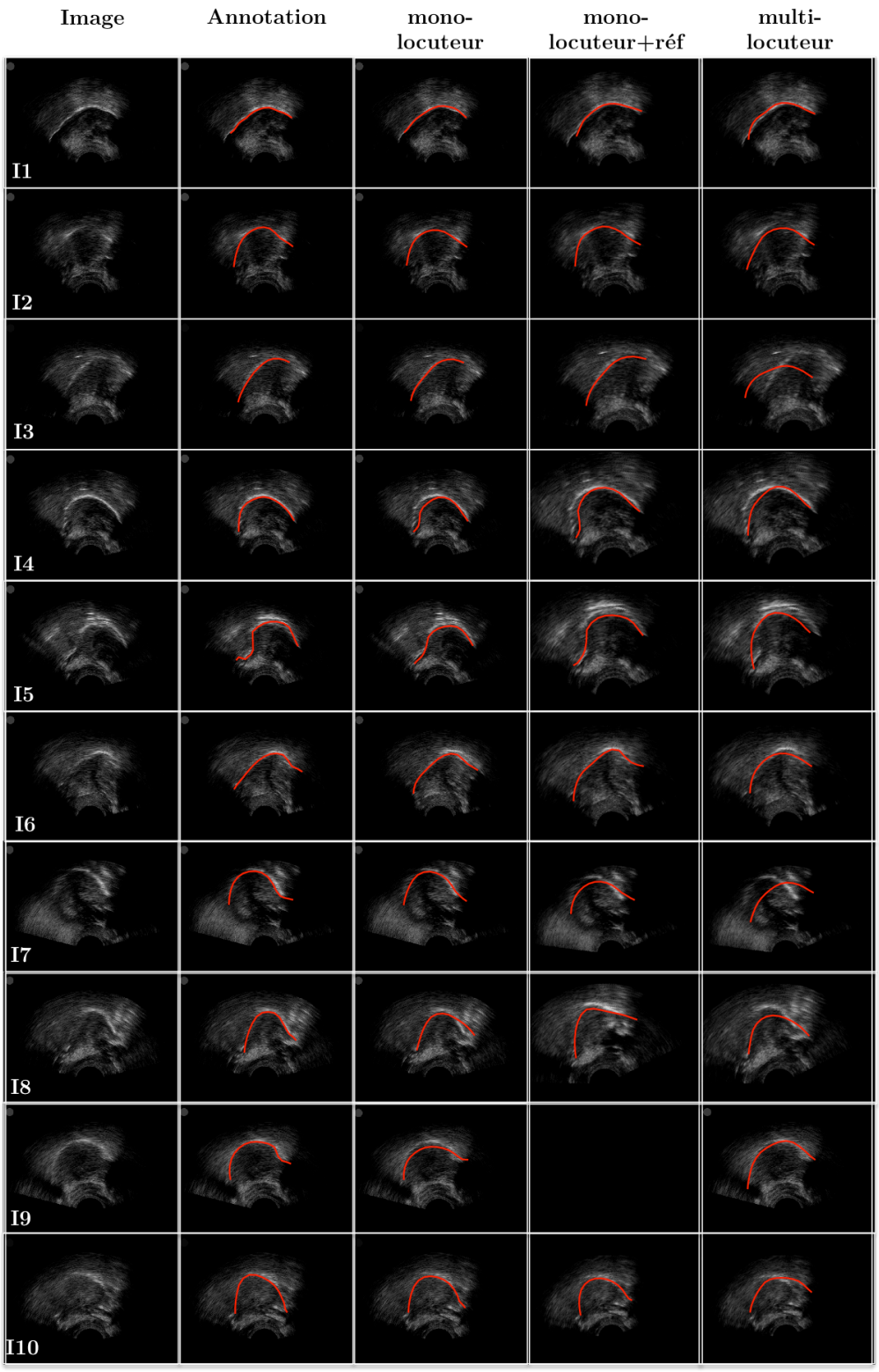


Figure 22. Exemples de contours estimés pour différentes configurations de la langue, différents locuteurs et différents scénarios.

2.4.2. Comparaison avec la méthode état de l'art EdgeTrak

Nous avons confronté les performances de notre méthode avec celles du logiciel EdgeTrak Li *et al.* (2005) présenté en section 2.1.1. Nous avons utilisé le corpus Slovaque à notre disposition pour cette évaluation (présenté à la section 2.3.1.2). Ce corpus est composé de phrases entièrement annotées, et donc d'images successives, indispensable pour EdgeTrak qui utilise l'initialisation manuelle sur la première image d'une séquence pour estimer le contour des images suivantes, chaque image étant ainsi initialisée avec le contour de l'image d'avant. Les annotations manuelles réalisées sur le corpus Slovaque serviront de vérité de terrain pour comparer les performances des différentes techniques.
Pour chaque locuteur de ce corpus, une trentaine de phrases, pour un total de 600 images environ, sont écartées pour réaliser le test. Parmi les phrases restantes, 100 images sont sélectionnées (cf. 2.3.5) pour construire notre modèle mono-locuteur.

Notons qu'EdgeTrak ne requiert aucune phase d'apprentissage, mais que la première image de chaque séquence évaluée doit être annotée manuellement. Pour chaque séquence, nous avons réajusté cette annotation manuelle par EdgeTrak pour limiter un biais éventuel. La Figure 23 rapporte les résultats obtenus pour les quatre locuteurs du corpus Slovaque, composé de séquences d'images pour les trois articulations {e,l,r}, ainsi que pour un locuteur non inclus dans le corpus MultiLoc et noté ici MG pour lequel nous avons à notre disposition huit phrases entièrement annotées. Pour ce dernier, nous avons sélectionné pour l'apprentissage 100 images comme décrit en section 2.3.6 parmi quatre phrases (soit environ 800 images), les quatre autres phrases ont été utilisées pour le test (soit 739 images). Ces résultats sont représentés sous forme de *boxplots*. Pour un ensemble de points, une *boxplot*, ou boîte à moustache, se compose d'une marque centrale en rouge, indiquant la médiane. Le haut et le bas de la boîte indiquent respectivement le 25^{e} et le 75^{e} centile. Les moustaches intègrent tous les points non considérés comme des points aberrants, qui sont pour leur part représentés avec des croix.

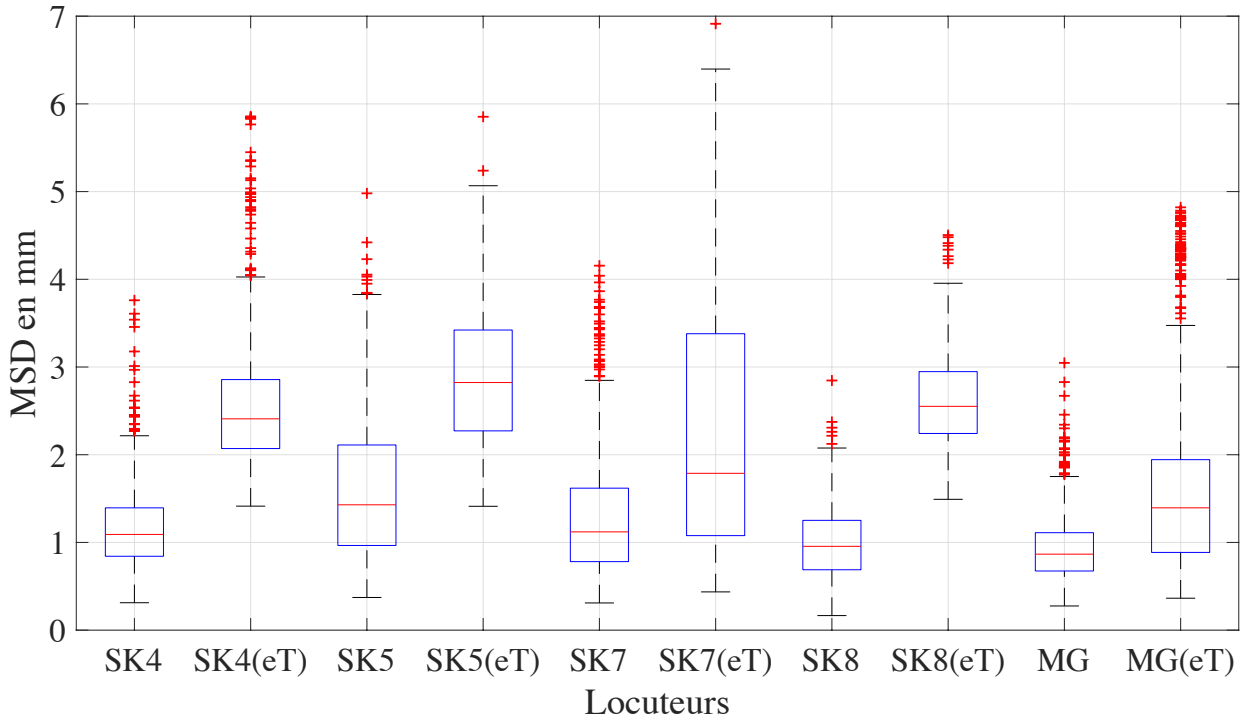


Figure 23. Résultats du suivi de contour par notre méthode et par EdgeTrak (noté eT), pour les locuteurs SK4, SK5, SK7, et SK8 de la base de données « Slovaque » et pour le locuteur MG.

Le Tableau 4 présente ces résultats en termes de RMSE de la MSD pour les mêmes locuteurs.

Tableau 4. $RMSE_{MSD}$ (en mm) pour les quatre locuteurs du corpus Slovaque et le locuteur MG.

$RMSE_{MSD}$ (mm)	**SK4**	**SK5**	**SK7**	**SK8**	**MG**
Notre méthode	1,26	1,81	1,49	1,08	1,00
EdgeTrak	2,67	3,05	2,75	2,66	1,98

Pour l'ensemble des locuteurs du corpus Slovaque, notre méthode fournit donc des meilleurs résultats que EdgeTrak, pour les 4 locuteurs du corpus Slovaque et pour le locuteur MG de notre base de données MultiLoc (avec une erreur parfois presque deux fois moins importante). Nous pouvons de plus observer en Figure 24, image par image, la différence entre la valeur de la MSD obtenue avec notre méthode et celle obtenue avec EdgeTrak, pour le locuteur MG. En général, EdgeTrak présente des performances très proches de notre méthode (différence proche de zéro) au début de la phrase. Cependant, quand la détection du contour s'éloigne de la réalité, l'estimation devient vite très

mauvaise et la différence augmente drastiquement, comme nous pouvons le voir pour la troisième phrase.

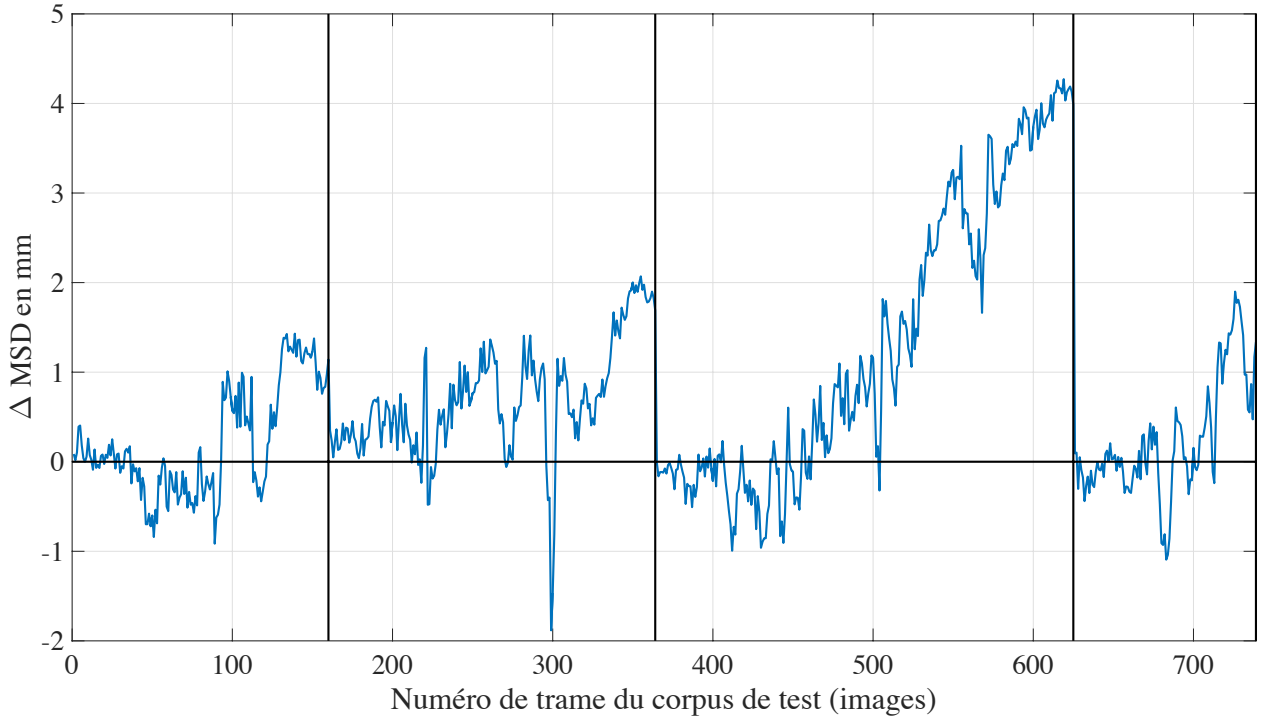


Figure 24. Différence entre l'estimation de notre méthode et celle d'EdgeTrak pour les quatre phrases du locuteur MG. Plus la différence est importante et plus le contour fourni par notre méthode peut être jugé plus précis que celui fourni par EdgeTrak. Les barres verticales séparent les phrases.

Nous voyons donc sur cette évaluation que notre méthode de suivi de contour fournit de meilleurs résultats qu'EdgeTrak. Les performances obtenues par notre modèle sur ces corpus étant similaires à celles observées sur le corpus MultiLoc, nous pouvons donc les comparer. Rappelons cependant que cette performance (erreur de 2 mm ou moins) s'atteint au prix d'un étiquetage d'au moins 40 images en moyenne pour le scenario mono-locuteur et 20 images en moyenne pour le scénario multi-locuteur.

2.5. Conclusions et perspectives

Dans ce chapitre, nous avons présenté une méthode pour le suivi de contour de langue sur des images échographiques. Dans la lignée des travaux de Fasel & Berry (2010), nous avons modélisé la relation entre l'intensité des pixels de l'ensemble de l'image

échographique et le contour de la langue à l'aide d'un réseau de neurones artificiel (ANN). Guidée par notre contexte applicatif, notre approche vise à trouver le meilleur compromis entre niveau d'intervention manuelle d'une part, et qualité du suivi d'autre part. En s'appuyant sur un modèle pré-entraîné sur une base multi-locuteur, nous avons montré qu'une performance satisfaisante pour un système de retour visuel et meilleure que la méthode état de l'art EdgeTrak, pouvait être atteinte dès 20 images annotées manuellement, soit une intervention manuelle inférieure à 3 minutes. Nous pensons donc que la méthode proposée est une bonne candidate pour un système de retour visuel par échographie augmentée.

Plusieurs améliorations peuvent cependant être apportées à notre approche. Tout d'abord, pour chaque évaluation, nous comparons la différence entre le contour estimé et le contour dit réel. Or, ce contour réel est annoté par des non-experts et n'a pas été évalué. Il se peut que, dans certains cas, le système estime correctement le contour alors que la personne ayant annoté s'est trompée. Cette situation peut être rencontrée dans un contexte clinique, où l'orthophoniste n'est pas forcément expert, et cela constitue un problème pour notre évaluation. Il faudrait donc être en mesure d'évaluer l'annotation proposée en croisant différentes propositions. Nous pourrions aussi utiliser une autre modalité (comme l'EMA) pour connaître le contour exact de la langue et proposer un système multi-locuteur plus précis. Ces données pourraient également aider à l'estimation des points manquants à la place de l'algorithme p-PCA (Porta *et al.* (2005)) que nous avons utilisé.

Nous pouvons aussi discuter du recalage sur un locuteur de référence. Les différences anatomiques entre les locuteurs sont parfois trop importantes pour que la matrice de transformation puisse recaler correctement un locuteur sur ce locuteur de référence. De plus, la sonde échographique n'était pas fixée lors de l'acquisition des données, donc les trois images sélectionnées pour le recalage ne reflètent pas forcément l'inclinaison globale des images. Cependant, malgré toutes ces limites, le recalage des locuteurs sur un locuteur de référence améliore les résultats obtenus, surtout lorsque le locuteur courant est inconnu du modèle, en permettant de réduire la variance dans les données.

Au vu des différences morphologiques existant entre les hommes et les femmes, il pourrait être pertinent de créer deux modèles distincts afin d'augmenter encore la précision de l'estimation. Nous pourrions aussi réaliser un modèle pour les enfants, à différents âges, en utilisant des corpus comme celui de Barbier, Perrier *et al.* (2015).

Le système présenté dans ce chapitre peut fonctionner en temps-réel. Cependant, nous pourrions profiter du fait que les données exploitées dans le cadre de la rééducation

orthophonique sont forcément des images consécutives pour ajouter une information temporelle à nos données d'entrée. Une architecture de type *Recurrent Neural Network* (Pineda (1987)) pourrait être profitable pour modéliser explicitement la structure temporelle des mouvements linguaux.

Chapitre 3. L'échographie linguale augmentée : animation d'un modèle de langue

Dans le chapitre précédent, nous avons développé une méthode permettant d'augmenter une image échographique linguale en rendant plus visible le contour de la surface supérieure de la langue. Cependant, les autres structures de la cavité buccale, comme le palais, les dents ou le pharynx, restent absentes de l'image obtenue. En outre, seule une partie de la langue est reconstituée. Dans ce chapitre, nous proposons donc une méthode permettant de faire apparaître ces structures, en faisant l'hypothèse que ces ajouts amélioreraient la compréhension de l'image pour un locuteur non-expert. Pour cela, nous souhaitons animer automatiquement, à partir d'images échographiques brutes, le modèle de langue de la tête parlante articulatoire développé au GIPSA-lab par Badin *et al.* (2008). Nous proposons une approche par apprentissage statistique supervisé, basée sur la modélisation par mélange de gaussiennes (*Gaussian Mixture Model* ou *GMM*). Toujours dans l'optique de développer des systèmes transférables à une utilisation clinique, nous avons fixé deux contraintes :

- Trouver un bon compromis entre performance et quantité de données d'apprentissage, ce qui détermine directement le délai à partir duquel l'utilisateur peut se servir du système
- Concevoir un système capable de s'adapter aux progrès d'un patient notamment en généralisant correctement à des configurations articulatoires non vues pendant l'apprentissage.

Pour ce faire, notre système s'appuie sur la technique *Cascaded-Gaussian Mixture Regression*, récemment proposée par Hueber *et al.* (2015), que nous décrirons en détail dans la suite.

3.1. Etat de l'art

L'utilisation d'une tête parlante comme retour visuel est proposée par Engwall & Bälter (2007) dans le contexte de l'apprentissage d'une L2. Afin d'enseigner à des locuteurs français la prononciation de phonèmes suédois qui n'existent pas dans leur langue maternelle, une tête parlante développée à partir d'un locuteur suédois est utilisée dans un

paradigme de type Magicien d'Oz : plutôt que d'animer la langue de la tête parlante en fonction de l'articulation de la langue de façon automatique, un phonéticien expert choisit, à chaque tentative de l'apprenant, une séquence d'animation pré-enregistrée parmi dix séquences de cette articulation, correspondant au plus près à l'articulation qu'il estime à partir de la production acoustique du locuteur. Les productions échographiques sont enregistrées pour une analyse ultérieure. Il ne s'agit donc pas d'un *retour visuel* via une tête parlante, à partir d'échographies du locuteur, mais plutôt d'une *illustration* de l'articulation réalisée et de l'articulation souhaitée. Engwall (2012) compare ensuite cette méthode à un retour par inversion acoustico-articulatoire, et souligne le besoin d'un modèle articulatoire adapté aux particularités du locuteur.

Roxburgh, Scobbie *et al.* (2015) réalisent une étude sur deux patients présentant une fente palatine. La rééducation proposée se découpe en deux temps : une première étape avec illustration par une tête parlante, suivi d'une deuxième étape où le patient visualise sa propre articulation grâce à un dispositif échographique. Les résultats montrent une plus grande progression avec la tête parlante, ce qui peut suggérer que les patients sont plus réactifs à ce mode de représentation. Cleland *et al.* (2013) et Roxburgh *et al.* (2015) font donc l'hypothèse qu'un retour visuel par une tête parlante, plus complet et plus facile à interpréter, pourrait être bénéfique pour la rééducation orthophonique.
Dans la continuité de ces études, nous proposons un système de retour visuel basé sur l'animation d'une tête parlante à partir de l'image échographique. Nous avons donc développé un algorithme permettant **l'animation automatique de la langue** de cette tête parlante à partir d'images échographiques brutes de n'importe quel locuteur. Nous représentons donc ici le mouvement de la langue d'un locuteur dans l'espace articulatoire de la tête parlante. Notons que nous n'avons pas ici pour objectif de modifier la géométrie de la tête parlante pour l'adapter à la géométrie du conduit vocal de l'utilisateur.

3.2. Méthode proposée

3.2.1. Principe général

Le système proposé utilise la tête parlante articulatoire développée au Gipsa-Lab par Badin *et al.* (2008), illustrée en Figure 25. Comme les autres têtes parlantes présentées au Chapitre 1, la tête parlante du GIPSA-lab intègre un modèle géométrique de la langue construit à partir de données statiques et cinématiques enregistrées sur un locuteur, que nous appellerons dans la suite de ce chapitre *locuteur de référence*. Elle se compose d'un jeu de modèles articulatoires 3D des différents articulateurs de la parole. Les modèles de la langue, du velum, et de la mâchoire ont été construits à partir de scans 3D IRM statiques

du locuteur de *référence*, tandis que ceux des lèvres l'ont été à partir de données vidéo stéréoscopiques.

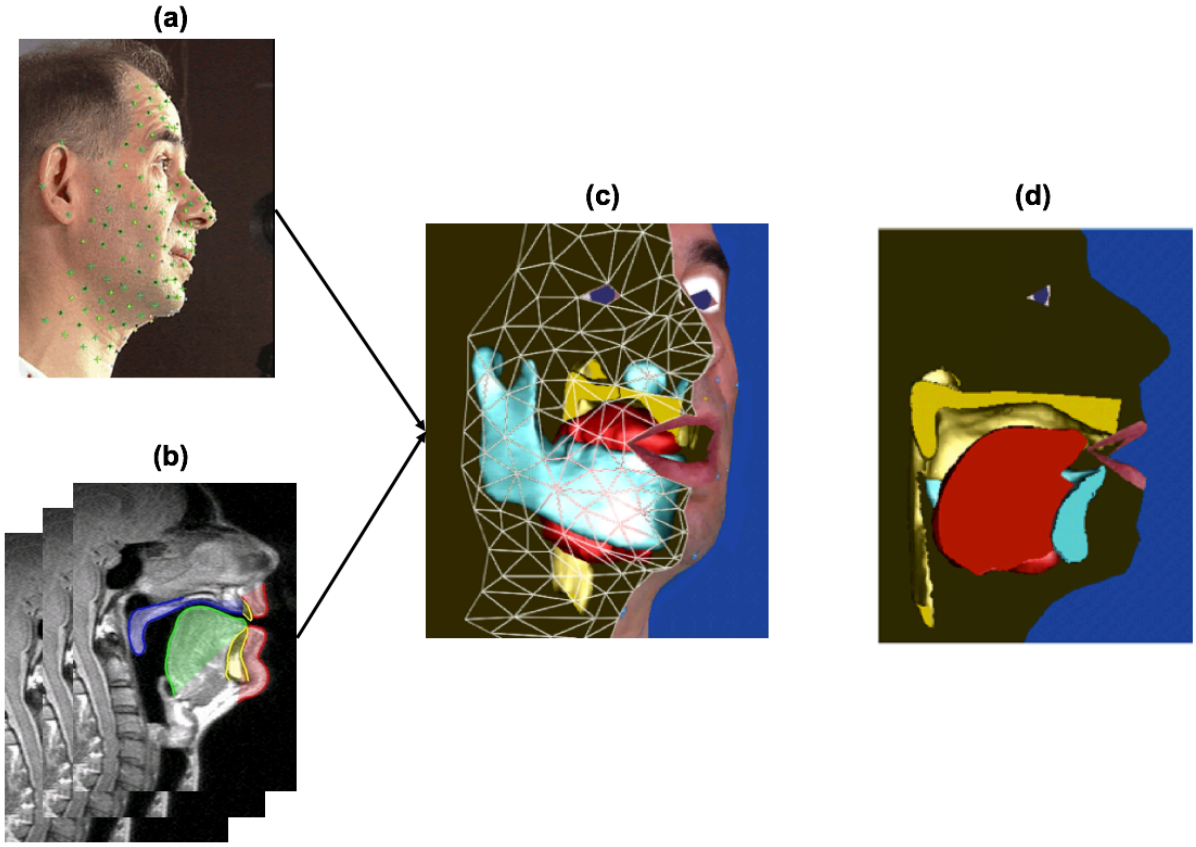


Figure 25. Tête parlante articulatoire construite à partir des données IRM du locuteur de référence, développée au Gipsa-lab par Badin *et al.* (2008) (a) Locuteur de référence et marqueurs utilisés pour les mouvements du visage (billes collées) ; (b) images IRM avec segmentation manuelle des articulateurs pour 46 différentes articulations tenues ; (c) tête parlante articulatoire 3D ; (d) mode de visualisation choisi pour le système de retour visuel proposé.

Badin *et al.* (2010) montrent que les modèles de la langue, des lèvres et de la mâchoire de cette tête parlante peuvent être animés à partir d'un flux de données articulatoires, enregistrées sur le locuteur de référence (*i.e.* le même locuteur que celui utilisé pour fixer la géométrie de la tête parlante) par EMA. Dans la suite de ce chapitre, nous appellerons donc *paramètres de contrôle EMA* un vecteur composé des coordonnées 2D (plan médio-sagittal) de 3 bobines EMA, placées sur la surface de la langue (arrière, corps et apex). Ces trois points sont illustrés sur la Figure 26.

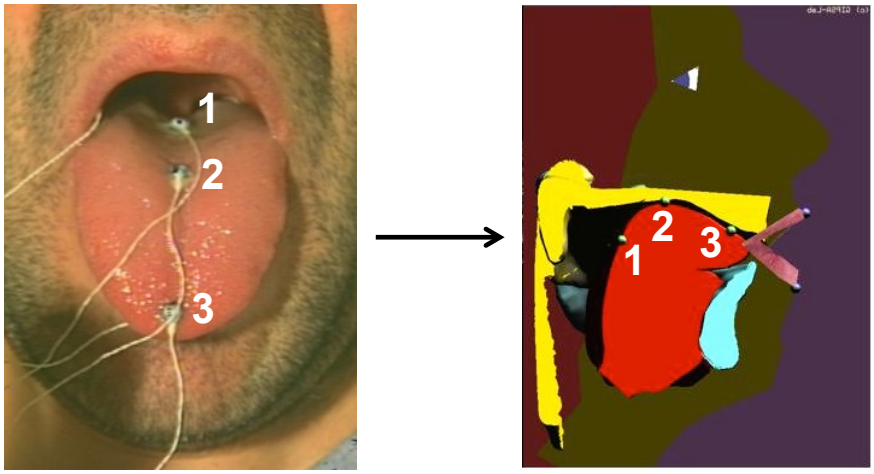


Figure 26. Bobines EMA sur la langue du locuteur de référence numérotées de 1 à 3 de l'arrière à la pointe de la langue (gauche), et visualisation sur la tête parlante articulatoire des paramètres de contrôle EMA associés (à droite).

L'objectif est donc de déterminer les paramètres de contrôle EMA du modèle de langue de la tête parlante articulatoire à partir de l'image échographique de l'utilisateur. Il s'agit donc d'estimer des trajectoires EMA dans l'espace articulatoire du locuteur de référence à partir de l'image échographique d'un locuteur que nous appellerons par la suite le *locuteur source*.

