Apprentissage auto-supervisé des relations entre sons, gestes articulatoires et unités de la parole pour le contrôle de la production : vers un agent apprenant à parler

Marc-Antoine Georges

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : **ISCE - Ingénierie pour la Santé, la Cognition et l'Environnement**

Spécialité : **CIA - Ingénierie de la Cognition, de l'Interaction, de l'Apprentissage et de la création**

Unité de recherche : **Laboratoire Grenoble Images Parole Signal & Automatique**

Apprentissage auto-supervisé des relations entre sons, gestes articulatoires et unités de la parole pour le contrôle de la production : vers un agent apprenant à parler

Self-supervised learning of the relationships between sounds, gestures and units for the control of speech production: towards an agent learning to speak

Présentée par :

Marc-Antoine GEORGES

Direction de thèse :

Jean-Luc SCHWARTZ Directeur de Recherche, CNRS Délégation Alpes	Directeur de thèse
Thomas HUEBER Directeur de Recherche, CNRS Délégation Alpes	Co-directeur de thèse
Laurent GIRIN Professeur des Universités, Grenoble INP	Co-encadrant de thèse

Thèse soutenue publiquement le **31 mai 2023,** devant le jury composé de :

Xavier HINAUT Chargé de recherche, INRIA Bordeaux	Rapporteur
Yves LAPRIE Directeur de recherche, CNRS Délégation Centre-Est	Rapporteur
Emmanuel DUPOUX Directeur d'Études, EHESS Paris	Examinateur
Clément MOULIN-FRIER Chargé de Recherche, INRIA Bordeaux	Examinateur
Anne GUÉRIN-DUGUÉ Professeure des Universités, Université Grenoble Alpes	Présidente
Jean-Luc SCHWARTZ Directeur de Recherche, CNRS Délégation Alpes	Directeur de thèse
Thomas HUEBER Directeur de Recherche, CNRS Délégation Alpes	Co-directeur de thèse

Remerciements

Ce manuscrit n'aurait pas vu le jour sans les contributions directes et indirectes de nombreuses personnes que j'ai eu la chance de côtoyer au cours de ces dernières années. Je souhaite aujourd'hui vous adresser toute ma gratitude.

En premier lieu aux membres de mon jury de thèse, merci à chacun d'entre vous d'avoir accepté de vous plonger dans mon travail, merci pour vos retours, et merci pour les discussions passionnantes le jour de ma soutenance.

Ensuite, bien entendu, merci à vous Jean-Luc et Thomas. Merci Jean-Luc d'avoir toujours été présent pour répondre à mes questions, merci de m'avoir fait confiance dès le départ et surtout merci d'avoir toujours écouté mon avis, même lorsqu'il était à contresens de la majorité. Merci Thomas pour ton énergie, merci de m'avoir aidé à sortir de moult galères et surtout merci de m'avoir autant poussé à donner le meilleur de moi-même. Merci également à vous, Laurent et Julien. Malheureusement l'équipe originale n'aura pas pu rester éternellement au complet mais j'ai quand même beaucoup apprécié ces moments car même si c'était le chaos, ça restait un chaos stimulant. Chacun de vous quatre est vraiment au top, autant scientifiquement qu'humainement, je n'aurai pas pu rêver mieux comme encadrement. Mes modèles internes de vous vont m'être utile pendant encore longtemps !

Je souhaiterais ensuite remercier toutes les personnes avec qui j'ai pu partager un moment au laboratoire. Plus spécialement merci à Pierre et Olivier, d'avoir toujours gardé la porte de votre bureau ouverte à mes questions. Et surtout merci Hélène, la meilleure co-bureau ! Merci pour ces discussions, ces donuts et ces rires partagés, c'était vraiment cool. Merci aussi d'avoir été un indicateur météo des états psychologiques que j'allais traverser lors de ma fin de thèse.

Merci à ma mère de m'avoir toujours fait confiance, je ne serai jamais allé aussi loin sans ça. Merci à mes frères et sœurs, beaux et pas beaux, on aura bien rigolé et ça m'a beaucoup aidé à garder le cap. Merci à toi Cyril, de m'avoir écouté parler avec insistance des vocodeurs neuronaux et des auto-encodeurs variationnels. Merci à toi Mickaël pour toutes ces soirées à tuer des zombies et pour le soutient lors des moments difficiles. Merci à toi Mathis pour nos discussions qui m'ont toujours poussé à voir plus grand et qui ont contribué à mon arrivée jusqu'ici. Enfin, merci à toi Magali, je ne sais pas si j'aurai pu aller jusqu'au bout de cette thèse sans tes enseignements.

Pour conclure cette section qui commence à ressembler de plus en plus à un article dédicace sur Skyblog, je termine en remerciant toute l'équipe du French Coffee Shop, dans lequel la majorité des lignes de ce manuscrit ont été écrites.

Table des matières

Table des figures

Table des sigles et acronymes

AE	Auto-encodeur
EMA	*Electromagnetic articulography*
HMM	*Hidden Markov model*
LSTM	*Long short-term memory*
MUSHRA	*Multiple Stimuli with Hidden Reference and Anchor*
PCA	*Principal component analysis* (Analyse en composantes principales)
REQM	Racine de l'erreur quadratique moyenne
VAE	*Variational autoencoder* (Auto-encodeur variationnel)
VQ-VAE	*Vector quantized variational autoencoder* (Auto-encodeur variationnel quantifié vectoriel)

Introduction

La production de la parole est un processus moteur complexe qui nécessite un contrôle fin de l'appareil vocal, et notamment des différents articulateurs de la parole que sont les lèvres, la langue, la mâchoire et le voile du palais pour transformer une source acoustique en une parole intelligible. Pour permettre ce contrôle et interagir avec son environnement pendant les premiers stades de son développement, l'enfant doit apprendre à inférer un geste articulatoire à partir d'un but acoustique qu'il souhaite atteindre. Il tente donc de résoudre un problème dit d'inversion acoustico-articulatoire et plus globalement sensori-motrice, qui consiste à retrouver un geste moteur à partir d'un son de parole perçu. Il s'agit d'un problème complexe car mal posé, en raison du caractère non-linéaire et surtout non bijectif de la relation entre le contenu spectral du signal de parole d'une part, et la configuration de l'appareil vocal associée d'autre part (un même son peut être obtenu à l'aide de plusieurs configurations différentes). Ce problème classique du traitement de la parole a notamment été abordé à l'aide de techniques de modélisation par apprentissage automatique (*machine learning*), très souvent à l'aide d'un paradigme d'apprentissage dit « supervisé », sur la base de données acquises *in vivo*, associant, à chaque instant, observations acoustique et articulatoire. Ce paradigme d'apprentissage supervisé semble très éloigné de celui de l'humain qui découvre ces relations complexes uniquement à partir des sons qu'il perçoit, *a priori* sans informations articulatoires lui indiquant comment piloter son appareil vocal (à l'exception d'informations visuelles sur les mouvements faciaux des interlocuteurs, si elles sont disponibles). Par conséquent, du point de vue de la modélisation statistique, l'apprentissage de la relation acoustique-vers-articulatoire peut être plutôt considéré comme un processus faiblement supervisé.

L'étude de ce processus a généré un grand nombre de travaux dans le domaine de la modélisation des processus physiques et cognitifs de la communication parlée, et a notamment mené à des développements computationnels dans trois directions principales, avec la création : (1) de synthétiseurs articulatoires, permettant de générer des formes de conduit vocal et des sons de parole à partir d'un ensemble restreint de paramètres interprétables phonétiquement ; (2) de modèles statistiques apprenant la relation articulatoire-acoustique directe ou inverse à partir d'enregistrements articulatoires *in vivo* ; et (3) d'architectures de contrôle moteur de la parole cherchant à modéliser la façon dont le cerveau humain exploite les représentations internes acoustiques et articulatoires dans la production de la parole. Cependant, la plupart de ces modèles de contrôle sont entraînés et testés sur des données élémentaires, non représentatives de la complexité de la parole réelle, telles que des voyelles isolées, des syllabes simples ou des données synthétiques.

Une autre ligne de recherche récente s'est concentrée sur l'apprentissage auto-supervisé de représentations de la parole à partir de grands ensembles de données audio non étiquetées. Ces représentations permettent de former des modèles génératifs de parole qui peuvent être utilisés pour reproduire et étendre un ensemble de *stimuli* audio. Cependant, les systèmes développés n'intègrent pas explicitement de connaissances articulatoires sur le processus de production de la parole. Par conséquent, bien qu'ils soient puissants et efficaces pour générer des sons

vocaux, ils ne peuvent pas fournir beaucoup d'informations sur les mécanismes sous-jacents impliqués dans la production de la parole.

Le travail présenté dans cette thèse propose des avancées à la croisée de ces deux axes de recherche, via une série de simulations focalisées sur le lien entre les modalités acoustique et articulatoire de la parole. L'ensemble de ces simulations est effectué sur des données de parole « réelles » et est basé sur des réseaux de neurones profonds. La contribution principale de ce travail est la création d'un agent capable d'apprendre « à parler », en répétant des *stimuli* auditifs à l'aide de son « appareil vocal virtuel ».

Pour concevoir et implémenter cet agent, nous utilisons dans ce travail la modélisation computationnelle et la simulation, à l'aide notamment de techniques d'apprentissage automatique profond auto-supervisé. Concrètement, la mise en œuvre d'un tel agent s'appuie sur trois contributions principales. D'abord, afin de doter cet agent d'un « appareil vocal virtuel », la première contribution est l'élaboration complète d'un synthétiseur articulatoire, générant des sons proches de ceux d'un locuteur de référence et contrôlé par des paramètres articulatoires interprétables. Ensuite, l'agent étant capable de se construire, de façon auto-supervisée, un dictionnaire d'unités phonétiques discrètes en exploitant l'acoustique de la parole perçue et, le cas échéant, les gestes articulatoires sous-jacents, la seconde contribution vise à mieux comprendre comment ce processus peut s'effectuer avec un minimum de supervision – c'est-à-dire sans disposer d'une segmentation temporelle du flux audio de parole – et comment l'inférence de représentations articulatoires peut le faciliter. Enfin, sur ces bases, une architecture complète d'agent a été conçue, intégrant un modèle inverse estimant les paramètres de contrôle du synthétiseur afin d'imiter des *stimuli* auditifs perçus, ou de les répéter au sens du contenu des représentations phonétiques inférées. Cette architecture s'inspire de certains modèles théoriques du contrôle moteur de la parole (et du mouvement en général). Le modèle inverse (ainsi que d'autres composantes de l'agent) est entraîné de façon auto-supervisée, uniquement à partir de *stimuli* acoustiques de parole naturelle, issus de différents locuteurs. La troisième contribution porte ainsi sur une description précise des fondements de cette architecture et de ses performances d'imitation et de répétition, sur plusieurs locuteurs cibles.

