UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

HABILITATION A DIRIGER DES RECHERCHES DE L'UNIVERSITE DE GRENOBLE

Spécialité : **Signal, Image, Parole, Telecom**

Arrêté ministériel : 7 août 2006

Présentée par
Thomas HUEBER

Thèse préparée au sein du
laboratoire Grenoble Images Parole Signal Automatique (GIPSA-lab, UMR 5216)

Traitement automatique de la parole multimodale : application à la suppléance vocale et à la rééducation orthophonique

Soutenue publiquement le 3 juillet 2019,
devant le jury composé de :

Dr. Frédéric BIMBOT
IRISA, Rennes, Rapporteur
Pr. Olov ENGWALL
KTH Royal Institute of Technology, Stockholm, Rapporteur
Dr. Frédéric BEVILACQUA
IRCAM Paris, Rapporteur
Dr. Emmanuel DUPOUX
LSCP, Paris, Examinateur
Dr. Cécile FOUGERON
LPP, Paris, Examinateur
Pr. Eric GAUSSIER
LIG, Grenoble, Examinateur, Président du Jury

UNIVERSITÉ DE GRENOBLE ALPES

ÉCOLE DOCTORALE EEATS

Electronique, Electrotechnique, Automatique, Traitement du signal

T H È S E

pour obtenir

Habilitation à diriger des recherches

de l'Université de Grenoble Alpes

Mention : Signal, Image, Parole, Telecom

Présentée et soutenue par

Thomas HUEBER

Traitement automatique de la parole multimodale : application à la suppléance vocale et à la rééducation orthophonique

Thèse préparée au laboratoire Grenoble Images Parole Signal Automatique (GIPSA-lab, UMR 5216)

soutenue le 3 juillet 2019

Jury :

Rapporteurs :	Dr. Frédéric BIMBOT	-	IRISA, Rennes
	Pr. Olov ENGWALL	-	KTH Royal Institute of Technology, Stockholm
	Dr. Frédéric BEVILACQUA	-	IRCAM, Paris
Examinateurs :	Dr. Emmanuel DUPOUX	-	LSCP, Paris
	Dr. Cécile FOUGERON	-	LPP, Paris
Président :	Pr. Eric GAUSSIER	-	LIG, Grenoble

Table des matières

Table des sigles et acronymes

GIPSA-lab	Laboratoire Grenoble, Image, Parole, Signal Automatique
SSI	*Silent Speech Interface*
BCI	*Brain-Computer Interface*
GMM	*Gaussian Mixture Model*
GMR	*Gaussian Mixture Regression*
C-GMR	*Cascaed Gaussian Mixture Regression*
HMM	*Hidden Markov Model*
MLLR	*Maximum Likelihood Linear Regression*
MAP	*Maximum a Posteriori*
EM	*Expectation-Maximization*
ACP	*Analyse en Composantes Principales*
PCA	*Principal Component Analysis*
EMA	*Electromagnetic articulography*
IRM	Imagerie à Résonance Magnétique
ANN	*Artificial Neural Network*
DNN	*Deep Neural Network*
SVM	*Support Vector Machine*
ECoG	*Electrocorticography*
GAN	*Generative Adversarial Network*
C-GAN	*Conditional Generative Adversarial Network*
RNN	*Recurrent Neural Network*
LSTM	*Long Short-Term Memory*
AE	*Autoencoder*
VAE	*Variational Autoencoder*
POS	*Part-Of-Speech*

TTS *Text-To-Speech*

VADS *Voies Aéro-Digestive Supérieures*

Introduction

Mes activités de recherche portent sur le traitement automatique de la parole multimodale, avec un intérêt particulier pour la capture, l'analyse et la modélisation de signaux décrivant les différentes activités physiologiques impliquées lors de sa production. Par activités physiologiques, on distingue ici l'activité motrice, c'est-à-dire l'ensemble des gestes respiratoires, laryngés et articulatoires [1] qui façonnent le signal acoustique, et l'activité électrique du système nerveux (central et périphérique) qui coordonne ces gestes [2]. Mes travaux visent à développer des technologies vocales qui exploitent ces différentes activités physiologiques, pour la reconnaissance automatique et la synthèse de la parole, à destination notamment des personnes présentant un trouble de la communication parlée. Plus spécifiquement, ces technologies ont pour objectif :

- de rétablir la capacité à communiquer oralement lorsqu'une partie de la chaîne de production de la parole est défaillante, en raison d'une pathologie touchant soit le système nerveux, soit l'appareil phonatoire. On parlera dans la suite du manuscrit de *suppléance vocale*.
- d'améliorer la rééducation orthophonique d'un trouble phonétique [3], ou d'un trouble phonologique [4], apparaissant par exemple chez l'enfant au cours du développement du langage, ou chez l'adulte après une lésion du système nerveux central ou une chirurgie carcinologique de la cavité orale. On parlera dans la suite de *rééducation articulatoire assistée*.

La méthodologie générale sur laquelle je m'appuie est basée sur :

- La mise en place de dispositifs et de protocoles expérimentaux pour l'enregistrement des différentes activités physiologiques impliquées dans la production de la parole. Ces dispositifs sont basés sur différents capteurs, comme par exemple l'échographie ou l'articulographie électromagnétique (EMA), qui permettent l'acquisition des mouvements des articulateurs non-visibles comme la langue.
- La modélisation par apprentissage automatique (*machine learning*) des relations entre ces différentes activités physiologiques, principalement dans un but prédictif. On cherchera par exemple à prédire le contenu spectral d'un signal de parole à partir des mouvements articulatoires, ou bien à décoder l'activité cérébrale au niveau lexical.

[1] L'activité motrice est caractérisée par le mouvement des différents articulateurs de la parole que sont principalement les lèvres, la mâchoire, la langue, et le voile du palais

[2] Les gestes co-verbaux qui accompagnent la communication parlée comme les expressions faciales, les mouvements de tête, les gestes de pointage, ne sont pas étudiés dans le cadre de mes travaux.

[3] Un trouble phonétique ou articulatoire est caractérisé par une erreur systématique dans la réalisation de certains phonèmes.

[4] Un trouble phonologique est caractérisé par l'incapacité à utiliser et à agencer correctement les phonèmes [Bri+11].

- L'utilisation de ces modèles prédictifs dans des systèmes "temps-réel" qui viennent s'insérer dans la boucle de contrôle sensori-moteur de la parole, ce qui permet donc de questionner certains mécanismes cognitifs qui sous-tendent ce contrôle. A titre d'exemple, la "parole silencieuse", c'est-à-dire une parole produite en articulant sans vocaliser, est un paradigme intéressant pour quantifier l'exploitation du retour somatosensoriel (auditif, proprioceptif, kinesthésique) pour le contrôle moteur de la parole. De même, un système de rééducation articulatoire par "retour visuel" qui permet au patient de visualiser ses propres mouvements linguaux, renseigne sur la manière dont notre cerveau intègre différentes modalités de la parole.

Mes travaux de recherches sont de fait trans-disciplinaires, et se positionnent à l'intersection du traitement du signal, du traitement automatique de la parole, de la vision par ordinateur, de l'apprentissage statistique, mais également des sciences du langage, de la phonétique clinique et des sciences cognitives. Ce manuscrit présente mes travaux réalisés depuis 2009 (obtention de mon doctorat). La recherche étant un sport individuel qui se pratique en équipe, ces travaux ont fait l'objet de multiples collaborations, tout d'abord avec des chercheurs, enseignants-chercheurs et ingénieurs de recherche du GIPSA-lab [5], et d'autres laboratoires dont INRIA (Montbonnot, X. Alameda-Pineda, équipe Perception), le Laboratoire de Psychologie et NeuroCognition (LPNC, Grenoble, H. Loevenbruck, M. Baciu), le laboratoire BrainTech (INSERM, Grenoble, Blaise Yvert), le laboratoire *Dynamique du langage* (DDL, Lyon, Mélanie Canault et N. Bedoin), l'Institut Langevin (Paris, B. Denby), le *Cognitive System Laboratory* (Brême, Allemagne, T. Schultz), et plusieurs acteurs de santé dont le CHU de Grenoble, le CHU de Lyon, le centre médicale Rocheplane à Saint-Martin d'Hères, et enfin un réseau de plusieurs orthophonistes libérales. Enfin, ces travaux ont en partie été réalisés par des chercheurs post-doctorants et doctorants, des étudiants de Master, des étudiants d'écoles d'ingénieurs et d'écoles d'orthophonie, que j'ai eu le plaisir et la responsabilité de co-encadrer (voir section 4.5).

Ce manuscrit est organisé comme suit. Le premier chapitre est consacré à mes travaux sur la suppléance vocale. Le second chapitre porte sur la rééducation articulatoire assistée. Des perspectives directes, dans la continuité de ces deux axes, seront présentées à la fin de chaque sous-section. Le troisième chapitre présentera un nouveau projet de recherche portant sur l'apprentissage faiblement supervisé d'espaces de représentation de la parole (acoustique, moteur, linguistique) à l'aide de réseaux neuronaux génératifs, pour la modélisation computationnelle de la production de la parole et la restauration d'un signal (audio) de parole pathologique. Un *curriculum vitæ* détaillé, incluant la liste des mes encadrements et productions scientifiques conclue ce document.

[5] L. Girin, P. Badin, G. Bailly, J-L Schwartz, P. Perrier, D. Beautemps, G. Feng, F. Eliséi, C. Savariaux, C. Vilain, M. Garnier.

Chapitre 1

Systèmes de suppléance vocale

Sommaire

Cet axe de recherche porte sur la conception de systèmes visant à rétablir la capacité à communiquer oralement lorsqu'une partie de la chaine de production de la parole est défaillante. Cette chaîne, illustrée à la figure 1.1, débute par le système nerveux central et périphérique, se poursuit par l'action coordonnée d'un ensemble de muscles pilotant la respiration, la vibration des plis vocaux, et l'articulation (langue, lèvres, voile du palais, mâchoire), et se termine par la production acoustique. Certaines pathologies peuvent interrompre le bon fonctionnement de ce processus moteur complexe. Une insuffisance respiratoire chronique ou aiguë peut être à l'origine d'une hypophonie plus ou moins importante. Une ablation totale du larynx dans le cadre d'un traitement du cancer peut laisser le patient dans l'incapacité de vocaliser le son car les voix aériennes se voient alors "déconnectées" du conduit vocal. Une maladie neurodégénérative comme la sclérose en plaque amyotrophique ou la maladie de Parkinson peut provoquer un dysfonctionnement des muscles nécessaires à la mise en mouvement des plis vocaux (on parle alors de dysphonie) et des articulateurs (on parle alors de dysarthrie) pouvant conduire à une perte de la communication orale.

Les systèmes de suppléance vocale sur lesquels je travaille cherchent à exploiter la partie encore fonctionnelle de la chaîne de production de la parole et convertir une des activités physiologiques impliquées (cérébrale, musculaire ou articulatoire) soit en une séquence de mots - on parle alors de "reconnaissance" - soit en un signal acoustique intelligible - on parle alors de "conversion". En fonction de l'activité physiologique considérée, on distingue deux principales catégories de systèmes :

- les interfaces de communication en parole silencieuse ou *silent speech interface* (SSI) qui visent à capturer, de façon la moins invasive possible, l'activité musculaire ou les mouvements articulatoires, puis à les convertir automatiquement, soit en texte, soit en un signal de parole synthétique.
- les interfaces cerveau-ordinateur (ou *brain-computer interface*, BCI) qui s'appuient sur la capture de l'activité électrique des aires cérébrales impliquées dans la production de la parole et qui cherchent soit à décoder cette activité au niveau phonétique ou lexical, soit à reconstruire un signal de parole intelligible, si possible en temps-réel (contrôle en boucle fermée).

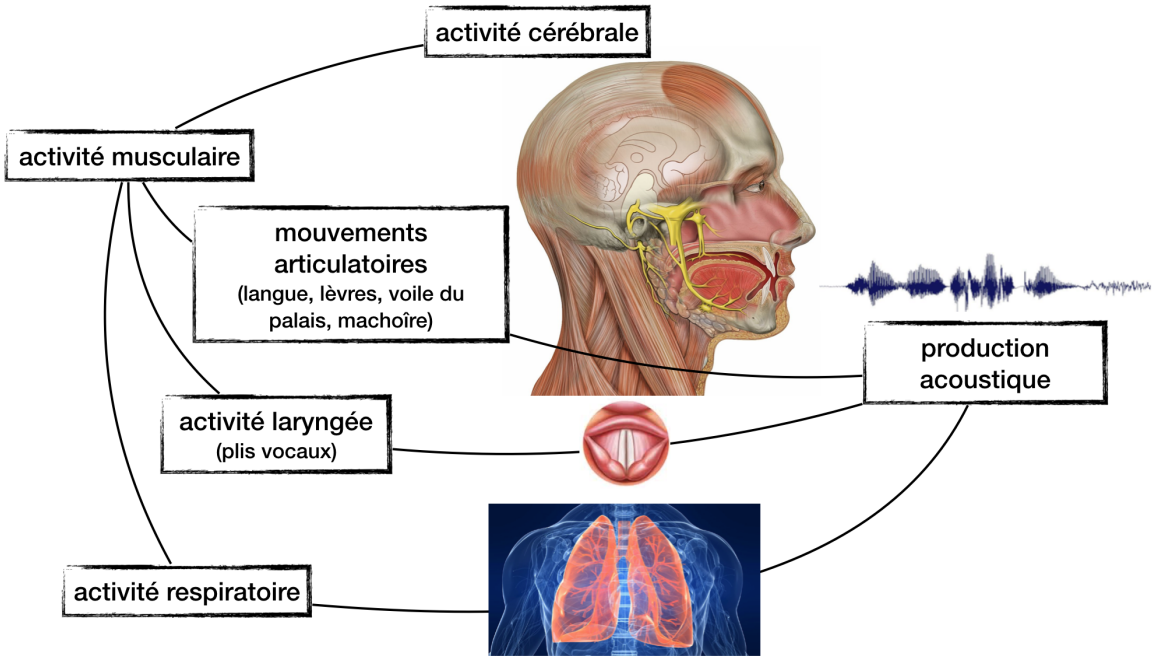


FIGURE 1.1 – Activités physiologiques impliquées dans la production de la parole

Ce chapitre est organisé comme suit. Mes travaux sur les SSI sont décrits à la section 1.1. Mon activité sur le développement d'un BCI pour la restauration de la parole est présentée à la section 1.2. A la section 1.3, je présente un autre axe de recherche qui porte sur l'amélioration de l'interactivité des systèmes *Text-to-speech* (TTS) lorsque ces derniers sont utilisés comme voix de substitution, au travers du paradigme de "synthèse vocale incrémentale". Enfin, la section 1.4 présentera une contribution récente sur la reconnaissance automatique de la

langue Française parlée complétée (LPC ou *cued-speech* en anglais), qui est une méthode de communication augmentée utilisée par certaines personnes sourdes ou malentendantes.

1.1 Interface de communication en parole silencieuse

1.1.1 Contexte et état de l'art

Depuis mon doctorat [Hue09 ; Hue+10], je mène des travaux sur la conception de SSI [HBD12 ; HB16 ; TH17]. Je travaille notamment sur une approche basée sur la capture des mouvements articulatoires par échographie et vidéo, et leur conversion automatique, soit en texte (reconnaissance), soit en un signal de parole synthétique (conversion). A terme, le dispositif envisagé pourrait être une alternative intéressante aux différentes voix de substitution mises en place après une laryngectomie totale, comme par exemple les voix œsophagienne et tracheo-œsophagienne, ainsi qu'à l'électrolarynx. Il pourrait également être utile à des personnes présentant une aphonie (baisse de l'intensité de la voix sans dysarthrie sévère associée), liée par exemple à une insuffisance respiratoire ou à une paralysie des cordes vocales. Il permettrait également à toute personne de communiquer silencieusement, c'est-à-dire en articulant normalement mais sans vocaliser, puisque le système n'exploite pas l'activité acoustique. Une autre application possible est la communication dans des environnements très bruyants, les mouvements articulatoires étant moins sensible au bruit que le signal acoustique.

Le concept de SSI apparaît au début des années 2000. Cependant, leur conception s'appuie sur des recherches plus anciennes, portant sur l'exploitation de la modalité articulatoire dans les technologies vocales, pour la reconnaissance comme pour la synthèse. Après un état de l'art sur ces domaines, je présenterai différents travaux menés dans le cadre du projet *Ultraspeech2* (voir section 4.10), du projet de fin d'études d'ingénieur de Maël Pouget, et de la thèse de Florent Bocquelet (voir section 4.5).

1.1.1.1 Reconnaissance visuelle et audiovisuelle de la parole

De nombreuses études se sont penchées sur l'exploitation des mouvements labiaux pour la reconnaissance vocale, soit en complément du signal acoustique pour améliorer la robustesse au bruit - on parle alors de reconnaissance audiovisuelle de la parole (*audiovisual speech recognition, AVSR*) - soit en se substituant au signal acoustique - on parle alors de reconnaissance visuelle de la parole (*automatic lipreading* ou *visual speech recognition*, VSR). Depuis les travaux princeps de Petajan [Pet84] en 1984, différentes techniques ont été proposées pour extraire d'une image d'un visage parlant des caractéristiques visuelles discriminantes. Une première approche consiste à extraire et paramétrer les contours des lèvres (externe et interne), à l'aide de techniques de segmentation automatique du type *active shape model* [LTB96], *active appearance model* [BSC16], ou plus récemment de réseaux neuronaux comme les *Constrained Local Neural Field* [BRM13]. Une seconde approche consiste à modéliser l'intensité de l'ensemble des pixels d'une région d'intérêt (englobant les lèvres), à l'aide de différentes techniques de ré-

duction de dimensions, comme par exemple l'analyse en composante principale (ACP) [BK94] ou la transformée en cosinus discret *Discrete consine transform* (DCT) [Hec+02 ; Mat+01]. Le décodage de ces caractéristiques au niveau phonétique et/ou lexical s'effectue classiquement avec les techniques utilisées en reconnaissance automatique de la parole "acoustique". Jusqu'au début des années 2010, ces dernières étaient principalement basées sur des modèles probabilistes de type HMM-GMM [1] [Pot+03]. Différentes stratégies de fusion des modalités acoustique et visuelle ont été proposées, dont les HMM-GMM multiflux (*multistream HMM*) [DL00], les HMM-GMM couplés (*coupled HMM*) [GPN02]), où des réseaux Bayésiens dynamiques à l'architecture plus complexe [Gow+04]. Depuis 2012 et l'avènement des techniques d'apprentissage profond (*deep learning*), ces approches laissent la place aux réseaux de neurones profonds [MMG15], et aux réseaux de neurones profonds "récurrents", notamment ceux de type LSTM (*Long Short-term Memory*) [WKS16 ; WS17]. Cette évolution est directement inspirée par les récentes approches en reconnaissance automatique de la parole acoustique [2]. La tendance actuelle semble être au développement de systèmes dit *end-to-end*, tel que le système *LipNet* proposé en 2017 [Chu+17] et constitué d'un unique réseau construisant une série de représentations (abstractions) intermédiaires, directement depuis les "pixels" (entrée du réseau) vers les "caractères" (sortie du réseau). Ce type d'approche vise à s'affranchir d'une extraction explicite et préalable de toute caractéristique visuelle haut-niveau (*hand-crafted feature*), d'une représentation phonétique intermédiaire, voir même de l'utilisation d'un modèle de langage entraîné indépendamment des modèles visuels. Pour plus de détails sur ces approches, le lecteur pourra consulter [Zho+14] et [Sch+17b].

Dans plusieurs articles récents (notamment ceux publiés dans des conférences de vision par ordinateur), les systèmes de lecture labiale automatique sont présentés comme potentiel moyen de communication alternative. Si les performances de ces systèmes ne cessent d'être repoussées, notamment grâce à l'apprentissage profond, elles risquent néanmoins de se heurter à une limite

[1] L'acronyme HMM-GMM fait ici référence à un modèle de Markov caché (HMM) dont la densité de probabilité d'émission associée à un état est un modèle de mélange Gaussien (*Gaussian Mixture Model*, *GMM*.

[2] Jusqu'au début des années 2010, l'approche privilégiée en reconnaissance automatique de la parole est en effet basée sur la modélisation, pour chaque contexte phonétique (triphone, quinphone, etc.) de séquences d'observations acoustiques (vecteur de coefficients MFCC, PLP, etc.), par un modèle de type HMM-GMM (voir [Hin+12] pour une revue de la littérature). Le décodage au niveau lexical est obtenu par l'ajout d'informations *a priori* (indépendantes des observations acoustiques) fournies par le "modèle de langage". Ce dernier peut prendre de multiples formes (modèles de type *n-gram*, modèles basés sur les réseaux de neurones, voir [DBM15] pour une revue de la littérature) et est entraîné sur un large corpus de texte. Depuis le début des années 2010, l'apprentissage profond (*deep learning*) est à l'origine d'un changement de paradigme en reconnaissance vocale. L'étape d'extraction de caractéristiques acoustiques expertes semble disparaître des systèmes récents au profit d'un apprentissage par le système de ses propres représentations internes directement à partir du signal brut ou de sa représentation temps-fréquence, à l'aide notamment de réseaux convolutionnels [Abd+14]. Par ailleurs, l'approche par HMM-GMM pour la modélisation acoustique semble laisser la place aux approches hybrides de type HMM-DNN pour lesquelles un réseau de neurones approxime la densité de probabilité d'émission associée à chaque état (à la place du GMM) [Maa+17]. Enfin, les approches *end-to-end* visant à résoudre le problème de reconnaissance à l'aide d'un unique réseau de neurones profond, en s'affranchissement de toute représentation phonétique intermédiaire, et d'un modèle de langage entraîné indépendamment du décodeur acoustique, semblent prometteuses. On citera par exemple le système *DeepSpeech2* de la société *Baidu* [Amo+16] qui combine couches de convolution, couches récurrentes (bidirectionnelles) avec une couche de sortie de type CTC (*connectionnist temporal classification*, [GJ14]) et qui prédit une transcription orthographique (et non phonétique) directement à partir de la représentation temps-fréquence du signal de parole.

naturelle. En effet, les mouvements des lèvres ne portent qu'une partie limitée de l'information phonétique. Un ordre de grandeur de cette quantité d'information nous est donné par les travaux sur la lecture labiale chez l'humain, comme ceux de Bernstein et Auer en 1996 [BA96]. Ces derniers montrent que dans le cas de syllabes isolées (ne pouvant s'apparenter à mot comme par exemple "oti" qui n'est pas un mot Français), le pourcentage de bonne reconnaissance au niveau phonétique à partir uniquement de la perception des mouvements labiaux, était d'à peine 30% (environ 70% d'erreur, voir Table 1 de [BA96]). Ce score augmente lorsqu'il s'agit d'une série de phrases, mais uniquement chez les personnes mal-entendantes. Ces personnes feraient appel à une capacité accrue de suppléance mentale, leur permettant de décoder un "mot visuel" ambiguë (par exemple "jambe" vs. "lampe") en s'appuyant sur le contexte. On peut raisonnablement penser que les systèmes récents de lecture labiale automatique, et notamment ceux s'appuyant sur des réseaux de neurones récurrents, cherchent à fonctionner de la même manière en exploitant des corrélations linguistiques à long-terme. Cependant, les performances des systèmes de lecture labiale automatique sont à mon sens insuffisantes pour envisager leur utilisation comme moyen alternatif de communication. A titre d'exemple, le taux de bonne reconnaissance (au niveau lexical) du système développé par *Google DeepMind* et l'université d'Oxford [CZ16], entraîné sur 4960 heures de vidéo (soit 118 116 phrases), et qui a fait l'objet d'une forte médiatisation, n'est que de 53%, soit environ un mot sur deux décodé de façon incorrect.

D'autres travaux comme ceux Soquet et coll. [SSL99], puis ceux de Wrench et Richmond [WR00], ont proposé d'exploiter pour la reconnaissance vocale, les mouvements de l'ensemble des articulateurs (langue, lèvres, vélum, acquis par EMA, mais également larynx, acquis par électroglottographie ou EGG). Les performances obtenues était similaires (voir légèrement supérieures) à celle obtenue en considérant le signal acoustique[3]. Certes, ces travaux ne visait pas la communication en parole silencieuse, dont nous verrons plus tard qu'elle soulève d'autres défis, mais ils confirment la faisabilité d'un décodage robuste au niveau phonétique et/ou lexical des mouvements articulatoires.

1.1.1.2 Synthèse articulatoire

Les recherches sur les interfaces de communication en parole silencieuse peuvent également être mises en relation avec celles sur la synthèse articulatoire, c'est-à-dire la synthèse vocale pilotée à partir de paramètres décrivant les formes et positions successives des différents articulateurs dans un conduit vocal donné. Ce conduit vocal est généralement construit à partir de l'anatomie d'un véritable locuteur (voir par exemple [BJK06]), à partir d'images ciné-radiographiques ou IRM. Les mouvements articulatoires sont classiquement enregistrés par articulographie électromagnétique (EMA) ou par IRM dynamique [Nar+14]. Plusieurs approches sont possibles pour la construction d'un synthétiseur articulatoire :

- L'approche dite "géométrique" qui consiste à construire un modèle de la configuration articulatoire (position et déformation des différents articulateurs), par analyse statistique

[3] dans le cas d'un système mono-locuteur, basé sur une approche par *HMM-GMM* [WR00]

de données anatomiques acquises, par exemple, par imagerie ciné-radiographique ou IRM [Mae90 ; Bad+02]. La synthèse sonore est ici obtenue par estimation, pour une configuration donnée, de la fonction de transfert acoustique du conduit vocal, puis par convolution (ou filtrage) d'une source glottique synthétique (voir par exemple [Mae82 ; MS87]). Dans le cadre des approches dites "biomécaniques", on cherchera de plus à construire des paramètres de contrôle encodant explicitement les commande musculaires à l'origine des mouvements de chaque articulateurs [BPP09 ; DH04].

- L'approche dite "par apprentissage" qui consiste à construire par apprentissage statistique supervisé un regresseur articulatoire-vers-acoustique. Les paramètres acoustiques sont généralement issues d'une modélisation de l'enveloppe spectrale, par exemple par *LPC* (*Linear Predictive Coding*) ou par décomposition mel-cepstrale généralisée. De nombreuses techniques de régression ont été proposées, basées par exemple sur des modèles de mélanges [TBT08] des modèles de Markov cachés (HMM) [ZNT11], des réseaux de neurones [Boc+14]. La synthèse sonore est ici réalisée à l'aide d'un *vocodeur*. Un filtre numérique dont les paramètres sont déduits de l'enveloppe estimée (par exemple le filtre *MLSA* [ISF83] dans le cas de l'analyse mel-cepstrale), est excité par un signal modélisant la source glottique. Dans sa forme la plus simple, ce dernier est constitué d'un bruit blanc pour les segments non-voisés, et d'un train d'impulsion à la fréquence fondamentale souhaitée pour les segments voisés (voir par exemple [TBT08]). Il peut aussi être construit à l'aide d'un modèle plus sophistiqué d'onde glottique, comme par exemple celui décrit dans [Deg+13].

1.1.1.3 Interface de communication en parole silencieuse (SSI)

Nous présentons à présent un état de l'art des études portant spécifiquement sur la conception de SSI. Au début des années 2000, on pouvait identifier trois techniques principales utilisées dans ces études :

- l'imagerie ultrasonore (ou échographie) proposée par le Professeur Bruce Denby (directeur de ma thèse de doctorat) à l'ESPCI ParisTech. Cette approche utilise une sonde échographique placée sous la mâchoire du sujet pour capturer de façon non-invasive et inoffensive les mouvements linguaux. Dans mes travaux, cette technique est complétée par une caméra vidéo placée devant le visage du locuteur afin de capturer également les mouvements labiaux. Mes travaux sur ce sujet seront décrits à la section 1.1.2.

- l'electromyographie (EMG), développée dans l'équipe du Professeur Tanja Schultz (Université de Karlsruhe puis Université de Brême en Allemagne), et qui est basée sur la capture de l'activité électrique des muscles de la face et du cou lors de leur contraction. Un pic d'activité électromyographique est généralement observé 60ms avant le début du mouvement articulatoire [MP04]. Dans la plupart des études sur les interfaces de communication en parole silencieuse, des électrodes de surface sont utilisées. Elles sont placées sur la peau ou niveau de la joue et du cou [Jou+06] et sont éventuellement organisées en matrice [Wan+13]. L'approche par EMG a initialement été mise en œuvre

pour la reconnaissance de la parole silencieuse [Mai+05 ; Jou+06 ; SW10 ; WS16] et plus récemment pour la conversion articulatoire-vers-acoustique, c'est-à-dire la synthèse d'un signal acoustique sans chercher à passer explicitement par une représentation intermédiaire de type phonétique ou lexicale [TWS09 ; JD17] (plus formellement, il s'agit d'un problème de régression entre deux espaces de données multivariées, qui sera détaillé à la section 1.1.2.3).

- la microphonie stéthoscopique ou *Non-audible murmur (NAM) microphone*, développée au sein du laboratoire du Professeur Shikano au Japon (avec de nombreux contributeurs comme Yoshitaka Nakajima, Panikos Heracleous, Tomoki Toda, Nick Campbell, etc.), et qui est basée sur la capture de la très faible activité acoustique intra-buccale, c'est-à-dire non rayonnée aux lèvres, présente lors d'un murmure dit "non-audible". Bien qu'il ne semble pas avoir de consensus sur ce concept (voir néanmoins la section II.B de [Sch+17a] pour une définition possible), on peut raisonnablement penser à une production non-audible par un auditeur placé à environ un mètre de distance du locuteur. Cette technique utilise un microphone passif placé souvent près de l'oreille et fonctionnant comme un stéthoscope médical, c'est-à-dire en capturant les ondes acoustiques (ici de très faible amplitude) se propageant dans les tissus. L'approche NAM nécessite donc la présence d'un léger flux d'air dans la cavité orale pour la production du murmure et ne permet donc pas *stricto sensu* une communication orale totalement silencieuse. L'approche NAM a été utilisée dans le cadre de la reconnaissance vocale [Nak+03], et de la conversion de voix [TNS12] (dans ce cas, la voix source est l'enregistrement NAM et la voix cible un enregistrement de voix modale ou chuchotée capturée par un microphone classique, éventuellement dans une session ultérieure). Par ailleurs, l'utilisation de ce dispositif par une personne ayant subi une ablation totale du larynx n'est a priori pas possible car cette dernière ne peux plus expirer de l'air en provenance des poumons vers les cavités orales (buccale et nasale) suite à la séparation de la trachée avec le conduit vocal. Des pistes pour ce type d'application sont néanmoins avancées dans [Nak+12].

D'autres techniques ont depuis été proposées pour capturer et décoder l'articulation silencieuse. Un axe de recherche très actif s'appuie sur la technique "d'articulographie magnétique portable" ou PMA (*Portable Magnetic Articulography*). Cette technique consiste à coller ou implanter par chirurgie des petites aimants permanents sur la langue et les lèvres, et à capturer, à l'aide de capteurs généralement placés sur une paire de lunette, les variations du champ magnétique ambiant (autour de la tête) lors de l'articulation. Cette technique a été mise en œuvre pour la reconnaissance de la parole silencieuse [Fag+08] ainsi que pour la conversion articulatoire-vers-acoustique [Gon+16]. Enfin, Birkholz et coll. développe l'opto-electro-palatographie, une technique basée sur l'intégration, dans un faux-palais en résine, d'un ensemble de capteurs optiques permettant de mesurer, pendant l'articulation, leur distance relative avec la langue, le voile du palais et les lèvres [SB17]. Ce même groupe travaille également sur l'utilisation de deux antennes dites "Vivaldi" (2-12 GHz), une jouant le rôle d'émetteur et collées sur la joue, la seconde celui de récepteur et placée sous le menton. Cette technique a pour l'instant été évaluée sur une tâche de reconnaissance de phonèmes [Bir+18].

Aussi, la conception de SSI sont devenues à ce jour un domaine de recherche bien identifié

et actif, réunissant des travaux en instrumentation, phonétique expérimentale, traitement du signal et de l'image et modélisation par apprentissage statistique. Un premier état des lieux des connaissances dans ce domaine a été effectué en 2010 sous la forme d'un numéro spéciale de la revue *Speech Communication* qui comprend notamment une première revue de la littérature dans ce domaine [Den+10]. En 2017, j'ai co-écrit une nouvelle revue de la littérature [Sch+17b] dans le cadre du numéro spécial de la revue *IEEE/ACM Trans. Audio, Speech and Language Processing*, intitulé *Biosignal-based Spoken Communication*, que j'ai co-édité [Sch+17a].

1.1.2 Travaux réalisés

Mes travaux portent sur la conception d'une SSI basée sur la capture des mouvements articulatoires par imagerie échographique et vidéo. Dans le dispositif envisagé, la sonde échographique est placée sous la mâchoire du locuteur et fournit une image d'une partie de la cavité buccale [4]. La sonde échographique est couplée à une caméra vidéo, placée en face des lèvres du locuteur. Comme l'illustre la figure 1.2, ce double système d'imagerie permet de capturer de façon inoffensive et non-invasive les mouvements de deux des principaux articulateurs : la langue et les lèvres [5]. Une analyse des images permet l'extraction de caractéristiques visuelles qui sont ensuite converties, soit en texte (reconnaissance), soit en un signal acoustique synthétique (conversion). La reconnaissance comme la conversion s'appuient sur l'apprentissage supervisé d'un modèle statistique associant trajectoires articulatoires d'une part, et cibles linguistiques (phonétiques et/ou lexicales) ou acoustiques d'autre part.

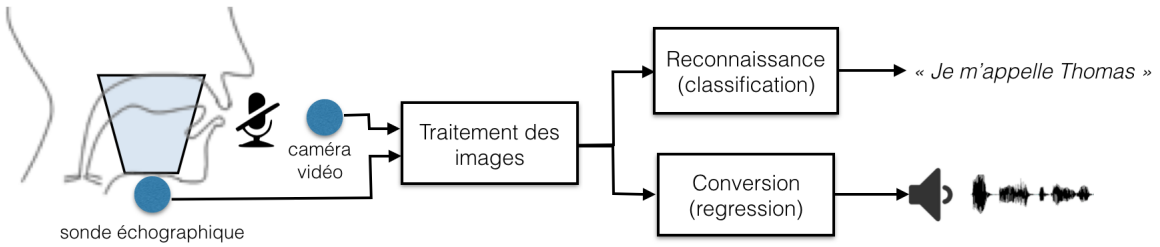


FIGURE 1.2 – Interface de communication en parole silencieuse basée sur la capture des mouvements articulatoires par imagerie ultrasonore et vidéo.

1.1.2.1 Dispositif expérimental

Le dispositif expérimental est présenté à figure 1.3. Il est basé sur une version modifiée d'un casque développé par la société *Articulate Instruments*, permettant de maintenir la sonde échographique en contact avec la mâchoire pendant l'articulation. La sonde est de type micro-

[4] Pour plus d'information sur l'échographie du conduit vocal, le lecteur est invité à consulter [HD09].

[5] Le voile du palais n'est visible sur les images échographiques que rarement (en cas de contact avec la langue) et très partiellement.

convexe, contient 64 ou 128 éléments en fonction de l'échographe utilisé [6], émettant une onde ultrasonore dans une bande de fréquence comprise typiquement entre 3 et 5 MHz (afin d'assurer un bon compromis entre résolution spatiale et profondeur d'exploration). Une profondeur d'exploration de 7cm, bien adaptée pour imager les mouvements linguaux, permet d'obtenir environ 80 images par seconde. Les mouvements des lèvres sont acquis à l'aide d'une caméra industrielle [7] permettant de contrôler précisément le temps d'exposition (typiquement fixé à 1/128 seconde dans mes expériences) et fournissant un flux vidéo à 60 images par seconde.

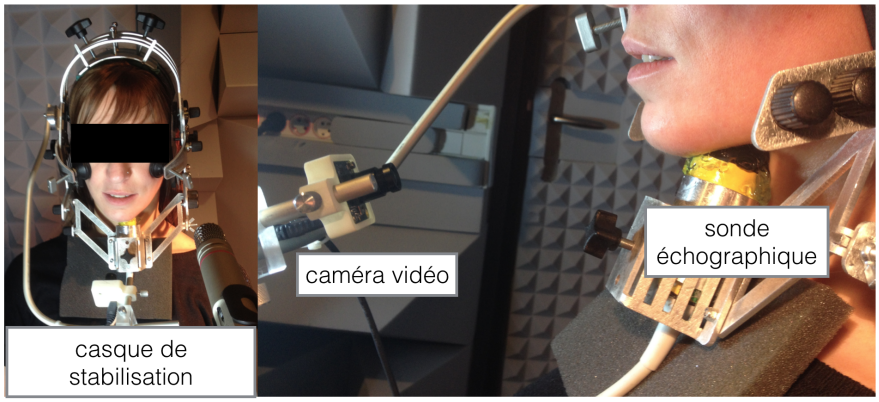


FIGURE 1.3 – Dispositif expérimental pour l'acquisition conjointe des mouvements de la langue et des lèvres par imagerie ultrasonore et vidéo. Extrait de [HB16].

Les flux d'images ultrasonores et vidéos sont enregistrés de façon simultanée, et synchronisés avec le signal acoustique (capturé à l'aide d'un microphone), à l'aide du logiciel *Ultraspeech* que je développe depuis 2008 [Hue+08a]. Actuellement dans sa version 1.3, ce logiciel, dont une capture d'écran est présentée à la figure 1.4, est aujourd'hui compatible avec plusieurs échographes, est téléchargeable gratuitement[8], et est utilisé par une douzaine de laboratoires [9].

1.1.2.2 Reconnaissance automatique de la parole silencieuse

Le décodage d'un flux d'image échographique et vidéo est un problème de reconnaissance de la parole visuelle, dont la résolution s'inspire naturellement des travaux sur la lecture labiale automatique mentionnés à la section 1.1.1.1. Dans ce domaine, la tendance actuelle est aux approches par apprentissage profond et aux systèmes *end-to-end* (comme par exemple [CZ16]), qui convertissent une séquence d'images brutes directement en une suite de mots, sans passer par une étape d'extraction de caractéristiques visuelles explicite, d'un décodage au niveau phonétique, et d'une régularisation basé sur l'utilisation d'un modèle de langage

[6] *Terason T3000*, *Telemed Echoblaster* ou *Telemed MicroUS*

[7] dans mon cas, il s'agit de la caméra *Imaging Source DFM 22BUC03-ML*

[8] `www.ultraspeech.com`

[9] Arizona State University (USA), Macquarie University, Center for Cognitive Science (Australie), Max Planck Institute for Evolutionary Antropolgy (Allemagne), Tuabjin University, School of Computer Science (Chine), University of Ottawa, Dept. of Linguistics, (Canada), etc.

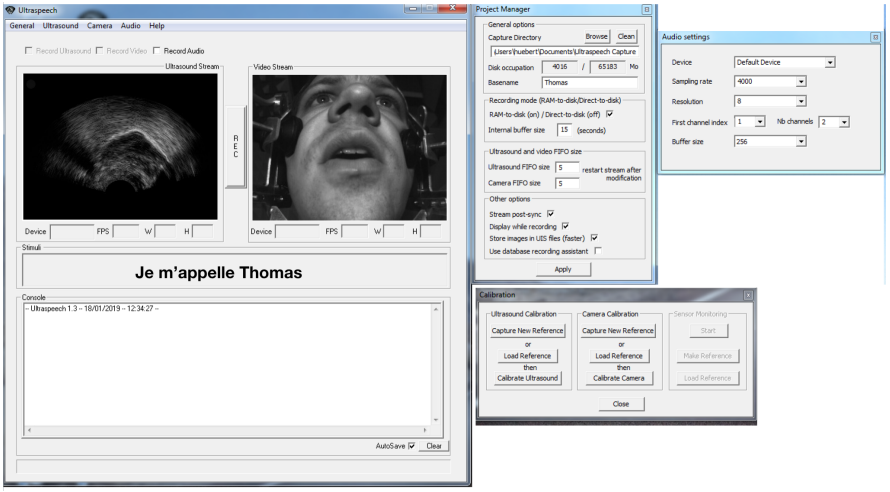


FIGURE 1.4 – Logiciel *Ultraspeech* développé pour la capture simultanée et synchrone de données échographiques, vidéos, acoustiques, et l'aide à l'enregistrement de grandes bases de données multimodales (version 1.3).

permettant d'introduire des informations linguistiques a priori. Cependant, si ces approches semblent donner des résultats prometteurs, elles sont extrêmement gourmandes en quantité de données d'apprentissage. A titre d'exemple, le système de reconnaissance vocale *Deep Speech 2* est entraîné sur un corpus de 11 940 heures de parole (soit environ 8 millions de phrases), et le système de lecture labiale décrit dans [CZ16] est entraîné sur 4960 heures de vidéo (soit 118 116 phrases). Ce passage à l'échelle est difficile dans notre cas au regard de la difficulté d'acquisition de très larges bases de données acoustico-articulatoires. Aussi, mes travaux ne se sont portés à ce jour que sur des systèmes dépendant du locuteur, entraînés sur des corpus de moins d'une heure de parole (soit entre 500 et 1000 phrases). De plus, les approches que j'ai proposées restent basées sur une première étape d'extraction de caractéristiques visuelles, suivie d'une seconde étape de décodage des caractéristiques visuelles extraites au niveau phonétique et lexical, à l'aide de modèles de type HMM-GMM.

Pour la première étape d'extraction des caractéristiques visuelles, j'ai proposé en 2017 une approche basée sur les réseaux de neurones à convolution (*convolutional neural network* ou CNN) [TH17] [10].

[10] Proposé par Lecun et coll. [LBH15], un CNN est une architecture dédiée à la classification des images. Les CNNs sont aujourd'hui utilisés dans de très nombreuses tâches de vision par ordinateur, telles que la détection d'objet [Sze+15], la reconnaissance de gestes [Bac+12 ; Ji+13 ; Kar+14 ; SZ14b], mais également la reconnaissance visuelle de la parole [Nod+14 ; CZ16]. Techniquement, un CNN est un réseau de neurones multi-couches profond composé de couches de convolution (*convolutional layer*), (ii) de couches de sous-échantillonnage (*pooling layer*), (iii) de couches dites "pleinement connectées ou denses" (*fully connected/dense layer*), et d'une couche de sortie (*output layer*). Une couche de convolution "convolue" l'image d'entrée ou la sortie de la couche précédente avec un filtre 2D dont les coefficients sont des paramètres libres estimés lors de la phase d'apprentissage. Cette opération de convolution permet de construire des représentations robustes aux translations. Une transformation non-linéaire est ensuite classiquement appliquée sur le résultat de cette convolution. La sortie de cette transformation est ensuite sous-échantillonnée afin de construire des représentations robustes au changement d'échelle. Cette séquence de couches "convolution/transformation non-linéaire-/sous-échantillonnage" est répétée un certain nombre de fois puis est généralement concaténée avec une architecture de type *percep-*

Dans le cadre du post-doctorat de Eric Tatulli, nous avons évalué différents manières de fusionner les deux modalités visuelles au sein d'un CNN. L'architecture la plus performante est présentée à la figure 1.5. Elle comporte notamment une couche de fusion permettant de combiner les représentations internes extraites des deux flux visuels (et se rapproche donc d'une stratégie dite de *middle fusion* d'un système de reconnaissance de la parole audiovisuelle [Pot+03]). Ce réseau est entraîné de façon supervisée par rétropropagation du gradient à partir de séquences d'images échographiques et vidéo, segmentées au niveau phonétique (typiquement par alignement forcé de la chaîne phonétique sur le signal acoustique, lorsque ce dernier est disponible). Après entraînement, les caractéristiques visuelles, utilisables ensuite pour la reconnaissance, sont définies comme la sortie de l'avant-dernière couche du réseau.

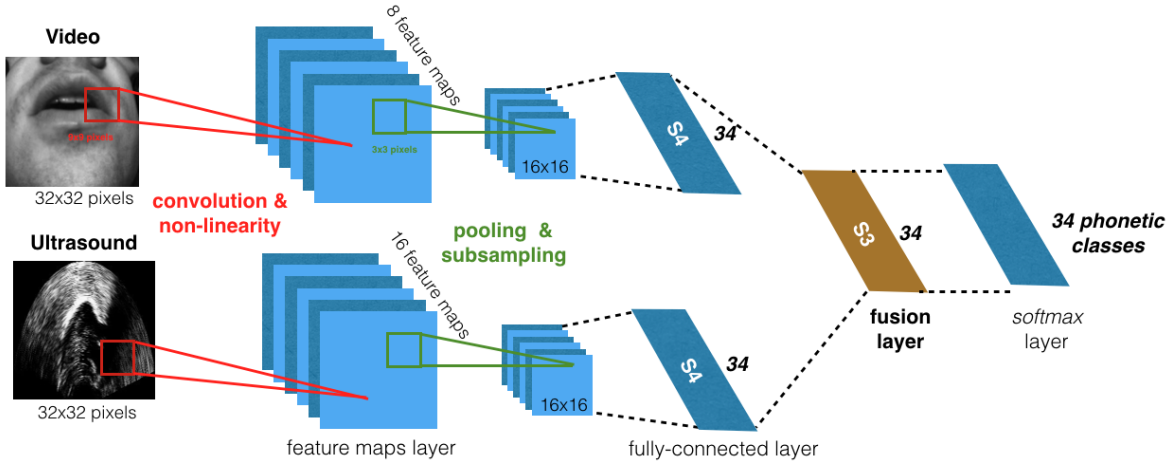


FIGURE 1.5 – Extraction de caractéristiques visuelles à partir d'images échographiques de la langue et d'images vidéo des lèvres à l'aide d'un réseau de neurones à convolution traitant conjointement les deux flux visuels (architecture dite "multimodale"). Extrait de [TH17].

Un décodeur phonétique de type HMM-GMM est utilisé pour décoder les trajectoires de caractéristiques visuelles extraites par le CNN au niveau phonétique. L'approche par CNN est comparée d'une part avec l'approche par ACP (*EigenLips*/*EigenTongues*) utilisée dans mes précédents travaux [Hue+07b ; Hue+08b ; Hue+10 ; HB16], et d'autre part à un décodage au niveau phonétique du signal acoustique (paramétré par analyse MFCC) associé aux séquences d'images échographiques et vidéo. La quantité d'information dans les flux visuels étant a priori plus limitée que celle dans le signal acoustique, la performance du décodeur acoustico-phonétique est considérée comme la borne supérieure de celle du décodeur visuo-phonétique. De façon importante, afin d'évaluer uniquement la quantité d'information que l'on peut décoder à partir des flux d'images échographiques et vidéo, et ce, indépendamment de toute information linguistique *a priori*, ni modèle de langage, ni dictionnaire, ni ensemble de

tron multi-couches, c'est-à-dire une ou plusieurs couches denses, dans lesquelles chaque neurone est connecté à l'ensemble des neurones de la couche précédente et effectue une transformation non-linéaire d'une combinaison linéaire de ses entrées. Enfin, la couche de sortie est également dense, et sa fonction d'activation est généralement de type *softmax* pour une tâche de classification (afin d'obtenir, pour chacune des classes considérées, la probabilité *a posteriori* d'observer cette classe, étant donnée l'image présentée à l'entrée du réseau), et linéaire pour une tâche de régression.

TABLE 1.1 – Performance des décodeurs visuo-phonétiques (VSR) basés sur une extraction des caractéristiques visuelles par ACP ou par CNN (erreurs d'insertion, d'omission et de substitution comprises). Comparaison avec un décodeur acoustico-phonétique standard (ASR), entrainé et évalué sur les mêmes corpus d'apprentissage et de test. Pour toutes les expériences, l'intervalle de confiance à 95% est d'environ 1.5%. Extrait de [TH17].

	ASR - MFCC	
T_p (%)	84	
	VSR - PCA	VSR - CNN
T_p (%)	74.7	**80.4**

règles phonotactiques ne sont ici utilisés. Une partie des résultats présentés dans [TH17] est rappelée à la table 1.1.

Ces résultats montrent une amélioration significative (de plus de 6%) des performances avec l'approche par CNN, par rapport à l'approche par ACP. A ce jour, plus de 80% des phonèmes sont correctement décodés uniquement à partir d'un flux d'images échographiques et vidéo du conduit vocal. L'étude de la matrice de confusion (voir [HB16]-figure 6 dans le cas de la configuration B2) montrent, de façon attendu, que la majorité des erreurs de décodage sont (i) des erreurs d'insertion ou d'omission de phonème, (ii) des erreurs liés au manque d'information sur la configuration du voile du palais et donc sur la nasalité, et (iii) des erreurs liées à l'absence d'activité laryngée (notamment en parole silencieuse) et donc sur le voisement. Une façon de combler ce manque est de s'appuyer sur des informations linguistiques a priori, telles que celles fournies par un modèle de langage au niveau lexical [11].

1.1.2.3 Conversion articulatoire-vers-acoustique

Comme indiqué précédemment, le second paradigme classiquement étudié dans le cadre des interfaces de communication en parole silencieuse est celui de la conversion directe de données articulatoires (ou electromyographiques) en une voix de synthèse intelligible. La notion de "direct" fait référence à deux contraintes. La première est d'effectuer cette conversion en *temps-réel*, c'est-à-dire avec un délai court et constant (typiquement $\leq 50ms$) entre le geste articulatoire et la synthèse sonore associée [Sch+17b]. La seconde contrainte est de synthétiser une voix dont le timbre est proche de celui de la voix de l'utilisateur. Cela permettrait notamment de préserver la voix d'une personne sur le point de la perdre, par exemple dans le cas d'une laryngectomie totale ou d'une maladie neuro-dégénérative.

La conversion directe est donc un problème de régression entre espaces de données multivariées. Dans notre cas, il s'agit d'une régression entre caractéristiques visuelles d'une part, et caractéristiques acoustiques d'autre part. Pour extraire ces dernières, mes travaux se sont pour l'instant appuyés sur une décomposition source-filtre du signal de parole afin d'extraire

[11] C'est notamment ce qui a été évalué dans [Cai+11] dans le cadre d'un système basé sur une extraction des caractéristiques visuelles par DCT et d'un décodage au niveau lexical par HMM-GMM.

et d'encoder l'enveloppe spectrale qui est le type d'information à mettre naturellement en regard de la configuration articulatoire capturée (partiellement) par échographie et vidéo [12]. Comme précédemment indiqué à la section 1.1.1.2, de multiples approches sont possibles pour aborder le problème de la régression articulatoire-vers-acoustique. J'ai étudié principalement la régression par modèle de mélange Gaussien, par réseau de neurones profond, et enfin par un modèle de type HMM-GMM. Ces approches sont brièvement décrites dans les paragraphes suivants.

Régression par modèle de mélange Gaussien (GMM)

Les bases théoriques de ce modèle, qui sera également largement utilisé dans mes travaux sur la conversion acoustico-articulatoire décrits au chapitre suivant (voir section 2.2.2.2), sont brièvement rappelées ici. On note $\mathbf{X}$ and $\mathbf{Y}$ deux vecteurs (colonne) aléatoires de dimensions respectives D_X and D_Y. On note $\mathbf{J}$ la concaténation de $\mathbf{X}$ et $\mathbf{Y}$ dans un nouveau vecteur colonne, tel que $\mathbf{J} = [\mathbf{X}^\top, \mathbf{Y}^\top]^\top$, avec $^\top$ l'opérateur de transposition. On note $p(\mathbf{x}|\mathbf{\Theta_X})$[13] la densité de probabilité (*probability density function* ou PDF) de $\mathbf{X}$, paramétrée par $\mathbf{\Theta_X}$. On note $\mathcal{N}(\mathbf{x}|\mu_\mathbf{X}, \mathbf{\Sigma_{XX}})$ une distribution Gaussienne sur $\mathbf{X}$ de vecteur de moyennes $\mu_\mathbf{X}$ et de matrice de covariance $\mathbf{\Sigma_{XX}}$. On note $\mathbf{\Sigma_{XY}}$ la matrice de covariance croisée entre $\mathbf{X}$ et $\mathbf{Y}$. Un modèle de mélange Gaussien (*Gaussian Mixture Model* ou GMM) sur ($\mathbf{X}$, $\mathbf{Y}$) est une somme pondérée de densités de probabilités Gausiennes, définie par :

$$p(\mathbf{j}|\mathbf{\Theta_J}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{j}|\mu_{\mathbf{J},m}, \mathbf{\Sigma_{JJ},}_m), \tag{1.1}$$

où M est le nombre de composantes dans le mélange. Pour chaque composante m, $\pi_m = p(m)$ est une 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 moyennes et $\mathbf{\Sigma_{JJ},}_m$ est la matrice de covariance définie par :

$$\mathbf{\Sigma_{JJ},}_m = \begin{bmatrix} \mathbf{\Sigma_{XX},}_m & \mathbf{\Sigma_{XY},}_m \\ \mathbf{\Sigma_{YX},}_m & \mathbf{\Sigma_{YY},}_m \end{bmatrix}. \tag{1.2}$$

Ces différents paramètres sont classiquement estimés par apprentissage supervisé à l'aide de l'algorithme *Expectation-Maximization* pour les GMM (non rappelé ici mais détaillé par exemple dans [Bis06], voir ch. 9).