Ce problème de régression, qui implique deux locuteurs (source et référence) et deux modalités (échographie et EMA) est abordé en utilisant une approche de *machine learning* supervisée. Nous proposons le scénario suivant, illustré en Figure 27. Durant une étape dite d'*enrôlement*, il est demandé au locuteur source de prononcer quelques phrases ou logatomes. Les séquences échographiques associées sont d'abord traitées pour en extraire des paramètres. Ensuite, chaque séquence d'images est automatiquement alignée avec une séquence de trajectoires EMA, enregistrées sur le locuteur de référence prononçant la même phrase ou logatome. Un modèle de conversion est ensuite entraîné de manière supervisée entre les données échographiques du locuteur source et les données EMA du locuteur de référence. Dans la phase d'utilisation, les paramètres de contrôle de la tête parlante sont estimés automatiquement à partir des images échographiques du locuteur source qui reçoit ainsi un retour visuel complet sur son articulation linguale, dans lequel la langue est maintenant présentée en contexte (par rapport au palais, aux dents, etc.).

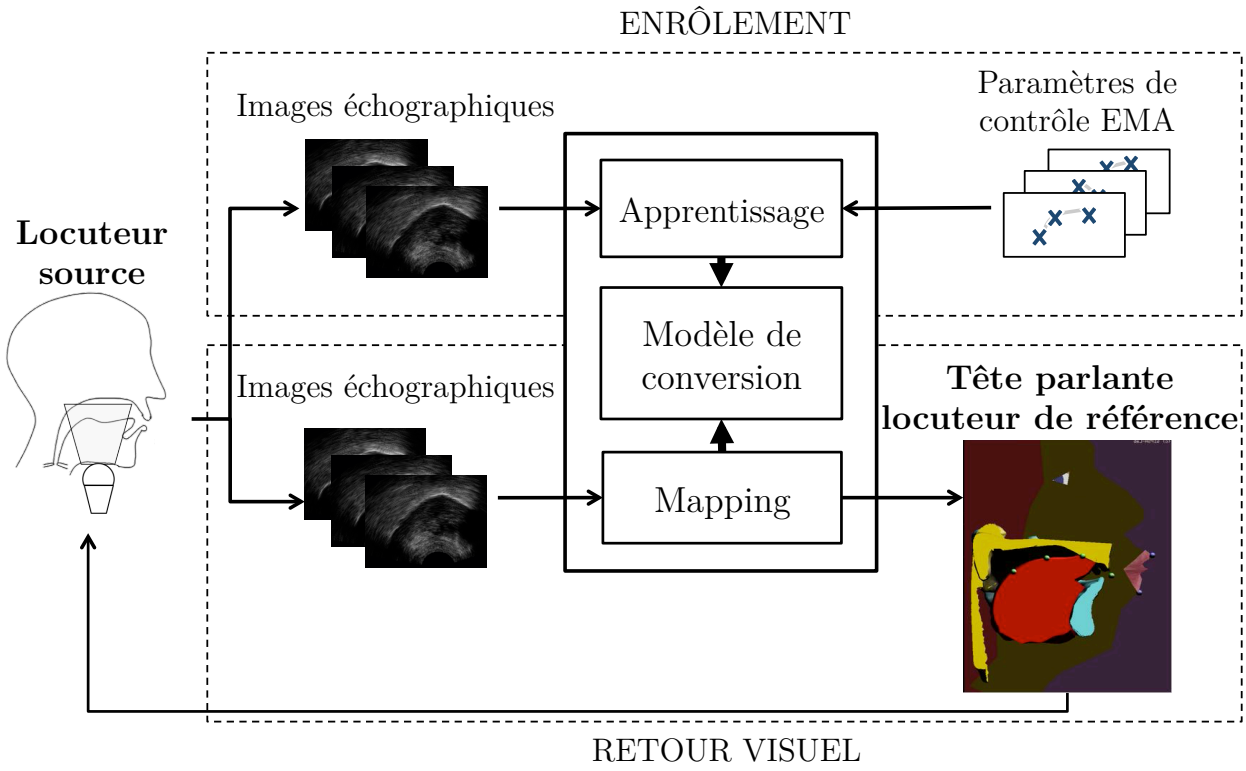


Figure 27. Animation automatique du modèle de langue d'une tête parlante articulatoire à partir d'images échographiques

Ce problème de régression est traité ici par apprentissage automatique, les paramètres du modèle de régression étant estimés sur le corpus acquis en phase d'enrôlement. Deux approches sont proposées :

- Etablir un modèle direct entre données échographiques du locuteur source et paramètres de contrôle de la tête parlante. Dans ce cas, nous utilisons une approche de type GMM que nous appelons dans la suite D-GMR (*Direct Gaussian Mixture Regression*).
- Adapter au locuteur source (l'utilisateur) un modèle existant entraîné sur les données échographiques du locuteur de référence ayant fixé la géométrie de la tête parlante. Dans ce cas, nous étudions l'approche *Cascaded Gaussian Mixture Regressions* (C-GMR), développée par Hueber *et al.* (2015) dans le contexte de l'inversion acoustico-articulatoire. Il s'agit ici d'adapter un modèle de régression de type GMR, entraîné sur les données du locuteur de référence, au locuteur source. Nous faisons l'hypothèse que l'exploitation de connaissance *a priori* sur le locuteur de référence permettra de minimiser la quantité de données d'enrôlement, et d'améliorer les capacités de généralisation du modèle de conversion par rapport à

l'approche directe mentionnée précédemment (D-GMR). Les approches D-GMR et C-GMR sont développées dans la partie qui suit.

3.3. Méthodologie

Nous rappelons dans cette section le principe général de la régression par mélange de Gaussiennes (*Gaussian Mixture Regression* ou *GMR*). Nous détaillons ensuite les particularités de l'algorithme choisi dans le cadre de notre étude.

3.3.1. Gaussian Mixture Model (GMM) et Gaussian Mixture Regression (GMR)

Soit $\mathbf{X}$ et $\mathbf{Y}$ deux vecteurs colonnes aléatoires de dimension D_X et D_Y respectivement. On note $\mathbf{J}$ la concaténation de $\mathbf{X}$ et $\mathbf{Y}$ telle que $\mathbf{J}=[\mathbf{X}^\top,\mathbf{Y}^\top]^\top$ où $^\top$ représente la transposée. Soit $p(\mathbf{x}|\Theta_\mathbf{X})$ la densité de probabilité (PDF) de $\mathbf{X}$, paramétrée par un jeu de paramètres $\Theta_\mathbf{X}$, avec $\mathbf{x}$ une réalisation de la variable aléatoire $\mathbf{X}$. Soit la distribution gaussienne de $\mathbf{X}$ avec un vecteur moyen $\mu_\mathbf{X}$ et une matrice de covariance $\Sigma_{\mathbf{XX}}$. Un modèle de mélange de Gaussiennes (GMM) sur est défini par :

$$p(\mathbf{j}|\mathbf{\Theta_J}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{j}|\mu_{\mathbf{J},m}, \mathbf{\Sigma_{JJ},}_m) \tag{1}$$

où M est le nombre de composantes du mélange. Pour chaque composante (indicée par m), $\pi_m = p(m)$ est la probabilité *a priori* satisfaisant $\sum_{m=1}^{M}\pi_m = 1$, $\mu_{\mathbf{J},m} = [\mu_{\mathbf{X},m}^\top, \mu_{\mathbf{Y},m}^\top]^\top$ est le vecteur de moyenne et $\Sigma_{\mathbf{JJ},m}$ la matrice de covariance donnée par :

$$\mathbf{\Sigma_{JJ},}_m = \begin{bmatrix} \mathbf{\Sigma_{XX},}_m & \mathbf{\Sigma_{XY},}_m \\ \mathbf{\Sigma_{YX},}_m & \mathbf{\Sigma_{YY},}_m \end{bmatrix} \tag{2}$$

Notons la présence de $\Sigma_{\mathbf{XY}}$, la matrice de covariance croisée entre $\mathbf{X}$ et $\mathbf{Y}$ qui va jouer un rôle important dans la régression. L'algorithme classique Expectation-Maximization (EM) (C. Bishop (2007)) pour les GMM peut être utilisé pour estimer ces paramètres à partir d'un jeu d'apprentissage d'observations jointes $(\mathbf{x},\mathbf{y})$. C'est un algorithme itératif composé d'une étape E d'estimation des variables latentes et d'une étape M de maximisation permettant de mettre à jour les paramètres du modèle Θ pour maximiser sa vraisemblance sur les données d'apprentissage : probabilité *a priori*, moyennes et matrices de covariance des M gaussiennes. Ainsi, à la n$^{\text{ème}}$ itération de l'algorithme, pour le modèle

$\mathbf{\Theta}^n$, la probabilité *a posteriori* qu'une observation x soit générée par la m$^{\text{ième}}$ gaussienne, notée m est estimée par la phase E, tel que :

$$p(m|\mathbf{x}, \mathbf{\Theta}^n) = \frac{\pi_m \mathcal{N}(\mathbf{x}, \mu_m^n, \mathbf{\Sigma}_m^n)}{\sum_{i=1}^{M} \pi_i \mathcal{N}(\mathbf{x}, \mu_i^n, \mathbf{\Sigma}_i^n)} \tag{3}$$

Les paramètres du modèle $\mathbf{\Theta}^{n+1}$ sont donc définis comme suit pour la composante m et pour T observations :

$$\alpha_m^{n+1} = \frac{1}{T} \sum_{t=1}^{T} p(m|\mathbf{j}_t, \mathbf{\Theta}^n) \tag{4}$$

$$\mu_m^{n+1} = \frac{\sum_{t=1}^{T} p(m|\mathbf{j}_t, \mathbf{\Theta}^n)\mathbf{j}_t}{\sum_{t=1}^{T} p(m|\mathbf{j}_t, \mathbf{\Theta}^n)} \tag{5}$$

$$\Sigma_m^{n+1} = \frac{\sum_{t=1}^{T} p(m|\mathbf{j}_t, \mathbf{\Theta}^n)(\mathbf{j}_t - \mu_m^n)(\mathbf{j}_t - \mu_m^n)^\top}{\sum_{t=1}^{T} p(m|\mathbf{j}_t, \mathbf{\Theta}^n)} \tag{6}$$

Le GMR utilisé pour associer $\mathbf{x}$ à la valeur $\hat{\mathbf{y}}$ estimée de $\mathbf{y}$ est défini dans Ghahramani & Jordan (1994) :

$$\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{x}, \mathbf{\Theta_J}] = \sum_{m=1}^{M} p(m|\mathbf{x}, \mathbf{\Theta_X}) \mu_{\mathbf{Y}|\mathbf{x},m} \tag{7}$$

avec

$$\mu_{\mathbf{Y}|\mathbf{x},m} = \mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YX},m} \mathbf{\Sigma}_{\mathbf{XX},m}^{-1} (\mathbf{x} - \mu_{\mathbf{X},m}) \tag{8}$$

$$p(m|\mathbf{x}, \mathbf{\Theta_X}) = \frac{\pi_m \mathcal{N}(\mathbf{x}|\mu_{\mathbf{X},m}, \mathbf{\Sigma}_{\mathbf{XX},m})}{\sum_{i=1}^{M} \pi_i \mathcal{N}(\mathbf{x}|\mu_{\mathbf{X},i}, \mathbf{\Sigma}_{\mathbf{XX},i})} \tag{9}$$

Ce mapping revient à minimiser l'erreur quadratique moyenne (MSE) entre $\mathbf{y}$ et $\hat{\mathbf{y}}$ en supposant une indépendance statistique et une distribution identique des observations.

3.3.2. Approche directe D-GMR

Soit $\mathbf{Z}$ un vecteur aléatoire de paramètres échographiques dérivés des images brutes du locuteur de référence (l'extraction de ces paramètres est détaillée en section 2.2.1). Soit $\mathbf{Y}$ le vecteur correspondant aux paramètres de contrôle EMA, décrivant la position de la langue du locuteur de référence lors de la production du même phonème (la mise en

correspondance entre les données du locuteur source et du locuteur de référence est décrite en section 3.4.3). Soit N_0 le nombre d'images échographiques enregistrées par le locuteur source durant la session d'enrôlement, et $\{\mathbf{z}_n\}_{n=1}^{N_0} = \mathbf{z}_{1:N_0}$ le jeu de paramètres correspondant dérivant de ces images.

Une première méthode directe pour aborder le problème de régression considéré dans cette étude consiste à modéliser directement les relations statistiques entre les vecteurs de paramètres échographiques du locuteur source et les vecteurs de paramètres de contrôle EMA du locuteur de référence **Y**. Avec un GMM **Z**-**Y** nous pouvons directement dériver la régression **Z**-vers-**Y** telle que décrite dans la section précédente. Cette approche aborde simultanément le problème du mapping de données inter-locuteurs et inter-modalités. Comme dans l'article de Hueber *et al.* (2015), nous appelons cette approche par la suite régression directe ou *D-GMR*.

Il est important de noter que l'entraînement de ce modèle requiert d'associer les données d'enrôlement $\mathbf{z}_{1:N_0}$ du locuteur source avec les données EMA correspondantes $\mathbf{y}_{1:N_0}$ du locuteur de référence. Cet alignement est réalisé temporellement par un algorithme de Dynamic Time Warping (DTW). Cependant, les données échographiques et les données EMA étant des modalités différentes, nous ne pouvons appliquer directement l'algorithme de DTW : la sémantique et la dimension des descripteurs sont différentes d'une modalité à l'autre, rendant impossible le calcul direct d'une distance. Nous proposons donc d'effectuer cet alignement dans l'espace acoustique, en exploitant le signal vocal audio enregistré avec les images échographiques pour le locuteur source et avec les données EMA pour le locuteur de référence. Des informations complémentaires sur les détails de cet alignement sont données dans la section 3.4.3. Après cette procédure, nous considérons que les données du locuteur source $\mathbf{z}_{1:N_0}$ sont alignées avec celles du locuteur de référence $\mathbf{y}_{1:N_0}$. L'algorithme EM pour le GMM peut ensuite être appliqué au jeu de données $\{\mathbf{z}_{1:N_0}, \mathbf{y}_{1:N_0}\}$. Enregistrer le signal audio en même temps que les données échographiques constitue un des inconvénients principaux de l'approche D-GMR, sur lequel nous reviendrons.

3.3.3. Approche par adaptation d'un modèle de conversion existant par C-GMR

3.3.3.1 Adaptation MAP et MLLR

Dans la section précédente, un GMR est estimé à partir des données d'adaptation échographiques du locuteur source d'une part, et des données EMA du locuteur de

référence associées d'autre part. On peut à ce stade supposer que la qualité de ce modèle, et donc des performance de régression, dépendra fortement de la quantité de données d'apprentissage disponible, et donc de la longueur de la phase d'enrôlement. Une autre approche envisageable est de considérer le problème posé comme **un problème d'adaptation**. L'idée est d'adapter un GMR pré-appris sur une grande quantité de données articulatoires du locuteur de référence.

L'adaptation de modèles probabilistes de type GMM a fait l'objet d'une littérature importante. Dans le cadre de la reconnaissance automatique de la parole, pour l'adaptation des modèles acoustiques HMM (pour lesquels les densités de probabilités d'émission sont des GMM) à un nouveau locuteur, deux des méthodes les plus classiquement utilisées sont la *Maximum-a-Posteriori* (MAP) (Gauvain & Lee (1994)) et la *Maximum-Likelihood-Linear-Regression* (Leggetter & Woodland (1995)). Ces méthodes reposent sur une transformation des paramètres du modèle de référence afin de maximiser la vraisemblance des données d'adaptation. Dans le cas du MAP, le vecteur de moyenne μ_X est mis à jour en utilisant la formule suivante :

$$\mu_{\mathbf{X},m}^{MAP} = \frac{\tau\mu_{\mathbf{X},m} + \sum_{n=1}^{N_0} p(m|\mathbf{z}_n, \mathbf{\Theta_Z})\mathbf{z}_n}{\tau + \sum_{n=1}^{N_0} p(m|\mathbf{z}_n, \mathbf{\Theta_Z})} \tag{10}$$

où τ est un hyperparamètre partagé par les GMM et correspondant à un degré d'adaptation entre les données.

Dans notre contexte de régression par GMM, il s'agit donc d'adapter les paramètres de moyenne et de covariance du GMR **X**-vers-**Y**. Cependant, dans notre cas, nous ne disposons pas des données **y** pour le locuteur source car l'acquisition de données EMA pendant une séance d'orthophonie est naturellement exclue. Les approches MAP et MLLR ne permettent donc d'adapter que les paramètres de moyenne et de covariance liés à la modalité d'entrée **X** du GMR **X**-**Y**. Dans le cas de l'inversion acoustico-articulatoire, Hueber *et al.* (2015) montrent que cette adaptation partielle n'est pas optimale et fournit des performances limitées. Dans sa formulation standard, cette méthode ne nous a pas semblé adaptée à notre problème, elle n'est donc pas évaluée ici[6]. L'approche C-GMR que nous discutons maintenant est une alternative pour traiter le problème d'adaptation d'un GMR.

[6] Des résultats préliminaires confirment cependant ceux de Hueber *et al.* (2015)

3.3.3.2 Approche par C-GMR

L'idée principale de la structure C-GMR présentée dans Hueber *et al.* (2015) est d'exploiter les informations *a priori* de l'espace articulatoire du locuteur de référence. Ces informations sont fournies par un GMM entraîné sur un jeu d'observations jointes $\{(\mathbf{x}_n,\mathbf{y}_n)\}_{n=1}^{N}=\{\mathbf{x}_{1:N},\mathbf{y}_{1:N}\}$, où $\mathbf{X}$ est un vecteur de paramètres dérivé de l'image échographique du locuteur de référence (rappelons que $\mathbf{Y}$ se réfère aux paramètres de contrôle EMA de la tête parlante articulatoire). Ce jeu de données peut être obtenu en synchronisant les mouvements de la langue du locuteur de référence avec l'échographie et l'EMA. Cette synchronisation sera décrite à la section 3.4.3.
Par conséquent, le GMM **X**-**Y** modélise les relations statistiques entre le même mouvement de langue, pour le même locuteur, mais enregistré avec deux dispositifs différents (*i.e.* l'échographe et l'EMA). De plus, étant donné que le modèle de référence n'implique pas le locuteur source, il est possible de l'entraîner en laboratoire sur un large jeu de données. En pratique, cela signifie que le nombre de paires de vecteurs dans ce jeu de données peut être choisi significativement plus grand que le nombre N_0 de $\mathbf{z}$ vecteurs du jeu de données enregistré dans la session d'enrôlement ($N_0 \leq N$ et idéalement $N_0 \ll N$). Le modèle de référence devrait donc être correctement estimé et décrire finement l'espace articulatoire du locuteur de référence, dans les deux modalités utilisées (échographie et EMA).

L'approche C-GMR exploite le modèle de référence **X**-**Y** en découpant la régression **Z**-vers-**Y** en deux étapes :

- une étape de mapping **Z**-vers-**X** qui modélise, dans notre cas, les relations statistiques entre les données échographiques du locuteur source et du locuteur de référence (mapping monomodal inter-locuteur ou *cross-speaker monomodal mapping*)
- une étape de mapping **X**-vers-**Y**, dérivée du modèle de référence **X**-**Y**, qui modélise les relations statistiques entre les données échographiques et les données EMA du locuteur de référence (mapping mono-locuteur inter-modalités ou *cross-modal single-speaker mapping*).

Hueber *et al.* (2015) proposent deux versions de cette approche C-GMR, illustrées en Figure 28 : le Split C-GMR et le Integrated C-GMR. Ces deux versions sont détaillées dans les paragraphes qui suivent. La première version est le **Split C-GMR** (SC-GMR), composé de l'enchaînement de deux GMRs. La deuxième version du C-GMR est le

Integrated C-GMR (IC-GMR), qui intègre les deux GMR du SC-GMR en un seul modèle probabiliste.

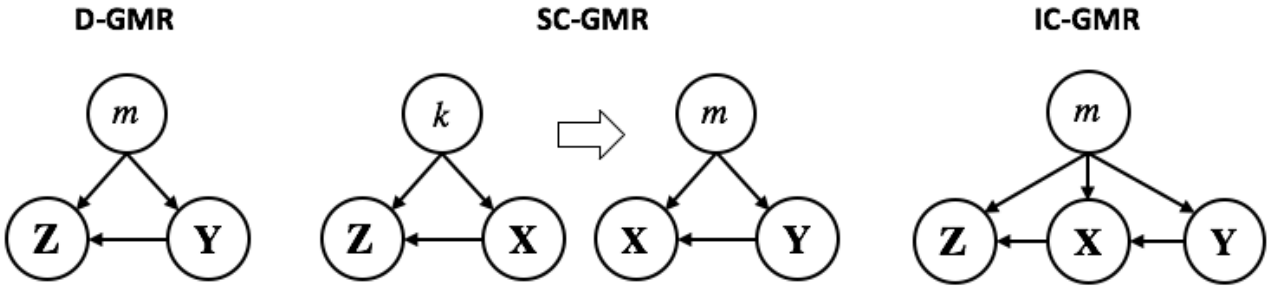


Figure 28. Représentation graphique du Direct-GMR (D-GMR), Split Cascaded GMR (SC-GMR) et Integrated Cascaded GMR (IC-GMR). Z et X sont respectivement les données échographiques des locuteurs source et référence. Y correspond aux paramètres de contrôle EMA. m et k indicent les composantes des modèles de mélange.

3.3.3.3 Split C-GMR (SC-GMR)

Dans le cas du SC-GMR, la sortie du premier GMR estimée est utilisée comme entrée du deuxième GMR, soit $\hat{\mathbf{y}} = E[\mathbf{Y}|\hat{\mathbf{x}},\Theta_J]$ avec $\hat{\mathbf{x}} = E[\mathbf{X}|\mathbf{z},\Theta_I]$ (sachant que $\mathbf{I} = [\mathbf{Z}^\top, \mathbf{X}^\top]^\top$), définis selon l'équation (8) avec leurs paramètres respectifs. Dans le SC-GMR, les deux GMR peuvent avoir un nombre différent de composantes. Comme mentionné précédemment, le GMR de référence **X**-**Y** est entraîné à partir des N observations jointes du locuteur de référence, alors que le GMR **Z**-**X** est entraîné sur un plus petit jeu de données, constitué des vecteurs de paramètres échographiques $\mathbf{z}_{1:N_0}$ et du sous-ensemble $\mathbf{x}_{1:N_0}$ correspondant.

Notons que dans l'approche SC-GMR le nombre de composantes du GMR **Z**-**X** peut être différent de celui du GMR **X**-**Y**. L'approche SC-GMR peut être considérée comme *flexible* en ce sens où la structure du GMR Z-X (*i.e.* k) peut s'adapter à la structure de l'ensemble $\{z_n; x_n\}_{n=1}^{N_0} = \{\mathbf{z}_{1:N_0}; \mathbf{x}_{1:N_0}\}$ qui est susceptible de varier en fonction de N_0 (taille de la quantité de données d'adaptation). Comme détaillé plus tard, le choix du nombre de composante k peut s'optimiser par validation croisée.

3.3.3.4 Integrated C-GMR (IC-GMR)

Le IC-GMR intègre les deux GMR précédemment cités en un seul GMR comme illustré en Figure 28. Les régresseurs **Z**-vers-**X** et **X**-vers-**Y** partagent la même composante m. Le but est de décrire les données d'adaptation avec le même partitionnement que celui utilisé

pour décrire l'espace articulatoire du locuteur de référence (*i.e.* **X**-**Y**). Rappelons que ce dernier est supposé être bien estimé sur un large jeu de données. Mathématiquement, les dépendances statistiques entre **X**, **Y** et **Z** sont modélisées comme suit :

$$p(\mathbf{x},\mathbf{y},\mathbf{z}|\mathbf{\Theta}) = \sum_{m=1}^{M} p(m)p(\mathbf{y}|m,\mathbf{\Theta}_{\mathbf{Y},m})p(\mathbf{x}|\mathbf{y},m,\mathbf{\Theta}_{\mathbf{X}|\mathbf{Y},m})p(\mathbf{z}|\mathbf{x},m,\mathbf{\Theta}_{\mathbf{Z}|\mathbf{X},m}) \qquad (11)$$

où pour chaque composante $\pi_m = p(m)$ est la distribution *a priori*, $p(\mathbf{y}|m,\Theta_{\mathbf{Y},m})$ est une distribution gaussienne, et les PDFs conditionnelles $p(\mathbf{x}|\mathbf{y},m,\Theta_{\mathbf{X}|\mathbf{Y},m})$ et $p(\mathbf{z}|\mathbf{x},m,\Theta_{\mathbf{Z}|\mathbf{X},m})$ sont des distributions Linear-Gaussian (*i.e.* une distribution gaussienne avec une moyenne étant une fonction affine de la variable conditionnelle). Toutes ces distributions ont des matrices de covariance pleines.

Les paramètres de ces distributions sont estimés à l'aide d'un algorithme de type EM dédié. Cet algorithme est entièrement décrit dans Hueber *et al.* (2015) et ne sera donc pas détaillé ici. Notons que contrairement à MAP et à MLLR, le processus d'adaptation prend donc ici la forme de l'apprentissage d'un nouveau modèle, à partir de l'ensemble des N observations du locuteur de référence (**X** et **Y**), et des N_0 observations du locuteur source (**Z**) (avec classiquement $N_0<N$). Une représentation schématique des données d'apprentissage du IC-GMR est proposée à la Figure 29. Nous mentionnons ici les spécificités de l'algorithme d'apprentissage du modèle IC-GMR :

- L'initialisation de $p(\mathbf{x}|\mathbf{y},m,\Theta)$ s'effectue à partir du modèle de référence, supposé bien estimé sur un large corpus du locuteur de référence. Le modèle IC-GMR hérite donc de la structure de ce modèle (*i.e.* m).
- L'étape d'enrôlement étant généralement limitée dans le temps, la quantité de données acquises sur le locuteur source (*i.e.* N_0) peut être relativement réduite, notamment par rapport à N. Un partitionnement trop fin de ces données issues des données du locuteur source peut conduire à un corpus dit clairsemé (sparse), avec peu ou pas d'observations à associer à certaines composantes du mélange. Ce phénomène peut également s'accentuer si le locuteur source ne peut pas prononcer un ou plusieurs phonèmes, comme dans le cas du contexte clinique envisagé. Pour pallier ces manques, l'algorithme d'apprentissage du IC-GMR exploite la méthode générale des *données manquantes* (missing data) décrite dans Ghahramani & Jordan (1994). L'algorithme d'apprentissage étant de type EM, le principe général est d'inférer pendant la phase E les données $\mathbf{z}_{N_0+1:N}$ (illustrées schématiquement en Figure 29) à partir des données existantes et des données du locuteur de référence. Les données manquantes, que nous noterons z', sont estimées telles que pour n=[N_0+1:N] et m=[1:M], $z'_{n,m} = E[Z|x_n,m,\Theta_{IC}]$, c'est-à-dire via un GMR **X**-**Z**.

Ensuite, les responsabilités $p(m|z_n)$ sont classiquement estimées, avant la mise à jour en phase M des paramètres des GMR qui décrivent les relations statistiques en **Z** et **X** d'une part, et entre **X** et **Y** d'autre part.

Contrairement aux approches D-GMR et SC-GMR, notons ici que les cibles articulatoires dans l'espace du locuteur de référence (*i.e.* **Y**) sont susceptibles d'être modifiées pendant l'apprentissage du modèle IC-GMR, à partir des données du locuteurs source (*i.e.* **Z**). La régression par IC-GMR s'effectue à l'aide de la formule suivante :

$$\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{z}] = \sum_{m=1}^{M} p(m|\mathbf{z}, \mathbf{\Theta_Z})(\mu_{\mathbf{Y},m} + \mathbf{\Sigma_{YX},}_m \mathbf{\Sigma}^{-1}_{\mathbf{XX},m} \mathbf{\Sigma_{XZ},}_m \mathbf{\Sigma}^{-1}_{\mathbf{ZZ},m}(\mathbf{z} - \mu_{\mathbf{Z},m})) \quad (12)$$

Les poids des composantes sont obtenus en appliquant l'équation classique présentée en (7) avec les distributions marginales $p(\mathbf{z}|m,\Theta_{\mathbf{z},m})$, obtenues à partir des distributions définies ci-dessus. L'équation (12) fait apparaître une cascade de matrices de covariance permettant un conversion de **Z** à **Y** en passant par **X**. Cette cascade donne le nom à la méthode. Notons que, comme pour le D-GMR et le SC-GMR, la conversion peut s'effectuer en temps-réel puisque l'observation **y** ne s'estime qu'à partir de l'observation **z** au même instant, et des paramètres du modèle.