Ce manuscrit de thèse est organisé comme suit :

- Chapitre 1 : Revue de la littérature sur les différentes théories et modèles qui sous-tendent la perception et le contrôle moteur de la parole, la modélisation acoustico-articulatoire (inversion et synthèse), et l'apprentissage auto-supervisé de représentations.
- Chapitre 2 : Présentation des différents jeux de données utilisés dans nos simulations.
- Chapitre 3 : Construction d'un synthétiseur articulatoire basé sur une architecture neuronale, pilotable par un nombre restreint de paramètres interprétables d'un point de vue phonétique (Georges et al., 2020).
- Chapitre 4 :
  - Régularisation de l'espace latent d'un auto-encodeur variationnel par des contraintes articulatoires, et évaluation de celui-ci sur une tâche de débruitage de la parole (Georges et al., 2021).
  - Découverte auto-supervisée d'unités de la parole à partir de données acoustiques

et/ou articulatoires à l'aide d'un auto-encodeur variationnel quantifié vectoriel (VQ-VAE), et caractérisation de ces unités au niveau phonétique (Georges, Schwartz et al., 2022).

— Chapitre 5 : Architectures neuronales pour la création d'un agent communiquant, intégrant le synthétiseur articulatoire neuronal construit au chapitre 3, deux modèles internes, l'un direct et l'autre inverse, ainsi qu'un mécanisme de découverte d'unités de la parole, appris de façon auto-supervisée (Georges, Schwartz et al., 2022).

— Chapitre 6 : Ce manuscrit se termine par une discussion résumant les différentes contributions et présentant des perspectives qui permettraient de poursuivre leur développement.

Cette thèse a été menée au sein du pôle Parole et Cognition du GIPSA-lab (équipes CRISSP et PCMD) à Grenoble, et a fait l'objet d'une collaboration avec le LPNC, également à Grenoble. Ce travail a été financé par l'institut d'intelligence artificielle 3IA MIAI (ANR-19-P3IA-0003) au sein de la chaire *Bayesian Cognition and Machine Learning for Speech Communication*.

Ce travail a fait l'objet des publications suivantes :

— Georges, M.-A., Badin, P., Diard, J., Girin, L., Schwartz, J.-L. & Hueber, T. (2020). Towards an articulatory-driven neural vocoder for speech synthesis. *International Seminar on Speech Production (ISSP)*

— Georges, M.-A., Girin, L., Schwartz, J.-L. & Hueber, T. (2021). Learning robust speech representation with an articulatory-regularized variational autoencoder. *Conference of the International Speech Communication Association (Interspeech)*, 3345-3349

— Georges, M.-A., Diard, J., Girin, L., Schwartz, J.-L. & Hueber, T. (2022). Repeat after Me : Self-Supervised Learning of Acoustic-to-Articulatory Mapping by Vocal Imitation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8252-8256

— Georges, M.-A., Schwartz, J.-L. & Hueber, T. (2022). Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE. *Conference of the International Speech Communication Association (Interspeech)*, 774-778

Chapitre 1

État de l'art

Sommaire

Notre projet de développement d'un agent autonome capable d'apprendre à parler à partir d'exemples fournis par son environnement s'appuie sur un ensemble de concepts, de modèles et d'outils provenant à la fois de la cognition humaine (avec les processus de contrôle moteur de la parole), du traitement automatique de la parole et des mécanismes d'apprentissage par réseaux de neurones profonds. Nous allons dans ce premier chapitre présenter une revue de la littérature relative à ce contexte large, revue de laquelle découlent nos choix méthodologiques.

1.1 Contrôle de la parole

Les modèles théoriques du contrôle moteur font généralement appel à trois éléments : la tâche, le contrôleur et le *plant*, c'est-à-dire le système physique à contrôler. Dans ce formalisme, la tâche à réaliser (« quoi faire ? ») est prise en charge par le contrôleur, dont la responsabilité

est la planification de la variation de variables de contrôle (« comment faire ? ») servant à piloter le *plant* (« avec quoi le faire ? »).

Dans le cadre de la production de la parole la tâche est la transmission d'informations à l'interlocuteur, le *plant* est l'appareil phonatoire, et le contrôleur est l'ensemble des processus en charge de l'évolution temporelle des variables de contrôle du *plant* pour réaliser la tâche, typiquement implémentés au niveau du système nerveux central. Cette définition étant générale, la suite de cette section a pour but de détailler davantage les enjeux liés à la tâche et au contrôleur. Nous renvoyons pour la description précise du *plant* à la section suivante 1.2.

1.1.1 La tâche

En production de la parole, la tâche est la transmission d'informations à un interlocuteur. Afin de définir plus précisément celle-ci, il convient de se demander : de quelle nature sont les représentations mises en jeu lorsque nous percevons et produisons de la parole ?

Les théories de la perception Historiquement, c'est d'abord dans le domaine de la perception de la parole qu'a été posée la question de la nature des représentations partagées entre le locuteur et l'auditeur en s'intéressant à la façon dont la parole est décodée. Deux grandes familles de théories se sont ainsi peu à peu construites en opposition : les théories motrices et les théories auditives.

En partant du délicat problème de la recherche d'invariants associés aux unités phonologiques (phonèmes ou traits), les théories motrices se sont construites autour de Liberman depuis le début des années 60 (voir par exemple Liberman et Mattingly, 1985), sur l'hypothèse forte que l'invariance et les corrélats phonétiques ne devaient pas être cherchés directement dans l'espace acoustique. Dans ce cadre théorique, des processus cognitifs automatiques seraient en charge de retrouver les gestes articulatoires à l'origine des sons perçus, et ce seraient ces gestes qui caractériseraient les unités composant le message. L'exemple classique cité à l'appui des théories motrices est celui de la variabilité acoustique des consonnes plosives en fonction du contexte vocalique précédent et suivant, qui contrasterait avec l'invariance des gestes articulatoires associés (voir un exemple figure 1.1).

Les théories auditives (par exemple Diehl et Kluender, 1989 ou Kuhl, 2000) se placent en contradiction avec les théories motrices. L'hypothèse principale de ces théories est qu'il est possible de s'appuyer sur les caractéristiques spectro-temporelles du signal de parole pour décoder les unités phonétiques, et que le recours aux gestes articulatoires ou aux propriétés motrices n'est donc pas nécessaire à la perception. Un argument important utilisé par les défenseurs de ces théories concerne la précédence de capacités de catégorisation des sons de parole chez le nourrisson avant l'apparition de capacités minimales de production de la parole. Cette précédence, qui montre que l'on peut percevoir sans savoir produire, semble en contradiction avec les théories motrices.

Depuis les années 2010, pour tenter d'intégrer les phénomènes dont ces deux familles de

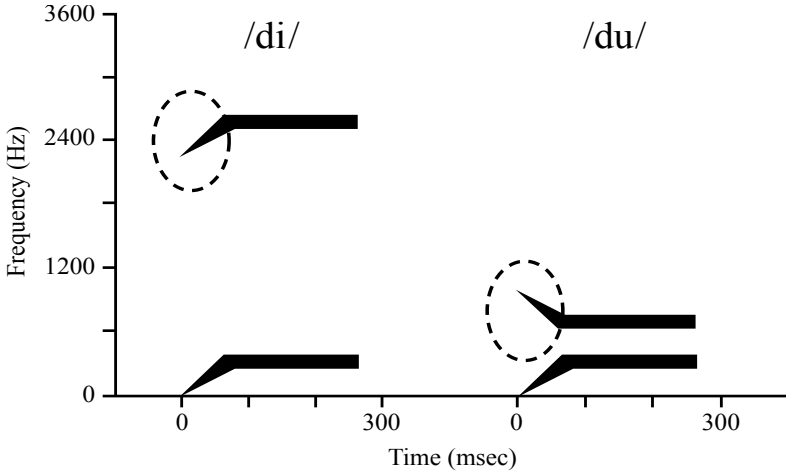

FIGURE 1.1 – **Trajectoires des deux premiers formants de deux syllabes synthétiques /di/ et /du/** Les ovales mettent en évidence la différence de transition. Figure issue de Galantucci et al., 2006.

théories rendent compte, un troisième cadre théorique intégrant les connaissances acoustiques et motrices a progressivement vu le jour, celui des théories perceptuo-motrices. Un exemple de ces théories est la PACT, pour *perception for action control theory*, de Schwartz et al., 2012. Pour cette théorie, l'apprentissage de la parole, faisant intervenir à la fois des aspects de perception et de production, résulterait en l'acquisition d'unités co-structurées autour de ces deux modalités. Ce cadre théorique permet ainsi d'associer représentations acoustiques, articulatoires et motrices pour bénéficier au cas par cas de leurs propriétés intrinsèques dans le décodage et l'accès aux unités phonologiques. Il permet également de comprendre comment, dans le développement de la parole, des capacités perceptives pourraient préexister aux premiers pas de la production, puis être progressivement complétées par des connaissances articulatori-motrices au cours du développement.

Les théories de la production Parallèlement, les théories de la production se sont, elles, intéressées à la façon dont sont représentées les unités phonétiques dans les processus de production. Les réflexions dans ce domaine, qui ont démarré un peu plus tard que pour les théories de la perception, ont convergé vers le même triptyque des théories motrices, auditives et perceptuo-motrices.

Pour les théories motrices, la production des sons de parole serait guidée par une suite de buts caractérisés articulatoirement. C'est dans ce cadre théorique qu'a été développée, dans le même laboratoire que celui de Liberman (les Laboratoires Haskins), la phonologie articulatoire (Browman & Goldstein, 1992), qui caractérise les unités phonologiques pour la production en termes de gestes articulatoires. Ces gestes, définis par les caractéristiques des constrictions du conduit vocal (zones de rétrécissement du conduit vocal à l'intérieur de la bouche ou au niveau des lèvres) qui leur sont associées, constituent donc des actions coordonnées des articulateurs du conduit vocal, contrôlant le lieu et la taille de ces constrictions.