Par théorème, si $\mathbf{J}$ suit une distribution Gaussienne, alors la distribution marginale sur $\mathbf{X}$ ainsi que la distribution conditionelle de $\mathbf{Y}$ sachant $\mathbf{x}$ suivent également une distribution Gaussienne. Ce résultat s'applique également aux distributions de type GMM de telle sorte que :

$$p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta_J}) = \sum_{m=1}^{M} p(m|\mathbf{x}, \mathbf{\Theta_X}) \mathcal{N}(\mathbf{y}|\mu_{\mathbf{Y}|\mathbf{x},m}, \mathbf{\Sigma_{YY}}_{|\mathbf{x},m}), \tag{1.3}$$

[12] J'ai notamment privilégié une modélisation de cette enveloppe basée sur l'analyse mel-cepstrale généralisée et une synthèse par le filtre numérique associé, connu sous le nom de filtre *MLSA* [ISF83].

[13] $p(\mathbf{x}|\mathbf{\Theta_X})$ est un abus de notation signifiant $p(\mathbf{X} = \mathbf{x}|\mathbf{\Theta_X})$. Les majuscules $\mathbf{X}$ sont utilisées pour les vecteurs aléatoires et les minuscules $\mathbf{x}$ pour les réalisations de ces vecteurs aléatoires.

avec

$$\mu_{\mathbf{Y}|\mathbf{x},m} = \mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}^{-1}_{\mathbf{XX},m}(\mathbf{x} - \mu_{\mathbf{X},m}), \tag{1.4}$$

$$\mathbf{\Sigma}_{\mathbf{YY}|\mathbf{x},m} = \mathbf{\Sigma}_{\mathbf{YY},m} - \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}^{-1}_{\mathbf{XX},m}\mathbf{\Sigma}_{\mathbf{XY},m}, \tag{1.5}$$

$$p(m|\mathbf{x}, \mathbf{\Theta}_\mathbf{X}) = \frac{\pi_m \mathcal{N}\left(\mathbf{x}|\mu_{\mathbf{X},m}, \mathbf{\Sigma}_{\mathbf{XX},m}\right)}{\sum_{i=1}^{M} \pi_i \mathcal{N}\left(\mathbf{x}|\mu_{\mathbf{X},i}, \mathbf{\Sigma}_{\mathbf{XX},i}\right)}. \tag{1.6}$$

La densité de probabilité conditionnelle formulée par l'équation 1.3 peut être utilisée pour effectuer la régression de $\mathbf{x}$ vers une estimée $\hat{\mathbf{y}}$ de $\mathbf{y}$. Lorsque cette regression est faite au sens de la minimisation de l'erreur quadratique moyenne pour chaque observation $\mathbf{x}_t$ (indépendamment des autres observations), $\hat{\mathbf{y}}_t$ s'écrit :

$$\begin{aligned}\hat{\mathbf{y}}_t = \mathrm{E}[\mathbf{Y}_t|\mathbf{x}_t, \mathbf{\Theta}_\mathbf{J}] &= \sum_{m=1}^{M} p(m|\mathbf{x}_t, \mathbf{\Theta}_\mathbf{X})\mu_{\mathbf{Y}_t|\mathbf{x}_t,m} \\ &= \sum_{m=1}^{M} p(m|\mathbf{x}_t, \mathbf{\Theta}_\mathbf{X})(\mu_{\mathbf{Y}_t,m} + \mathbf{\Sigma}_{\mathbf{YX},m}\mathbf{\Sigma}^{-1}_{\mathbf{XX},m}(\mathbf{x}_t - \mu_{\mathbf{X},m})).\end{aligned} \tag{1.7}$$

Cette régression est connue sous le nom de régression par modèle de mélange Gaussien (*Gaussian Mixture Regression* ou GMR) [GJ94]. Nous l'appellerons GMR-MSE dans la suite de ce manuscrit. Cette technique a d'abord été évaluée pour la conversion articulatoire-vers-acoustique "hors-ligne" [Hue+11b], puis "en ligne", dans le cadre du projet *Ultraspeech2* et du projet de fin d'étude de Maël Pouget, à l'aide d'un prototype temps-réel, réalisé sous Max/MSP, et illustré à la figure 1.6.

Une autre approche pour la régression par GMM a été proposée par Toda et coll. en 2007 dans le cadre de la conversion de voix [TBT07]. La conversion s'effectue ici non pas par trame mais "par séquence" : T vecteurs d'entrée $[\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T]$ sont convertis en une fois en T vecteurs de sortie $[\mathbf{y}_1, \ldots, \mathbf{y}_t, \ldots, \mathbf{y}_T]$. Brièvement, cette régression s'obtient en résolvant l'équation suivante :

$$\tilde{\mathbf{y}}_{\text{seq}} = \left(\mathbf{W}^\top \mathbf{D}^{-1} \mathbf{W}\right)^{-1} \mathbf{W}^\top \mathbf{D}^{-1} \mathbf{E}, \tag{1.8}$$

avec $\tilde{\mathbf{y}}_{\text{seq}} = [\tilde{\mathbf{y}}_1^\top, \ldots, \tilde{\mathbf{y}}_t^\top, \ldots, \tilde{\mathbf{y}}_T^\top]^\top$ un vecteur colonne de dimension $D_Y T$, $\tilde{\mathbf{y}}_t = [\mathbf{y}_t^\top \Delta\mathbf{y}_t^\top]^\top$ un vecteur obtenu en concaténant $\mathbf{y}_t$ avec ses dérivées temporelles première et seconde, $\mathbf{W}$ est une matrice encodant les relations linéaires entre les observations et leurs dérivées temporelles, $\mathbf{E}$ est un vecteur construit par concaténation des estimations $\hat{\mathbf{y}}_t$ obtenues par régression GMR-MSE (équation 1.7), et $\mathbf{D}$ est une matrice diagonale par bloc construite à partir des matrices de covariance conditionnelles (équation 1.5) et des responsabilités (équation 1.6), et ce, pour chaque observation $\mathbf{x}_t$ de la séquence d'entrée (voir [TBT07] pour plus de détails sur cette méthode). Cette approche est appelée ici GMR-MLPG car elle est une adaptation de l'algorithme *maximum likelihood parameter generation* ou MLPG proposé par Tokuda et coll. dans le cadre de la synthèse paramétrique par HMM-GMM [Tok+00]. La régression GMR-MLPG permet la génération de trajectoires acoustiques généralement plus précises que l'approche GMR-MSE. Cependant, la régression s'effectuant séquence-par-séquence, son implémentation en temps-réel ne pourra se faire qu'au prix d'une certaine latence, contrairement à la régression GMR-MSE.

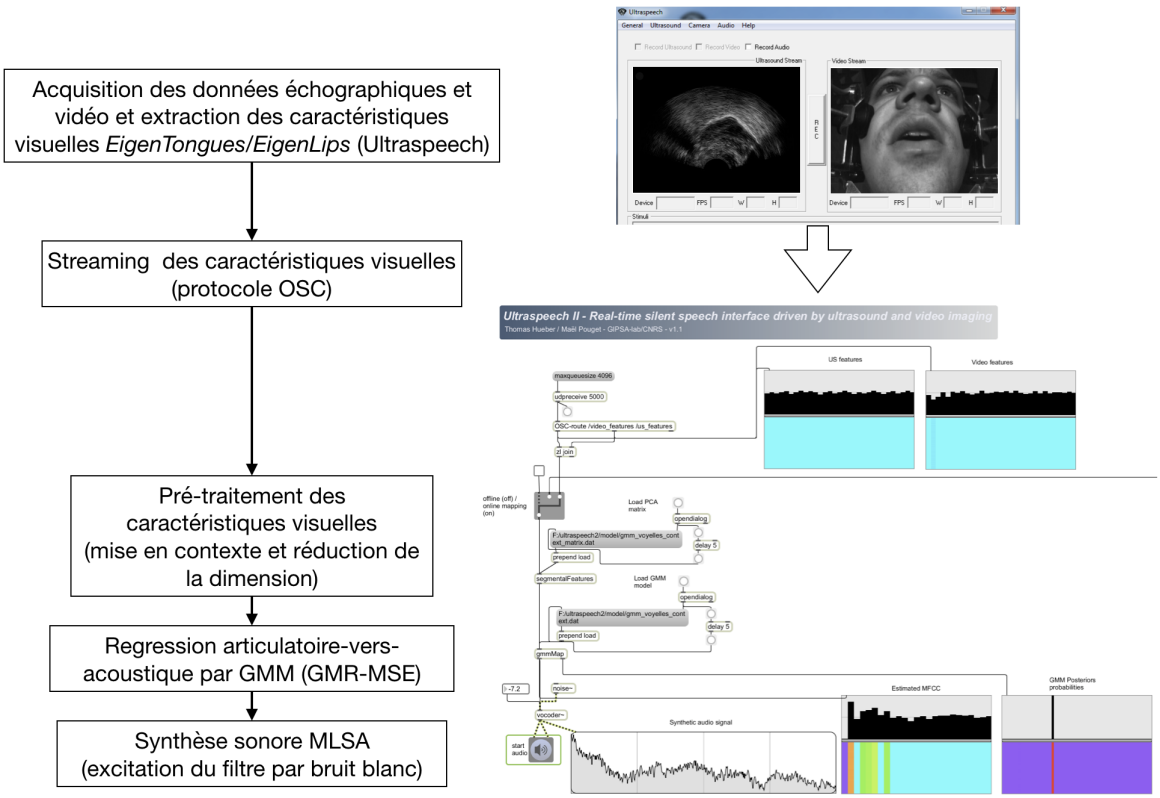


FIGURE 1.6 – Architectures du prototype de conversion articulatoire-acoustique en temps-réel Ultraspeech2. Une vidéo de ce prototype en fonctionnement est disponible sur `https://youtu.be/F2nslCgEa04`

Approche par HMM-GMM

Dans [HBD12] puis [HB16], j'ai proposé une extension de la régression par modèles de mélange Gaussiens (GMR-MSE et GMR-MLPG) aux modèles de Markov cachés [14]. L'objectif était double : d'une part, modéliser de façon plus explicite que dans l'approche par GMM l'évolution temporelle des relations acoustico-articulatoires, et d'autre part, de permettre l'introduction, dans la régression, d'informations linguistiques *a priori* afin de régulariser le problème mal-posé de la conversion de l'articulation silencieuse (dû principalement à l'absence d'activité laryngée). En s'appuyant sur les notations mathématiques introduites précédemment dans le cadre de la régression par GMM, il s'agit de modéliser la densité de probabilité conjointe de $(\mathbf{X}, \mathbf{Y})$ par un HMM dont la densité de probabilité d'émission par état est une loi normale (ou un GMM) et défini par :

$$p(\mathbf{j}|\mathbf{\Theta_J}) = \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{\Theta_J}) p(\mathbf{j}|\mathbf{q}, \mathbf{\Theta_J}) \tag{1.9}$$

avec

$$p(\mathbf{q}|\mathbf{\Theta_J}) = \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1}q_t} \tag{1.10}$$

$$p(\mathbf{j}|q_t, \mathbf{\Theta_J}) = \mathcal{N}(\mathbf{j}|\mu_{\mathbf{J},q_t}, \mathbf{\Sigma_{JJ}}_{,q_t}) \tag{1.11}$$

et $\mathbf{q} = [q_1, ..., q_T]$ une séquence de T états, π_{q_1} une probabilité a priori, $a_{q_{t-1}q_t}$ une probabilité de transition et $\mu_{\mathbf{J},q_t}$ et $\mathbf{\Sigma_{JJ}}_{,q_t}$, le vecteur de moyenne et la matrice de covariance associée à la probabilité d'émission de l'état q_t telle que définie à l'équation 1.1. De façon importante, cette matrice de covariance est définie comme pleine et ce, afin de capturer les corrélations locales (pour chaque état) entre observations articulatoires et observations acoustiques. Le modèle associé à cette approche est illustré à la figure 1.7. Les paramètres de ce modèle sont estimés par apprentissage supervisé à l'aide de l'algorithme *Baum-Welch* (voir [Bis06], chapitre 13).

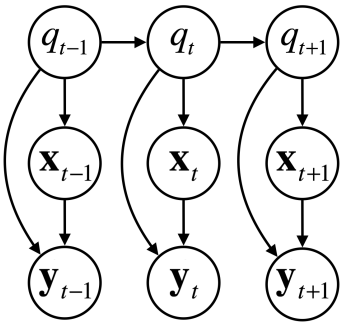


FIGURE 1.7 – Modèle graphique associé à la régression par HMM

Une estimée $\hat{\mathbf{y}}$ de la séquence $\mathbf{y}$, étant donnée une séquence $\mathbf{x}$, est définie comme celle maximisant la densité de probabilité conditionnelle $p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta_J})$, telle que :

$$p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta_J}) = \sum_{\forall \mathbf{q}} p(\mathbf{y}|\mathbf{x}, \mathbf{q}, \mathbf{\Theta_J}) p(\mathbf{q}|\mathbf{x}, \mathbf{\Theta_J}) \tag{1.12}$$

[14] Notons que des approches proches ont été proposées dans [HH04] puis [ZNT11] pour d'autres applications ; une étude des différences avec notre approche est proposée dans [HB16].

Cette densité de probabilité est approximée par $p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta_J}) \approx p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{q}}, \mathbf{\Theta_J})$ où $\hat{\mathbf{q}}$ est une séquence d'états définie telle que $\hat{\mathbf{q}} = \text{argmax}_{\mathbf{q}}\{p(\mathbf{q}|\mathbf{x}, \mathbf{\Theta_J}\}$ et obtenue à l'aide de l'algorithme de Viterbi (voir [Bis06], section 13.2.5). De façon similaire à la régression par GMM, deux estimateurs peuvent être définis. Le premier définit chaque trame de sortie $\hat{\mathbf{y}}_t$ selon le critère MSE tel que :

$$\begin{aligned}\hat{\mathbf{y}}_t = \mathrm{E}[\mathbf{Y}_t|\mathbf{x}_t, \mathbf{\Theta_J}] &= \mu_{\mathbf{Y}_t|\mathbf{x}_t,\hat{\mathbf{q}}_t} \\ &= \mu_{\mathbf{Y}_t,\hat{\mathbf{q}}_t} + \mathbf{\Sigma}_{\mathbf{YX},\hat{\mathbf{q}}_t}\mathbf{\Sigma}^{-1}_{\mathbf{XX},\hat{\mathbf{q}}_t}(\mathbf{x}_t - \mu_{\mathbf{X},\hat{\mathbf{q}}_t}).\end{aligned} \tag{1.13}$$

Le second (appelé HMM-MLPG) utilise l'algorithme MLPG pour estimer en une seule étape la séquence entière d'observations de sortie $\tilde{\mathbf{y}}_{\text{seq}}$ à l'aide de l'équation 1.8.

En pratique, et de façon similaire à un système de reconnaissance ou de synthèse vocale, chaque phrase du corpus d'apprentissage est modélisée par un HMM obtenu par concaténation des sous-modèles HMM associés aux phonèmes qui la composent. Il est ainsi possible, d'une part, de contraindre les modèles à respecter certaines séquences phonétiques imposées par un dictionnaire de prononciation, et d'autre part, d'introduire un *a priori* linguistique lors du décodage de la séquence d'états $\hat{\mathbf{q}}_t$ par l'intermédiaire d'un modèle de langage au niveau lexical. Cette approche est donc une sorte d'hybride entre reconnaissance visuelle et synthèse acoustique paramétrique. Les techniques de régression par GMR-MLPG (fournissant de meilleurs résultats que la technique GMR-MSE) et HMM-MLPG ont été évaluées et comparées dans [HB16] à l'aide de tests d'intelligibilité (l'auditeur doit retranscrire ce qu'il entend), et de tests d'identification de phonèmes dans des phrases porteuses (l'auditeur doit indiquer le phonème qu'il perçoit). Ces derniers permettent d'évaluer plus précisément la qualité segmentale de la synthèse. Les résultats des tests d'identifications sont présentés à la figure 1.8. On constate des taux d'identification d'un peu plus de 70% pour les voyelles (en langue Française) et d'environ 60% pour les consonnes (résultats obtenus sans dictionnaire de prononciation ni modèle de langage). Aucune différence statistiquement significative entre les approches par GMM et HMM ne sont observées pour les voyelles, mais l'approche par HMM fournit de meilleurs résultats que l'approche par GMM pour les consonnes. Ceci semble valider la nécessité d'une modélisation explicite de l'évolution temporelle des relations acoustico-articulatoires, ce que vise la technique de régression proposée.

Adaptation au locuteur et au mode de production, approche par DNN

Nous pouvons trouver deux limitations majeures aux travaux sur la conversion de l'articulation silencieuse présentés précédemment. Tout d'abord, il s'agit de systèmes dépendant du locuteur. Pour commencer à utiliser le système, un nouvel utilisateur doit donc enregistrer plusieurs centaines de phrases, ce qui n'est pas toujours facile à mettre en œuvre en pratique. De plus, les approches de régression par GMM et HMM fonctionnant par apprentissage supervisé, il est nécessaire de disposer d'un enregistrement de la voix cible pour estimer les paramètres des modèles de régression. Dans le cas des applications médicales envisagées, cela empêche *a priori* l'utilisation du système par les personnes ayant déjà perdu leur voix. Enfin, la plupart des études sur les interfaces de communication en parole silencieuse font mention d'une baisse importante des performances, en reconnaissance ou en conversion, lorsque le mo-

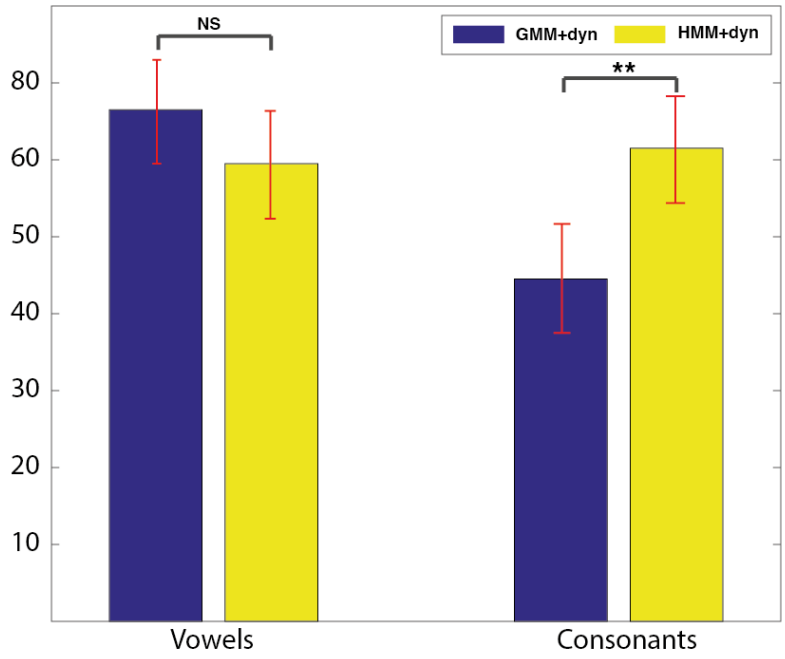


FIGURE 1.8 – Résultats des tests d'identification de phonèmes dans des phrases porteuses, synthétisées à partir de l'articulation silencieuse (capturée par échographie et vidéo), à l'aide des techniques de régression par GMM (GMR-MLPG en bleu) et HMM (HMM-MLPG en jaune). Précision de l'identification en %. Extrait de [HB16].

dèle articulatoire, préalablement entraîné sur des données acquises chez un locuteur vocalisant "normalement" (de façon audible), est utilisé pour décoder une articulation "silencieuse". Ce phénomène est observé à la fois par Janke et coll. dans le cas des approches par EMG [JWS11], mais également dans notre approche par échographie et vidéo où nous constatons une dégradation d'environ 15% du taux de reconnaissance phonétique entre une articulation "audible" et une articulation "silencieuse" [HB16]. Ces résultats semblent donc suggérer une modification des stratégies articulatoires entre parole modale et parole silencieuse. Pour quantifier plus finement ces différences, nous avons mené en 2011 une première étude pilote utilisant l'EMA pour caractériser ces modifications [Hue+11a]. Nous avons notamment observé une réduction de la coarticulation en parole silencieuse. D'autres différences ont été révélées par une étude de plus grande ampleur, menée par Dromey et coll., et publiée en 2017 [DB17]. De façon surprenante, les auteurs rapportent très peu de différences au niveau articulatoire entre parole chuchotée et parole modale. Les auteurs suggèrent que les modifications articulatoires en parole silencieuse sont principalement dues à l'absence de retour auditif. Aussi, la conversion articulatoire-vers-acoustique, si elle s'effectue en temps-réel et avec à une latence très faible (quelques dizaines de millisecondes), pourrait permettre à l'utilisateur de compenser ce phénomène, en s'appuyant sur le retour auditif synthétique.

Ces différents aspects (adaptation au locuteur, absence de données sur la voix cible et conversion en temps-réel) ont été abordés dans le cadre de la thèse de Florent Boquelet (voir section 4.5). Cette thèse s'inscrit dans un projet plus général portant sur le pilotage d'un synthétiseur vocal en temps-réel à partir de l'activité cérébrale (ces travaux seront détaillés à la section 1.2). Dans ce cadre, un synthétiseur vocal adaptable à un locuteur tiers (après une courte phase d'enrôlement), et pilotable en temps-réel à partir de données articulatoires acquises en parole silencieuse a été développé [Boc+16b]. Dans cette étude, les mouvements articulatoires sont acquis par EMA (3D). Tout comme le prototype *Ultraspeech2* décrit à la figure 1.6, la synthèse du signal sonore s'appuie sur un *vocodeur* de type MLSA [ISF83] (l'enveloppe spectrale est paramétrée par 25 coefficients mel-cepstraux). Une large base de données a été enregistrée sur un locuteur de référence à l'aide d'un système EMA 3D [15]. La régression articulatoire-vers-acoustique s'effectue ici à l'aide d'un DNN [16] [Boc+14]. Lors de la phase d'enrôlement, le nouvel utilisateur prononce une courte liste de phrase en parole silencieuse. Puis une régression linéaire est estimé entre ses trajectoires articulatoires et celles du locuteur de référence prononçant les même phrases. En phase de contrôle du synthétiseur, les positions articulatoires de ce nouvel utilisateur sont tout d'abord converties en temps-réel en positions articulatoires dans l'espace du locuteur de référence, puis en caractéristiques acoustique (toujours dans l'espace du locuteur de référence). Ces différentes étapes sont illustrées à la figure 1.9. La chaîne de traitement étant effectuée en moins de 30ms, l'utilisateur est susceptible d'exploiter ce retour auditif pour réguler sa production et maximiser l'intelligibilité de la parole de synthèse. Les résultats expérimentaux, évalués à l'aide de tests perceptifs (retranscription et identification) montrent qu'une bonne intelligibilité peut être obtenue pour les voyelles ainsi

[15] Les mouvements articulatoires sont capturés à l'aide de 3 bobines sur la langue, 4 bobines sur les lèvres, 1 bobine sur le voile du palais et 1 bobine sur les incisives inférieures pour capturer le mouvement de la mâchoire, soit un vecteur d'observation articulatoire résultant de dimension 27.

[16] De façon similaire au prototype *Ultraspeech2* (figure 1.6), la régression est effectuée à partir de plusieurs observations articulatoires consécutives afin d'exploiter des informations contextuelles.

que pour la plupart des consonnes. La synthèse de phrases complètes (mais plutôt simples et courtes) semble également possible.

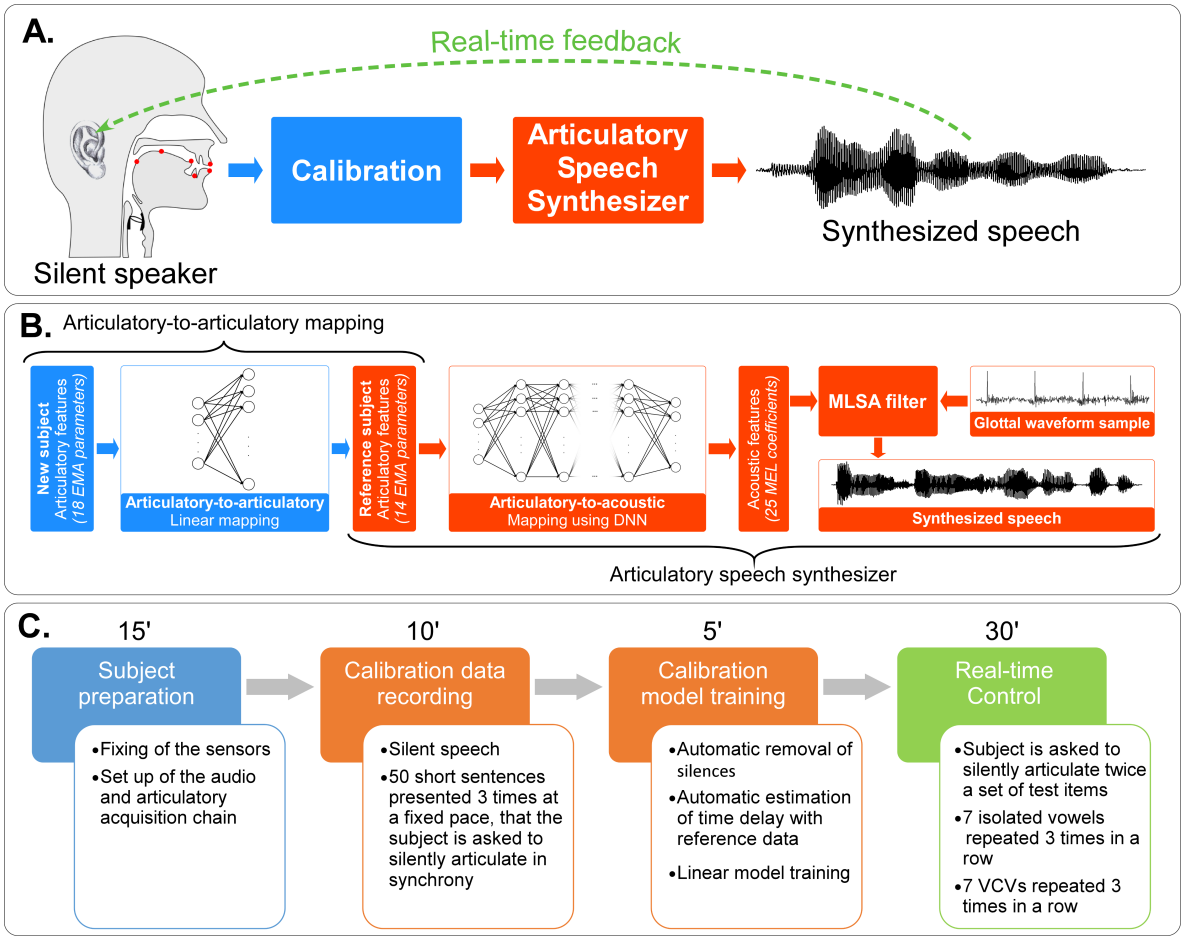


FIGURE 1.9 – Contrôle d'une synthétiseur articulatoire en temps réel a) schéma général, b) détail de la chaîne de traitement basée sur un réseau de neurones profonds et un *vocoder* de type MLSA, c) protocole expérimental pour évaluer la capacité d'un locuteur à contrôler le synthétiseur en temps-réel à partir de son articulation silencieuse. Une vidéo de ce système en action est disponible sur `https://doi.org/10.1371/journal.pcbi.1005119.s009`. Extrait de [Boc+16b].

1.1.3 Bilan et perspectives

Depuis mes travaux de doctorat portant sur le développement d'une SSI basée sur l'échographie (linguale) et la vidéo (labiale), différents travaux ont été menés, d'une part pour la reconnaissance de l'articulation silencieuse (décodage au niveau lexical), et d'autre part pour sa conversion temps-réel en un signal acoustique. Pour les résumer, on mentionnera l'utilisation des réseaux à convolution CNN pour la reconnaissance [TH17], la conception d'un prototype temps-réel basé sur la conversion visuo-acoustique par GMR (projet *Ultraspeech2*, prix Chris-

tian Benoît, voir section 4.10), la technique de régression par HMM-GMM [HB16 ; Hue+12], et enfin l'adaptation à un nouveau locuteur et le contrôle en temps-réel d'un synthétiseur articulatoire basé sur un DNN [Boc+14 ; Boc+15b ; Boc+16b]. Mes perspectives de recherche sont décrites ci-après.

Parole silencieuse et prosodie

Une dimension cruciale de la parole qui n'a pas encore été abordée ici est la prosodie. Cette dernière est définie par différents paramètres dont l'évolution de la fréquence fondamentale (f0), l'intensité sonore, la durée syllabique, la qualité vocale, et le degré d'articulation. La prosodie encode l'intonation globale de la phrase et une partie importante de l'expressivité et du contenu émotionnel de la parole [BS96 ; Sch03]. A ce jour, aucun système basé sur la conversion directe articulatoire-vers-acoustique ne permet la génération d'un contenu prosodique correct et il s'agit d'un problème majeur de cette approche, sur lequel je souhaite travailler. Je décris ci-après quelques pistes.

Tout d'abord, on peut faire l'hypothèse que certaines caractéristiques acoustiques de la prosodie pourraient être inférées à partir de l'activité articulatoire, comme par exemple l'intensité sonore, relativement corrélée à l'ouverture de la mâchoire [Geu01]. Ensuite, les durées segmentales, obtenues lors du décodage des mouvements articulatoires au niveau phonétique, pourraient être utilisées lors de la synthèse sonore, à condition cependant d'être adaptées afin de tenir compte des phénomènes d'anticipation articulatoire [SS14]. Enfin, nous avons montré expérimentalement dans [HBD12] que la caractéristique de voisement pouvait être inférée à partir des mouvements articulatoires avec une précision d'environ 80%. Cette inférence exploite probablement sur des corrélations "indirectes". Par exemple, le degré de compression labiale est plus important sur des consonnes occlusives bilabiales non voisées /p/ que voisées /b/ [LP70], la vitesse de réouverture des lèvres est plus rapide pour /p/ que pour /b/ [Fuj61], etc.

En revanche, la prédiction, à partir de l'articulation silencieuse uniquement, de la fréquence fondamentale "souhaitée" par l'utilisateur (c'est-à-dire celle obtenue en considérant la même phrase prononcée en voix modale) semble plus complexe. Cette inférence pourrait s'appuyer sur certains ajustements formantiques qui participent à la réalisation de traits prosodiques en parole chuchotée (voir par exemple dans le cas de la restitution des tons en Mandarin [Mey56]), et donc peut-être en parole silencieuse. Une autre possibilité serait de laisser l'utilisateur contrôler lui même, en temps-réel, la fréquence fondamentale lors de la synthèse. C'est ce que permet par exemple certains electrolarynx pour lesquels la fréquence d'excitation est asservie au débit d'air expiré [Has+07], est contrôlée directement par l'utilisateur à l'aide d'un bouton sensible à la pression [Tak+05], ou bien par l'activité des muscles du cou [Gol+03]. Une autre approche serait de s'inspirer des travaux sur le contrôle gestuel de la voix chantée, dans lesquels différents dispositifs comme des gants, claviers, pédales [DAl+05], ou tablettes graphiques [DRL11 ; Feu+17 ; PD16] ont été utilisées pour le contrôle de la hauteur et de la qualité vocale. Enfin, une dernière piste, en lien avec mes travaux sur la synthèse vocale incrémentale (qui seront décrit à la section 1.3), serait d'effectuer un décodage incrémental de l'articulation silencieuse au niveau lexical (voir par exemple [SA11] pour un résumé des différents algorithmes

proposés pour la reconnaissance vocale incrémentale), et de faire synthétiser les mots décodés au fil de l'eau, regroupés par exemple en syntagme, par un TTS (incrémental). La parole synthétique serait alors délivrée avec une latence variant entre un et quelques mots par rapport à leur articulation, mais avec un contenu prosodique plus acceptable que celui obtenu avec les techniques de conversions actuelles.

De façon générale, il y a un compromis à trouver entre qualité de la parole synthétisée (au niveau segmental et prosodique) d'une part, et réactivité de l'interface d'autre part. En effet, une conversion "à la trame" peut s'effectuer en moins de 10ms, mais ne permet pas d'inférer un contenu prosodique correct. A l'inverse, une synthèse par syntagme améliorera probablement la qualité prosodique, mais au prix d'une réactivité moindre. La recherche du compromis optimal qualité/réactivité devra faire l'objet de nouveaux travaux visant à étudier comment une personne utilisant une SSI comme voix de substitution, s'accommode d'une latence dans la conversion de son articulation silencieuse (on observera par exemple la manière dont son débit de parole évolue en fonction de la réactivité du système).

Ergonomie du système

Une critique récurrente sur le prototype actuel est son manque d'ergonomie, qui devra être améliorée. Dans le cadre d'une approche par échographie, cette amélioration passe par une miniaturisation des capteurs et leur intégration dans un dispositif portatif, de type *smartphone*. Des chercheurs de l'Université de Colombie Britannique ont récemment réussi à fabriquer des transducteurs ultrasonores non pas à partir de cristaux piezo-éléctriques (qui est la technique utilisée pour la conception des sondes médicales que j'utilise actuellement), mais à partir de minuscules membranes vibrantes faites d'une résine en polymère spécifique [GCR18]. Ces nouveaux transducteurs appelés polyCMUTs, pour *polymer capacitive micro-machined ultrasound transducers*, laissent envisager, selon les auteurs de l'étude, la création de sondes ultrasonores médicales très fines, à peine de la taille et de l'épaisseur d'un petit pansement, et donc tout-à-fait intégrable dans un *smartphone*. En attendant l'arrivée de cette nouvelle technologie, nous prévoyons à court terme d'intégrer les capteurs existants dans un casque léger et ergonomique, réalisé par impression 3D [Der+18]. Ce nouveau prototype est indispensable pour l'évaluation de notre SSI en contexte clinique.

1.2 Interface cerveau-ordinateur pour la restauration de la parole

1.2.1 Contexte et état de l'art

Depuis 2013, je suis impliqué dans un projet portant sur la conception d'une interface cerveau-ordinateur (BCI) dédiée à la restauration de la parole. L'objectif est de piloter en temps réel un synthétiseur vocal, uniquement à partir de l'activité cérébrale. Le locuteur articule "dans sa tête" (ou s'imagine en train d'articuler), mais ne bouge pas ses articulateurs, ni ne vocalise. Cette activité est enregistrée à l'aide d'électrodes disposées directement à la surface du cortex par voie chirurgicale (il s'agit donc d'une interface cerveau-ordinateur dite invasive). Le dispositif envisagé est illustré à la figure 1.10. Ce projet est porté par Blaise Yvert du laboratoire *BrainTech* de l'INSERM. Les travaux présentés ci-après ont été menés dans le cadre de la thèse de Florent Bocquelet, et des projets *Brainspeak* et *BrainCom* (voir section 4.10). Dans ces projets, je suis principalement impliqué dans le développement du synthétiseur vocal qui sera piloté, à terme, à partir de l'activité cérébrale.

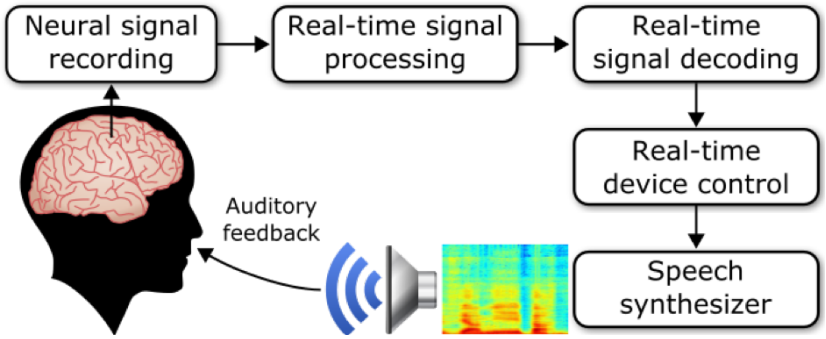


FIGURE 1.10 – Principe général de l'interface cerveau-ordinateur pour la restauration de la parole, envisagée dans le projet *Brainspeak*. Extrait de [Boc+16a].

Il s'agit d'un projet ambitieux et compliqué à mettre en œuvre (conception et certification des implants, autorisation éthique, recrutement des patients, etc.). Les résultats obtenus pour l'instant sont encore très préliminaires. En revanche, la conception de ce type d'interface est un domaine de recherche actuellement très actif. Dans cette section, je présenterai donc un état de l'art sur ce domaine, puis la démarche générale du projet *BrainSpeak*, et enfin certains résultats de la thèse de Florent Bocquelet [Boc17].

On peut distinguer deux type de BCI, les BCI "non-invasifs" qui utilisent le signal provenant d'un enregistrement electroencéphalographique (EEG), et les BCI "invasifs" enregistrant l'activité corticale et/ou sous-corticale à l'aide d'électrodes placées par voie chirurgicale à la surface du cortex ou implantées plus en profondeur dans le cerveau. Les premiers BCIs non-invasifs proposés pour restaurer une forme de communication exploite un processus de sélection de caractères par décodage du potentiel évoqué cognitif P300 [FD88]. La technique consiste à afficher sur un écran un tableau de caractères et à demander à l'utilisateur de se concentrer sur une lettre (une case du tableau). Les lignes et les colonnes sont alors successi-

vement mises en surbrillance. Lorsque c'est la ligne (ou la colonne) contenant la lettre choisie par le patient qui est mise en surbrillance, alors un potentiel évoqué de type P300 est censé être observé. D'autres caractéristiques du signal EEG peuvent être exploitées, comme par exemple les potentiels d'états stables (*steady-state potentials*) [Mid+00]. Si ces systèmes sont aujourd'hui utilisables en pratique, ils ne permettent cependant la saisie que d'une vingtaine de caractères par minute [Jar+15].

Deux types de capteurs sont souvent utilisés dans le cadre d'approches plus invasives, 1) l'électrocorticographie ou ECoG, qui utilise une matrice d'électrodes (de l'ordre de la centaine, espacées d'environ 10mm dans le cas de l'ECoG standard, à plusieurs centaines dans le cas de la micro-ECoG, espacées de moins d'1 mm), déposée directement à la surface du cortex, sous la dure-mère, et enregistrant les potentiels de champ local (*local field potential*, LFP), ou 2) les matrices de micro-électrodes (*microelectrode array* ou MEA), comme par exemple les *Utah array*, couvrant une surface beaucoup plus petite (de l'ordre de $1cm^2$), enfoncées très légèrement dans le cortex (sur une profondeur d'1 à 2 mm) et capables d'enregistrer directement les potentiels d'action à l'échelle d'un ou de quelques neurones.

Plusieurs études portent sur le décodage hors-ligne (non temps-réel) des données ECoG ou MEA enregistrées pendant la production de la parole "vocalisée" (*overt speech*), ou "imaginée" (*covert speech*), pour laquelle le locuteur s'imagine en train d'articuler un ou plusieurs phonèmes. Ainsi, le décodage au niveau phonétique a notamment été proposé par Mugler et coll. [Mug+14] dans le cas de données ECoG (parole vocalisée, 4 sujets, précision moyenne 36%), et par Brumberg et coll. [Bru+11] (1 patient, parole intérieure, précision moyenne entre 16 et 20%) puis par Tankus et coll. [TFS12] (parole vocalisée, 11 sujets, 93% de précision sur 5 voyelles) dans le cas de micro-électrodes. D'autres études proposent un décodage de données ECoG au niveau lexical, comme Kellis et coll. [Kel+10] (1 patient, parole vocalisée, 10 mots isolés). Herff et coll. [Her+15] ont récemment proposé un système de reconnaissance de la parole continue à partir de données ECoG directement inspiré de la reconnaissance automatique de la parole (classique), mixant donc décodage phonétique et lexical, avec régularisation par ajout d'information linguistique *a priori* par l'intermédiaire d'une réduction du vocabulaire autorisé et d'un modèle de langage statistique au niveau lexical. Cette étude est basée sur 7 sujets, un dictionnaire de 10 mots, et les performances obtenues au niveau phonétique sont de l'ordre de 50%.

Enfin, une seule étude démontre la faisabilité d'un contrôle en boucle fermée d'un synthétiseur vocal à partir de l'activité cérébrale. Il s'agit de l'étude de référence de Guenther et coll. [Gue+09] publiée en 2009. Un synthétiseur à formant (synthétiseur de Klatt) est contrôlé en temps-réel, à partir de l'activité cérébrale d'un patient souffrant d'un syndrome d'enfermement total (*locked-in syndrome*). Cette activité est capturée à l'aide d'une unique électrode placée dans le cortex moteur (gyrus précentral gauche). Les paramètres de contrôle du synthétiseur sont les fréquences des deux premiers formants F1 et F2. Ils sont estimés à partir de l'activité cérébrale à l'aide d'un filtre de Kalman. En partant d'une voyelle neutre, cette étude a montré que le sujet était capable, après 25 sessions d'entrainement et de test sur une période de 5 mois, d'amener le synthétiseur vers chacune des 3 voyelles extrêmes (/i/ comme dans *heat*, /u/ comme dans *hoot*, ou /a/ comme dans *hot*), avec une précision de 89% au bout de la dernière

(et 25ème) session.

Ces études pionnières montrent qu'un décodage de la parole à partir de l'activité cérébrale et le contrôle d'un synthétiseur est envisageable, à condition d'une part de capturer finement et de façon la plus complète possible l'activité du cortex, et d'autre part de choisir le bon espace de représentation de la parole en fonction de la position des électrodes. Ces principales contraintes motivent le projet *BrainSpeak* dont la démarche générale et les premiers résultats sont résumés dans les paragraphes suivants.

1.2.2 Travaux réalisés

1.2.2.1 Développement du synthétiseur vocal

La synthèse articulatoire est l'approche privilégiée à ce jour dans le cadre du projet *BrainSpeak*. Nous faisons ici l'hypothèse qu'une description articulatoire du signal de parole fait sens dans un contexte de BCI car elle s'appuie d'une part sur un ensemble restreint de paramètres (typiquement moins d'une dizaine), et d'autre part sur des paramètres « compatibles » avec l'activité du cortex moteur dont la somatotopie semble précisément être organisée par articulateurs [Bou+13]. En d'autres termes, nous faisons l'hypothèse, qui reste à valider, qu'une représentation articulatoire est potentiellement plus facile à mettre en relation avec l'activité cérébrale observée dans le cortex moteur qu'une représentation acoustique (construite par exemple à partir des paramètres d'un modèle d'enveloppe spectrale).

Par analogie avec les BCIs basés sur le contrôle d'autres actionneurs, comme par exemple un bras robotisé [Col+13], ce synthétiseur devra *a priori* satisfaire les contraintes suivantes :

- Il doit pouvoir être contrôlé à partir d'un nombre réduit de paramètres (à titre d'exemple, Collinger et al. rapportent dans [Col+13] le contrôle par une personne tétraplégique d'un bras artificiel à 7 degrés de liberté). Dans ce cadre, dans [Boc+14], un auto-encodeur [17] est utilisé pour extraire un nombre restreint de paramètres à partir des coordonnées 3D des bobines EMA (soit 27 coefficients dans notre cas). Des tests perceptifs ont permis de montrer qu'une bonne intelligibilité pouvait être obtenue à partir de 7 paramètres seulement (ce résultat est notamment cohérent avec d'autres études dont [BBB01]).
- Chaque paramètre doit être relativement dé-corrélé des autres et être associé à une dimension perceptive bien identifiable par l'utilisateur. Ce point n'a pas encore été abordé et des perspectives sont proposée à la section 1.2.3.
- La synthèse doit être réalisée en temps-réel (latence constante et probablement inférieure

[17] Un auto-encodeur (AE) est un réseau de neurones (de type perceptron multi-couche ou réseau récurrent) entrainé à prédire en sortie les observations qui lui sont présentées en entrée [Ben09]. Il est constitué d'un encodeur, qui convertit une observation d'entrée $\mathbf{x}$ en sa représentation dans l'espace latent (*embeddings*) $\mathbf{z}$ et d'un décodeur qui effectue la transformation inverse. Une structure type d'AE est présentée à la figure 1.11. Lorsque la dimension L de l'espace latent (couche de sortie de l'encodeur) est plus faible que celle des observations (notée F), l'AE peut être utilisé comme technique (non-linéaire) de réduction de dimension.

à 50ms) pour une utilisation compatible avec un contrôle du BCI en boucle fermée. C'est cette contrainte qui a motivé le développement du synthétiseur basé sur un réseau de neurones profond et contrôlable en temps-réel à partir de l'articulation silencieuse, décrit à la section 1.1.2.3 [Boc+16b].

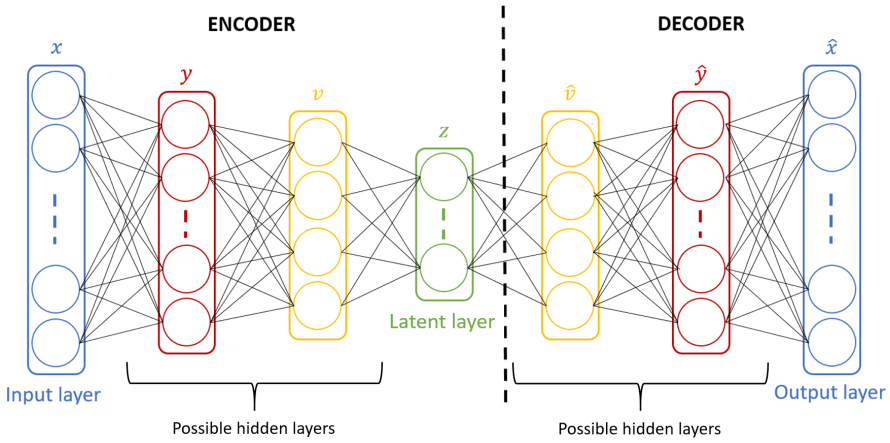


FIGURE 1.11 – Architecture d'un auto-encodeur (standard). Extrait de [Roc+19].

1.2.2.2 Décodage de l'activité cérébrale

Des enregistrements simultanés de données cérébrales (par ECoG ou MEA) et acoustiques ont été réalisés depuis 2014 chez 4 patients, au CHU de Grenoble, avec l'aide du Professeur Stephan Chabardès (neurochirurgien), pendant une chirurgie éveillée (pendant laquelle le patient peut réaliser une tâche, comme ici prononcer un ensemble de phrases). Un enregistrement simultané de données ECoG et acoustiques est présenté à la figure 1.12. Dans la lignée des différentes études présentées à la section 1.2.1, différentes expériences sont actuellement menées pour tenter de décoder ces données, en comparant différents espaces de représentation (acoustique, articulatoire), et en testant différentes techniques de classification et de régression. Sur une tâche de classification "parole *vs.* non-parole" chez un premier patient, par machine à vecteur support (SVM), une précision de l'ordre de 80% à la fois pour la parole vocalisée et, de façon intéressante, pour la parole imaginée également, a été obtenue (voir [Boc17], chapitre 7). Pour un second patient pour lequel davantage de données ont été enregistrées (en parole vocalisée uniquement), et avec une matrice ECoG plus dense, cette précision atteint les 96%. Sur ce même second patient, une précision de l'ordre de 75% a également été obtenue pour un décodage de la caractéristique de voisement.

1.2.3 Bilan et perspectives

Nos travaux sur la conception d'une interface cerveau-machine pour la restauration de la parole ont à ce jour porté sur la conception d'un synthétiseur articulatoire adapté à un contrôle par BCI [Boc+14 ; Boc+15b ; Boc+16b] et sur le décodage des premiers enregistrements de données

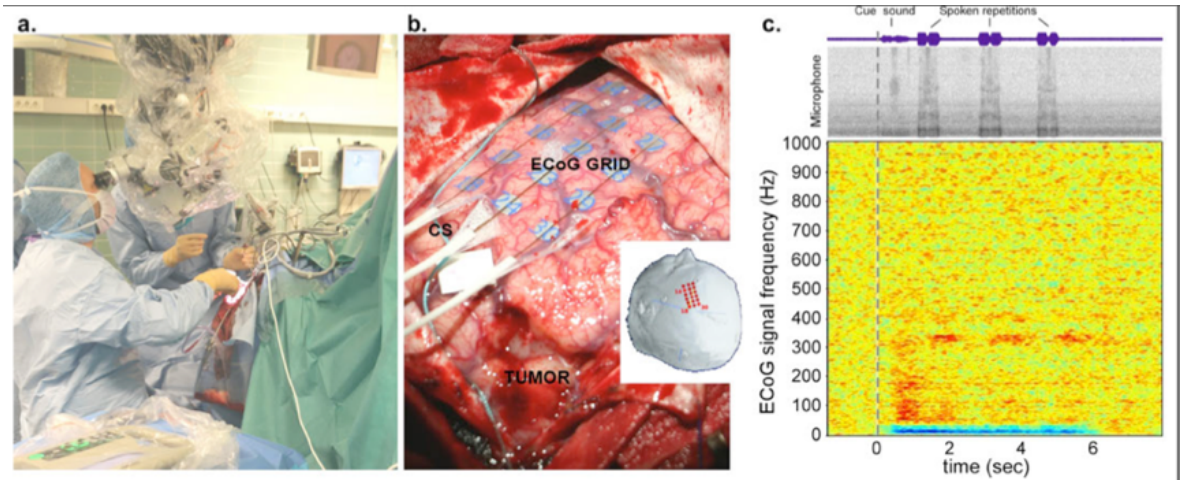


FIGURE 1.12 – Données préliminaires montrant l'activité électrique cérébrale enregistrée pendant la production de parole (CHU Grenoble Alpes) a) Salle d'opération du CHU de Grenoble Alpes équipée pour la chirurgie cérébrale éveillée (Pr. Chabardès) b) capteur ECoG à la surface du cortex c) enregistrement simultané du signal acoustique de parole et de l'activité cérébrale. Extrait de [Boc17].

ECoG acquis lors de chirurgies cérébrales éveillées. La suite des recherches sur les BCI pour la restauration de la parole s'inscrivent dans le cadre des projets ANR BrainSpeak et H2020-FETPROACT Braincom (voir section 4.10) et portent sur l'amélioration du synthétiseur vocal et sur le décodage de l'activité cérébrale. Nos pistes de recherche pour ces deux axes, dont certaines sont abordées dans le cadre de la thèse de Gaël Le Godais, sont résumées dans les paragraphes suivants.

Amélioration du synthétiseur articulatoire

A ce jour, chaque paramètre de contrôle du synthétiseur, issu de la compression des coordonnées EMA par auto-encodeur [Boc+14], ne contrôle pas une seule dimension articulatoire (par exemple, la variation d'un paramètre impliquera une modification à la fois de la position de la langue, et de la mâchoire). Ceci est potentiellement problématique dans le cadre d'un contrôle par BCI pour lequel il semble important que l'utilisateur puisse bien identifier les différents degrés de liberté du système à piloter [WW12]. Une piste d'amélioration serait donc de construire un synthétiseur dont les paramètres contrôlent des dimensions articulatoires orthogonales, comme par exemple la protrusion des lèvres, leur ouverture, la hauteur de la machoire, etc. Ceci peut s'obtenir soit à l'aide de méthodes linéaires de type ACP guidée [BBB01], soit, en poursuivant une approche par réseau de neurones, par conditionnement d'un GAN [18] ou d'un auto-encodeur variationnel (VAE) [19].

[18] Le principe des GAN est présenté à la section 3.

[19] Un auto-encodeur variationnel ou VAE [KW14] est une version probabiliste d'un AE. A la place d'une conversion déterministe entre le vecteur d'entrée $\mathbf{x}$ et sa représentation dans l'espace latent $\mathbf{z}$, l'encodeur d'un VAE convertit $\mathbf{x}$ en les paramètres d'une densité de probabilité conditionnelle $q_\phi(\mathbf{z}|\mathbf{x})$. De façon symétrique, le décodeur converti une observation dans l'espace latent $\mathbf{z}$ en les paramètres d'une autre distribution condi-

Une autre piste de travail porte sur la prise en compte du contexte articulatoire passé et/ou futur pour la régression articulatoire-vers-acoustique. Différents études rapportent que la prise en compte d'informations contextuelles améliore la régression (voir par exemple [TBT08 ; Hue+11b ; Boc+14]). Cependant, la taille optimal de ce contexte, et notamment celle du contexte futur, dont la prise en compte est susceptible de retarder la synthèse sonore dans le scénario d'utilisation envisagé (contrôle en boucle fermé par BCI), n'a, à notre connaissance, jamais été précisément déterminée.

Décodage de l'activité cérébrale

A ce jour, nous disposons de plusieurs enregistrements simultanés de données ECoG et acoustiques acquis en chirurgie éveillée, acquis dans le cadre du protocole clinique *BrainSpeak*. Bien que nous privilégions la piste d'un décodage au niveau articulatoire, piste récemment confortée par les travaux similaires de Chartier et coll. [Cha+18], nous envisageons également un décodage au niveau phonétique, en ligne avec les récents travaux de Herff et coll. [Her+15], ou acoustique tel que réalisé récemment par Angrik et coll. [Ang+19]) à l'aide de réseaux convolutionnels. Ces différentes pistes ont été décrites dans [Boc+16a] et sont illustrées à la figure 1.13.