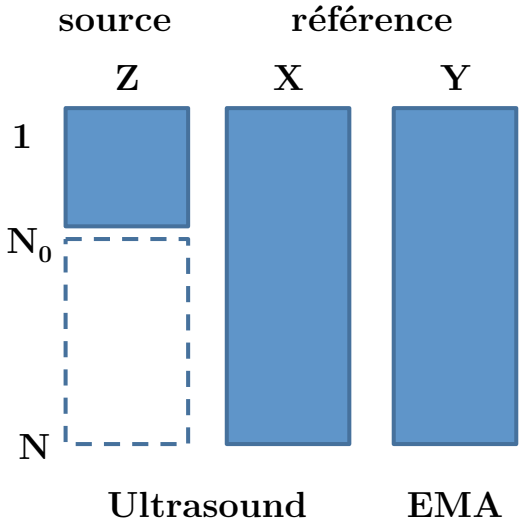


Figure 29. Représentation schématique des variables X Y et Z utilisées pour les C-GMR. En pointillé, les données manquantes

3.4. Dispositif expérimental et évaluation des méthodes

3.4.1. Acquisition des données

Le corpus utilisé dans cette étude est celui présenté dans Badin *et al.* (2008). Une large base de données en audio et en EMA a été enregistrée pour le locuteur de référence grâce au système EMA Carstens AG200 (*cf* Figure 8). Six bobines sont attachées aux organes orofaciaux du locuteur dans le plan médio-sagittal : trois sur la langue, une sur chaque lèvre et une sur la mâchoire. Deux bobines complémentaires, sur le nez et les incisives supérieures, permettent d'établir une référence. Le corpus est constitué de 1108 productions qui se détaillent comme suit : deux répétitions de 266 séquences VCV, avec C = une des 19 consonnes/semi-voyelles françaises et V = une des 14 voyelles françaises orale et nasales ; deux répétitions des 109 paires de CVC de mots français, différant dans une paire d'un seul phonème (Peckels & Rossi (1973)) ; 68 phrases courtes et 9 des 10 phrases de la première liste de Combescure (1981) ; 11 longues phrases arbitraires (pour un total d'environ 17 minutes de parole, longues pauses exclues).

Le corpus acquis sur les images échographiques est constitué des mêmes 1108 productions que le corpus EMA. Enregistré avec le logiciel Ultraspeech 1.3 de Hueber *et al.* (2008), il inclut le signal audio ainsi que les images échographiques de la langue et les vidéos des lèvres synchronisées. Trois locuteurs ont été enregistrés : le locuteur de référence, ainsi qu'un homme (différent du locuteur de référence) et une femme sans aucun trouble articulatoire, que nous nommerons par la suite respectivement M1 et F1. Les images de la langue sont enregistrées à 60 Hz avec 640x480 pixels avec le même dispositif échographique que celui décrit en section 2.3. Durant l'acquisition des données, la sonde échographique est fixée à la tête du locuteur avec un casque de stabilisation (conçu par la société Articulate Instruments). Enfin, le signal acoustique est enregistré simultanément à 44.1 kHz et 32 bits.

3.4.2. Paramétrisation du signal audio et des images échographiques

Le contenu spectral du signal de parole est paramétré par une décomposition en MFCC, réalisée avec la boîte à outils HTK (Young, Woodland *et al.* (2013)) et une configuration standard (sous-échantillonnage du signal audio à 16kHz, 16 bits, fenêtre d'analyse de 20ms, 5ms de décalage). L'analyse MFCC crée des vecteurs de 26 coefficients incluant les coefficients statiques et leur dérivée temporelle.

Les images échographiques brutes ont été encodées par des paramètres EigenTongues résultant d'une analyse en composantes principales (section 2.2.1). Le nombre D de coefficients retenus afin de conserver 80 % de la variance s'élève ici à D = 30.

3.4.3. Alignement des données

3.4.3.1 Alignement des données échographiques et EMA du locuteur de référence

La mise en place du modèle de référence (**X-Y**) nécessite l'enregistrement simultané de données échographiques et EMA sur le locuteur de référence. Cet enregistrement est délicat car il y a un risque d'interférence de l'échographe et du casque métallique avec le champ magnétique du système de capture EMA. Nous avons donc réalisé deux enregistrements distincts que nous avons ensuite synchronisés *a posteriori*. Cette synchronisation s'effectue à l'aide de l'algorithme DTW de Myers & Rabiner (1981) en exploitant les productions acoustiques enregistrées simultanément avec les données échographiques d'une part, et les données EMA d'autre part. Nous illustrons cette procédure en Figure 30.

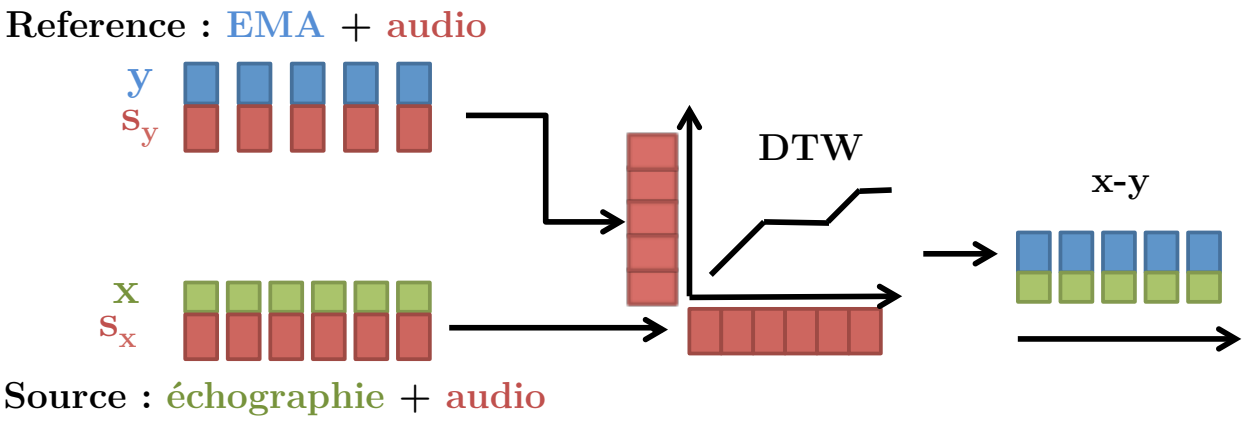


Figure 30. Alignement par DTW d'une séquence d'images échographiques (paramétrées par l'approche EigenTongues) avec une séquence de paramètres EMA pour le locuteur de référence, en exploitant les signaux audio s_x et s_y associés respectivement aux données échographiques et aux données EMA.

3.4.3.2 Alignement des données échographiques du locuteur source avec les données échographiques du locuteur de référence

Deux approches sont proposées pour aligner les données échographiques du locuteur source avec les donnes échographiques et/ou EMA du locuteur de référence, pour la phase d'enrôlement.

La première approche exploite la production acoustique telle que décrite précédemment dans le cas du locuteur de référence. Il s'agit d'aligner par DTW le signal vocal (audio) du locuteur source avec celui du locuteur de référence prononçant la même phrase ou logatome. Le chemin d'alignement DTW est ensuite utilisé pour interpoler les séquences d'images échographiques du locuteur source pour s'adapter au rythme d'articulation du locuteur de référence. Cette procédure est utilisée dans l'approche D-GMR (conversion directe Z-Y) pour mettre en correspondance les séquences échographiques du locuteur source avec des séquences EMA du locuteur de référence et construire ainsi $\left\{\mathbf{z}_{1:N_0}, \mathbf{y}_{1:N_0}\right\}$. Dans l'approche C-GMR (SC-GMR et IC-GMR), cet alignement permet d'obtenir un corpus parallèle $\left\{\mathbf{z}_{1:N_0}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}\right\}$, c'est-à-dire une mise en correspondance des données échographiques et EMA du locuteur source et du locuteur de référence.

La seconde approche proposée vise à s'affranchir du signal acoustique pour aligner les données échographiques du locuteur source avec celle du locuteur de référence, dans le cas de l'approche par C-GMR (SC-GMR et IC-GMR). Ceci peut être utile dans un contexte clinique, notamment pour traiter des patients qui peuvent bouger leur langue mais qui ne peuvent plus vocaliser (c'est notamment le cas de certains patients victimes d'un AVC).

Pour ce faire, l'alignement DTW est réalisé directement dans l'espace échographique (EigenTongues) plutôt que dans l'espace acoustique (MFCC). Afin de pouvoir calculer une distance euclidienne entre deux vecteurs de descripteurs visuels (coordonnées d'une image dans l'espace des EigenTongues), la décomposition en EigenTongue est effectuée sur une base de données d'apprentissage composée des images des deux locuteurs (source et référence) de manière équilibrée (1000 images de chacun sélectionnées aléatoirement).
Par la suite, nous nous référerons à cette approche en utilisant les noms de **SC-GMR (no audio)** et **IC-GMR (no audio)**.

Pour les différentes procédures d'alignement, les données échographiques (à 60Hz à l'origine) des locuteurs sources et du locuteur de référence, les séquences de vecteurs

MFCC, et les séquences de paramètres EMA (locuteur de référence) sont ré-échantillonnés à 100Hz par interpolation linéaire.

3.4.4. Choix du nombre de composantes des modèles D-GMR et C-GMR.

Après l'alignement des différents corpus, les modèles **Z-Y** (D-GMR) ainsi que le GMR **X-Y**, c'est-à-dire le modèle associé au locuteur de référence utilisé pour l'initialisation des modèles SC-GMR et IC-GMR), sont entrainés à l'aide de l'algorithme EM. L'algorithme k-means est utilisé pour l'initialisation de l'EM. L'hyperparamètre principal d'un GMM est son nombre de composantes. Dans notre implémentation, ce dernier est estimé sur un corpus de validation (typiquement 20% du corpus d'apprentissage). Cette procédure permet d'adapter automatiquement, en fonction de la quantité (et de la structure) des données d'enrôlement, le nombre de composantes du D-GMR (Z-Y), celui du premier GMR Z-X de l'approche SC-GMR, et celui du GMR X-Y associé au locuteur de référence (rappelons que le IC-GMR hérite par définition de la structure du modèle de référence **X-Y**). En effet, accroître la quantité de données d'enrôlement est susceptible de complexifier la structure de l'espace articulatoire, nécessitant un plus grand nombre de composantes pour le modéliser.

3.4.5. Corpus d'apprentissage, de validation et de test

Une procédure de validation croisée à cinq partitions a été utilisée pour évaluer les approches D-GMR et C-GMR (SC-GMR et IC-GMR). Le corpus de données (présenté à la section 3.4.1) est divisé en cinq partitions de taille quasi-identique (la division s'effectuant sur le nombre de phrases et non sur le nombre de trames). Quatre partitions sont utilisées afin de constituer le corpus d'apprentissage et le corpus de validation pour estimer les paramètres et hyper-paramètres (nombre de composantes) des différents GMR. La partition restante est utilisée pour le test. Cette procédure est répétée en considérant les cinq permutations possibles. Cette procédure permet donc de constituer un ensemble de test de taille égale à celle du corpus de données.

3.4.6. Métrique de comparaison des paramètres EMA

La performance des différents modèles est évaluée en calculant l'erreur quadratique (RMSE) moyenne entre les observations EMA réelles et celles estimées à partir des données échographiques.

$$RMSE(i) = \sqrt{\frac{1}{N_v(i)} \frac{1}{D_Y} \sum_{n=1}^{N_v(i)} \sum_{d=1}^{D_Y} (y_{dn} - \hat{y}_{dn})^2} \tag{13}$$

$$RMSE_{Moyenne} = \frac{1}{N_S} \sum_{i=1}^{N_S} RMSE(i) \tag{14}$$

où i est l'index de l'élément du corpus, $N_v(i)$ est le nombre total de vecteurs d'un élément i, D_Y et y_{dn} sont respectivement les dimensions et les entrées des vecteurs $\mathbf{y}$, et N_s est le nombre total d'éléments du jeu de données de test.

3.5. Résultats et interprétations

3.5.1. Performance du modèle de référence X-Y

Nous avons tout d'abord évalué la préférence du modèle de référence **X-Y** qui sera utilisé notamment pour l'initialisation des modèles C-GMR. Rappelons qu'il s'agit ici d'évaluer la conversion « données échographiques vers données EMA » pour le locuteur de référence. Il s'agit donc d'un problème *a priori* plus simple que la conversion impliquant deux modalités et deux locuteurs différents. La performance de ce modèle **X-Y** peut donc être considérée comme le meilleur résultat possible pour les approches D-GMR et C-GMR.
Pour le modèle **X-Y**, le nombre optimal de composantes estimé par validation croisée (tel que décrit précédemment) s'est révélé être M = 16. Dans le Chapitre 2, nous avions estimé qu'il y avait une vingtaine d'articulations possibles pour la langue. Nous pouvons faire l'hypothèse que ces 16 composantes représentent les différentes articulations de la langue dans le corpus.

Les performances du modèle de référence sont présentées en Tableau 5.

Tableau 5. Performance du GMR $\mathbf{X}$-vers-$\mathbf{Y}$ (mapping du locuteur de référence) : RMSE et intervalle de confiance à 95% (CI), pour chaque paramètre de contrôle EMA (tip, mid, back)

	RMSE (en mm)	Coeff. corrélation Pearson
$\mathbf{Tip}_x$	2,31 ± 0,07	0,89
$\mathbf{Tip}_y$	2,01 ± 0,05	0,85
$\mathbf{Mid}_x$	1,90 ± 0,05	0,89
$\mathbf{Mid}_y$	1,74 ± 0,06	0,90
$\mathbf{Back}_x$	1,65 ± 0,05	0,94
$\mathbf{Back}_y$	2,25 ± 0,06	0,79

La moyenne de la RMSE de chaque séquence de test s'élève à 2,2 mm. La performance de ce modèle est globalement bonne, et meilleure que celle rapportée dans la littérature sur l'inversion acoustico-articulatoire (cf. section 1.3.4) comme dans Toda *et al.* (2008). Ce résultat était attendu car le mapping échographie/EMA est vraisemblablement moins complexe que le mapping acoustico-articulatoire. En effet, les données échographiques portent explicitement et quasi uniquement les informations sur la langue, contrairement aux données acoustiques qui peuvent être décrites comme la conséquence des mouvements combinés de tous les articulateurs (langue, lèvres, etc.) et des sources d'excitation acoustique du conduit vocal (vibrations de la glotte et bruits de frictions).

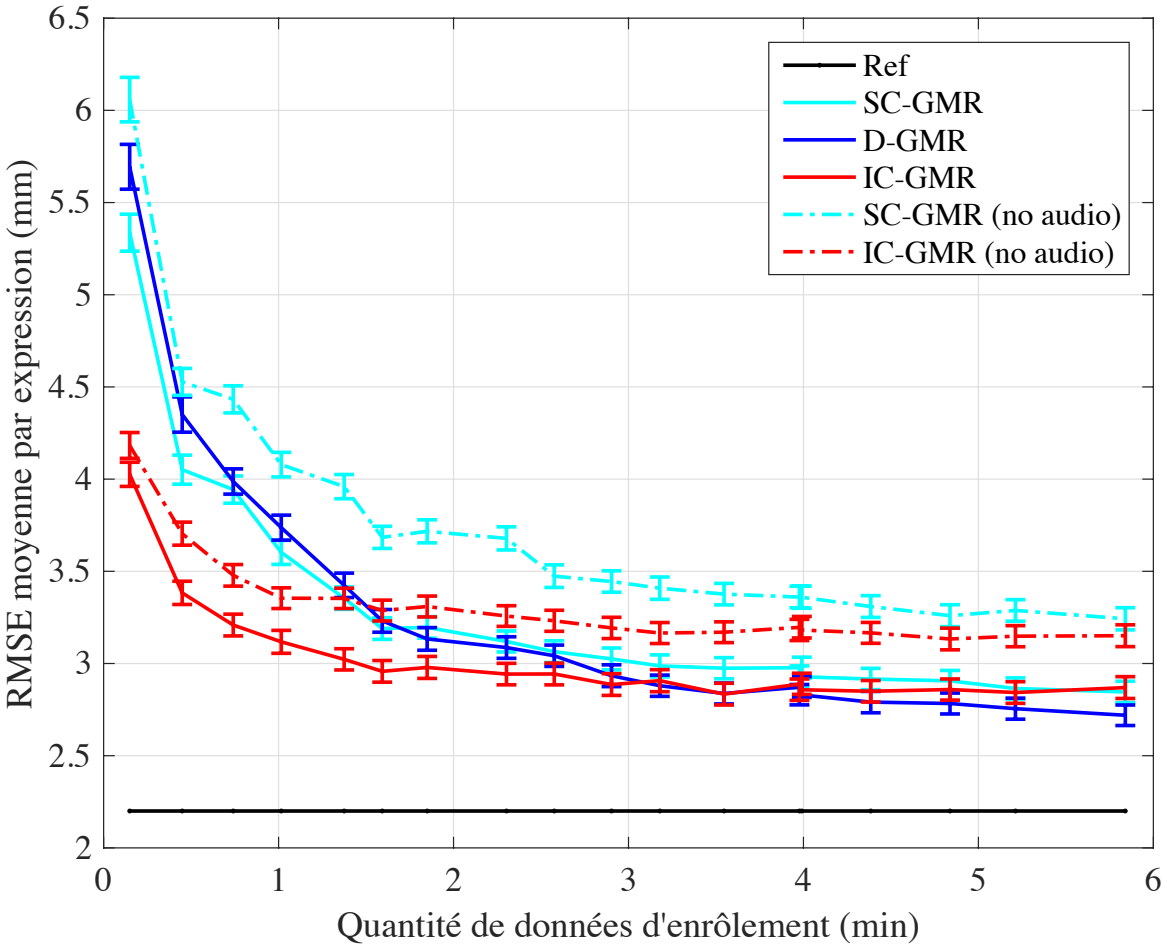
Le détail des résultats pour chaque coordonnée des bobines EMA estimée est présenté dans le Tableau 5. La RMSE est calculée indépendamment pour chaque dimension de $\mathbf{y}$, sans contrairement à ce qui est proposé dans l'équation (13) où la RMSE est sommée sur l'ensemble des dimensions. Le point du milieu de la langue (mid) est estimé avec plus de précision que la pointe et l'arrière de la langue. Ces différences peuvent être expliquées par le fait que ces deux extrémités sont parfois cachées par des ombres de la mâchoire et de l'os hyoïde.

3.5.2. Performance des approches D-GMR et C-GMR

3.5.2.1 Performance en fonction de la quantité de données d'enrôlement

Dans la perspective d'une future application en contexte clinique, nous évaluons la performance des modèles en fonction de la quantité de données d'enrôlement (que nous cherchons à limiter au maximum), c'est-à-dire N_0. Pour ce faire, un jeu de phrases

d'enrôlement est sélectionné aléatoirement parmi les phrases disponibles dans les quatre partitions d'apprentissage de la procédure de validation croisée à cinq partitions. La taille du jeu de données d'enrôlement varie de 1/20 à 1/2 de la taille du jeu de données d'apprentissage, avec **17** tailles intermédiaires. Cette procédure est répétée pour les cinq permutations possibles de la procédure de validation croisée, pour un total de **85** expériences pour chacun des locuteurs source F1 et M1.



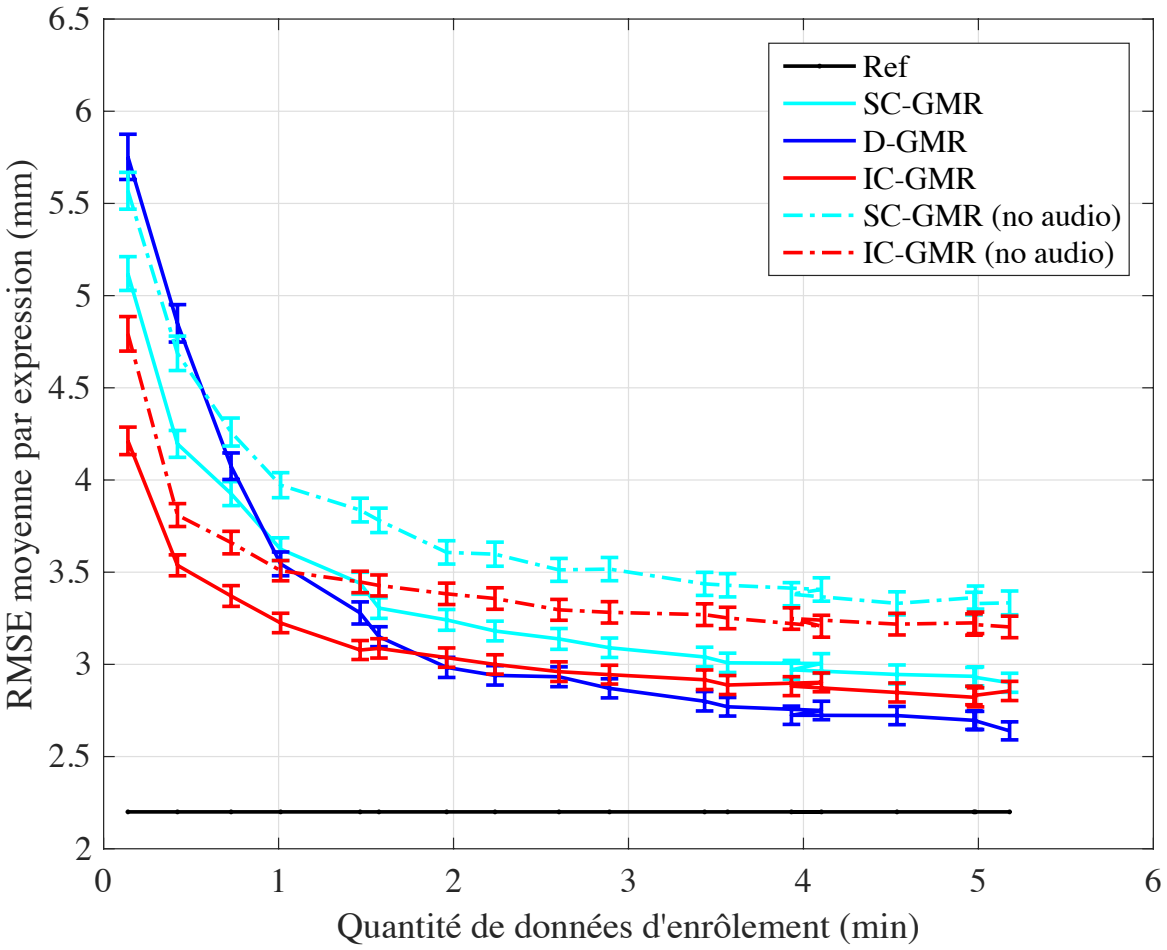


Figure 31. Performance du mapping Z-vers-Y en RMSE (mm), avec un intervalle de confiance à 95 %, en fonction de la quantité de données d'enrôlement, pour le locuteur source F1 (en haut) et M1 (en bas)

Nous observons tout d'abord que les performances de tous les modèles sont significativement plus faibles que celles observées pour le locuteur de référence (cf Tableau 5), même pour une quantité relativement élevée de données d'enrôlement (environ 1 mm de plus). Ce résultat était attendu et peut être en partie expliqué par les idiosyncrasies articulatoires : cela correspond aux différences entre stratégies articulatoires de deux locuteurs pour la prononciation d'un même phonème. Nous ne prenons pas en compte ce phénomène dans l'évaluation et considérons que l'articulation du locuteur de référence est la *vérité terrain*. Ainsi, un vecteur de paramètres EMA estimé à partir des images du locuteur source peut différer de celui observé pour le locuteur de référence sans que le retour visuel proposé soit erroné.

Nous voyons ensuite que la performance de toutes les techniques de mapping considérées dans cette étude (D-GMR, SC-GMR, IC-GMR) augmente avec la quantité de données d'enrôlement. Pour une quantité relativement importante de données d'enrôlement (plus de 3min30 pour le locuteur M1 et plus de 4min pour le locuteur F1), la meilleure performance est atteinte en utilisant l'approche D-GMR. Comme attendu, pour une quantité suffisante de données sur le locuteur source, il n'est pas nécessaire d'exploiter les informations *a priori* sur le locuteur de référence. Cependant, avec peu de données d'enrôlement, la meilleure performance est obtenue avec les techniques C-GMR. Ce résultat est celui qui nous intéresse dans cette étude. Cette tendance est observée pour les deux locuteurs pour le IC-GMR, et surtout pour le locuteur M1 pour le SC-GMR. Pour le locuteur F1, la différence avec le D-GMR n'est significative que pour les deux jeux de données les plus réduits. De plus, le gain obtenu avec le IC-GMR est bien plus important que celui du SC-GMR sur le D-GMR. Par exemple, pour le locuteur source F1, la performance du IC-GMR avec environ 1 min de données d'enrôlement est proche de celle obtenue avec plus de 6 min en terme de RMSE en fonction de N_0, alors que cette différence est bien plus importante pour le D-GMR. **Ces résultats démontrent l'efficacité de l'approche C-GMR et l'intérêt d'exploiter les informations *a priori* du locuteur de référence pour pallier le manque de connaissance sur le locuteur source**. Le IC-GMR surpasse systématiquement et significativement le SC-GMR pour les deux locuteurs.

En effet, pour le SC-GMR, un nouveau modèle est recréé à chaque fois pour entraîner le GMR **Z**-vers-**X** à partir de zéro, le GMR de référence **X**-vers-**Y** restant inchangé d'une expérience à l'autre. Pour le IC-GMR, les relations statistiques entre toutes les données disponibles des locuteurs source et référence (*i.e.* **Z**, **X** et **Y**) sont exploitées conjointement, les cibles articulatoires sont donc susceptibles de varier pendant l'apprentissage.

Enfin, nous discutons des performances des modèles SC-GMR (no audio) et IC-GMR (no audio), pour lesquels l'alignement entre les données des locuteur source et référence a été réalisé directement sur les images plutôt qu'en exploitant le signal audio. La performance est plus faible que celle obtenue avec les modèles SC-GMR et IC-GMR dont nous venons de discuter. Nous observons pour le SC-GMR une différence moyenne de 10 %, contre 5% pour le IC-GMR. Cette baisse de performance peut s'expliquer par le fait d'encoder à la fois la variabilité inter-locuteur et intra-locuteur avec un simple modèle linéaire comme le modèle EigenTongue proposé. Cependant, le IC-GMR (no audio) demeure significativement meilleur que le D-GMR pour moins d'une minute de données

d'adaptation. Dans un scénario clinique où le patient est incapable de vocaliser, le IC-GMR demeure la meilleure approche pour proposer un retour visuel satisfaisant.

3.5.2.2 Capacité de généralisation du C-GMR

Toujours dans la perspective d'une application en contexte clinique, nous nous intéressons par ailleurs à l'évaluation de la capacité de généralisation des approches D-GMR et C-GMR, en évaluant leur performance sur une configuration articulatoire non vue pendant la phase d'enrôlement, simulant ainsi le cas d'un phonème que le patient n'arrive pas à articuler au début de sa rééducation. Le protocole expérimental est le suivant.

Nous avons utilisé certaines séquences VCV du corpus décrit en section 3.4.1, avec V = {a, i, u, ɛ, o} et C = {t, k, ʁ, l, s, ʃ}, et deux répétitions de chaque séquence. Ces deux jeux de phonèmes ont été sélectionnés afin de maximiser la couverture des articulations linguales en français tout en minimisant la quantité de données pour être applicable en contexte clinique et constituent notre *corpus d'apprentissage* (près d'une minute soit environ 6000 trames).

Pour chaque simulation, nous avons généré un jeu de données d'enrôlement composé des VCV du corpus d'apprentissage, sauf un phonème qui a été retiré. Ainsi, nous avons soit V composé d'un sous-ensemble de quatre voyelles parmi {a, i, u, ɛ, o}, soit C sous-ensemble de cinq consonnes parmi {t, k, ʁ, l, s, ʃ}. Cette opération a été réalisée indépendamment pour chacun des onze phonèmes de {a, i, u, ɛ, o, t, k, ʁ, l, s, ʃ}, résultant en un total de onze simulations par locuteur.