Là encore, en réaction à cette proposition faisant des gestes articulatoires la référence des modèles de contrôle, les théories auditives postulent que les mouvements des articulateurs et les contrôles musculaires sous-jacents sont planifiés en fonction de trajectoires dans l'espace acoustique, et que la référence du contrôle est une référence auditive (Guenther et al., 1998). Ce cadre théorique est appuyé notamment par les résultats de l'expérience du « lip tube » Savariaux et al., 1995. Dans cette expérience, les expérimentateurs perturbent la production du /u/ de leurs sujets français en plaçant un tube entre leurs lèvres, qui les empêche de rapprocher et protrure leurs lèvres, geste nécessaire à la production de cette voyelle postérieure arrondie. Ce tube labial perturbe notamment la valeur des résonances du conduit vocal (formants) en les empêchant d'atteindre leurs valeurs basses requises pour le /u/. Or, après plusieurs essais, plusieurs sujets réorganisent complètement leur stratégie de production en reculant leur langue, ce qui leur permet de faire rediminuer les formants vers des valeurs caractéristiques de la voyelle. Cette réorganisation articulatoire est interprétée par les auteurs comme une preuve que le but de communication est défini en termes acoustiques et non articulatoires.

Enfin, les propositions récentes dans le domaine convergent vers un cadre de théories perceptuo-motrices dans lesquelles les modèles de contrôle impliquent des cibles à la fois auditives et somesthésiques associées aux propriétés articulatoires des gestes, ainsi que des répertoires moteurs stockés permettant d'automatiser le processus de production (voir par exemple Guenther, 2006 ; Perrier, 2005). Les représentations de la parole seraient donc de nature mixte, et ces modèles associant dans les représentations des cibles de la production les modalités auditive, articulatoire et motrice, fournissent pour la production de la parole un cadre unificateur aux théories motrices et auditives.

1.1.2 Le contrôleur

La tâche étant définie, le contrôleur est en charge de spécifier les coordinations musculaires et les dynamiques de ces coordinations permettant d'atteindre la séquence adéquate de cibles (acoustiques, articulatoires, motrices ou mixtes) dans leur déroulement temporel. Pour ce faire, nous introduisons dans ce qui suit deux enjeux qui sont classiques dans le domaine, et centraux pour la suite de notre travail.

1.1.2.1 La modélisation des relations sensori-motrices, un problème mal posé

Lorsque la relation entre l'espace moteur et l'espace sensoriel n'est pas bijective – comme c'est le cas, nous le verrons plus tard, dans le cadre de la production de la parole – le contrôleur doit implémenter des mécanismes qui lui permettent de sélectionner un geste parmi tous ceux possibles pour atteindre sa cible.

La première famille de critères porte sur la minimisation d'un coût énergétique associé à la trajectoire de commandes musculaires. Dans ce cadre, on fait l'hypothèse que le contrôleur sélectionne la trajectoire de mouvements permettant d'assurer la réalisation de la tâche avec le plus faible coût. Il est possible de définir ce coût de plusieurs manières (voir Nelson, 1983), et il

existe un consensus s'accordant sur la minimisation du *jerk* (le *jerk* correspondant à la dérivée de l'accélération) comme critère d'économie d'énergie le plus plausible – en l'occurrence, la minimisation de l'intégrale du carré de la norme du *jerk* le long de la trajectoire permettant d'exécuter la tâche (Hogan, 1984 ; Nelson, 1983).

Plus récemment, un autre critère a été introduit par Harris et Wolpert, 1998 portant sur la robustesse des trajectoires par rapport au « bruit neural ». Ce critère repose sur l'hypothèse que les signaux neuronaux de contrôle sont corrompus par du bruit dont la variance augmente avec l'intensité de ces signaux de contrôle. Le critère proposé consiste donc à sélectionner, parmi les trajectoires permettant de réaliser la tâche, celle qui minimise la variance de la position finale du *plant* induite par la corruption du signal de contrôle. Ce modèle prédit efficacement un large ensemble de données expérimentales sur le mouvement de l'œil et du bras, voir notamment Harris et Wolpert, 1998 ; Wolpert et Ghahramani, 2000.

1.1.2.2 Les modèles internes

Afin de planifier convenablement l'évolution des variables de contrôle du *plant* pour réaliser la tâche, il peut être utile au contrôleur d'avoir accès à des connaissances sur le comportement du *plant*. Une hypothèse de la littérature est que ces connaissances prennent la forme d'un modèle interne. Pour Kawato, 1999, un modèle interne est un mécanisme neuronal qui prédit la relation entrée-vers-sortie, ou son inverse, du *plant*. Dans le sens entrée-vers-sortie, un modèle interne est qualifié de modèle direct, et dans le sens contraire, de modèle inverse. Le concept de modèle interne est intéressant car il donne au contrôleur la possibilité de faire du calcul prédictif permettant de mettre en œuvre des processus d'optimisation de trajectoire comme ceux que nous avons introduits précédemment. Ainsi, dans leur présentation des « principes computationnels en neurosciences du mouvement », Wolpert et Ghahramani, 2000 montrent comment le Système Nerveux Central (SNC) pourrait structurer les mécanismes de contrôle autour de trois types de modèles internes, respectivement « inverse » (associant la sortie sensorielle désirée à une action motrice adéquate), « direct dynamique » (associant une commande à un état dynamique du *plant*) et « direct sensoriel » (prédisant la sortie sensorielle pour un état donné du système).

Ces concepts sont largement discutés dans la littérature, et pour une revue de ceux-ci dans le contexte du contrôle moteur en production de la parole le lecteur peut se référer à Perrier, 2006. Néanmoins, l'existence et le contenu des modèles internes reste un sujet de discussion et de questionnement. Pour le montrer, nous allons présenter, comme illustration, deux articles récents qui, tout en portant sur des paradigmes proches (impliquant l'apprentissage d'un système bracchio-manuel de génération de sons), conduisent à des résultats plutôt contradictoires sur le sujet.

Dans un premier article, van Vugt et Ostry, 2018 proposent un protocole expérimental, illustré figure 1.2, cherchant spécifiquement à évaluer les capacités d'interpolation des modèles internes. Dans leur expérience, les sujets pilotent un synthétiseur de sons à l'aide d'un bras robotique tout en ayant les yeux bandés. L'angle de la position de ce bras robotique à l'intérieur

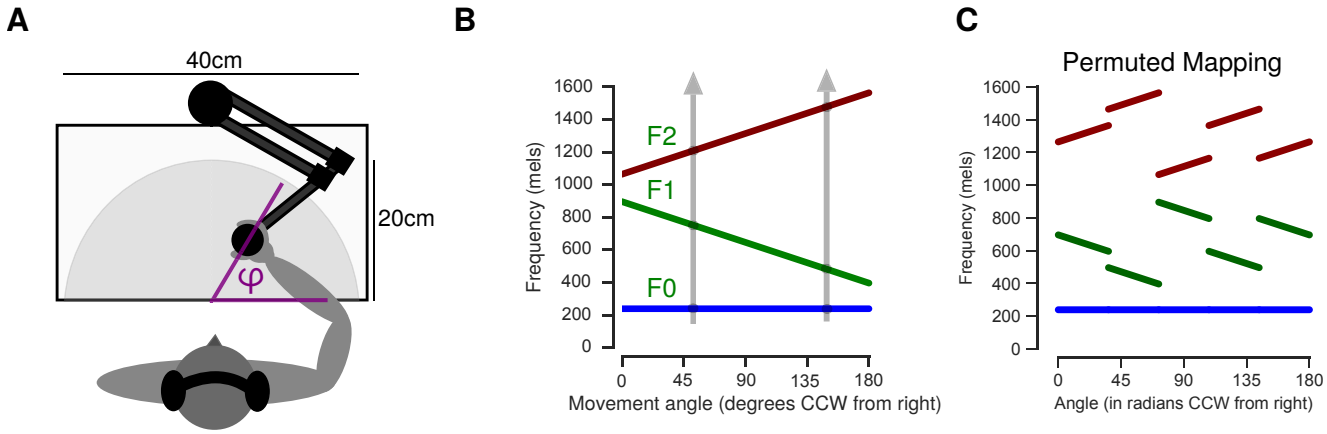

FIGURE 1.2 – **Illustration du protocole expérimental de van Vugt et Ostry, 2018** (A) Les participants étaient assis devant le bras robotique à mouvement de bras planaire et effectuaient des mouvements à partir d'une position centrale vers des points situés sur un demi-cercle. (B) Relation entre l'angle du mouvement effectué et la fréquence des 3 oscillateurs de sons purs (F0, F1 et F2) composant le son émis à la fin du mouvement. (C) Relation entre l'angle du mouvement et la fréquence des 3 oscillateurs dans la condition ordre local et désordre global. Figure adaptée de van Vugt et Ostry, 2019.

d'un demi-cercle centré sur le sujet est relié linéairement aux caractéristiques des sons synthétisés. Les auteurs testent spécifiquement plusieurs hypothèses portant sur l'apprentissage sensori-moteur, en contrastant des conditions respectivement compatibles ou incompatibles avec ces hypothèses. De manière remarquable, ils montrent que les participants sont capables d'apprendre pour partie la correspondance entre gestes et sons, même dans des conditions où la relation sensori-motrice est non régulière, avec un ordre local mais un désordre global obtenu par permutation aléatoire des relations locales dans l'espace de contrôle (figure 1.2C). Les auteurs concluent que la formation de cartes sensori-motrices s'appuierait non pas sur un modèle interne complet mais sur une collection d'expériences sensori-motrices individuelles accumulées au cours de l'apprentissage.

Dans un second article publié quelques mois plus tard, pour tester les capacités de généralisation des modèles internes, Thompson et al., 2019 montent un protocole expérimental dans lequel des sujets pilotent un synthétiseur de voyelles à l'aide d'un écran tactile. Pour produire une voyelle, les sujets doivent toucher l'écran, et les positions verticale et horizontale du toucher sur l'écran permettent de contrôler respectivement les formants F1 et F2 d'une voix synthétique. Avec ce dispositif qui demande à des sujets naïfs une tâche *a priori* complexe – se construire une « géométrie des sons » – les auteurs mettent en place deux expériences. Dans la première, les auteurs montrent que, une fois apprise par essai et erreur la correspondance geste-son dans une zone de l'espace (voyelles antérieures fermée /i/ et mi-ouverte /ɛ/ et voyelle centrale ouverte /ɑ/), les sujets transfèrent partiellement cette connaissance vers de nouvelles régions non apprises, aussi bien dans une région contenue dans la zone d'apprentissage (voyelle antérieure mi-fermée /ɪ/) que dans une région clairement différente (voyelle postérieure arrondie fermée /ʊ/). Plus frappant encore, dans une seconde expérience dans

laquelle les expérimentateurs changent légèrement le son correspondant à un geste donné, les participants sont capables de compenser partiellement cette perturbation sensori-motrice en rectifiant leur geste en conséquence. Les auteurs concluent sur la capacité des participants à apprendre *de novo* une carte sensori-motrice associant des gestes et des sons et postulent que cette carte pourrait s'appuyer sur un modèle interne articulatori-acoustique pré-existant, associant configurations articulatoires (géométrie linguale) et acoustiques (valeurs des formants dans l'espace (F1, F2)).