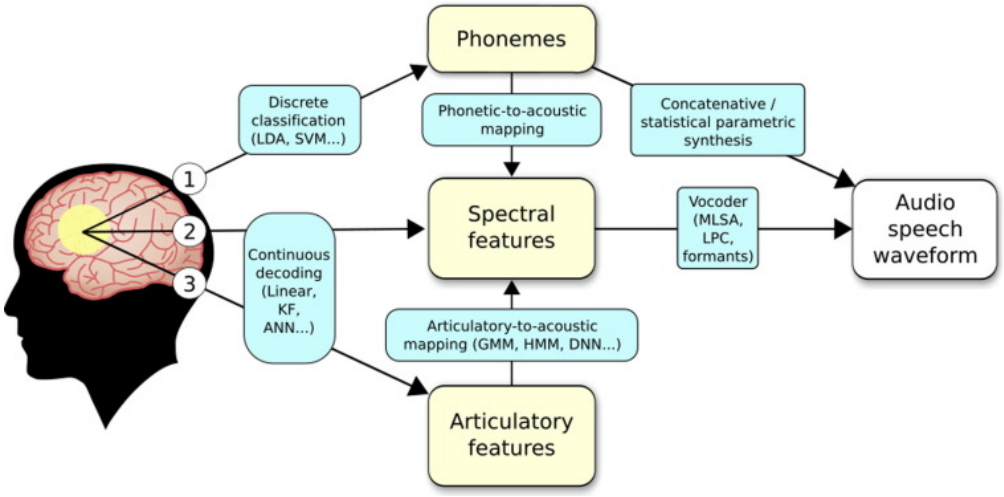


FIGURE 1.13 – Pistes envisagées pour le décodage des données cérébrales intra-corticales dans le cadre du projet *BrainSpeak*. Extrait de [Boc+16a].

Enfin, le projet européen *Braincom*, pour *Brain high-density cortical implants for cognitive neuroscience*, qui a débuté fin 2016, a pour objectif le développement d'implants plus denses et plus flexibles, pour l'enregistrement et la stimulation neuronale chez l'homme. Basés sur l'utilisation de nanomatériaux, afin notamment d'augmenter la densité d'électrodes et la flexibilité

tionnelle $p_\theta(\mathbf{x}|\mathbf{z})$. Plus formellement, un VAE modélise la densité de probabilité conjointe sur $\mathbf{x}$ et $\mathbf{z}$ qui s'écrit $p_\theta(\mathbf{x},\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$ avec θ l'ensemble des paramètres du modèle. La distribution a priori $p_\theta(\mathbf{z})$ permet de régulariser la structure de l'espace latent. On utilise classiquement une distribution Gaussienne tel que $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_L)$, ($\mathbf{I}_L$ is la matrice identité de taille L), ce qui favorise des dimensions latentes orthogonales.

des matrices ECoG, ces implants devraient également permettre d'améliorer les performances des interfaces cerveau-ordinateur pour la restauration de la parole.

1.3 Synthèse vocale incrémentale

Cette section présente les travaux que je mène depuis 2014 sur l'amélioration de la réactivité des systèmes *Text-to-speech* (TTS) au travers du paradigme de synthèse « incrémentale ». Cette section s'appuie notamment sur les travaux menés dans le cadre de la thèse de Maël Pouget (voir section 4.5) [Pou17] et du projet *SpeakRightNow* (voir section 4.10).

1.3.1 Contexte et état de l'art

Les systèmes TTS ont aujourd'hui atteint une qualité suffisante pour être déployés dans des applications tout public. Ils sont utilisés dans les GPS, pour la diffusion de messages dans les lieux publics, dans les robots humanoïdes, les assistants virtuels des *smartphones* ou encore des systèmes d'assistance aux personnes atteintes de déficiences visuelles (via par exemple le *voicemail* ou lecture automatique de pages Internet etc.). Par ailleurs, les systèmes TTS, éventuellement couplés à des interfaces d'aide à la saisie de texte (pictogramme, saisie prédictive, etc.), constituent un système complet de suppléance vocale, utilisé par exemple par des personnes ayant perdu l'usage de leur voix [20]).

Cependant, un synthétiseur TTS fonctionne très souvent à l'échelle de la phrase. En effet, l'analyse du texte et la synthèse sonore sont déclenchées à chaque fois que l'utilisateur a terminé la saisie d'une phrase complète, indiquée classiquement par la présence d'un marqueur de fin de phrase (tel que le point). La connaissance des limites de début et de fin de phrase est importante pour son analyse linguistique, notamment pour déterminer sa structure syntaxique, c'est-à-dire la fonction grammaticale de chacun des mots qui la constitue ainsi que les relations d'ordre et de dominance entre ces mots. Une bonne connaissance de la structure syntaxique de la phrase est primordiale pour que la parole de synthèse ait une bonne qualité segmentale, c'est-à-dire une restitution correcte de la chaîne phonétique cible, et suprasegmentale, c'est-à-dire une prosodie naturelle.

Dans ces travaux, nous faisons l'hypothèse que ce paradigme de synthèse "phrase-à-phrase" est peu adapté à une personne utilisant un TTS comme voix de substitution. Il introduit dans la communication une latence importante et proportionnelle à la longueur de la phrase. Cette latence peut être à l'origine d'une certaine frustration pour le destinataire de la communication qui est contraint d'attendre la fin de la saisie de chaque phrase pour comprendre et réagir aux propos de son interlocuteur, comme pour l'utilisateur du TTS, qui peut être tenté de simplifier son discours pour limiter cette attente, et maintenir ainsi une certaine fluidité dans l'interaction. Certains utilisateurs préfèrent donc déclencher la synthèse vocale après la saisie de chaque phonème ou de chaque mot. C'est notamment une option du TTS *Lightwriter SL40* de la société *Toby Churchill*, dédié aux personnes en situation de handicap. Cette stratégie diminue évidemment la latence entre la saisie du texte et sa synthèse et améliore donc la qualité de l'interaction. En revanche, cela se fait au détriment de la qualité de la parole de synthèse,

[20] par exemple dans le cas de certaines maladies neurodégénératives comme la sclérose latérale amyotrophique, voir par exemple `http://www.arsla.org/la-sla-en-chiffres/`

dont la génération ne s'appuie que sur la connaissance du mot à synthétiser, indépendamment de son contexte linguistique (sa position dans la phrase, sa fonction grammaticale, etc.).

Le paradigme de synthèse "incrémentale" vise à améliorer l'interactivité d'une communication orale effectuée par l'intermédiaire d'un TTS, en délivrant, au fur et à mesure de la saisie du texte, une parole de synthèse à la qualité proche de celle obtenue à l'aide d'un TTS classique travaillant à l'échelle de la phrase. La synthèse de la parole accompagne ainsi la saisie du texte, comme illustrée à la figure 1.14.

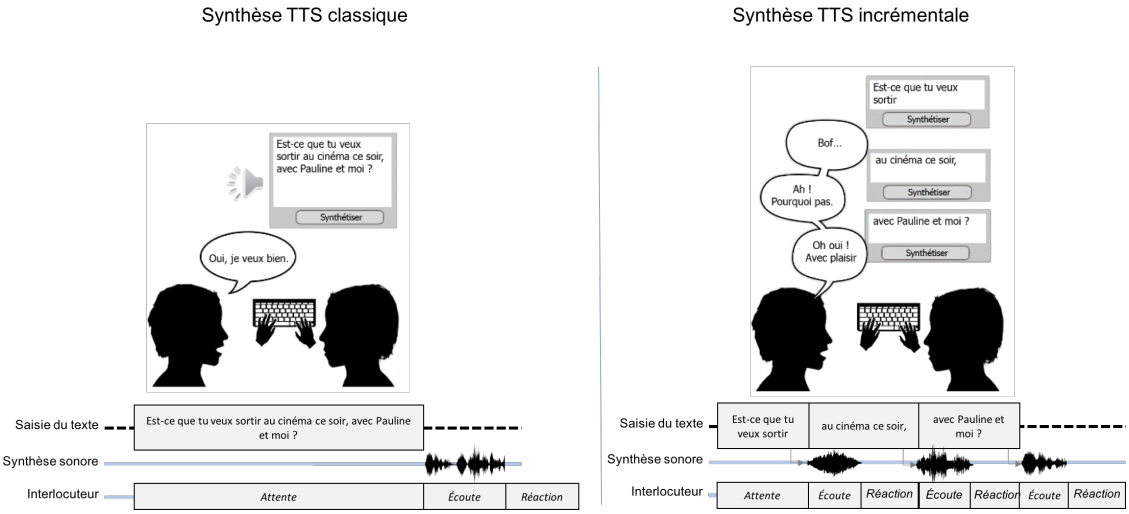


FIGURE 1.14 – Synthèse *Text-to-speech* classique (à gauche) et incrémentale (à droite). Le synthétiseur incrémental se comporte comme un lecteur humain, capable de débuter « en ligne » la vocalisation d'une phrase sans connaître cette dernière entièrement, avec une prosodie adaptée. Extrait de [Pou17].

Ce paradigme pose de multiples défis, car la plupart des modules d'analyse linguistique et de synthèse sonore d'un TTS classique [21] exploitent, lors de la synthèse d'un mot, des informations sur les mots qui le précèdent et lui succèdent - on parlera alors respectivement de contexte "passé" et "futur". Or le contexte futur n'est, par définition, pas accessible en synthèse incrémentale.

Par ailleurs, couplé à un système de reconnaissance (incrémental) de la parole, la synthèse TTS incrémentale pourrait être appliquée pour différentes tâches impliquant une conversion "parole-vers-parole", comme par exemple la traduction automatique (reconnaissance incrémentale dans la langue source, traduction dans la langue cible, synthèse incrémentale), mais également, et c'est ce qui m'a initialement motivé à travailler sur ce sujet, toute forme de conversion "inter-modale" gagnant à s'effectuer à faible latence, comme la conversion de l'articulation silencieuse, de l'activité cérébrale, de la Langue Parlée Complétée (LPC, voir section 1.4), etc.

[21] phonétiseur, analyseur morphosyntaxtique, génération de prosodie, choix de l'unité pour la synthèse concaténative, estimation des paramètres des modèles acoustiques pour la synthèse paramétrique statistique

Dans le cadre de la thèse de Maël Pouget, nous avons travaillé sur deux aspects de la synthèse TTS afin de la rendre incrémentale, à savoir, d'une part, un algorithme d'analyse morpho-syntaxique (*Part-of-speech tagging*) dit "à latence adaptative" [Pou+16] (voir section 1.3.2.1), et d'autre part, une procédure de création d'une voix de synthèse par HMM intégrant une incertitude sur le contexte futur [Pou+15] (voir section 1.3.2.2). Ces travaux sont brièvement décrits dans les sections suivantes, après une présentation de l'état de l'art sur ce domaine de recherche relativement récent et encore assez peu exploré.

Le concept de synthèse vocale incrémentale semble être formulé pour la première fois en 2008 dans un article de *Edlund* dans le cadre de l'amélioration de la réactivité d'un système de dialogue [Edl08]. L'idée est d'adapter à des événements extérieurs les instants auxquels est délivrée la parole de synthèse. Par exemple, on interrompt le flux audio lorsqu'un bruit potentiellement masquant est détecté et on le reprend, éventuellement quelques mots avant, une fois ce bruit disparu. Dans le même idée, *Buschmeier et Koll* proposent d'interrompre la parole de synthèse lorsque l'utilisateur ne regarde plus ou s'éloigne de l'écran [BK11]. Cependant, dans ces deux études, le texte est connu avant la génération (et la diffusion) de la parole de synthèse. A notre connaissance, c'est *Baumann et Schlangen* qui décrivent en 2012 dans [BS12] l'implémentation d'un TTS incrémental et en 2014 dans [Bau14] sa véritable évaluation. Ces travaux pionniers sont utilisés comme référence et seront brièvement décrits à la section 1.3.2.2.

Indépendamment de la synthèse vocale, le traitement "incrémental" de texte a également fait l'objet de plusieurs contributions. Ces dernières portent principalement sur l'analyse morpho-syntaxique incrémentale [BKM11], et sur l'analyse structurelle (en anglais *syntactic parsing* ou *prosodic phrasing*), qui vise à construire des "unités prosodiques" au fur et à mesure que le texte se dévoile[22] [MMI01 ; RRB02 ; KM14].

1.3.2 Travaux réalisés

1.3.2.1 Analyse morpho-syntaxique incrémentale

L'analyse syntaxique (ou analyse morpho-syntaxique) vise à déterminer de façon univoque la classe lexicale de chaque mot de la phrase à synthétiser. Cette étape vient traditionnellement après l'étape d'analyse morphologique qui fournit un ensemble de classes possibles pour chaque mot. Cette analyse va notamment s'appuyer sur l'analyse du contexte de chaque mot, qui va aider à lever une ambiguïté éventuelle sur sa classe lexicale. Une approche classique repose sur la modélisation *n-gram* (voir par exemple l'analyseur décrit dans [Bra00]). Il s'agit d'estimer et d'exploiter la probabilité conditionnelle d'observer une classe lexicale pour un mot donné, sachant les classes lexicales des n mots précédents. Tout comme certains modèles de langage au niveau lexical, ces probabilités sont estimées par comptage (et régularisation) de co-occurrences de classes lexicales sur des grandes bases de données textuelles. D'autres ap-

[22] Les unités prosodiques peuvent prendre la forme d'un groupe accentuel, c'est-à-dire un regroupement de mots par unité de sens, et d'un groupe intonatif, qui est constitué de groupes accentuels appartenant à une même phrase prosodique, délimités par exemple par des pauses [Ros+81].

proches sont possibles, comme l'utilisation de modèles discriminants comme les SVM (*support vector machine*) [GM04] ou les CRF (*conditional random field*) [Sok+10]. Plus récemment, des études exploitant l'apprentissage profond ont proposé de traiter simultanément plusieurs tâches telles que le prétraitement, l'analyse morpho-syntaxique, l'analyse structurelle, la phonétisation, etc., à l'aide d'un unique réseau de neurones censé extraire automatiquement la structure morphologique, syntaxique, voire sémantique d'une chaîne de caractère [Col+11 ; SZ14a].

Dans le cadre d'une analyse morpho-syntaxique incrémentale, l'objectif est de garantir, au fur et à mesure que le texte se dévoile, une classe lexicale correcte pour chaque mot, en l'absence d'information sur leur contexte futur (les mots non encore saisis). Pour aborder ce problème, nous avons proposé une première approche basée sur la prédiction en ligne, étant données les classes lexicales des mots précédents déjà estimés, du nombre de mots supplémentaires à saisir pour garantir la stabilité de la classe lexicale. Cette prédiction est réalisée à l'aide d'un ensemble d'arbres de décision dont les paramètres sont estimés par apprentissage supervisé à partir d'un grand corpus de texte. Lors de cette phase d'apprentissage, on réalise l'analyse morpho-syntaxique en dévoilant chaque phrase mot après mot, et on estime, pour chacun des mots, la taille du contexte futur qui stabilise la valeur de la classe lexicale. L'analyse morpho-syntaxique résultante est donc à latence variable (typiquement 1, 2 ou 3 mots). Cette approche, qui vise à trouver un compromis entre réactivité et précision de l'analyse, est illustrée à la figure 1.15.

Nous avons évalué cette approche en s'appuyant sur un analyseur *n-gram* du Français [BA92][23] et un apprentissage des arbres de décision à partir d'un corpus de texte contenant 20154 phrases extraites du "Tour du monde en 80 jours" et de "Notre-Dame de Paris". En moyenne, cette approche permet d'estimer correctement la classe lexicale d'un mot avec une précision moyenne de $92,5\%$ et une latence moyenne de $1,4$ mots [Pou+16]. Plus précisément, comme illustré à la figure 1.16, environ 60% des mots sont synthétisés dès leur saisie, avec un taux d'erreur (pourcentage de faux positifs) de 5%. Dans 35% des cas, un mot supplémentaire (contexte droit) est nécessaire (taux d'erreur égal à 3%), et deux mots supplémentaires dans 4% des cas (taux d'erreur inférieur à 1%).

1.3.2.2 Synthèse paramétrique statistique incrémentale

Notre seconde contribution sur la synthèse incrémentale porte sur l'apprentissage et l'utilisation des modèles acoustiques dans le cadre de la synthèse paramétrique statistique par HMM. La synthèse paramétrique statistique est une technique de synthèse vocale utilisant d'une part, un modèle statistique (HMM-GMM, réseaux de neurones, réseaux récurrents, etc.) pour mettre en relation, par apprentissage automatique supervisé, une séquence d'observations acoustiques (source glottique et enveloppe spectrale) avec une séquence d'observations linguistiques (descripteurs symboliques), et d'autre part, un *vocoder* pour la synthèse du signal sonore à partir des observations acoustiques inférées (parmi les *vocoder* les plus populaires, on citera *MLSA*

[23] Un décodage de type Viterbi classiquement utilisé en analyse hors-ligne a cependant été remplacé par une approche sous-optimale de type *forward*, plus adaptée à un décodage en ligne.

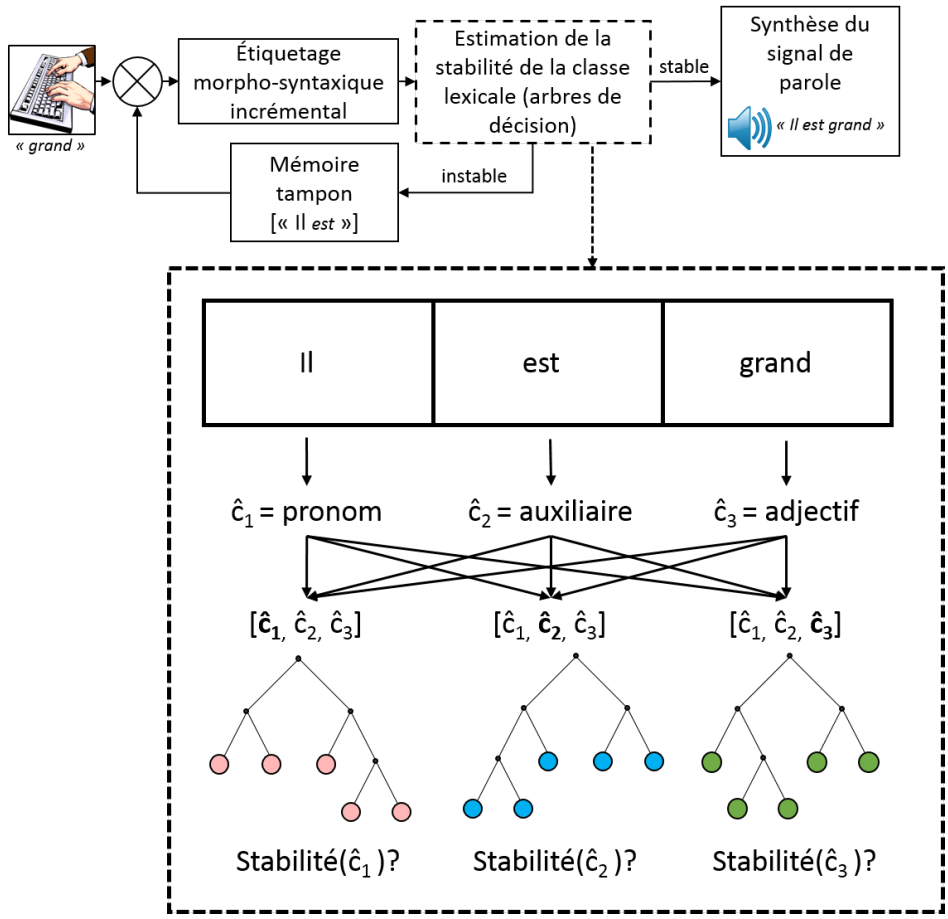


FIGURE 1.15 – Analyse morpho-syntaxique à latence adaptative pour la synthèse TTS incrémentale. La phrase en cours de saisie est "Il est grand et sympathique.". L'utilisateur a déjà entré les mots "Il" et "est" mais leur classe lexicale a été jugée comme susceptible d'évoluer lorsque d'autres mots seront saisis - la synthèse est donc retardée. Lors de la saisie du prochain mot ("grand"), la sortie des arbres de décision indique que la classe lexicale des mots "Il", "est" et "grand" ne fait plus l'objet d'ambiguïté - les trois mots sont donc synthétisés ensemble. Extrait de [Pou17].

[ISF83], *STRAIGHT* [Kaw06], *WORLD* [MYO16], ou encore *GlotDNN* [Air+16]).

L'unité linguistique de base de la synthèse paramétrique statistique est le phonème. Aussi, une observation linguistique encode classiquement la nature du phonème, mais également le contexte linguistique de ce phonème. On distingue le niveau segmental - par exemple la nature des 2 phonèmes précédents et des 2 phonèmes suivants, la position du phonème dans la syllabe ou dans le mot parent, etc. - et le niveau supra-segmental - c'est-à-dire la position du mot parent dans la phrase, la fonction grammaticale du mot parent, la position du mot parent dans la phrase, dans le syntagme, etc. Si la connaissance du contexte segmental joue principalement un rôle pour la génération de la structure formantique, les informations contextuelles au niveau supra-segmental sont principalement exploitées pour la définition de la prosodie.

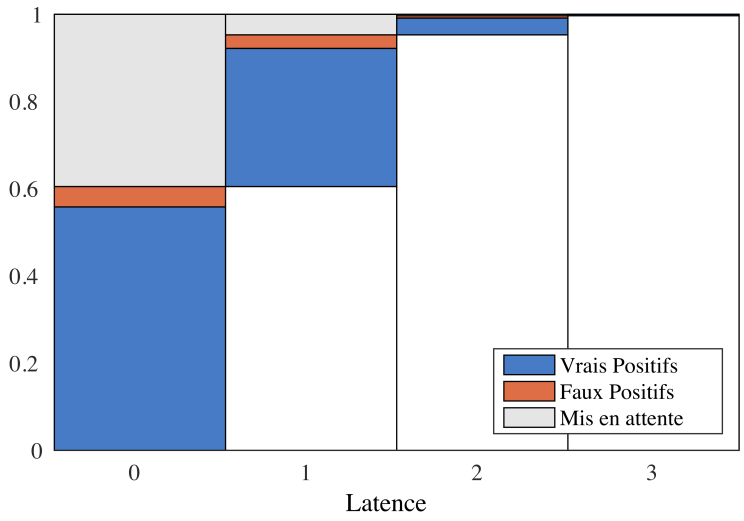


FIGURE 1.16 – Performance de l'analyse morpho-syntaxique incrémentale, à latence adaptative. Pourcentage de vrais positifs (bleu), de faux positifs (orange) et de mots dont l'analyse est reportée (gris), en fonction de la taille du contexte futur considéré (de gauche à droite : 0, 1 ou 2 mots). Extrait de [Pou17]

Le modèle le plus utilisé dans la période 2000-2012 a été le modèle semi-markovien ou HSMM (pour *hidden semi-markov model*) que l'on peut rapidement décrire comme un modèle HMM-GMM dont les probabilités de transition sont remplacées par une loi normale qui décrit le "temps" moyen d'occupation de chaque état lors de la génération d'un phonème. La synthèse paramétrique dite par HSMM a notamment été popularisée par le groupe de travail HTS[24]. Les modèles semi-markoviens ont depuis laissé leur place aux réseaux de neurones profonds [ZS14], aux réseaux récurrents [Lin+15], et tout récemment aux systèmes *end-to-end* basés sur les architectures neuronales auto-régressives de type *WaveNet*, tel que les systèmes *Tacotron* [Wan+17] et *Tacotron2* [She+18].

L'entraînement des modèles HSMM contextuels pour la synthèse vocale suit une procédure proche de celle utilisée pour la reconnaissance automatique de la parole par HMM-GMM. Il s'agit d'estimer, par apprentissage supervisé à l'aide de l'algorithme de Baum-Welch un modèle HSMM pour chaque classe phonétique mais également pour chaque contexte linguistique (segmental et supra-segmental). Cependant, compte tenu du nombre de contextes possibles (plusieurs milliers), le nombre d'occurrences de chaque contexte dans un corpus d'apprentissage, même grand, est souvent trop faible pour permettre une estimation robuste des paramètres des modèles décrivant chaque contexte. Aussi, on procède classiquement à une étape dite de "partage d'états" (*state-tying*) qui consiste à regrouper les états des modèles dont certaines informations contextuelles sont identiques et qui sont donc susceptibles de décrire des observations acoustiques proches [YOW94]. Ce regroupement des états s'effectue classiquement à l'aide d'arbres de décision binaire, où chaque nœud interroge sur la valeur d'une information contextuelle. Par exemple : "le phonème suivant est-il une plosive ?", "le mot pré-

[24] qui a produit le *toolkit* du même nom, voir `http://hts.sp.nitech.ac.jp`.

cédent est-il un adjectif ?", etc. Différents critères peuvent être utilisés pour la construction de ces arbres (maximum de vraisemblance, *minimum description length*, voir [Jen05]). En phase de synthèse, la génération d'une séquence d'observations acoustiques à partir d'une séquence d'observations linguistiques s'effectue classiquement à l'aide de l'algorithme MLPG [Tok+13] (déjà détaillé précédemment, voir équation 1.8).

Le problème majeur posé par le paradigme incrémental est le calcul des informations contextuelles nécessitant la connaissance des mots non encore saisis par l'utilisateur (par exemple la position du mot courant dans la phrase). Pour quantifier l'impact de ce manque d'information dans la synthèse du Français, une quantification fine du taux d'exploitation de chaque descripteur linguistique contextuel, lors de l'étape de partage des états (phase d'apprentissage des modèles) a tout d'abord été réalisée (non détaillée ici, voir par exemple la figure 4 de [Pou+15]). Baumann et coll. ont proposé dans [Bau14] de remplacer, lors de la synthèse incrémentale, chaque descripteur contextuel manquant (non calculable) par une valeur moyenne par défaut, estimée sur le corpus d'apprentissage. Nous avons proposé une approche alternative basée sur l'adaptation de la procédure de partage d'états décrite précédemment et utilisée lors de l'entraînement des modèles acoustiques [Pou+15]. L'objectif est de regrouper les états des modèles qui partagent la même incertitude sur un ou plusieurs descripteurs linguistiques contextuels. Cette procédure augmente à tord la variance des probabilités d'émission des états regroupés, et aboutit donc à des modèles moins précis. Elle permet cependant d'inférer, lors de la synthèse incrémentale, une observation acoustique potentiellement plus neutre que si un choix erroné avait été commis sur la valeur du descripteur manquant, tel que proposé dans [Bau14]. Nous faisons l'hypothèse que la parole de synthèse résultante sera plus naturelle. Cette hypothèse a été confirmée par un test perceptif comparant les deux approches, dont les résultats sont rappelés à la figure 1.17.

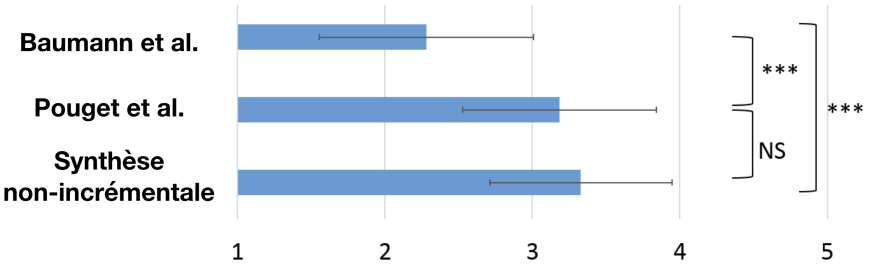


FIGURE 1.17 – Résultats du test perceptif évaluant l'approche proposée pour la synthèse TTS incrémentale par HMM, par comparaison avec la technique de référence décrite dans [Bau14] et la synthèse TTS classique (non-incrémentale). De gauche à droite, du moins au plus naturel sur une échelle MOS. Extrait de [Pou+15].

Un premier prototype d'un synthétiseur *Text-to-speech* incrémental combinant les deux techniques décrites respectivement dans les sections 1.3.2.1 (analyse morpho-syntaxique à latence adaptative) et 1.3.2.2 (procédure d'entraînement des modèles HSMM prenant en compte une incertitude sur la valeur de certains descripteurs linguistiques contextuels) a été implémenté dans le cadre du projet *SpeakRightNow*. Une vidéo de démonstration de ce prototype est accessible sur `https://youtu.be/eLsl1qx14JA`.

1.3.3 Bilans et perspectives

Dans le cadre du projet *SpeakRightNow* et de la thèse de Maël Pouget, nous avons développé un prototype complet de synthétiseur TTS incrémental, capable de délivrer la parole de synthèse au fur et à mesure de sa saisie. Ce système est basé sur un algorithme d'analyse morpho-syntaxique à latence adaptative [Pou+16] et sur une procédure d'entraînement des modèles acoustiques adaptée au contexte incrémental, dans le cadre de la synthèse paramétrique statistique par HMM [Pou+15].

L'algorithme d'analyse morpho-syntaxique à latence adaptative a pour effet de retarder la synthèse d'un mot dont la classe lexicale est susceptible de changer lors de la saisie des mots suivants par l'utilisateur. Ceci a pour effet de délivrer la parole de synthèse par groupes de 1 à 3 mots maximum, dont la stabilité de la classe lexicale (et donc notamment la phonétisation) est garantie. Si cette approche permet une synthèse réactive, certains regroupements ne sont pas satisfaisants. Par exemple, la phrase "Ces légendes me rappellent les temps anciens." sera décomposée en "Ces légendes" - "me rappellent" - "les temps" - "anciens". Un regroupement par syntagme (nominale, verbale, adverbiale, prépositionnel) de type "Ces légendes" - "me rappellent" - "les temps anciens" - pourrait être une alternative intéressante. Le découpage automatique d'une phrase en syntagme, également appelé *chunking* est une tâche classique du traitement automatique de la langue naturelle (TALN). Historiquement, cette tâche a d'abord été abordée à l'aide d'approches "par règles" s'appuyant sur la distinction "mot lexical/mot grammatical", certains mots lexicaux marquant la fin d'un groupe prosodique. L'algorithme le plus connu est probablement l'algorithme *chinks 'n' chunks* [LC96]. Cet algorithme est basée sur une analyse locale et ne nécessite pas la connaissance de l'intégralité de la phrase pour la construction d'un nouveau syntagme. Il est donc bien adapté à la synthèse incrémentale.

Les approches par apprentissage automatique, plus performantes, ont progressivement remplacées ces approches par règles. Sur la base de large corpus étiquetés, un modèle est entrainé à inférer automatiquement les frontières des groupes prosodiques, à partir d'un ensemble de descripteurs linguistiques (pour un mot donné, on considérera par exemple sa catégorie grammatical, celles des mots adjacents, la longueur de la phrase, etc.). Jusqu'au début des années 2000, le modèle le plus utilisé était l'arbre de classification et de régression (CART) (voir par exemple [SHY92]). Depuis quelques années, ce modèle semble aujourd'hui laisser la place aux réseaux de neurones profonds. Ces derniers permettent notamment de résoudre simultanément, et avec un unique modèle, différentes tâches d'analyse linguistique, comme l'analyse syntaxique (*part-of-speech tagging*) et l'analyse structurelle (*chunking*/*syntactic parsing*) [CW08 ; Col+11 ; SG16] (mais également la recherche d'entité nommée, la phonétisation, etc.). On parlera alors de modèles multi-tâches.

Un avantage majeur de ces modèles est leur capacité à exploiter le niveau sémantique en s'appuyant sur la technique de plongement lexical. Cette technique vise à associer à un mot, un vecteur de nombres réels, appelé *word embedding*, encodant son sens (autrement dit, deux mots sémantiquement proches seront proches dans l'espace des *word embeddings*). Différentes techniques ont été proposées pour construire, à partir très large bases de données, l'*embedding*

d'un mot à partir de son contexte, parmi lesquelles la technique *word2vec* [25]. L'extraction d'un *word embedding* peut être le premier étage d'un réseau plus profond, entrainés (de façon supervisé) à extraire d'autres informations linguistiques comme la classe lexicale, la position du mot dans le syntagme, etc. Cette prédiction s'appuie largement sur le contexte adjacent de chaque mot, grâce à des architectures récurrentes de type LSTM bi-directionnels, qui exploitent à la fois le contexte gauche (mots précédents) et le contexte droit d'un mot (mots suivants). Ces techniques ne sont donc a priori pas directement applicables pour la synthèse incrémentale.

Une première piste serait, de façon similaire à nos précédents travaux sur l'analyse morpho-syntaxique incrémentale [Pou+16], d'entrainer un réseau récurrent uni-directionnel à estimer, en ligne et pour un contexte gauche donné, la taille du contexte droit nécessaire (*lookahead*) pour réaliser l'analyse structurelle. Une seconde piste, plus ambitieuse, serait d'entrainer un réseau à estimer, sur la base des mots déjà saisis (contexte gauche) et de la sémantique associée (encodée via les *word embeddings*), certaines informations sur les mots non encore saisis, et qui sont nécessaires, à la fois pour l'analyse linguistique, et pour la synthèse sonore (comme par exemple la classe lexicale probable du prochain mot, le nombre de mots restants dans le syntagme, le nombre de syntagmes restants, voir même la longueur de la phrase). Ces différentes pistes devraient permettre l'amélioration du synthétiseur TTS incrémental existant, et son utilisation concrète pour la suppléance vocale.

[25] Cette technique est basée sur la reconstruction d'un mot à partir de son contexte à l'aide d'un réseau de neurones. Proposée par Mikolov et al. en 2013, elle a été publiée sous la forme d'un *preprint* (arXiv :1301.3781).

1.4 Reconnaissance automatique de la Langue Parlée Complétée

1.4.1 Contexte et état de l'art

La *Langue française Parlée Complétée (LPC)* [LCL10] (ou *Cued speech* (CS) en anglais) est une technique de communication à destination des personnes mal-entendantes, inventée par Cornett en 1967 [Cor67] et basée sur le complément de l'articulation normale de la parole par des gestes spécifiques de la main. A l'aide de cinq positions et huit formes distinctives, la main vient compléter l'information portée par les lèvres et permet de coder l'ensemble des phonèmes d'une langue. Cette technique a été adaptée dans plus de 60 langues et est utilisée partout de le monde. C'est une alternative (ou un complément) à la langue des signes. Un système de reconnaissance automatique de la langue Française parlée complétée, éventuellement couplé à un système de synthèse TTS, constituerait un système complet de suppléance vocale pour une personne maîtrisant cette technique de communication.

La reconnaissance automatique de la LPC a fait l'objet de plusieurs travaux [Cap+04 ; Abo07 ; HBA10 ; HBH12]. Des artifices y sont utilisés afin de faciliter la détection, le suivi et la paramétrisation du mouvement des lèvres et de la main (lèvres colorées, marqueurs collés sur la main, gant instrumenté, etc.). L'étude de Caplier et coll. [Cap+04] se focalise sur le décodage des positions et des formes de la main uniquement. Le décodage des gestes labiaux et manuels au niveau phonétique a été proposé en 2010 par Heracleous et coll. [HBA10]. Dans cette étude, des artifices placés sur les lèvres et la main facilitent l'extraction de descripteurs de forme et de position, et un décodeur phonétique de type HMM-GMM est utilisé pour la reconnaissance. Cependant, dans cette étude, la segmentation temporelle du flux visuel au niveau phonétique est connue au moment de la reconnaissance. Une technique similaire est mise en œuvre dans [HBH12] pour la reconnaissance de mots isolés.

1.4.2 Travaux réalisés

Dans [Liu+18b], nous avons proposé un système permettant, d'une part, de s'affranchir de tout artifice visant à faciliter la paramétrisation des mouvements labiaux et manuels, et d'autre part, de décoder de la parole continue (et non des phonèmes ou mots isolés). Le système proposé, inspiré de mes travaux sur la reconnaissance visuelle de la parole (voir section 1.1.2.2), fait intervenir trois étapes : i) l'extraction automatique de deux régions d'intérêt contenant respectivement la main et les lèvres, ii) l'extraction de descripteurs caractéristiques à l'aide de CNNs, et iii) un décodage au niveau phonétique et/ou lexical par HMM-GMM. Ces étapes sont illustrées à la figure 1.18. L'évaluation expérimentale a montré qu'une précision de l'ordre de 60% au niveau phonétique était atteinte, en l'absence de toute information linguistique a priori, ce qui est une amélioration significative par rapport à l'état de l'art. Une partie importante des erreurs de décodage restantes sont des erreurs d'insertion ou d'omission de phonèmes qui devraient être limitées par l'utilisation d'un dictionnaire de prononciation et d'un modèle de langage.

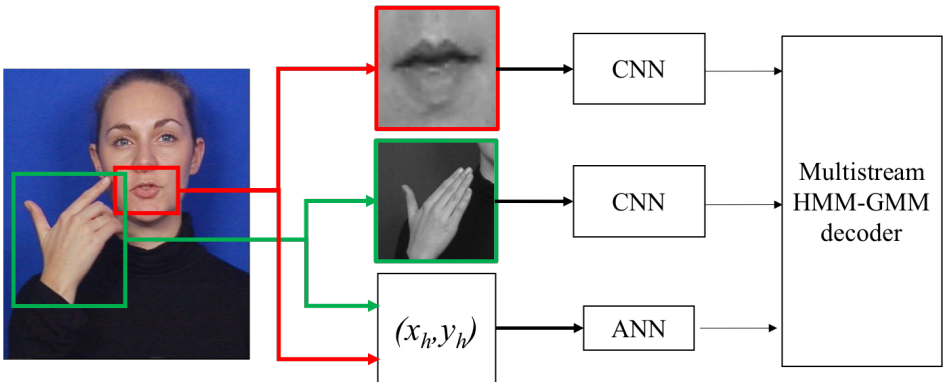


FIGURE 1.18 – Système de reconnaissance automatique de la langue Française Parlée Complétée basé sur les réseaux de neurones convolutionnels. Extrait de [Liu+18b].

Chapitre 2

Rééducation articulatoire assistée

Sommaire

Cet axe de recherche porte sur le développement d'outils d'aide à la rééducation orthophonique des troubles de l'articulation et sur leur évaluation clinique. Ces outils visent à améliorer la conscience qu'un locuteur a de ses propres mouvements articulatoires, et notamment ceux de sa langue. En effet, il est parfois difficile pour un patient d'associer une consigne verbale lui indiquant comment articuler un son (exemple : « pour produire le son [k], pressez le dos de la langue contre le palais puis relâchez ... ») à un but articulatoire et sensoriel concret qu'il doit planifier et atteindre. L'objectif est de rendre visible au patient les mouvements des articulateurs comme la langue dont il n'a, a priori, peu ou pas conscience. Je distingue deux paradigmes, l'illustration visuelle d'une part, et le retour visuel d'autre part (ou *visual feedback*). L'illustration visuelle vise à présenter au patient des mouvements linguaux cibles. Ces mouvements seront par exemple pré-enregistrés sur un locuteur de référence. Cette approche implique pour le patient de transférer dans son propre espace articulatoire ce qu'il voit "sur un autre". Dans le cadre du retour visuel, le patient visualise directement, et si possible en temps-réel, ses propres mouvements articulatoires.

Ces paradigmes d'illustration et de retour visuel exploitent la nature multimodale de la parole, et s'appuient sur les théories sensorimotrices qui supposent un lien étroit entre les mécanismes cognitifs de perception et de production. Nous faisons notamment l'hypothèse que la vision des articulateurs internes comme la langue pourrait compléter le retour somatosensoriel (auditif, kinesthésique, et proprioceptif), et améliorer (par exemple en accélérant) la rééducation d'un trouble de l'articulation. Ces considérations théoriques qui motivent les développements technologiques réalisés sont présentées à la section 2.1. A la section 2.2, je décrits mes différents travaux sur la conception de systèmes d'illustration et de retour visuel articulatoire, ainsi que différentes perspectives pour leur amélioration. A la section 2.3, je présente différentes études portant sur l'évaluation de ces systèmes, principalement pour la rééducation orthophonique, mais également pour l'apprentissage de la prononciation d'une langue seconde.

Ces différents travaux s'appuient sur la thèse de doctorat de Diandra Fabre [Fab16], sur le mémoire de Master Recherche en Sciences du Langage de Marion Girod-Roux [Gir17], sur celui de Xiaoou Wang [WHB14], ainsi que sur les mémoires d'orthophonie de Camille Bach et de Lorene Lambourion [BL15] (voir section 4.5).

2.1 Cadre théorique

2.1.1 Théories de la perception de la parole

Plusieurs théories sur les mécanismes cognitifs qui sous-tendent la perception de la parole ont été proposées. Les théories motrices (voir par exemple [LM85]) postulent que le but de la perception de la parole est de retrouver les gestes intentionnels du locuteur qui "structurent la matière sonore" [Sch01] et donc que le décodage de traits phonétiques fait appel à des mécanismes inférentiels de simulation motrice. L'objectif étant de retrouver les commandes motrices contrôlant le mouvement des articulateurs. En d'autre terme, percevoir de la parole, c'est percevoir des gestes, et la perception de la parole impliquerait donc une forme de production inconsciente. Cette théorie s'oppose aux théories dites "auditives" (voir par exemple [DLH04] pour une revue de la littérature) qui supposent que le décodage phonétique se baserait uniquement sur l'identification d'indices caractéristiques dans le signal acoustique (invariants acoustiques), qui seront décodés exclusivement au niveau du cortex auditif (sans faire appel donc à aucune simulation motrice). Cette hypothèse de simulation motrice inconsciente a été renforcée suite à la découverte d'un système de « neurones miroirs » [Pel+92 ; Riz+96]. Initialement observés chez les singes, il s'agit de neurones qui s'activent aussi bien lors de la production d'actions orientées vers un but, que lors de la perception, visuelle ou auditive, de ces mêmes actions, produites par un autre singe. Dans une étude utilisant l'IRM fonctionnel, Iacoboni et coll. ont rapporté un phénomène similaire chez l'homme, en observant l'activation des zones motrices de la main lors de la perception d'une action de préhension [Iac+05].

Cependant, les théories purement motrices ou purement auditives de la perception de la parole n'arrivent pas, à elles seules, à fournir une explication satisfaisante à certains phéno-

mènes. Par exemple, le phénomène d'équivalence motrice, c'est-à-dire le fait qu'un même son peut être produit par différentes configurations articulatoires[1], oblige un éventuel processus de simulation motrice à résoudre un problème mathématiquement mal-posé (relation non-injective ou *many-to-one* entre configuration articulatoire et enveloppe spectrale). A l'inverse, le phénomène de co-articulation, c'est-à-dire l'influence des phonèmes voisins sur le contenu acoustique d'un phonème complique fortement la tâche d'un mécanisme de perception basée exclusivement sur la recherche d'invariants acoustiques, comme le sous-tend les théories auditives de la perception de la parole (à moins de disposer d'un modèle acoustique pour chaque phonème, et pour chaque contexte phonétique possible).

Face à cette difficulté des théories motrices et perceptives à rendre compte de la complexité de la parole, Schwartz et coll. ont formulé une autre théorie, intitulée "théorie de la perception par le contrôle de l'action" (PACT) [Sch+12]. Cette théorie propose une unité de base associant systématiquement des aspects sensoriels et moteurs. Le décodage de traits phonétiques fait appel conjointement à des mécanismes d'extraction d'informations des signaux auditifs et visuels (*perceptual shaping*), et des connaissances procédurales motrices (*motor procedural knowledge*). Cette co-construction des représentations sensorielles et motrices aurait lieu dans les premières années de vie, au cours de processus d'imitation ou d'apprentissage sensorimoteur comme le babillage. Cette hypothèse d'un lien entre perception et simulation motrice est également soutenue par Grabski et coll. à l'aide d'une étude en IRM fonctionnelle montrant une activation des mêmes zones cérébrales lors de la production et de la perception de voyelles, et un possible recrutement plus important du système moteur en condition d'écoute difficile [Gra+13].

Une rééducation articulatoire basée sur la visualisation en temps-réel des mouvements articulatoires sollicite le couplage entre systèmes sensoriel et moteur en apportant au locuteur une nouvelle source (visuelle) d'informations, et pourrait renforcer, voir corriger, des relations sensorimotrices perturbées par la pathologie (par exemple après une glossectomie, qui change les propriétés biomécaniques de la langue et implique donc un ré-apprentissage de la relation entre la commande motrice et le but articulatoire ou acoustique [Ach14]).

2.1.2 Contrôle moteur de la production de la parole

Les mécanismes de simulation motrice joueraient également un rôle important dans le contrôle en ligne de la production de la parole. En effet, les principaux modèles de contrôle moteur du mouvement, et donc de la parole, qui est un ensemble de gestes articulatoires, mettent en jeu différentes boucles de régulation sensorimotrice qui permettent la bonne exécution des mouvements et leur réajustement, notamment en cas de perturbation externe [2]. Ces boucles exploitent le concept de "modèle interne", introduit dans le cadre du contrôle moteur chez

[1] un bon exemple est donné par les expériences dites de *liptube* de Savariaux et coll., qui montrent qu'un sujet réorganise la forme de son conduit vocal pour produire un /u/ lorsqu'un tube rigide placé dans la bouche empêche l'arrondissement des lèvres (mouvement caractéristique de l'articulation de cette voyelle) [SPO95 ; Sav+99]

[2] voir [Per12] pour une revue complète sur le contrôle moteur du mouvement en général et sur celui de la parole en particulier.

l'humain par Kawato et coll. [KFS87]. Les modèles internes encodent dans le cerveau les relations entre but moteur (par exemple la cible acoustique à atteindre dans le cas de la parole), commandes motrices (activités des muscles concernés) et propriétés du mouvements (position, vitesse, etc.). On distingue classiquement deux type de modèles : le modèle interne direct (*forward internal model*) et le modèle interne inverse (*inverse internal model*). Le modèle inverse prédit l'ensemble des commandes motrices permettant d'atteindre un but moteur (par exemple la cible acoustique). Le modèle interne direct est en charge de la simulation motrice, c'est-à-dire la prédiction inconsciente des conséquences somatosensorielles d'une exécution motrice, à partir d'une copie des commandes motrices (prédites par le modèle inverse) appelée "copie d'efférence". Le modèle direct est construit au cours du développement et de l'apprentissage d'un nouveau geste en comparant le but moteur moteur souhaité avec le but moteur réalisé, estimé par le système somatosensoriel (système auditif et proprioceptif). Une fois le modèle interne directe construit, et en l'absence de perturbation externe trop importante, il est possible de contrôler un geste déjà acquis uniquement à partir des simulations internes, sans s'appuyer sur le retour du système somatosensoriel.

L'intérêt du modèle direct est double. La simulation interne étant beaucoup plus rapide que l'intégration du retour somatosensoriel, il permet d'une part, l'exécution rapide de gestes déjà maîtrisés (en l'absence de perturbation trop importante). Il permet d'autre part, au cours de l'apprentissage, de simuler la conséquence d'un nombre important de gestes, sans avoir à les exécuter réellement, afin d'affiner le modèle interne inverse (voir par exemple Figure 4 de [Per12]). Une fois l'acquisition d'un nouveau geste et de son modèle interne (directe et inverse) effectuée, le contrôle de ce geste s'effectue en exploitant les modèles internes en cas de perturbation peu importante. En cas d'inadéquation persistente entre les retours somatosensoriels prédit et vécu, ce dernier peut être recruté pour corriger les commandes motrices et mettre à jour les modèles internes. Ces différentes boucles sensorimotrices impliquées dans le contrôle de la parole sont illustrées par la figure 2.1.

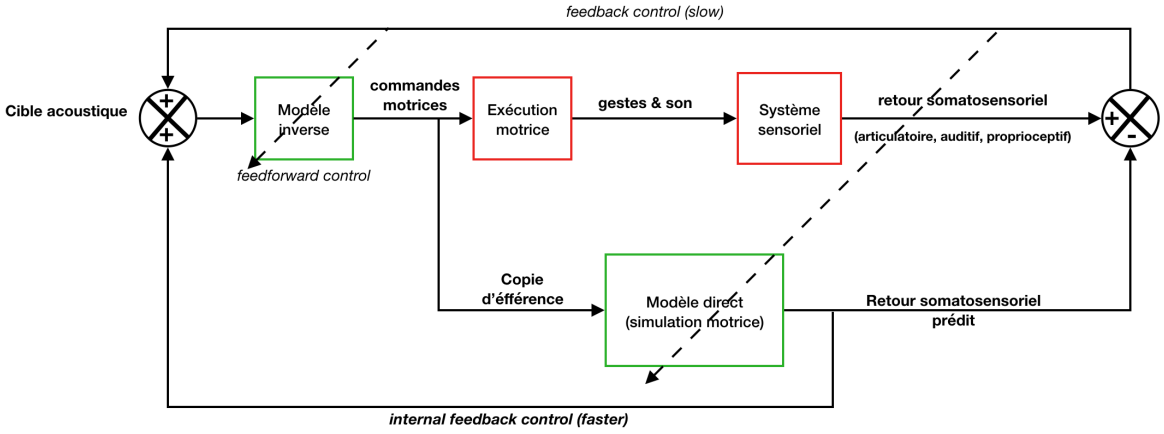


FIGURE 2.1 – Schéma simplifié du contrôle moteur de la parole. Les flèches obliques en pointillés signifient un processus de mise à jour des modèles internes. Adapté de [Per12].

.

Dans le cadre de l'étude Revison (qui sera décrite à la section 2.3.4), nous nous sommes in-

téressés aux troubles de l'articulation apparaissant après une ablation d'une partie de la langue dans le cadre du traitement du cancer (glossectomie). Comme le suggère [Ach+14 ; Ach14], l'apparition de ces troubles pourraient s'expliquer en partie par une inadéquation entre les modèles internes et le système moteur après la chirurgie. En effet, du fait d'une modifications des propriétés biomécaniques de la langue, les commandes motrices prédites à partir de la cible acoustique par le modèle inverse ne sont plus adaptées, tout comme les conséquences somatosensorielles simulées par le modèle direct. Une recalibration de ces modèles est nécessaire. C'est un des objectifs de la rééducation orthophonique. Tout comme pour l'acquisition d'un nouveau geste, l'objectif est d'amener le patient à ajuster ses commandes motrices afin d'atteindre la cible acoustique désirée. Les modèles internes n'étant plus à jour, le patient s'appuie donc a priori principalement sur son retour somatosensoriel (vécu), et doit comprendre, avec l'aide de l'orthophoniste, comment ajuster sa commande motrice pour minimiser l'erreur entre ce retour et la cible désirée. Dans un protocole de rééducation standard, on peut faire l'hypothèse qu'il s'appuie d'une part sur les consignes verbales de l'orthophoniste lui expliquant la cible acoustique en terme de cible articulatoire ("Pour faire le son [t], il faut presser le bout de la langue contre le palais, derrière les dents, puis relâcher"), et d'autre part, sur ses retours auditif, proprioceptif et kinesthésique ("en sentant le bout de sa langue contre le palais").

L'objectif d'une rééducation par illustration visuelle vise principalement à compléter les consignes de l'orthophoniste afin d'amener le patient à visualiser et comprendre la relation entre configurations articulatoires d'une part, et résultat acoustique d'autre part. Le paradigme de retour visuel vise, quant à lui, à fournir une information supplémentaire au système sensoriel, afin d'enrichir le retour somatosensoriel. En combinant illustration et retour visuel, on peut espérer fournir au patient à la fois la cible articulatoire à atteindre [3], et l'écart entre son propre geste et la cible, lui permettant d'ajuster encore plus finement et plus rapidement ses commandes motrices, puis recalibrer ses modèles internes. En s'inspirant du modèle de contrôle moteur de la parole décrit précédemment, nous illustrons ces mécanismes d'intégration des informations visuelles articulatoires à la figure 2.2.

2.1.3 Lecture linguale

Le modèle présenté à la figure 2.2 suppose cependant que le système sensoriel du patient est capable d'interpréter une information visuelle concernant un organe comme la langue, dont il n'a que très peu conscience de la forme et de la position dans le conduit vocal, et de la mettre en relation avec ses retours auditif, kinesthésique et proprioceptif. Cette hypothèse est reliée à la question de l'intégration, par notre cerveau, des différentes modalités de la parole. Cette question a largement été abordée en étudiant comment la visualisation des articulateurs externes, comme les lèvres et la mâchoire, à partir d'un visage parlant, interagissait avec la perception auditive. Un bon exemple de cette interaction et de cette influence mutuelle est le célèbre effet McGurk [MM76] qui apparait lors de la perception d'un stimulus audiovisuel synthétique construit à partir de stimuli auditifs et visuels non congruents (entraînant le

[3] Cette dernière étant présentée dans un autre espace articulatoire que celui du patient, une étape de normalisation, visant à adapter cette cible à la géométrie de son propre espace articulatoire, est a priori nécessaire.