Pour chaque jeu de données d'enrôlement, un jeu de données de test correspondant a été généré. Nous avons sélectionné parmi les VCV restantes, non-incluses dans le corpus d'apprentissage, les séquences où V ou C appartient à l'ensemble {a, i, u, ɛ, o, t, k, ʁ, l, s, ʃ}. De ces séquences VCV, seulement les vecteurs correspondant au phonème testé (donc soit les V, soit le C de la séquence) ont été conservés pour composer le jeu de données de test. De plus, dans le cas des consonnes, nous avons élargi notre jeu de données de test en ajoutant les consonnes ayant le même lieu d'articulation que le phonème évalué, soit C dans {t, d, n, k, g, ʁ, l, s, ʃ, z, ʒ}. Ceci permet d'augmenter la taille des données de test, tout en conservant une cohérence phonétique. Prenons un exemple : si nous testons la capacité d'un modèle à généraliser au phonème /t/, le corpus d'enrôlement est donc constitué de tous les vecteurs correspondant aux séquences VCV avec V = {a, i, u, ɛ, o} et

C = {k, ʁ, l, s, ʃ}. Les données de test sont composées des vecteurs associés aux consonnes des séquences VCV avec V *non inclus* dans {a, i, u, ɛ, o} et C = {t, d, n}.

Pour chacun des onze phonèmes {a, i, u, ɛ, o, t, k, ʁ, l, s, ʃ}, les D-GMR, SC-GMR et IC-GMR ont été entraînés avec les différents jeux de données d'enrôlement, suivant la même procédure que celle décrite en section 3.3, et ont été testés sur le jeu de données de test correspondant. Pour la technique C-GMR, l'apprentissage a été réalisé sur le même modèle de référence que celui considéré dans la section précédente.

Afin d'évaluer les résultats obtenus, nous avons aussi produit une expérience afin de déterminer une *baseline*. Pour cela, nous avons utilisé l'ensemble des VCV du corpus d'apprentissage incluant les onze phonèmes, sans en exclure aucun. Le test a été réalisé sur le même corpus de test que lorsque la capacité de généralisation est évaluée. Nous soulignons de plus qu'au vu de la faible taille du corpus de travail, aucune procédure de validation croisée n'a été réalisée dans les expériences qui suivent.

La Figure 32 affiche les *boxplots* des RMSE obtenues sur l'ensemble des onze expériences de généralisation ainsi que pour la *baseline*, et ce pour chaque locuteur M1 et F1. Comme détaillé en section 2.4.2, une *boxplot*, ou boîte à moustache, se compose d'une médiane matérialisée par une marque rouge. Les limites haute et basse de la boîte indiquent respectivement le 25^{e} et le 75^{e} centile, les croix rouges représentent des points aberrants.

De façon attendue, nous constatons tout d'abord que dans ce scénario visant à évaluer leur capacité de généralisation, les performances sont moins bonnes que celles de la *baseline* : l'exclusion d'une catégorie phonétique du corpus d'apprentissage a un impact négatif sur les performances globales. Comme dans la section précédente, nous constatons par ailleurs que le D-GMR est moins performant que les C-GMR dans n'importe quelle configuration.

Dans les expériences de généralisation, le D-GMR est particulièrement mauvais pour le locuteur F1, alors que sa performance équivaut celle du SC-GMR pour le locuteur M1. Pour les deux locuteurs, le IC-GMR surpasse les deux autres modèles en généralisation : pour le locuteur F1, l'erreur médiane obtenue par le IC-GMR est à peu près 2,5mm plus basse que celle obtenue avec le D-GMR ; pour le locuteur M1, cette différence est d'environ 1,2mm avec l'erreur médiane du SC-GMR. De plus, les estimations du IC-GMR sont moins dispersées que celles des deux autres modèles (moins de points aberrants). Nous voyons ici à nouveau l'influence positive de l'exploitation de connaissances *a priori* sur l'espace articulatoire du locuteur de référence.

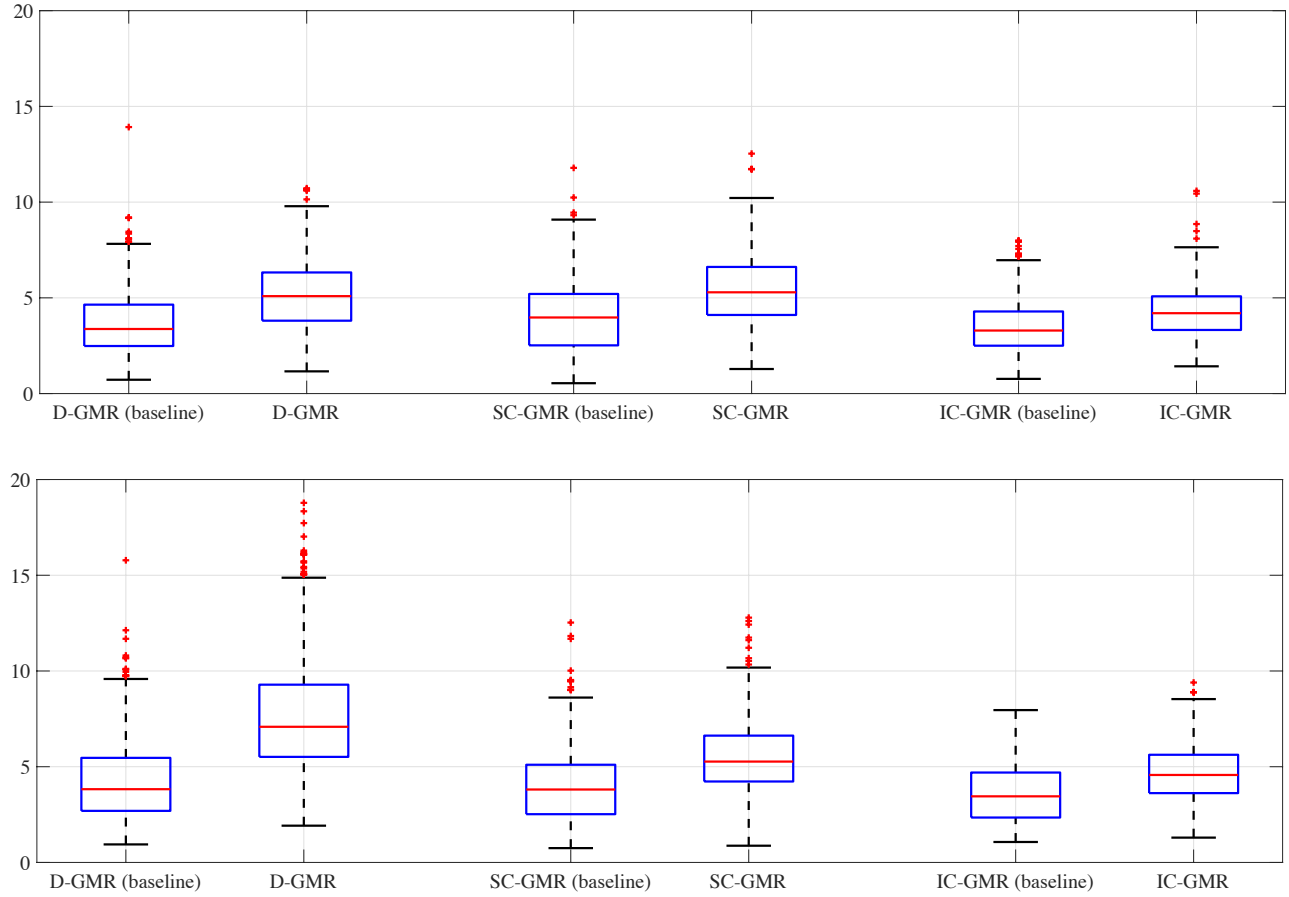


Figure 32. Boxplots des RMSE obtenues pour la *baseline* et en généralisation sur le corpus des 11 phonèmes de test, pour M1 (en haut) et F1 (en bas), pour les modèles D-GMR, SC-GMR et IC-GMR

Ces résultats généraux semblent confirmer la tendance observée en section 3.5.2.1. Nous allons maintenant regarder en détail les performances des modèles par phonème. La Figure 33 présente les résultats sur les consonnes pour les deux locuteurs, et la Figure 34 ceux obtenus pour les voyelles. La RMSE moyenne de la *baseline* est représentée par un rectangle noir transparent, superposé sur la RMSE moyenne du modèle en généralisation, représentée en couleur. Pour chaque phonème ou groupe de phonème, nous présentons les performances des trois modèles.

Dans le cas des consonnes, nous constatons une très mauvaise généralisation de tous les modèles pour le groupe phonétique {ʃ,ʒ}. Le modèle IC-GMR généralise très bien pour {t,d,n} et {l}. Le D-GMR est pour sa part très mauvais dans l'estimation du {l}, pour les deux locuteurs. Pour le locuteur M1, le {r} est bien estimé par le IC-GMR contrairement

au D-GMR. Dans le cas des voyelles, de manière générale, les modèles présentent des *baselines* très proches. Cependant, le IC offre une meilleure capacité de généralisation avec des erreurs plus faibles que les deux autres modèles. La généralisation du {a} est globalement mauvaise. Pour le {i}, la performance des modèles lorsque ce phonème est connu (*baseline*) est la meilleure d'entre tous les phonèmes. Cependant, la performance se dégrade considérablement pour devenir au moins aussi mauvaise que celle des autres phonèmes lorsque nous évaluons la généralisation du modèle IC-GMR au {i}. Le D-GMR est particulièrement mauvais, avec des erreurs de généralisation supérieures à 7 mm. Le /i/ est un phonème dont l'articulation place la langue parallèle aux faisceaux échographiques, générant ainsi des images de très mauvaise qualité, avec un contour difficile à visualiser comme illustré en Figure 1. Nous pouvons supposer que la particularité des images leur permet d'être représentées par des gaussiennes spécifiques au sein du GMM, mais que dans le cas de la généralisation, le modèle est incapable d'estimer ces images, très différentes de celles qu'il a appris.

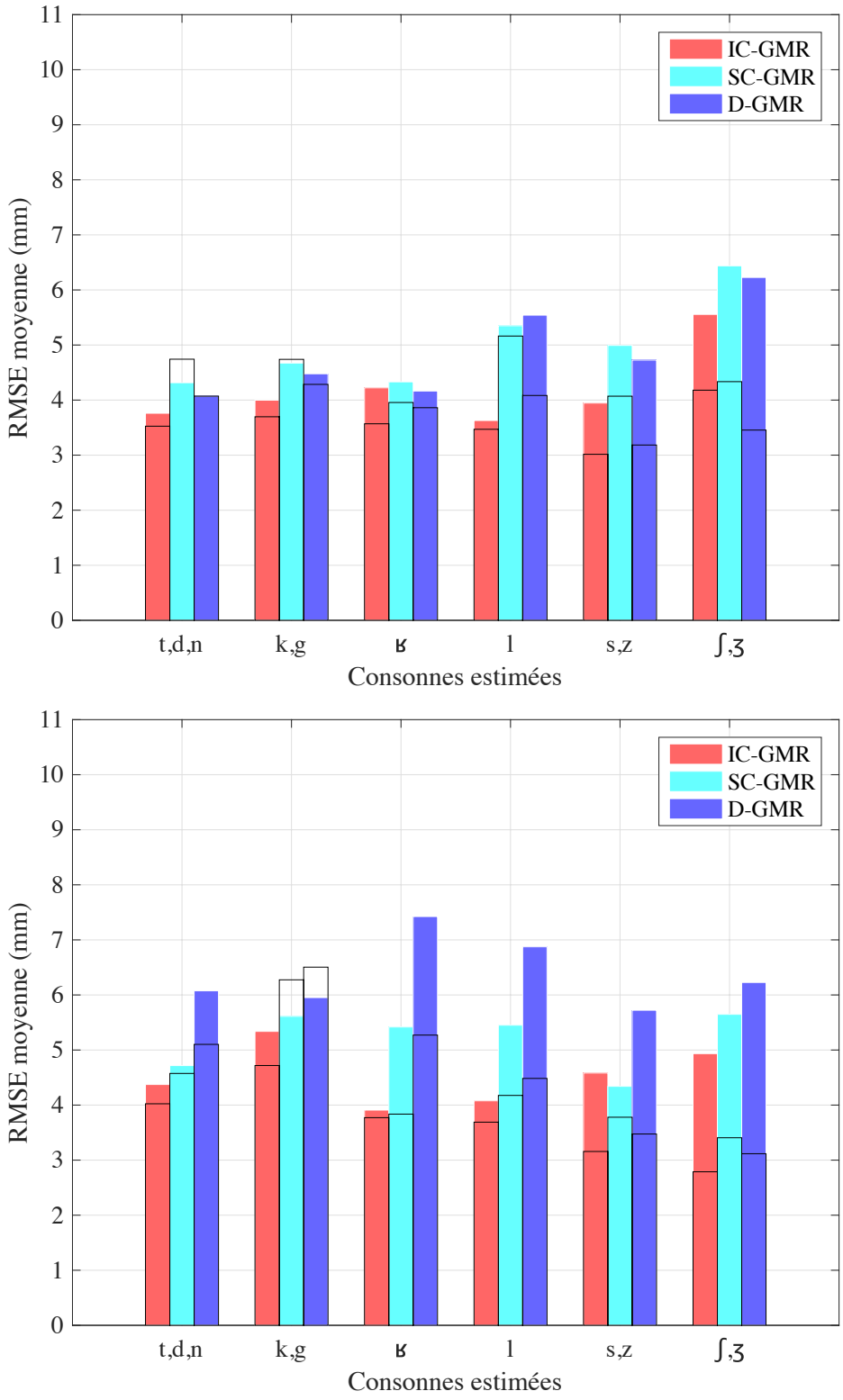


Figure 33. Moyenne des RMSE pour la généralisation sur les consonnes des trois modèles IC-GMR, SC-GMR, D-GMR, pour le locuteur M1 (en haut) et F1 (en bas)

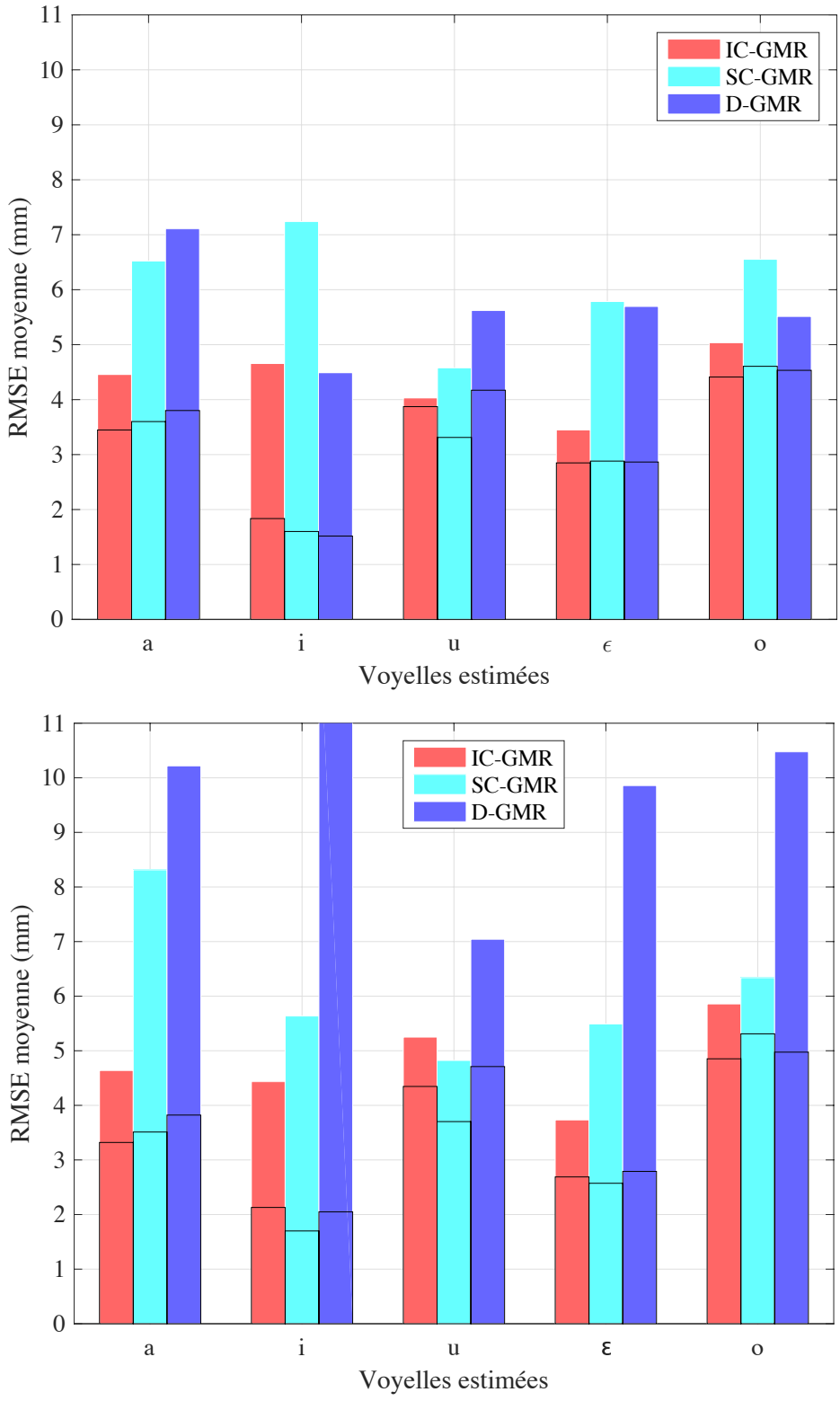


Figure 34. Moyenne des RMSE pour la généralisation sur les voyelles des trois modèles IC-GMR, SC-GMR, D-GMR, pour le locuteur M1 (en haut) et F1 (en bas).

3.5.3. Exemples illustratifs

Afin d'illustrer les expériences réalisées dans ce chapitre, nous proposons en Figure 35 deux exemples de VCV. L'image échographique de la langue a subi une rotation manuelle afin de pouvoir la comparer à son estimation dans la tête parlante. L'ouverture des lèvres a été déterminée de manière fixe (et non estimée) pour chaque séquence à partir les données acquises sur le locuteur de référence lors de la production d'un /a/.

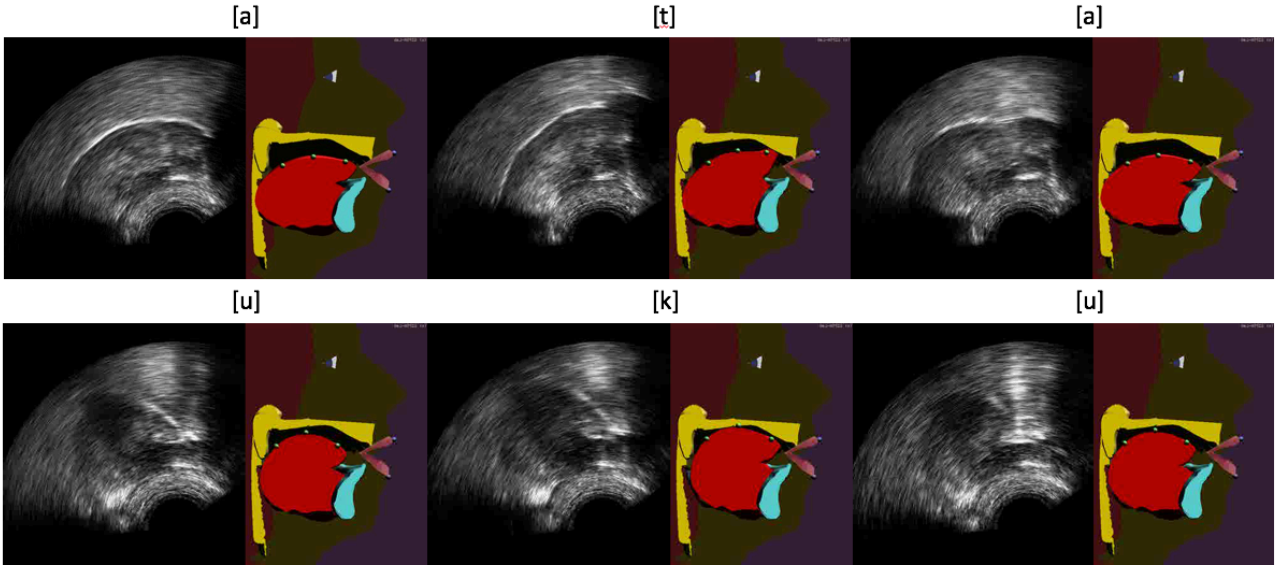


Figure 35. Exemples illustratifs d'animation du modèle de langue à partir d'images échographiques, estimées par le modèle IC-GMR.

3.6. Animation du modèle de lèvres et de la mâchoire de la tête parlante

Dans cette section, nous présentons des résultats préliminaires pour l'animation du modèle de lèvres de la tête parlante, à partir de vidéos du visage. Il s'agit de construire un système de retour visuel complet, renseignant à la fois sur les mouvements de la langue et sur ceux des lèvres et de la mâchoire, qui sont complémentaires. Une illustration du système complet est proposée en Figure 36.

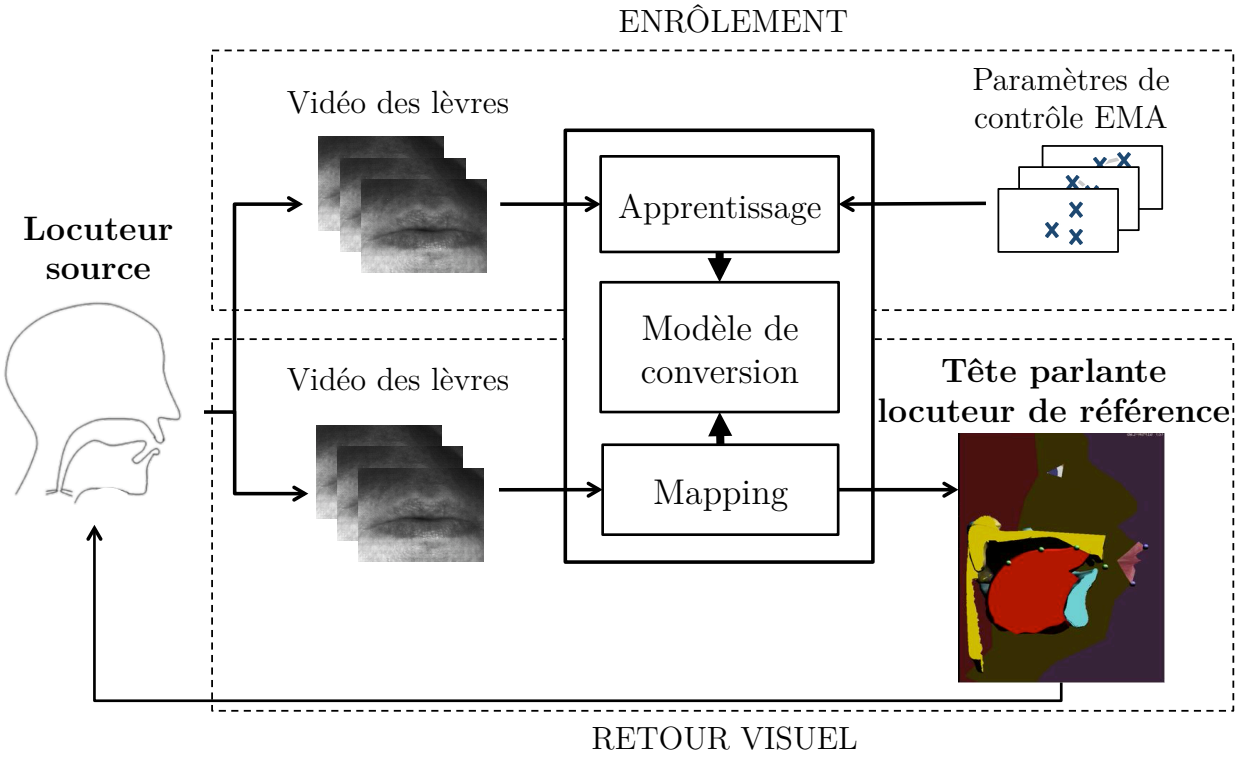


Figure 36. Animation automatique du modèle de lèvres et de mâchoire d'une tête parlante articulatoire à partir de vidéos des lèvres.

Nous adoptons ici la même approche que celle mise en place pour la langue dans les sections précédentes. Un enregistrement synchrone des données de visage est réalisé avec Ultraspeech. Nous réalisons ensuite un prétraitement des images des lèvres avec une sélection d'une région d'intérêt, puis une extraction de paramètres EigenLips, en réalisant une ACP sur un ensemble d'images d'apprentissage, de la même manière que pour les EigenTongues. Nous conservons 13 paramètres EigenLips, représentant 80 % de la variance.

Ayant pour objectif final d'estimer l'ensemble des paramètres de contrôle de la tête parlante, nous présentons un protocole expérimental en deux parties :

- Construction de deux GMM disjoints, un pour les lèvres+mâchoire et un pour la langue, dont les sorties peuvent être ensuite associées afin d'animer l'ensemble de la tête parlante
- Construction d'un GMM conjoint lèvres / langue / EMA avec en sortie les 12 paramètres de contrôle de la tête parlante (6 pour la langue, 4 pour les lèvres et 2 pour la mâchoire)

L'expérience est réalisée avec une validation croisée à cinq partitions sur l'ensemble du corpus comme détaillé en section 3.4.5. Tous les modèles sont des D-GMR construits avec M=16 gaussiennes.

Le corpus utilisé est celui du locuteur de *référence* des sections précédentes, auquel sont ajoutées les vidéos des lèvres pour chaque élément du corpus. Nous proposons donc ici de premiers résultats très préliminaires sur un seul locuteur. Les performances des modèles disjoints et du modèle conjoint sont comparées pour chaque paramètre. Le Tableau 6 répertorie les résultats obtenus en RMSE pour les trois modèles. Notons que dans ce cas, le modèle disjoint sur la langue correspond au modèle du locuteur de référence présenté en section 3.5.1.

Tableau 6. Performances du modèle de référence en RMSE moyenne et intervalle de confiance (à 95%) dans trois situations : colonne de gauche, en haut, modèle construit entre les échographiques de la langue et les paramètres de contrôle de la langue ; colonne de gauche, en bas, modèle construit entre les images des lèvres et les paramètres de contrôle de la mâchoire et des lèvres ; colonne de droite, modèle conjoint construit entre les données jointes échographies + lèvres et l'ensemble des paramètres de contrôle de la tête parlante. TT = Tongue Tip ; TD = Tongue Dorsum ; TB = Tongue Back ; J = Jaw ; UL = Upper Lip ; LL = Lower Lip.

		Modèles disjoints	Modèle conjoint
Paramètres de contrôle de la langue	TTx	2,31 ± 0,07	2,28 ± 0,06
	TTy	2,01 ± 0,05	1,88 ± 0,05
	TDx	1,90 ± 0,05	1,90 ± 0,06
	TDy	1,74 ± 0,06	1,75 ± 0,06
	TBx	1,65 ± 0,05	1,66 ± 0,05
	TBy	2,25 ± 0,06	2,25 ± 0,06
Paramètres de contrôle des lèvres et de la mâchoire	Jx	0,53 ± 0,01	0,52 ± 0,01
	Jy	1,41 ± 0,05	1,24 ± 0,04
	ULx	1,85 ± 0,07	2,04 ± 0,07
	ULy	1,64 ± 0,04	1,73 ± 0,04
	LLx	1,20 ± 0,04	1,13 ± 0,03
	LLy	2,08 ± 0,07	2,21 ± 0,07

Le modèle des lèvres + mâchoire présente une moyenne des RMSE de 1,63 mm. Pour rappel, le modèle de langue est à 2,2 mm. Le modèle conjoint présente quant à lui une

moyenne des RMSE de 1,91 mm. Nous constatons que les performances sont équivalentes sur les paramètres de contrôle de la langue lorsqu'on construit uniquement un modèle de langue ou lorsqu'on construit un modèle conjoint. La même observation peut être faite pour les lèvres et la mâchoire. Nous remarquons cependant un gain sur l'estimation de la coordonnée y de la pointe de langue. Nous pouvons faire l'hypothèse que des informations sur ce point peuvent être estimées à partir des vidéos des lèvres qui laissent entrevoir la pointe de la langue pour certaines articulations.

Ces premiers résultats sont intéressants et ouvrent des perspectives sur l'animation de l'ensemble de la tête parlante à partir des images échographiques et des vidéos des lèvres d'un locuteur. Cette étude doit être poursuivie dans le cas de l'adaptation à un nouveau locuteur pour confirmer sur l'ensemble des paramètres de contrôle de la tête parlante les résultats obtenus tout au long de ce chapitre.