Comme nous le verrons dans une prochaine section décrivant des exemples emblématiques de modèles de contrôle de la parole, cette notion de modèle interne est à la base de la plupart des travaux dans ce domaine. Elle est également fondamentale dans les travaux présentés dans ce manuscrit.

1.1.2.3 Apprentissage des modèles internes

Les modèles internes reposent sur des principes d'apprentissage des relations sensori-motrices fournies par la réponse de l'environnement aux actions du sujet (Wolpert & Ghahramani, 2000). Ainsi, on peut retrouver dans les travaux de modélisation computationnelle de l'apprentissage de la parole un nombre important d'algorithmes d'apprentissage permettant de simuler ces processus de construction de modèles internes. Cette sous-section présente quatre approches générales, pour une revue plus complète le lecteur peut se référer à Barnaud, 2018 ; Moulin-Frier et Oudeyer, 2013 ; Pagliarini et al., 2020.

Babillage moteur aléatoire Le babillage moteur aléatoire consiste à explorer de manière systématique l'espace des actions possibles en associant les actions produites par exploration aléatoire, et les résultats sensoriels correspondants (voir par exemple Bullock et al., 1993). Cette méthode peut être rapprochée du comportement de vocalisation spontanée des enfants leur permettant d'explorer le fonctionnement de leur appareil phonatoire. Elle permet d'apprendre un modèle interne direct, qui encode une relation générique des commandes motrices aux sons, à partir d'exemples fournis par le babillage. L'algorithme comporte trois étapes, répétées en boucle jusqu'à ce que l'entraînement soit considéré comme terminé. La première étape est le choix aléatoire d'un geste articulatoire. La seconde est l'exécution de ce geste grâce au *plant*. La dernière est la mise à jour du modèle interne direct à l'aide de la paire geste effectué/résultat acoustique. L'utilisation de cette méthode a pour résultat une exploration relativement complète de l'espace moteur mais ne garantit pas une exploration exhaustive de l'espace acoustique.

Babillage à but aléatoire Le babillage à but aléatoire, originellement proposé par Rolf et al., 2010, reprend les mêmes principes que ceux du babillage moteur aléatoire mais avec une procédure différente pour la sélection des gestes à produire. Il présente la particularité de nécessiter l'existence d'un modèle interne de la relation inverse, appris en même temps que celui de la relation directe. Dans cet algorithme, la première étape est la sélection aléatoire

d'un son cible. Le modèle inverse est ensuite utilisé pour estimer quel geste aurait pu être à l'origine de ce son cible, et c'est ce geste qui est exécuté par le *plant*. Pour finir, la paire geste effectué/résultat acoustique est utilisée pour mettre à jour les modèles internes. De manière importante, même si les buts acoustiques sont tirés de façon aléatoire tout au long de l'apprentissage, l'estimation des gestes qui leur correspond étant liée à la performance du modèle inverse, elle varie avec l'amélioration de celui-ci. Ce mode d'apprentissage permet une exploration plus complète de l'espace acoustique que le babillage moteur aléatoire et induit une limitation de l'emploi de gestes donnant un résultat acoustique similaire à celui d'autres gestes déjà connus (les premiers résultats satisfaisants dans l'espace articulatoire ou moteur sont progressivement renforcés).

Apprentissage par accommodation L'algorithme d'apprentissage par babillage à but avec accommodation appartient à la catégorie des algorithmes d'apprentissage avec « guidage social » (Nguyen & Oudeyer, 2012). Dans cet algorithme, qui est comme le précédent un algorithme orienté vers le but (et nécessite donc également l'existence d'un apprentissage conjoint des modèles direct et inverse), au lieu d'un tirage aléatoire, la cible acoustique est fournie au modèle par un tuteur extérieur possédant lui-même une connaissance de la parole et des cibles (par exemple celles de sa langue maternelle). Ici aussi la dynamique d'exploration varie au cours de l'apprentissage en commençant par une sélection quasi-aléatoire de gestes du fait de l'ignorance du modèle inverse, puis en évoluant vers des choix de gestes de plus en plus adéquats au fil de l'amélioration de ce dernier. Ce type d'apprentissage présente l'avantage, contrairement à ceux présentés précédemment, de permettre le guidage de l'exploration de l'espace acoustique vers les mêmes régions que celles où se situent les cibles données par le tuteur, et donc de ne pas explorer celles moins utiles à la production de la langue utilisée par le tuteur.

Babillage par curiosité L'algorithme de babillage par curiosité appartient à la catégorie des algorithmes d'exploration active (Moulin-Frier & Oudeyer, 2013). Il repose sur l'idée que l'exploration des vocalisations serait un jeu de découverte pour les enfants, dans lequel leur but serait la recherche de la nouveauté et de l'amélioration de leurs compétences de production. Là encore, cet algorithme reprend les principes du babillage à but aléatoire mais en en modifiant le choix des cibles. À chaque répétition, la cible choisie est celle supposée améliorer le plus possible les connaissances des modèles internes. Le résultat est une exploration de l'espace acoustique qui favorise les zones inconnues et qui limite la répétition des sons déjà connus du modèle. Un exemple d'implémentation de ce mode d'exploration peut être trouvé dans les travaux de Moulin-Frier et al., 2017.

1.1.3 Mise en œuvre de ces principes dans quelques modèles de la production de la parole

De nombreux modèles, plus ou moins complets et détaillés, de la production de la parole ont été proposés dans la littérature. Nous allons, pour finir cette revue rapide des principes du

contrôle moteur de la parole, en présenter quatre, illustratifs de réponses variées aux questions portant sur la nature de la tâche, la structure du contrôleur et les mécanismes d'apprentissage envisagés.

1.1.3.1 DIVA

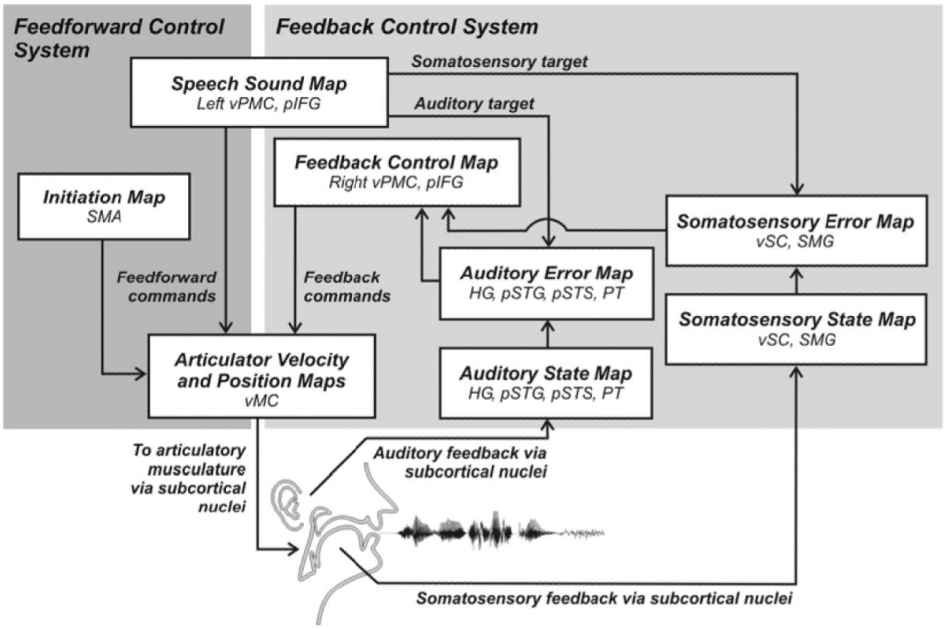

FIGURE 1.3 – **Schéma du modèle DIVA** Figure issue de Guenther et Vladusich, 2012.

DIVA (pour *Directions Into Velocities of Articulators*) est l'un des modèles computationnels de référence de la production de la parole. Ce modèle a connu plusieurs évolutions et nous nous concentrerons ici sur la version de Tourville et Guenther, 2011.

Illustré figure 1.3, le modèle DIVA s'organise autour de « cartes » qui contiennent des représentations des unités phonétiques dans divers espaces (auditif, somatosensoriel, articulatoire, moteur), et qui sont reliées par des modèles internes permettant le passage d'une représentation à une autre, et *in fine* la génération d'actions motrices assurant l'exécution de la séquence phonétique souhaitée.

Ainsi, la production d'un son commence par l'activation de la *Speech Sound Map*, qui active dans le *Feedforward Control System*, via ce que l'on peut considérer comme un modèle interne inverse global (*Feedforward Commands*), les commandes motrices adéquates (*Articulatory Velocity and Position Maps*) pour produire le son cible correspondant à l'unité phonologique souhaitée, en l'occurrence un phonème ou une syllabe. En retour, le *Feedback Control System* scrute le résultat de la production dans les espaces de représentation auditif (*Auditory Error Map*) et somatosensoriel (*Somatosensory Error Map*) et envoie, en cas d'erreur, des commandes motrices correctives (*Feedback Control Map*), en utilisant des modèles inverses locaux.

L'apprentissage des modules du modèle se fait en trois étapes. La première est une étape d'exploration sensori-motrice par babillage moteur aléatoire, où le modèle collecte des informations sur le lien entre les modalités auditive (correspondant à des propriétés temps-fréquence dans l'espace acoustique), somato-sensorielle et motrice. Ces informations sont utilisées pour mettre à jour le *Feedback Control System*. Dans la seconde étape, le modèle est exposé à des sons de parole et apprend les régions cibles dans l'espace acoustique pour chaque phonème. Enfin, dans la troisième étape, le modèle imite les sons perçus. Les imitations sont au départ très mauvaises du fait de la pauvreté du *Feedforward Control System*. Cependant, le modèle se sert des transformations acoustico-articulatoires apprises dans le *Feedback Control System* lors du babillage pour produire des commandes motrices correctives, et finit par converger vers un état où le *Feedforward Control System* est assez précis pour produire des commandes ne nécessitant pas de corrections *feedback*, sauf dans le cas de perturbations où le *Feedback Control System* est à nouveau mis à contribution à partir des erreurs auditives ou somatosensorielles détectées par le modèle.