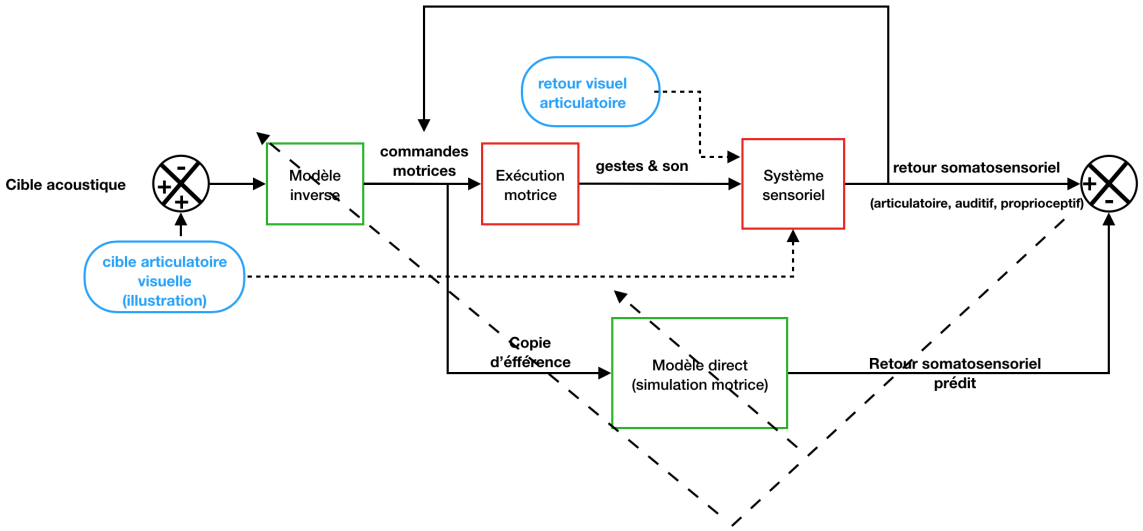


FIGURE 2.2 – Modèle théorique d'intégration des informations articulatoires fournies par les paradigmes d'illustration et de retour visuel, pour la réhabilitation d'un trouble de l'articulation impliquant un ré-apprentissage des modèles internes.

.

décodage d'un phonème tiers). Un autre exemple est celui du rôle plus important de la modalité visuelle lorsque que nous sommes en situation de communication perturbée, comme lorsque nous communiquons dans le bruit [SP54 ; BMK94 ; Oun+07]. D'autres travaux ont étudié notre capacité à décoder une information visuelle sur la langue. Dans [Bad+10], des sujets non-experts, plongés dans le bruit, essayent de décoder des voyelles et de consonnes en visualisant les mouvements de la langue d'une tête parlante artificielle [4]. Au travers d'une première étude basée sur l'IRM fonctionnel [Tre+17], puis d'une seconde basée sur l'EEG [Tre+18], nous avons recherché les corrélats neuronaux de cette capacité de "lecture linguale" [5]. Les résultats montrent d'une part, un large recouvrement des zones du cortex moteur activées lors de la visualisation d'un mouvement lingual et labial - suggérant des réseaux communs de décodages pour ces deux articulateurs - et d'autre part, une activation plus forte des zones motrices lors de la visualisation d'un geste lingual, suggérant une simulation motrice mentale accrue pour la perception d'une image d'un articulateur habituellement non-visible. Fort de ce cadre théorique, mes travaux ont consisté d'une part à développer des systèmes d'illustration et de retour visuel de l'articulation linguale, puis à les évaluer en contexte clinique, pour l'aide à la rééducation d'un trouble de l'articulation.

[4] Cette tête parlante a été utilisée dans plusieurs de mes travaux et est décrite plus en détail à la section 2.2.1.1

[5] Cette recherche a été principalement conduite par Marc Sato

2.2 Systèmes d'illustration et de retour visuel articulatoire

2.2.1 Contexte et état de l'art

2.2.1.1 Illustration visuelle articulatoire

Comme mentionné précédemment, l'illustration visuelle articulatoire vise à rendre explicite la configuration articulatoire associée à la cible acoustique à atteindre. Il n'est pas rare que les orthophonistes fassent appel à différents outils pour apprendre au patient les bases de la phonétique articulatoire. Il peut s'agir de dessins du conduit vocal, voire même de petites poupées que l'on peut animer avec les doigts pour illustrer différentes configurations articulatoires. Des logiciels ont également été développés dans ce but. On citera par exemple le logiciel *Diadolab* [MS12] et les sites Internet "L'outil vocal en action" de Canault [6] qui proposent des animations stylisées des mouvements articulatoires pour une visualisation intuitive. D'autres logiciels exploitent des enregistrements échographiques, vidéo ou IRM comme le logiciel *Ultraspeech-player* [Hue13] que j'ai developpé en 2013 et qui sera décrit à la section 2.2.2.1, ainsi que le site Internet *SeeingSpeech*[7] développé par l'université QMU en Écosse.

Une autre approche possible est d'utiliser une tête parlante dite "articulatoire", car permettant la visualisation en 3D des articulateurs normalement cachés comme la langue et le voile du palais. Construites à partir de données articulatoires acquises sur des locuteurs de référence, ces têtes parlantes permettent une visualisation complète et réaliste de la sphère orofaciale. De nombreux modèles ont été proposés et évalués dans différents contextes. Chen et coll. propose une revue de la littérature sur ces techniques et sur leur utilisation en orthophonie [Che+16]. Les têtes parlantes articulatoires proposées par Massaro et coll. [ML04], Bälter et coll. [Bäl+05], Fagel et coll. [FM08] et Badin et coll. [Bad+08] sont présentées à la figure 2.3.

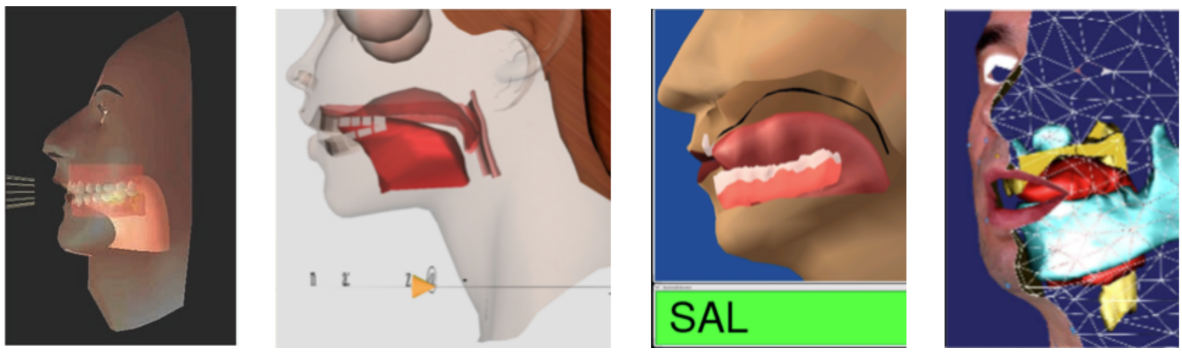


FIGURE 2.3 – Têtes parlantes articulatoires. De gauche à droite : [ML04] [FM08], [Bäl+05], [Bad+08]

.

Dans [Bäl+05], une tête parlante construite à partir de données IRM et EMA est animée

[6] http://anatomie3d.univ-lyon1.fr/webapp/website/website.html?id=3346735&pageId=223201
[7] https://www.seeingspeech.ac.uk/

"manuellement" par l'expérimentateur qui écoute la parole produite par le patient (un enfant présentant un trouble de l'articulation) et sélectionne l'animation pré-enregistrée qui lui semble être la plus en adéquation avec les mouvements produits par le patient. Cette procédure d'animation manuelle est également utilisée dans [FM08], pour le traitement de troubles articulatoires chez l'enfant et plus récemment dans [Che+19] pour le traitement des troubles articulatoires liés aux syndromes autistiques.

2.2.1.2 Retour visuel articulatoire

Le second paradigme étudié dans mes travaux sur la rééducation articulatoire assistée est celui du retour visuel ou *biofeedback*. De façon générale, ce paradigme vise à faire prendre conscience à un patient une fonction biologique ou physiologique. Le *biofeedback* est utilisé dans de nombreux domaines, comme par exemple la rééducation musculaire et celle des troubles neurologiques (voir [GPC13] pour une revue de littérature sur le *biofeedback*). Dans le domaine de la rééducation des troubles articulatoires, le *biofeedback* consiste à faire visualiser au patient les mouvements de sa propre langue, en temps réel. L'objectif est d'améliorer la conscience articulatoire du patient pour l'aider à corriger son trouble de prononciation.

La littérature récente sur le retour visuel articulatoire mentionne deux techniques principales : l'électropalatographie (EPG) et l'échographie. L'EPG consiste à capturer et à montrer au patient les points de contact de sa langue avec son palais, à l'aide d'un palais artificiel muni d'électrodes. Cette technique a notamment été utilisée pour la rééducation des troubles articulatoires chez les enfants trisomiques [Woo+09] et chez les enfants atteints de fentes palatines [Gib+01]. Cependant, l'EPG ne permet pas de capturer la position de la langue pour les phonèmes n'impliquant pas de contact avec le palais (comme les voyelles par exemple).

L'autre technique très étudiée ces dix dernières années est l'échographie, qui comme nous l'avons vu dans la section 1.1 permet de suivre en temps-réel les mouvements de la langue de façon non-invasive et inoffensive (et à l'aide d'un équipement aujourd'hui accessible pour moins de 3k€). Plusieurs études se sont penchées sur la rééducation des troubles phonétiques ou phonologiques chez l'enfant comme par exemple [PBL13] qui présente 6 études de cas d'enfants présentant différents troubles liés à une apraxie de parole, [CSW15] qui rapporte les résultats de 7 études de cas d'enfants présentant différents troubles persistants, [RSC15] qui présente 2 études de cas d'enfants présentant une fente palatine. La rééducation du /ɹ/ anglais, qui peut se réaliser de deux manières (soit la pointe de la langue est relevée et rétrofléchie, soit elle est vers le bas et rétractée) et qui est parfois remplacé chez l'enfant par le phonème /w/ (voire également une fricative ou une occlusive) fait également l'objet de plusieurs études. On citera par exemple [Mod+08] (une étude de cas), [Adl+07], (deux études de cas), et [Ber+08] (étude de groupe sur 13 enfants). Le retour visuel articulatoire par échographie a également été testé pour l'apprentissage de la parole chez les personnes sourdes [Ber+03 ; BBG07] et la rééducation d'un patient aphasique post-AVC [Ach+16] (étude de cas). Enfin, certaines études se sont intéressées à la rééducation des troubles apparaissant suite à une chirurgie carcinologique linguale. Blyth et coll. présentent en 2016 deux études de cas sur deux sujets glossectomisés adultes, ayant suivi un protocole de rééducation par retour visuel articulatoire

basé sur l'échographie linguale (à raison de trois séances de 30 minutes par semaine, durant quatre semaines) [Bly+16]. Les résultats de cette étude ont montré une amélioration significative de la production des consonnes travaillées, une généralisation de cette amélioration à des consonnes non travaillées (pour l'un des deux participants), ainsi qu'une persistance des effets après le traitement. La rééducation des patients glossectomisés fait l'objet de l'étude Revison décrite à la section 2.3.4.

2.2.2 Travaux réalisés

2.2.2.1 Illustration visuelle articulatoire

Ma contribution principale sur le développement des systèmes d'illustration visuelle articulatoire est le logiciel *Ultraspeech-player* (son utilisation dans différents contextes sera décrit à la section 2.3). Ce logiciel permet de visualiser simultanément, pour différents phonèmes ou logatomes d'une langue, les mouvements de la langue et des lèvres, acquis sur plusieurs locuteurs de référence par échographie et vidéo à l'aide du logiciel *Ultraspeech* (décrit à la section 1.1.2.1), tout en écoutant le signal sonore de la parole vocalisée. Une capture d'écran du logiciel est présentée à la figure 2.4.

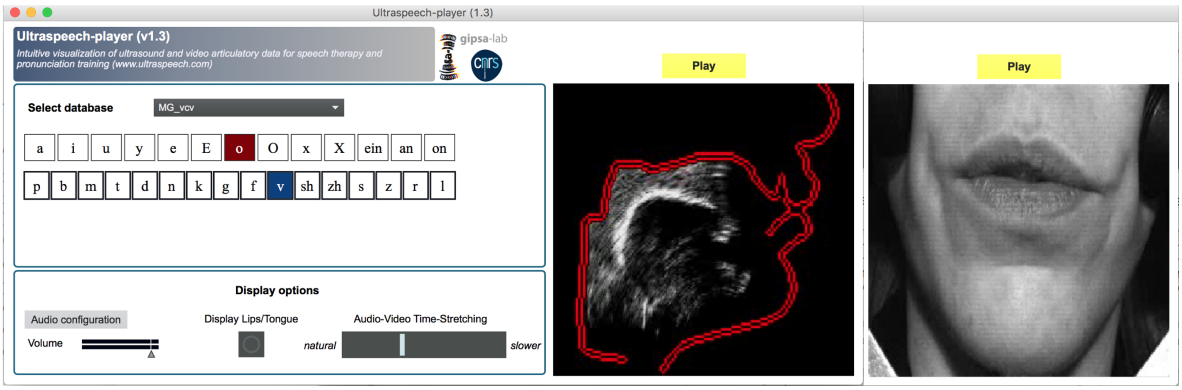


FIGURE 2.4 – Capture d'écran du logiciel *Ultraspeech-player* [Hue13] développé pour l'aide à la rééducation articulatoire par illustration visuelle (à gauche, visualisation des mouvements de la langue, à droite, des mouvements des lèvres). Le logiciel et ses différentes bases de données sont téléchargeables gratuitement sur `www.ultraspeech.com`.

Ce logiciel permet notamment au patient de ralentir le son et la vidéo, ce qui permet une meilleure visualisation des enchaînements articulatoires et de la co-articulation [8].

[8] Un *vocoder* de type "Harmonique+Bruit" (HNM) [Sty01] est utilisé pour le *time-stretching* du son, la vitesse de lecture des séquences d'images échographiques et vidéo est ajustée en fonction.

2.2.2.2 Retour visuel par conversion acoustico-articulatoire

Je développe depuis 2010 des systèmes de retour visuel articulatoire basés sur la tête parlante articulatoire développée au GIPSA-lab par Badin et coll. [Bad+08]. Cette dernière intègre un modèle géométrique 3D de la langue, des lèvres, de la mâchoire et du voile du palais . Ces modèles sont construits à partir de données IRM 3D statiques, de données vidéo stéréoscopiques, et d'un moulage du palais et des dents d'un locuteur de référence. Badin et coll. décrivent dans [Bad+08] un technique pour animer ces différents modèles 3D à partir de données EMA enregistrées sur ce même locuteur de référence. Dans la suite de ce manuscrit, les coordonnées 2D de 6 bobines EMA (3 sur la langue, 2 sur les lèvres et 1 sur la mâchoire [9]) sont considérés comme les paramètres de contrôle de cette tête parlante (soit un vecteur de dimension 12). Les étapes de construction de cette tête parlante, ainsi que le positionnement des bobines EMA sur le locuteur de référence, sont résumées sur la figure 2.5.

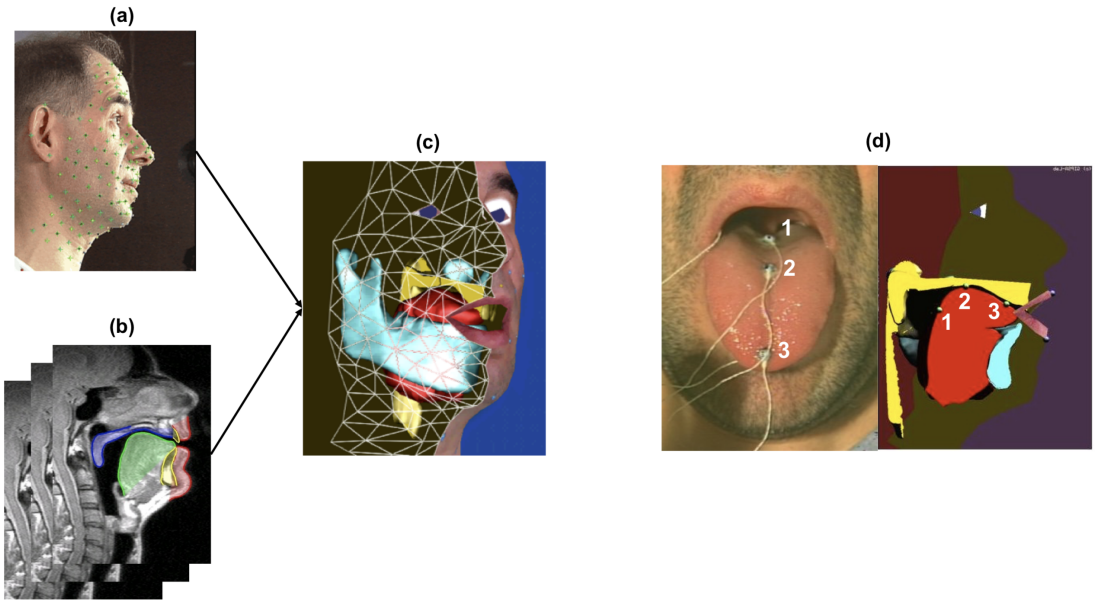


FIGURE 2.5 – Tête parlante articulatoire de Badin et coll. [Bad+08] construite à partir de données vidéo stéréoscopiques (a) de données IRM 3D statiques (b), acquises sur un locuteur de référence (c). Vues 3D des modèles de langue, lèvres, mâchoire et vélum construits par analyse statistique de la forme des articulateurs lors de l'articulation de différents phonèmes, après segmentation manuelle des images IRM. (d) Bobines EMA collées sur la langue, les lèvres et la mâchoire du locuteur de référence (et position de ces bobines sur le modèle 3D de langue) dont les coordonnées peuvent être utilisées pour animer la tête parlante. Extrait de [Fab+17].

L'objectif est ici d'animer cette tête parlante automatiquement à partir du signal audio de parole de l'utilisateur, que nous appellerons par la suite le "locuteur source", et ce, en temps-réel. Il s'agit donc d'estimer les mouvements articulatoires de ce locuteur source, et

[9] Le modèle de voile du palais n'est pas considéré ici.

de les représenter dans l'espace articulatoire de la tête parlante, qui est construite à partir de l'anatomie d'un locuteur que nous appellerons par la suite le locuteur de référence. De façon importante, notons que nous ne cherchons pas à adapter la tête parlante à l'anatomie du locuteur source. Il s'agit plutôt de montrer à ce dernier, quel mouvement le locuteur de référence ferait, s'il devait prononcer le même son que lui. Cette approche ne garantit donc pas une estimation des véritables mouvements articulatoires du locuteur source et un retour (visuel) fidèle sur sa propre articulation. Cependant, cette approche présente un double intérêt. D'une part, elle ne nécessite pas d'équipement couteux comme un échographe (un simple microphone et un ordinateur suffit), et d'autre part, elle permet au patient de visualiser, dans un même espace, son geste (incorrect) et la cible (correcte) articulatoire à atteindre. Cette cible peut être pré-enregistrée en laboratoire par le locuteur de référence ou bien estimée à partir de la voix de l'orthophoniste. En d'autres termes, il s'agit donc ici de disposer d'un système permettant de faire de l'illustration et du retour visuel avec le même outil de visualisation. Un schéma général du système est proposé à la figure 2.6.

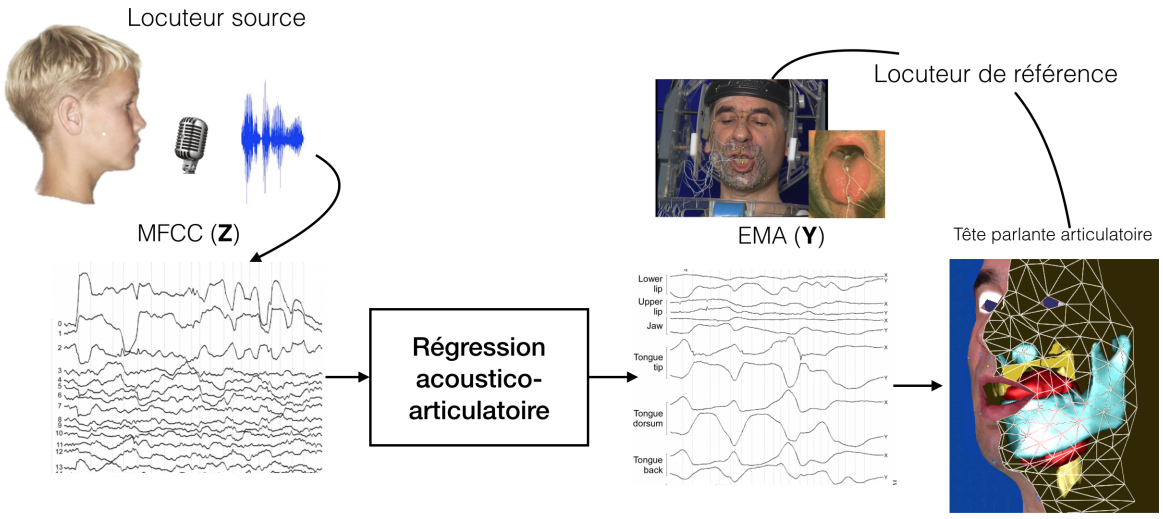


FIGURE 2.6 – Système de retour visuel basé sur une tête parlante articulatoire animée à partir de la voix de l'utilisateur.

Comme mentionné ci-avant, la tête parlante articulatoire peut être animée *via* une séquence de vecteurs contenant les cordonnées 2D de 6 bobines EMA, collées sur les articulateurs du locuteur de référence. Il s'agit donc d'estimer ces coordonnées à partir du contenu spectral de la voix du locuteur source. Ben-Youssef et coll. ont proposé en 2011 d'effectuer cette régression en 2 étapes disjointes : (i) une décodage acoustico-phonétique à l'aide d'un ensemble de modèles acoustiques de type HMM-GMM, entraînés sur la voix du locuteur de référence puis adaptés à la voix du locuteur source à l'aide de la technique MLLR [LW95], (ii) la génération des trajectoires articulatoires par HMM-GMM à partir de la séquence phonétique décodée (à l'aide de l'algorithme MLPG) [Ben+11]. Nous avons étendu cette approche en 2012 à l'aide de la technique de régression par HMM-GMM [Hue+12], déjà décrite à la section 1.1.2.3. Ici, la génération des trajectoires articulatoires prend en compte à la fois la séquence phonétique décodée, mais également la production acoustique. Cette approche permet la génération, pour

un même phonème décodé, de différentes trajectoires articulatoires. Cependant, l'étape de décodage phonétique impliquée dans la régression par HMM-GMM empêche d'effectuer la conversion en temps-réel. Pour lever cette limitation, je me suis donc penché sur différentes approches ne faisant pas appel de façon explicite au niveau phonétique, et s'appuyant sur les modèles de mélange Gaussiens. Les différentes techniques proposées sont résumées dans les paragraphes suivants.

Approche par GMR

On note $\mathbf{Z}$ le vecteur aléatoire associé à l'espace acoustique du locuteur source (on utilisera typiquement une paramétrisation de l'enveloppe spectrale par décomposition mel-cepstrale, aboutissant à un vecteur d'observation acoustique de dimension 13) et $\mathbf{z_n}$ une réalisation associée (vecteur colonne). On note $\mathbf{X}$ un vecteur aléatoire associé à l'espace acoustique du locuteur référence (paramétré également par décomposition mel-cepstrale) et $\mathbf{x_n}$ une réalisation associée. On note $\mathbf{Y}$ un vecteur aléatoire associé à l'espace articulatoire du locuteur référence (vecteur de coordonnées EMA de dimension 12 dans le cas de la tête parlante articulatoire) et $\mathbf{y_n}$ une réalisation associée. On note $\mathcal{D}_\mathbf{z} = \{\mathbf{z}_n\}_{n=1}^{N_0}$, l'ensemble des N_0 observations acoustiques disponibles (locuteur source). On suppose que $\mathcal{D}_\mathbf{z}$ peut être mis en regard avec un ensemble $\mathcal{D}_\mathbf{xy}$ de N_0 observations acoustico-articulatoires enregistrées chez le locuteur de référence. En pratique, on demande au locuteur source de prononcer quelques phrases également enregistrées par le locuteur de référence, puis pour chaque phrase, on aligne les signaux acoustiques des deux locuteurs à l'aide de l'algorithme *Dynamic Time Warping* (DTW). Un schéma des différentes variables utilisées est fourni à la figure 2.7.

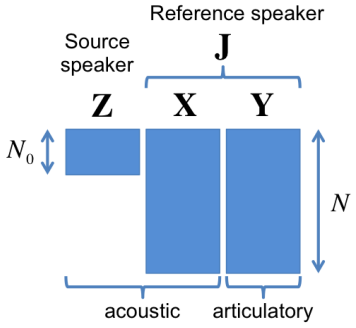


FIGURE 2.7 – Variables impliquées dans l'approche C-GMR. Extrait de [Hue+15].

Une première approche pour piloter la tête parlante à partir de la voix du locuteur source est de modéliser la densité de probabilité jointe sur $(\mathbf{Z}, \mathbf{Y})$ par un GMM tel que décrit par l'équation 1.1, puis de déduire le GMR $\mathbf{Z}$-vers-$\mathbf{Y}$ tel que décrit par l'équation 1.7 (qui permet une conversion $\mathbf{z}_n$-vers-$\mathbf{y}_n$ en temps-réel). Cette approche est appelée conversion directe ou D-GMR et est illustrée à la figure 2.8a .

Approche par adaptation d'un GMM

Une seconde approche possible serait d'exploiter les techniques d'adaptation de type MLLR

[LW95] ou MAP [GC94] traditionnellement utilisées en reconnaissance automatique de la parole pour adapter au locuteur source ($\mathbf{Z}$), un GMM sur ($\mathbf{X}$, $\mathbf{Y}$) (locuteur de référence) dont les paramètres sont estimés à partir d'une large base de données présentant une couverture phonétique la plus optimale possible. Cependant, ne disposant que des productions acoustiques du locuteur source et non des trajectoires articulatoires associées (leur acquisition en contexte clinique est difficilement envisageable), seuls les paramètres du GMM sur ($\mathbf{X}$, $\mathbf{Y}$) relatifs à $\mathbf{X}$ peuvent être adaptés. Pour chaque composante m du mélange, il s'agit de $\mu_{\mathbf{X},m}$ et $\mathbf{\Sigma}_{\mathbf{XX},m}$ (voir équation 1.2), c'est-à-dire une partie seulement des vecteurs de moyenne et des matrices de covariance. Nos expériences ont montré que cette adaptation partielle est très problématique pour la régression (voir par exemple les figures 4 et 5 de [Hue+15]), probablement car elle fait apparaitre une incohérence entre les paramètres adaptés et ceux qui ne le sont pas, comme par exemple les sous-matrices de covariance croisée $\mathbf{\Sigma}_{\mathbf{XY},m}$.

Cascaded Gaussian Mixture Regression (C-GMR)

En 2013, j'ai proposé la première version d'une technique intitulée "*Cascaded Gaussian Mixture Regression*" ou C-GMR. Cette dernière exploite un modèle GMM sur ($\mathbf{X}$, $\mathbf{Y}$) (entrainé à partir d'une large base de données) et du GMR ($\mathbf{X}$-vers-$\mathbf{Y}$) associé. Dans une première version de cette technique, la régression ($\mathbf{Z}$-vers-$\mathbf{Y}$) s'effectue en "cascadant" deux sous-regressions : (i) une régression $\mathbf{Z}$-vers-$\mathbf{X}$ de l'espace acoustique du locuteur source vers l'espace acoustique du locuteur de référence [10] à l'aide d'un GMM sur ($\mathbf{Z}$,$\mathbf{X}$) dont les paramètres sont estimés sur $\mathcal{D}_{\mathbf{zx}} = \{\mathbf{z}_n, \mathbf{x}_n\}_{n=1}^{N_0}$, et (ii) une étape d'inversion acoustico-articulatoire [11] $\mathbf{X}$-vers-$\mathbf{Y}$ à l'aide du GMM sur ($\mathbf{X}$,$\mathbf{Y}$) dont les paramètres sont estimés sur un plus large ensemble de données $\mathcal{D}_{\mathbf{xy}} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$ avec classiquement $N_0 \ll N$. Formellement, l'inférence s'effectue ensuite en calculant successivement $\hat{\mathbf{x}} = \mathrm{E}[\mathbf{X}|\mathbf{z}]$ et $\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\hat{\mathbf{x}}]$, à l'aide de l'équation 1.7. Les deux GMMs étant ici appris de façon disjointes [12], cette approche est appelée *Split-C-GMR* ou SC-GMR [Hue+13] et est illustrée à la figure 2.8b.

En 2015, nous avons proposé une variante de l'approche C-GMR, intitulée *Integrated C-GMR* [Hue+15][13], dans laquelle les deux étapes sont combinées dans un unique modèle probabiliste dont une représentation graphique est fournie à la figure 2.8c. La densité de probabilité

[10] Cette régression peut être vue comme une étape de "conversion de voix", qui est une tâche à part entière ayant donnée lieu à de nombreux travaux. On citera par exemple les approches par GMM [SCM98 ; KM98 ; TBT08], les approches par réseaux de neurones profonds [Lin+14], par réseaux de neurones récurrents [Min+16], par auto-encodeurs variationnels (VAE) [Hsu+16], par réseaux antagonistes génératifs (GAN) [Hsu+17], par réseaux de neurones auto-regressifs de type *Wavenet* [Kob+17]. Une revue de la littérature sur ce domaine a été proposée en 2017 par Mohammadi et Kain [MK17].

[11] L'inversion acoustico-articulatoire est également un problème classique du traitement automatique de la parole. Comme indiqué à la section 2.1.1, il s'agit d'un problème mal-posé car un même son peut être associé à plusieurs configurations articulatoires, et qui nécessite donc d'être régularisé en s'appuyant par exemple sur des informations contextuelles et/ou sur des informations linguistiques *a priori*. Depuis les travaux pionniers de Atal et coll. [Ata+78], ce problème a donné lieu à de très nombreux travaux, voir par exemple [OL05] pour une approche par *codebook*, [Ric02 ; Ric06 ; Uri+12 ; LLD16] pour des approches par réseaux de neurones artificiels (ANN/DNN/RNN-LSTM), [TBT08 ; AE11] pour une approche par GMM, [LR08 ; You+09 ; ZNT11] pour une approche par HMM-GMM.

[12] Notons que les deux GMRs d'un SC-GMR peuvent avoir un nombre différent de composantes.

[13] Ce travail a fait l'objet d'une collaboration avec l'équipe Perception de INRIA.

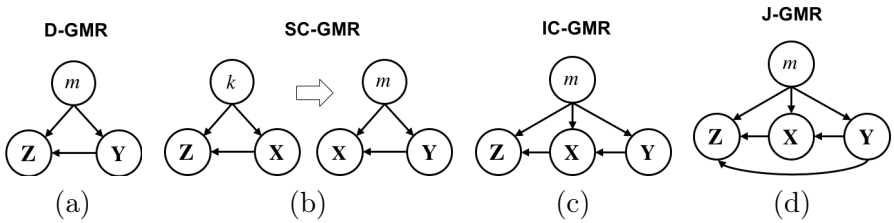


FIGURE 2.8 – (a) Approche D-GMR (b) Approche SC-GMR [Hue+13] - (c) Modèle IC-GMR [Hue+15] - (d) Modèle J-GMR [GHA17b]

conjointe sur ($\mathbf{Z}$, $\mathbf{X}$,$\mathbf{Y}$) est ici modélisée par :

$$p(\mathbf{o}) = \sum_{m=1}^{M} p(m)p(\mathbf{y}|m)p(\mathbf{x}|\mathbf{y},m)p(\mathbf{z}|\mathbf{x},m), \tag{2.1}$$

avec $\mathbf{O} = [\mathbf{X}^\top, \mathbf{Y}^\top, \mathbf{Z}^\top]^\top$. Les différentes densités de probabilité conditionnelle ont une forme linéaire gaussienne :

$$p(m) = \pi_m, \tag{2.2}$$
$$p(\mathbf{y}|m, \mathbf{\Theta}_{\mathbf{Y},m}) = \mathcal{N}\left(\mathbf{y}|\mathbf{e}_m, \mathbf{R}_m\right), \tag{2.3}$$
$$p(\mathbf{x}|\mathbf{y}, m, \mathbf{\Theta}_{\mathbf{X}|\mathbf{Y},m}) = \mathcal{N}\left(\mathbf{x}|\mathbf{A}_m\mathbf{y} + \mathbf{b}_m, \mathbf{U}_m\right), \tag{2.4}$$
$$p(\mathbf{z}|\mathbf{x}, m, \mathbf{\Theta}_{\mathbf{Z}|\mathbf{X},m}) = \mathcal{N}\left(\mathbf{z}|\mathbf{C}_m\mathbf{x} + \mathbf{d}_m, \mathbf{V}_m\right). \tag{2.5}$$

avec, pour chaque composante du mélange, la probabilité a priori π_m, $\mathbf{e}_m$ et $\mathbf{R}_m$ étant respectivement le vecteur de moyenne et la matrice de covariance de la densité de probabilité (Gaussienne) marginale sur $\mathbf{Y}$, $\mathbf{A}_m$, $\mathbf{b}_m$ et $\mathbf{U}_m$ étant respectivement les matrices de passage de $\mathbf{x}$-vers-$\mathbf{y}$, le vecteur de biais et la matrice de covariance de la densité de probabilité conditionnelle sur $(\mathbf{X}, \mathbf{Y})$, et *idem* pour $\mathbf{C}_m$, $\mathbf{d}_m$ et $\mathbf{V}_m$ par rapport à $(\mathbf{Z}, \mathbf{X})$. Ces paramètres s'expriment en fonction des vecteurs de moyenne et des matrice de covariance des GMMs conjoints sur les différents variables, et se déduisent directement de l'équation 1.3 tel que :

$$\mathbf{A}_m = \mathbf{\Sigma}_{\mathbf{XY},m}\mathbf{\Sigma}^{-1}_{\mathbf{YY},m} \tag{2.6}$$
$$\mathbf{b}_m = \mu_{\mathbf{X},m} - \mathbf{A}_m\mu_{\mathbf{Y},m} \tag{2.7}$$
$$\mathbf{U}_m = \mathbf{\Sigma}_{\mathbf{XX}|\mathbf{y},m} \tag{2.8}$$
$$\mathbf{C}_m = \mathbf{\Sigma}_{\mathbf{ZX},m}\mathbf{\Sigma}^{-1}_{\mathbf{XX},m} \tag{2.9}$$
$$\mathbf{d}_m = \mu_{\mathbf{Z},m} - \mathbf{C}_m\mu_{\mathbf{X},m} \tag{2.10}$$
$$\mathbf{V}_m = \mathbf{\Sigma}_{\mathbf{ZZ}|\mathbf{x},m} \tag{2.11}$$

De façon similaire à un régresseur GMR, l'estimateur $\hat{\mathbf{y}}$ de $\mathbf{y}$ sachant $\mathbf{z}$ minimisant le critère MSE est défini comme l'espérance conditionnelle $\mathrm{E}[\mathbf{Y}|\mathbf{z}]$ [14] :

$$\hat{\mathbf{y}} = \mathrm{E}[\mathbf{Y}|\mathbf{z}] = \int_{\mathbb{R}^{D_Y}} \mathbf{y}p(\mathbf{y}|\mathbf{z})\mathrm{d}\mathbf{y} \tag{2.12}$$

[14] Nous ne détaillons ici que les étapes clés du calcul, le lecteur est invité à consulter [Hue+15] pour une description complète.

avec

$$p(\mathbf{y}|\mathbf{z}) = \int_{\mathbb{R}^{D_X}} \sum_{m=1}^{M} p(\mathbf{x}, \mathbf{y}, m|\mathbf{z}) \mathrm{d}\mathbf{x}. \tag{2.13}$$

Dans le cas du modèle IC-GMR, nous avons :

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, m|\mathbf{z}) &= p(m|\mathbf{z}) p(\mathbf{y}|\mathbf{x}, \mathbf{z}, m) p(\mathbf{x}|\mathbf{z}, m) \\ &= p(m|\mathbf{z}) p(\mathbf{y}|\mathbf{x}, m) p(\mathbf{x}|\mathbf{z}, m), \end{aligned} \tag{2.14}$$

et puisque $\mathbf{Y}$ est indépendant de $\mathbf{Z}$ conditionnellement à $\mathbf{X}$ et m, on obtient :

$$p(\mathbf{y}|\mathbf{z}) = \sum_{m=1}^{M} p(m|\mathbf{z}) \int_{\mathbb{R}^{D_X}} p(\mathbf{y}|\mathbf{x}, m) p(\mathbf{x}|\mathbf{z}, m) \mathrm{d}\mathbf{x}. \tag{2.15}$$

En combinant l'équation 2.15 avec les équations 2.12 et 2.11, on peut montrer que :

$$\hat{\mathbf{y}} = \sum_{m=1}^{M} p(m|\mathbf{z}) (\mu_{\mathbf{Y},m} + \mathbf{\Sigma}_{\mathbf{YX},m} \mathbf{\Sigma}_{\mathbf{XX},m}^{-1} \mathbf{\Sigma}_{\mathbf{XZ},m} \mathbf{\Sigma}_{\mathbf{ZZ},m}^{-1} (\mathbf{z} - \mu_{\mathbf{Z},m})) \tag{2.16}$$

Cette équation est une forme particulière d'un GMR $\mathbf{Z}$-to-$\mathbf{Y}$ avec une contrainte sur la forme de la matrice de covariance $\mathbf{\Sigma}_{\mathbf{YZ},m}$, tel que :

$$\mathbf{\Sigma}_{\mathbf{YZ},m} = \mathbf{\Sigma}_{\mathbf{YX},m} \mathbf{\Sigma}_{\mathbf{XX},m}^{-1} \mathbf{\Sigma}_{\mathbf{XZ},m} \tag{2.17}$$

En d'autres termes, on contraint la conversion de l'espace acoustique du locuteur source à l'espace articulatoire du locuteur de référence à passer d'abord par l'espace acoustique de ce dernier. Les étapes de conversion de voix et d'inversion acoustico-articulatoire sont donc bien intégrées dans une même régression.

Pour ce modèle, nous avons dérivé l'algorithme *EM* exact permettant l'estimation des paramètres optimaux à partir des données $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, c'est-à-dire $\mathcal{D}_{\mathbf{z}} \cup \mathcal{D}_{\mathbf{xy}}$. De façon importante, cet algorithme exploite la stratégie dite des "données manquantes" (*missing data*) pour palier au manque de données sur le locuteur source. Brièvement, on suppose que l'ensemble des données d'apprentissage est composé d'un sous ensemble de N_0 triplets $\mathcal{D}_{\mathbf{o}} = \{\mathbf{z}_n, \mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N_0}$ et d'un autre sous-ensemble de $N - N_0 + 1$ couples $\mathcal{D}_{\mathbf{o}} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=N_0+1}^{N}$. L'ensemble $\mathcal{D}_{\mathbf{z}} = \{\mathbf{z}_n\}_{n=N_0+1}^{N}$ est donc considéré comme "manquant". Chaque observation manquante $\mathbf{z}_n$ est alors considérée comme la réalisation d'une variable aléatoire latente qui est estimée à chaque itération i de l'étape E de l'algorithme *EM*, à l'aide de l'équation :

$$\mathbf{z}'_{nm} = \mathbf{C}_m^{(i)} \mathbf{x}_n + \mathbf{d}_m^{(i)} \tag{2.18}$$

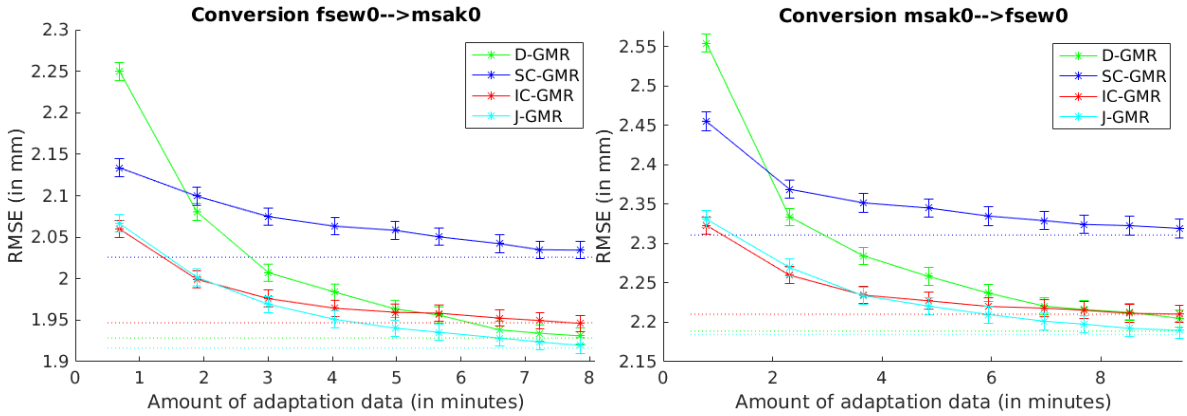
(notons que cette estimation est effectuée pour chaque composante du mélange). Une fois ces données manquantes et les responsabilités de chaque composante du mélange estimées, les paramètres des différentes densités de probabilités sont mis à jour à l'étape M, afin de maximiser la (log-) vraisemblance du modèle sur $\mathcal{D}_{\mathbf{z}} \cup \mathcal{D}_{\mathbf{xy}}$.

Contrairement à l'approche *Split C-GMR*, $\mathbf{Z}$, $\mathbf{X}$ et $\mathbf{Y}$ sont conditionnés par la même variable latente (variable indicatrice de composante m). Cette structure permet notamment de

"calquer" la structure de l'espace acoustico-articulatoire du locuteur de référence, bien estimée sur un corpus large et dense, sur celle du locuteur source. Comme nous le verrons ultérieurement, cela permet d'améliorer la capacité de généralisation du modèle à des phonèmes absents du corpus d'adaptation, ce qui est un des buts recherchés dans un contexte clinique (le patient ne maîtrise pas un phonème lors de l'enrôlement dans le système, puis l'acquiert grâce à la rééducation).

Cependant, le modèle IC-GMR n'exploite pas les éventuelles dépendances statistiques entre les variables $\mathbf{Z}$ et $\mathbf{Y}$, c'est-à-dire entre l'espace acoustique du nouveau locuteur et l'espace articulatoire du locuteur de référence. Bien entendu, le lien entre ces deux espaces ne peut qu'être indirect. Il ne s'appuie sur aucune relation physique (comme c'est le cas pour la relation acoustico-articulatoire intra-locuteur). Cependant, il peut s'appuyer sur une relative similitude de la structure de l'espace acoustico-articulatoire de chaque locuteur qui est relativement contraint par la phonétique. Aussi, nous avons cherché à quantifier l'intérêt d'exploiter les relations possibles entre $\mathbf{Z}$ et $\mathbf{Y}$. Nous avons donc proposé en 2017 une nouvelle variante de l'approche C-GMR ([GHA17a ; GHA17b]), basée sur un modèle probabiliste exploitant (si possible) un lien éventuel entre ces variables. Ce modèle est intitulé *Joint-GMR* (J-GMR) et est illustré à la figure 2.8d. Le terme "*Joint*" reflète le fait que ce modèle est équivalent à un GMM conjoint sur $(\mathbf{Z}, \mathbf{X}, \mathbf{Y})$. Aussi, on montre que l'équation permettant la régression $\mathbf{Z}$-vers-$\mathbf{Y}$ à l'aide du modèle J-GMR est identique à celle de l'approche D-GMR, c'est-à-dire un GMR standard entraîné sur un ensemble $\{\mathbf{z}_n, \mathbf{y}_n\}_{n=1}^{N_0}$ (à l'aide de l'équation 1.7). Cependant, tout comme pour le modèle IC-GMR, l'algorithme *EM* exact associé au modèle J-GMR (que nous avons également dérivé, voir le détail dans [GHA17b]) fait appel d'une part, à la stratégie de données manquantes pour faire face au manque de données d'apprentissage du locuteur source ($\mathbf{Z}$), et d'autre part, à l'ensemble des productions acoustiques ($\mathbf{X}$) et articulatoires ($\mathbf{Y}$) disponibles pour le locuteur de référence. Cela permet une estimation plus robuste des paramètres du modèle et notamment de la matrice de covariance $\mathbf{\Sigma}_{\mathbf{YZ},m}$ (qui n'est cependant plus contrainte, contrairement au modèle IC-GMR).

Ces différents modèles ont été évalués d'une part sur des données synthétiques obtenues à l'aide du synthétiseur articulatoire VLAM [Mén+07] permettant de simuler différents locuteurs en faisant changer la taille du conduit vocal, et d'autre part, sur des bases de données acoustico-articulatoires acquise *in-vivo* par EMA, en langue Française (base PB2007) et Anglaise (base MOCHA). Par soucis de concision, nous ne présentons à la figure 2.9 que les résultats obtenus sur la base de données MOCHA. Ces résultats montrent tout d'abord que l'approche D-GMR n'est performante que lorsqu'une grande quantité de données acoustiques sur le locuteur source (N_0) est disponible. Pour N_0 faible ($\leq$ 3mn, ce qui est une cible raisonnable pour une utilisation en contexte clinique), les modèles IC-GMR et J-GMR fournissent de bien meilleurs résultats. Ceci confirme l'intérêt d'exploiter l'espace acoustique du locuteur de référence lors de la conversion. Cependant, les performances de l'approche SC-GMR sont meilleures que celles du modèle D-GMR uniquement pour la plus petite valeur de N_0 et sont bien moins bonnes que celles des modèles IC-GMR et J-GMR. Ceci montre que l'étape de conversion de voix, lorsqu'elle est réalisée indépendamment de celle d'inversion acoustico-articulatoire (lors de la phase d'apprentissage comme lors de la phase de conversion), est susceptible d'agir comme un maillon faible dans la chaîne de conversion. Autrement dit, une erreur commise lors de la



(a) (b)

FIGURE 2.9 – Précision de la conversion acoustico-articulatoire entre un locuteur source (a : fsew0, b : msak0) et un locuteur de référence (a : msak0, b : fsew0), en fonction de la quantité de données acoustiques utilisée (N_0), pour les modèles D-, SC-, IC- et J-GMR (RMSE en mm, les barres d'erreurs représentent les intervalles de confiance à 95%, les lignes en pointillées illustrent pour chacun des modèles la performance obtenue lorsque toutes les données acoustiques du locuteur source disponibles sont exploitées). Extrait de [GHA17b].

régression **Z**-vers-**X** ne sera pas "rattrapée" lors de la régression **X**-vers-**Y**. Un chaînage de ces deux étapes au sein de la même régression, combiné avec l'exploitation conjointe de toutes les données acoustico-articulatoires disponibles lors de la phase d'apprentissage, tel qu'effectué par les modèles IC-GMR et J-GMR, semble donc être une approche intéressante. Enfin, les modèles IC-GMR et J-GMR donnent globalement des résultats proches, ce qui confirme la faible quantité d'information apportée par le lien **Z**-**Y** lors de l'apprentissage. Cependant, le modèle J-GMR donne des résultats légèrement meilleurs que le modèle IC-GMR dès que la quantité de données acoustiques pour le locuteur source devient plus grande ($\geq$ 7mn). Il semble donc que les contraintes qu'impose le modèle IC-GMR sur la conversion deviennent ici pénalisantes, alors qu'elles étaient bénéfiques pour de plus faibles quantités de données d'apprentissage. Aussi, le modèle J-GMR qui présente de bonnes performances quelle que soit la quantité de données acoustiques du locuteur source exploitée, semble plus flexible.

L'approche de conversion acoustico-articulatoire basée sur les modèles IC-GMR a été implémentée dans un prototype permettant l'animation en temps-réel de la tête parlante articulatoire du GIPSA-lab à partir de la voix d'un locuteur tiers. Une vidéo de ce prototype (réalisé sous l'environnement Max/MSP/Jitter) est disponible à l'adresse : `https://www.youtube.com/watch?v=Cv9ONXyoOTk`. L'ensemble du code source nécessaire à l'apprentissage et à l'utilisation des modèles IC-GMR et J-GMR est disponible sur `https://github.com/thueber/cgmr`.

2.2.2.3 Échographie linguale augmentée

L'approche par inversion acoustico-articulatoire fournit des résultats globalement satisfaisants mais échoue quasi-systématiquement pour certains phonèmes, et notamment les occlusives comme /k/ ou /t/. En effet, au moment de l'occlusion et jusqu'au relâchement le problème d'inversion est particulièrement mal posé, et les modèles existants, même ceux exploitant une information contextuelle sur les instants qui précèdent l'occlusion et suivent le relâchement, peinent à estimer correctement les mouvements articulatoires. Afin de fournir un retour fiable pour ses sons, nous avons proposé, dans le cadre de la thèse de Diandra Fabre, une nouvelle approche basée sur la capture directe du mouvement de la langue par échographie.

Nous partons du constat qu'une image échographique de la langue est parfois difficile à interpréter par un patient non-expert pour les raisons suivantes :

- L'image est d'assez mauvaise qualité en raison de la présence du bruit dit de *speckle*.
- Certaines parties du contour correspondant à la surface supérieure de la langue disparaissent lorsque l'orientation de la langue s'éloigne de la perpendiculaire au faisceau ultrasonore incident [HD09].
- Les autres structures du conduit vocal comme le palais, la paroi pharyngée et les dents ne sont pas visibles.

Aussi, nous avons cherché à "augmenter" l'image échographique tout d'abord en segmentant automatiquement le contour de la langue, puis en la remplaçant par une tête parlante articulatoire. Ces deux pistes sont détaillées dans les paragraphes suivants.

Extraction automatique du contour de la langue

Différentes approches ont été proposées dans la littérature pour la segmentation de la surface supérieure de la langue dans les images ultrasonores. Certaines sont basées sur la technique des contours actifs [LKS05 ; Xu+16], d'autres sur les modèles actifs d'apparence [RKM09]. Lossvelt et coll. proposent dans [LVB14] de s'appuyer sur un modèle biomécanique de la langue pour guider et régulariser la segmentation de son contour dans les images ultrasonores. En 2010, Fasel et Berry décrivent une technique basée sur un réseau de neurones profond, dont les performances sont présentées comme comparables à celles d'un humain effectuant la tâche de segmentation manuellement. Plus précisément, un *translational deep belief network* est entraîné à régresser l'ensemble des pixels d'une image de langue vers les pixels qui composent son contour (annoté manuellement) [FB10]. Bien que difficile à reproduire, les résultats semblent prometteurs. Mais cette performance ne s'atteint qu'au prix d'une quantité très importante de données d'apprentissage (environ 8000 images segmentées manuellement). En 2015, nous avons proposé une approche intermédiaire, basée également sur l'apprentissage supervisé d'un réseau de neurones, mais exploitant une représentation plus compacte des images échographiques que les pixels bruts [Fab+15]. Cette représentation est obtenue à

l'aide de la technique dite des *EigenTongues* [Hue+07a]. Il s'agit d'une simple adaptation de la technique dite des *EigenLips* [BK94]) qui est basée sur une analyse en composante principale d'un ensemble (idéalement phonétiquement équilibré) d'images échographiques. Cette technique permet d'encoder les principales structures anatomiques visibles dans l'image (dont la langue) dans un vecteur d'une trentaine de coefficients. Nous avons montré que l'utilisation de cette représentation compacte permet de maintenir une bonne qualité de segmentation tout en limitant la quantité de données nécessaire à son apprentissage (quelques dizaines d'images segmentées manuellement, au lieu de plusieurs milliers). Un schéma général de cette technique ainsi que deux exemples d'images segmentées sont présentés respectivement sur les figures 2.10a et 2.10b.