3.7. Discussion et perspectives

Nous avons évalué différentes approches pour animer le modèle de langue d'une tête parlante articulatoire à partir d'une séquence d'images échographiques enregistrées sur un locuteur. La première est une conversion directe entre une représentation compacte des images (approche EigenTongues) et des paramètres de contrôle de la tête parlante, à l'aide de la technique GMR (approche D-GMR). La seconde consiste à adapter un modèle pré-entraîné sur le locuteur ayant servi à la construction de la tête parlante (locuteur de référence). Il s'agit de l'approche Cascaded-GMR, dont nous avons évalué les deux formes, le Split C-GMR et le Integrated C-GMR. Nous avons évalué les performances de ces différentes approches en fonction de la quantité de données d'enrôlement, et en évaluant sa capacité de généralisation à des phonèmes non vus durant l'enrôlement.

L'IC-GMR apparaît comme le modèle réalisant le meilleur compromis entre performance d'une part, avec une plus grande précision des mouvements linguaux estimés dans l'espace articulatoire de la tête parlante, quantité de données d'enrôlement et capacité de généralisation d'autre part.

Afin de poursuivre cette étude, il reste plusieurs angles d'amélioration. Tout d'abord, en enregistrant simultanément les données EMA et les données échographiques du locuteur de référence, nous nous libérerions de la contrainte de l'alignement par l'audio et garantirions une articulation rigoureusement similaire dans les deux modalités. Ceci pourra être rendu possible par la conception d'un casque de soutien de la sonde échographique réalisé par

impression 3D dans un matériau ne perturbant pas le champ magnétique du système EMA.

A travers une première expérience sur les données du locuteur de référence, nous pouvons considérer qu'il est pertinent d'animer également le modèle de lèvres et la mâchoire de la tête parlante à partir des mouvements du visage acquis par vidéo. Cela fournirait un système de retour visuel encore plus complet, qui permettrait au patient de situer ses mouvements linguaux par rapport aux autres articulateurs.

Enfin, il sera intéressant de valider ce système avec des patients dans le cadre d'une rééducation orthophonique présentant des troubles de l'articulation, ce qui n'est pas le cas ici. C'est ce à quoi nous nous attelons dans le prochain chapitre. Nous y utilisons cependant un système plus simple de retour visuel par échographie augmentée, le développement de l'approche basée sur la tête parlante n'ayant pu être abouti avant le début de l'essai clinique décrit dans le chapitre suivant.

Chapitre 4. Application clinique du retour visuel par échographie : études de cas sur des patients glossectomisés

Dans le premier chapitre, nous avons présenté différentes techniques de retour visuel pour l'orthophonie. Parmi ces techniques, nous avons fait le choix de l'échographie, qui semble aujourd'hui être une technique privilégiée pour la rééducation de certains troubles de l'articulation.
Ce travail vise à évaluer l'apport du retour visuel par échographie augmentée. Cependant, à la différence des études déjà réalisées à ce jour, nous souhaitons évaluer l'usage du retour visuel en temps réel de la langue du patient par rapport à l'usage de l'illustration visuelle, en nous appuyant sur des bilans orthophoniques classiques pratiqués entre chaque série de séances. **Nous faisons l'hypothèse que le retour visuel serait plus efficace que l'illustration pour la rééducation orthophonique des troubles de la parole et de l'articulation.** En effet, alors que l'illustration permet de visualiser l'articulation ciblée réalisée par un autre locuteur, et donc de comprendre l'exercice demandé, le retour visuel permet de visualiser sa propre articulation, de comprendre l'erreur et donc de la corriger en conséquence.
Dans ce chapitre, nous dresserons dans ce chapitre un état de l'art de l'utilisation de l'échographie pour la rééducation orthophonique. Nous présenterons ensuite une application de cette technologie à des patients ayant subi une glossectomie. Nous détaillerons pour cela le protocole mis en place et décrirons les résultats obtenus pour les cinq premiers patients inclus dans l'étude.

4.1. Rééducation orthophonique par échographie : état de l'art

La première étude référencée dans le domaine de la rééducation par échographie est celle de Shawker & Sonies (1985). Les auteurs proposent pour la première fois d'utiliser des images échographiques de la langue pour la rééducation du /ɹ/ anglais. Cette étude est une étude de cas, menée avec une patiente âgée de 9 ans. Un enregistrement de l'articulation cible et de l'audio correspondant sont diffusés en boucle de manière synchrone, pendant

que l'enfant répète et visualise sa propre articulation sur un second écran. Au vu des résultats obtenus, les auteurs émettent l'hypothèse que ce retour pourrait améliorer la parole des populations malentendantes, sous deux conditions : a) que le patient soit capable d'interpréter ce qu'il voit et b) pouvoir y consacrer suffisamment de séances.

Depuis cette première étude de cas, de nombreuses études ont été conduites, principalement au cours des dix dernières années. Ces travaux, essentiellement anglophones s'intéressent pour la plupart aux troubles articulatoires isolés, en particulier le /ɹ/ anglais, aux troubles de la parole chez les personnes malentendantes et à l'apprentissage d'une langue seconde. Nous insisterons ici sur les deux premiers domaines, le troisième ne relevant pas de la rééducation orthophonique même si le travail réalisé est similaire, comme en témoignent les plus récents travaux de Pillot-Loiseau *et al.* (2015) ou Wu *et al.* (2015). Le Tableau 7 détaille les études portant sur la rééducation orthophonique des troubles articulatoires et de la parole par échographie.

4.1.1. Troubles articulatoires isolés

Chez l'enfant, les troubles de la parole auxquels nous nous intéressons regroupent les troubles articulatoires et les troubles phonologiques définis dans le Chapitre 1. Tous concernent des enfants âgés de plus de 5 ans. Modha, Bernhardt *et al.* (2008), Jonathan L Preston, Brick *et al.* (2013), Cleland, Scobbie *et al.* (2015) et Jonathan L. Preston *et al.* (2016) traitent de la rééducation orthophonique de cette pathologie en adoptant des approches différentes. Par exemple, Cleland *et al.* (2015) mènent des études de cas auprès de sept enfants présentant des troubles de production de consonnes variées. Les progrès des enfants sont mesurés par des bilans orthophoniques réguliers. Cette étude souligne l'efficacité de ce retour visuel après 12 séances de rééducation, avec un maintien des performances plusieurs semaines après la fin des séances.

La rééducation du /ɹ/ anglais occupe une place importante parmi les demandes des patients chez les orthophonistes nord-américains. Ce phonème peut s'articuler de deux manières, provenant de divergences régionales. Adler-Bock, Bernhardt *et al.* (2007) nous informe plus particulièrement sur ces différences. En anglais nord-américain, il existe deux configurations possibles : soit la pointe de la langue est relevée et rétrofléchie, soit elle est vers le bas et rétractée. Les recherches montrent cependant que les locuteurs utilisent le plus souvent une configuration linguale se situant entre ces deux extrêmes. Souvent, au cours du développement de l'enfant, ce phonème est remplacé par le phonème /w/. Moins fréquemment, il peut être aussi remplacé par des fricatives ou des occlusives.

Adler-Bock *et al.* (2007) s'intéressent à deux adolescents de 12 et 13 ans présentant un trouble articulatoire résiduel sur cette consonne. Les patients de l'étude semblent remplacer le /ɹ/ par une approximante vélaire : le dos de la langue est trop haut et se rapproche trop du palais. Après 13 séances, ces patients sont parvenus à articuler correctement le /ɹ/ dans certains contextes, même si les productions spontanées ne présentent pas d'amélioration. Les auteurs préconisent donc une durée de rééducation plus longue. Bernhardt, Bacsfalvi *et al.* (2008) s'intéressent au même trouble pour un groupe de 13 enfants chez lesquels les méthodes traditionnelles de rééducation n'ont pas fonctionné. Chacun bénéficie d'une rééducation intensive : 3 à 4 séances réparties sur deux jours, pour un total de 2 à 3 heures de rééducation. Avant et après ces séances, les participants bénéficient de 7 à 8 séances traditionnelles chez un orthophoniste. Des progrès rapides sont observés lors des séances traditionnelles après échographie. Il en résulte une amélioration de la précision calculée sur l'ensemble des productions pour le phonème isolé ou dans un mot. Comme Adler-Bock *et al.* (2007), les auteurs suggèrent d'augmenter le nombre de séances afin d'améliorer la production en parole spontanée. Il est cependant difficile de confirmer que le retour visuel est à l'origine de ces progrès, plutôt que le nombre important de séances sur une courte durée. Il existe d'autres études de cas portant sur ce trouble, nous pouvons notamment citer celles de Jonathan L Preston & Leaman (2014) auprès d'une patiente de 59 ans victime d'un AVC ayant entraîné une aphasie de Broca, mais aussi celle de Byun, Hitchcock *et al.* (2014) portant sur un enfant de 11 ans ou encore celle de Cavin (2015) pour un homme de 22 ans, dont les seuls troubles articulatoires concernent ce phonème. Des progrès ont été observés pour chacun de ces patients.

4.1.2. Troubles de la parole : cas des personnes malentendantes de naissance

Les personnes sourdes, quel que soit leur mode de communication (signé ou oral) semblent plus sensibles aux indices visuels que les personnes entendantes (Muir & Richardson (2005)). En effet, habituées à la lecture labiale pour comprendre les personnes entendantes, les personnes sourdes peuvent aussi apprendre d'autres moyens de communication, en particulier les Langues des Signes et le Langue Parlé Complété (LPC). Les Langues des Signes, spécifiques à chaque pays ou chaque communauté linguistique, sont des langues visuelles dont il a été observé qu'elles améliorent la perception des mouvements du visage chez les apprenants (Muir & Richardson (2005)). Le LPC (Cornett (1967)) permet de compléter les mouvements des lèvres par ceux la main, dont la forme et la position par rapport au visage permettent de désambigüiser les syllabes prononcées.

A cause de l'altération de leur perception acoustique, les personnes sourdes implantées ou malentendantes qui souhaitent oraliser affichent un retard important dans leur acquisition du langage. Hudgins & Numbers (1942) étudient la parole de 192 enfants sourds moyens à profonds âgés de 8 à 20 ans. Les troubles de la parole observés chez cette population sont multiples. Cinq classes d'erreurs sur les voyelles sont ainsi recensées : nasalisation, substitution, voyelles neutres, voyelles articulées en deux voyelles, diphtongues articulées en deux voyelles distinctes ; ils notent également neuf classes d'erreurs sur les consonnes : suppression des consonnes en début de mot, insertions et suppressions des consonnes finales, dénasalisation, substitutions et insertions de voyelles, non voisement de /b/, /g/, /d/.

Plusieurs études ont été menées impliquant l'utilisation du retour visuel par échographie pour l'apprentissage de la parole chez les personnes sourdes dans le cadre orthophonique. Gallagher (2013) propose une critique intéressante de ces travaux, dans laquelle elle valide l'apport positif de l'échographie pour la population concernée en analysant les travaux de Adler-Bock *et al.* (2007), Bernhardt, Gick *et al.* (2003) et Bacsfalvi (2007). Gallagher (2013) souligne cependant la nécessité de poursuivre ces travaux sur des populations plus larges (> 7 personnes), en alternant les séances avec et sans retour visuel. L'auteur indique que l'échographie possède un important potentiel comme technique de retour visuel pour la rééducation des personnes sourdes, mais souligne la difficulté de réaliser des évaluations cohérentes d'une étude à l'autre. Ainsi, l'auteur propose par exemple de réaliser des études plus longues (> 10 séances), pour valider l'intelligibilité du discours à travers des mots, des phrases, et enfin des conversations. Les progrès pourraient être évalués par un auditeur naïf, la famille du patient ou encore par une auto-évaluation.

4.1.3. Résumé de la littérature

En résumé, la littérature sur le retour visuel par échographie reflète une grande diversité dans l'organisation et le contenu des séances de rééducation. Suivant les auteurs et les troubles étudiés, plusieurs procédures sont proposées : des sessions uniques, par exemple pour corriger un trouble articulatoire très spécifique, ou une série de sessions réparties sur 20 semaines lorsque le trouble est plus important. La durée d'une séance varie elle aussi, de 30 à 60 min, avec un temps consacré purement au retour visuel plus ou moins important. Les patients sont de tous âges - enfants, adolescents ou adultes - et la progression de leurs performances est évaluée sous forme d'études de cas, de petits groupes, ou de groupes plus conséquents pouvant inclure jusqu'à 30 participants. L'évaluation des performances est réalisée majoritairement à travers des bilans orthophoniques. Des analyses acoustiques (enregistrements audio) ou articulatoires (tracés

sur des images) complètent parfois ces bilans (Cleland *et al.* (2015)). La diversité des études se manifeste aussi dans les approches adoptées. Quand Bernhardt *et al.* (2008) alternent rééducation traditionnelle et retour visuel par échographie, Roxburgh *et al.* (2015) proposent une étude comparative entre le retour visuel échographique et l'illustration par une tête parlante articulatoire. Les travaux de Cleland *et al.* (2013) et Bacsfalvi, Bernhardt *et al.* (2007) s'orientent vers la comparaison des deux dispositifs de retour visuel principaux : l'EPG et l'échographie. Il est cependant encourageant de constater que l'ensemble de la littérature converge vers le même constat : l'utilisation de l'échographie semble accentuer l'impact positif de la rééducation orthophonique, comme le soulignent par exemple Jonathan L Preston *et al.* (2013). Cet apport reste cependant discret et difficile à mesurer. Il convient donc de mener des études de plus grande envergure et centrées sur des troubles spécifiques pour espérer recueillir des arguments convaincants, susceptibles de confirmer plus fermement l'apport du retour visuel par échographie.

Nous voyons à travers cette revue de la littérature que l'utilisation de l'échographie pour la rééducation orthophonique s'est accrue au cours des dernières années, notamment dans les pays anglo-saxons. Nous relevons en France un seul usage de cette technologie faisant l'objet d'une recherche publiée : le travail de Acher, Fabre *et al.* (2016) portant sur la rééducation chez des patients aphasiques à la suite d'un AVC. Les autres travaux impliquant l'échographie concernent l'apprentissage d'une L2, par exemple avec l'étude de Wu *et al.* (2015) ou encore celle de Pillot-Loiseau *et al.* (2015). Les différences entre les troubles traités et la façon d'aborder la rééducation par les orthophonistes dans ces travaux rendent difficiles la sélection des meilleurs paramètres. Pour notre part, nous avons sélectionné une population souffrant d'un déficit ciblé afin d'évaluer aussi spécifiquement que possible l'impact du retour échographique sur l'application clinique : les personnes ayant subi une ablation partielle d'une partie de la langue ou du plancher de la bouche. Les personnes incluses dans cette étude ne présentaient, avant opération, aucun trouble de la parole. Nous avons mené une étude dont les participants alternent une rééducation avec un retour visuel par échographie et une rééducation avec une illustration par Ultraspeech-player de l'articulation à réaliser (Chapitre 1 section 1.2.2) afin de comparer les effets de ces deux modes de visualisation. Ce chapitre présente une étude de cas sur cinq patients glossectomisés. Nous détaillons le protocole mis en place et les mesures réalisées avant de présenter les premiers résultats.

Tableau 7. Etat de l'art sur la rééducation des troubles de la parole par échographie. N représente la taille de l'échantillon

Auteurs	Design de l'étude	Phonème	Session	N	Participants	Mesures	Résultats
Shawker & Sonies (1985)	Etude de cas	/ɹ/	4 sessions 1 session = 1 h	1	9F substitution /r/ /w/	PCC sur 20 mots	Résultats encourageants, besoin d'ajuster les conditions de rééducation. Nécessite du patient une adaptation (interprétation de retours visuels)
Bernhardt *et al.* (2003)	Etudes de cas	/s/, /ʃ/, /l/, /ɹ/	14 sessions 1 session/semaine 1 session = 30 mn	4	Adolescents avec des troubles de l'audition	PCM et PVM	Avantage de la technologie peu convainquant
Adler-Bock *et al.* (2007)	Etude de cas multiples	/ɹ/	13 sessions 1 session = 1h	2	12 & 14 ans Canadiens anglophones	Analyse de formants acoustiques et jury d'écoute	Pas de gain sur le /r/ en conversation mais seulement dans des mots et phrases. Besoin de plus de séances
Bernhardt *et al.* (2008)	Etude Clinique de type expérimental	/ɹ/	3-4 avec échographie, sur deux jours, pour un total de 2h	13	7-15 (médiane 9)	% de réponses correctes	Progrès après seulement 1 à 3h de rééducation par échographie
Modha *et al.* (2008)	Etude de cas avec alternance de 3 types de prise en charge	/ɹ/	9 sessions 1 session = 30-45 min	1	13H	% de réponses correctes + analyse acoustique	Progrès sur le /r/ observés. Besoin d'une étude à plus grande échelle
Bacsfalvi & Bernhardt (2011)	Follow-up study Within-subject	fricatives, voyelles et /ɹ/ rhotique		7	14-19 ans avec troubles auditifs	Evaluation perceptive	Généralisation ou maintien des performances observé

	design Etude longitudinale de cas multiples						post-traitement
Jonathan L Preston *et al.* (2013)	Comparaison de traitements avec un objectif par bilan	Troubles articulatoires (CAS).	18 sessions 2 sessions / semaine 1 session = 1h	6	9-15 ans ; troubles articulatoires associé à l'apraxie de la parole chez l'enfant	Validité de 64 mots évalués à chaque session (focus sur le phonème traité)	Progrès significatifs et maintenus 2 mois post-traitement. Retour visuel par échographie pertinent pour les troubles articulatoires chez l'enfant
Jonathan L Preston & Leaman (2014)	Etude de cas	/ɹ/	12 sessions 2 sessions / semaine 1 session = 1h	1	59F avec aphasie de Broca (apraxia de la parole)	Bilans réguliers sur 60 mots (24 traités/36 inconnus) avec jury d'écoute	Résultats prometteurs, mais le clinicien manquait d'expérience dans l'utilisation de la sonde
Byun *et al.* (2014)	Etude de cas multiples	/ɹ/	14 sessions sur 8 semaines	8	Study 1 : 4 (2M, 2F, 6-11 yo) Study 2 : 4 ??	% de réponses correctes (jury d'écoute)	Retour visuel par échographique semble efficace avec des enfants pour lesquels rien d'autre ne fonctionne. Besoin d'optimiser l'utilisation, trouver les bons paramètres
Cleland *et al.* (2015)	Etude de cas multiples	/k/, /ɹ/, /ʃ/, /t/	12 sessions 1 session / semaine	7	6-11 ans troubles articulatoires	Validité de la réponse pour une articulation donnée : tracé sur l'image et jury d'écoute DEAP	Nouveauté bénéfique pour les enfants quand d'autres traitements ont échoué, plus explicite.

Roxburgh *et al.* (2015)	Etudes de cas	/n,s/ /g,k/	2 séances d'évaluation, 8 avec tête parlante, 2 d'évaluation, 8 avec la sonde, 2 d'évaluation	2	6 et 9 ans fente palatine	PCC PTCC DEAP	Progrès globaux sur le PTCC, progrès plus importants avec la tête parlante (Speech Trainer 3D). A confirmer avec une étude plus large échelle.
Jonathan L. Preston *et al.* (2016)	Comparaison de deux traitements (position de syllabe)	/ɹ/	14 sessions 1 session = 1h	3	10-13 ans troubles articulatoires (apraxie chez l'enfant)	PCC & PRC	L'échographie pourrait compliquer l'acquisition ou la généralisation des rhotiques pour certains enfants (CAS)
Cavin (2015)	Etude de cas	/ɹ/	9 sessions 1 session = 1h	1	22H	Analyse articulatoire, acoustique et auditive	Evidences qualitatives et quantitatives de progrès sur le /r/ Pas de période sans sonde pour comparer la progression
Blyth, McCabe *et al.* (2016)	Etudes de cas	/s,t/ /s,l,t,ʃ/	B-Phase: 12 sessions (3sessions/semaine) A-Phase: 4 sessions (1 session/semaine)	2	Glossectomie 53M et 59F	PCC Jury d'écoute (AsIDS) Commentaires des participants	Effet positif du retour visuel couplé aux principes moteurs sur la précision articulatoire. Nécessité de comprendre, entendre et aussi savoir interpréter l'image comme le retour auditif

4.2. Rééducation orthophonique de patients glossectomisés

4.2.1. Définition de la glossectomie

La glossectomie est une intervention chirurgicale à l'issue de laquelle une partie de la langue ou du plancher de la bouche est enlevée. Cette opération est généralement réalisée à la suite d'une lésion cancéreuse, et s'accompagne souvent de séances de radiothérapie parfois très lourdes pour les patients. Dans certains cas, la zone retirée est remplacée par un bout de muscle prélevé sur une autre partie du corps. A l'issue de cette opération, outre les difficultés à s'alimenter, le patient est confronté à un appareil vocal souvent fortement modifié, tant au niveau de sa géométrie, de ses muscles que de ses repères physiques ou sensoriels. La langue, après ce type d'opération, présente par ailleurs une perte de tonicité, de précision et de vitesse, qui peut nuire à la production de certains phonèmes (Perrier, Savariaux *et al.* (1999)) et générer des difficultés de déglutition. La thèse de Acher (2014) nous informe sur les altérations possibles associées à cette opération. Ainsi, les consonnes telles que le /s/, / ʃ/, /t/ et /k/ de même que /d/ et /n/ sont affectées par ces changements anatomiques. En effet, un bruit important est perceptible lors de la phase de tenu de ces consonnes occlusives. Une rééducation est alors nécessaire, afin d'aider le patient à maitriser ce nouvel espace articulatoire.

4.2.2. Littérature sur la rééducation de patients glossectomisés

De nombreuses approches sont utilisées par les orthophonistes pour la prise en charge de la rééducation articulatoire suite à des glossectomies. Certaines sont centrées sur la fonction motrice de la langue. Elles visent à compenser ses déficits par un travail musculaire (intensité, précision, rapidité du mouvement) : travail de recul de la langue au moyen d'exercices fonctionnels ou analytiques (articulation de mots contenant des phonèmes ciblés mais aussi recul de langue, balayage du palais, exercices de contre-résistance), travail de praxies linguales et plus largement bucco-linguo-faciales, sur commande et en imitation (miroir) (Acher (2009)). D'autres ciblent les fonctions sensorielles et proprioceptives. Woisard-Bassols & Puech (2011) développent différentes prises en charge. Ces approches peuvent passer par la thermothérapie, le travail des sensibilités (tactiles, gustatives, thermiques) ou celui, plus général, de l'oralité. La thérapie manuelle, les électrostimulations de la sphère ORL, l'explication par l'orthophoniste du geste articulatoire, la méthode Feldenkrais (1993) sont également utilisées dans la prise en charge des glossectomisés et permettent aux patients de mieux appréhender leur nouvelle morphologie intrabuccale et linguale.

En ce qui concerne le retour visuel par échographie, seule l'étude de Blyth, McCabe *et al.* (2016) semble à ce jour traiter de la rééducation par échographie chez des patients glossectomisés. Elle est réalisée sur deux participants, un homme et une femme d'une cinquantaine d'années. Leur rééducation se déroule en trois phases. Durant la première phase, des mesures sont réalisées une fois par semaine sur le patient afin d'établir une base de référence qui servira à évaluer ses progrès (cinq mesures en trois semaines pour un patient, et trois mesures en une semaine pour l'autre patient). Ensuite, 12 sessions, réparties sur quatre semaines, sont consacrées à la rééducation par échographie, avec des mesures hebdomadaires pour contrôler les progrès. Enfin, quatre sessions réparties sur quatre semaines permettent de mesurer la persistance des effets de la rééducation. L'évaluation était basée sur le pourcentage de consonnes correctes, et montre des progrès au cours des 12 séances. L'intelligibilité a aussi été mesurée, même si nous pouvons nous interroger sur la pertinence de cette mesure puisque le score obtenu par les patients était très élevé avant le début de la rééducation. Les auteurs soulignent que la pertinence de cette étude reste à démontrer sur un plus grand échantillon de participants. Les deux patients présentent de plus des profils différents. Outre la chirurgie qui diffère, la durée entre la chirurgie et la rééducation n'a pas été la même pour les deux patients : l'un a commencé 6 jours après, l'autre 5 mois après. Les patients, interrogés sur l'intérêt de la méthode, ont jugé utile le retour visuel pour améliorer leur articulation.

Cette recherche souligne un concept intéressant concernant le retour fourni par l'orthophoniste. Il en existe deux types : le retour par la connaissance sur l'exécution (*Knowlegde of Performance* ou KP) et le retour par la connaissance sur le résultat (*Knowledge of Results* ou KR). Jonathan L Preston & Leaman (2014) utilisent aussi cette distinction dans leur étude de cas sur l'apraxie de la parole à la suite d'un AVC.

Dans le cas du retour articulatoire par échographie, Ballard, Smith *et al.* (2012) semblent être les premiers à discuter de ces deux types de retour pour l'apprentissage ou la rééducation d'habiletés motrices en parole. Le retour par la *connaissance sur l'exécution* (*Knowledge of Performance (KP) feedback*), encore appelé *retour cinématique*, fournit des informations détaillées *externes* sur la nature du mouvement, comme le déplacement des articulateurs, et sur l'écart d'exécution par rapport au mouvement cible en termes d'amplitude ou de qualité. Ces informations permettent ainsi de guider un apprentissage ou une réhabilitation dynamique des habiletés motrices. En revanche, le retour par la *connaissance sur le résultat* (*Knowledge of Results (KR) feedback*) indique seulement si l'objectif de la tâche a été atteint ou non, par exemple, pour un phonème spécifique, si la production est correcte ou incorrecte. KP facilite l'acquisition de compétences motrices qui

concernent la parole ou d'autres actions, alors que KR aide généralement à l'apprentissage moteur (Maas, Robin *et al.* (2008)). Blyth *et al.* (2016) rappellent l'hypothèse de Ballard *et al.* (2012) et Newell, Carlton *et al.* (1990) qui suggèrent que le retour KP est bénéfique quand l'apprenant n'a pas une solide représentation interne du mouvement nécessaire pour atteindre le résultat désiré, comme un but acoustique en parole. Blyth *et al.* (2016) font l'hypothèse que c'est le cas des patients ayant subi une glossectomie partielle qui ont des difficultés à retrouver une parole claire, leur représentation sensori-motrice interne ne correspondant plus au nouvel état de leur langue.
Notre étude vise à évaluer l'utilisation de l'imagerie échographique de la langue comme outil de retour de type KP. Nous proposons ainsi, à la suite de cet état de l'art une nouvelle étude sur la rééducation des personnes glossectomisées par échographie. Nous détaillons le protocole de rééducation proposé, ainsi que les résultats obtenus sur cinq patients.

4.3. Protocoles de rééducation proposés

Cette étude est réalisée en collaboration avec le Centre Médical Rocheplane de Saint-Martin-d'Hères avec l'accord du Comité de Protection des Personnes (CPP) - Lyon Sud Est II (69HCL15_0736). Deux orthophonistes participent à ce projet. Elles sont responsables par alternance de la rééducation des patients, en fonction de leurs emplois du temps.

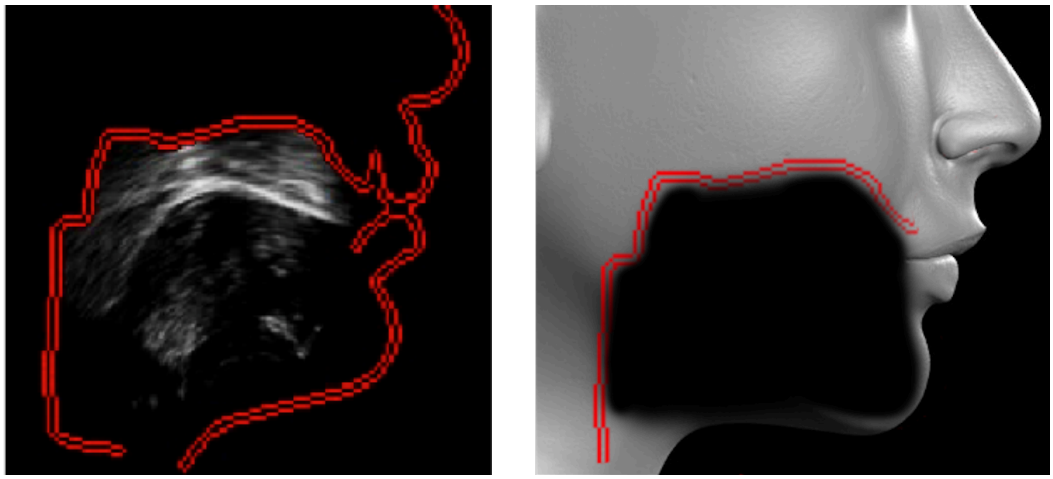
4.3.1. Dispositifs d'aide à la rééducation : sonde échographique et Ultraspeech-player

Dans cette étude, nous proposons aux patients une alternance entre deux protocoles lors des séances de rééducation : un protocole RETOUR VISUEL (R) et un mode ILLUSTRATION VISUELLE (I).