Notons donc que ce modèle se caractérise par (1) un espace de la tâche qui est, au départ, auditif (défini dans le *Speech Sound Map*) et intègre peu à peu, après apprentissage, des spécifications articulatoires (par le biais de la *Somatosensory Error Map*) ; (2) l'apprentissage de modèles internes inverses (*feedforward* et *feedback*) sans passer par des modèles internes directs prédisant le son à partir de la configuration motrice ; et (3) un processus d'apprentissage de type babillage moteur aléatoire.

1.1.3.2 SFC (*State feedback control*)

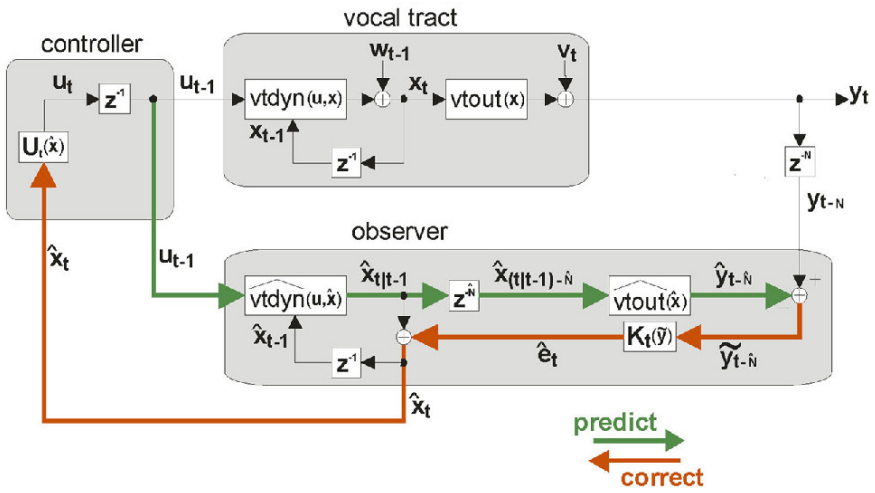

FIGURE 1.4 – **Schéma du modèle State Feedback Control** Figure issue de Houde et Nagarajan, 2011.

Houde et Nagarajan, 2011 proposent un modèle, illustré figure 1.4, de la production de la parole intégrant les principes de la théorie du contrôle par retour d'état (en anglais *state feedback control*, SFC). La SFC est une théorie reposant sur le concept d'état dynamique, qui est défini comme une description du *plant* (e.g. la position et la vélocité des différents

articulateurs) suffisante pour prédire son comportement futur. Elle postule que le système nerveux central contrôle les mouvements en estimant, à chaque pas de temps, l'état du *plant* et en générant des commandes basées sur ces états estimés. Ce modèle permet d'intégrer la connaissance de la dynamique du *plant* et de son contrôle à la fois dans le processus de commande *feedforward* et dans le processus de rétro-action *feedback* nécessaire en cas de perturbations.

Le modèle SFC s'intéresse principalement à la gestion des retours auditif et somatosensoriel et ne spécifie pas comment sont déterminées les cibles ni comment sont construits les différents modèles internes qu'il met en jeu. Son fonctionnement consiste en une série d'étapes répétées à chaque pas de temps. La première étape est l'envoi en parallèle des commandes motrices par le *controller* au tractus vocal ainsi qu'à un module nommé *observer*. L'*observer* prédit à partir des commandes motrices l'état dynamique du tractus vocal qu'elles vont engendrer via un premier modèle interne, puis prédit quelles seront les conséquences sensorielles (auditives et somatosensorielles) de cette dynamique via un second modèle interne. L'*observer* compare alors les conséquences sensorielles anticipées avec les conséquences sensorielles produites par le tractus vocal, et se sert de cette comparaison pour améliorer son estimation de l'état dynamique. Cette estimation est ensuite envoyée au *controller* qui s'en sert pour corriger les commandes qu'il enverra au pas de temps suivant.

Ainsi, le modèle SFC (1) s'inscrit clairement dans une approche audio-articulatoire de la description de la tâche, via la double entrée auditive et somatosensorielle de son canal sensoriel ; (2) repose sur l'intégralité de la structure de commande proposée par Wolpert et Ghahramani, 2000, avec modèles directs dynamique et sensoriel et modèle inverse ; mais (3) ne propose pas de processus d'apprentissage dans ce cadre global (on trouvera cependant des propositions sur l'apprentissage et l'implémentation de ces principes de contrôle dans le modèle FACTS – pour *Feedback-Aware Control of Tasks in Speech*, Parrell et al., 2018, 2019).

1.1.3.3 COSMO

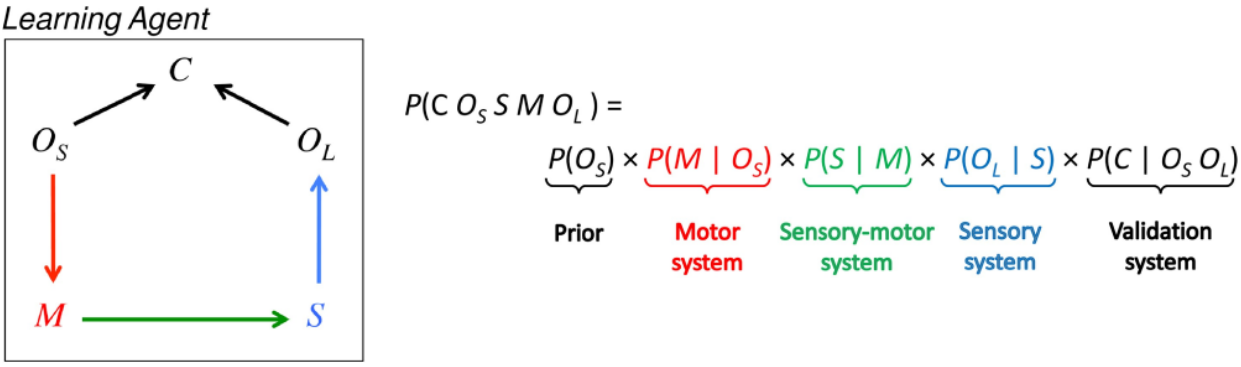

FIGURE 1.5 – **Décomposition de l'équation conjointe du modèle COSMO et illustration graphique correspondante de la structure de dépendance probabiliste entre les variables** C : succès de la communication, O_S et O_L : objets linguistiques, M : représentation motrice, S : représentation sensorielle. Figure adaptée de Barnaud et al., 2019.

COSMO, pour *Communication about Objects using Sensory-Motor Operations*, est une famille de modèles probabilistes des processus cognitifs de la parole. Dans un souci de clarté, nous nous cantonnerons ici à la présentation du modèle dans une forme abstraite et simplifiée, permettant de mieux en comprendre la structure. COSMO a été conçu autour de l'hypothèse d'internalisation de la situation de communication entre deux locuteurs. Cette hypothèse implique qu'un agent communicant, locuteur ou auditeur, dispose d'un modèle interne complet de la situation de communication intégrant capacités de production et de perception, et s'appuie sur ce modèle complet pour simuler aussi bien des tâches de perception que de production. Ce modèle fournit ainsi un cadre général qui s'intègre dans les théories perceptuo-motrices aussi bien de la perception que de la production de la parole (Moulin-Frier et al., 2015).

Ces principes se retrouvent dans les variables du modèle, qui décrivent chacune un aspect de la communication orale. La figure 1.5 montre ces variables et leurs relations de dépendance. Pour mieux les comprendre, imaginons le cas d'un agent COSMO souhaitant transmettre un concept à un autre agent. Pour ce faire il va, à partir de l'objet O_S [1], représentant un objet linguistique à transmettre, trouver le geste moteur M de son tractus vocal associé à cet objet. L'agent peut anticiper le son S que va provoquer l'exécution de ce geste M grâce à son modèle interne direct représenté par la relation M-vers-S. À partir du son qu'il prédit, il peut inférer quel sera l'objet O_L compris par son auditeur. La variable C, qui ne peut prendre que deux états, est conditionnée par la correspondance de O_S et O_L. Si les deux sont identiques, alors elle aura la valeur *vrai* et *faux* dans le cas contraire. L'état de cette variable sert à l'agent à déterminer le succès de la communication. Le modèle est mis en œuvre dans un cadre d'implémentation bayésienne, et s'appuie sur une description simplifiée de la structure de dépendance entre variables, également représentée dans la figure 1.5. Le modèle comporte ainsi trois modèles internes, respectivement un modèle direct de la relation sensori-motrice $P(S|M)$, un modèle moteur associant objets et gestes $P(M|O_S)$ et un modèle sensoriel associant sons et objets $P(O_L|S)$. De manière importante, COSMO n'intègre pas explicitement de modèle inverse $P(M|S)$, le processus d'inversion du modèle direct $P(S|M)$ étant pris en charge par les mécanismes d'inférence bayésienne. COSMO traite alors les tâches de production et de perception de la parole comme des questions probabilistes adressées à sa structure de dépendance $P(CO_SSMO_L)$.

L'apprentissage des relations de dépendance dans COSMO se fait par l'exécution de l'algorithme d'apprentissage par accommodation (présenté en section 1.1.2.3), avec l'hypothèse supplémentaire que le tuteur fournit explicitement à l'agent apprenant l'objet O_L en plus du signal acoustique S le représentant. L'agent COSMO peut ainsi apprendre directement le modèle sensoriel $P(O_L|S)$. Il utilise les gestes qu'il infère au cours de son processus de babillage à but avec accommodation pour apprendre le modèle moteur $P(M|O_S)$ et le modèle direct $P(S|M)$. Il peut alors réaliser une tâche de production en associant une voie motrice directe $P(M|O_S)$ (répertoire moteur) et une voie auditive inférée $P(M|O_L)$. Ainsi COSMO (1) s'intègre naturellement dans le cadre des théories perceptuo-motrices associant description auditive et motrice de la tâche ; (2) n'intègre pas de modèle inverse explicite, utilisant des processus d'inférence bayésienne à partir d'un modèle direct $P(S|M)$; et (3) s'appuie sur un

1. Les objets O_S et O_L ont pour indice S pour *Speaker* et L pour *Listener* afin de respecter la terminologie utilisée lors des présentations de COSMO dans les revues internationales (e.g. Moulin-Frier et al., 2015).

apprentissage à base de babillage avec accommodation.