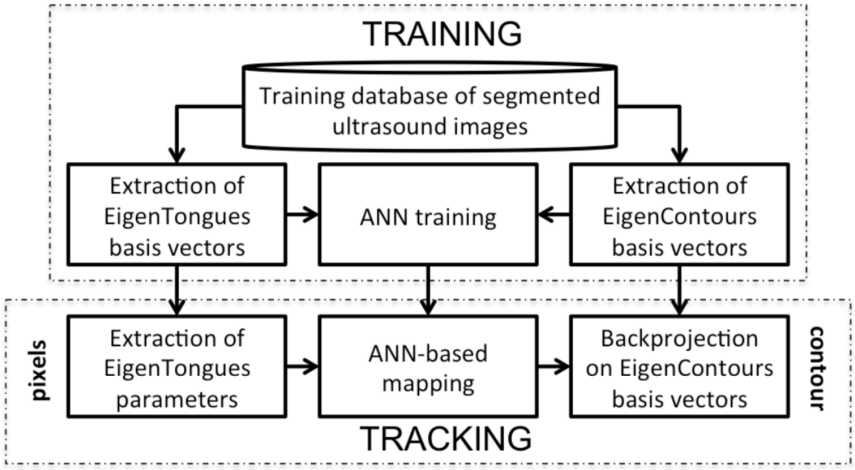
Animation du modèle de langue de la tête parlante articulatoire

La seconde approche proposée pour "augmenter" l'échographie linguale est d'animer automatiquement le modèle de langue de la tête parlante articulatoire décrite à la section 2.2.2.2, à partir du flux d'images ultrasonores [FHB14 ; Fab16 ; Fab+17]. Notons que contrairement à l'image échographique, la tête parlante articulatoire permet un affichage de la langue "en contexte", c'est-à-dire avec les autres structures du conduit vocal comme le palais, la paroi pharyngée, les dents, etc. Dans l'approche proposée, la mise en correspondance des images échographiques avec les paramètres de contrôle du modèle de langue de la tête parlante s'effectue de nouveau par apprentissage statistique. De façon similaire au système par conversion acoustico-articulatoire décrit à la section 2.2.2.2, les paramètres du modèle de conversion sont estimés à partir d'un corpus de données d'enrôlement, acquis en demandant à l'utilisateur (toujours nommé "locuteur source") de prononcer un ensemble de phrases ou de logatomes également prononcés (en laboratoire) par le locuteur de référence, et pour lesquelles nous disposons des trajectoires articulatoires associées (EMA). Une fois les paramètres du modèle de conversion estimés lors de la phase d'apprentissage, chaque image échographique est convertie en un vecteur de coordonnées EMA décrivant la forme et la position de la langue dans le conduit vocal de la tête parlante. Un schéma général de fonctionnement du système proposé est présenté à la figure 2.11.

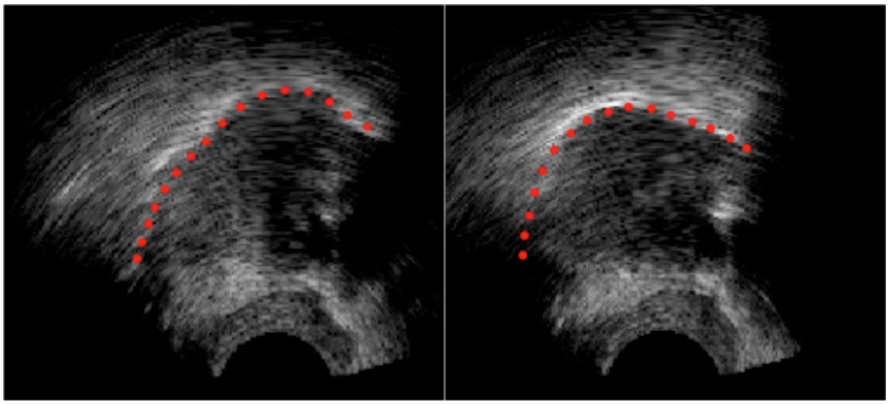
L'approche *EigenTongues*, déjà décrite à la section **??** et basée sur une analyse en composantes principales des images échographiques [Hue+07a], est utilisée pour extraire un vecteur de descripteurs visuels de chaque image échographique. Différents techniques de régression peuvent ensuite être utilisées pour convertir ce vecteur de descripteurs visuels en un vecteur de coordonnées EMA permettant l'animation du modèle de langue de la tête parlante. Dans [FHB14], nous avons évalué une approche par GMR. Le problème rencontré est le même que pour la conversion acoustico-articulatoire décrite précédemment, à savoir la nécessité de disposer d'une (trop) grande quantité de données d'enrôlement pour commencer à obtenir des performances satisfaisantes. Aussi, nous avons ensuite évalué l'approche C-GMR présentée à la section 2.2.2.2, et notamment les modèles SC-GMR et IC-GMR [Fab+17]. L'idée est toujours de faire précéder la conversion inter-modalités (image échographique vers EMA) par une conversion inter-locuteur (locuteur source vers locuteur de référence). L'approche C-GMR exploîte ici un modèle GMM associant (en adaptant les notations utilisées à la section 2.2.2.2),

FIGURE 2.10 – Segmentation automatique du contour de la surface supérieure de la langue dans les images échographiques à partir d'un ensemble limité de données d'apprentissage, basée sur une réduction de la dimension des images et des contours par analyse en composantes principales et sur une régression non-linéaire à l'aide d'un réseau de neurones. Extrait de [Fab+15].

(a) Principe générale de fonctionnement



(b) Exemples d'images échographiques segmentées automatiquement (images de la langue de la plan sagittal médian, l'apex est situé à droite)



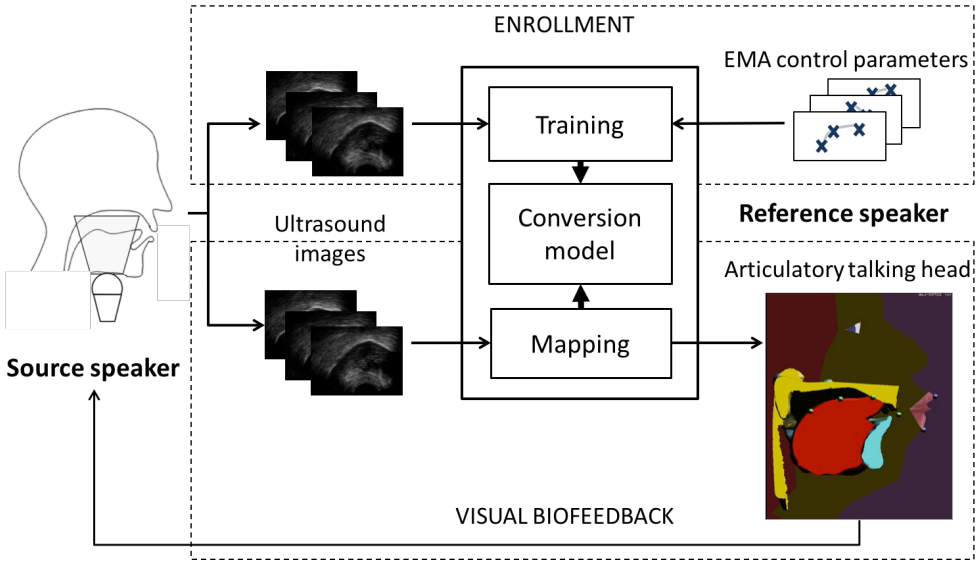


FIGURE 2.11 – Animation automatique d'une tête parlante articulatoire à partir d'images échographiques linguales. Extrait de [Fab+17].

chez le locuteur de référence et pour une même position de la langue, les descripteurs visuels issus de l'analyse de l'image échographique $\mathbf{X}$ avec les coordonnées EMA $\mathbf{Y}$. Les données d'apprentissage de ce modèle s'obtiennent en demandant au locuteur de référence d'enregistrer un large corpus de données, en laboratoire, et en enregistrant ses mouvements de langue à la fois par échographie et par EMA [15]. Ce GMM sur $(\mathbf{X},\mathbf{Y})$ et son régresseur GMR associé, sont ensuite utilisés pour initialiser les modèles SC-GMR, IC-GMR ou J-GMR dont les paramètres sont enfin estimées à partir des données échographiques d'enrôlement du locuteur source $\mathbf{Z}$, ainsi que de l'ensemble des données échographiques $\mathbf{X}$ et EMA $\mathbf{Y}$ disponibles pour le locuteur de référence.

Ces différents modèles ont été évalués sur une base de données multi-locuteurs [16] (le protocole expérimental est décrit dans [Fab+17]). Leurs performances en fonction de la quantité de données échographiques d'enrôlement considérée sont présentées à la figure 2.12. Ces résultats sont cohérents avec ceux présentés à la figure 2.9 dans le cadre de la conversion acoustico-articulatoire. Pour de faibles quantités de données d'enrôlement, l'approche IC-GMR fournit de bien meilleures performances que l'approche directe D-GMR (et également que l'approche SC-GMR). C'est résultats semblent confirmer l'intérêt d'estimer de façon robuste la structure de l'espace articulatoire du locuteur de référence et d'utiliser cette espace pour décrire celui du locuteur source, comme proposé par l'approche IC-GMR.

Dans [Fab+17 ; Fab16], nous nous sommes ensuite intéressés à évaluer la capacité de l'approche C-GMR à traiter une configuration linguale non-vue pendant la phase d'enrôlement.

[15] Une autre stratégie consiste à aligner temporellement *a posteriori* un enregistrement échographique et un enregistrement EMA d'une même phrase, sur la base des productions acoustiques. C'est ce qui a été fait dans [Fab+17]

[16] Base de données rendue publique et disponible sur `http://doi.org/10.5281/zenodo.1194786`

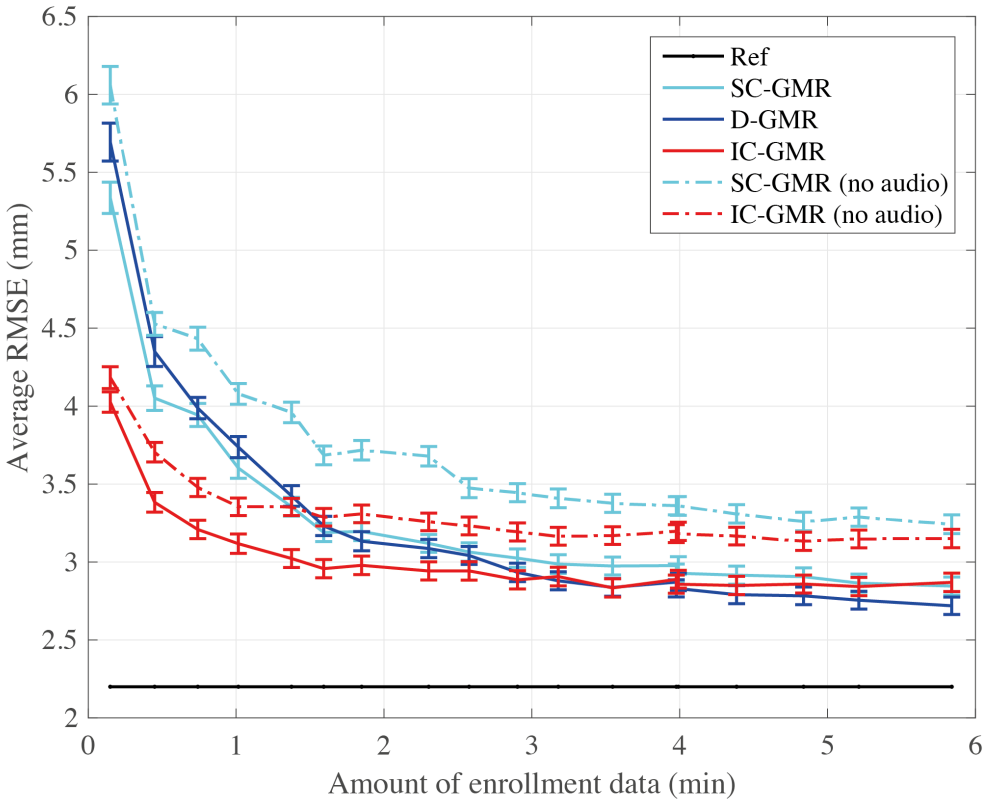


FIGURE 2.12 – Précision de l'animation du modèle de langue de la tête parlante articulatoire à partir d'une image échographique acquise chez un locuteur tiers (F1), en fonction de la quantité de données d'enrôlement utilisées pour l'estimation des modèles D-GMR, SC-GMR, et IC-GMR. RMSE en mm, les barres d'erreur représentent l'intervalle de confiance à 95%. *Ref* represente l'erreur de prédiction lorsque l'estimation est effectuée à l'aide du GMR **X**-vers-**Y** à partir des données échographiques du locuteur de référence (cette valeur fournit une borne inférieure de l'erreur). Pour les courbes "SC-GMR (no audio)" et "IC-GMR (no audio)", la mise en correspondance des données échographiques du locuteur source avec celles du locuteur de référence est effectuée sans exploiter les productions acoustiques. Cette condition expérimentale simule une utilisation du système par un patient incapable de vocaliser. Extrait de [Fab+17].

Dans la perspective d'une utilisation de ce système en contexte clinique, l'objectif est ici de simuler le cas d'un patient n'arrivant pas à articuler un certain phonème au début de sa rééducation. Nous faisons l'hypothèse que l'exploitation d'un modèle complet de l'espace articulatoire du locuteur source peut contribuer à améliorer cette capacité de généralisation. Le protocole expérimental consiste à construire des corpus d'enrôlement contenant des séquences de type VCV en omettant une voyelle ou une consonne, et des corpus de test contenant également des séquences VCV mais incluant cette voyelle ou cette consonne en particulier. Par exemple, lorsque nous testons la capacité d'un modèle à généraliser au phonème /t/, le corpus d'enrôlement est constitué de séquences VCV avec V choisie parmi {a i u ɛ o} et C parmi {k ʁ l s ʃ}, et le corpus de tests de séquences VCV avec C parmi {t d n}(le protocole expérimental est détaillé dans [Fab+17 ; Fab16]). Les résultats expérimentaux sont présentés à la figure 2.13. Ils confirment la meilleure capacité de généralisation du modèle IC-GMR, dont l'écart à la *baseline* (performances obtenues pour un corpus d'enrôlement phonétiquement dense) est bien plus faible que pour les autres modèles.

Enfin, à titre d'exemple, deux animations de la tête parlante à partir d'images échographiques du locuteur source M1, à l'aide de l'approche IC-GMR sont présentées à la figure 2.14. Une vidéo de démonstration du système est accessible à l'adresse `https://youtu.be/u8jb4b0fMsE`.

2.2.3 Bilan et perspectives

A ce jour, un logiciel d'illustration visuelle articulatoire intitulé *Ultraspeech-player* a été développé. Deux systèmes de retour visuel articulatoire basés sur une tête parlante articulatoire ont également été proposés. Le premier est basé sur l'animation automatique de cette tête parlante à partir du signal acoustique (voir section 2.2.2.2), le second sur l'animation du modèle de langue uniquement, à partir d'images échographiques (voir section 2.2.2.3). Ces deux systèmes s'appuient sur la technique C-GMR que nous avons développée pour l'adaptation d'un regresseur de type GMR à partir d'une faible quantité de données d'entrée seulement [Hue+15 ; GHA17b].

Ces systèmes de retour visuel n'ont à ce jour été testés qu'en laboratoire [17], et uniquement à partir d'enregistrements de locuteurs ne présentant pas de troubles articulatoires. Dans la perspective d'une utilisation en contexte clinique, une contrainte imposée par l'approche C-GMR doit d'abord être levée. Dans sa formulation actuelle, les phrases enregistrées lors de l'enrôlement d'un nouvel utilisateur sont en effet supposées avoir été prononcées de façon correcte, c'est-à-dire présenter le contenu phonétique attendu, ce qui n'est pas réaliste dans un contexte de rééducation orthophonique.

Différents cas de figure peuvent se présenter :

- Le nouveau locuteur présente un trouble articulatoire (phonétique) localisé qui l'amène

[17] Le logiciel d'illustration visuelle *Ultraspeech-player* a été utilisé pour différentes applications cliniques, détaillées à la section 2.3.

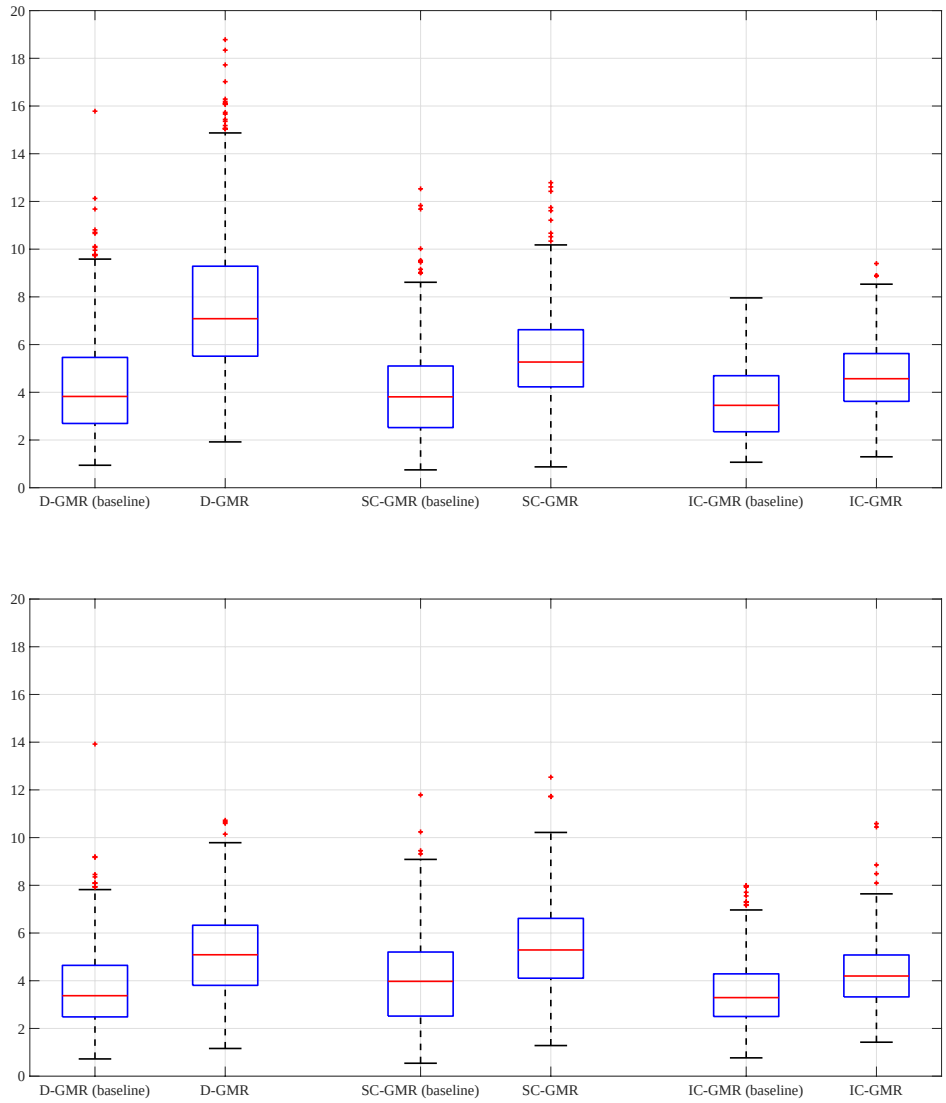


FIGURE 2.13 – Capacité des modèles D-GMR, SC-GMR et IC-GMR à traiter une configuration articulatoire non-vue pendant l'enrôlement (erreur en mm, locuteur F1 (haut) et M1 (bas)). Les *baselines* sont les performances moyennes obtenues lorsque le corpus d'enrôlement est phonétiquement dense. Extrait de [Fab+17].

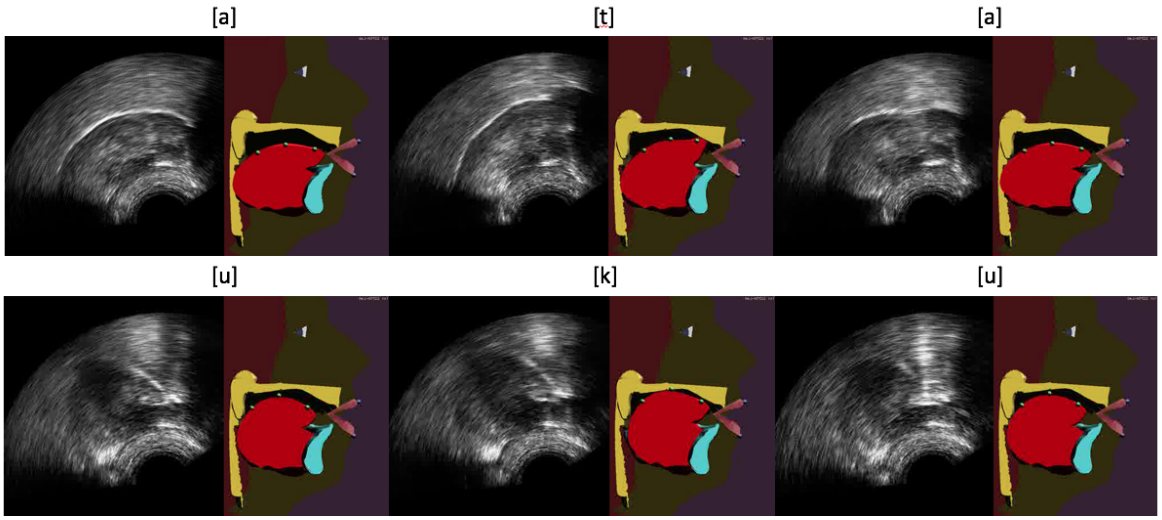


FIGURE 2.14 – Exemples d'animation du modèle de langue de la tête parlante articulatoire à partir d'image échographique pour les séquence [ata] (haut) et [uku] (bas) (locuteur source M1, approche IC-GMR). Extrait de [Fab+17].

à produire un son qui n'est pas dans l'inventaire phonétique du locuteur de référence (comme par exemple dans le cas du sigmatisme).

- Le nouveau locuteur présente un trouble phonologique. Certains phonèmes sont mal-prononcés mais uniquement dans certains contextes, et il existe des productions correctes de ces mêmes phonèmes dans d'autres contextes [18].

Dans les deux cas, il est impossible d'aligner "directement" les enregistrements acoustiques du nouveau locuteur avec ceux du locuteur de référence (enregistré en train de prononcer la même phrase). Une première piste pour éviter d'introduire une incohérence dans le modèle de conversion consiste à détecter automatiquement les phonèmes mal-prononcés et les disfluences dans le signal acoustique. Cette tâche a été étudiée soit dans le cadre de l'apprentissage d'une langue seconde [WY00 ; CSB00 ; LQM17], soit dans celui des troubles de la parole et du langage [LFM15 ; Dud+18]. Les principales approches proposées sont basées sur la création d'une mesure de qualité de prononciation (*goodness of pronuncation*, GOP) construite par alignement forcé du signal acoustique avec la séquence phonétique cible, à l'aide d'un système de reconnaissance automatique de la parole, basé soit sur des modèles de type HMM-GMM [19], soit sur un réseau de neurones profonds [LQM17].

Une fois l'erreur de prononciation détectée, son traitement dépend du type de trouble.

[18] Dans [SMJ04], les auteurs rapportent différents types de troubles phonologiques, liés à une déformation (ex. [caraval] pour carnaval), à une instabilité des productions (ex. le mot armoire produit [amwa] ou [abwa]), de la présence d'homophones (deux mots ayant la même prononciation), ou encore de l'absence de certains type de phonèmes comme les fricatives ou les voyelles nasales.

[19] La mesure de qualité de prononciation est alors construite à partir de la log-vraisemblaisance de modèles phonétiques, éventuellement normalisée par la durée des phonèmes observés et pondérée par une information a priori sur les erreurs classiques de prononciation dans un contexte donnée (voir par exemple [Dud+18]).

S'il s'agit d'un trouble phonétique localisé et systématique (premier scénario), une piste serait d'associer au segment correspondant, pendant la phase d'enrôlement, un mouvement articulatoire adapté, obtenu par exemple en enregistrant, en laboratoire, le locuteur de référence en train de "mimer" cette disfluence. Dans le cas d'un trouble phonologique (second scénario), un alignement des productions acoustiques du locuteur source avec celles du locuteur de référence, présupposant un contenu phonétique commun, n'est plus possible. Une nouvelle approche est donc nécessaire. Plusieurs pistes peuvent être données par les travaux sur la conversion de voix à partir de corpus "non-parallèles", c'est-à-dire une conversion pour laquelle on ne dispose pas d'enregistrements d'une même phrase par le locuteur source et le locuteur cible. Plusieurs approches ont été proposées dans la littérature parmi lesquelles : l'utilisation d'un système de reconnaissance automatique et la mise en correspondance a posteriori des segments correspondants à un même phonème [XSL16], l'adaptation, à une nouvelle paire de locuteurs, d'un GMR pré-entrainé sur un corpus contenant des enregistrements parallèles pour une autre paire de locuteurs [MVM06], l'utilisation de la technique dite des *EigenVoices* [Oht+09][20], et plus récemment les modèles neuronaux génératifs de type Cycle-GAN [KK18]. La technique C-GMR incluant une étape de conversion de voix (de façon explicite dans le cas du SC-GMR et implicite dans le cas de modèles IC-GMR et J-GMR), elle pourrait a priori être combiner avec ces techniques (notamment celle basée sur la régression par GMR) afin de disposer d'un système de retour visuel véritablement utilisable en contexte clinique.

2.3 Applications cliniques

Dans cette section, je décris les différentes applications des systèmes d'illustration et de retour visuel articulatoire décrits précédemment pour la rééducation orthophonique et l'aide à la prononciation d'une langue seconde.

2.3.1 Rééducation d'un trouble de substitution

Dans le cadre du mémoire d'orthophonie de Claire Bach et de Lorene Lambourion, nous avons testé l'apport de l'illustration articulatoire visuelle par échographie, pour la prise en charge d'un trouble phonologique chez l'enfant [BL15]. Le trouble considéré est la substitution de /k/ et /t/ en contexte /ʁ/ (l'enfant prononce "krotinette" au lieu de "trotinette") que l'on observe fréquemment dans le retard de parole chez les jeunes enfants. Quatorze enfants âgés de 5 ans 1 mois, à 7 ans 5 mois, ont été répartis en deux groupes homogènes de sept participants. Chaque enfant a été suivi pendant trois semaines consécutives au rythme de 30 minutes hebdomadaires, par les étudiantes en orthophonie. Chaque séance consistait en un entraînement visant à

[20] L'idée générale de l'approche *EigenVoices* est d'exploiter un ensemble de modèles de conversion de type GMR entre plusieurs paires de locuteurs dont les paramètres sont préalablement estimés à partir de corpus parallèles. Un "modèle de modèle" (EV-GMM) est ensuite construit par analyse en composante principale des vecteurs de moyenne de l'ensemble des composantes de ces différents GMR. En phase de conversion, à partir d'un ensemble de données d'adaptation, un modèle de conversion adapté à la voix du nouveau locuteur est construit par pondération des modèles GMR pré-entrainés sur d'autres locuteurs.

développer la conscience proprioceptive et kinesthésique de la langue (par exemple en insérant le groupe diconsonantique problématique entre 2 voyelles, avec éventuellement d'abord un court silence entre les 2 consonnes, [at-ra],[ut-ru], etc., puis en amenant l'enfant à enlever le silence). Chaque enfant a bénéficié d'une phase d'entrainement avec l'illustration visuelle via le logiciel *Ultraspeech-player* décrit à la section 2.2.2.1, et d'une autre phase d'entrainement sans illustration. Cette étude a l'objet d'une communication au congrès national de recherche en orthophonie [Fab+16]. Une partie des résultats de cette étude est présentée à la figure 2.15.

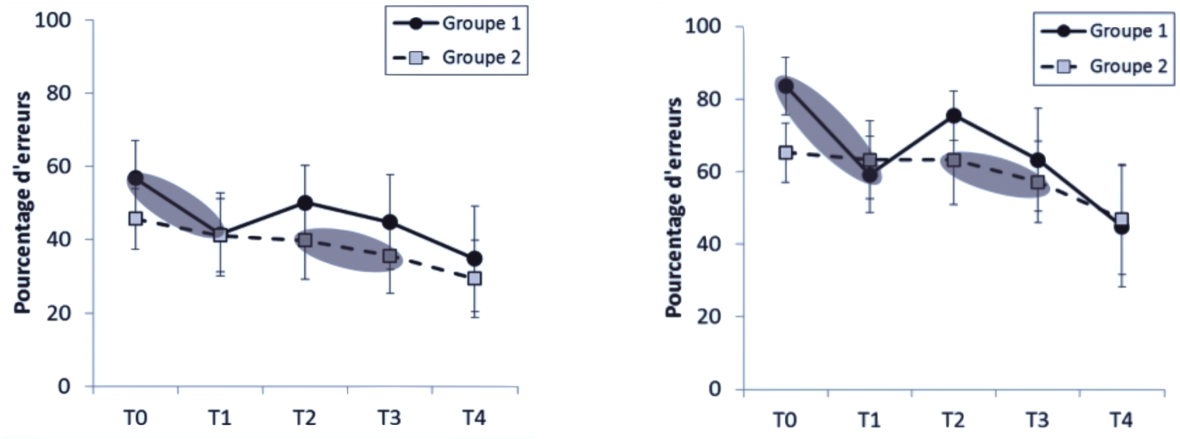


FIGURE 2.15 – Pourcentages d'erreurs de production du groupe diconsonantique /tʁ/ pour le groupe 1 entraîné avec *Ultraspeech-player* entre T0 et T1 et sans *Ultraspeech-player* entre T2 et T3, et pour le groupe 2 entraîné sans *Ultraspeech-player* entre T0 et T1 et avec *Ultraspeech-player* entre T2 et T3, sur l'ensemble des stimuli (gauche) ; idem pour les stimuli avec tʁ en position initiale uniquement (droite). Les barres représentent les erreurs-type. Les effets de l'entraînement avec *Ultraspeech-player* sont mis en relief par les ellipses. Extrait de [Fab+16].

Bien que le nombre de patients impliqués dans cette étude reste limité, cette première étude semble montrer que la rééducation par illustration visuelle est à l'origine d'un bénéfice significatif et immédiat sur la production de la séquence tʁ en position initiale notamment, par rapport à une rééducation classique.

2.3.2 Rééducation des troubles articulatoires liés à une aphasie non fluente chronique post-AVC

Dans le cadre d'une collaboration avec le Laboratoire de Psychologie et NeuroCognition (LPNC), le logiciel *Ultraspeech-player* a été utilisé dans une étude de cas impliquant une patiente présentant une aphasie non fluente chronique apparue après un accident vasculaire cérébral (AVC) [Hal+18] [21]. La rééducation a commencé par une première session pendant laquelle l'orthophoniste expliquait les bases de l'articulation en s'appuyant sur les illustrations visuelles fournies par le logiciel (voyelles isolées, consonnes isolées, puis séquences de type

[21] Cette étude a été autorisée par le comité de protection des personnes (CPP-ISIS 07PHR04, DCIC/06/25).

voyelle-consonne-voyelle ou VCV). Puis la patiente effectuait 13 séances en autonomie avec le logiciel, suivies de 3 séances de nouveau avec l'orthophoniste. Les résultats expérimentaux ont montré une amélioration des capacités de dénomination, de lecture, de répétition de mots, mais également de la précision articulatoire pour les voyelles et les consonnes [22]. Cette étude de cas suggère que la vision explicite d'une cible articulatoire permet d'améliorer la mise en relation d'un but moteur avec le retour somatosensoriel.

2.3.3 Apprentissage de la prononciation des voyelles du Français par des locuteurs Chinois Mandarins

Dans le cadre du mémoire de Master 2 Recherche en Sciences du Langage de Xiaoou Wang, nous avons évalué une méthode d'aide à l'apprentissage de la prononciation du Français pour les apprenants Chinois basé d'une part sur la tête parlante articulatoire décrite à la section 2.2.1.1, et d'autre part sur un paradigme de *close shadowing*, qui consiste à répéter le stimulus dès qu'on commence à le percevoir (c'est-à-dire avant qu'il soit terminé). Ce paradigme expérimental exploite explicitement les liens entre perception et production de la parole mentionnés à la section 2.1.1 [PC80 ; Mar85 ; Fow+03]. Ce paradigme a été évalué dans le cas de la perception de la parole audiovisuelle [RMG87], puis a été utilisé pour l'apprentissage d'une langue seconde [DK98]. Dans [WHB14], nous proposons d'utiliser ce paradigme pour la perception d'une parole "augmentée", c'est-à-dire en faisant visualiser à l'apprenant des mouvements linguaux cibles, présentés à l'aide de la tête parlante articulatoire. Nous nous sommes intéressés à la prononciation par des locuteurs Chinois Mandarins des voyelles du Français /e o ɔ øœ/. Ces locuteurs ont en effet tendance à assimiler les trois premières aux diphtongues /ei/, /ʏʊ/ et les deux dernières à /ʏ/. L'étude a impliqué 14 participants de moins de 25 ans, ayant étudié le Français pendant 6 mois. L'entraînement reposait sur un principe de comparaison, montrant successivement, pour une des voyelles à acquérir, les autres voyelles qui pouvaient lui être assimilées (voir [WHB14] pour plus de détails). En ligne avec les considérations théoriques sur le contrôle moteur de la production de la parole décrites précédemment, qui suggèrent que l'apprentissage d'un nouveau geste passe par son exécution lente, les stimuli de type VCV enregistrés à vitesse d'élocution normale par le locuteur de référence ayant servi pour la création de la tête parlante, ont été ralentis à l'aide d'un *vocoder* HNM [Sty01]. Une partie des résultats de cette étude sont rappelés à la figure 2.16.

Ces résultats montrent que le groupe ayant reçu un entraînement basé sur le son et l'illustration visuelle articulatoire converge plus vers la cible canonique que le groupe ayant reçu un entraînement basé uniquement sur le son. Bien que les résultats restent contrastés, cette étude tend à montrer un apport positif de l'illustration visuelle pour l'apprentissage de la prononciation d'une langue seconde.

[22] Cette étude implique aussi une analyse de l'activité cérébrale de la patiente avant et après la rééducation, qui ne sera pas détaillée ici

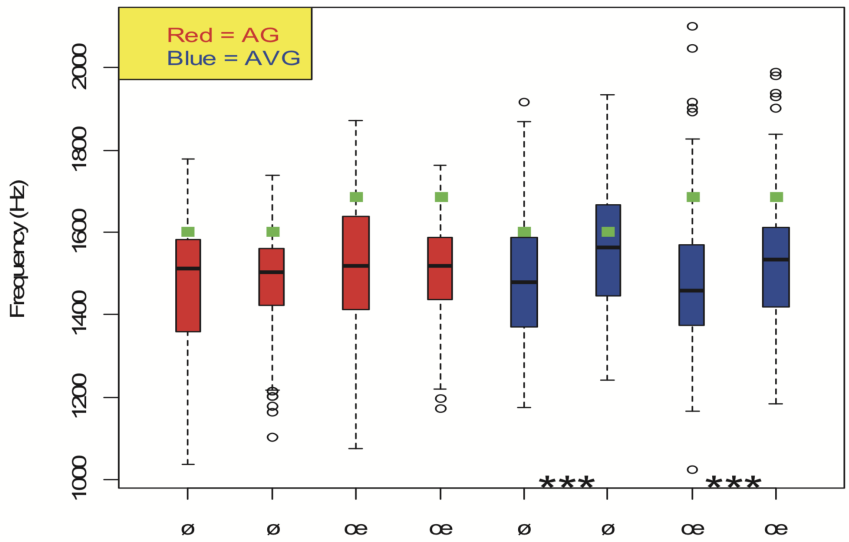


FIGURE 2.16 – Apprentissage des voyelles du Français par des locuteurs Chinois à l'aide d'une méthode basée sur une illustration visuelle articulatoire exploitant une tête parlante. Distributions du second formant F2 pour les voyelles ø et æ, avant et après apprentissage (première et seconde barre, pour les 2 voyelles). Les carrés verts fournissent une valeur canonique pour le Français [Cal89]). Extrait de [WHB14].

2.3.4 Etude clinique Revison

Le cancer de la langue fait partie de la classe des cancers des voix aéro-digestives supérieures, qui regroupe les cancers de la cavité orale et du pharynx (lèvres, bouche, pharynx) et ceux du larynx. En France, en 2015, l'Institut National du Cancer (INCa) estimait à 11610 le nombre de nouveaux cas de cancers de la cavité orale et du pharynx. La chirurgie, la radiothérapie et la chimiothérapie constituent les trois traitements de référence des cancers des VADS La chirurgie d'exérèse permet de retirer la tumeur et/ou les aires ganglionnaires touchées. La chirurgie buccale entraîne souvent une perturbation de la mobilité et de la sensibilité de la langue qui sont nécessaires à la déglutition et à l'articulation de la parole. La rééducation orthophonique vise à redéfinir les buts moteurs afin d'améliorer la précision articulatoire.

Le projet Revison (pour Retour Visuel par imagerie ultraSonore) est la première étude de groupe visant à quantifier l'apport du retour visuel par échographie pour la rééducation articulatoire des personnes ayant subi une glossectomie [23]. Cette étude a été menée dans le cadre de la thèse de Diandra Fabre [Fab16] et du mémoire de Master Recherche de Marion Girod-Roux [Gir17], orthophoniste, en collaboration avec le Centre Médical Rocheplane, à St Martin d'Hères (38) qui prend en charge les patients dans les premiers jours qui suivent leur opération. Cette étude a reçu l'accord du Comité de Protection des Personnes (CPP) Lyon Sud-Est 2 (numéro 69HCL15 0736) et s'est déroulée sur une période de 2 ans, de mai 2016 à mai 2018. En lien avec le cadre théorique décrit à la section 2.1, nous proposons une

[23] La glossectomie est l'ablation d'une partie de la langue suivie d'une reconstruction éventuelle à l'aide d'un lambeau musculocutané.

méthode de rééducation qui combine illustration visuelle et retour visuel par échographie. L'objectif est de permettre au patient de comparer une cible articulatoire visuelle avec ses propres mouvements [24]. Ce paradigme "illustration+retour" est comparé à une rééducation basée uniquement sur l'illustration visuelle. Dans ce cas, le patient ne peut s'appuyer que sur son retour somatosensoriel pour ajuster son geste. L'illustration visuelle est fournie par le logiciel *Ultraspeech-player* (voir section 2.2.2.1) et le retour visuel est effectué à l'aide d'un logiciel développé à cet effet et intitulé *Ultraspeech-biofeedback* [25]. Le protocole de rééducation Revison est illustré à la figure 2.17.

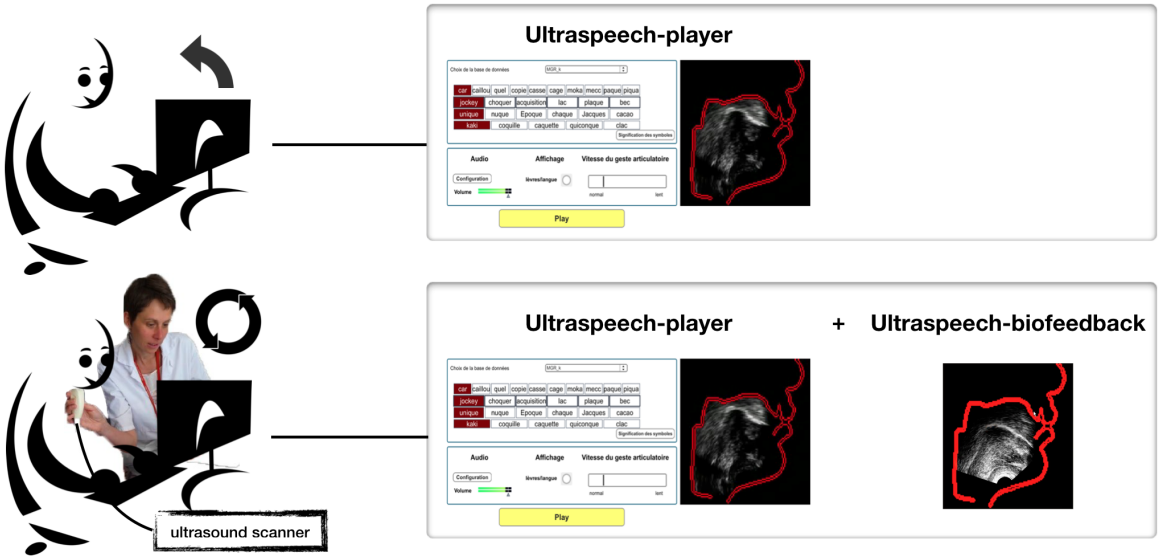


FIGURE 2.17 – Protocoles de rééducation articulatoire par illustration visuelle (haut) et illustration+retour visuel (bas) utilisés dans l'étude *Revison*.

10 patients ont été retenus pour participer à cette étude (7 hommes et 3 femmes, âge moyen 59 ans, inclusion environ 20 jours après la chirurgie). Le tableau clinique à respecter pour l'inclusion ainsi que les détails des chirurgies et des traitements de chaque participant sont disponibles dans [Gir17]. Ce groupe de 10 patients est réparti en 2 sous-groupes (répartition aléatoire lors de l'inclusion d'un nouveau patient dans le protocole). Après inclusion à T0, les patients du groupe nommé "*I+R puis I*" sont rééduqués pendant 10 sessions, à raison d'une session par jour, à l'aide du protocole "illustration+retour visuel". Leur progrès sont évalués après ces 10 séances (T1), puis 10 nouvelles séances de rééducation sont conduites à l'aide du protocole "illustration visuelle seule". Leurs progrès sont de nouveau évalués à l'issue de ces séances (T2). Un protocole symétrique est utilisé pour les patients de l'autre groupe, nommé "*I puis I+R*". Deux bilans sont utilisés pour évaluer la progression des patients : le protocole MBLF (Motricité Bucco-Linguo-Facial) [GL11] qui regroupe un ensemble de praxies

[24] Cette cible étant acquise sur un locuteur de référence, une forme de normalisation mentale de la part du patient est cependant nécessaire pour la projeter mentalement dans son propre espace articulatoire.

[25] Tout comme *Ultraspeech-player*, ce logiciel permet à l'utilisateur de replacer grossièrement l'image échographique de la langue du patient dans un conduit vocal générique, afin de rendre la visualisation plus intuitive (voir figure 2.17).

visant à quantifier la coordination musculaire, et le BECD (Batterie d'Evaluation Clinique de la Dysarthrie) [PR06], qui permet d'évaluer le trouble d'articulation et de parole. Ces bilans sont réalisés d'une part par l'orthophoniste ayant réalisé la rééducation des patients (Marion Girod-Roux) ainsi que par une autre orthophoniste n'ayant pas participé à l'expérience. Par soucis de concision, nous ne présentons ici qu'un sous-ensemble du BECD, à savoir le test d'intelligibilité phonétique ou TPI [26]. Il s'agit d'un test d'identification à choix multiple dans lequel le patient lit à voix haute 52 mots (13 séries de 4 mots qui s'opposent deux à deux par un ou deux contrastes phonétiques). L'orthophoniste sélectionne le mot perçu le plus proche dans chaque série de 4 mots (un mot cible et trois distracteurs).

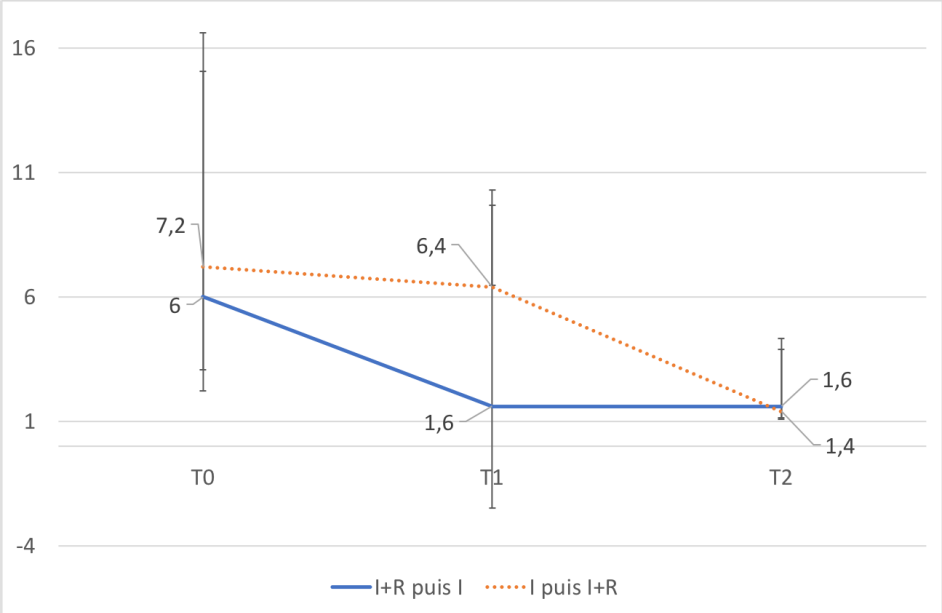


FIGURE 2.18 – Résultats du test d'intelligibilité phonétique (TPI), en terme de nombre d'erreurs commises (moyenne et écart type par groupe), effectué dans le cadre de l'étude *Revison*. En traits pointillés, les patients du groupe "*I puis I+R*" ayant d'abord été traités par illustration visuelle seule entre T0 et T1 (10 séances), puis par illustration et retour visuel combinés entre T1 et T2 (10 séances). En traits pleins, les patients du groupe "*I+R puis I*" ayant d'abord été traités par illustration et retour visuel combinés entre T0 et T1, puis par illustration visuelle seule entre T1 et T2.

L'évolution du nombre moyen d'erreurs commises pour les 10 patients, au fur et à mesure de la rééducation (c'est-à-dire aux temps T0, T1 et T2) et en fonction du paradigme de rééducation mis en œuvre (*I puis I+R* vs. *I+R puis I*) est représentée sur la figure 2.18. Tout d'abord, on constate une forte variabilité inter-individuelle. Cette variabilité peut éventuellement s'expliquer par les différences entre les chirurgies pratiquées d'un patient à l'autre. Par ailleurs, on observe d'une part une certaine homogénéité entre les 2 groupes au début et à la fin de l'étude, et d'autre part une progression plus importante lorsque le retour visuel est

[26] L'ensemble des résultats de cette étude ont été présentés dans un article en cours de révision au moment de l'écriture de ce manuscrit. Un *preprint* est cependant disponible à `https://hal.archives-ouvertes.fr/hal-01977670/`.

utilisé en complément de l'illustration visuelle (passage de 6 à 1.6 pour le groupe "*I+R puis I*" entre T0 à T1 et passage de 6.4 à 1.4 pour le groupe "*I puis I+R*" entre T1 et T2). Cet apport positif du retour visuel est confirmé par les analyses statistiques (non reportées ici), basées sur une régression ordinale avec effet aléatoire. Les résultats de cette première étude de groupe semblent donc montrer un effet bénéfique de la méthode de rééducation proposé, qui combine illustration et retour visuel articulatoire. Cependant, ces résultats doivent être pris avec précaution compte tenu de la taille limitée de la cohorte de patients impliqués dans l'étude et de l'hétérogénéité des troubles de ces derniers.

Chapitre 3

Projet de recherche

Sommaire

En plus des différentes perspectives de recherche détaillées aux chapitres précédents (voir sections 1.1.3, 1.2.3, 1.3.3 et 2.2.3), je souhaite développer un nouvel axe de recherche portant sur l'apprentissage non ou faiblement supervisé d'espaces de représentation de la parole (acoustique, moteur/articulatoire, linguistique), à l'aide de réseaux neuronaux génératifs [1]. Les applications visées sont la modélisation computationnelle de certaines hypothèses cognitives qui sous-tendent la perception et le contrôle de la production de la parole, et le traitement automatique de la parole pathologique.

3.1 Contexte et objectifs

Le projet de recherche proposé est motivé par les considérations suivantes :

- Comme mentionné à la section 2.1, certaines théories sur la perception de la parole [LM85 ; Sch+12] font référence à des mécanismes inférentiels de simulation motrice pour le décodage de traits phonétiques à partir d'un stimulus auditif ou audiovisuel. L'accès à des connaissances procédurales motrices est également un concept central dans les modèles de contrôle de la production [HN11 ; GV12 ; Per12], au travers notamment du concept de modèle interne (voir section 2.1). De nombreuses études ont discuté de l'intérêt d'exploîter des connaissances motrices et articulatoires dans les systèmes artificiels de perception de la parole [RSS96 ; Moo96 ; DRS97 ; Kin+07]. Ce projet s'inscrit

[1] Cet axe de recherche sera en partie développé dans le cadre de la chaire "Parole" de l'institut d'intelligence artificiel MIAI, dans laquelle je suis impliqué.

dans cette ligne de recherche. Je fais également l'hypothèse qu'un accès explicite à des connaissances motrices peut être bénéfique, et je souhaite tester cette hypothèse dans le cadre du traitement automatique de la parole pathologique.

- Par ailleurs, les théories sur la perception et la production de la parole font appel à un autre concept intitulé "codage prédictif". L'idée d'un codage prédictif de l'information par le cerveau (*predictive brain*) est apparue dans les années 50 [Att54 ; Bar61]. Elle suppose que notre cerveau implémente un algorithme d'inférence en ligne des entrées sensorielles à venir, à partir des entrées sensorielles passées. Cette prédiction s'appuierait sur la recherche de régularités dans le signal acoustique, mais recruterait également le niveau moteur et le niveau linguistique. De plus, cette théorie suppose que le cerveau n'encoderait pas explicitement chaque nouvelle entrée sensorielle, mais plutôt l'erreur entre cette dernière et la prédiction qu'il en fait. Friston et coll. ont formalisé cette théorie dans un cadre Bayésien [Fri05], autour du concept de minimisation de l'énergie libre [FKH06 ; Fri10] qui est directement en rapport avec la minimisation des erreurs de prédiction d'un système de codage. De récentes études ont identifié les corrélats neuronaux d'un codage prédictif de la parole [AG12] [OMH18]. Par ailleurs, on retrouve ce même concept de codage prédictif dans différentes technique d'analyse et de codage du signal de parole, comme l'analyse prédictive linéaire LPC (*Linear Predictive Coding*). Dans ce projet, je propose d'étudier le codage prédictif par le biais de la modélisation computationnelle, en utilisant l'apprentissage profond pour évaluer la quantité d'information a priori prédictible dans la parole auditive et audiovisuelle, à l'échelle de la syllabe voir du mot (environ 150ms). Je souhaite étudier dans ce cadre l'intérêt de baser cette prédiction sur des représentations motrices (articulatoires) et linguistiques.

- Enfin, la plupart des systèmes artificiels conçus pour résoudre de façon automatique certaines tâches de perception et de production de la parole, et notamment ceux développés dans mes travaux précédents, exploitent des modèles statistiques dont les paramètres sont généralement estimés par apprentissage automatique strictement supervisé. Par exemple, un système de reconnaissance automatique de la parole exploite une base de données associant un enregistrement sonore d'un locuteur avec une séquence d'étiquettes phonétiques correspondant à la phrase prononcée par ce dernier ; un système d'inversion articulatoire exploite une base de données mettant en regard, à chaque instant, une configuration articulatoire avec le contenu spectral de la parole associée, etc. Cette apprentissage supervisé des liens entre les différents espaces de représentation de la parole (acoustique, articulatoire, linguistique) semble *a priori* éloigné de la manière dont un humain perçoit, traite et produit de la parole. En effet, lors des premières années de sa vie, un enfant commence à comprendre et à produire de la parole avec un minimum de supervision, par exemple sans que lui soit pré-découpé explicitement le flux de parole en une suite de phonèmes ou de syllabes. L'enfant semble capable de construire des représentations linguistiques de façon faiblement supervisée, à partir uniquement des productions acoustiques "brutes" de ses interlocuteurs (dont certains indices acoustiques peuvent néanmoins être renforcés pour en faciliter la segmentation [THS05]). De même, les modèles internes impliqués dans la perception et le contrôle de la production de la parole semblent également être appris avec un minimum de supervision. En effet,

l'enfant apprend différents liens sensorimoteurs sans même avoir accès aux trajectoires articulatoires associées aux stimuli auditifs qu'il perçoit. Cette apprentissage s'appuie sur la perception de ses propres productions acoustiques, de celles de son environnement (qu'il peut chercher à imiter), de la statistique de sa langue, mais probablement également de la découverte de son propre conduit vocal, de sa géométrie, de ses propriétés biomécaniques, et de sa réponse acoustique, qu'il peut explorer pendant les premières vocalisations puis pendant la phase de babillage. L'enfant semble donc capable d'apprendre les liens entre différents espaces de représentation de la parole - acoustique, articulatoire, linguistique - à partir de données hétérogènes et de façon non ou faiblement supervisé.

Dans ce projet, je propose de développer des systèmes artificiels de perception et de production de la parole, qui s'inspirent sur ces hypothèses cognitives, à savoir : une exploitation conjointe des représentations acoustiques, motrices et linguistique pour la perception et la production, l'implémentation d'un mécanisme de codage prédictif basé sur ces représentations, et enfin un apprentissage faiblement supervisé de ces représentations [2]. Ces systèmes seront appris à partir de données "réelles" (non simulées), "multimodales" (signaux acoustiques, données articulatoires, textes), et avec un minimum de connaissances expertes pour leur description. Dans ce cadre, je souhaite privilégier une approche par apprentissage automatique, basée notamment sur les réseaux de neurones génératifs. Les applications visées sont d'une part, l'étude, par le biais de la modélisation computationnelle, des hypothèses cognitives mentionnées précédemment (théorie (perceptuo-)motrice de la perception, codage prédictif, apprentissage de la parole chez l'enfant), et d'autre part, le traitement automatique de la parole pathologique, et notamment son rehaussement. Sur ce point, je propose d'adapter à la parole un paradigme utilisé en vision par ordinateur appelé *inpainting*, qui regroupe un ensemble de techniques de reconstruction d'images détériorées, par remplissage des parties manquantes [Yeh+17 ; Yu+18]. En considérant que la parole pathologique (et notamment la parole dysarthrique) contient à la fois des segments intelligibles et d'autres qui ne le sont pas (les parties manquantes), je fais l'hypothèse qu'un modèle computationnel tel que celui envisagé dans ce projet, implémentant un mécanisme de codage prédictif basé sur un accès à des représentations articulatoires et linguistiques, devrait être capable de restaurer, par synthèse, certains segments inintelligibles.