Durant les séances ILLUSTRATION, l'orthophoniste comme le patient n'ont aucune information visuelle directe sur l'articulation réalisée, comme lors des séances de rééducation traditionnelles. A la place, une illustration de l'articulation cible à produire est proposée. Dans ce but, nous avons enregistré une base de données de mots prononcés par une orthophoniste. Ces données audiovisuelles, qui servent de référence pour l'illustration, peuvent être présentées grâce au logiciel Ultraspeech-player présenté au Chapitre 1.

Durant les séances de rééducation avec le protocole de RETOUR VISUEL, le patient est informé par l'image échographique de la position et de la forme de sa langue. Afin d'améliorer le rendu proposé par la sonde, plusieurs images peuvent être superposées à

l'image échographique (logiciel intitulé Ultraspeech-biofeedback). Ces images incluent un contour délimitant la cavité buccale identique à celui affiché par Ultraspeech-player. Un espace transparent au niveau de la cavité buccale permet par ailleurs la visualisation de l'image échographique du patient. Cependant, la morphologie du patient peut différer de celle affichée. En effet, le contour de la cavité vocale est tracé à partir de l'IRM d'un locuteur avec une anatomie qui lui est propre. L'image échographique doit donc être normalisée par rapport à ce locuteur. Pour chaque patient, au cours de la première séance, des modifications sur le positionnement de l'image par rapport à ce conduit vocal sont donc réalisées grâce à plusieurs paramètres (orientation, zoom, translation). Les paramètres optimaux de transformation peuvent alors être enregistrés et ainsi réutilisés à chaque séance. Ils peuvent de plus être modifiés, par exemple si le patient tient la sonde différemment d'une fois à l'autre. Ces deux modes de visualisation sont illustrés à la Figure 37.



(a) Illustration : Ultraspeech-player (b) Retour : Ultraspeech-biofeedback

Figure 37. Représentation des deux modes de visualisation disponibles (a) l'illustration proposée par Ultraspeech-player avec l'échographie de la langue de l'orthophoniste (b) un des contours proposés par Ultraspeech-biofeedback : la langue du patient est placée dans la zone transparente et normalisée par des transformations pour s'adapter à la forme du conduit vocal. Le même contour de palais est affiché dans les deux protocoles.

Dans notre protocole, afin d'avoir exactement le même affichage lors de la session ILLUSTRATION et de la session RETOUR, plutôt que d'utiliser le contour avec le visage illustré en Figure 37b, les orthophonistes ont choisi comme contour pour Ultraspeech-biofeedback le même que celui affiché dans Ultraspeech-player illustré en Figure 37a. De plus, durant les sessions RETOUR, il était toujours proposé au patient d'afficher le logiciel d'illustration en complément de son propre retour visuel.

4.3.2. Protocoles de rééducation

Les séances de rééducation sont (bi)quotidiennes, week-end exclus, et durent 30 min. Au cours de ces séances, l'orthophoniste travaille durant 15 minutes sur des praxies bucco-linguo-faciales, puis 15 minutes avec l'un des supports visuels. Des mots, issus d'une banque de données détaillée en Annexe A, permettent de travailler, à chaque séance, les phonèmes ou groupes de phonèmes suivants : /t/, /k/, /kt/, /kl/, /st/, /sk/, /b d g/, /p t k b d g/, ainsi que des phrases contenant les phonèmes /p t k/, /b d g/ et /p t k b d g/.
Blyth *et al.* (2016) alternent de manière planifiée les retours de type KP et KR. Dans notre étude, les orthophonistes, indépendamment du support utilisé, ont gardé les habitudes de pratique prises dans le cadre des rééducations traditionnelles. Le type de retour fourni au patient dépend de ce dont il a besoin au cours de la rééducation, avec une alternance possible au sein d'une séance.

4.3.3. Evaluation des progrès par des bilans orthophoniques

Deux types de bilans orthophoniques sont réalisés pour évaluer les progrès des patients : le Bilan de Motricité Bucco-Linguo-Faciale (MBLF) adulte, extension de celui utilisé chez l'enfant (Gatignol, Troadec *et al.* (2013)) et les épreuves de la Batterie d'Evaluation Clinique de la Dysarthrie (BECD) de Auzou & Rolland-Monnoury (2006).

Le MBLF est un bilan des praxies bucco-linguo-faciales permettant d'observer la coordination musculaire et le degré d'atteinte des muscles linguaux et faciaux. Il est basé sur la reproduction, sur ordre verbal et / ou visuel, de mouvements de la face, des yeux, des lèvres, des joues, des mandibules et de la langue. Ce bilan permet d'évaluer les muscles nécessaires à l'articulation, la mimique et la déglutition. Chacun des mouvements est noté sur une échelle discrète de 0 à 3, où 3 correspond à une exécution correcte du geste.

La BECD propose plusieurs épreuves permettant d'évaluer le degré de trouble de l'articulation : évaluation de la sévérité, analyse perceptive, analyse phonétique, examen moteur, auto-évaluation, analyse acoustique. Nous nous intéressons en particulier dans ce bilan aux performances du patient sur différents phonèmes isolés et sur des mots. Notons que les productions sonores des patients lors de ces épreuves sont enregistrées pour une utilisation ultérieure dans des analyses acoustiques et dans des évaluations par un jury d'écoute.

Ces deux bilans nous permettent de mesurer l'efficacité des séances de rééducation en calculant la différence entre les scores aux bilans orthophoniques avant et après des

séances RETOUR et ILLUSTRATION. Nous avons relevé en particulier trois scores donc les variations pouvaient être pertinentes pour caractériser les progrès réalisés plus spécifiquement sur la langue :
- le score de la langue pour le bilan de MBLF (/39) : il s'agit de 13 mouvements, notés de 0 à 3. Ce score est une variable discrète qui dépend de l'opinion experte de l'orthophoniste. Tous les mouvements présentent la même difficulté et peuvent être considérés comme des tâches de difficultés équivalentes. Ainsi, il est possible de calculer la moyenne par bilan et par participant des scores obtenus pour l'ensemble des mouvements.
- l'intelligibilité sur des mots, phrases et en conversation de la BECD (/24) : score global sur 8 dépendant du jugement de l'orthophoniste pour trois catégories : un ensemble de mots, un ensemble de phrases et une conversation.
- le score au Test Phonétique d'Intelligibilité (TPI) de la BECD (/52) : 52 mots, l'orthophoniste note le mot qui, parmi une liste de quatre mots, se rapproche le plus de la production du patient. Les mêmes mots sont prononcés d'un bilan à l'autre ; il est donc possible de comparer les scores. De plus, ce bilan nous informe sur les types d'erreurs, réparties en treize catégories (notées A à M), réalisées par le patient : pour chaque mot, deux types d'erreurs peuvent être réalisées.

Deux orthophonistes s'occupent de la prise en charge des patients, mais une seule des deux réalise l'ensemble des bilans par souci d'homogénéité dans l'évaluation.

4.3.4. Déroulement de l'étude clinique

Le Tableau 8 représente le déroulement des séances de rééducation pour les patients. Les patients sont répartis en 2 groupes : le groupe RI (RETOUR puis ILLUSTRATION) et le groupe IR (ILLUSTRATION puis RETOUR). Pour chaque protocole, des séries de dix séances sont consacrées à la rééducation, suivies d'un bilan orthophonique permettant d'évaluer les progrès. Les mêmes bilans orthophoniques sont donc effectués à T0, T0 + 10 séances, et T0 + 20 séances. Les séances de rééducation orthophonique seront réalisées à raison de 4 à 5 séances par semaine. Les patients sont affectés au groupe RI ou au groupe IR en fonction de leur score au bilan à T0, en essayant d'avoir une moyenne équivalente dans les deux groupes.

Tableau 8. Déroulement des séances de rééducation. Le premier bilan est réalisé à T0 et marque le début de la prise en charge.

	T0-3 jours	**T0** (début)		**T1** (T0+10)		**T2** (T0+20)
Examen clinique	X					
Antécédents (CRF)	X					
Information du patient	X					
Bilan orthophonique		X		X		X
Consentement éclairé		X				
Groupe RI			**RETOUR**		**ILLU**	
Groupe IR			**ILLU**		**RETOUR**	

Dans la suite de ce manuscrit, une série de dix séances pour un protocole donné sera appelée « session RETOUR » ou « session ILLUSTRATION ».

4.4. Résultats sur cinq patients ayant subi une chirurgie bucco-pharyngienne

Nous rappelons ici notre hypothèse initiale : le patient progresserait mieux avec un retour visuel sur sa propre articulation qu'avec une illustration de l'articulation d'un locuteur de référence. Ouni (2014) constate qu'une première étape d'apprentissage par retour visuel avant un apprentissage de nouveaux mouvements par illustration semble améliorer les performances des apprenants. Pour que ce constat soit répliqué, il faudrait que nous observions une progression plus importante au bilan T1 après une première session RETOUR qu'après une session ILLUSTRATION. Il pourra être intéressant de voir ensuite comment l'évolution se poursuit lorsque les sessions sont inversées. Cependant, gardons à l'esprit que nous analysons les résultats obtenus auprès de cinq patients, et que seule une étude plus large permettra de tirer de véritables conclusions sur ces effets d'ordre.

4.4.1. Population de l'étude

Nous recrutons donc les participants à notre étude parmi les patients du Centre Médical Rocheplane. Afin de limiter les biais dans l'étude, nous avons défini des critères d'inclusion et d'exclusion de ces patients. Le Tableau 9 détaille ces critères.

Tableau 9. Critères d'inclusion et d'exclusion des patients.

CRITERES D'INCLUSION	Présence de troubles articulatoires de la parole pour lesquels une rééducation orthophonique est prescrite.
	Langue maternelle française
	Patients adultes [18 ≥ âge < 80 ans] ayant subi une chirurgie bucco-pharyngée dans le cadre d'un cancer de la langue et/ou du plancher de bouche
CRITERES D'EXCLUSION	Incapacité à comprendre facilement les consignes de l'orthophoniste.
	Mineur ou majeur protégé par la loi
	Personne privée de liberté par décision judiciaire ou administrative, personne faisant l'objet d'une mesure de protection légale
	Allergie potentielle au gel utilisé pour la sonde
	Surpoids si grosse masse adipeuse en sous-mentonnier
	Œdème sous- mentonnier
	Trouble vision / audition non corrigé
	Trouble neurologique d'origine centrale ou périphérique (aphasie, dysarthrie, paralysie faciale) / maladie neurodégénérative / psychiatrique
	Antécédents de chirurgie ou radiothérapie de la sphère ORL
	Trouble articulatoire isolé ou trouble de la parole antérieur rééduqué ou non, trouble de la fluence (bégaiement)
	Dyspraxie orofaciale massive, en particulier sur la langue
	Trouble de la posture
	Problème moteur aux membres supérieurs

A ce jour, seuls deux patients ont été exclus car ne rentrant pas dans les critères. Cinq patients, nommés par la suite IR001, RI002, RI003, IR004 et IR005 par souci d'anonymat, ont participé à l'ensemble de l'étude. Les informations sur ces patients sont présentées dans le Tableau 10. Le patient IR001 a subi une Bucco-Pharyngectomie trans-mandibulaire, soit l'ablation d'une partie du fond de la bouche ; le patient RI002 une hémiglossectomie, soit l'ablation de la moitié de la langue, avec reconstruction par lambeau musculocutané ; le patient RI003 une pelvectomie, soit l'ablation du plancher buccal ; le patient IR004 une glossectomie partielle avec reconstruction par lambeau ; enfin, le patient IR005 a subi une pelvectomie antérieure gauche avec reconstruction par lambeau. La rééducation au centre commence trois semaines environ après l'opération. La diversité des opérations nous amène à réaliser cinq études de cas.

Tableau 10. Informations générales sur les patients et scores des bilans orthophoniques obtenus à T0 avant le début de la rééducation. I = ILLUSTRATION ; R = RETOUR VISUEL.

Infos générales	**IR001**	**IR004**	**IR005**	**RI002**	**RI003**	
Âge et sexe	47H	58F	58H	54H	52F	
Ordre de rééducation	I-R	I-R	I-R	R-I	R-I	
Total MBLF + BECD	**259**	**269**	**273**	**274**	**253**	**/332**
MBLF	**62**	**77**	**76**	**74**	**60**	**/105**
Dont langue	16	13	13	19	19	/33
BECD - synthèse globale	**197**	**192**	**197**	**200**	**193**	**/227**
BECD - Intelligibilité	12	14	13	13	14	**/24**
Mots	3	4	4	3	4	/8
Phrases	4	4	4	4	4	/8
Conversation	5	6	5	6	6	/8
BECD - TPI	**41**	**46**	**45**	**43**	**46**	**/52**
BECD - Analyse des erreurs phonétiques	**144**	**132**	**139**	**144**	**133**	**/151**
BECD - Erreurs phonétiques isolées	32	28	32	28	28	/33
BECD - Mots simples	86	82	82	87	81	/88
BECD - Mots complexes	26	22	25	29	24	/30

L'ensemble des résultats pour le bilan à T0, ainsi que pour les bilans à T1 et T2 est présenté en Annexe B.

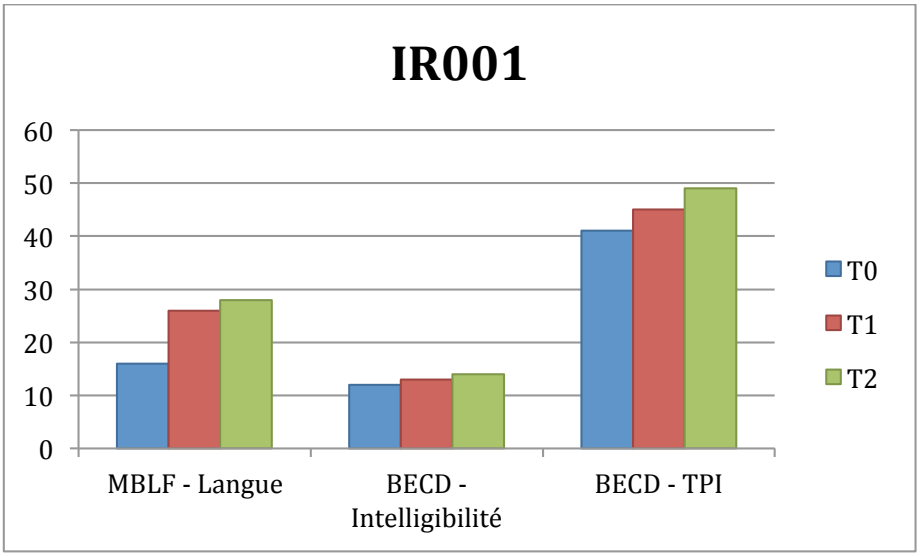
4.4.2. Analyse descriptive des bilans

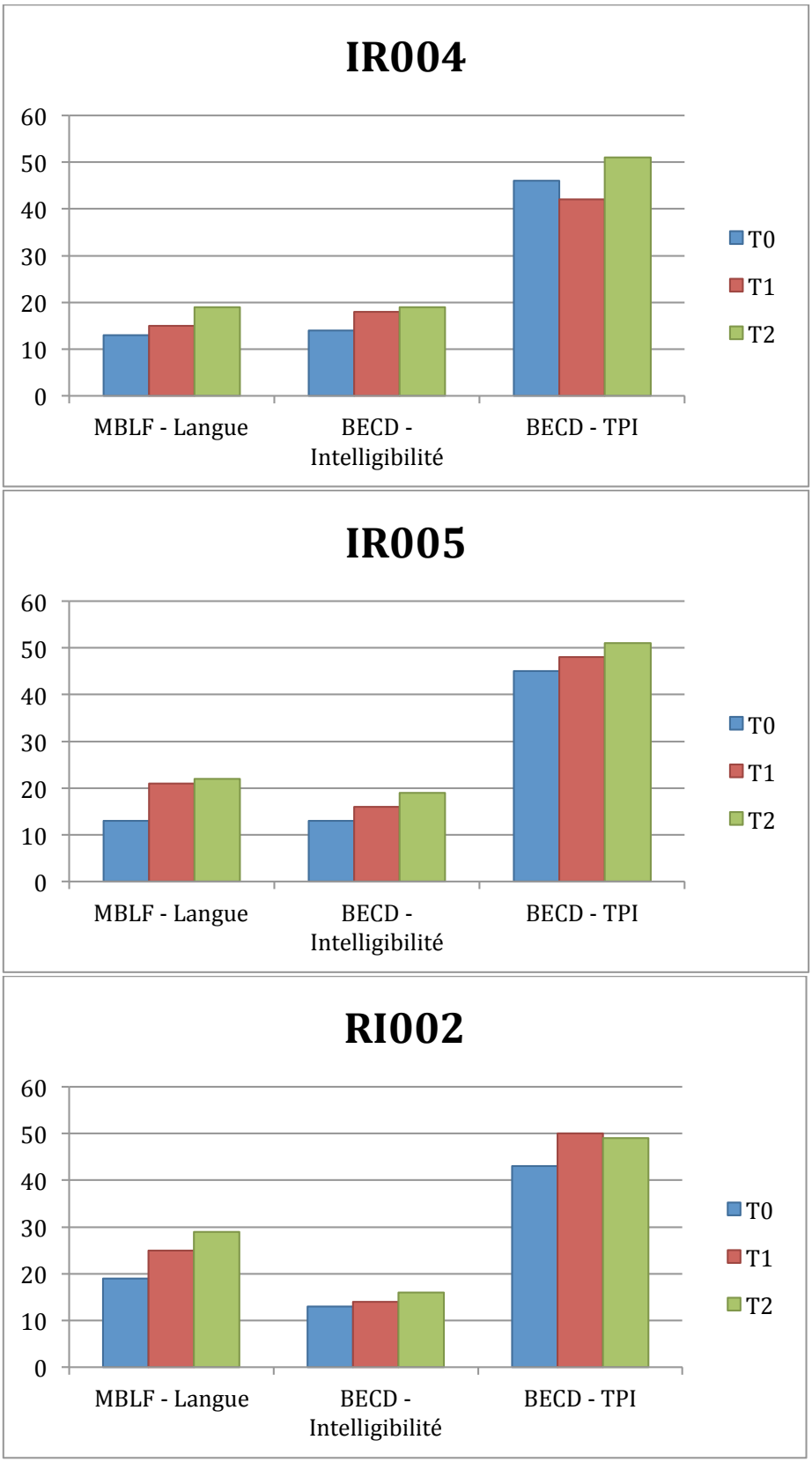
Dans cette section, nous présentons les performances des cinq patients, à partir de leurs bilans à T0, T1 et T2. Le Tableau 11 reporte les scores totaux par bilan et par patient.

Tableau 11. Score total par bilan et par patient.

Score /338	T0	T1	T2
IR001	259	285	300
IR004	269	280	296
IR005	273	296	305
RI002	274	293	301
RI003	253	285	308

Nous voyons dans ce tableau que le patient RI002, dont le score est le plus élevé à T0, présente la plus faible progression globale. RI003 est le patient présentant la progression la plus importante. Comme mentionné en section 4.3.3, nous avons sélectionné en particulier trois bilans pour mieux quantifier les progrès sur l'articulation de la langue. La Figure 38 reporte ces scores pour les trois bilans : à gauche, le score de la langue pour le bilan de MBLF ; au milieu, l'intelligibilité sur des mots, phrases et en conversation de la BECD ; à droite, le score au TPI de la BECD.




IR004
60
50
40
30
20
10
0
MBLF - Langue
BECD - Intelligibilité
BECD - TPI
T0
T1
T2
IR005
60
50
40
30
20
10
0
MBLF - Langue
BECD - Intelligibilité
BECD - TPI
T0
T1
T2
RI002
60
50
40
30
20
10
0
MBLF - Langue
BECD - Intelligibilité
BECD - TPI
T0
T1
T2

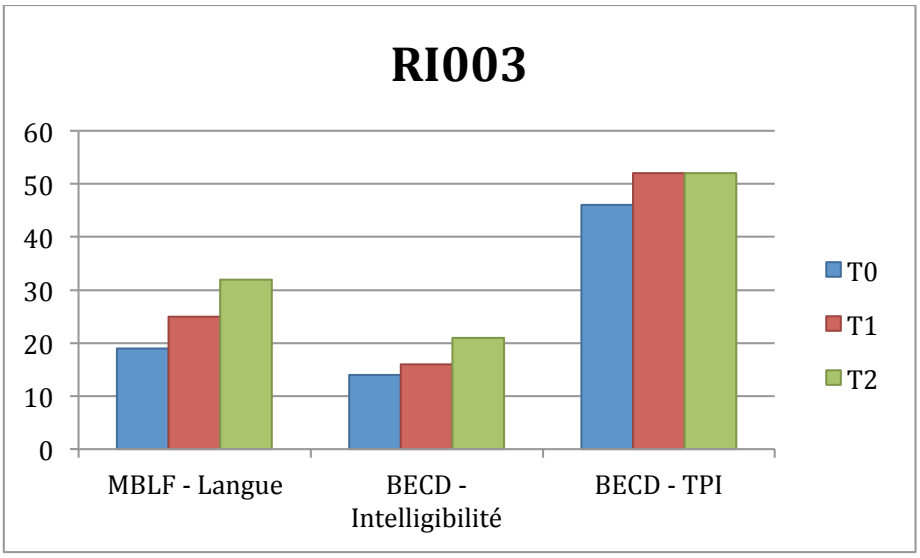


Figure 38. Scores obtenus pour trois bilans de T0 à T2 par IR001, RI002 et RI003 (IR = ILLUSTRATION puis RETOUR ; RI = RETOUR puis ILLUSTRATION).

Tous les patients affichent des progrès au cours de la rééducation et pour chaque type de bilan. Afin de quantifier au mieux les progrès réalisés, nous allons étudier ces trois bilans plus en détail dans la suite de ce chapitre.

Ainsi, la Figure 39 reporte les performances moyennes sur la langue de chacun des patients à chaque bilan. La moyenne est calculée sur les 13 éléments (comme « tirer la langue » ou « élever la pointe dans la bouche »), notés chacun entre 0 et 3, et l'intervalle de confiance est calculé avec un test de Student pour α=0,05.

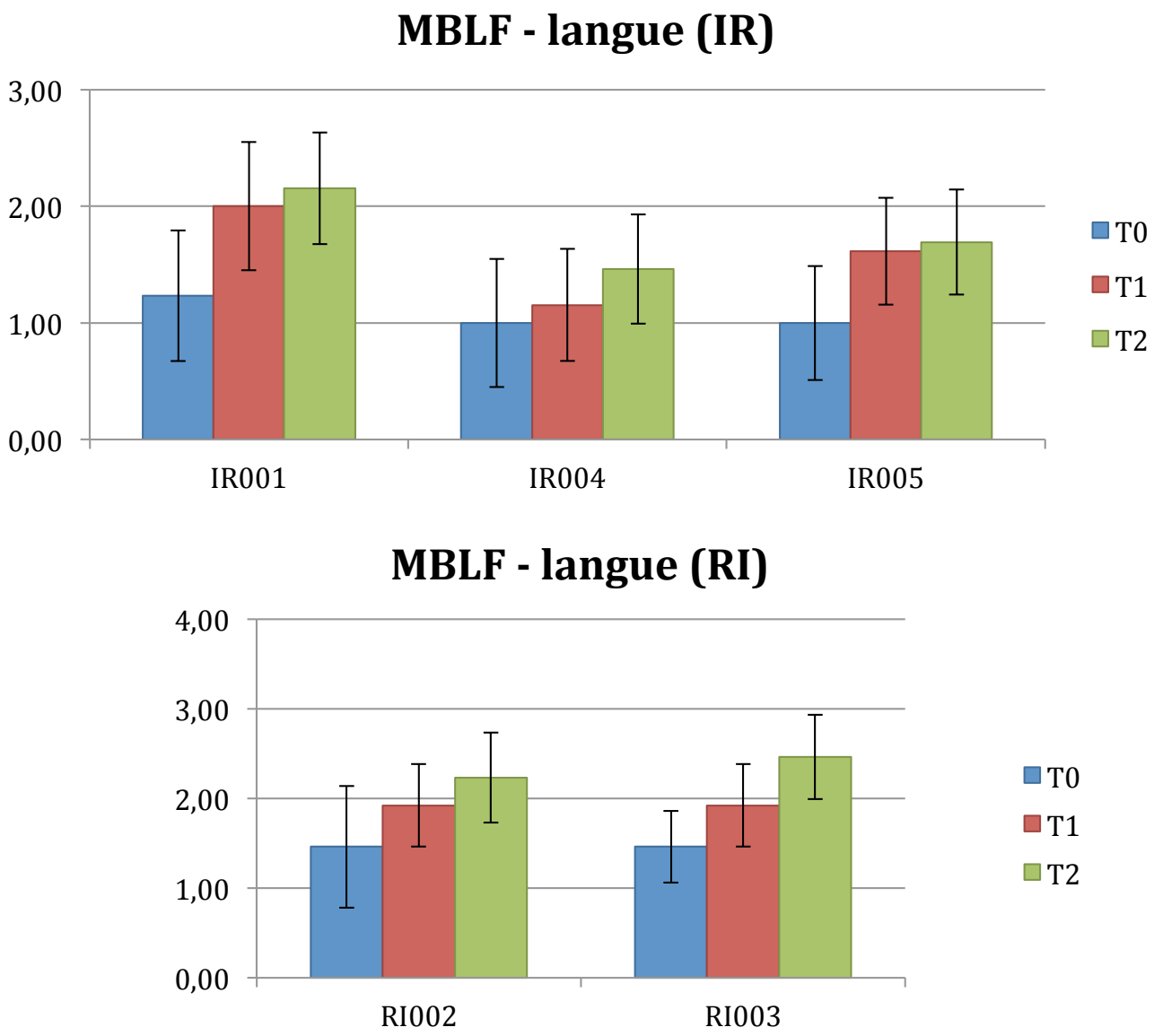


Figure 39. Score moyen par bilan et par patient au MBLF pour la langue. Les intervalles de confiance sont indiqués pour chaque valeur (α=0,05).

Les cinq patients présentent une progression sur la moyenne des scores obtenus au bilan de MBLF – langue au cours des bilans. IR001 présente une progression importante entre T0 et T1, avec une moyenne de 2 à l'issue de la session ILLUSTRATION. Les performances de ce patient se maintiennent ensuite sans réel progrès, avec une moyenne de 2,15 à T2, à l'issue de la session RETOUR. Les patients IR004 et IR005, appartenant au même groupe IR, présentent tous deux une moyenne de 1 à T0 et leurs performances à T2 sont très en dessous de celles de IR001 (1,46 pour IR004 et 1,69 pour IR005). RI002 et RI003, tous deux bénéficiant de l'ordre inverse de rééducation (RETOUR puis ILLUSTRATION) affichent une progression plus régulière. Leurs moyennes augmentent 1,46 à 1,92 pour tous

les deux entre T0 à T1. RI002 atteint une moyenne de 2,23 à T2. RI003 est le patient qui présente la meilleure moyenne à T2 (2,46/3). L'intervalle de confiance pour les cinq patients diminue entre T0 et T2, preuve d'une plus grande constance dans les scores obtenus pour chaque tâche. Nous avons utilisé des comparaisons multiples pour évaluer la significativité statistique des différences entre des moyennes du bilan MBLF-langue à T0, T1 et T2 pour chaque patient. Avec une p-value $< 0,05$, nous observons des progrès significatifs entre T0 et T2 pour IR001 ($p = 0,039$) IR005 ($p = 0,049$) et RI003 ($p = 0,006$). Le Tableau 12 contient les scores obtenus pour le test global d'intelligibilité de la BECD. Pour chaque patient et pour chaque bilan, le score sur 8 est reporté pour les mots, les phrases et la conversation. Nous pouvons voir que RI003 est globalement meilleur que les autres, quels que soient le bilan et l'épreuve considérés, avec une intelligibilité presque parfaite obtenue à T2.