1.1.3.4 GEPPETO

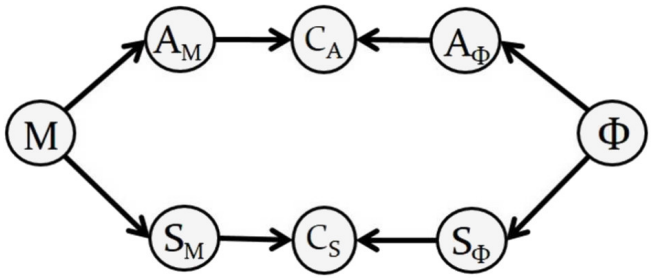

FIGURE 1.6 – **Schéma du modèle GEPPETO** Figure issue de Patri et al., 2019.

GEPPETO (*GEstures shaped by the Physics and by a PErceptually oriented Targets Optimization*) est un modèle de la production de la parole ayant connu plusieurs évolutions (Patri et al., 2015) et dont nous présentons ici la dernière mouture (Patri et al., 2019), illustrée figure 1.6. Contrairement aux modèles présentés précédemment, GEPPETO suppose que les trajectoires articulatoires et acoustiques, qui sont en partie le résultat des caractéristiques physiques de l'appareil phonatoire, ne sont pas complètement contrôlées en tout point dans le temps. Dans ce modèle, les trajectoires sont définies dans le cadre de l'hypothèse du point d'équilibre (Feldman, 1986). Cette théorie du contrôle moteur représente les trajectoires comme une succession de points d'équilibre, qui sont une description de l'activation des muscles du *plant* et qui sont espacés dans le temps. Entre chaque point d'équilibre, les mouvements du *plant* sont régis par ses propriétés bio-mécaniques gérées par des boucles sensori-motrices réflexes à un niveau sous-cortical. Ainsi, dans GEPPETO, les trajectoires sont définies par un patron d'activations musculaires du système orofacial à certains instants clés et la bio-mécanique se charge de faire la transition entre ces instants clés.

Comme son nom l'indique, GEPPETO considère des cibles sensorielles (« orientées vers la perception »). Ceci se traduit dans le schéma de la figure 1.6 par l'existence de variables auditives (A) et somatosensorielles (S) dépendant à la fois de la commande motrice M et de l'objectif phonétique à atteindre Φ (le phonème). Comme dans le cadre COSMO défini précédemment et auquel GEPPETO peut se rattacher, les variables C sont des « variables de cohérence » faisant le lien entre commandes motrices M et objectifs phonétiques Φ définis en termes auditifs et somatosensoriels. Afin de prendre en compte les caractéristiques physiques, le modèle repose sur l'utilisation d'un modèle bio-mécanique de la langue (Payan & Perrier, 1997 ; Perrier et al., 2003), et le niveau de représentation des commandes motrices qu'il adopte est relatif à l'activation des muscles dans ce modèle biomécanique. Enfin, comme dans COSMO, GEPPETO ne comporte pas de modèle inverse mais deux modèles directs associant commandes et sorties sensorielles auditives A et somatosensorielles S, et appris par exploration motrice aléatoire, les modèles inverses s'en déduisant par des processus d'inférence bayésienne.

Ainsi, GEPPETO (1) s'inscrit là encore dans la catégorie des modèles à cible sensorielle à la fois auditive et articulatoire par le biais de son canal somatosensoriel ; (2) utilise comme COSMO des modèles directs mais pas de modèles inverses, estimés simplement par inversion bayésienne des modèles directs ; et au contraire de SFC n'intègre pas de modèle interne dynamique mais des processus automatiques de génération de trajectoire par modèle biomécanique ; et (3) s'appuie sur une exploration motrice aléatoire pour son apprentissage.

1.1.4 Conclusion

Ainsi munis de ces éclairages sur les principes et principaux modèles de contrôle moteur en production de la parole, nous allons maintenant dans la section suivante nous centrer sur le système à contrôler – le *plant* – caractérisé essentiellement, dans la suite de notre travail, par les relations entre configurations articulatoires et acoustiques dont nous allons présenter les contenus respectifs et les différents types de processus et d'algorithmes permettant de caractériser et le cas échéant de modéliser ces relations.

1.2 Modélisation acoustico-articulatoire

1.2.1 Théorie source-filtre

La théorie source-filtre (Fant, 1970) permet de considérer la production de la parole comme un processus en deux étapes : la génération d'un ou plusieurs sons par une ou plusieurs sources, puis la modulation de ces sons lors de leur passage par le conduit vocal.

On distingue deux types de sources, qui chacune produisent du son grâce à une perturbation du flux d'air traversant le conduit vocal. Le premier type de source est la vibration des plis vocaux, entraînée par le passage de l'air lors de leur mise en tension. Le résultat est un son quasi-périodique, dont le contenu spectral est défini principalement par une fréquence fondamentale et les harmoniques de cette dernière. Le deuxième type de source est produit par les perturbations du flux d'air provoquées par une constriction locale du conduit vocal. Cette constriction entraîne une accélération de l'air expiré, provoquant un son apériodique (source de bruit). Elle peut avoir lieu au niveau glottique, ou bien au niveau supra-glottique avec par exemple la constriction des lèvres ou celle de la langue contre le palais. Les sons de parole ayant pour source la vibration des plis vocaux sont dits « voisés », les autres sont dits « non voisés ».

Les sons émis par les sources traversent ensuite le conduit vocal, qui fait office de filtre, accentuant ou réduisant certaines composantes de son spectre. Les propriétés acoustiques du conduit vocal dépendent de la forme de la cavité orale (composée des cavités buccale et nasale) contrôlée par la position des articulateurs.

En conclusion, la théorie source-filtre nous indique que les mouvements et déformations

du conduit vocal sont au cœur de la production de la parole. Pour étudier ce processus, il faut donc un moyen d'observer ces phénomènes.

1.2.2 Dispositifs d'acquisition de données articulatoires

Alors qu'un microphone suffit à capturer de façon satisfaisante le son émis par un locuteur produisant de la parole, enregistrer les configurations adoptées par son conduit vocal requiert des dispositifs plus sophistiqués.

1.2.2.1 Imagerie par résonance magnétique

L'imagerie par résonance magnétique (IRM) est une technique d'imagerie médicale reposant sur l'utilisation de champs magnétiques puissants. Elle permet de capturer des images en 3 dimensions des tissus biologiques des sujets enregistrés. Dans le cadre de l'enregistrement de la production de parole, la grande force de l'IRM réside en sa capacité à capturer la forme complète du conduit vocal avec une grande résolution spatiale. Cependant, si l'on souhaite enregistrer le son produit par le locuteur en parallèle de ses gestes, l'IRM n'est pas idéale du fait du bruit important causé par son fonctionnement. Plus d'informations concernant l'usage de cette technique appliquée à l'étude de la parole sont disponibles dans Demolin et Metens, 2009.

1.2.2.2 Imagerie ultrasonore

L'imagerie ultrasonore, ou échographie, est une technique s'appuyant sur la mesure du temps de parcours d'un faisceau d'ondes ultrasonores se propageant dans les tissus, et subissant à l'interface entre deux milieux d'impédances acoustiques différentes, un ensemble de réflexions et de réfractions. En plaçant la sonde émettrice d'ultrasons sous la mâchoire du sujet, il est possible de capturer de manière non invasive ses mouvements linguaux. Cette acquisition se fait généralement sur le plan médio-sagittal et présente une bonne résolution temporelle (de l'ordre de 80 images/seconde) et spatiale (de l'ordre de 1 mm). Cependant, comme la sonde est solidaire de la mâchoire, cette méthode ne permet pas de capturer les mouvements de celle-ci. Pour plus de détails sur cette méthode d'enregistrement, voir Hueber et Denby, 2009.

1.2.2.3 Articulographie électromagnétique

L'articulographie électromagnétique, en anglais *electromagnetic articulography* (EMA), est une méthode pour enregistrer la position de bobines collées en différents points des articulateurs et plongées dans un champ magnétique (Schönle et al., 1987). Ces capteurs sont typiquement positionnés sur la langue, les lèvres, l'incisive inférieure (pour enregistrer les mouvements de la mâchoire) et le velum. Ainsi, contrairement aux autres méthodes citées précédemment, l'EMA ne capture pas une image intégrale du conduit vocal mais la position de « points de

chair ». Ceci présente l'avantage, malgré une acquisition bien moins complète du contour du conduit vocal, de mieux suivre les déformations des tissus lors de leurs mouvements. L'enregistrement de la position des bobines peut se faire, selon l'équipement, en 2 ou 3 dimensions, avec une résolution de l'ordre du millimètre et avec une fréquence élevée (typiquement entre 200 et 500 Hz) permettant de capturer les mouvements rapides. De plus, l'EMA est silencieuse et est adaptée à un enregistrement parallèle acoustique et articulatoire. Cependant l'EMA est invasive, et à cause des capteurs présents dans le conduit vocal et des fils qui y sont accrochés, l'articulation du sujet peut être gênée.

1.2.2.4 Conclusion

En conclusion, il existe plusieurs méthodes permettant de capturer les mouvements des différents articulateurs de la parole et les configurations successives du conduit vocal. Ces méthodes diffèrent par leur précision et leur invasivité. Les données ainsi acquises sont souvent dans des espaces dont la dimension est importante et peuvent être difficiles à analyser directement. Néanmoins, il est possible de dresser des modèles de conduit vocal permettant de projeter ces données dans des espaces aux dimensions moins nombreuses et interprétables d'un point de vue phonétique. Ces modèles sont classiquement appelés « modèles articulatoires ». Dans la section suivante, nous présentons les principales approches proposées dans la littérature pour leur mise en œuvre.

1.2.3 Modélisation articulatoire

Un modèle articulatoire est un système permettant de générer une forme de conduit vocal en 2 ou 3 dimensions, à partir de paramètres décrivant l'état des articulateurs. Ces paramètres, dits « paramètres articulatoires », peuvent être par exemple « ouverture des lèvres », « avancée de la langue » ou encore « hauteur de la mâchoire ». À la lecture de la littérature, on distingue deux grands types de modèles articulatoires, chacun décrit dans les prochaines sous-parties.