Après un bref état de l'art sur les modèles computationnels de perception et de production de la parole, je présenterai des résultats préliminaires sur l'utilisation des réseaux convolutionnels pour la modélisation du codage prédictif de la parole auditive et audiovisuelle. Je détaillerai ensuite les pistes envisagées pour le développement d'un modèle de perception/production basé sur les réseaux antagonistes génératifs (GAN), implémentant certaines des hypothèses cognitives mentionnées précédemment. Je terminerai en présentant une possible utilisation de ce modèle pour le rehaussement de la parole pathologique par *speech inpainting*.

[2] Ce dernier point est en lien avec l'axe de recherche dit "*Zero Ressource*", visant à concevoir des systèmes de reconnaissance et de synthèse de la parole avec un minimum de supervision [Ver+15].

3.1.1 État de l'art

La modélisation computationnelle de la perception de la parole a fait l'objet de nombreuses études. On peut distinguer deux types de modèles : les modèles utilisés dans les systèmes de reconnaissance automatique de la parole, et dont certains peuvent faire appel à des représentations motrices pour le décodage des unités phonétiques (voir [Kin+07] pour revue), et les modèles moins focalisés sur la performance de ce décodage, mais dont l'architecture est construite à partir d'hypothèses sur les mécanismes cognitifs qui sous-tendent la perception, et notamment sur la nature des représentations des unités phonétiques. Dans cette seconde catégorie, certains modèles s'appuient sur une représentation purement auditive [Cla+08], et d'autres font appel à des représentations motrices. Par exemple, [Cas+11] propose un modèle de perception implémentant l'estimation par réseau de neurones, à partir du signal acoustique, d'invariants phonétiques moteur, construits à partir d'une base acoustico-articulatoire multi-locuteurs. Les auteurs montrent que cette représentation motrice est plus compacte que la représentation auditive, et qu'elle améliore la reconnaissance, notamment en présence de bruit.

Du coté de la modélisation computationnelle de la production (et du contrôle de cette dernière), le modèle le plus connu est probablement le modèle DIVA de Guenther et coll. [Gue95 ; GV12]. Il s'agit d'un modèle à la fois computationnel et anatomique, qui fait appel à la notion de cartes, représentant différentes zones du cerveau. Ce modèle, illustré à la figure 3.1a, implémente deux types de contrôle : le contrôle *feedforward*, qui génère le geste de production à partir d'une cible phonétique (à l'aide des *Articulator Velocity Maps* et des *Position Maps*, dont le rôle peut être rapproché de celui du modèle interne inverse), et le contrôle *feedback*, qui analyse le retour sensoriel de cette production et le compare avec le retour sensoriel prédit (par les *Auditory Target Map* et les *Somatosensory Target Map*, assimilable au modèle interne direct). Contrairement au modèle de contrôle moteur décrit à la section 2.1 et inspiré de [KFS87] et [Per12], dans DIVA, la correction des commandes motrices ne s'effectue qu'après comparaison entre le retour sensoriel prédit et le retour sensoriel vécu. Cependant, ce dernier n'étant disponible qu'après l'exécution du mouvement, un ajustement des commandes motrices par contrôle *feedback* uniquement semble trop lent pour être réaliste. Houde et coll. proposent donc un autre modèle, illustré à la figure 3.1b, basé sur la notion de *State Feedback Control* [HN11] et qui permet un ajustement des commandes motrices, uniquement à partir du résultat de la simulation de l'exécution du mouvement par le modèle interne direct.

Un autre modèle proposé par Jean-Luc Schwartz, Julien Diard et Pierre Bessière est le modèle COSMO pour "*Communicating about Objects using Sensory–Motor Operations*" [Mou+15 ; Lau+17]. Ce modèle peut tout d'abord être décrit comme une transcription, sous un formalisme Bayésien, d'une situation de communication entre un locuteur et un auditeur. Brièvement, un locuteur veut produire une unité distinctive (de type phonème ou syllabe) appelé "objet moteur" et représenté par une variable aléatoire discrète O_L, à laquelle il associe un geste moteur décrit par la variable aléatoire continue M, qui a pour conséquence une production acoustique vue comme la réalisation d'une variable aléatoire continue S. L'auditeur perçoit cette production acoustique et décode une unité distinctive ou "objet sensoriel" O_S. La communication est un succès si $O_S = O_L$. De façon importante COSMO s'appuie sur

l'hypothèse d'internalisation de cette situation de communication dans le cerveau d'un même agent communiquant, comme illustré à la figure 3.1c. Aussi, est associé à un agent un modèle Bayésien associant l'ensemble de ces variables : objet moteur O_L, objet sensoriel O_s, représentations motrices M, représentations sensorielles S, ainsi qu'une variable booléenne notée C encodant le succès ou l'échec de la communication. En factorisant la probabilité conjointe sur ces différentes variables en un produit de probabilités conditionnelles, ce formalisme permet notamment de retrouver différents éléments des modèles de perception et de production mentionnés précédemment, tel que :

$$P\left(CO_LSMO_S\right) = P\left(O_S\right) P\left(M|O_S\right) P(S|M) P\left(O_L|S\right) P\left(C|O_SO_L\right) \tag{3.1}$$

avec par exemple $P(S|M)$ qui fait directement référence au modèle interne (direct), $P(O_L|S)$ qui est un décodeur acoustico-phonétique, $P(M|O_S)$ qui encode les commandes motrices à effectuer pour produire un certain phonème, etc. A chacun de ces termes est associé une densité de probabilité a priori, fixée au moment de la définition du modèle et dont la forme et la technique d'estimation des paramètres dépend des versions du modèle. Pour plus de détails, le lecteur est invité à consulter [Lau+17] et [Bar18].

3.2 Codage prédictif de la parole auditive et audiovisuelle

Dans le cadre d'une première étude sur le développement d'un modèle computationnel de la perception de la parole implémentant un mécanisme de codage prédictif, nous avons récemment cherché à quantifier le gain potentiel apporté par cette stratégie, pour la parole auditive et la parole audiovisuelle [3]. Par soucis de concision, nous ne rappelons que les éléments clés de cette étude [4]. Si la prédiction à court terme (de l'ordre de 20 à 30ms) du signal de parole a largement été abordée dans le cadre du codage, et notamment dans celui de l'analyse prédictive linéaire LPC, aucune étude n'a (à notre connaissance) évalué la possibilité de réaliser une prédiction à plus long terme, par exemple à l'échelle du phonème ou de la syllabe (c'est-à-dire sur un empan temporel allant de 25ms à 150ms). C'est ce que nous avons cherché à faire, dans le cadre de la parole auditive et audiovisuelle, à l'aide d'une approche par apprentissage profond.

Plus formellement, le problème du codage prédictif de la parole peut s'écrire :

$$\widehat{\mathbf{x}}_{t+\tau_f} = f(\mathbf{x}_t, \mathbf{I}_t, \mathbf{x}_{t-1}, \mathbf{I}_{t-1}, ..., \mathbf{x}_{t-\tau_p}, \mathbf{I}_{t-\tau_p}) \tag{3.2}$$

avec $\mathbf{x}_t$ une observation acoustique décrivant classiquement le contenu spectral d'un segment du signal acoustique de parole jugé stationnaire [5], $\mathbf{I}_t$ l'image d'un visage parlant correspondant à cette observation, et τ_p /τ_f l'empan temporel passé/futur considéré pour la prédiction,

[3] Ce travail a débuté dans le cadre du projet ERC *SpeechUnit* de Jean-Luc Schwartz (voir section 4.10).

[4] Cette étude fait l'objet d'un article de revue soumis au moment de l'écriture de ce manuscrit. Un *preprint* est disponible sur `https://www.biorxiv.org/content/10.1101/471581v1`.

[5] Dans notre étude, une observation acoustique est soit le spectre d'amplitude à court terme (257 coefficients), soit un modèle de l'enveloppe spectrale de type MFCC (13 coefficients). Dans les deux cas, une fenêtre d'analyse de 25ms est utilisée. Le décalage de cette fenêtre d'analyse est également fixé à 25ms afin de ne pas introduire une redondance artificielle entre deux observations acoustiques consécutives, et biaiser ainsi la prédiction.

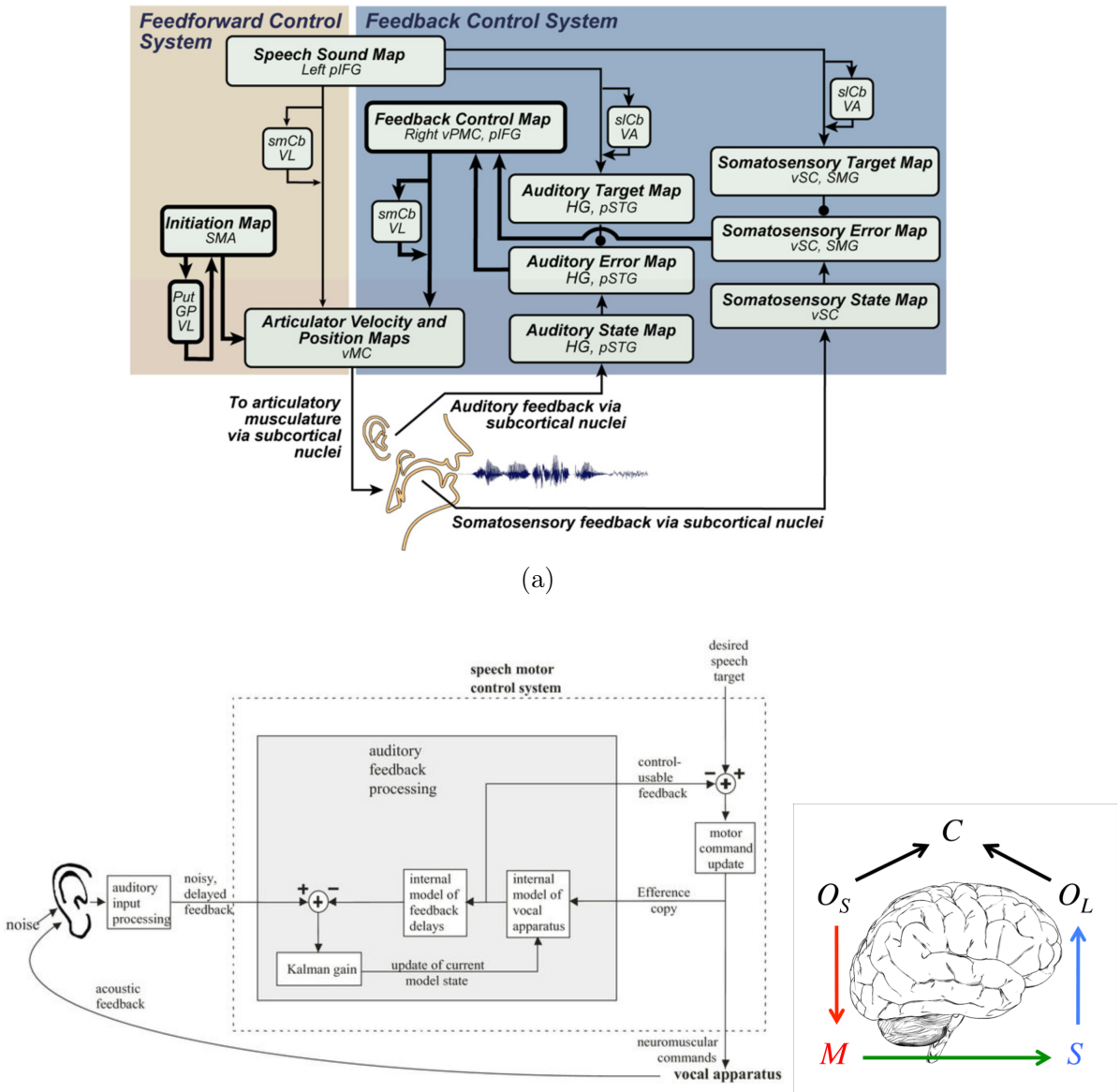


FIGURE 3.1 – (a) Modèle DIVA (extrait de [TRG08]) - (b) Modèle *State Feedback Control* (extrait de [HN11]) - (c) Modèle d'un agent COSMO (extrait de [Bar18])

en nombre de trames [6]. Nous présentons à la figure 3.2 les différents modèles proposés pour approximer cette fonction f dans le cas de la parole auditive et dans celui de la parole audiovisuelle. Ces modèles sont notamment basés sur des réseaux convolutionnels (CNN), ce qui permet de traiter directement des représentations "brutes" (bas-niveau) du son et des images (spectre d'amplitude et image de visage).

Les expériences sont menées sur la base de données NTCD-TIMIT [Abd17] [7]. Deux métriques sont utilisées pour mesurer la qualité de la prédiction, l'erreur quadratique moyenne (*Mean Square Error* ou MSE_{τ_p,τ_f}, qui est également la fonction de coût à minimiser lors de l'apprentissage des modèles) et la variance expliquée (*Explained Variance score* ou EV_{τ_p,τ_f}) évaluant la proportion avec laquelle les observations prédites contribuent à expliquer la variance des observations observées, et qui est directement liée au gain de codage G_{τ_p,τ_f} avec $G_{\tau_p,\tau_f} = \frac{1}{1-EV_{\tau_p,\tau_f}}$.

Les premiers résultats, présentés sur les figures 3.3 et 3.4, suggèrent que :

- Une stratégie de codage prédictif basée sur l'analyse en ligne des entrées sensorielles auditives et visuelles (sans accès explicite à d'autres espaces de représentation, articulatoire et/ou linguistique) est intéressante uniquement pour une prédiction jusqu'à 100 ms. A 25 ms, un gain de codage de 2.5 est observé (ce qui signifie qu'un encodage par le cerveau de l'erreur de prédiction serait 2.5 plus efficace qu'un encodage de chaque nouvelle observation acoustique). Cependant la précision de la prédiction diminue rapidement (e.g. avec un gain moyen de 1.5 at 50 ms, 1.25 à 75 ms et finalement 1, c'est-à-dire aucun gain autour de 100 ms) (voir figure 3.3).

- L'ajout de la modalité visuelle améliore la prédiction, mais de façon limitée. Le gain maximum est observé pour une prédiction à 75 ms avec un amélioration de la variance expliquée d'à peine 0.1 (voir figure 3.4 (droite)). De plus, et contrairement à ce qui est parfois avancé par certaines études (voir par exemple [Cha+09]), la performance du modèle visuel ne corrobore pas l'hypothèse d'une avancée systématique des lèvres sur le son (la meilleur précision de prédiction à partir de la modalité seule est obtenue pour $\tau_f = 0$ (voir 3.4 (gauche)).

- Les meilleurs prédictions sont obtenues en considérant un contexte passé de 50 ms à la fois pour le modèle auditif et le pour modèle audiovisuel.

Cette première étude propose de quantifier le gain potentiel apporté par une stratégie de codage prédictif de la parole auditive et audiovisuelle. Cependant, les modèles proposés ne font pas appel de façon explicite à des représentations motrices et/ou linguistiques. Or, on peut faire l'hypothèse que ces dernières pourraient améliorer la prédiction). Des pistes pour

[6] Par exemple, $\tau_f = 3$ et $\tau_p = 2$ correspond à la prédiction de l'observation acoustique $x_{t+(25\times3=75ms)}$ à partir des observations acoustiques $[x_t, x_{t-25ms}, x_{t-50ms}]$ et éventuellement visuelle $[I_t, I_{t-25ms}, I_{t-50ms}]$.

[7] Cette base contient les enregistrements audio et vidéo de 59 locuteurs prononçant chacun le même ensemble de 98 phrases. Une procédure de type *K-fold cross validation* est utilisée pour le partitionnement des données en ensembles d'apprentissage, de validation et de test.

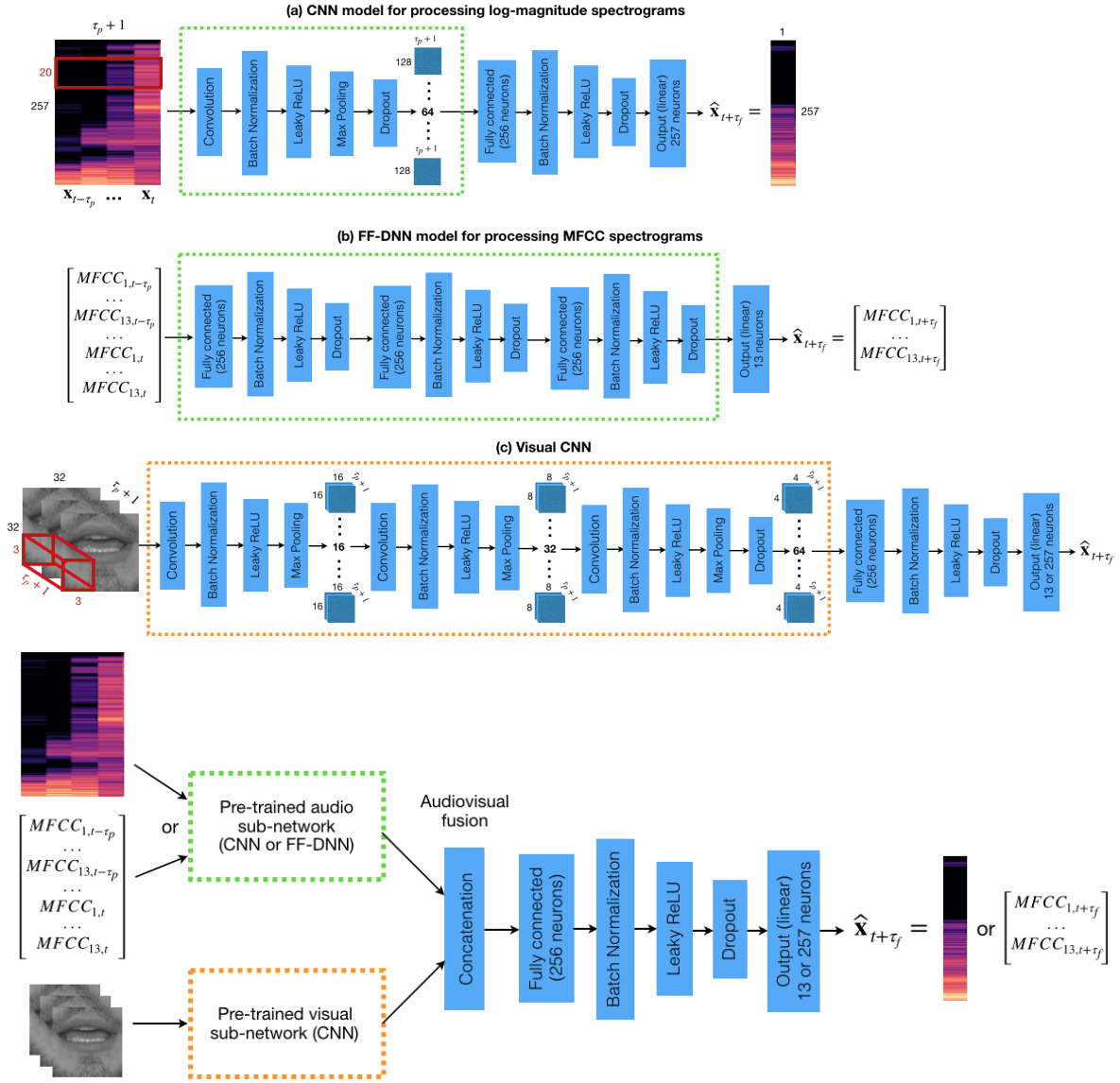


FIGURE 3.2 – Architecture des modèles neuronaux proposés pour mesurer le gain apporté par un codage prédictif de la parole auditive et audiovisuelle. De haut en bas, les deux modèles proposés pour traiter la parole auditive en fonction du type d'observation acoustique considéré (spectre d'amplitude à court terme ou vecteur de coefficients MFCC), le modèle proposé pour traiter la modalité visuelle (seule), et enfin, le modèle proposé pour traiter la parole audiovisuelle, obtenu par pré-entraînement, fusion et ré-apprentissage des modèles auditif (seul) et visuel (seul). Ces différentes architectures sont issues de procédures de sélection de modèles et d'optimisation d'hyperparamètres non détaillées ici.

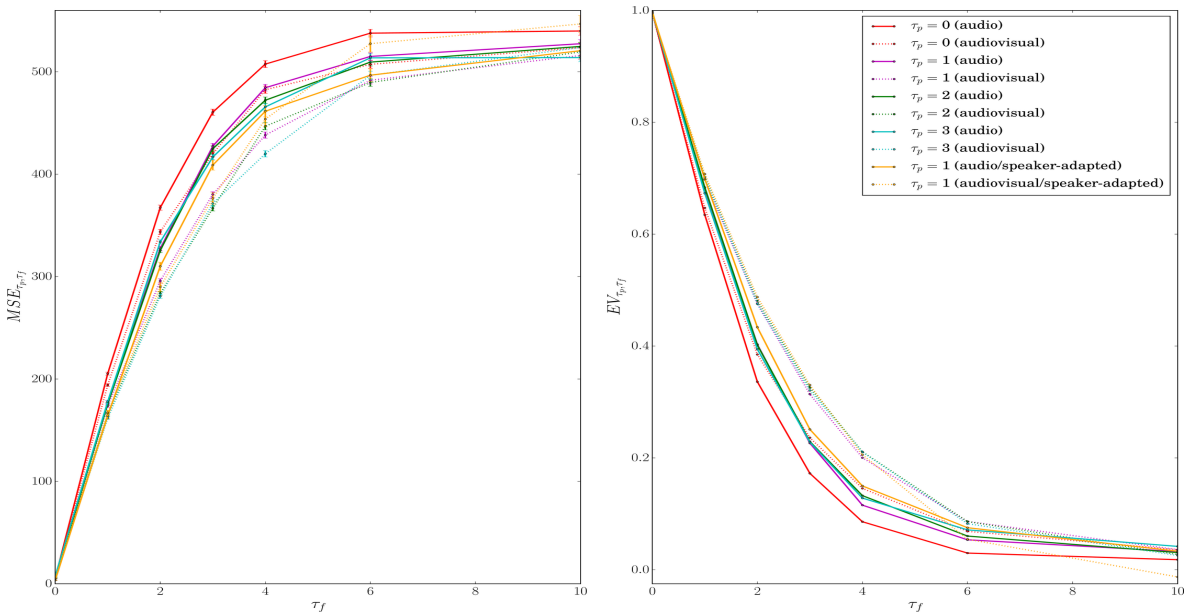


FIGURE 3.3 – Précision de la prédiction à l'aide du modèle auditif (basé sur des observations acoustiques de type MFCC) et du modèle audiovisuel, en termes de MSE_{τ_p,τ_f} (gauche) et de variance expliquée EV_{τ_p,τ_f} (droite), pour différents horizons temporels (τ_f) et différentes tailles de contexte passé (τ_p). Les barres d'erreur représentent les intervalles de confiance à 95%.

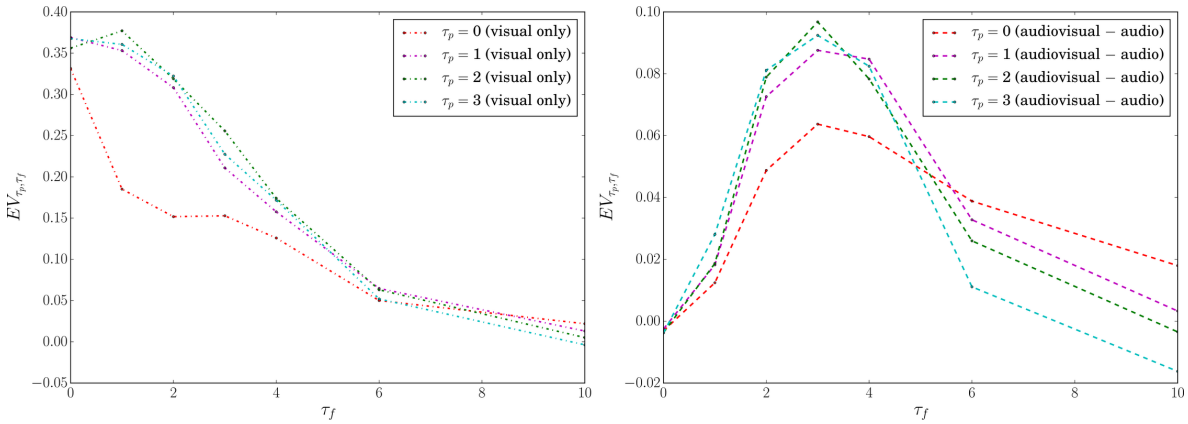


FIGURE 3.4 – Précision de la prédiction à l'aide du modèle visuel uniquement (gauche) et différence de précision entre les modèles audiovisuel et auditif (droite).

l'exploitation de ces espaces de représentation dans le cadre de modèles basés sur les réseaux antagonistes génératifs sont présentées à la section suivante.

3.3 Modèles computationnels de perception/production basés sur les GAN

Les réseaux de neurones génératifs regroupent différentes techniques, basées sur les réseaux de neurones artificiels, utilisables pour apprendre de façon non ou faiblement supervisée un espace de représentation dit "latent" à partir d'un ensemble d'observations, puis générer de nouvelles observations, par échantillonnage de cette espace latent. Deux techniques sont actuellement très étudiées, les auto-encodeurs variationnels ou VAE (brièvement présentés à la page 30) et les réseaux antagonistes génératifs ou GAN, introduits en 2014 par Goodfellow et al. [Goo+14]. L'étude des VAE pour l'apprentissage d'espaces de représentation de textures sonores et la synthèse musicale fait actuellement l'objet de la thèse de Fanny Roche que je co-encadre (voir section 4.5) et qui a donné lieu à une première publication [Roc+19]. Même si les VAE et les GAN seront tous deux étudiés dans le cadre de ce projet de recherche, je me focaliserai dans ce manuscrit sur les GAN, par soucis de concision. Les bases théoriques des GAN sont brièvement rappelées dans la section suivante.

3.3.1 Principe général d'un GAN

Un GAN est un modèle constitué de deux réseaux de neurones (potentiellement convolutifs, récurrents, etc.) : un générateur G et un discriminateur D [Goo+14]. G produit une observation $\mathbf{y}$ (par exemple une image) à partir d'une observation $\mathbf{z}$ tirée d'une distribution aléatoire. D est un classifieur binaire qui détermine si l'observation qui lui est présentée est issue du générateur G ou bien si elle est issue (de la distribution) des données d'apprentissage. Les paramètres de G sont optimisés dans le but de faire augmenter le taux d'erreur de classification de D (en produisant des observations qui se rapprochent de plus en plus de la distribution réelle des données d'apprentissage), tel que :

$$\min_G V(G) = \min_G(\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]) \tag{3.3}$$

et le discriminateur est entraîné à distinguer une observation issue des données d'apprentissage, d'une observation générée par G, tel que :

$$\max_D V(D) = \max_D(\mathbb{E}_{\mathbf{y}\sim p_{\text{data}}\,(\mathbf{y})}[\log D(\mathbf{y})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}(\mathbf{z})[\log(1 - D(G(\mathbf{z})))]) \tag{3.4}$$

On parle donc de réseaux antagonistes car leur apprentissage peut être vu comme une compétition pendant laquelle chacun des deux réseaux joue contre l'autre. En combinant les équations 3.4 et 3.4, l'apprentissage d'un GAN revient à optimiser un problème de type *minimax*, tel que :

$$\min_G \max_D V(D,G) = \min_G \max_D (\mathbb{E}_{\mathbf{y}\sim p_{\text{data}}(\mathbf{y})}[\log D(\mathbf{y})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))) \quad (3.5)$$

Dans sa formulation originale, les observations $\mathbf{z}$ fournies au générateur sont tirées d'une distribution aléatoire. Dans une extension des GAN intitulée C-GAN (pour *conditionnal GAN*) proposée par Mirza et Osindero en 2014 [8], la génération d'une observation $\mathbf{y}$ est conditionnée par une autre observation $\mathbf{x}$, issue d'un autre espace de données (structuré). L'apprentissage d'un C-GAN revient à optimiser :

$$\min_G \max_D V(D,G) = \min_G \max_D (\mathbb{E}_{\mathbf{y}\sim p_{\text{ data }}(\mathbf{y})}[\log D(\mathbf{y},\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z},\mathbf{x})))) \quad (3.6)$$

De façon importante, et contrairement à un réseau de neurones standard, l'apprentissage du générateur d'un C-GAN ne fait pas intervenir un calcul d'erreur entre chaque observation $\mathbf{y}$ générée à partir de $\mathbf{x}$ (et $\mathbf{z}$) d'une part, et une vérité terrain d'autre part. Autrement dit, le générateur ne "voit" pas jamais les données d'apprentissage $\mathbf{y}$, mais uniquement la réponse du discriminateur à ses prédictions. Cette propriété peut être intéressante dans notre cas si on considère une perception de la parole basée sur un mécanisme d'accès à des représentations motrices, qui s'est construit au cours du développement sans avoir un accès aux trajectoires articulatoires correspondant aux stimuli auditifs perçus. Ce point est développé dans la section suivante.

3.3.2 Approche par Cycle-GAN

Nous décrivons ici la possible première brique d'un modèle computationnel de la perception (motrice ou perceptuo-motrice) de la parole, basé sur les GAN. Nous modélisons ici le processus d'accès aux représentations motrices, dans l'espace articulatoire de l'agent, par le générateur d'un C-GAN ($G_{X\to Y}$). Le modèle proposé est présenté à la figure 3.5. Les paramètres du discriminateur D_Y sont estimés à partir d'une base de données acoustico-articulatoires, appelée ici "répertoire sensori-moteur", acquise soit *in vivo* par EMA, soit *in vitro* à l'aide d'un synthétiseur articulatoire (par exemple *VLAM* [Mén+07] ou *VocalTractLab* [BJK06]). L'utilisation d'un synthétiseur articulatoire permet notamment de contrôler la structure de la base de données acoustico-articulatoires sur laquelle sont estimés les paramètres du discriminateur. Pour améliorer cette densité (couverture phonétique), il pourra être intéressant d'intégrer pendant la phase d'apprentissage une phase d'exploration par l'agent de son conduit vocal. Différentes stratégies d'exploration seront étudiées, de la plus simple comme une exploration aléatoire, à des stratégies plus évoluées, comme l'apprentissage "par curiosité" [MNO14].

Par ailleurs, le modèle proposé fait intervenir une étape de normalisation du stimulus auditif perçu qui a pour but de le rendre "compatible" avec l'espace acoustique de l'agent. Ce module pourra prendre différentes formes, comme par exemple une normalisation de type VTLN

[8] *Preprint* disponible sur `https://arxiv.org/pdf/1411.1784.pdf`

[LR98], ou bien l'extraction d'un *embedding* acoustique à l'aide d'un système de reconnaissance automatique de la parole multi-locuteur (basé par exemple sur des réseaux convolutionnels [Abd+14]).

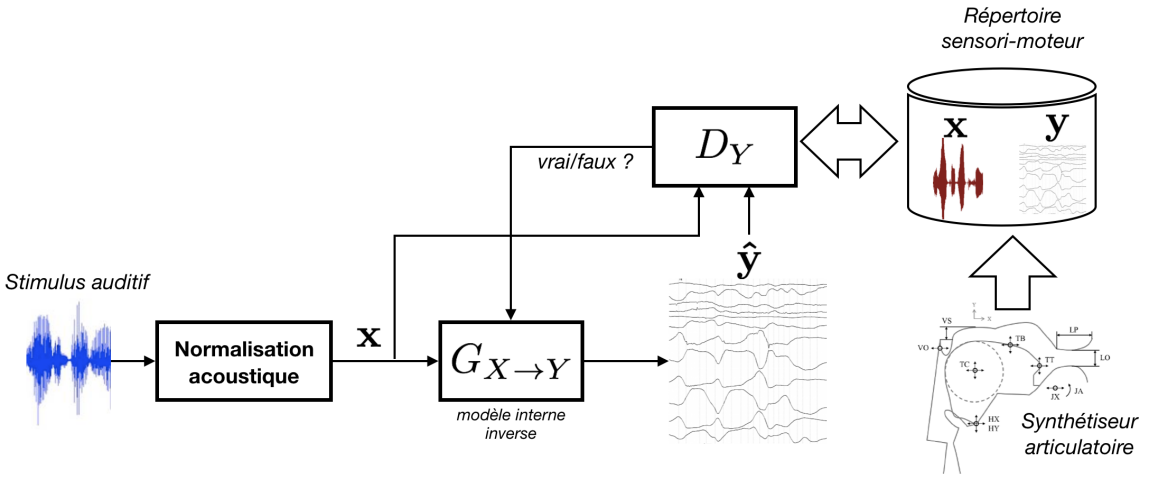


FIGURE 3.5 – Modélisation, à l'aide d'une architecture de type C-GAN, de l'accès à des représentations motrices pendant la perception d'un stimuli auditif.

Ce modèle de perception sera ensuite étendu à un modèle de production, afin de permettre à l'agent de renforcer ses représentations sensori-motrices par imitation des stimuli auditifs qu'il perçoit. Une autre extension des GAN, intitulée Cycle-GAN et proposée en 2017 par Zhu et coll. [Zhu+17] sera étudiée dans ce cadre. Brièvement, un Cycle-GAN est constitué de deux GANs, dont les générateurs sont mis en cascade. Le premier générateur $G_{X\to Y}$ effectue la conversion d'une observation $\mathbf{x}$ (ici acoustique) en une observation $\mathbf{y}$ (ici articulatoire), le second $G_{Y\to X}$ effectue la transformation inverse. Pour l'apprentissage d'un Cycle-GAN, un terme supplémentaire est ajouté à la fonction de coût d'un C-GAN (équation 3.6). Ce terme intitulé *Cycle Consistency Loss* et noté $\mathcal{L}_{\text{cyc}}$ garantit la cohérence d'un cycle complet de conversion, c'est-à-dire $G_{Y\to X}(G_{X\to Y}(\mathbf{x})) = \mathbf{x}$ (dans notre contexte, il s'agit de forcer l'agent à reproduire le stimulus auditif perçu). Plus formellement, il est défini par l'équation suivante :

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{\mathbf{x}\sim p_{\text{ data}}(x)}||G_{Y\to X}(G_{X\to Y}(\mathbf{x})) - \mathbf{x}||_1 \\ + \mathbb{E}_{\mathbf{y}\sim p_{\text{ data}}(\mathbf{y})}||G_{X\to Y}(G_{Y\to X}(\mathbf{y})) - \mathbf{y}||_1\end{aligned} \tag{3.7}$$

Le différent générateurs et discriminateurs impliqués dans ce modèle sont entraînés de façon conjointe en optimisant :

$$\begin{aligned}\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G_{X\to Y}, D_Y, X, Y) \\ + \mathcal{L}_{\text{GAN}}(G_{Y\to X}, D_X, Y, X) \\ + \lambda\mathcal{L}_{\text{cyc}}(G_{X\to Y}, G_{Y\to X})\end{aligned} \tag{3.8}$$

avec $\mathcal{L}_{\text{GAN}}(G_{X\to Y}, D_Y, X, Y)$ la fonction de coût d'un C-GAN définie à l'équation 3.6 (et λ un hyperparamètre à ajuster empiriquement).

Le modèle proposé, basé sur cette architecture, est illustré à la figure 3.6.

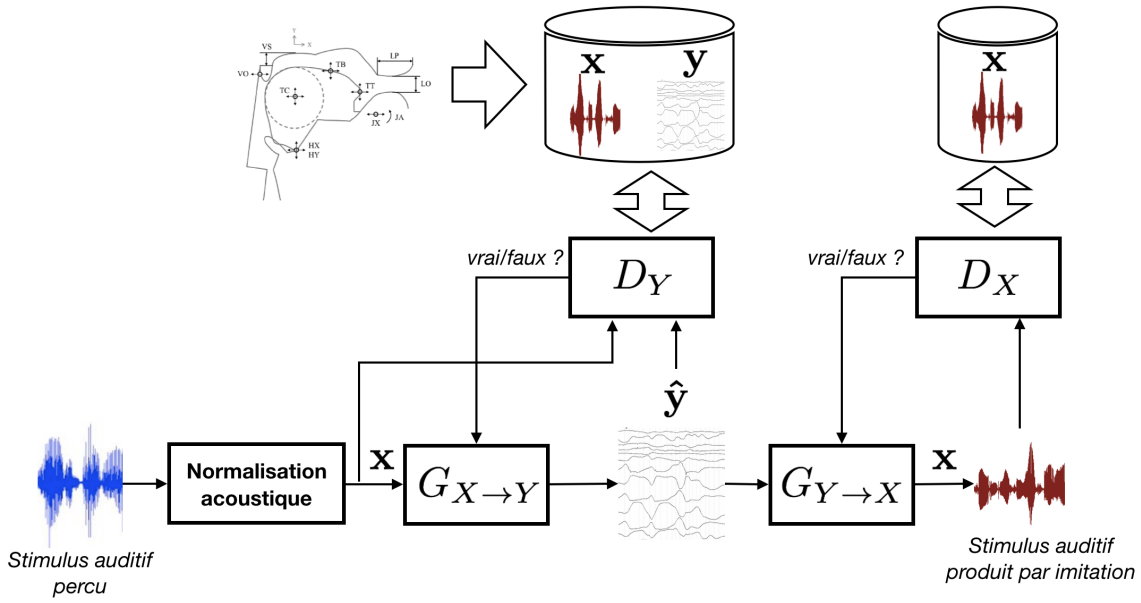


FIGURE 3.6 – Modélisation computationnelle d'un apprentissage sensori-moteur de la parole par imitation, basée sur un réseau de neurones génératif de type Cycle-GAN.

Ce modèle pourrait de plus être utilisé pour simuler une interaction communicationnelle avec un locuteur "maître" qui fournirait un stimulus auditif et indiquerait à l'agent si l'imitation qu'il en fait est correcte ou non. Cette information pourrait guider à son tour l'exploration de l'espace sensori-moteur (D_Y) et permettre un meilleur apprentissage des relations sensori-motrices.

3.4 Restauration de la parole pathologique

Je propose d'utiliser le modèle de perception/production basé sur les réseaux génératifs, présenté à la section précédente, à la restauration automatique des productions acoustiques d'une personne présentant un trouble de la voix et/ou de la parole (parole dysarthrique[9], dysphonique, etc.). On considère qu'un enregistrement audio d'une parole pathologique est constitué de segments intelligibles et de segments peu ou non-intelligibles (phonème mal-prononcé, baisse soudaine d'intensité, etc.). On considère de plus que la taille de ces segments "corrompus" est variable, et qu'elle peut être de l'ordre d'un phonème, d'une syllabe, voire d'un mot entier.

En faisant l'hypothèse qu'un humain décode ce type de production en faisant appel à différents mécanismes de suppléance mentale, et que ces mécanismes se basent sur des représentations acoustiques, motrices et linguistiques, je souhaite développer une technique de res-

[9]La dysarthrie est classiquement définie comme un trouble de la réalisation motrice de la parole, qui fait suite à des lésions du système nerveux central et/ou périphérique [DAB75].

tauration basée sur une prédiction du contenu acoustique "attendu" d'un segment corrompu, exploitant ces différentes représentations, puis son remplacement par un signal synthétique. Cette approche est inspirée du paradigme dit d'*inpainting* utilisé en vision par ordinateur et qui consiste à remplacer automatiquement les parties manquantes d'une image, en s'appuyant sur des informations contextuelles, à la fois locale (texture des segments voisins) et globale (sémantique de l'image). On parle ici de *speech inpainting* car on cherche à remplacer un segment de parole non-intelligible par un signal de synthèse, construit sur la base des segments intelligibles adjacents (passés et/ou futurs).

L'utilisation du paradigme d'*inpainting* pour la restauration d'un signal audio a été proposée pour la première fois par Adler et coll. en 2012 [Adl+12]. Cette étude vise à supprimer des distorsions du signal audio (*declipping*) dont la longueur varie entre 1 et 10ms. A ma connaissance, une seule autre étude propose une technique pour restaurer des segments de parole plus longs (de plusieurs secondes). Il s'agit de l'étude de Prablanc et coll. en 2016 [Pra+16]. Dans cette étude, les limites de la zone à traiter ainsi que le texte associé à cette zone sont connus. Un synthétiseur TTS, combiné avec un système de conversion de voix, est utilisé pour générer la parole manquante. L'approche proposée présente certains points communs avec cette dernière étude (notamment pour la combinaison "TTS/conversion de voix") mais se distingue par : (i) l'absence de connaissances a priori sur les limites de la zone à traiter et le texte associé, (ii) une prédiction du contenu acoustique basée sur l'exploitation conjointe de représentations acoustiques, articulatoires et linguistiques, (iii) une évaluation dans le cas de la parole pathologique.

Comme illustré à la figure 3.7, l'approche proposée fait intervenir 3 étapes :

1. La détection des segments "corrompus" dans un enregistrement audio de parole pathologique : la technique de détection mise en œuvre dépendra du type de trouble de la parole. Dans le cas d'une hypophonie, on cherchera par exemple à détecter les zones présentant une baisse anormale de l'intensité. Dans le cas d'un trouble de l'articulation impliquant une prononciation déviante d'un phonème, on pourra utiliser une mesure de type (*goodness of pronuncation*, GOP) issue d'un décodeur acoustico-phonétique [LFM15 ; Dud+18].

2. La prédiction d'une séquence d'observations acoustiques $\hat{\mathbf{x}}_{t,\tau_f} = [\hat{\mathbf{x}}_t, ..., \hat{\mathbf{x}}_{t+\tau_f}]$ (en reprenant les notations de la section 3.2) pour le segment corrompu détecté à l'étape 1, à partir des observations acoustiques des segments intelligibles passés et/ou futurs (par soucis de simplification, nous ne considérons ici que les τ_p dernières observations notées $\mathbf{x}_{t,\tau_p}$) : c'est ici qu'intervient un modèle (de codage) prédictif qui pourra exploiter des représentations acoustiques, articulatoires et linguistiques.

3. La synthèse d'un nouveau segment acoustique qui remplacera le segment corrompu, à partir de la séquence d'observations acoustiques prédite à l'étape 2. J'envisage ici une synthèse à l'aide d'un *vocoder* neuronal de type *Wavenet* [Kob+17], conditionné par la séquence d'observations acoustiques prédite [She+18], et adapté à la voix du locuteur traité [Liu+18a].

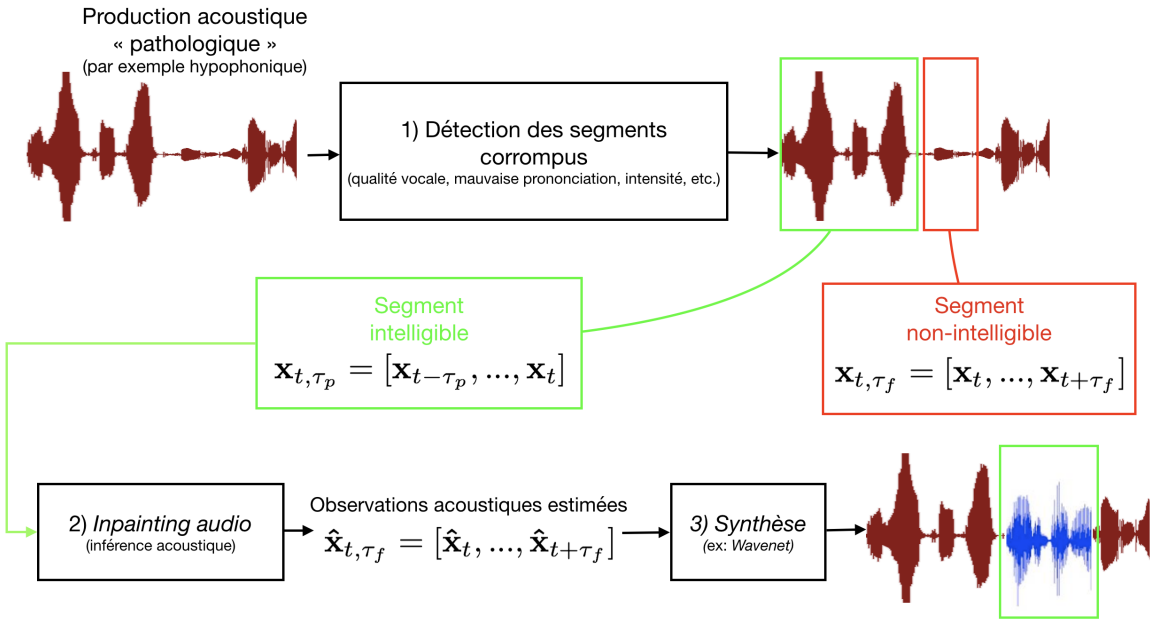


FIGURE 3.7 – Méthode proposée pour le rehaussement de la parole pathologique, basée sur un paradigme de *speech inpainting*.

Je présente à la figure 3.8 l'approche envisagée pour l'étape de prédiction acoustique (étape 2). L'objectif est de prédire les τ_f futures observations acoustiques (notées $\mathbf{x}_{t,\tau_f}$), à partir des τ_p dernières (notées $\mathbf{x}_{t,\tau_p}$). Les trajectoires articulatoires associées, notées $\hat{\mathbf{y}}_{t,\tau_p}$, sont obtenues à l'aide du modèle de type Cycle-GAN proposé à la section 3.3.2. Une représentation linguistique est obtenue par décodage au niveau lexical et/ou phonétique de $\mathbf{x}_{t,\tau_p}$ à l'aide d'un système de reconnaissance automatique de la parole. Une estimation du mot ou des phonèmes suivants $\hat{\mathbf{w}}_{t,\tau_f}$/ $\hat{\mathbf{p}}_{t,\tau_f}$ les plus probables est réalisée à l'aide d'un modèle de langage, basé *n-gram* ou par plongement lexical (technique *word2vec*, brièvement décrite à la page 40) [10]. Le problème de l'inférence acoustique, avec accès à ces différents espaces de représentation s'écrit $\hat{\mathbf{x}}_{t,\tau_f} = f(\mathbf{x}_{t,\tau_p}, \hat{\mathbf{y}}_{t,\tau_p}, \hat{\mathbf{w}}_{t,\tau_f}, \hat{\mathbf{p}}_{t,\tau_f})$. De façon similaire à nos premiers travaux sur la prédiction acoustique par réseaux de neurones décrits à la section 3.2, la fonction f pourra être approchée par un DNN ou un CNN. L'utilisation d'un GAN est également envisagée. Ce dernier serait conditionné par les observations acoustiques $\mathbf{x}_{\mathbf{t},\tau_\mathbf{p}}$, articulatoires $\hat{\mathbf{y}}_{\mathbf{t},\tau_\mathbf{p}}$, ainsi que par le mot ou les phonèmes suivants les plus probables ($\hat{\mathbf{w}}_{t,\tau_f}$ et $\hat{\mathbf{p}}_{t,\tau_f}$), et sera entraîné à prédire les observations acoustiques futures $\hat{\mathbf{x}}_{t,\tau_f}$. Bien que de nombreux problèmes "techniques" doivent être résolus avant son implémentation concrète (relatifs par exemple à la difficulté connue d'entrainer des modèles de type GAN, qui n'a pas été détaillée ici), l'architecture proposée est un nouveau pas vers la conception de systèmes artificiels de perception/production de la parole qui exploitent explicitement des représentations sensori-motrices apprises de façon non ou faiblement supervisée.

[10] Cette estimation pourrait également être réalisée de façon implicite par le module d'inférence acoustique qui serait alors conditionné par les mots précédents $\hat{\mathbf{w}}_{t,\tau_p}$

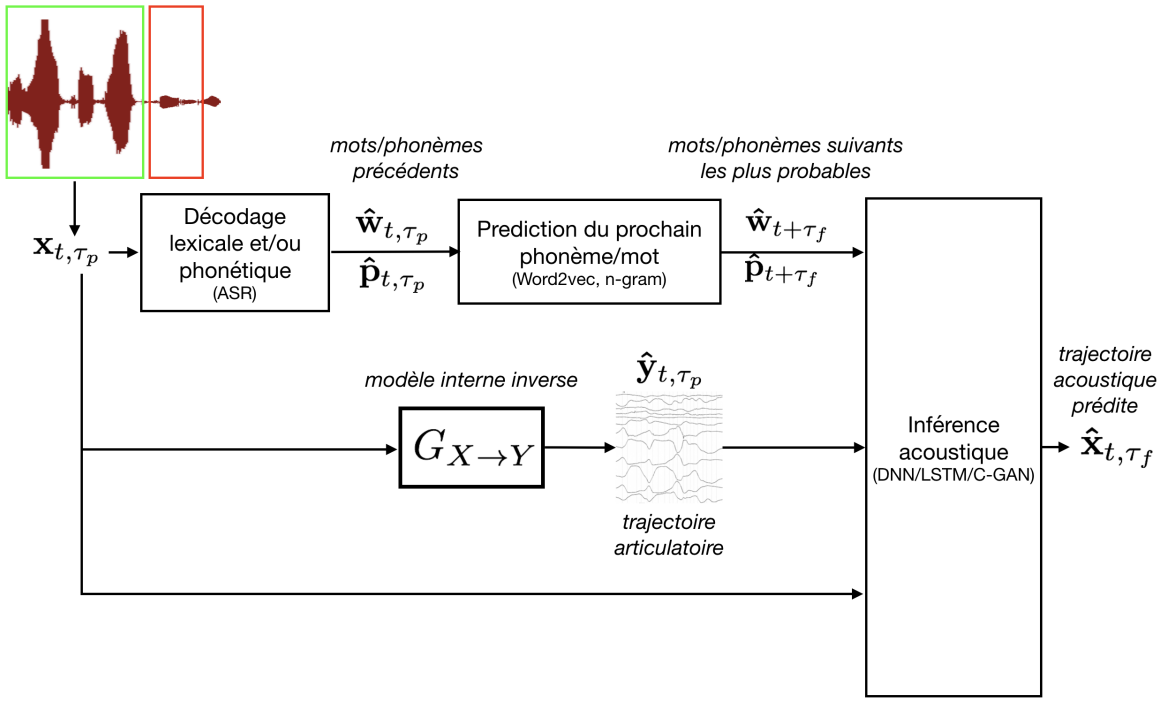


FIGURE 3.8 – Prédiction acoustique avec accès à des représentations articulatoires et linguistiques pour le *speech inpainting*.

Chapitre 4

Curriculum vitæ détaillé

Sommaire

4.1 Etat civil

Nationalité Française, 36 ans (né le 09/11/1982 à Cannes), Marié, 2 enfants (Clément et Manon Hueber)

4.2 Expériences professionnelles

- 2011–2017 : Chargé de Recherche CNRS (CR1 depuis 2015, section 07, GIPSA-lab, UMR 5216, Grenoble)

  - Co-responsable de l'équipe *CRISSP: Cognitive Robotics, Interactive Systems and Speech Processing* (depuis octobre 2014)
  - Enseignant vacataire à l'ENSIMAG et à PHELMA (Grenoble-INP)
- 2010-2011 : Chercheur post-doctorant, GIPSA-lab, Grenoble
  - Travaux de recherche en modélisation acoustico-articulatoire (projet ANR Artis)
- 2006-2009 : Chercheur doctorant, ESPCI/Telecom ParisTech, Paris
  - Travaux de recherche au laboratoire SIGMA/ESPCI ParisTech et au LTCI/Telecom ParisTech. Sujet de thèse : "Reconstitution de la parole à partir d'image ultrasonore et video de l'appareil vocal, vers une communication parlée silencieuse", thèse dirigée par Bruce Denby (ESPCI-ParisTech, Pr. à Univ. Pierre et Marie Curie) et Gérard Chollet (DR CNRS, LTCI Telecom ParisTech).
  - Moniteur en Electronique à l'Université Pierre et Marie-Curie.
- 2005 : Année de césure - Ingénieur, IRCAM, Paris, Synthèse sonore par concaténation d'unités pour la parole et la musique.