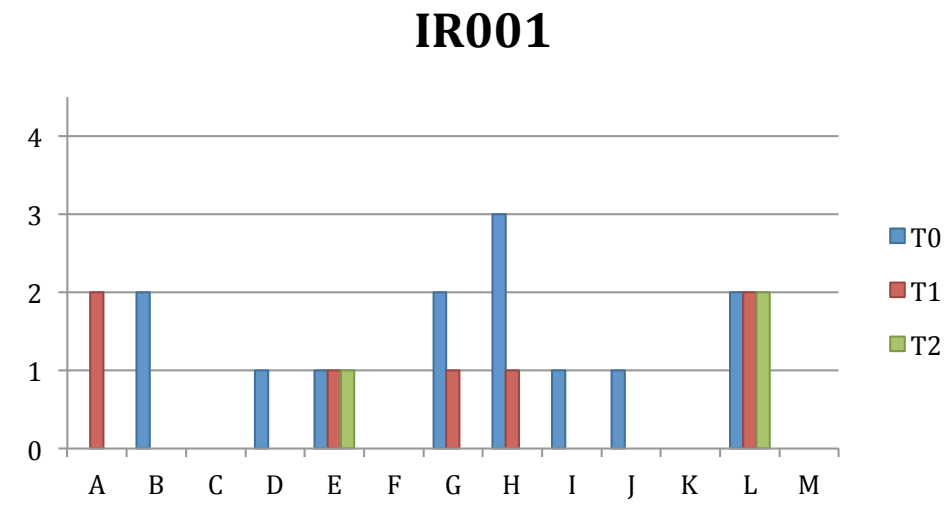
Tableau 12. Score au test d'intelligibilité pour les cinq patients.

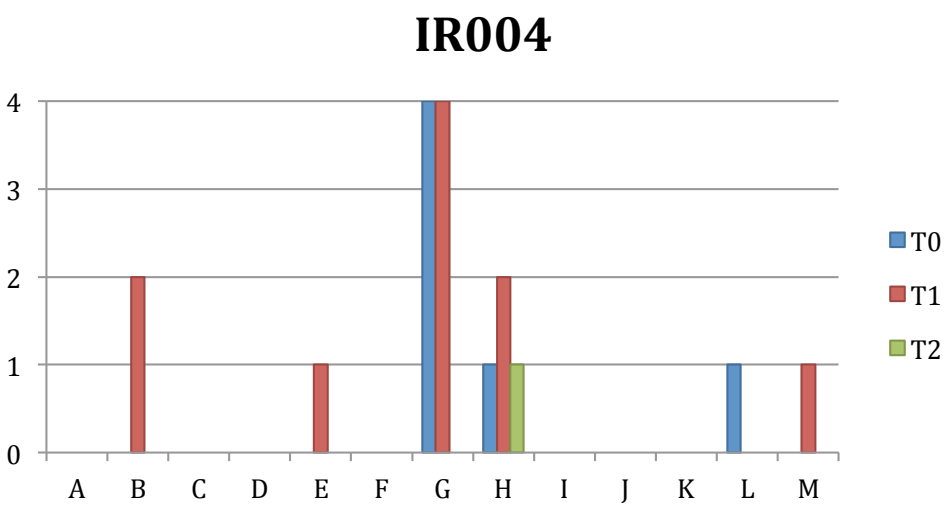
	IR001			**IR004**			**IR005**			**RI002**			**RI003**		
Bilan	**T0**	**T1**	**T2**	**T0**	**T1**	**T2**	**T0**	**T1**	**T2**	**T0**	**T1**	**T2**	**T0**	**T1**	**T2**
Mots (/8)	3	4	4	4	6	6	4	4	6	3	4	6	4	4	7
Phrases (/8)	4	4	4	4	6	7	4	6	7	4	4	4	4	6	7
Conversation (/8)	5	5	6	6	6	6	5	6	6	6	6	6	6	6	7

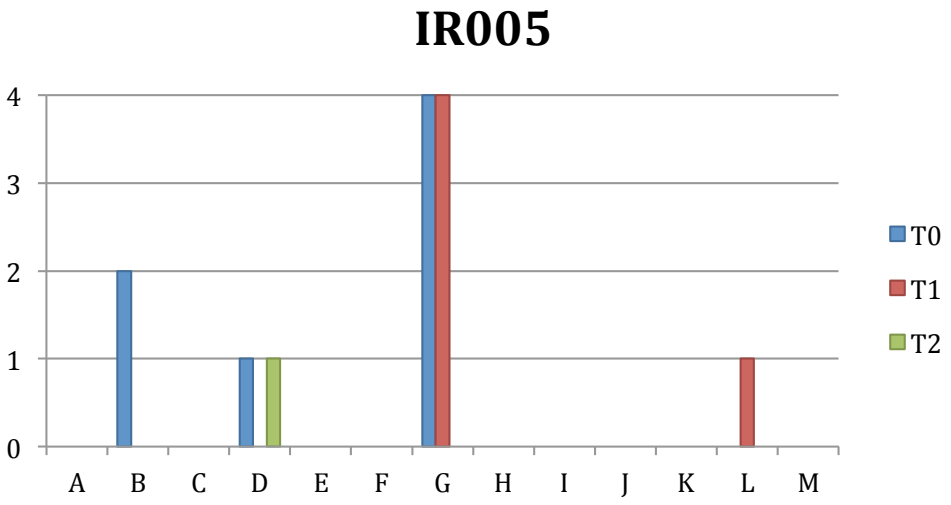
La Figure 40 reporte l'ensemble des performances des cinq patients sur le bilan TPI. Pour le TPI, les erreurs sont étiquetées de A à M, et la signification de ces lettres est détaillée dans le Tableau 13. Il s'agit généralement de confusions entre deux configurations articulatoires.

Tableau 13. Contenu des catégories d'erreurs A à M pour le bilan TPI.

A	voyelle antérieure - postérieure	**H**	changement de lieu d'articulation des occlusives
B	voyelle ouverte - fermée	**I**	occlusive - fricative
C	voyelle labiale - non labiale	**J**	occlusive - nasale
D	voyelle orale - nasale	**K**	/R/ - /w/
E	consonne initiale sourde - sonore	**L**	initiale complexe - simple
F	consonne médiane sourde-sonore	**M**	consonne finale - rien
G	changement de lieu d'articulation des fricatives	**Pour chaque catégorie, score /8**	


IR001
4
3
2
1
0
A B C D E F G H I J K L M
T0
T1
T2


IR004
4
3
2
1
0
A B C D E F G H I J K L M
T0
T1
T2


IR005
4
3
2
1
0
A B C D E F G H I J K L M
T0
T1
T2

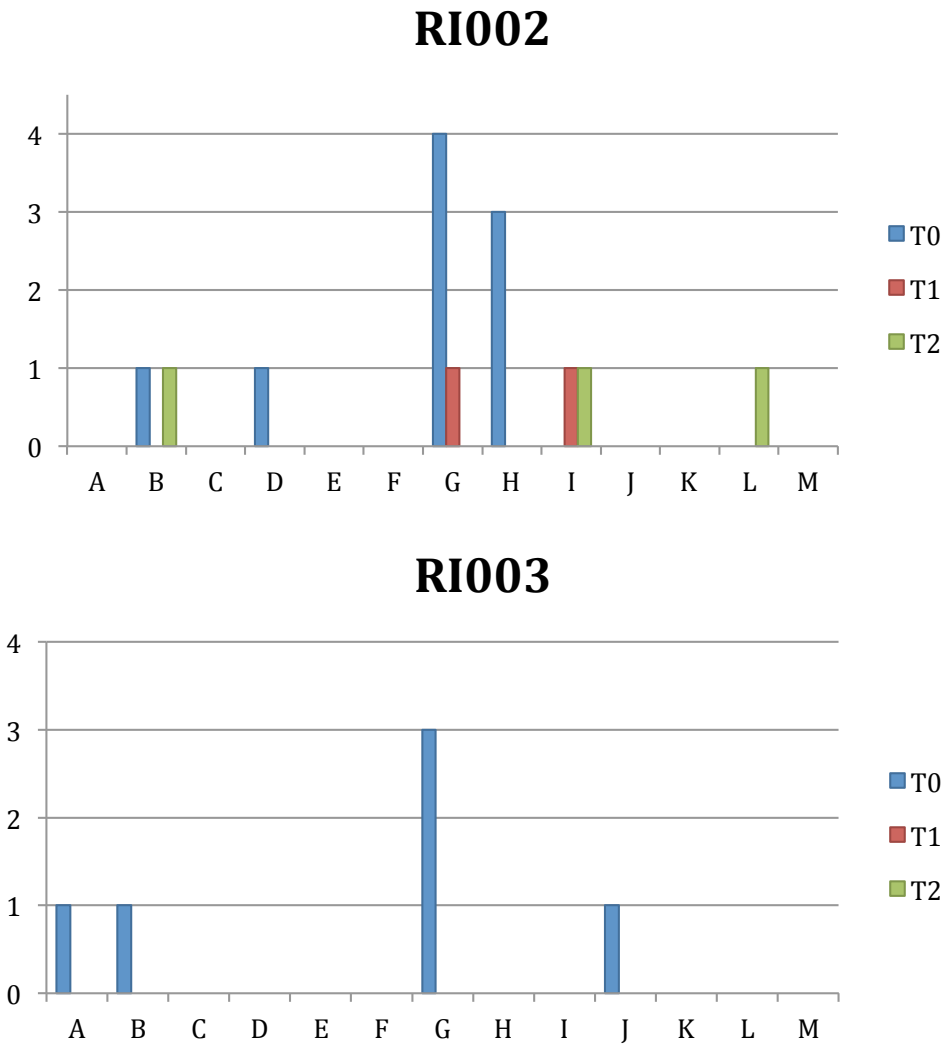


Figure 40. Résultats pour le TPI par patient et par bilan. En ordonnée, le nombre d'erreurs pour chaque catégorie.

IR001 est le patient ayant commis le plus d'erreurs à T0. La première session ILLUSTRATION ne suffit pas à corriger l'ensemble de ces erreurs. Les erreurs entre voyelle ouverte et fermée (B), voyelle orale et nasale (D), occlusive et fricative (I) et occlusive et nasale (J) sont corrigées, mais une nouvelle confusion apparaît entre voyelle antérieure et postérieure (A). Après 10 séances RETOUR, tout est corrigé sauf deux confusions, constantes tout au long de la rééducation : sur une consonne initiale sourde – sonore (E) et sur une initiale complexe – simple (L).
À T0, IR004 et IR005 présentent quatre erreurs sur le lieu d'articulation des fricatives qui ne sont corrigées qu'à T2 après la session RETOUR. Pour IR004, de nouvelles confusions apparaissent à T1, à la suite de la session ILLUSTRATION, comme entre voyelle ouverte et fermée (B), consonne initiale sourde et sonore (E). Seule une confusion est corrigée, entre initiale simple / complexe (L). À T2, il ne reste pour chacun de ces patients qu'une seule erreur.

RI002 présente beaucoup d'erreurs de changement de lieu d'articulation des fricatives et du changement de lieu d'articulation des occlusives, erreurs qui disparaissent en fin de rééducation. Deux confusions, entre occlusive et fricative (à T1) et initiale complexe et simple (à T2) apparaissent et se maintiennent. RI003, pour lequel nous pouvons observer un total de 6 erreurs, dont 3 sur le changement de lieu d'articulation des fricatives, résout tous ses problèmes sur la première session RETOUR.

Nous pouvons remarquer qu'aucun des cinq patients ne fait de confusion sur C, K ou M (exception faite d'une erreur M à T1 par IR004). L'absence d'erreur C peut sembler logique puisque cette erreur concerne la différence entre consonnes labiales et non-labiales. Cependant, si le patient présente de nombreuses faiblesses lors du bilan MBLF sur des positions extrêmes de lèvres, les mouvements de lèvres nécessaires pour la parole peuvent être altérés à la suite de la plupart des opérations. Pour IR001, l'opération subie a nécessité de couper une partie de la lèvre et de la mâchoire. Notons qu'en outre plusieurs dents ont dû être enlevées à chaque patient au cours de l'opération. L'absence d'erreur K signifie que le patient ne fait aucune erreur entre /R/ et /w/. Enfin, l'erreur M concerne l'ajout d'une consonne finale ou sa suppression.

Ces résultats sont très généraux et concernent l'ensemble des articulateurs de la cavité buccale. Or, l'échographie ne permet de visualiser que les mouvements de la langue. Il faut donc regarder, parmi les catégories du TPI, celles correspondant à un mauvais positionnement de la langue. Pour le lieu d'articulation (placement horizontal), il s'agit des catégories A, G et H. Pour l'élévation de la langue (placement vertical), il s'agit des

catégories B, I et M. Nous avons analysé plus en détail la nature des erreurs pour ces deux traits, reportées dans le Tableau 14.

Tableau 14. Erreurs réalisées au TPI par les patients à chaque bilan sur le lieu d'articulation et l'élévation de la langue. Le nombre d'erreurs est indiqué par la couleur des caractères.

4 erreurs 3 erreurs 2 erreurs 1 erreur

Bilan	T0	T1	T2
IR001	e/i t/s ʃ/s td/kg	ʃ/s y/u td/kg	
IR004	ʃ/s td/kg	e/i ʃ/s td/kg	td/kg
IR005	e/i ʃ/s	ʃ/s	
RI002	e/i ʃ/s td/kg		
RI003	e/i ʃ/s y/u	ʃ/s	e/i

Pour le lieu d'articulation, à T0, les cinq patients réalisent au moins deux confusions entre les phonèmes /ʃ/ et /s/. Ces erreurs diminuent pour disparaître totalement à T2. IR001 et RI002 présentent également à T0 trois erreurs entre les groupes /g k/ et /t d/, qui sont corrigées à T1. IR004 maintient quant à lui cette erreur, non corrigée à T2. Il en est de même pour RI003 dont l'erreur entre /y/ et /u/ est corrigée après la première session.

Pour l'élévation de la langue, quatre patients présentent des erreurs entre les voyelles /e/ et /i/ à T0. Ces erreurs disparaissent à T1, sauf pour RI002 pour qui une erreur est notée

à T2. Dans le cas de IR004, les deux erreurs qui apparaissent à T1 sont corrigées à T2. IR001 réalise aussi une confusion à T0 entre /s/ et /t/. Cette erreur peut être considérée comme une erreur d'élévation de la langue. Elle est cependant corrigée à T1.

Nous pouvons à ce stade proposer une première analyse par groupe. En effet, une tendance se dessine sur ces premiers résultats. Dans les groupes RI comme IR, les patients présentent à T0 un nombre comparable d'erreurs (entre 5 et 8 erreurs) Pour le groupe RI, quasiment toutes les erreurs sont corrigées à T1. Au contraire pour le groupe IR, il reste encore une majorité d'erreurs à T1, avec quelques nouvelles. Ces erreurs sont toutes corrigées après les séances RETOUR. Nous pouvons donc ici voir une tendance se dessiner, en faveur de l'hypothèse selon laquelle le retour visuel permettrait des progrès plus rapides pour un nombre de séances donné.

Les erreurs concernant le lieu d'articulation de la langue comme son élévation sont donc corrigées au cours des séances de rééducation. Les erreurs subsistantes ne concernent pas le mouvement de la langue et relèvent plutôt des mouvements des lèvres ou du velum. Notons cependant que IR001 présente pour tous les bilans deux erreurs lors de la production du /kʁ/, qui est réalisé comme un /ʁ/ en début de mot (catégorie L). Cette erreur est aussi présente pour RI002 à T2, mais jamais pour RI003. Nous pouvons de plus souligner que ces premiers résultats semblent montrer des progrès plus importants réalisés lors des séances RETOUR, qui restent à confirmer à plus large échelle.

4.4.3. Discussion

Le retour visuel comme l'illustration par échographie ne renseignent le patient que sur les mouvements de sa langue. Les différents bilans réalisés ont permis de mettre en avant les progrès réalisés pour cet articulateur.

Pour le Test Phonétique d'Intelligibilité de la BECD, les cinq patients présentent en fin de rééducation au plus trois erreurs sur les 52 mots prononcés. Cependant, il faut noter que dans ce test, le choix de l'orthophoniste est limité à quatre possibilités, ce qui laisse une chance sur quatre au patient d'obtenir la bonne réponse par hasard. Les erreurs concernant la position de la langue, que ce soit en termes d'élévation ou de lieu d'articulation, sont corrigées à la fin des deux sessions de rééducation. Elles concernent principalement les confusions entre /ʃ/ et /s/, /g k/ et /t d/ et /e/ et /i/. Les consonnes mentionnées sont répertoriées parmi les erreurs généralement commises par des personnes glossectomisées, dans Acher (2014) et dans Blyth *et al.* (2016). L'image échographique ne fournit de plus aucune information sur l'abaissement mandibulaire, cette information doit

donc être apportée par l'orthophoniste lorsque le patient articule. Certaines erreurs persistantes pourraient ainsi être liées à la l'attention accrue du patient accordée au mouvement de la langue, négligeant la précision du mouvement de la mâchoire.

Le test d'intelligibilité de la BECD est quant à lui moins concluant concernant les progrès des patients. Seul RI003, déjà meilleur que les autres à T0, atteint une intelligibilité presque maximale à T2.

A partir de ces premières observations, nous pouvons constater qu'après 20 séances, soit environ quatre semaines de rééducation, les performances des patients se sont sensiblement améliorées sur le plan articulatoire. Le nombre de patients, très réduit, et les faibles différences de performance entre les deux protocoles ne permettent pas de conclure sur la différence d'efficacité de ces protocoles de retour et d'illustration. Nous observons cependant une tendance sur ces premiers résultats, qui semble conforter l'hypothèse selon laquelle le retour visuel permettrait des progrès plus rapides que l'illustration. Nous pourrions supposer que l'ordre d'enchaînement des deux protocoles aurait une importance. Nous pouvons faire l'hypothèse que l'illustration permet de progresser et d'avoir une idée visuelle de l'articulation à atteindre, mais ne permet pas au patient d'associer ce qu'il voit aux mouvements de sa propre langue. Pour les patients suivant le protocole RI, les progrès seraient plus constants parce que l'illustration rappelle une connaissance antérieure sur leur propre articulation. Nous verrons si ces premières observations se confirment sur les prochains patients inclus.

Pour ces premières études de cas, les retours des patients comme des orthophonistes sont très positifs. Les orthophonistes ont bien intégré l'utilisation de ces nouveaux outils dans leur pratique. Un bémol est mis cependant sur la visualisation du phonème /ʃ/, pour lequel les patients présentent un trouble important et dont l'image échographique est en général très mauvaise. Du côté des patients, un effort est réalisé dans la compréhension de l'image, et ils ont le sentiment que cette visualisation les aide à progresser.

4.5. Perspectives

L'objectif d'une rééducation orthophonique est d'améliorer l'intelligibilité d'un patient dans la vie de tous les jours. Pour une personne ayant subi une ablation d'une partie de la langue, il ne s'agit pas exactement d'apprendre ou de corriger une articulation, mais plutôt de se réapproprier une articulation et un nouvel espace articulatoire. Pour évaluer cette intelligibilité, nous ferons appel, à la fin de l'étude clinique (prévue fin 2017), à un jury d'écoute pour comparer les productions des patients lors du TPI de la BECD, dont les

résultats, comme nous l'avons dit plus haut, manquent de précision. A chaque bilan, nous enregistrerions les productions sonores des patients. Pour chaque mot du bilan, le jury, constitué d'une vingtaine de personnes, mesurerait l'évolution de la qualité au cours du temps afin d'évaluer les progrès du patient. Nous utiliserions un test similaire à celui proposé par Pouget, Hueber *et al.* (2015) : les différentes productions à chaque bilan seraient proposées à chaque juge, qui devrait les classer suivant une échelle perceptive (de mauvais à excellent, de 1 à 5 par exemple). Il serait ainsi possible, par ce test, de classer d'une part les productions les unes par rapport aux autres, et d'autre part d'évaluer leur qualité individuellement. Nous pourrions aussi réaliser une analyse acoustique (détection formantique) des productions acoustiques de voyelles, syllabes et mots des patients enregistrées lors des bilans de la BECD (Analyse des erreurs phonétiques). Cette analyse nous permettrait de nous situer par rapport à Blyth *et al.* (2016) sur les phonèmes altérés à la suite de l'opération en validant les observations faites sur les résultats du TPI. En effet, Blyth *et al.* (2016) ont travaillé sur l'anglais australien, dont les bilans orthophoniques sont différents de ceux du français et ne nous permettent pas de comparer les performances des patients. Ce travail acoustique nous permettrait d'affiner notre analyse des progrès des patients. Notons qu'il peut y avoir un biais lié au fait que l'orthophoniste réalisant les bilans soit aussi un des orthophonistes traitant les patients. En utilisant un jury d'écoute indépendant de l'étude, nous pourrions ainsi éviter cet éventuel biais.

L'étude se poursuit et nous envisageons d'inclure au total une trentaine de patients, répartis en deux groupes équilibrés. Dans le Tableau 7, nous avons vu que les études réalisées à ce jour ne permettaient pas de démontrer avec certitude l'efficacité du retour visuel par échographie par rapport à la rééducation traditionnelle, mais que ce retour n'était en aucun cas un frein aux progrès des patients. Les patients comprennent ce qu'ils voient et savent agir en conséquence. Dans notre étude, nous nous plaçons dans un paradigme différent. Nous souhaitons évaluer le retour visuel par échographie par rapport à une illustration de l'articulation cible acquise dans la même modalité sur un locuteur de référence (à savoir l'une des orthophonistes des patients). À travers ces cinq études de cas, nous ne pouvons cependant confirmer ou infirmer notre hypothèse initiale. Nous observons cependant que l'un et l'autre des protocoles permettent aux patients de progresser, avec une bonne compréhension de ce qu'ils voient et de l'exercice demandé. Pour cette population de patients glossectomisés, d'autres paramètres doivent aussi être pris en compte. Les patients ont subi une chirurgie qui rend la rééducation difficile à cause de la douleur. Ils suivent souvent en parallèle des traitements parfois lourds comme la chimiothérapie. Lorsqu'on leur montre l'articulation cible, elle correspond à celle d'un

locuteur dont la morphologie et l'articulation n'ont pas subi cette même opération. Pour le patient, il peut donc être difficile de s'approprier l'illustration.

En poursuivant cette étude sur un groupe de patients plus large, nous espérons faire ressortir des différences plus marquées entre les sessions ILLUSTRATION et les sessions RETOUR. Si l'illustration s'avérait être le protocole de rééducation le plus performant, son développement auprès des orthophonistes serait facile, puisque ne nécessitant pas d'équipement échographique. Cependant, le retour visuel pourrait s'avérer plus adapté à la population concernée, dans la mesure où la morphologie de ces patients a été parfois considérablement modifiée.

La précision de l'échographie augmentée et les informations qu'elle apporte à la fois au patient et à l'orthophoniste en font un outil dont il reste à explorer les multiples possibilités.

Chapitre 5. Conclusion générale et perspectives

Dans cette thèse, nous nous sommes intéressée à l'utilisation de l'échographie linguale pour proposer un retour visuel en temps-réel dans le cadre de la rééducation orthophonique des troubles de l'articulation.
Un état de l'art des techniques de retour visuel lingual et d'illustration linguale nous a permis de mettre en avant les liens entre production et perception de la parole, et d'appuyer notre volonté d'exploiter une information visuelle sur la langue d'un patient pour améliorer son articulation. L'échographie fournit une image en temps-réel de la véritable articulation du patient sans en entraver le mouvement. Cependant, l'image est souvent rendue difficilement lisible par plusieurs contraintes. En plus d'un bruit de *speckle*, cette image ne fournit qu'une information sur le contour supérieur de la langue, parfois incomplet, et dans un plan 2D, sans aucune information sur les limites de la cavité orale et les autres articulateurs.
Ce travail de thèse s'est donc articulé autour de deux principaux objectifs. D'une part, nous avons développé des méthodes d'échographie linguale augmentée afin d'améliorer la lisibilité des images échographiques pour le patient et praticien. D'autre part, nous avons évalué le bénéfice du retour visuel et de l'illustration dans le cadre d'une étude clinique sur la prise en charge orthophonique des patients glossectomisés.

Suivi du contour de langue
Nous avons développé une méthode de segmentation des images échographiques visant à rendre plus visible le contour de la surface supérieure de la langue, parfois très mal imagé. Nous avons mis en place une méthode qui minimise l'intervention humaine, tout en étant aussi robuste qu'une méthode de l'état de l'art. Nous avons fait l'hypothèse qu'une partie manquante du contour pouvait être estimée non seulement à partir de la connaissance des autres parties de ce contour, mais également sur la base des autres structures présentes dans l'image. Pour cela, nous avons proposé une méthode s'appuyant sur un encodage compact d'une région d'intérêt (approche EigenTongues) et sur une modélisation des relations pixels-contours par réseau de neurones artificiels. Nous avons notamment proposé une approche multi-locuteur afin d'évaluer la capacité de généralisation de notre méthode à un locuteur inconnu.

Animation d'un modèle de langue

La deuxième méthode d'échographie linguale augmentée développée nous amène plus loin dans l'idée d'ajouter des informations. Nous animons le modèle de langue d'une tête parlante articulatoire développée au Gipsa-lab. Toujours pour faciliter son application clinique, nous cherchons pour cela une méthode permettant un bon compromis entre performance et quantité de données d'enrôlement acquises sur un nouveau locuteur. De plus, nous avons souhaité concevoir un système capable de s'adapter aux progrès d'un patient notamment en généralisant correctement à des configurations articulatoires non vues pendant la l'apprentissage. Nous proposons une méthode d'adaptation à partir d'un modèle de référence en utilisant l'algorithme *Cascaded Gaussian Mixture Regression*, dont la version *Integrated* combine dans un même modèle graphique deux régressions de type GMR. Cette approche réalise un bon compromis entre performance d'une part, avec une plus grande précision des mouvements linguaux estimés dans l'espace articulatoire de la tête parlante, quantité de données d'enrôlement et capacité de généralisation d'autre part. Les résultats obtenus démontrent l'intérêt d'exploiter des informations *a priori* sur un locuteur de référence pour pallier le manque de connaissances sur l'utilisateur.

Application clinique de l'échographie linguale

Nous avons mis en place un protocole pour la rééducation orthophonique par retour visuel chez des personnes ayant subi une ablation d'une partie de la langue (glossectomie) ou du plancher de la bouche. Ce protocole vise à comparer les progrès d'un patient lors de séances de rééducation avec une illustration visuelle et avec un retour visuel. L'illustration visuelle consistant en une visualisation des mouvements cibles, enregistrés sur un autre locuteur à l'articulation non-pathologique. Pour le retour visuel, nous avons utilisé une version simplifiée des approches décrites en chapitre 2 et 3, non finalisées pour le début de l'essai clinique. Ce protocole a pour objectif de déterminer si la visualisation de sa propre langue a un impact important pour l'acquisition d'une nouvelle articulation, ou si visualiser le geste correct suffit. Nous avons fait l'hypothèse que le retour visuel serait plus efficace que l'illustration dans notre cas. Nous avons présenté cinq études de cas sur des patients pour lesquels nous avons alterné, dans des ordres différents, dix séances de rééducation avec illustration et dix séances avec retour visuel. Pour trois des cinq patients, nous avons observé des progrès significatifs après vingt séances de rééducation, quel que soit l'ordre choisi. Les erreurs concernant la position de la langue, que ce soit en termes d'élévation ou de lieu d'articulation, ont été corrigées à la fin des deux sessions de rééducation pour tous les patients. Nous avons cependant remarqué une tendance à des progrès plus rapides avec l'utilisation du retour visuel.

Perspectives

Parmi les deux méthodes d'échographie augmentée, celle basée sur la tête parlante (Chapitre 3) nous apparaît comme la plus prometteuse. Elle permet une visualisation intuitive de l'ensemble des structures du conduit vocal. Sa mise en œuvre dans le cas d'une étude clinique reste une perspective majeure de ce travail, notamment en comparaison avec une approche basée sur l'image échographique brute. Par ailleurs, les deux méthodes d'échographie augmentée proposées dans ce travail ne modélisent pas explicitement la structure temporelle des mouvements linguaux. L'utilisation de réseaux récurrents (dont les architectures de type *Long Short-Term Memory*) permettrait cette prise en compte explicite dans le cas de la méthode de segmentation décrite au Chapitre 2.