1.2.3.1 Les modèles géométriques et statistiques

Les formes de conduit des modèles géométriques ainsi que les paramètres articulatoires régissant leur génération sont principalement créés de deux façons. La première repose sur leur définition manuelle *a priori* se basant sur les connaissances de l'anatomie du conduit vocal (e.g. Birkholz et al., 2006 ; Birkholz et al., 2010 ; Mermelstein, 1973). Cette définition peut ensuite être adaptée *a posteriori* pour mieux correspondre à des données articulatoires enregistrées. La seconde façon consiste à extraire les formes de conduit et les paramètres articulatoires associés directement depuis un corpus de données articulatoires à l'aide de méthodes statistiques (e.g. Alexander et al., 2019 ; Engwall, 2002 ; Laprie et al., 2018 ; Maeda, 1990 ; Serrurier et al., 2019 ; Story et al., 2018). Les modèles créés de cette façon sont parfois classés dans une catégorie à part entière, appelée « modèles statistiques ».

Un modèle qui prendra une importance particulière dans ce travail est celui de Maeda, 1990, modèle pionnier qui s'est avéré d'une grande importance dans le domaine. Ce modèle, dont une description complète sera donnée section 3.1.1, repose sur une analyse linéaire permettant de définir successivement des « paramètres articulatoires » associés respectivement aux mouvements des principaux articulateurs de la parole, la mâchoire, le dos, le corps et la pointe de la langue, le larynx, les lèvres, et de contrôler à partir de ces paramètres les formes du conduit vocal, en lien avec un corpus pré-existant de données cinéradiographiques (méthode d'imagerie du conduit vocal qui n'est plus autorisée à présent).

Le succès qu'a connu ce modèle, plus de 30 ans après sa création, est frappant. Il a été utilisé par de très nombreuses équipes à travers le monde, a été repris dans plusieurs autres modèles de production comme DIVA, décrit dans la section 1.1.3.1, et adapté fréquemment, pour en proposer des versions sur de nouvelles données (Beautemps et al., 2001), adaptées à l'âge et la morphologie des locuteurs (Callan et al., 2000 ; Ménard et al., 2004), ou développé vers des composants complémentaires comme les mouvements de l'épiglotte (Laprie et al., 2018). Ce succès est dû principalement au fait que la technique d'analyse linéaire sur laquelle il s'appuie, et qui sera décrite section 3.1.1, permet de dégager des dimensions explicatives qui sont à la fois efficaces pour décrire les données, mais aussi phonétiquement interprétables, résultant en des actions phonétiquement bien connues, de la mâchoire, de la langue et des lèvres, et même compatibles avec des données neurophysiologiques sur le contrôle des mouvements orofaciaux (Maeda & Honda, 1994). Pour ces raisons, ce modèle constituera un pivot de nos propres travaux, nous le verrons.

Globalement, la génération de formes de conduit à partir d'un modèle géométrique présente l'avantage d'avoir un faible coût de calcul, et d'être relativement facile à piloter grâce aux paramètres articulatoires explicites. Cependant, ces modèles ne prennent pas en compte les propriétés bio-mécaniques des tissus, et leurs paramètres articulatoires ne sont qu'une représentation très indirecte de l'activité des muscles du conduit vocal.

1.2.3.2 Les modèles bio-mécaniques

Les modèles bio-mécaniques se basent généralement sur la méthode des éléments finis (FEM, pour *finite element method* en anglais) pour simuler les tissus mous du conduit vocal et leurs propriétés bio-mécaniques (e.g. Buchaillard et al., 2009 ; Dang et Honda, 2004 ; Perrier et al., 2003). Ils intègrent également une structure musculaire qui peut être pilotée, permettant de déformer ces tissus mous. Ainsi, les paramètres de ces modèles sont les commandes envoyées à ces muscles.

Les modèles articulatoires appartenant à cette famille sont adaptés pour étudier les mécanismes bio-mécaniques qui sous-tendent la production de la parole, tels que la relation entre l'activation des muscles et les mouvements articulatoires. Cependant, ils sont gourmands en ressources de calcul et leur contrôle est difficile du fait de leur grand nombre de degrés de liberté (Calka et al., 2021).

1.2.3.3 Conclusion

Les modèles articulatoires permettent d'obtenir des formes de conduit à partir de paramètres interprétables. À partir de ces formes de conduit, nous pouvons calculer, comme nous le verrons dans la section suivante, le son qu'elles pourraient produire ou ses caractéristiques spectrales grâce à des modèles d'un autre type.

1.2.4 Lien acoustique-articulatoire

1.2.4.1 Modélisation acoustique physique

Les fonctions de transfert correspondant aux formes de conduit vocal obtenues à partir de modèles articulatoires peuvent être calculées à l'aide de modèles acoustiques physiques. Pour la plupart des modèles de cette famille, ce calcul se fait en deux temps : d'abord, la forme du conduit est simplifiée en un ensemble de formes géométriques élémentaires, puis les propriétés acoustiques de cet ensemble sont calculées à l'aide d'équations décrivant les phénomènes acoustiques de propagation.

L'approche traditionnelle se base sur une géométrie à 1 dimension, qui résume la forme de conduit donnée par le modèle articulatoire en une suite de tubes (Fant, 1970). Pour ajuster la longueur et le diamètre de ces tubes, des mesures d'aires sont effectuées à différentes profondeurs de la forme de conduit utilisée. Ensuite, cette suite de tubes est traduite en un modèle de ligne de transmission dont est dérivée une fonction de transfert (méthode décrite dans Badin et Fant, 1984). Cette approche a pour avantage d'être peu coûteuse en ressources de calcul et de fournir des résultats assez réalistes, principalement dans les basses fréquences (en dessous de 5 kHz).

D'autres modèles acoustiques physiques se basent sur d'autres techniques, comme la méthode de la matrice de ligne de transmission à 2 dimensions (Mullen et al., 2007), la méthode des éléments finis (Lu et al., 1993 ; Niikawa et al., 2000) ou la méthode multimodale (Blandin et al., 2021). Les modèles acoustiques physiques n'étant pas employés dans ce travail, ces techniques ne seront pas présentées plus en détail.

1.2.4.2 Modèles acoustico-articulatoires statistiques

Les modèles acoustico-articulatoires statistiques exploitent les corpus de données acoustiques et articulatoires enregistrées, et utilisent des techniques d'apprentissage automatique pour faire le lien entre ces deux modalités. La création d'un modèle appris revient donc à effectuer une régression entre des données articulatoires et des données acoustiques. Cette régression peut s'effectuer dans les deux sens et ces modèles peuvent ainsi représenter soit la relation directe articulatoire-vers-acoustique, soit la relation inverse acoustique-vers-articulatoire.

Ces relations possèdent toutes deux de hautes non-linéarités locales (par exemple autour

de configurations articulatoires proches de l'occlusion, où un petit mouvement peut engendrer une plus forte variation du son produit). Cependant, il existe une différence fondamentale entre elles, causée par le fait qu'une configuration articulatoire ne puisse engendrer qu'un seul son mais qu'un même son puisse être produit avec différentes formes de conduit (Atal et al., 1978 ; Neiberg & Ananthakrishnan, 2008 ; Qin & Carreira-Perpinán, 2007). Autrement dit, dans le sens articulatoire-vers-acoustique, un point de l'espace articulatoire ne peut mener qu'à un seul point de l'espace acoustique et dans le sens acoustique-vers-articulatoire, un point de l'espace acoustique peut être rejoint depuis plusieurs points de l'espace articulatoire. La relation articulatoire-vers-acoustique peut ainsi être qualifiée de relation *many-to-one* et la relation acoustique-vers-articulatoire de relation *one-to-many*.

Modélisation trame à trame Les modèles statistiques cherchent typiquement à mettre en correspondance des paramètres spectraux issus d'une analyse spectrale à court terme, avec des paramètres articulatoires issus d'une des méthodes d'imagerie présentées précédemment. L'analyse spectrale fournit des données sur des trames, typiquement à des cadences d'analyse de l'ordre de 50 Hz, et les méthodes d'acquisition de données articulatoires fonctionnent dans des gammes similaires ou à des fréquences plus élevées (notamment pour l'EMA). La première étape consiste donc à synchroniser les données spectrales et articulatoires sur les trames d'analyse acoustique.

Sur cette base, une première catégorie de modèles articulatori-acoustiques statistiques fonctionne « trame à trame », en supposant que chaque instant d'un signal de parole est indépendant de ceux qui l'entourent. De ce fait, ces modèles sont davantage adaptés pour représenter la relation articulatoire-vers-acoustique (pour laquelle en effet le son dépend exclusivement de la configuration articulatoire instantanée) que la relation inverse, qui peut présenter des ambiguïtés sur un instant isolé du fait de sa nature *one-to-many*.

Les premières méthodes proposées s'appuyaient sur des outils de classification et régression linéaire simples (Shiga & King, 2004), avant de céder la place à des outils de régression non linéaire plus conformes à la nature du lien articulatori-acoustique, impliquant notamment l'utilisation de réseaux de neurones artificiels (des travaux pionniers de Kello et Plaut, 2004 à de nombreux travaux récents comme ceux de Gosztolya et al., 2019). Ainsi, un travail récent de Saha et Fels, 2020 cherche à construire une représentation conjointe entre données articulatoires et acoustiques de voyelles en utilisant des réseaux de neurones inversibles, tout en préservant simultanément les caractéristiques spécifiques à chaque domaine. Ils montrent comment une architecture de type auto-encodeur (voir section 1.3.1.2) permet d'obtenir des mises en relation directe et inverse de manière semi-supervisée entre la géométrie du conduit vocal et le spectrogramme de sons vocaliques de synthèse.