4.3 Formation

- 2006-2009 : Doctorat en Informatique, Université Pierre et Marie Curie, Paris
- 2005-2006 : Master Recherche « Traitement du Signal et de l'Image », INSA Lyon, mention très bien
- 2006 : Diplôme d'Ingénieur en Electronique, CPE Lyon
- 2000-2002 : Classes préparatoires mathématiques supérieures et spéciales, Lycée Masséna, Nice (MPSI & PSI)
- 2000 : Baccalauréat scientifique, Lycée Carnot, Cannes, mention très bien

4.4 Distinctions

- 2015 : *Best paper award* décerné par *European Association for Signal Processing* (EURASIP) et *International Speech Communication Association* (ISCA)
- 2011 : 6ème Prix Christian Benoît décerné par la *International Speech Communication Association* (ce prix international récompense un jeune chercheur pour ces travaux accomplis dans le domaine des sciences de la communication parlée ainsi que pour son projet de recherche)
- 2008 : *Best Student Paper Award* dans la catégorie *Speech Communication* décerné lors du *Meeting of the Acoustical Society of America*

4.5 Activités d'encadrement

4.5.1 Thèses de doctorat

4.5.1.1 Thèses en cours

- **Fanny Roche**, "Apprentissage automatique pour la génération de textures sonores", thèse en cours (débutée le 01/12/2016), co-encadrée avec L. Girin (GIPSA-lab, directeur) et Maëva Garnier (GIPSA-lab), taux d'encadrement 25%, Samuel Limier (Arturia), financement : contrat CIFRE avec l'entreprise Arturia, publication associée : [Roc+19].
- **Gaël le Godais**, "Développement de méthodes d'apprentissage automatique pour la synthèse de la parole pilotée par l'activité cérébrale", thèse en cours (débutée le 01/03/2017), co-encadrée avec B.Yvert (INSERM, directeur) et L. Girin (GIPSA-lab), taux d'encadrement 50%, financement : projet ANR *BrainSpeak*.
- **Lorenzo Dall'amico**, "Analyse des performances de l'apprentissage automatique : une approche à la frontière entre matrices aléatoires et physique statistique, et application au traitement de la parole", thèse en cours (débutée le 01/09/2018), co-encadrée avec M. Romain Couillet (GIPSA-lab, directeur), taux d'encadrement 20%, financement : bourse ministérielle ED EEATS.

4.5.1.2 Thèses soutenues

- **Maël Pouget**, Synthèse incrémentale de la parole à partir du texte, thèse débutée le 01/10/2013 et soutenue le 23/06/2017, co-encadrée avec Gérard Bailly (GIPSA-lab, directeur), taux d'encadrement 60%, publications associées : [Pou+15 ; Pou+16], financement : bourse doctorale ministérielle fléchée (ED EEATS)
- **Florent Bocquelet**, "Interface cerveau-machine pour la restauration de la parole", thèse débutée le 01/02/2014 et soutenue le 23/04/2017, co-encadré avec B.Yvert (INSERM, directeur), L.Girin (GIPSA-lab), taux d'encadrement 30%, financement : bourse ministérielle (ED EDISCE), publications associées : [Boc+16a ; Boc+14 ; Boc+15b ; Boc+15a ; Boc+16b ; Boc+16a]
- **Diandra Fabre**, "Système de retour visuel des mouvements articulatoires pour l'orthophonie", thèse débutée le 01/10/2013 et soutenue le 16/12/2016, co-encadrée avec Pierre Badin (GIPSA-lab, directeur), Mélanie Canault et Nathalie Bedoin (Laboratoire Dynamique du Langage, Lyon), taux d'encadrement : 40%, publications associées : [FHB14 ; Fab+15 ; Fab+17 ; Fab+16], financement : bourse région Rhône-Alpes (ARC 6)

4.5.2 Chercheurs post-doctorants

J'ai été responsable scientifique de deux chercheurs post-doctorants :

- **Eric Tatulli**, "Réseaux de neurones à convolution pour la modélisation de la parole audiovisuelle", octobre 2015 - juin 2017, financements :projet *Living Book of Anatomy* (Labex Persyval) et *ERC SpeechUnit(e)s de Jean-Luc Schwartz* (voir section 4.10), publication associée : [TH17]

- **Olha Nahorna**, "Synthèse vocale incrémentale", octobre 2014 - octobre 2015, financement : Projet AGIR (pôle MSTIC, UGA) *SpeakRightNow*, publication associée : [Pou+16]

4.5.3 Mémoire de *Master* Recherche et Projets fin d'études

- **Marion Girod-Roux**, Master Recherche Science du Langage, Université Paris Ouest Nanterre), "Retour articulatoire visuel par échographie pour la rééducation des troubles de la production liés à une glossectomie", juin 2016-juin 2017

- **Xiaoou Wang** : Master Recherche Phonétique, Université Aix-Marseille, "Développement et évaluation d'un protocole d'aide à l'apprentissage du français par des chinois, basé sur une tête parlante virtuelle", avril-septembre 2013, publication associée : [WHB14]

- **Maël Pouget**, Projet fin d'études, PHELMA/Grenoble-INP, "Implémentation temps-réel d'une interface de communication en parole silencieuse", février-juillet 2012

4.5.4 Mémoires d'orthophonie

- **Camille Bach et Lorene Lambourion** (binôme), Ecole d'orthophonie de Lyon (ISTR, Université Claude Bernard Lyon 1), "Impact de la visualisation du geste articulatoire, acquis par imagerie ultrasonore, pour la rééducation du trouble phonologique chez l'enfant", septembre 2013-2014

4.5.5 Autres

Stages encadrés au GIPSA-lab (par soucis de concision, les stages encadrés avant ma nomination au CNRS ne seront pas listés ici) :

- **Gina Yang**, Stage 2ème année PHELMA/Grenoble-INP (3 mois), "Segmentation des images ultrasonores par modèles actifs", 2014

- **Remi Vincent**, Stage 2ème année PHELMA/Grenoble-INP (3 mois), "Codage Harmonique+Bruit pour la synthèse de la parole par HMM", 2011

- **Nicolas Asin**, Stage 2ème année, License Statistique, Université de Besançon (3 mois), "Synthèse Text-to-speech du francais par HMM à partir de livres audio", 2011

4.6 Activités d'enseignements

Mes activités d'enseignement sont détaillées à la table 4.1 (par soucis de concision, les enseignements effectués à l'université Paris VI dans le cadre d'un monitorat, avant ma nomination au CNRS, ne seront pas listés ici).

TABLE 4.1 – Activités d'enseignement sur la période 2011-2019

Intitulé	**Établissement**	**Filière**	**Niveau**	**Type**	**Volume**	**Période**	**Contenu**
Bayesian methods in signal and image processing	Univ. Grenoble-Alpes	M2 SIGMA	Bac+5	CM	4h/an	2017-2019	Modèles GMM et HMM, application à la reconnaissance automatique de la parole
Traitement signal temps-réel	PHELMA	SICOM	Bac+5	CM/TP	20h/an	2014-2019	Concepts principaux des systèmes temps-réel (modèles théoriques, aspects pratiques liés au traitement audio sur un OS non-temps-réel, buffers circulaires, filtrage en temps-réel par *overlap-add*, etc.). Implémentation d'effets audio temps-réel de type *reverb* à convolution et compresseur.
Parole et langage	ENSIMAG	MMIS	Bac+5 (3A)	CM	9h/an	2014-2017	Analyse, reconnaissance, synthèse et conversion de la parole.
Traitement de la parole	PHELMA	SICOM	Bac+4 (2A)	TP	24h/an	2014-2015	Débruitage d'un signal de parole à l'aide du filtre de Wiener et d'un modèle LPC, transformation vocale basée sur la technique TD-PSOLA.
Projets d'électronique	PHELMA	SICOM	Bac+3 (1A)	TP	14h/an	2011-2012	Co-encadrement de plusieurs groupes d'étudiants sur des projets autour de l'utilisation de matrices de capteurs ultrasonores pour l'étude des mouvements labiaux.
Communication augmentée	Ecole doctorale EEATS		Doctorants	CM	3h/an	2012-2013	Reconnaissance et synthèse audiovisuelle de la parole.

4.7 Activité éditoriale, organisation de conférences

- **Co-édition d'un numéro spécial dans la revue *IEEE/ACM Trans. Audio Speech and Language Processing (TASLP)* (2017)** intitulé *Biosignal-based Spoken Communication* avec T. Schultz (Université de Breme, Allemagne) D. Krusienski (Old Dominion University, USA), et J. Brumberg (Kansas University, USA)
- **Organisation d'une session spéciale à la conférence internationale Interspeech (2017)** intitulée *Incremental Processing and Responsive Behaviour*, avec Timo Baumann (Carnegie Mellon University, USA) et David Schlangen (Université de Bielefeld, Allemagne)
- **Co-organisation de la conférence SLaTE 2013** (*Speech and Language Technology in Education*, satellite de la conférence *Interspeech* 2013), avec Pierre Badin, Gérard Bailly (GIPSA-lab) et Francoise Raby (Lidilem, Grenoble), qui s'est tenue du 30/08/2013 au 01/09/2013 à Grenoble, a accueilli environ 80 participants. Responsable du processus de soumission et de relecture des articles, de la conception des actes, et des événements sociaux, membre du comité de relecture.

4.8 Travaux d'expertise

- **Expertise de projets pour l'Agence Nationale de la Recherche (ANR)** - 2018-2019 dans le cadre de l'appel "CE23 Données, Connaissances, Contenus, Big Data, Simulation numérique, HPC"
- **Membre du jury du concours INRIA** (CR2-CR1, Centre Nancy) en 2015.
- **Examinateur** de la thèse de Mme. Aurore Hakun, soutenue le 5 septembre 2016 à l'ESPCI ParisTech.
- **Membre du comité de programme** des Journées d'Etudes sur la Parole (JEP) 2012
- **Membre régulier du jury du prix Christian Benoît** (décerné par ACB/AFCP/ISCA) depuis 2012.
- **Relecteur pour les revues** internationales *IEEE/ACM Trans. Audio Speech and Language Processing* (IEEE), *Computer Speech and Language* (Elsevier), *Speech Communication* (Elsevier) et *Journal of Acoustical Society of America*
- **Relecteur pour plusieurs conférences** internationales dont *ICASSP*, *Interspeech* et *ISSP*

4.9 Responsabilités, management de la recherche

- **Co-responsable de l'équipe CRISSP** (2014-à ce jour) (*Cognitive Robotics, Interactive Systems and Speech Processing*, anciennement équipe MAGIC) (avec Gérard

Bailly[1]). Cette équipe regroupe 5 chercheurs CNRS (2 DR, 3 CR), 2 ingénieurs de recherche CNRS, 1 enseignant-chercheur (Pr. Grenoble-INP) ainsi que plusieurs chercheurs post-doctorants et doctorants. Je m'occupe de tous les aspects liés à la vie de l'équipe (budget, animation des réunions, entretien annuel, lien avec la direction, etc.).

- **Animateur du groupe de travail *MLSpeech*** (2013-2015) : J'ai créé puis animé un groupe de travail intitulé *MLSpeech*, regroupant permanents et non-permanents du GIPSA-lab autour de la modélisation par apprentissage (*machine learning*) et de son utilisation pour le traitement automatique de la parole.

- Membre de la commission informatique du GIPSA-lab (2014-2015)

- Responsable de l'équipe séminaire du Département Parole et Cognition du GIPSA-lab (2011-2013) (équipe composé d'un chercheur permanent et de plusieurs doctorants, en charge de l'organisation d'une vingtaine de séminaires par an).

4.10 Activité contractuelle et responsabilités dans des projets de recherche

Je présente ici, par ordre chronologique, les différents contrats de recherche auxquels j'ai participé depuis mon recrutement au CNRS.

- **Ultraspeech II** : 2011-2013, GIPSA-lab, porté par Thomas Hueber, projet financé par le prix Christian Benoît, budget 7.5k€, implémentation temps-réel de la chaine de traitement d'une interface de communication en parole silencieuse basée sur l'imagerie ultrasonore et video du conduit vocal (voir section **??**).

- **Vizart3D** : 2012-2014, GIPSA-lab, porté par Thomas Hueber, financé par le pôle CSVSB de l'université Joseph Fourier, budget 10k€, développement d'un prototype temps-réel de retour visuel articulatoire (voir section 2.2.2.2).

- **Living book of anatomy** : 2013-2016, porté par Jocelyne Troccaz (TIMC), financé par le Labex Persyval-lab, partenaires : laboratoires TIMC, GIPSA-lab, LJK, LIG, INRIA, réalité augmentée pour l'apprentissage de l'anatomie des membres inférieurs et de la sphère orofaciale. Ce projet a permis de financer le post-doctorat d'Eric Tatulli portant sur le décodage robuste de mouvements linguaux à partir d'images échographiques à l'aide de réseaux de neurones à convolution (voir section 1.1.2.2).

- **SpeakRightNow** : 2015-2017, porté par Thomas Hueber et financé par le pôle MSTIC de l'Université Grenoble-Alpes, budget 60k€, financement du post-doctorat d'Olha Nahorna, synthèse vocale incrémentale à partir du texte (voir section 1.3).

[1] Les équipes du GIPSA-lab sont toutes animées par un binôme.

- **Speech Unit(e)s** : 2013-2018, *ERC advanced* de Jean-Luc Schwartz (FP7-IDEAS-ERC, numéro 339152), visant à comprendre comment la parole unifie et intègre les flux d'information sensoriels (auditif, visuel, tactile) et moteurs, et comment les unités de la parole émergent des interactions perceptuo-motrices. Ma contribution porte sur l'utilisation de l'apprentissage profond pour quantifier la quantité d'information prédictible à partir des modalités auditives et visuelles de la parole, dans le cadre des théories du codage prédictif en perception, financement du post-doctorat d'Eric Tatulli.

- **BrainSpeak** : 2016-2020, partenaires : INSERM U1205 (Blaise Yvert, coordinateur), GIPSA-lab (T. Hueber, L. Girin, P. Perrier), CHU Grenoble Alpes (Pr. S. Charbardes, Pr. P. Kahane), financeur : ANR (AAP Technologies pour la santé, ANR-16-CE19-0005) et Fondation pour la recherche médicale (FRM), budget : 845k€. Ce projet porte sur le développement d'une interface cerveau-machine pour la restauration de la parole basée sur la capture de l'activité cérébrale à l'aide de techniques de type ECoG et *micro-electrode array* (MEA, Utah array), son décodage à l'aide de techniques d'apprentissage automatique, et le pilotage en temps-réel d'un synthétiseur vocal (voir section 1.2). Je suis responsable scientifique pour le GIPSA-lab et en charge du lot "*Machine learning algorithms for improved speech synthesis and neural-to-speech conversion*" (WP3).

- **BrainCom** (*Brain high-density cortical implants for cognitive neuroscience*) : 01/12/2016 - 01/12/2021, contrat européen H2020-FETPROACT-2016-2017 numéro 723032, budget : 8,359,862.50 €, partenaires : Catalan Institute of Nanoscience and Nanotechnology (porteur), Agencia Estatal Consejo Superior de Investigaciones Científicas (CSIC), Université Grenoble-Alpes (qui regroupe dans ce projet le BrainTech-lab et le GIPSA-lab), CHU Grenoble Alpes, École des Mines de Paris, MultiChannel Systems (PME), Université de Genève, Université d'Oxford, LMU (Munich, Allemagne), Wavestone (gestion de projet). Ce projet porte sur le développement d'une nouvelle génération d'implants ECoG pour l'enregistrement et la stimulation neuronale chez l'homme. Ces implants sont basés sur l'utilisation de nano-matériaux qui permettent d'en augmenter la densité et la flexibilité, et sont mis en œuvre pour le décodage de l'activité cérébrale.

Projets non-financés

- Projet européen ITN (*International training network*) BASIC et ExoSpeech, soumis en 2011 et 2012, portant sur les interfaces de communication en parole silencieuse et visant à mettre en relation différents acteurs académiques (Université de Karlsruhe, GIPSA-lab, Univ. Sheffield, Univ. Hull.) avec des industriels dont Samsung Europe et la société Tobii Churchill.

- Projet ANR TecSan MaVoix : soumis en 2013, portant sur la personnalisation des voix de synthèses pour les personnes handicapées, et l'interactivité des systèmes TTS (au travers du paradigme de synthèse incrémentale), partenaires : société *Voxygen* (porteur), GIPSA-lab, LPP-HEGP, société Ergonotics.

4.11 Productions scientifiques

4.11.1 Publications

Par soucis de concision, les communications dans des congrès nationaux, les communications courtes (*abstract*) et les communications dans des congrès ou revues sans comité de lecture ne seront pas listées ici [2].

[2]Une liste complète de mes publications est disponible en ligne sur `http://www.gipsa-lab.fr/~thomas.hueber`.

Articles dans des revues internationales avec comité de lecture

[Hal+18] Célise Haldin, Audrey Acher, Louise Kauffmann, Thomas Hueber, Emilie Cousin et al. "Speech recovery and language plasticity can be facilitated by Sensori-Motor Fusion training in chronic non-fluent aphasia. A case report study". In : *Clinical Linguistics and Phonetics* 32.7 (2018), p. 595-621.

[Tre+18] Avril Treille, Coriandre Vilain, Jean Luc Schwartz, Thomas Hueber et Marc Sato. "Electrophysiological evidence for Audio-visuo-lingual speech integration". In : *Neuropsychologia* 109 (2018), p. 126-133.

[Fab+17] Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda et Pierre Badin. "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract". In : *Speech Communication* 93 (2017), p. 63-75.

[GHA17b] Laurent Girin, Thomas Hueber et Xavier Alameda-Pineda. "Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping". In : *IEEE/ACM Transactions on Audio Speech and Language Processing* 25.3 (2017), p. 662-673.

[Sch+17a] T. Schultz, T. Hueber, D. J. Krusienski et J. S. Brumberg. "Introduction to the Special Issue on Biosignal-Based Spoken Communication". In : *IEEE/ACM Transactions on Audio Speech and Language Processing* 25.12 (2017), p. 2254-2256.

[Sch+17b] T Schultz, M Wand, T Hueber, D J Krusienski, C Herff et al. "Biosignal-Based Spoken Communication : A Survey". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (2017), p. 2257-2271.

[Tre+17] Avril Treille, Coriandre Vilain, Thomas Hueber, Laurent Lamalle et Marc Sato. "Inside speech : Multisensory and modality-specific processing of tongue and lip speech actions". In : *Journal of Cognitive Neuroscience* 29.3 (2017), p. 448-466.

[Boc+16a] Florent Bocquelet, Thomas Hueber, Laurent Girin, Stéphan Chabardès et Blaise Yvert. "Key considerations in designing a speech brain-computer interface". In : *Journal of Physiology Paris* 110.4 (2016), p. 392-401.

[Boc+16b] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux et Blaise Yvert. "Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces". In : *PLoS Computational Biology* 12.11 (2016), e1005119.

[HB16] Thomas Hueber et Gérard Bailly. "Statistical conversion of silent articulation into audible speech using full-covariance HMM". In : *Computer Speech and Language* 36 (2016), p. 274-293.

[Hue+15] Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda et Gerard Bailly. "Speaker-Adaptive Acoustic-Articulatory Inversion Using Cascaded Gaussian Mixture Regression". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.12 (déc. 2015), p. 2246-2259.

[Den+10] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, James Gilbert et al. "Silent Speech Interfaces". In : *Speech Communication* 52.4 (2010), p. 270-287.

[Hue+10] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus et al. "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips". In : *Speech Communication* 52.4 (avr. 2010), p. 288-300.

Articles dans des conférences internationales avec comité de lecture

[Roc+19] F. Roche, T. Hueber, S. Limier et L. Girin. "Autoencoders for music sound modeling : a comparison of linear, shallow, deep, recurrent and variational models". In : *Proc. of SMC*. Malaga, Spain, 2019, to appear.

[Liu+18b] L. Liu, T. Hueber, G. Feng et D. Beautemps. "Visual Recognition of Continuous Cued Speech Using a Tandem CNN-HMM Approach". In : *Proc. of Interspeech*. Hyderabad, India : ISCA, 2018, p. 2643-2647.

[GHA17a] Laurent Girin, Thomas Hueber et Xavier Alameda-Pineda. "Adaptation of a Gaussian Mixture Regressor to a New Input Distribution : Extending the C-GMR Framework". In : *Proc. of International Conference on Latent Variable Analysis and Signal Separation*. Springer. Grenoble, France, 2017, p. 459-468.

[TH17] Eric Tatulli et Thomas Hueber. "Feature extraction using multimodal convolutional neural networks for visual speech recognition". In : *Proc. of ICASSP*. New Orleans, LA, USA : IEEE, 2017, p. 2971-2975.

[Pou+16] Maël Pouget, Olha Nahorna, Thomas Hueber et Gérard Bailly. "Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis". In : *Proc. of Interspeech*. San Francisco, CA, USA, 2016, p. 2846-2850.

[Boc+15b] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux et Blaise Yvert. "Real-time control of a DNN-based articulatory synthesizer for silent speech conversion : A pilot study". In : *Proc. of Interspeech*. Dresden, Germany : ISCA, 2015, p. 2405-2409.

[Fab+15] Diandra Fabre, Thomas Hueber, Florent Bocquelet et Pierre Badin. "Tongue tracking in ultrasound images using Eigentongue decomposition and artificial neural networks". In : *Proc. of Interspeech*. Dresden, Germany : ISCA, 2015, p. 2410-2414.

[FHB14] D Fabre, T Hueber et P Badin. "Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression". In : *Proc. of Interspeech*. Singapore : ISCA, 2014, p. 2293-2297.

[Tre+14b] Avril Treille, Coriandre Emmanuel Vilain, Thomas Hueber, Jean-Luc Schwartz, Laurent Lamalle et al. "The sight of your tongue : neural correlates of audio-lingual speech perception". In : *International Seminar on Speech Production (ISSP)*. Koln, Germany, 2014, p. 429-432.

[WHB14] Xiaoou Wang, Thomas Hueber et Pierre Badin. "On the use of an articulatory talking head for second language pronunciation training : the case of Chinese learners of French". In : *Proc. of International Seminar on Speech Production (ISSP)*. Cologne, Germany, 2014, p. 449-452.

[Bar+13] Adela Barbulescu, Thomas Hueber, Gérard Bailly et Rémi Ronfard. "Audio-Visual Speaker Conversion using Prosody Features". In : *Proc. of International Conference on Auditory-Visual Speech Processing (AVSP)*. Annecy, France, 2013, p. 11-16.

[Hue+12b] Thomas Hueber, AB Youssef, Gérard Bailly, Badin Pierre. et Frédéric Elisei. "Cross-speaker Acoustic-to-Articulatory Inversion using Phone-based Trajectory HMM for Pronunciation Training". In : *Proc. of Interspeech*. T. 1. Portland, OR, USA, 2012, p. 3-6.

[Ben+11b] Atef Ben Youssef, Thomas Hueber, Pierre Badin et Gérard Bailly. "Toward a multi-speaker visual articulatory feedback system". In : *Proc. of Interspeech*. Florence, Italy : ISCA, 2011, p. 589-592.

[Cai+11] Jun Cai, Thomas Hueber, Bruce Denby, Elie-Laurent Benaroya, Gérard Chollet et al. "A visual speech recognition system for an ultrasound-based silent speech interface ". In : *Proc. of International Congress of Phonetic Sciences (ICPhS)*. Hong Kong, China, 2011, p. 384-387.

[Den+11b] Bruce Denby, Jun Cai, Thomas Hueber, Pierre Roussel, Gérard Dreyfus et al. "Towards a Practical Silent Speech Interface Based on Vocal Tract Imaging". In : *Proc. of International Seminar on Speech Production (ISSP)*. Montr{é}al, Canada, 2011, p. 89-94.

[Hue+11e] Thomas Hueber, Pierre Badin, Christophe Savariaux, Coriandre Vilain et Gérard Bailly. "Differences in articulatory strategies between silent, whispered and normal speech ? A pilot study using ElectroMagnetic Articulography". In : *International Seminar on Speech Production (ISSP)*. Montréal, Canada, 2011.

[Hue+11f] Thomas Hueber, Elie-Laurent Benaroya, Bruce Denby et Gérard Chollet. "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface." In : *Proc. of Interspeech*. Florence, Italy : ISCA, 2011, p. 593-596.

[Bad+10b] Pierre Badin, Atef Ben Youssef, Gérard Bailly, Frédéric Elisei, Thomas Hueber et al. "Visual articulatory feedback for phonetic correction in second language learning". In : *Speech and Language Technology in Education (SLaTE)*. Tokyo, Japan, 2010, p. 1-10.

[Flo+10] Victoria-M. Florescu, Lise Crevier-Buchman, Bruce Denby, Thomas Hueber, Antonia Colazo-Simon et al. "Silent vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface". In : *Proc. of INTERSPEECH*. Makuhari, Chiba, Japan, 2010, p. 450-453.

[Hue+09] Thomas Hueber, Elie Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus et al. "Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface". In : *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2009, p. 640-643.

[Hue+08a] Thomas Hueber, G. Chollet, B. Denby et M. Stone. "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application". In : *Proc. of International Seminar on Speech Production (ISSP)*. Strasbourg, France, 2008, p. 365-369.

[Hue+08b] Thomas Hueber, Gérard Chollet, Bruce Denby, Gérard Dreyfus et Maureen Stone. "Phone recognition from ultrasound and optical video sequences for a silent speech interface". In : *Proc. of Interspeech*. Brisbane, Australia : ISCA, 2008, p. 2032-2035.

[Hue+08c] Thomas Hueber, Gérard Chollet, Bruce Denby, Gérard Dreyfus et Maureen Stone. "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips". In : *Proc. of Interspeech*. Brisbane, Australia : ISCA, sept. 2008, p. 2028-2031.

[Hue+07c] T Hueber, G Chollet, B Denby, M Stone et L Zouari. "Ouisper : Corpus Based Synthesis Driven by Articulatory Data". In : *International Congress of Phonetic Sciences (ICPHS)*. Saarbrücken - Germany, 2007, p. 2193-2196.

[Bel+06] Grégory Beller, Diemo Schwarz, Thomas Hueber et Xavier Rodet. "Speech Rates in French Expressive Speech". In : *Proc. of Speech Prosody*. Dresden, 2006, p. 672-675.

[Bel+05] Grégory Beller, Diemo Schwarz, Thomas Hueber et Xavier Rodet. "Hybrid Concatenative Synthesis In The Intersection of Speech and Music". In : *Proc. of JIM*. T. 12. 2005, p. 41-45.

Chapitres d'ouvrages

[Ach+16] Audrey Acher, Diandra Fabre, Thomas Hueber, Pierre Badin, Olivier Detante et al. "Retour visuel en aphasiologie : résultats comportementaux, acoustiques et en neuroimagerie". In : *Actes des XVIèmes Rencontres Internationales d'Orthophonie : Orthophonie et technologies innovantes*. Sous la dir. de Silvia Topouzkhanian Nathalie Joyeux. Orthophonie et technologies innovantes. Paris, France : Ortho Edition, 2016. Chap. 12, p. 227-260.

[Fab+16] Diandra Fabre, Thomas Hueber, Mélanie Canault, Nathalie Bedoin, Audrey Acher et al. "Apport de l'échographie linguale à la rééducation orthophonique". In : *XVIèmes Rencontres Internationales d'Orthophonie*. Sous la dir. de Silvia Topouzkhanian Nathalie Joyeux. Orthophonie et technologies innovantes. Paris, France : Ortho Edition, 2016. Chap. 11, p. 199-225.

[Ryb+14] Yves Rybarczyk, Tiago Cardoso, João Rosas, LuisM. Camarinha-Matos, Nicolas D'Alessandro et al. "Reactive Statistical Mapping : Towards the Sketching of Performative Control with Data". In : *Innovative and Creative Developments in Multimodal Interaction Systems*. T. 425. IFIP Advances in Information and Communication Technology. Springer Berlin Heidelberg, 2014, p. 20-49.

[HD09] T Hueber et B Denby. "Analyse du conduit vocal par imagerie ultrasonore". In : *L'imagerie médicale pour l'étude de la parole*. Sous la dir. d'A Marchal et C Cavé. IC2, Hermes Science, 2009, p. 147-174.

[Cho+07] Chollet G., R Landais, T Hueber, H Bredin, C Mokbel et al. “Some Experiments in Audio-Visual Speech Processing”. In : *Advances in Nonlinear Speech Processing*. T. 4885. Springer, 2007, p. 28-56.

Brevet

[Hue+11c] T Hueber, R Dubois, P Roussel, B Denby et G Dreyfus. "Device for reconstructing speech by ultrasonically probing the vocal apparatus". WO/2011/032688. 2011.

Thèse de doctorat/master

[Hue09] Thomas Hueber. "Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal : vers une communication parlée silencieuse". French. PhD thesis. Paris : Université Pierre et Marie Curie - Paris VI, 2009, p. 200.

[Hue06] Thomas Hueber. "Synthèse de la Parole à partir d'Imagerie Ultrasonore et Optique de l'Appareil Vocal". Master thesis. Univ. Claude Bernard - Ecole Centrale - INSA (Lyon), 2006.

4.11.2 Logiciels

- **Ultraspeech** (2008-2017) : Enregistrement simultané et synchrone de données échographiques, vidéo, audio et inertiels. Logiciel utilisé par une dizaine de laboratoires dont Univ. Ottawa (Canada), Tuabjin Univ, (Chine), Max Plank Inst. for Evol. Antropology (Allemagne), Macquarie Univ. (Australie). Mise à disposition de la version 1.3 le 11/05/2016 sur (*www.ultraspeech.com*).
- **Ultraspeech-player** (2014-2016) : Visualisation intuitive des mouvements articulatoires pour la rééducation orthophonique et l'apprentissage d'une langue seconde - environ 150 téléchargements depuis son lancement en 2014. Mise à disposition de la version 1.3 le 28/04/2017 sur (*www.ultraspeech.com/player*).

Bibliographie

[Ang+19] Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky et al. "Speech synthesis from ECoG using densely connected 3D convolutional neural networks". In : *Journal of Neural Engineering* (2019) (cf. p. 30).

[Che+19] Fei Chen, Lan Wang, Gang Peng, Nan Yan et Xiaojie Pan. "Development and evaluation of a 3-D virtual pronunciation tutor for children with autism spectrum disorders". In : *PLOS ONE* 14.1 (2019), e0210858 (cf. p. 50).

[Roc+19] F. Roche, T. Hueber, S. Limier et L. Girin. "Autoencoders for music sound modeling : a comparison of linear, shallow, deep, recurrent and variational models". In : *Proc. of SMC*. Malaga, Spain, 2019, to appear (cf. p. 28, 84, 93).

[Bar18] Marie-Lou Barnaud. "Modélisation bayésienne du développement conjoint de la perception, l'action et la phonologie". Thèse de doct. Univ. Grenoble-Alpes, 2018, p. 244 (cf. p. 79, 80).

[Bir+18] Peter Birkholz, Simon Stone, Klaus Wolf et Dirk Plettemeier. "Non-Invasive Silent Phoneme Recognition Using Microwave Signals". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.12 (déc. 2018), p. 2404-2411 (cf. p. 9).

[Cha+18] Josh Chartier, Gopala K. Anumanchipalli, Keith Johnson et Edward F. Chang. "Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex". In : *Neuron* 98.5 (juin 2018), 1042-1054.e4 (cf. p. 30).

[Der+18] D. Derrick, C. Carignan, W. R. Chen, M. Shujau et C. T. Best. "Three-dimensional printable ultrasound transducer stabilization system". In : *The Journal of the Acoustical Society of America* 144.5 (2018), EL392-EL398 (cf. p. 24).

[Dud+18] Shiran Dudy, Steven Bedrick, Meysam Asgari et Alexander Kain. "Automatic analysis of pronunciations for children with speech sound disorders". In : *Computer Speech & Language* 50 (juil. 2018), p. 62-84 (cf. p. 67, 88).

[GCR18] Carlos D. Gerardo, Edmond Cretu et Robert Rohling. "Fabrication and testing of polymer-based capacitive micromachined ultrasound transducers for medical imaging". In : *Microsystems & Nanoengineering* 4.1 (déc. 2018), p. 19 (cf. p. 24).

[Hal+18] Célise Haldin, Audrey Acher, Louise Kauffmann, Thomas Hueber, Emilie Cousin et al. "Speech recovery and language plasticity can be facilitated by Sensori-Motor Fusion training in chronic non-fluent aphasia. A case report study". In : *Clinical Linguistics and Phonetics* 32.7 (2018), p. 595-621 (cf. p. 69).

[KK18] Takuhiro Kaneko et Hirokazu Kameoka. "CycleGAN-VC : Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks". In : *Proc. of EUSIPCO*. Rome, Italy : IEEE, sept. 2018, p. 2100-2104 (cf. p. 68).

[Liu+18a] Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou et Li-Rong Dai. "WaveNet Vocoder with Limited Training Data for Voice Conversion". In : *Interspeech 2018*. Hyderabad, India : ISCA, 2018, p. 1983-1987 (cf. p. 88).

[Liu+18b] L. Liu, T. Hueber, G. Feng et D. Beautemps. "Visual Recognition of Continuous Cued Speech Using a Tandem CNN-HMM Approach". In : *Proc. of Interspeech*. Hyderabad, India : ISCA, 2018, p. 2643-2647 (cf. p. 41, 42).

[OMH18] Kayoko Okada, William Matchin et Gregory Hickok. "Neural evidence for predictive coding in auditory cortex during speech production". In : *Psychonomic Bulletin & Review* 25.1 (fév. 2018), p. 423-430 (cf. p. 76).

[She+18] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly et al. "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions". In : *Proc. of ICASSP*. Calgary, Canada : IEEE, avr. 2018, p. 4779-4783 (cf. p. 37, 88).

[Tre+18] Avril Treille, Coriandre Vilain, Jean Luc Schwartz, Thomas Hueber et Marc Sato. "Electrophysiological evidence for Audio-visuo-lingual speech integration". In : *Neuropsychologia* 109 (2018), p. 126-133 (cf. p. 48).

[Yu+18] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu et al. "Generative Image Inpainting with Contextual Attention". In : *Proc. of CVPR*. Salt Lake City, Utah, USA : IEEE, juin 2018, p. 5505-5514 (cf. p. 77).

[Abd17] Ahmed Hussen Abdelaziz. "NTCD-TIMIT : A new database and baseline for noise-robust audio-visual speech recognition". In : *Proc. of Interspeech*. Stockholm, Sweden : ISCA, 2017, p. 3752-3756 (cf. p. 81).

[Boc17] Florent Bocquelet. "Toward a brain-computer interface for speech restoration". Thèse de doct. Université Grenoble-Alpes, 2017 (cf. p. 25, 28, 29).

[Chu+17] Joon Son Chung, Andrew W Senior, Oriol Vinyals et Andrew Zisserman. "Lip Reading Sentences in the Wild." In : *Proc. of CVPR*. Honolulu, Hawaii, USA, 2017, p. 3444-3453 (cf. p. 6).

[DB17] Christopher Dromey et Katherine M. Black. "Effects of Laryngeal Activity on Articulation". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (déc. 2017), p. 2272-2280 (cf. p. 21).

[Fab+17] Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda et Pierre Badin. "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract". In : *Speech Communication* 93 (2017), p. 63-75 (cf. p. 52, 61, 63-67, 93).

[Feu+17] Lionel Feugère, Christophe D'Alessandro, Boris Doval et Olivier Perrotin. "Cantor Digitalis : chironomic parametric synthesis of singing". In : *EURASIP Journal on Audio, Speech, and Music Processing* 2017.1 (déc. 2017), p. 2 (cf. p. 23).

[GHA17a] Laurent Girin, Thomas Hueber et Xavier Alameda-Pineda. "Adaptation of a Gaussian Mixture Regressor to a New Input Distribution : Extending the C-GMR Framework". In : *Proc. of International Conference on Latent Variable Analysis and Signal Separation*. Springer. Grenoble, France, 2017, p. 459-468 (cf. p. 58).

[GHA17b] Laurent Girin, Thomas Hueber et Xavier Alameda-Pineda. "Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping". In : *IEEE/ACM Transactions on Audio Speech and Language Processing* 25.3 (2017), p. 662-673 (cf. p. 56, 58, 59, 65).

[Gir17] Marion Girod-Roux. "Retour visuel par échographie linguale chez les sujets glossectomisés : Intérêt du biofeedback par rapport à l'illustration en rééducation de l'articulation". Thèse de Master Recherche. Université Paris X, 2017, p. 86 (cf. p. 44, 71, 72).

[Hsu+17] Chin Cheng Hsu, Hsin Te Hwang, Yi Chiao Wu, Yu Tsao et Hsin Min Wang. "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks". In : *Proc. of Interspeech*. Stockholm, Sweden : ISCA, 2017, p. 3364-3368 (cf. p. 55).

[JD17] Matthias Janke et Lorenz Diener. "EMG-to-Speech : Direct Generation of Speech from Facial Electromyographic Signals". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 12 (déc. 2017), p. 2375-2385 (cf. p. 9).

[Kob+17] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori et Tomoki Toda. "Statistical Voice Conversion with WaveNet-Based Waveform Generation". In : *Proc. of Interspeech*. Stockholm, Sweden : ISCA, août 2017, p. 1138-1142 (cf. p. 55, 88).

[Lau+17] Raphaël Laurent, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière et Julien Diard. "The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception." In : *Psychological Review* 124.5 (oct. 2017), p. 572-602 (cf. p. 78, 79).

[LQM17] Kun Li, Xiaojun Qian et Helen Meng. "Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (jan. 2017), p. 193-207 (cf. p. 67).

[Maa+17] Andrew L. Maas, Peng Qi, Ziang Xie, Awni Y. Hannun, Christopher T. Lengerich et al. "Building DNN acoustic models for large vocabulary speech recognition". In : *Computer Speech & Language* 41 (jan. 2017), p. 195-213 (cf. p. 6).

[MK17] Seyed Hamidreza Mohammadi et Alexander Kain. "An overview of voice conversion systems". In : *Speech Communication* 88 (avr. 2017), p. 65-82 (cf. p. 55).

[Pou17] Maël Pouget. "Synthèse incrémentale de la parole à partir du texte". Thèse de doct. Université Grenoble-Alpes, 2017 (cf. p. 32, 33, 36, 37).

[Sch+17a] T. Schultz, T. Hueber, D. J. Krusienski et J. S. Brumberg. "Introduction to the Special Issue on Biosignal-Based Spoken Communication". In : *IEEE/ACM Transactions on Audio Speech and Language Processing* 25.12 (2017), p. 2254-2256 (cf. p. 9, 10).

[Sch+17b] T Schultz, M Wand, T Hueber, D J Krusienski, C Herff et al. "Biosignal-Based Spoken Communication : A Survey". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (2017), p. 2257-2271 (cf. p. 6, 10, 14).

[SB17] Simon Stone et Peter Birkholz. "Angle Correction in Optopalatographic Tongue Distance Measurements". In : *IEEE Sensors Journal* 17.2 (jan. 2017), p. 459-468 (cf. p. 9).

[TH17] Eric Tatulli et Thomas Hueber. "Feature extraction using multimodal convolutional neural networks for visual speech recognition". In : *Proc. of ICASSP*. New Orleans, LA, USA : IEEE, 2017, p. 2971-2975 (cf. p. 5, 12-14, 22, 94).

[Tre+17] Avril Treille, Coriandre Vilain, Thomas Hueber, Laurent Lamalle et Marc Sato. "Inside speech : Multisensory and modality-specific processing of tongue and lip speech actions". In : *Journal of Cognitive Neuroscience* 29.3 (2017), p. 448-466 (cf. p. 48).

[WS17] Michael Wand et Jurgen Jürgen Schmidhuber. "Improving Speaker-Independent Lipreading with Domain-Adversarial Training". In : *Proc. of Interspeech*. Stockholm, Sweden : ISCA, août 2017, p. 3662-3666 (cf. p. 6).

[Wan+17] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss et al. "Tacotron : Towards End-to-End Speech Synthesis". In : *Proc. of Interspeech*. Stockholm, Sweden : ISCA, août 2017, p. 4006-4010 (cf. p. 37).

[Yeh+17] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson et al. "Semantic Image Inpainting with Deep Generative Models". In : *Proc. of CVPR*. Hawaï, Honolulu, Hawaï, USA : IEEE, juil. 2017, p. 6882-6890 (cf. p. 77).

[Zhu+17] Jun-Yan Zhu, Taesung Park, Phillip Isola et Alexei A. Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In : *Proc. of ICCV*. Venice, Italy : IEEE, oct. 2017, p. 2242-2251 (cf. p. 86).

[Ach+16] Audrey Acher, Diandra Fabre, Thomas Hueber, Pierre Badin, Olivier Detante et al. "Retour visuel en aphasiologie : résultats comportementaux, acoustiques et en neuroimagerie". In : *Actes des XVIèmes Rencontres Internationales d'Orthophonie : Orthophonie et technologies innovantes*. Sous la dir. de Silvia Topouzkhanian Nathalie Joyeux. Orthophonie et technologies innovantes. Paris, France : Ortho Edition, 2016. Chap. 12, p. 227-260 (cf. p. 50).

[Air+16] Manu Airaksinen, Bajibabu Bollepalli, Lauri Juvela, Zhizheng Wu, Simon King et al. "GlottDNN — A Full-Band Glottal Vocoder for Statistical Parametric Speech Synthesis". In : *Proc. of Interspeech*. San Francisco, CA, USA : ISCA, sept. 2016, p. 2473-2477 (cf. p. 36).

[Amo+16] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg et al. "Deep speech 2 : End-to-end speech recognition in english and mandarin". In : *Proc. of International Conference on Machine Learning*. New York, New York, USA, 2016, p. 173-182 (cf. p. 6).

[BSC16] Astik Biswas, P K Sahu et Mahesh Chandra. "Multiple cameras audio visual speech recognition using active appearance model visual features in car environment". In : *International Journal of Speech Technology* 19.1 (2016), p. 159-171 (cf. p. 5).

[Bly+16] Katrina M. Blyth, Patricia Mccabe, Catherine Madill et Kirrie J. Ballard. "Ultrasound visual feedback in articulation therapy following partial glossectomy". In : *Journal of Communication Disorders* 61 (mai 2016), p. 1-15 (cf. p. 51).

[Boc+16a] Florent Bocquelet, Thomas Hueber, Laurent Girin, Stéphan Chabardès et Blaise Yvert. "Key considerations in designing a speech brain-computer interface". In : *Journal of Physiology Paris* 110.4 (2016), p. 392-401 (cf. p. 25, 30, 93).

[Boc+16b] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux et Blaise Yvert. "Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces". In : *PLoS Computational Biology* 12.11 (2016), e1005119 (cf. p. 21-23, 28, 93).

[Che+16] Yi-Ping Phoebe Chen, Caddi Johnson, Pooia Lalbakhsh, Terry Caelli, Guang Deng et al. "Systematic review of virtual speech therapists for speech disorders". In : *Computer Speech & Language* 37 (mai 2016), p. 98-128 (cf. p. 49).

[CZ16] J S Chung et A Zisserman. "Lip Reading in the Wild". In : *Proc. Asian Conference on Computer Vision*. Taipei,Taiwan, 2016 (cf. p. 7, 11, 12).

[Fab16] Diandra Fabre. "Retour articulatoire visuel par échographie linguale augmentée : développements et application clinique". Thèse de doct. Université Grenoble-Alpes, 2016, p. 165 (cf. p. 44, 61, 63, 65, 71).

[Fab+16] Diandra Fabre, Thomas Hueber, Mélanie Canault, Nathalie Bedoin, Audrey Acher et al. "Apport de l'échographie linguale à la rééducation orthophonique". In : *XVIèmes Rencontres Internationales d'Orthophonie*. Sous la dir. de Silvia Topouzkhanian Nathalie Joyeux. Orthophonie et technologies innovantes. Paris, France : Ortho Edition, 2016. Chap. 11, p. 199-225 (cf. p. 69, 93).

[Gon+16] Jose A. Gonzalez, Lam A. Cheah, James M. Gilbert, Jie Bai, Stephen R. Ell et al. "A Silent Speech System based on Permanent Magnet Articulography and Direct Synthesis". In : *Computer Speech & Language* 39 (sept. 2016), p. 67-87 (cf. p. 9).

[Hsu+16] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao et Hsin-Min Wang. "Voice conversion from non-parallel corpora using variational auto-encoder". In : *Proc. of APSIPA*. Jeju Island, Korea : IEEE, déc. 2016, p. 1-6 (cf. p. 55).

[HB16] Thomas Hueber et Gérard Bailly. “Statistical conversion of silent articulation into audible speech using full-covariance HMM”. In : *Computer Speech and Language* 36 (2016), p. 274-293 (cf. p. 5, 11, 13, 14, 18-21, 23).

[LLD16] Zheng-Chen Liu, Zhen-Hua Ling et Li-Rong Dai. “Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks”. In : *Proc. of Interspeech*. San Francisco, CA, USA : ISCA, 2016, p. 1502-1506 (cf. p. 55).

[Min+16] Huaiping Ming, Dongyan Huang, Lei Xie, Jie Wu, Minghui Dong et al. “Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion”. In : *Proc. of Interspeech*. San Francisco, CA, USA, sept. 2016, p. 2453-2457 (cf. p. 55).

[MYO16] Masanori MORISE, Fumiya YOKOMORI et Kenji OZAWA. “WORLD : A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications”. In : *IEICE Transactions on Information and Systems* E99.D.7 (2016), p. 1877-1884 (cf. p. 36).

[PD16] Olivier Perrotin et Christophe D’alessandro. “Seeing, Listening, Drawing”. In : *ACM Transactions on Applied Perception* 14.2 (oct. 2016), p. 1-19 (cf. p. 23).

[Pou+16] Maël Pouget, Olha Nahorna, Thomas Hueber et Gérard Bailly. “Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis”. In : *Proc. of Interspeech*. San Francisco, CA, USA, 2016, p. 2846-2850 (cf. p. 34, 35, 39, 40, 93, 94).

[Pra+16] Pierre Prablanc, Alexey Ozerov, Ngoc Q. K. Duong et Patrick Perez. “Text-informed speech inpainting via voice conversion”. In : *Proc. of EUSIPCO*. Budapest, Hungary : IEEE, août 2016, p. 878-882 (cf. p. 88).

[SG16] Anders Sogaard et Yoav Goldberg. “Deep multi-task learning with low level tasks supervised at lower layers”. In : *Proc. of Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA, 2016, p. 231-235 (cf. p. 39).

[WKS16] Michael Wand, Jan Koutník et Jürgen Schmidhuber. “Lipreading with Long Short-Term Memory”. In : *Proc. of ICASSP*. Shanghai, China : IEEE, 2016, p. 6115-6119 (cf. p. 6).

[WS16] Michael Wand et Jürgen Schmidhuber. “Deep Neural Network Frontend for Continuous EMG-based Speech Recognition”. In : *Proc. of Interspeech*. San Francisco, CA, USA : ISCA, sept. 2016, p. 3032-3036 (cf. p. 9).

[XSL16] Feng Long Xie, Frank K. Soong et Haifeng Li. “A KL divergence and DNN-based approach to voice conversion without parallel training sentences”. In : *Proc. of Interspeech*. San Francisco, CA, USA : ISCA, 2016, p. 287-291 (cf. p. 68).

[Xu+16] Kele Xu, Yin Yang, Maureen Stone, Aurore Jaumard-Hakoun, Clémence Leboullenger et al. “Robust contour tracking in ultrasound tongue image sequences”. In : *Clinical Linguistics and Phonetics* (2016) (cf. p. 60).

[BL15] Camille Bach et Lorene Lambourion. “Impact de la visualisation du geste articulatoire, acquis par imagerie ultrasonore, pour la rééducation du trouble phonologique chez l’enfant”. Certificat de capacité d’orthophoniste. Université Claude Bernard Lyon I, 2015, p. 100 (cf. p. 44, 68).

[Boc+15a] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux et Blaise Yvert. *Real-time articulatory speech synthesis for brain-computer interfaces*. Society for Neuroscience Annual Meeting (abstract and oral presentation). Chicago, USA, 2015 (cf. p. 93).

[Boc+15b] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux et Blaise Yvert. “Real-time control of a DNN-based articulatory synthesizer for silent speech conversion : A pilot study”. In : *Proc. of Interspeech*. Dresden, Germany : ISCA, 2015, p. 2405-2409 (cf. p. 23, 28, 93).

[CSW15] Joanne Cleland, James M Scobbie et Alan A Wrench. “Using ultrasound visual biofeedback to treat persistent primary speech sound disorders”. In : *Clinical Linguistics and Phonetics* 29.8-10 (2015), p. 575-597 (cf. p. 50).

[DBM15] Wim De Mulder, Steven Bethard et Marie-Francine Moens. “A survey on the application of recurrent neural networks to statistical language modeling”. In : *Computer Speech & Language* 30.1 (mar. 2015), p. 61-98 (cf. p. 6).

[Fab+15] Diandra Fabre, Thomas Hueber, Florent Bocquelet et Pierre Badin. “Tongue tracking in ultrasound images using Eigentongue decomposition and artificial neural networks”. In : *Proc. of Interspeech*. Dresden, Germany : ISCA, 2015, p. 2410-2414 (cf. p. 60, 62, 93).

[Her+15] Christian Herff, Dominic Heger, Adriana de Pesters, Dominic Telaar, Peter Brunner et al. “Brain-to-text : decoding spoken phrases from phone representations in the brain”. In : *Frontiers in Neuroscience* 9.217 (juin 2015) (cf. p. 26, 30).

[Hue+15] Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda et Gerard Bailly. “Speaker-Adaptive Acoustic-Articulatory Inversion Using Cascaded Gaussian Mixture Regression”. In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.12 (déc. 2015), p. 2246-2259 (cf. p. 54-56, 65).

[Jar+15] Beata Jarosiewicz, Anish A. Sarma, Daniel Bacher, Nicolas Y. Masse, John D. Simeral et al. “Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface”. In : *Science Translational Medicine* 7.313 (nov. 2015), p. 1-11 (cf. p. 26).

[LFM15] Imed Laaridh, Corinne Fredouille et Christine Meunier. “Automatic Detection of Phone-Based Anomalies in Dysarthric Speech”. In : *ACM Transactions on Accessible Computing* 6.3 (mai 2015), p. 1-24 (cf. p. 67, 88).

[LBH15] Y LeCun, Y Bengio et G Hinton. “Deep Learning”. In : *Nature* 521 (2015), p. 436-444 (cf. p. 12).

[Lin+15] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster et al. "Deep Learning for Acoustic Modeling in Parametric Speech Generation : A systematic review of existing techniques and future trends". In : *IEEE Signal Processing Magazine* 32.3 (mai 2015), p. 35-52 (cf. p. 37).

[Mou+15] Clément Moulin-Frier, Julien Diard, Jean-Luc Schwartz et Pierre Bessière. "COSMO ("Communicating about Objects using Sensory–Motor Operations") : A Bayesian modeling framework for studying speech communication and the emergence of phonological systems". In : *Journal of Phonetics* 53 (nov. 2015), p. 5-41 (cf. p. 78).

[MMG15] Y Mroueh, E Marcheret et V Goel. "Deep multimodal learning for Audio-Visual Speech Recognition". In : *Proc. of ICASSP*. Brisbane, Queensland, Australia : IEEE, 2015, p. 2130-2134 (cf. p. 6).

[Pou+15] Maël Pouget, Thomas Hueber, Gérard Bailly et Timo Baumann. "HMM training strategy for incremental speech synthesis". In : *Proc. of Interspeech*. Dresden, Germany, 2015, p. 1201-1205 (cf. p. 34, 38, 39, 93).

[RSC15] Zoe Roxburgh, James M Scobbie et Joanne Cleland. "Articulation Therapy for Children With Cleft Palate Using Visual Articulatory Models and Ultrasound Biofeedback". In : *Proc. of ICPhS*. 1. Glasgow, Scotland, 2015 (cf. p. 50).

[Sze+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed et al. "Going deeper with convolutions". In : *Proc. of CVPR*. Boston, Massachusetts, USA : IEEE, 2015, p. 1-9 (cf. p. 12).

[Ver+15] Maarten Versteegh, Roland Thiollière, Thomas Schatz, Xuan Nga Cao, Xavier Anguera et al. "The Zero Resource Speech Challenge 2015". In : *Proc. of Interspeech*. Dresden, Germany : ISCA, 2015, p. 3169-3173 (cf. p. 77).

[Abd+14] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn et al. "Convolutional neural networks for speech recognition". In : *IEEE/ACM Transactions on audio, speech, and language processing* 22.10 (2014), p. 1533-1545 (cf. p. 6, 86).

[Ach14] Audrey Acher. "Corrélats cérébraux de l'adaptation de la parole après exérèse au niveau de la cavité orale". Thèse de doct. Université Grenoble-Alpes, 2014 (cf. p. 45, 47).

[Ach+14] Audrey Acher, Pascal Perrier, Christophe Savariaux et Cecile Fougeron. "Speech production after glossectomy : Methodological aspects". In : *Clinical Linguistics and Phonetics* (2014) (cf. p. 47).

[Bau14] Timo Baumann. "Decision tree usage for incremental parametric speech synthesis". In : *Proc. of ICASSP*. Florence, Italy : IEEE, 2014, p. 3819-3823 (cf. p. 34, 38).

[Boc+14] Florent Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin et Blaise Yvert. "Robust articulatory speech synthesis using deep neural networks for BCI applications". In : *Proc. of Interspeech*. Singapore, 2014, p. 2288-2292 (cf. p. 8, 21, 23, 27-30, 93).

[FHB14] D Fabre, T Hueber et P Badin. "Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression". In : *Proc. of Interspeech*. Singapore : ISCA, 2014, p. 2293-2297 (cf. p. 61, 93).

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley et al. "Generative Adversarial Nets". In : *Proc. of NIPS*. Montreal, Quebec, Canada, 2014, p. 2672-2680 (cf. p. 84).