Dans le cas de l'approche basée sur l'animation de la tête parlante (Chapitre 3), une extension de la méthode C-GMR basée sur une architecture de type HMM pourrait être envisagée. Enfin, dans les deux cas, l'extraction automatique de descripteurs robustes à partir des images pourrait s'effectuer à l'aide de réseaux à convolution, ou *CNN* (LeCun, Bengio *et al.* (2015)) qui sont aujourd'hui une méthode privilégiée pour la classification de gestes à partir de séquences d'images naturelles (Karpathy, Toderici *et al.* (2014) ; Simonyan & Zisserman (2014) ; Noda, Yamaguchi *et al.* (2014)).

Pour l'aspect applicatif en situation clinique, les premiers résultats obtenus sur cinq patients sont encourageants pour la poursuite de l'étude, pour laquelle nous prévoyons d'en inclure une trentaine. En plus des bilans orthophoniques, nous évaluerons l'intelligibilité des patients à l'aide de jurys d'écoute. De plus, à ce stade de l'étude, nous n'avons pas analysé l'ensemble des résultats des bilans et d'autres informations sur les erreurs phonétiques peuvent en être extraites. Nous pourrions aussi regarder plus en détail le travail réalisé par les orthophonistes, et déterminer s'il existe un lien entre le mode de visualisation choisi et le type de retour fourni par l'orthophoniste. Nous pourrions aussi nous intéresser à la façon dont le patient interagit avec les différents outils, afin d'évaluer de manière plus pratique le protocole le plus adapté aux patients, en fonction de leur aisance à comprendre l'image qui leur est fournie, ainsi qu'en fonction des troubles qu'ils présentent.

Comme nous l'avons indiqué dans ce manuscrit, la littérature met en évidence plusieurs manques constatés dans différentes études. Ainsi, Eriksson *et al.* (2005) souligne dans son article l'importance de proposer des outils simples, facilement modulables et utilisables en dehors des séances de rééducation. L'utilisation d'un logiciel comme Ultraspeech-player au quotidien en complément d'un retour visuel par échographie adapté au locuteur au cours des séances de rééducation avec un orthophoniste pourrait parfaitement répondre à ces exigences. Engwall (2012) utilise un retour en utilisant une tête parlante articulatoire dans

le cadre de l'inversion acoustico-articulatoire pour l'apprentissage d'une nouvelle langue. Il souligne l'importance de trouver une méthode permettant l'adaptation à tout nouveau locuteur à partir de données acquises sur ce dernier. Notre adaptation de la tête parlante à partir de données articulatoires acquises par échographie pourrait être intéressante dans ce cas de figure pour résoudre ce problème d'affichage de l'articulation. En poursuivant ces travaux, nous pourrions ainsi essayer de confirmer l'hypothèse émise par Cleland *et al.* (2013) et Roxburgh *et al.* (2015) qu'un retour visuel basé sur l'animation d'une tête parlante pourrait être bénéfique pour la rééducation orthophonique, étant plus complet et plus facile à interpréter.

Contributions

Revues

D. Fabre, T. Hueber, L. Girin, X. Alameda-Pineda, P. Badin, "Automatic Animation of an Articulatory Tongue Model from Ultrasound Images of the Vocal Tract", *Speech Communication* (soumis)

Conférences internationales

D. Fabre, T. Hueber, and P. Badin, "Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression," in *Interspeech 2014*, Singapour, 2014, pp. 2293-2297.

D. Fabre, T. Hueber, F. Bocquelet, and P. Badin, "Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks," in *Interspeech 2015*, Dresden, Germany, 2015, pp. 2410-2414.

Conférences nationales

D. Fabre, T. Hueber, M. Canault, N. Bedoin, A. Acher, C. Bach, *et al.*, "Apport de l'échographie linguale à la rééducation orthophonique," in *XVI*^{èmes} *Rencontres Internationales d'Orthophonie. Orthophonie et technologies innovantes*, Paris, France, Ortho Edition, pp. 199-225, Orthophonie et technologies innovantes.

A. Acher, **D. Fabre**, T. Hueber, P. Badin, O. Detante, E. Cousin, *et al.*, "Retour visuel en aphasiologie : résultats comportementaux, acoustiques et en neuroimagerie," in *XVI*^{èmes} *Rencontres Internationales d'Orthophonie. Orthophonie et technologies innovantes*, Paris, France, 2016, p. in press.

D. Fabre, T. Hueber, M. Canault, N. Bedoin, and P. Badin, "Retour articulatoire visuel pour la rééducation orthophonique : approches basées sur l'échographie linguale augmentée," in *6*^{èmes} *Journées de Phonétique Clinique (JPC)*, Montpellier, France, 2015.

A. Acher, **D. Fabre**, T. Hueber, S. Amen, C. Lagarde, P. Badin*, et al.*, "Retour visuel en rééducation orthophonique : étude d'un cas d'aphasie non-fluente," in *6èmes Journées de Phonétique Clinique (JPC)*, Montpellier, France, 2015.

A. Acher, **D. Fabre**, T. Hueber, O. Detante, E. Cousin, C. Pichat*, et al.*, "Bilan orthophonique en neuroimagerie pour l'évaluation de la production du langage oral chez des patients aphasiques : mise en place auprès d'une population de sujets sains," in *6èmes Journées de Phonétique Clinique (JPC)*, Montpellier, France, 2015.

Bibliographie

Acher, A. (2009). Etude perceptive et articulatoire de la parole à partir de données échographiques en 2D : comparaison de la parole normale et de la parole pathologique de patients glossectomisés.

Acher, A. (2014). *Corrélats cérébraux de l'adaptation de la parole après exérèse de la cavité orale.* Grenoble.

Acher, A., Fabre, D., Hueber, T., Badin, P., Detante, O., Cousin, E., Pichat, C. & Baciu, M. (2016). Retour visuel en aphasiologie : résultats comportementaux, acoustiques et en neuroimagerie. In *XVIèmes Rencontres Internationales d'Orthophonie. Orthophonie et technologies innovantes* (N.a.T. Joyeux, Sylvia, editor), pp. In press.

Adler-Bock, M., Bernhardt, B.M., Gick, B. & Bacsfalvi, P. (2007). The use of ultrasound in remediation of north American English /r/ in 2 adolescents. *American Journal of Speech-Language Pathology,* **16**, 128 - 139.

Atal, B.S., Chang, J.J., Mathews, M.V. & Tukey, J.W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America,* **63**(5), 1535-1555.

Auzou, P. & Rolland-Monnoury, V. (2006). *BECD: batterie d'évaluation clinique de la dysarthrie*: Ortho éditions.

Bach, C. & Lambourion, L. (2014). *L'illustration visuelle échographique en orthophonie : un entraînement pour la prise en charge du trouble phonologique fonctionnel chez l'enfant.* .Unpublished manuscript.

Bacsfalvi, P. (2007). *Visual feedback technology with a focus on ultrasound: The effects of speech habilitation for adolescents with sensorineural hearing loss.* University of British Columbia.

Bacsfalvi, P. & Bernhardt, B.M. (2011). Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography. *Clinical Linguistics & Phonetics,* **25**(11-12), 1034-1043.

Bacsfalvi, P., Bernhardt, B.M. & Gick, B. (2007). Electropalatography and ultrasound in vowel remediation for adolescents with hearing impairment. *Advances in Speech Language Pathology,* **9**(1), 36-45.

Badin, P., Elisei, F., Bailly, G. & Tarabalka, Y. (2008). An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's

articulatory data. In *Articulated Motion and Deformable Objects*, pp. 132-143. Springer.

Badin, P., Tarabalka, Y., Elisei, F. & Bailly, G. (2010). Can you 'read'tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication,* **52**(6), 493-503.

Ballard, K.J., Smith, H.D., Paramatmuni, D., McCabe, P., Theodoros, D.G. & Murdoch, B.E. (2012). Amount of kinematic feedback affects learning of speech motor skills. *Motor control,* **16**(1), 106-119.

Bälter, O., Engwall, O., Öster, A.-M. & Kjellström, H. (2005). Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, vol., pp. 36-43. ACM.

Barbier, G., Perrier, P., Ménard, L., Payan, Y., Tiede, M. & Perkell, J. (2015). Speech planning in 4-year-old children versus adults: Acoustic and articulatory analyses. In *16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, vol., pp.

Ben Youssef, A. (2011). *Control of talking heads by acoustic-to-articulatory inversion for language learning and rehabilitation (Contrôle de têtes parlantes par inversion acoustico-articulatoire pour l'apprentissage et la réhabilitation du langage).* Unpublished Thèse doctorale : EEATS, Signal, Image, Parole, Telecoms, Grenoble University, Grenoble.

Benoît, C. & Le Goff, B. (1998). Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication,* **26**(1), 117-129.

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.

Bernhardt, B.M., Gick, B., Bacsfalvi, P. & Ashdown, J. (2003). Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. *Clinical Linguistics & Phonetics,* **17**(3), 199-216.

Bezard, M. (2015). *Conscience articulatoire et illustration visuelle: effet d'un entraînement pour l'amélioration de l'intelligibilité de l'enfant déficient auditif.*

Bishop, C. (2007). Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn: Springer, New York.

Bishop, C.M. (1995). *Neural networks for pattern recognition*: Oxford university press.

Blyth, K.M., McCabe, P., Madill, C. & Ballard, K.J. (2016). Ultrasound visual feedback in articulation therapy following partial glossectomy. *Journal of Communication Disorders,* **61**, 1-15.

Byun, T.M. & Hitchcock, E.R. (2012). Investigating the use of traditional and spectral biofeedback approaches to intervention for/r/misarticulation. *American Journal of Speech-Language Pathology,* **21**(3), 207-221.

Byun, T.M., Hitchcock, E.R. & Swartz, M.T. (2014). Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research,* **57**(6), 2116-2130.

Cavin, M. (2015). The use of ultrasound biofeedback for improving English/r. *Working Papers of the Linguistics Circle,* **25**(1), 32-41.

Chen, Y.-P.P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A. & Doube, W. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech & Language,* **37**, 98-128.

Cleland, J., McCron, C. & Scobbie, J.M. (2013). Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds. *Clinical Linguistics & Phonetics,* **27**(4), 299-311.

Cleland, J., Scobbie, J.M. & Wrench, A.A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics*(0), 1-23.

Combescure, P. (1981). 20 listes de dix phrases phonétiquement équilibrées. *Revue d'acoustique,* **56**, 34-38.

Cootes, T.F., Edwards, G.J. & Taylor, C.J. (1998). Active appearance models. In *Computer Vision—ECCV'98*, pp. 484-498. Springer.

Cornett, R.O. (1967). *Cued speech*: publisher not identified.

Cowie, R., Douglas-Cowie, E. & Kerr, A. (1982). A study of speech deterioration in post-lingually deafened adults. *The Journal of Laryngology & Otology,* **96**(02), 101-112.

Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning,* **25**(1), 37-64.

Engwall, O. & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Computer Assisted Language Learning,* **20**(3), 235-262.

Epstein, M.A. (2005). Ultrasound and the IRB. *Clinical Linguistics & Phonetics,* **19**(6-7), 567-572.

Erber, N.P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders,* **40**(4), 481-492.

Eriksson, E., Bälter, O., Engwall, O., Öster, A.-M. & Kjellström, H. (2005). Design recommendations for a computer-based speech training system based on end-user interviews. In, vol., pp.

Fabre, D., Hueber, T., Canault, M., Bedoin, N., Acher, A., Bach, C., Lambourion, L. & Badin, P. (2016). Apport de l'échographie linguale à la rééducation orthophonique. In *XVIèmes Rencontres Internationales d'Orthophonie. Orthophonie et technologies innovantes* (N.a.T. Joyeux, Sylvia, editor), pp. In press.

Fagel, S. & Madany, K. (2008). A 3-D virtual head as a tool for speech therapy for children. In *9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, vol., pp. 2643-2646. Brisbane, Australia.

Fasel, I. & Berry, J. (2010). Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, vol., pp. 1493-1496. IEEE.

Feldenkrais, M. (1993). Énergie et bien-être par le mouvement. *Dangles, Saint Jean de Bray.*

Gallagher, L. (2013). The effectiveness of ultrasound technology as a visual biofeedback tool on the productive speech intelligibility of adolescents and young adults with a hearing impairment.

Gatignol, P., Troadec, J., Martel, C. & Robert-Jahier, A. (2013). *MBLF 4/8 ans (Motricité Bucco-Linguo-Faciale - Articulation - Déglutition).* Châteauroux: ADEPRIO.

Gauvain, J.-L. & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE transactions on speech and audio processing,* **2**(2), 291-298.

Ghahramani, Z. & Jordan, M.I. (1994). Supervised learning from incomplete data via an EM approach. *Advances in neural information processing systems*, 120-120.

Gibbon, F. (2011). Bibliography of electropalatographic (epg) studies in english (1957-2013): Staženo.

Gibbon, F., Hardcastle, W.J., Crampin, L., Reynolds, B., Razzell, R. & Wilson, J. (2013). Visual feedback therapy using electropalatography (EPG) for articulation disorders associated with cleft palate. *Asia Pacific Journal of Speech, Language and Hearing.*

Gibbon, F. & Lee, A. (2015). Electropalatography for Older Children and Adults with Residual Speech Errors. In *Seminars in speech and language*, vol. 36, pp. 271-282. Thieme Medical Publishers.

Hudgins, C.V. & Numbers, F.C. (1942). *An investigation of the intelligibility of the speech of the deaf*: The Journal Press.

Hueber, T. (2009). *Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal: vers une communication parlée silencieuse.* Université Pierre et Marie Curie-Paris VI.

Hueber, T. (2013). Ultraspeech-player: intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training. In *INTERSPEECH*, vol., pp. 752-753.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P. & Stone, M. (2007). Eigentongue feature extraction for an ultrasound-based silent speech interface. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, pp. I-1245-I-1248. IEEE.

Hueber, T., Chollet, G., Denby, B. & Stone, M. (2008). Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. of ISSP*, 365-369.

Hueber, T., Girin, L., Alameda-Pineda, X. & Bailly, G. (2015). Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on,* **23**(12), 2246-2259.

Hueber, T., Youssef, A.B., Bailly, G., Badin, P. & Elisei, F. (2012). Cross-speaker Acoustic-to-Articulatory Inversion using Phone-based Trajectory HMM for Pronunciation Training. In *13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, vol., pp. Citeseer.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, vol., pp. 1725-1732.

Katz, W.F. & Mehta, S. (2015). Visual Feedback of Tongue Movement for Novel Speech Sound Learning. *Frontiers in human neuroscience,* **9**.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature,* **521**(7553), 436-444.

Leggetter, C.J. & Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language,* **9**(2), 171-185.

Li, M., Kambhamettu, C. & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics,* **19**(6-7), 545-554.

Loosvelt, M., Villard, P.-F. & Berger, M.-O. (2014). Using a biomechanical model for tongue tracking in ultrasound images. In *Biomedical Simulation*, pp. 67-75. Springer.

Maas, E., Robin, D.A., Hula, S.N.A., Freedman, S.E., Wulf, G., Ballard, K.J. & Schmidt, R.A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology,* **17**(3), 277-298.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics,* **11**(2), 431-441.

Massaro, D.W. & Light, J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of speech, Language, and hearing research,* **47**(2), 304-320.

Meltzoff, A.N. & Moore, M.K. (1997). Explaining facial imitation: A theoretical model. *Early development & parenting,* **6**(3-4), 179.

Menin-Sicard, A. & Sicard, E. (2012). Intérêt de la visualisation de la position et du mouvement des articulateurs dans la prise en charge des troubles phonologiques.

Mills, A. (1987). The development of phonology in the blind child. In. B. Dodd & R. Campbell (Ed.). Hearing by eye: The psychology of lip reading (pp. 145-161): London, Lawrence Erlbaum Associates Publishers.

Modha, G., Bernhardt, B.M., Church, R. & Bacsfalvi, P. (2008). Case study using ultrasound to treat. *International Journal of Language & Communication Disorders,* **43**(3), 323-329.

Montgomery, D. (1981). Do dyslexics have difficulty accessing articulatory information? *Psychological Research,* **43**(2), 235-243.

Muir, L.J. & Richardson, I.E. (2005). Perception of sign language and its application to visual communications for deaf people. *Journal of Deaf Studies and Deaf Education,* **10**(4), 390-401.

Mukamel, R., Ekstrom, A.D., Kaplan, J., Iacoboni, M. & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current biology,* **20**(8), 750-756

Myers, C.S. & Rabiner, L.R. (1981). A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal,* **60**(7), 1389-1409.

Newell, K., Carlton, M. & Antoniou, A. (1990). The interaction of criterion and feedback information in learning a drawing task. *Journal of Motor Behavior,* **22**(4), 536-552.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G. & Ogata, T. (2014). Lipreading using convolutional neural network. In *15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, vol., pp. 1149-1153.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning,* **27**(5), 439-453.

Ouni, S. & Laprie, Y. (2005). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America,* **118**(1), 444-460.

Peckels, J.P. & Rossi, M. (1973). Filetest de diagnostic par paires minimales. Adaptation au francais du'Diagnostic Rhyme Test'de WD Voiers. *Revue d'Acoustique,* **27**, 245-262.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J. & Guiod, P. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech communication,* **22**(2), 227-250.

Perkell, J.S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics,* **25**(5), 382-407.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Tricarico, R. & Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics,* **28**(3), 233-272.

Perrier, P., Savariaux, C., Lebeau, J. & Magaña, G. (1999). Speech production after tongue surgery and tongue reconstruction. In *ICPhS*, vol. 99, pp. 1805-1808.

Pillot-Loiseau, C., Kamiyama, T. & Antolík, T.K. (2015). French/y/-/u/contrast in Japanese learners with/without ultrasound feedback: vowels, non-words and words. In, vol. 1, pp. 1-5. International Phonetic Association: London.

Pineda, F.J. (1987). Generalization of back-propagation to recurrent neural networks. *Physical review letters,* **59**(19), 2229.

Porta, J.M., Verbeek, J.J. & Kröse, B.J.A. (2005). Active appearance-based robot localization using stereo vision. *Autonomous Robots,* **18**(1), 59-80.

Pouget, M., Hueber, T., Bailly, G. & Baumann, T. (2015). Hmm training strategy for incremental speech synthesis. In *16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, vol., pp. 1201-1205.

Preston, J.L., Brick, N. & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology,* **22**(4), 627-643.

Preston, J.L. & Leaman, M. (2014). Ultrasound visual feedback for acquired apraxia of speech: A case report. *Aphasiology,* **28**(3), 278-295.

Preston, J.L., Maas, E., Whittle, J., Leece, M.C. & McCabe, P. (2016). Limited acquisition and generalisation of rhotics with ultrasound visual feedback in childhood apraxia. *Clinical Linguistics & Phonetics*, 1-17.

Richmond, K. (2002). *Estimating articulatory parameters from the acoustic speech signal.* University of Edinburgh.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research,* **3**(2), 131-141.

Roussos, A., Katsamanis, A. & Maragos, P. (2009). Tongue tracking in ultrasound images with active appearance models. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, vol., pp. 1733-1736. IEEE.

Roxburgh, Z., Scobbie, J.M. & Cleland, J. (2015). Articulation therapy for children with cleft palate using visual articulatory models and ultrasound biofeedback. *Proceedings of the 18th ICPhS, Glasgow*.

Shawker, T.H. & Sonies, B.C. (1985). Ultrasound Biofeedback for Speech Training: Instrumentation and Preliminary Results. *Investigative Radiology,* **20**(1), 90-93.

Shigemori, L.S.B., Pouplier, M. & Benuš, Š. (2015). Phonemic length and phrasal accent in slovak consonantal and vocalic nuclei.

Shuster, L.I., Ruscello, D.M. & Smith, K.D. (1992). Evoking [r] using visual feedback. *American Journal of Speech-Language Pathology,* **1**(3), 29-34.

Shuster, L.I., Ruscello, D.M. & Toth, A.R. (1995). The use of visual feedback to elicit correct/r. *American Journal of Speech-Language Pathology,* **4**(2), 37-44.

Simonyan, K. & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, vol., pp. 568-576.

Stone, M. & Davis, E.P. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *The Journal of The Acoustical Society of America,* **98**(6), 3107-3112.

Stone, M. & Shawker, T.H. (1986). An ultrasound examination of tongue movement during swallowing. *Dysphagia,* **1**(2), 78-83.

Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america,* **26**(2), 212-215.

Tang, L., Bressmann, T. & Hamarneh, G. (2012). Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves. *Medical image analysis,* **16**(8), 1503-1520.

Toda, T., Black, A.W. & Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication,* **50**(3), 215-227.

Turgeon, C., Prémont, A., Trudeau-Fisette, P. & Ménard, L. (2015). Exploring consequences of short-and long-term deafness on speech production: A lip-tube perturbation study. *Clinical Linguistics & Phonetics,* **29**(5), 378-400.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience,* **3**(1), 71-86.

Woisard-Bassols, V. & Puech, M. (2011). *La réhabilitation de la déglutition chez l'adulte: le point sur la prise en charge fonctionnelle*: Groupe de Boeck.

Wu, Y., Gendrot, C., Hallé, P. & Adda-Decker, M. (2015). On Improving the Pronunciation of French r in Chinese Learners by Using Real-time Ultrasound Visualization. In, vol., pp. 5.

Xu, K., Yang, Y., Stone, M., Jaumard-Hakoun, A., Leboullenger, C., Dreyfus, G., Roussel, P. & Denby, B. (2016). Robust contour tracking in ultrasound tongue image sequences. *Clinical Linguistics & Phonetics,* **30**(3-5), 313-327.

Yeung, F., Levinson, S.F., Fu, D. & Parker, K.J. (1998). Feature-adaptive motion tracking of ultrasound image sequences using a deformable mesh. *IEEE Transactions on Medical Imaging,* **17**(6), 945-956.

Young, S., Woodland, P., Evermann, G. & Gales, M. (2013). The HTK toolkit 3.4. 1: Cambridge, UK: Cambridge Univ. Eng Dept.

Zen, H., Nankaku, Y. & Tokuda, K. (2011). Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Transactions on Audio, Speech, and Language Processing,* **19**(2), 417-430.

Annexe A : corpus Ultraspeech-player

/t/	tapis - tirer - tulipe - tasse - tacher - tissu - tousse - toujours état - raté - petit - têtu - auto - matou - satin - patte - tomate - tête - imite - huit - suite tais-toi ! - Où étais-tu ? - Tu as tout ton temps. - À ta santé !
/tʃ/ /tl/	caoutchouc - scotch - sketch - Atlantique
/k/	car - caillou - quel - copie - casser - cage moka - meccano - paquet - piquant - jockey - choquer - acquisition lac - plaque - bec - unique - nuque - époque - chaque - Jacques
/k/+/k/	cacao - kaki - coquille - caquette - quiconque - clac
/k/+/t/	carotte - quêter - quitter - culotte - culbute - comptine - cantine - écouter - atomique
/kt/	dactylo - tracteur - tactique - directeur - tic-tac - détective Compte les tickets. - On entend le tic-tac. - Colette est coquette.
/kl/	clapoter - cliquetis - club - clôture - climatiser éclat - éclater - incliné - conclure - musclé - encyclique racle - oracle - miracle - article - boucle Éclaire-moi. - A-t-il la clé du club ? - Le climat est clément.
/p/+/t/+/k/	Ôte ton képi - Tu iras camper - Écoute tes parents
/b/+/d/+/g/	Hugues a bien dormi - Donne un bonbon à Guy - Odile désire une bague
/b/+/d/+/g/+ /p/+/t/+/k/	Le catalogue de Paul est tombé. - Tu cueilleras beaucoup de grappes. - Pose la guitare dans la cabine.
/st/	stalle - steak - stock - style - statut - stylo - sténotypie hostile - accoster - astiquer - dépister peste - piste - toast - caste C'est stupide ! - C'est astucieux. - Arrête-toi au stop. - Son stylo est dans sa trousse.
/sk/	scalp - scotch - script - scalaire oscar - bosquet - visqueux - escale - escarpolette casque - kiosque - basque - disque - risque Qu'est-ce que c'est ? - Il risque de se tromper. - Les esquimaux font du ski.

Annexe B : bilans orthophoniques

Afin de ne pas divulguer le contenu des bilans, nous avons simplement indiqué les scores obtenus aux différents sous-bilans sans préciser la nature des différents points évalués pour chaque sous-bilan.

IR001

	T0	T1	T2	
Total MBLF + BECD	**259**	**285**	**300**	**/338**
MBLF	**62**	**82**	**91**	**/111**
Face	5	5	5	/6
Œil	9	9	9	/9
Lèvres	17	20	24	/27
Joues et mandibule	15	22	25	/30
Langue	16	26	28	/39
BECD - synthèse globale	**197**	**203**	**209**	**/227**
BECD - Intelligibilité	**12**	**13**	**14**	**/24**
Mots	3	4	4	/8
Phrases	4	4	4	/8
Conversation	5	5	6	/8
BECD - TPI	**41**	**45**	**49**	**/52**
BECD - Analyse des erreurs phonétiques	**144**	**145**	**146**	**/151**
BECD - Erreurs phonétiques isolées	32	33	31	/33
BECD - Mots simples	86	84	86	/88
BECD - Mots complexes	26	28	29	/30

RI002

	T0	T1	T2	
Total MBLF + BECD	**274**	**293**	**301**	**/338**
MBLF	**74**	**83**	**92**	**/111**
Face	6	6	6	/6
Œil	9	9	9	/9
Lèvres	20	20	21	/27
Joues et mandibule	20	23	27	/30
Langue	19	25	29	/39
BECD - synthèse globale	**200**	**210**	**209**	**/227**
BECD - Intelligibilité	**13**	**14**	**16**	**/24**
Mots	3	4	6	/8
Phrases	4	4	4	/8
Conversation	6	6	6	/8
BECD - TPI	**43**	**50**	**49**	**/52**
BECD - Analyse des erreurs phonétiques	**144**	**146**	**144**	**/151**
BECD - Erreurs phonétiques isolées	28	30	30	/33
BECD - Mots simples	87	86	86	/88
BECD - Mots complexes	29	30	28	/30

RI003

	T0	T1	T2	
Total MBLF + BECD	**253**	**285**	**308**	**/338**
MBLF	**60**	**75**	**88**	**/111**
Face	6	6	6	/6
Œil	7	8	9	/9
Lèvres	18	20	24	/27
Joues et mandibule	10	16	17	/30
Langue	19	25	32	/39
BECD - synthèse globale	**193**	**210**	**220**	**/227**
BECD - Intelligibilité	**14**	**16**	**21**	**/24**
Mots	4	4	7	/8
Phrases	4	6	7	/8
Conversation	6	6	7	/8
BECD - TPI	**46**	**52**	**52**	**/52**
BECD - Analyse des erreurs phonétiques	**133**	**142**	**147**	**/151**
BECD - Erreurs phonétiques isolées	28	30	32	/33
BECD - Mots simples	81	87	87	/88
BECD - Mots complexes	24	25	28	/30

IR004

	T0	T1	T2	
Total MBLF + BECD	**269**	**280**	**296**	**/338**
MBLF	**77**	**84**	**88**	**/111**
Face	6	6	6	/6
Œil	9	9	9	/9
Lèvres	26	26	26	/27
Joues et mandibule	23	28	28	/30
Langue	13	15	19	/39
BECD - synthèse globale	**192**	**196**	**208**	**/227**
BECD - Intelligibilité	**14**	**18**	**19**	**/24**
Mots	4	6	6	/8
Phrases	4	6	7	/8
Conversation	6	6	6	/8
BECD - TPI	**46**	**42**	**51**	**/52**
BECD - Analyse des erreurs phonétiques	**132**	**136**	**138**	**/151**
BECD - Erreurs phonétiques isolées	28	31	30	/33
BECD - Mots simples	82	82	84	/88
BECD - Mots complexes	22	23	24	/30

IR005

	T0	T1	T2	
Total MBLF + BECD	**273**	**296**	**305**	**/338**
MBLF	**76**	**91**	**93**	**/111**
Face	6	6	6	/6
Œil	8	8	9	/9
Lèvres	23	26	26	/27
Joues et mandibule	26	30	30	/30
Langue	13	21	22	/39
BECD - synthèse globale	**197**	**205**	**212**	**/227**
BECD - Intelligibilité	**13**	**16**	**19**	**/24**
Mots	4	4	6	/8
Phrases	4	6	7	/8
Conversation	5	6	6	/8
BECD - TPI	**45**	**48**	**51**	**/52**
BECD - Analyse des erreurs phonétiques	**139**	**141**	**142**	**/151**
BECD - Erreurs phonétiques isolées	32	31	31	/33
BECD - Mots simples	82	85	85	/88
BECD - Mots complexes	25	25	26	/30