[GJ14] Alex Graves et Navdeep Jaitly. "Towards End-To-End Speech Recognition with Recurrent Neural Networks." In : *Proc. of International Conference on Machine Learning*. T. 14. Beijing, China, 2014, p. 1764-1772 (cf. p. 6).

[Kar+14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar et al. "Large-scale video classification with convolutional neural networks". In : *Proc. of CVPR*. Columbus, Ohio, USA : IEEE, 2014, p. 1725-1732 (cf. p. 12).

[KW14] D.P. Kingma et M. Welling. "Auto-Encoding Variational Bayes". In : *Proc. of ICLR*. Banff, Canada, 2014 (cf. p. 29).

[KM14] Arne Köhn et Wolfgang Menzel. "Incremental Predictive Parsing with TurboParser". In : *Proc. of Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA, 2014, p. 803-808 (cf. p. 34).

[Lin+14] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu et Li-Rong Dai. "Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.12 (déc. 2014), p. 1859-1872 (cf. p. 55).

[LVB14] Matthieu Loosvelt, Pierre-Frederic Villard et Marie-Odile Berger. "Using a biomechanical model for tongue tracking in ultrasound images". In : *Proc. of 6th International Symposium on Biomedical Simulation*. Strasbourg, France, 2014 (cf. p. 60).

[MNO14] Clément Moulin-Frier, Sao M. Nguyen et Pierre-Yves Oudeyer. "Self-organization of early vocal development in infants and machines : the role of intrinsic motivation". In : *Frontiers in Psychology* 4 (2014) (cf. p. 85).

[Mug+14] Emily M Mugler, James L Patton, Robert D Flint, Zachary A Wright, Stephan U Schuele et al. "Direct classification of all American English phonemes using signals from functional speech motor cortex". In : *Journal of Neural Engineering* 11.3 (juin 2014), p. 035015 (cf. p. 26).

[Nar+14] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim et al. "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research". In : *The Journal of the Acoustical Society of America* 136.3 (sept. 2014), p. 1307-1311 (cf. p. 7).

[Nod+14] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno et Tetsuya Ogata. "Lipreading using convolutional neural network." In : *Proc. of Interspeech*. Singapore, 2014, p. 1149-1153 (cf. p. 12).

[SZ14a] Cicero Nogueira dos Santos et Bianca Zadrozny. “Learning Character-level Representations for Part-of-Speech Tagging.” In : *Proc. of International Conference on Machine Learning*. Beijing, China, 2014, p. 1818-1826 (cf. p. 35).

[SS14] Jean-Luc Schwartz et Christophe Savariaux. “No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag”. In : *PLoS Comput Biol* 10.7 (juil. 2014), e1003743 (cf. p. 23).

[SZ14b] Karen Simonyan et Andrew Zisserman. “Two-stream convolutional networks for action recognition in videos”. In : *Proc. of NIPS*. Montreal, Quebec, Canada, 2014, p. 568-576 (cf. p. 12).

[WHB14] Xiaoou Wang, Thomas Hueber et Pierre Badin. “On the use of an articulatory talking head for second language pronunciation training : the case of Chinese learners of French”. In : *Proc. of International Seminar on Speech Production (ISSP)*. Cologne, Germany, 2014, p. 449-452 (cf. p. 44, 70, 71, 94).

[ZS14] Heiga Zen et Andrew Senior. “Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis”. In : *Proc. of ICASSP*. Florence, Italy : IEEE, 2014, p. 3844-3848 (cf. p. 37).

[Zho+14] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong et Matti Pietikäinen. “A review of recent advances in visual speech decoding”. In : *Image and Vision Computing* 32.9 (sept. 2014), p. 590-605 (cf. p. 6).

[BRM13] Tadas Baltrusaitis, Peter Robinson et Louis-Philippe Morency. “Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild”. In : *Proc. ICCV*. Sydney, Australia : IEEE, déc. 2013, p. 354-361 (cf. p. 5).

[Bou+13] Kristofer E. Bouchard, Nima Mesgarani, Keith Johnson et Edward F. Chang. “Functional organization of human sensorimotor cortex for speech articulation”. In : *Nature* 495.7441 (mar. 2013), p. 327-332 (cf. p. 27).

[Col+13] Jennifer L. Collinger, Brian Wodlinger, John E. Downey, Wei Wang, Elizabeth C. Tyler-Kabara et al. “High-performance neuroprosthetic control by an individual with tetraplegia”. In : *The Lancet* 381.9866 (2013), p. 557-564 (cf. p. 27).

[Deg+13] Gilles Degottex, Pierre Lanchantin, Axel Roebel et Xavier Rodet. “Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis”. In : *Speech Communication* 55.2 (fév. 2013), p. 278-294 (cf. p. 8).

[GPC13] Oonagh M. Giggins, Ulrik Persson et Brian Caulfield. “Biofeedback in rehabilitation”. In : *Journal of NeuroEngineering and Rehabilitation* 10.1 (2013), p. 60 (cf. p. 50).

[Gra+13] Krystyna Grabski, Jean-Luc Schwartz, Laurent Lamalle, Coriandre Vilain, Nathalie Vallée et al. “Shared and distinct neural correlates of vowel perception and production”. In : *Journal of Neurolinguistics* 26.3 (2013), p. 384-408 (cf. p. 45).

[Hue13] Thomas Hueber. "Ultraspeech-player : Intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training". In : *Proc. of Interspeech (Show&Tell)*. Lyon, France : ISCA, 2013, p. 752-753 (cf. p. 49, 51).

[Hue+13] Thomas Hueber, Gérard Bailly, Pierre Badin et Frédéric Elisei. "Speaker adaptation of an acoustic-articulatory inversion model using cascaded gaussian mixture regressions". In : *Proc. of Interspeech*. Lyon, France : ISCA, 2013, p. 2753-2757 (cf. p. 55, 56).

[Ji+13] Shuiwang Ji, Wei Xu, Ming Yang et Kai Yu. "3D convolutional neural networks for human action recognition". In : *IEEE Trans. Pattern Analysis and Machine Intelligence,* 35 (2013), p. 221-231 (cf. p. 12).

[PBL13] Jonathan L. Preston, Nickole Brick et Nicole Landi. "Ultrasound Biofeedback Treatment for Persisting Childhood Apraxia of Speech". In : *American Journal of Speech-Language Pathology* 22.4 (nov. 2013), p. 627-643 (cf. p. 50).

[Tok+13] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi et al. "Speech Synthesis Based on Hidden Markov Models". In : *Proceedings of the IEEE* 101.5 (mai 2013), p. 1234-1252 (cf. p. 38).

[Wan+13] Michael Wand, Christopher Schulte, Matthias Janke et Tanja Schultz. "Array-based Electromyographic Silent Speech Interface". In : *Proc. of International Joint Conference on Biomedical Engineering Systems and Technologies (Biosignals)*. Barcelona, Spain, 2013 (cf. p. 8).

[Adl+12] Amir Adler, Valentin Emiya, Maria G. Jafari, Michael Elad, Rémi Gribonval et al. "Audio Inpainting". In : *IEEE Transactions on Audio, Speech, and Language Processing* 20.3 (mar. 2012), p. 922-932 (cf. p. 88).

[AG12] Luc H. Arnal et Anne-Lise Giraud. "Cortical oscillations and sensory predictions". In : *Trends in Cognitive Sciences* 16.7 (juil. 2012), p. 390-398 (cf. p. 76).

[Bac+12] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia et Atilla Baskurt. "Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification." In : *Proc. of British Machine Vision Conference (BMVC)*. Surrey, England, 2012, p. 1-12 (cf. p. 12).

[BS12] Timo Baumann et David Schlangen. "INPRO_iSS : A component for just-in-time incremental speech synthesis". In : *Proc. of the ACL 2012 System Demonstrations*. Jeju Island, Korea, 2012, p. 103-108 (cf. p. 34).

[GV12] Frank H. Guenther et Tony Vladusich. "A neural theory of speech acquisition and production". In : *Journal of Neurolinguistics* 25.5 (sept. 2012), p. 408-422 (cf. p. 75, 78).

[HBH12] Panikos Heracleous, Denis Beautemps et Norihiro Hagita. "Continuous phoneme recognition in Cued Speech for French". In : *Proc. of EUSIPCO*. Bucharest, Romania : IEEE, 2012, p. 2090-2093 (cf. p. 41).

[Hin+12] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition : The Shared Views of Four Research Groups". In : *IEEE Signal Processing Magazine* 29.6 (nov. 2012), p. 82-97 (cf. p. 6).

[HBD12] Thomas Hueber, Gérard Bailly et Bruce Denby. "Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface". In : *Proc. of Interspeech*. T. 1. Portland, Oregon, USA, 2012, p. 723-726 (cf. p. 5, 18, 23).

[Hue+12] Thomas Hueber, AB Youssef, Gérard Bailly, Badin Pierre. et Frédéric Elisei. "Cross-speaker Acoustic-to-Articulatory Inversion using Phone-based Trajectory HMM for Pronunciation Training". In : *Proc. of Interspeech*. T. 1. Portland, OR, USA, 2012, p. 3-6 (cf. p. 23, 53).

[MS12] Anne Menin-Sicard et Etienne Sicard. "Diadolab : a simulation tool illustrating speech movements for handling articulatory and phonological disorders". In : *Glossa* 111 (2012), p. 98-116 (cf. p. 49).

[Nak+12] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari et Kiyohiro Shikano. "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech". In : *Speech Communication* 54.1 (2012), p. 134-146 (cf. p. 9).

[Per12] Pascal Perrier. "Gesture planning integrating knowledge of the motor plant's dynamics : A literature review from motor control and speech motor control". In : *Speech Planning and Dynamics*. Sous la dir. de Susanne Fuchs, Mélanie Weirich, Pape Daniel et Pascal Perrier. Speech Pro. Peter Lang Publishers, 2012, p. 191-238 (cf. p. 45, 46, 75, 78).

[Sch+12] Jean-Luc Schwartz, Anahita Basirat, Lucie Ménard et Marc Sato. "The Perception-for-Action-Control Theory (PACT) : A perceptuo-motor theory of speech perception". In : *Journal of Neurolinguistics* 25.5 (2012), p. 336-354 (cf. p. 45, 75).

[TFS12] Ariel Tankus, Itzhak Fried et Shy Shoham. "Structured neuronal encoding and decoding of human speech features". In : *Nature Communications* 3.1 (jan. 2012), p. 1015 (cf. p. 26).

[TNS12] Tomoki Toda, Mikihiro Nakagiri et Kiyohiro Shikano. "Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement". In : *IEEE Transactions on Audio, Speech, and Language Processing* 20.9 (nov. 2012), p. 2505-2517 (cf. p. 9).

[Uri+12] Benigno Uria, Iain Murray, Steve Renals et Korin Richmond. "Deep architectures for articulatory inversion". In : *Proc. of Interspeech*. Portland, OR, USA : ISCA, 2012, p. 866-870 (cf. p. 55).

[WW12] Jonathan Wolpaw et Elizabeth Winter Wolpaw. *Brain–Computer Interfaces-Principles and Practice*. Oxford University Press, jan. 2012 (cf. p. 29).

[AE11] G. Ananthakrishnan et Olov Engwall. "Mapping between acoustic and articulatory gestures". In : *Speech Communication* 53.4 (avr. 2011), p. 567-589 (cf. p. 55).

[Ben+11] Atef Ben Youssef, Thomas Hueber, Pierre Badin et Gérard Bailly. "Toward a multi-speaker visual articulatory feedback system". In : *Proc. of Interspeech*. Florence, Italy : ISCA, 2011, p. 589-592 (cf. p. 53).

[BKM11] Niels Beuck, Arne Köhn et Wolfgang Menzel. "Decision Strategies in Incremental PoS Tagging". In : *Proc. of Nordic Conference on Computational Linguistics*. Riga, Latvia, 2011, p. 26-33 (cf. p. 34).

[Bri+11] Frédérique Brin, Catherine Courrier, Emmanuelle Lederlé, Véronique Masy et Julien-Daniel Guelfi. *Mini DSM-IV-TR : Critères diagnostiques*. Elsevier Masson, 2011 (cf. p. 1).

[Bru+11] Jonathan S Brumberg, E Joe Wright, Dinal S Andreasen, Frank H Guenther, Philip R Kennedy et al. "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex". In : *Frontiers in Neuroscience* 5.65 (mai 2011), p. 65 (cf. p. 26).

[BK11] Hendrik Buschmeier et Stefan Kopp. "Towards Conversational Agents That Attend to and Adapt to Communicative User Feedback". In : *Lecture Notes in Computer Science*. Sous la dir. de Vilhjálmsson H.H., Kopp S., Marsella S. et Thórisson K.R. Intelligen. Berlin, Heidelberg : Springer, 2011, p. 169-182 (cf. p. 34).

[Cai+11] Jun Cai, Thomas Hueber, Bruce Denby, Elie-Laurent Benaroya, Gérard Chollet et al. "A visual speech recognition system for an ultrasound-based silent speech interface ". In : *Proc. of International Congress of Phonetic Sciences (ICPhS)*. Hong Kong, China, 2011, p. 384-387 (cf. p. 14).

[Cas+11] Claudio Castellini, Leonardo Badino, Giorgio Metta, Giulio Sandini, Michele Tavella et al. "The Use of Phonetic Motor Invariants Can Improve Automatic Phoneme Discrimination". In : *PLoS ONE* 6.9 (sept. 2011). Sous la dir. de Paul L. Gribble, e24055 (cf. p. 78).

[Col+11] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu et al. "Natural language processing (almost) from scratch". In : *Journal of Machine Learning Research* 12.Aug (2011), p. 2493-2537 (cf. p. 35, 39).

[DRL11] Christophe D'Alessandro, Albert Rilliard et Sylvain Le Beux. "Chironomic stylization of intonation". In : *The Journal of the Acoustical Society of America* 129.3 (mar. 2011), p. 1594-1604 (cf. p. 23).

[GL11] Peggy Gatignol et Elodie Lannadère. *MBLF : Bilan informatisé de la Motricité Bucco-linguo-Faciale*. Chateauroux, 2011 (cf. p. 72).

[HN11] John F. Houde et Srikantan S. Nagarajan. "Speech Production as State Feedback Control". In : *Frontiers in Human Neuroscience* 5 (2011) (cf. p. 75, 78, 80).

[Hue+11a] Thomas Hueber, Pierre Badin, Christophe Savariaux, Coriandre Vilain et Gérard Bailly. "Differences in articulatory strategies between silent, whispered and normal speech ? A pilot study using ElectroMagnetic Articulography". In : *International Seminar on Speech Production (ISSP)*. Montréal, Canada, 2011 (cf. p. 21).

[Hue+11b] Thomas Hueber, Elie-Laurent Benaroya, Bruce Denby et Gérard Chollet. "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface." In : *Proc. of Interspeech*. Florence, Italy : ISCA, 2011, p. 593-596 (cf. p. 16, 30).

[JWS11] Matthias Janke, Michael Wand et Tanja Schultz. "Impact of different feedback mechanisms in EMG-based speech recognition". In : *Proc. of Interspeech*. Florence, Italy : ISCA, 2011, p. 2686-2689 (cf. p. 21).

[SA11] Eo Selfridge et Iker Arizmendi. "Stability and accuracy in incremental speech recognition". In : *Proc. of SIGDIAL*. Sous la dir. d'Association for Computational Linguistics. Portland, Oregon, USA, 2011, p. 110-119 (cf. p. 23).

[ZNT11] Heiga Zen, Yoshihiko Nankaku et Keiichi Tokuda. "Continuous stochastic feature mapping based on trajectory HMMs". In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.2 (2011), p. 417-430 (cf. p. 8, 18, 55).

[Bad+10] Pierre Badin, Yuliya Tarabalka, Frédéric Elisei et Gérard Bailly. "Can you 'read' tongue movements ? Evaluation of the contribution of tongue display to speech understanding". In : *Speech Communication* 52.6 (juin 2010), p. 493-503 (cf. p. 48).

[Den+10] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, James Gilbert et al. "Silent Speech Interfaces". In : *Speech Communication* 52.4 (2010), p. 270-287 (cf. p. 10).

[FB10] Ian Fasel et Jeff Berry. "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech". In : *Proc. of ICPR*. Istanbul, Turkey : IEEE, août 2010, p. 1493-1496 (cf. p. 60).

[Fri10] Karl Friston. "The free-energy principle : a unified brain theory ?" In : *Nature Reviews Neuroscience* 11.2 (2010), p. 127 (cf. p. 76).

[HBA10] Panikos Heracleous, Denis Beautemps et Noureddine Aboutabit. "Cued Speech automatic recognition in normal-hearing and deaf subjects". In : *Speech Communication* 52.6 (2010), p. 504-512 (cf. p. 41).

[Hue+10] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus et al. "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips". In : *Speech Communication* 52.4 (avr. 2010), p. 288-300 (cf. p. 5, 13).

[Kel+10] Spencer Kellis, Kai Miller, Kyle Thomson, Richard Brown, Paul House et al. "Decoding spoken words using local field potentials recorded from the cortical surface." In : *Journal of neural engineering* 7.5 (oct. 2010), p. 056007 (cf. p. 26).

[LCL10] Carol J LaSasso, Kelly Lamar Crain et Jacqueline Leybaert. *Cued Speech and Cued Language Development for Deaf and Hard of Hearing Children*. Plural Publishing, 2010 (cf. p. 41).

[SW10] Tanja Schultz et Michael Wand. "Modeling coarticulation in EMG-based continuous speech recognition". In : *Speech Communication* 52.4 (avr. 2010), p. 341-353 (cf. p. 9).

[Sok+10] N Sokolovska, T Lavergne, O Cappe et F Yvon. "Efficient Learning of Sparse Conditional Random Fields for Supervised Sequence Labeling". In : *IEEE Journal of Selected Topics in Signal Processing* 4.6 (2010), p. 953-964 (cf. p. 35).

[Ben09] Y. Bengio. "Learning Deep Architectures for AI". In : *Foundations and Trends in Machine Learning* 2.1 (2009), p. 1-127 (cf. p. 27).

[BPP09] Stéphanie Buchaillard, Pascal Perrier et Yohan Payan. "A biomechanical model of cardinal vowel production : Muscle activations and the impact of gravity on tongue positioning". In : *The Journal of the Acoustical Society of America* 126.4 (2009), p. 2033 (cf. p. 8).

[Cha+09] Chandramouli Chandrasekaran, Andrea Trubanova, Sébastien Stillittano, Alice Caplier et Asif A Ghazanfar. "The natural statistics of audiovisual speech". In : *PLoS Computational Biology* 5.7 (2009), e1000436 (cf. p. 81).

[Gue+09] Frank H. Guenther, Jonathan S. Brumberg, E Joseph Wright, Alfonso Nieto-Castanon, Jason A. Tourville et al. "A wireless brain-machine interface for real-time speech synthesis". In : *PLOS ONE* 4.12 (déc. 2009), e8218 (cf. p. 26).

[HD09] T Hueber et B Denby. "Analyse du conduit vocal par imagerie ultrasonore". In : *L'imagerie médicale pour l'étude de la parole*. Sous la dir. d'A Marchal et C Cavé. IC2, Hermes Science, 2009, p. 147-174 (cf. p. 10, 60).

[Hue09] Thomas Hueber. "Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal : vers une communication parlée silencieuse". French. PhD thesis. Paris : Université Pierre et Marie Curie - Paris VI, 2009, p. 200 (cf. p. 5).

[Oht+09] Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari et Kiyohiro Shikano. "Many-to-many eigenvoice conversion with reference voice". In : *Proc. of Interspeech*. Brighton, England : ISCA, 2009, p. 1623-1626 (cf. p. 68).

[RKM09] Anastasios Roussos, Athanassios Katsamanis et Petros Maragos. "Tongue tracking in Ultrasound images with Active Appearance Models". In : *Proc. of International Conference on Image Processing (ICIP)*. Cairo, Egypt : IEEE, nov. 2009, p. 1733-1736 (cf. p. 60).

[TWS09] Arthur R. Toth, Michael Wand et Tanja Schultz. "Synthesizing Speech from Electromyography using Voice Transformation Techniques". In : *Proc. of Interspeech*. Brighton, UK, 2009, p. 652-655 (cf. p. 9).

[Woo+09] Sara Wood, Jennifer Wishart, William Hardcastle, Joanne Cleland et Claire Timmins. "The use of electropalatography (EPG) in the assessment and treatment of motor speech disorders in children with Down's syndrome : Evidence from two case studies". In : *Developmental Neurorehabilitation* 12.2 (jan. 2009), p. 66-75 (cf. p. 50).

[You+09] Atef Ben Youssef, Pierre Badin, Gérard Bailly et Panikos Heracleous. "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models". In : *Proc. of Interspeech*. Brighton, England : ISCA, 2009, p. 2255-2258 (cf. p. 55).

[Bad+08] Pierre Badin, Frédéric Elisei, Gérard Bailly et Yuliya Tarabalka. "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*. Sous la dir. de Perales F.J. et Fisher R.B. Lecture No. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 132-143 (cf. p. 49, 52).

[Ber+08] May B. Bernhardt, Penelope Bacsfalvi, Marcy Adler-Bock, Reiko Shimizu, Audrey Cheney et al. "Ultrasound as visual feedback in speech habilitation : Exploring consultative use in rural British Columbia, Canada". In : *Clinical Linguistics & Phonetics* 22.2 (jan. 2008), p. 149-162 (cf. p. 50).

[Cla+08] Meghan Clayards, Michael K. Tanenhaus, Richard N. Aslin et Robert A. Jacobs. "Perception of speech reflects optimal use of probabilistic speech cues". In : *Cognition* 108.3 (sept. 2008), p. 804-809 (cf. p. 78).

[CW08] Ronan Collobert et Jason Weston. "A unified architecture for natural language processing". In : *Proc. of ICML*. New York, New York, USA : ACM Press, 2008, p. 160-167 (cf. p. 39).

[Edl08] Jens Edlund. "Incremental speech synthesis". In : *Proc. of Swedish Language Technology Conference*. Stockholm, Sweden, 2008, p. 53-54 (cf. p. 34).

[Fag+08] M.J. Fagan, S.R. Ell, J.M. Gilbert, E. Sarrazin et P.M. Chapman. "Development of a (silent) speech recognition system for patients following laryngectomy". In : *Medical Engineering & Physics* 30.4 (mai 2008), p. 419-425 (cf. p. 9).

[FM08] Sascha Fagel et Katja Madany. "A 3-D virtual head as a tool for speech therapy for children". In : *Proc. of Interspeech*. Brisbane, Australia : ISCA, 2008, p. 2643-2646 (cf. p. 49, 50).

[Hue+08a] Thomas Hueber, G. Chollet, B. Denby et M. Stone. "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application". In : *Proc. of International Seminar on Speech Production (ISSP)*. Strasbourg, France, 2008, p. 365-369 (cf. p. 11).

[Hue+08b] Thomas Hueber, Gérard Chollet, Bruce Denby, Gérard Dreyfus et Maureen Stone. "Phone recognition from ultrasound and optical video sequences for a silent speech interface". In : *Proc. of Interspeech*. Brisbane, Australia : ISCA, 2008, p. 2032-2035 (cf. p. 13).

[LR08] Le Zhang et Steve Renals. "Acoustic-Articulatory Modeling With the Trajectory HMM". In : *IEEE Signal Processing Letters* 15 (2008), p. 245-248 (cf. p. 55).

[Mod+08] Geetanjalee Modha, B. May Bernhardt, Robyn Church et Penelope Bacsfalvi. "Case study using ultrasound to treat /r/". In : *International Journal of Language & Communication Disorders* 43.3 (jan. 2008), p. 323-329 (cf. p. 50).

[TBT08] Tomoki Toda, Alan W Black et Keiichi Tokuda. "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model". In : *Speech Communication* 50.3 (2008), p. 215-227 (cf. p. 8, 30, 55).

[TRG08] Jason A Tourville, Kevin J Reilly et Frank H Guenther. "Neural Mechanisms Underlying Auditory Feedback Control Of Speech". In : 32 (2008), p. 1429-1443 (cf. p. 80).

[Abo07] Noureddine Aboutabit. "Reconnaissance de la Langue Française Parlée Complété (LPC) : décodage phonétique des gestes main-lèvres." Thèse de doct. Institut National Polytechnique de Grenoble-INPG, 2007 (cf. p. 41).

[Adl+07] Marcy Adler-Bock, Barbara May Bernhardt, Bryan Gick et Penelope Bacsfalvi. "The Use of Ultrasound in Remediation of North American English /r/ in 2 Adolescents". In : *American Journal of Speech-Language Pathology* 16.2 (mai 2007), p. 128-139 (cf. p. 50).

[BBG07] Penelope Bacsfalvi, Barbara May Bernhardt et Bryan Gick. "Electropalatography and ultrasound in vowel remediation for adolescents with hearing impairment". In : *Advances in Speech Language Pathology* 9.1 (2007), p. 36-45 (cf. p. 50).

[Has+07] Mitsuo Hashiba, Yasunori Sugai, Takashi Izumi, Shuichi Ino et Tohru Ifukube. "Development of a wearable electro-larynx for laryngectomees and its evaluation". In : *Proc. of International Conference of the IEEE Engineering in Medicine and Biology Society*. Lyon, France : IEEE, août 2007, p. 5267-5270 (cf. p. 23).

[Hue+07a] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus et al. "Eigentongue feature extraction for an ultrasound-based silent speech interface". In : *Proc. of ICASSP*. T. 1. Honolulu, Hawaii, USA : IEEE, 2007, p. 2193-2196 (cf. p. 61).

[Hue+07b] T. Hueber, G. Chollet, B. Denby, G Dreyfus et M Stone. "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips". In : *Proc. of Interspeech*. T. 3. Antwerp, Belgium : ISCA, 2007, p. 658-661 (cf. p. 13).

[Kin+07] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond et al. "Speech production knowledge in automatic speech recognition". In : *The Journal of the Acoustical Society of America* 121.2 (fév. 2007), p. 723-742 (cf. p. 75, 78).

[Mén+07] Lucie Ménard, Jean Luc Schwartz, Louis Jean Boë et Jérôme Aubin. "Articulatory-acoustic relationships during vocal tract growth for French vowels : Analysis of real data and simulations with an articulatory model". In : *Journal of Phonetics* (2007) (cf. p. 58, 85).

[Oun+07] Slim Ouni, Michael M. Cohen, Hope Ishak et Dominic W. Massaro. "Visual Contribution to Speech Perception : Measuring the Intelligibility of Animated Talking Heads". In : *EURASIP Journal on Audio, Speech, and Music Processing* 2007 (2007), p. 1-12 (cf. p. 48).

[TBT07] Tomoki Toda, Alan W Black et Keiichi Tokuda. "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory". In : *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), p. 2222-2235 (cf. p. 16).

[BJK06] P. Birkholz, D. Jackel et B.J. Kroger. "Construction And Control Of A Three-Dimensional Vocal Tract Model". In : *Proc. of ICASSP*. T. 1. 2-3. Toulouse, France : IEEE, 2006, p. 873-876 (cf. p. 7, 85).

[Bis06] Christopher M Bishop. *Pattern Recognition and Machine Learning*. New York, USA : Springer-Verlag, 2006, p. 738 (cf. p. 15, 18, 19).

[FKH06] Karl Friston, James Kilner et Lee Harrison. "A free energy principle for the brain". In : *Journal of Physiology-Paris* 100.1-3 (2006), p. 70-87 (cf. p. 76).

[Jou+06] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft et Alex Waibel. "Towards Continuous Speech Recognition using Surface Electromyography". In : *Proc. of ICSLP*. Pittsburgh, PA, USA, 2006, p. 573-576 (cf. p. 8, 9).

[Kaw06] Hideki Kawahara. "STRAIGHT, exploitation of the other aspect of VOCODER : Perceptually isomorphic decomposition of speech sounds". In : *Acoustical Science and Technology* 27.6 (2006), p. 349-353 (cf. p. 36).

[MVM06] Athanasios Mouchtaris, J. Van der Spiegel et Paul Mueller. "Nonparallel training for voice conversion based on a parameter adaptation approach". In : *IEEE Transactions on Audio, Speech and Language Processing* 14.3 (mai 2006), p. 952-963 (cf. p. 68).

[PR06] Auzou Pascal et Véronique Rolland-Monnoury. *BECD 2006 - Batterie d'Evaluation Clinique de la Dysarthrie*. Ortho Edit. 2006 (cf. p. 73).

[Ric06] Korin Richmond. "A trajectory mixture density network for the acoustic-articulatory inversion mapping". In : *Proc. of Interspeech*. T. 2. Pittsburgh, Pennsylvania, USA : ISCA, 2006, p. 577-580 (cf. p. 55).

[Bäl+05] Olle Bälter, Olov Engwall, Anne-Marie Öster et Hedvig Kjellström. "Wizard-of-Oz test of ARTUR". In : *Proc. of International Conference on Computers and accessibility*. New York, New York, USA : ACM Press, 2005, p. 36 (cf. p. 49).

[DAl+05] C. D'Alessandro, N. D'Alessandro, S. Le Beux, J. Simko, F. Cetin et al. "The Speech Conductor : Gestural Control of Speech Synthesis". In : *Proc. of eNTERFACE*. Mons, Belgium, 2005, p. 52-61 (cf. p. 23).

[Fri05] Karl Friston. "A theory of cortical responses". In : *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 360.1456 (2005), p. 815-836 (cf. p. 76).

[Iac+05] Marco Iacoboni, Istvan Molnar-Szakacs, Vittorio Gallese, Giovanni Buccino, John C. Mazziotta et al. "Grasping the Intentions of Others with One's Own Mirror Neuron System". In : *PLoS Biology* 3.3 (fév. 2005), e79 (cf. p. 44).

[Jen05] Jen-Tzung Chien. “Decision tree State tying using cluster validity criteria”. In : *IEEE Transactions on Speech and Audio Processing* 13.2 (mar. 2005), p. 182-193 (cf. p. 38).

[LKS05] Min Li, Chandra Kambhamettu et Maureen Stone. “Automatic contour tracking in ultrasound images”. In : *Clinical Linguistics and Phonetics* (2005) (cf. p. 60).

[Mai+05] Lena Maier-Hein, Florian Metze, Tanja Schultz et Alex Waibel. “Session Independent Non-Audible Speech Recognition Using Surface Electromyography”. In : *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*. San Juan, Puerto Rico, 2005, p. 331-336 (cf. p. 9).

[OL05] Slim Ouni et Yves Laprie. “Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion”. In : *The Journal of the Acoustical Society of America* 118.1 (2005), p. 444-460 (cf. p. 55).

[Tak+05] Hirokazu Takahashi, Masayuki Nakao, Yataro Kikuchi et Kimitaka Kaga. “Alaryngeal speech aid using an intra-oral electrolarynx and a miniature fingertip switch”. In : *Auris Nasus Larynx* 32.2 (juin 2005), p. 157-162 (cf. p. 23).

[THS05] Erik D. Thiessen, Emily A. Hill et Jenny R. Saffran. “Infant-Directed Speech Facilitates Word Segmentation”. In : *Infancy* 7.1 (jan. 2005), p. 53-71 (cf. p. 76).

[Cap+04] Alice Caplier, Laurent Bonnaud, Sotiris Malassiotis et Michael G Strintzis. “Comparison of 2D and 3D analysis for automated cued speech gesture recognition”. In : *Proc. of Conference on Speech and Computer (SPECOM)*. Saint-Petersburg, Russia, 2004 (cf. p. 41).

[DH04] Jianwu Dang et Kiyoshi Honda. “Construction and control of a physiological articulatory model.” In : *The Journal of the Acoustical Society of America* 115.2 (fév. 2004), p. 853-70 (cf. p. 8).

[DLH04] Randy L Diehl, Andrew J Lotto et Lori L Holt. “Speech perception”. In : *Annu. Rev. Psychol.* 55 (2004), p. 149-179 (cf. p. 44).

[GM04] Jesús Giménez et Lluis Marquez. “SVMTool : A general POS tagger generator based on Support Vector Machines”. In : *Proc. of International Conference on Language Resources and Evaluation*. Lisbon, Portugal, 2004, p. 43-46 (cf. p. 35).

[Gow+04] John N Gowdy, Amarnag Subramanya, Chris Bartels et Jeff Bilmes. “DBN-based multi-stream models for audio-visual speech recognition”. In : *Proc. of ICASSP*. T. 1. Montreal, Quebec, Canada, 2004, p. 993-996 (cf. p. 6).

[HH04] Sadao Hiroya et Masaaki Honda. “Estimation of Articulatory Movements From Speech Acoustics Using an HMM-Based Speech Production Model”. In : *IEEE Transactions on Speech and Audio Processing* 12.2 (2004), p. 175-185 (cf. p. 18).

[ML04] Dominic W. Massaro et Joanna Light. “Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss”. In : *Journal of Speech, Language, and Hearing Research* 47.2 (avr. 2004), p. 304-320 (cf. p. 49).

[MP04] Roberto Merletti et Philip A Parker. *Electromyography : physiology, engineering, and non-invasive applications*. T. 11. John Wiley & Sons, 2004, p. 520 (cf. p. 8).

[SMJ04] M.A. Schelstraete, C. Maillart et A-C. Jamart. "Les troubles phonologiques : cadre théorique, diagnostic et traitement". In : *Les troubles du langage et du calcul chez l'enfant*. Sous la dir. de M.A Schelstraete et M.P Noel. EME. Inter. Fernelmont, Belgique, 2004, p. 81-112 (cf. p. 67).

[Ber+03] Barbara Bernhardt, Bryan Gick, Penelope Bacsfalvi et Julie Ashdown. "Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners". In : *Clinical Linguistics and Phonetics* (2003) (cf. p. 50).

[Fow+03] C Fowler, J. Brown, L. Sabadini et J Weihing. "Rapid access to speech gestures in perception : Evidence from choice and simple response time tasks". In : *Journal of Memory and Language* 49.3 (oct. 2003), p. 396-413 (cf. p. 70).

[Gol+03] E.A. Goldstein, J.T. Heaton, J.B. Kobler, G.B. Stanley et R.E. Hillman. "Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity". In : *Proc. of IEEE EMBS Conference on Neural Engineering*. Capri Island, Italy : IEEE, 2003, p. 169-172 (cf. p. 23).

[Nak+03] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano et Nick Campbell. "Non-Audible Murmur Recognition Input Interface using Stethoscopic Microphone Attached to the Skin". In : *Proc. of ICASSP*. Hong Kong : IEEE, 2003, p. 127-130 (cf. p. 9).

[Pot+03] G Potamianos, C Neti, G Gravier, A Garg et A W Senior. "Recent advances in the automatic recognition of audiovisual speech". In : *Proceedings of the IEEE* 91.9 (sept. 2003), p. 1306-1326 (cf. p. 6, 13).

[Sch03] K Scherer. "Vocal communication of emotion : A review of research paradigms". In : *Speech Communication* 40.1-2 (avr. 2003), p. 227-256 (cf. p. 23).

[Bad+02] Pierre Badin, Gérard Bailly, Lionel Revéret, Monica Baciu, Christoph Segebarth et al. "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images". In : *Journal of Phonetics* 30.3 (juil. 2002), p. 533-553 (cf. p. 8).

[GPN02] G Gravier, G Potamianos et C Neti. "Asynchrony modeling for audio-visual speech recognition". In : *Proc. of International Conference on Human Language Technology Research*. San Diego, California : Morgan Kaufmann Publishers Inc., 2002, p. 1-6 (cf. p. 6).

[Hec+02] Martin Heckmann, Kristian Kroschel, Christophe Savariaux, Frédéric Berthommier et al. "DCT-based video features for audio-visual speech recognition." In : *Proc. of Interspeech*. T. 3. Denver, Colorado, USA : ISCA, 2002, p. 1925-1928 (cf. p. 6).

[Ric02] Korin Richmond. "Estimating articulatory parameters from the speech signal". Thèse de doct. University of Edinburgh, 2002 (cf. p. 55).

[RRB02] Carolyn P Rosé, Antonio Roque et Dumisizwe Bhembe. "An efficient incremental architecture for robust interpretation". In : *Proc. of International Conference on Human Language Technology Research*. San Diego, California : Morgan Kaufmann Publishers Inc., 2002, p. 307-312 (cf. p. 34).

[BBB01] Denis Beautemps, Pierre Badin et Gérard Bailly. "Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling". In : *The Journal of the Acoustical Society of America* 109.5 (mai 2001), p. 2165-2180 (cf. p. 27, 29).

[Geu01] Anja Geumann. "Vocal intensity : acoustic and articulatory correlates". In : *Proc. of International Speech Motor Conference*. Nijmegen, Netherlands, 2001, p. 70-73 (cf. p. 23).

[Gib+01] Fiona Gibbon, William J Hardcastle, Lisa Crampin, Beverley Reynolds, Roz Razzell et al. "Visual feedback therapy using electropalatography (EPG) for articulation disorders associated with cleft palate". In : *Asia Pacific Journal of Speech, Language and Hearing* 6.1 (jan. 2001), p. 53-58 (cf. p. 50).

[Mat+01] Iain Matthews, Gerasimos Potamianos, Chalapathy Neti et Juergen Luettin. "A comparison of model and transform-based visual features for audio-visual LVCSR". In : *Proc. of International Conference on Multimedia and Expo*. Tokyo, Japan : IEEE, 2001, p. 825-828 (cf. p. 6).

[MMI01] Daisuke Mori, Shigeki Matsubara et Yasuyoshi Inagaki. "Incremental parsing for interactive natural language interface". In : *Proc. of International Conference on Systems, Man, and Cybernetics*. T. 5. Tucson, AZ, USA : IEEE, 2001, p. 2880-2885 (cf. p. 34).

[Sch01] Jean-Luc Schwartz. "Une théorie de la perception pour le contrôle de l'action". In : *Percevoir : monde et langage. Invariance et variabilité du sens vécu*. Sous la dir. de J-Fr Bonnot, JP Durafour, D Keller et R Sock. Editions M. Bruxelles, 2001. Chap. 14, p. 260-270 (cf. p. 44).

[Sty01] Yannis Stylianou. "Applying the harmonic plus noise model in concatenative speech synthesis". In : *IEEE Transactions on Speech and Audio Processing* 9.1 (2001), p. 21-29 (cf. p. 51, 70).

[Bra00] Thorsten Brants. "TnT : a statistical part-of-speech tagger". In : *Proc. of Conference on Applied natural language processing*. Association for Computational Linguistics. Seattle, WA, USA, 2000, p. 224-231 (cf. p. 34).

[CSB00] Catia Cucchiarini, Helmer Strik et Lou Boves. "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms". In : *Speech Communication* 30.2-3 (fév. 2000), p. 109-119 (cf. p. 67).

[DL00] S Dupont et J Luettin. "Audio-visual speech modeling for continuous speech recognition". In : *IEEE Transactions on Multimedia* 2.3 (sept. 2000), p. 141-151 (cf. p. 6).

[Mid+00] Matthew Middendorf, Grant McMillan, Gloria Calhoun et K.S. Jones. "Brain-computer interfaces based on the steady-state visual-evoked response". In : *IEEE Transactions on Rehabilitation Engineering* 8.2 (juin 2000), p. 211-214 (cf. p. 26).

[Tok+00] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi et Tadashi Kitamura. "Speech parameter generation algorithms for HMM-based speech synthesis". In : *Proc. of ICASSP*. T. 3. Istanbul, Turkey : IEEE, 2000, p. 1315-1318 (cf. p. 16).

[WY00] S.M Witt et S.J Young. "Phone-level pronunciation scoring and assessment for interactive language learning". In : *Speech Communication* 30.2-3 (fév. 2000), p. 95-108 (cf. p. 67).

[WR00] Alan Wrench et Korin Richmond. "Continuous speech recognition using articulatory data". In : *Proc. of International Conference on Spoken Language Processing (ICSLP)* (2000), p. 145-148 (cf. p. 7).

[Sav+99] Christophe Savariaux, Pascal Perrier, Jean-Pierre Orliaguet et Jean-Luc Schwartz. "Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis." In : *The Journal of the Acoustical Society of America* (1999) (cf. p. 45).

[SSL99] Alain Soquet, Marco Saerens et Véronique Lecuit. "Complementary cues for speech recognition". In : *Proc. of International Conference on Phonetic Sciences (ICPhS)*. San Francisco, CA, USA, 1999, p. 1645-1648 (cf. p. 7).

[DK98] Chris Davis et Jeesun Kim. "Repeating and Remembering Foreign Language Words : Does seeing help ?" In : *Proc. of International Conference on Auditory-Visual Speech Processing (AVSP)*. Sydney, Australia, 1998, p. 121-126 (cf. p. 70).

[KM98] A. Kain et M.W. Macon. "Spectral voice conversion for text-to-speech synthesis". In : *Proc. of ICASSP*. T. 1. Seattle, Washington, USA : IEEE, 1998, p. 285-288 (cf. p. 55).

[LR98] Li Lee et Richard Rose. "A frequency warping approach to speaker normalization". In : *IEEE Transactions on Speech and Audio Processing* 6.1 (1998), p. 49-60 (cf. p. 86).

[SCM98] Yannis Stylianou, Olivier Cappé et Eric Moulines. "Continuous probabilistic transform for voice conversion". In : *IEEE Transactions on Speech and Audio Processing* 6.2 (1998), p. 131-142 (cf. p. 55).

[DRS97] L. Deng, G. Ramsay et D. Sun. "Production models as a structural basis for automatic speech recognition". In : *Speech Communication* 22.2-3 (août 1997), p. 93-111 (cf. p. 75).

[BS96] Rainer Banse et Klaus R. Scherer. "Acoustic profiles in vocal emotion expression." In : *Journal of personality and social psychology* 70.3 (mar. 1996), p. 614-36 (cf. p. 23).

[BA96] Lynne E Bernstein et Edward T Auer. "Word Recognition in Speechreading BT - Speechreading by Humans and Machines : Models, Systems, and Applications". In : sous la dir. de David G Stork et Marcus E Hennecke. Berlin, Heidelberg : Springer Berlin Heidelberg, 1996, p. 17-26 (cf. p. 7).

[LC96] Mark Y Liberman et Kenneth W Church. "Text Analysis and Word Pronunciation in Text-to-speech Synthesis". In : *Advances in Speech Signal Processing*. Sous la dir. de S. Furui et MM. Sondhi. Dekker. New York, New York, USA, 1996, p. 791-831 (cf. p. 39).

[LTB96] Juergen Luettin, Neil A Thacker et Steve W Beet. "Visual speech recognition using active shape models and hidden Markov models". In : *Proc. of ICASSP*. T. 2. Atlanta, Georgia, USA, 1996, p. 817-820 (cf. p. 5).

[Moo96] Roger K. Moore. "Critique : The potential role of speech production models in automatic speech recognition". In : *The Journal of the Acoustical Society of America* 99.3 (mar. 1996), p. 1710-1713 (cf. p. 75).

[Riz+96] Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese et Leonardo Fogassi. "Premotor cortex and the recognition of motor actions". In : *Cognitive Brain Research* 3.2 (mar. 1996), p. 131-141 (cf. p. 44).

[RSS96] R. C. Rose, J. Schroeter et M. M. Sondhi. "The potential role of speech production models in automatic speech recognition". In : *The Journal of the Acoustical Society of America* 99.3 (mar. 1996), p. 1699-1709 (cf. p. 75).

[Gue95] Frank H. Guenther. "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production." In : *Psychological Review* 102.3 (1995), p. 594-621 (cf. p. 78).

[LW95] Legetter.C.J et Woodland.P.C. "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models". In : *Computer Speech and Language* 9.2 (1995), p. 171-185 (cf. p. 53, 55).

[SPO95] Christophe Savariaux, Pascal Perrier et Jean Pierre Orliaguet. "Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube : A study of the control space in speech production". In : *The Journal of the Acoustical Society of America* 98.5 (nov. 1995), p. 2428-2442 (cf. p. 45).

[BMK94] Christian Benoît, Tayeb Mohamadi et Sonia Kandel. "Effects of Phonetic Context on Audio-Visual Intelligibility of French". In : *Journal of Speech, Language, and Hearing Research* 37.5 (oct. 1994), p. 1195-1203 (cf. p. 48).

[BK94] C Bregler et Y Konig. "Eigenlips for robust speech recognition". In : *Proc. of ICASSP*. T. 2. Adelaide, South Australia, Australia : IEEE, 1994, p. 669-672 (cf. p. 6, 61).

[GC94] J.-L. Gauvain et Chin-Hui Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains". In : *IEEE Transactions on Speech and Audio Processing* 2.2 (avr. 1994), p. 291-298 (cf. p. 55).

[GJ94] Zoubin Ghahramani et Michael I Jordan. “Supervised learning from incomplete data via an EM approach”. In : *Advances in Neural Information Processing Systems* 6.1 (1994), p. 120-127 (cf. p. 16).

[YOW94] S J Young, J J Odell et P C Woodland. “Tree-based State Tying for High Accuracy Acoustic Modelling”. In : *Proc. of Workshop on Human Language Technology*. Stroudsburg, PA, USA, 1994, p. 307-312 (cf. p. 37).

[BA92] Gérard Bailly et Mamoun Alissali. “Compost : un serveur de synthèse de parole multilingue”. In : *Traitement du Signal* 9.4 (1992), p. 359-366 (cf. p. 35).

[Pel+92] G. di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese et G. Rizzolatti. “Understanding motor events : a neurophysiological study”. In : *Experimental Brain Research* 91.1 (oct. 1992), p. 176-180 (cf. p. 44).

[SHY92] Richard Sproat, Julia Hirschberg et David Yarowsky. “A Corpus-Based Synthesizer”. In : *Proc. of ICSLP*. Banff, Alberta, Canada : ISCA, 1992, p. 563-566 (cf. p. 39).

[Mae90] Shinji Maeda. “Compensatory Articulation During Speech : Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model”. In : *Speech Production and Speech Modelling*. Dordrecht : Springer Netherlands, 1990, p. 131-149 (cf. p. 8).

[Cal89] Calliope. *La parole et son traitement automatique*. Masson. Paris, 1989 (cf. p. 71).

[FD88] L.A. Farwell et E. Donchin. “Talking off the top of your head : toward a mental prosthesis utilizing event-related brain potentials”. In : *Electroencephalography and Clinical Neurophysiology* 70.6 (déc. 1988), p. 510-523 (cf. p. 25).

[KFS87] M. Kawato, Kazunori Furukawa et R. Suzuki. “A hierarchical neural-network model for control and learning of voluntary movement”. In : *Biological Cybernetics* 57.3 (oct. 1987), p. 169-185 (cf. p. 46, 78).

[MS87] Man Sondhi et Juergen Schroeter. “A hybrid time-frequency domain articulatory speech synthesizer”. In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35.7 (juil. 1987), p. 955-967 (cf. p. 8).

[RMG87] Daniel Reisberg, John McLean et Anne Goldfield. “Easy to hear but hard to understand : A lipreading advantage with intact auditory stimuli”. In : *Hearing by eye : The psychology of lipreading*. Sous la dir. de B. Dodd et R. Cambell. Lawrence E. London, 1987, p. 97-113 (cf. p. 70).

[LM85] Alvin M. Liberman et Ignatius G. Mattingly. “The motor theory of speech perception revised”. In : *Cognition* 21.1 (oct. 1985), p. 1-36 (cf. p. 44, 75).

[Mar85] William D. Marslen-Wilson. “Speech shadowing and speech comprehension”. In : *Speech Communication* 4.1-3 (1985), p. 55-73 (cf. p. 70).

[Pet84] E D Petajan. “Automatic Lipreading to Enhance Speech Recognition”. In : *Proc. of the IEEE Communication Society Global Telecommunications Conference*. 1984 (cf. p. 5).

[ISF83] Satoshi Imai, Kazuo Sumita et Chieko Furuichi. “Mel Log Spectrum Approximation (MLSA) filter for speech synthesis”. In : *Electronics and Communications in Japan (Part I : Communications)* 66.2 (1983), p. 10-18 (cf. p. 8, 15, 21, 36).

[Mae82] Shinji Maeda. “A digital simulation method of the vocal-tract system”. In : *Speech Communication* 1.3-4 (déc. 1982), p. 199-229 (cf. p. 8).

[Ros+81] Mario Rossi, Albert Di Cristo, Daniel Hirst, Philippe Martin et Yukihiro Nishinuma. *L'intonation : de l'acoustique à la sémantique*. Klincksiec. Paris, 1981 (cf. p. 34).

[PC80] Robert J. Porter et F. Xavier Castellanos. “Speech-production measures of speech perception : Rapid shadowing of VCV syllables”. In : *The Journal of the Acoustical Society of America* 67.4 (avr. 1980), p. 1349-1356 (cf. p. 70).

[Ata+78] Bishnu S Atal, J J Chang, Max V Mathews et John W Tukey. “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique”. In : *The Journal of the Acoustical Society of America* 63.5 (1978), p. 1535-1555 (cf. p. 55).

[MM76] Harry McGurk et John Macdonald. “Hearing lips and seeing voices.” In : *Nature* 264.5588 (déc. 1976), p. 691-811 (cf. p. 47).

[DAB75] F. L. Darley, A. E. Aronson et J. R. Brown. *Motor speech disorders*. Philadelphia, PA, USA : Saunders, 1975, p. 304 (cf. p. 87).

[LP70] James F. Lubker et Pamela J. Parris. “Simultaneous Measurements of Intraoral Pressure, Force of Labial Contact, and Labial Electromyographic Activity during Production of the Stop Consonant Cognates /p/ and /b/”. In : *The Journal of the Acoustical Society of America* 47.2B (fév. 1970), p. 625-633 (cf. p. 23).

[Cor67] Richard Orin Cornett. “Cued speech”. In : *American annals of the deaf* 112.1 (1967), p. 3-13 (cf. p. 41).

[Bar61] Horace B Barlow. “Possible principles underlying the transformations of sensory messages”. In : *Sensory Communication*. Sous la dir. de W A Rosenblith. Cambridge, MA : MIT press, 1961, p. 217-234 (cf. p. 76).

[Fuj61] Osamu Fujimura. “Bilabial Stop and Nasal Consonants : a Motion Picture Study and its Acoustical Implications”. In : *Journal of Speech and Hearing Research* 4.3 (sept. 1961), p. 233-247 (cf. p. 23).

[Mey56] Werner Meyer-Eppler. “Realization of Prosodic Features in Whispered Speech”. In : *The Journal of the Acoustical Society of America* 28.4 (juil. 1956), p. 760-760 (cf. p. 23).

[Att54] Fred Attneave. “Some informational aspects of visual perception.” In : *Psychological review* 61.3 (1954), p. 183 (cf. p. 76).

[SP54] W. H. Sumby et Irwin Pollack. “Visual Contribution to Speech Intelligibility in Noise”. In : *The Journal of the Acoustical Society of America* 26.2 (mar. 1954), p. 212-215 (cf. p. 48).

Résumé — Mes activités de recherche portent sur le traitement automatique de la parole, avec un intérêt particulier pour la capture, l'analyse et la modélisation des gestes articulatoires et des signaux électrophysiologiques impliqués lors de sa production. Mes travaux visent à développer des technologies vocales qui exploitent ces différents signaux, pour la reconnaissance automatique et la synthèse de la parole, à destination notamment des personnes présentant un trouble de la communication parlée. Plus spécifiquement, ces technologies visent soit à rétablir la capacité à communiquer oralement lorsqu'une partie de la chaîne de production de la parole est défaillante (suppléance vocale), soit à faciliter la rééducation orthophonique d'un trouble phonétique ou phonologique (rééducation articulatoire assistée). La méthodologie générale sur laquelle je m'appuie est principalement basée sur la mise en place de dispositifs et protocoles expérimentaux pour l'acquisition de signaux multimodaux (par exemple vidéo et audio), la construction par apprentissage automatique (*machine learning*) de modèles prédictifs permettant la mise en relation de ces différents signaux, et enfin l'utilisation de ces modèles dans des systèmes temps-réel qui viennent interagir avec les boucles sensorimotrices qui régulent la perception et le contrôle moteur de la parole.

Mots clés : parole, multimodale, signal, apprentissage automatique, systèmes temps-réel, handicap.

Abstract — My research activities deal with the automatic processing of speech, with a special interest in capturing and modeling the articulatory gestures and the different electrophysiological signals involved in speech production. My goal is to develop automatic speech recognition and synthesis systems that exploit these multimodal signals for people with communication disorders. More precisely, these systems aim either at restoring oral communication when parts of the speech production chain is damaged (speech prosthesis), or at facilitating the treatment of speech sound disorders (assisted speech therapy). To build such systems, my approach is to capture multimodal speech-related signals using a variety of experimental devices, to model the statistical relationships between those signals using machine learning, and finally to implement these models in real-time systems that can interact with the low-level sensorimotor loops involved in speech perception and speech motor control.

Keywords : speech processing, multimodal, signal, machine learning, real-time systems, handicap.

GIPSA-lab (UMR 5216), 11 rue des Mathématiques
Saint Martin d'Hères