Modélisation intégrant la dynamique Rapidement, les outils de modélisation statistique des liens articulatori-acoustiques se sont orientés vers l'utilisation de modèles prenant en compte la dynamique des signaux traités (voir par exemple Richmond et al., 2003). L'utilisation de séquences de trames peut aider à mieux modéliser la prédiction de paramètres acous-

tiques à partir de données articulatoires, et on utilise classiquement une séquence de quelques trames pour améliorer les performances (Bocquelet et al., 2016 ; Gosztolya et al., 2019). Mais le passage aux modèles dynamiques est particulièrement requis pour le lien acoustique-vers-articulatoire car il permet de réduire les incertitudes liées à la nature *one-to-many* du processus d'inversion. En effet, la production de la parole, nous l'avons vu, est contrainte par les propriétés bio-mécaniques du conduit vocal ou par des critères d'optimisation des séquences motrices, par exemple de minimisation de l'énergie dépensée (section 1.1.2.1). De ce fait, donner accès au contexte au modèle lors de l'inférence peut lui permettre de bénéficier des régularités induites par ces contraintes.

Ainsi, Toda et al., 2008 proposent de représenter la relation articulatoire-vers-acoustique ainsi que la relation inverse à l'aide de modèles de mélange gaussien (en anglais *Gaussian mixture models*, GMM) définis sur chaque trame. La dynamique articulatoire est prise en compte par la modélisation de la covariation des positions successives des articulateurs, de leur vitesse et de leur accélération. Cette covariation est prise en compte à l'apprentissage, mais également au moment de l'inférence, à l'aide d'un estimateur réalisant une conversion non pas « trame-à-trame » (comme c'est classiquement le cas dans la régression par mélange de gaussiennes) mais « séquence-à-séquence ». Leurs résultats montrent que cette contrainte permet de meilleures estimations par rapport à une approche trame à trame.

Zen et al., 2010 utilisent des modèles de Markov cachés (en anglais *Hidden Markov Model*, HMM) pour modéliser la relation acoustique-vers-articulatoire. Lors de l'inférence, leur modèle procède en deux étapes. La première est une étape de reconnaissance des phonèmes, suivie par une étape d'inversion du son vers des configurations articulatoires conditionnée par les phonèmes reconnus. Leur méthode, comparée à celle de Toda et al., 2008, montre une erreur de reconstruction articulatoire plus faible.

Modélisation avec intégration explicite de contraintes Les modèles d'inversion n'intégrant pas ou que peu de connaissances sur la dynamique peuvent donner des trajectoires articulatoires comprenant des discontinuités qui ne sont pas observées sur les trajectoires originales qu'ils cherchent à reconstruire. Confrontés à ce problème sur un modèle d'inversion constitué d'un GMM inversant les trames paquets par paquets, Toda et al., 2004 proposent d'appliquer un filtrage passe-bas aux trajectoires articulatoires estimées. Il apparaît que ce filtrage, dont les paramètres sont ajustés pour chaque articulateur de façon à minimiser l'erreur de reconstruction, permet de rapprocher les trajectoires estimées des trajectoires originales.

De même, Ouni et Laprie, 2005 ont développé une méthode d'inversion acoustique-vers-articulatoire exploitant un dictionnaire articulatori-acoustique structuré sous forme d'une hiérarchie d'hypercubes. Cette structure leur permet de mettre en œuvre une procédure d'inversion trame-à-trame exhaustive à partir de laquelle la trajectoire articulatoire est inférée par une technique de régularisation implémentant un algorithme de lissage non linéaire sur les paramètres bruts.

Ghosh et Narayanan, 2010 construisent un modèle d'inversion intégrant un principe similaire, en proposant un critère de lissage généralisé qui permet d'adapter les caractéristiques

du filtre à l'articulateur en jeu et aux propriétés de la trajectoire à lisser. Leur modèle est constitué d'un algorithme récursif inversant une par une les trames d'un son de parole à partir de la configuration articulatoire précédente et de la trame de son courante. À chaque pas de temps la configuration articulatoire est sélectionnée de façon à minimiser l'énergie d'un filtre passe-haut appliqué à la trajectoire formée par la concaténation de cette nouvelle configuration articulatoire avec les précédentes. Ils montrent que ce critère de lissage généralisé présente de meilleures performances que les filtrages fixes classiquement utilisés, en termes d'erreur de reconstruction des trajectoires articulatoires.

On s'oriente ainsi peu à peu vers des modèles prenant en compte les principes d'optimisation du mouvement introduits précédemment. On peut citer dans ce cadre Avni et Patil, 2016, qui s'inspirent du modèle de Hogan, 1984 stipulant que les mouvements humains sont organisés de façon à emprunter des trajectoires lisses qui minimisent le *jerk*. Pour leur modèle, les auteurs utilisent une régression non paramétrique pour estimer les configurations articulatoires à partir du son dans un processus trame à trame. Lors de l'inférence d'une trajectoire complète, les configurations articulatoires sont sélectionnées de façon à minimiser le *jerk* de chaque articulateur. Comparée à la méthode citée précédemment se basant sur des filtres passe-haut, cette méthode donne de meilleures estimations (erreur globale réduite).

Modélisation multi-locuteurs Les modèles acoustico-articulatoires appris se basent le plus souvent sur des corpus de données ne comprenant des enregistrements que d'un seul locuteur. À cause des différences inter-individuelles de voix et de forme de conduit, ils ne peuvent pas être utilisés directement sur des données d'un autre locuteur, ils sont dits « locuteurs-dépendants ».

Pour résoudre ce problème, une première approche consiste à explorer la possibilité de transférer les connaissances articulatoires acquises d'un locuteur à un autre. Ceci passe par la mise en œuvre d'outils de conversion de voix reposant sur la définition d'espaces acoustiques normalisés, permettant de cadrer dans le même espace acoustique les données de production de différents locuteurs (Ghosh & Narayanan, 2011). Ainsi, Hueber et al., 2015 proposent un modèle d'inversion articulatoire construit sur les enregistrements acoustiques et articulatoires d'un locuteur référence, et adaptable à un locuteur cible dont on ne dispose que d'une faible quantité d'enregistrements acoustiques. L'inversion articulatoire du locuteur cible se déroule en deux étapes, la première est la conversion de la voix du locuteur cible vers celle du locuteur de référence, et la seconde est l'inversion articulatoire de la voix convertie dans l'espace du locuteur de référence. Les auteurs comparent deux façons d'enchaîner ces deux étapes : la première consiste à construire le modèle de conversion et le modèle d'inversion séparément à l'aide de GMM puis de les chaîner, tandis que la seconde consiste à combiner ces deux étapes au sein d'un seul et même modèle probabiliste. Cette seconde méthode montre de meilleures performances que la première, notamment grâce à la combinaison de la conversion et de l'inversion qui permet de reconstruire les données acoustiques manquantes du locuteur cible.

Girin et al., 2017 étendent cette idée en ajoutant un lien direct entre données acoustiques du locuteur cible et données articulatoires de référence, et montrent que ce lien, qui exploite la

cohérence phonétique implicite aux flux articulatoire et acoustique, améliore les performances de transfert. Sur une tâche et des questions similaires, Tang et al., 2018 utilisent d'autres outils basés sur l'analyse canonique des corrélations et son implémentation dans des réseaux de neurones profonds (« *deep variational canonical correlation analysis with private variables* », VCCAP).

Une autre série de travaux s'attaque directement à la modélisation statistique de corpus articulatori-acoustiques multi-locuteurs. Ainsi, Turrisi et al., 2018 montrent comment le passage par un espace de catégorisation phonétique permet d'obtenir de bonnes performances de reconstruction de trajectoires articulatoires à partir de données acoustiques multi-locuteurs dans un réseau de neurones entraîné de façon faiblement supervisée. Sivaraman et al., 2019 passent par des techniques de normalisation de longueur du conduit vocal (*vocal tract length normalization*, VTLN) et mettent en œuvre un réseau de neurones dynamique (*Dynamical Neural Network*) produisant des performances de reconstruction articulatoire relativement bonnes sur un corpus de 10 locuteurs de la base de données rayons X de l'Université de Wisconsin (Westbury et al., 1994).

Notons enfin les travaux de Parrot et al., 2020 proposant une méthode d'évaluation des modèles d'inversion articulatoire multi-locuteurs. Leur méthode, qui consiste à évaluer si les trajectoires articulatoires inférées sont porteuses d'informations permettant la discrimination des phonèmes, permet une évaluation complémentaire des performances des modèles par rapport à une mesure de leur erreur de reconstruction. Elle présente également l'avantage de pouvoir être utilisée sur des corpus ne comportant que des données audio.

1.2.4.3 Conclusion

La relation articulatori-acoustique dans le sens articulatoire-vers-acoustique peut être représentée par des modèles physiques, qui intègrent des descriptions mathématiques des phénomènes acoustiques physiques à l'œuvre lors de la production de la parole. Ces modèles peuvent présenter différents degrés de réalisme mais souvent au prix d'un temps de calcul élevé.

Il est également possible de modéliser la relation articulatori-acoustique grâce à des modèles construits avec des méthodes d'apprentissage automatique appliquées à des corpus de données intégrant ces deux modalités. Ces modèles peuvent représenter aussi bien les relations articulatoire-vers-acoustique qu'acoustique-vers-articulatoire, mais cette seconde relation est bien évidemment plus difficile à décrire de façon satisfaisante du fait de sa nature *one-to-many*. Pour outrepasser cette difficulté, il est possible de créer des modèles intégrant des notions de dynamiques de production de la parole. Ces notions peuvent être apprises à partir des corpus utilisés pour la création des modèles, ou bien définies *a priori* sous forme de contraintes reflétant la bio-mécanique des articulateurs ou la minimisation de l'énergie dépensée au cours du mouvement. Néanmoins, les modèles appris sont souvent construits à partir de données d'un seul locuteur et nécessitent un travail supplémentaire d'adaptation pour pouvoir être généralisés à d'autres locuteurs.

Pour achever la description de la relation articulatoire-vers-acoustique, il faut maintenant

passer vers la dernière étape du processus, qui est celle de la génération de la forme d'onde dans les modèles de synthèse. C'est ce que nous allons aborder maintenant.

1.2.5 Reconstruction de la forme d'onde

Différentes approches sont possibles pour reconstruire un signal de parole à partir d'une représentation paramétrique du contenu spectral cible. Les approches classiques sont basées « signal », c'est-à-dire sur la construction d'un filtre numérique de synthèse (par exemple un filtre MLSA, *Mel-log spectrum approximation*, pour une représentation utilisant des coefficients cepstraux Imai, 1983), classiquement excité par un signal représentant l'activité glottique (la source). On a vu plus récemment fleurir des approches basées sur les « vocodeurs neuronaux », technique initialement développée dans le cadre de la synthèse de la parole à partir du texte. Dans ce travail de thèse, nous nous sommes focalisés sur cette seconde approche, et en particulier sur le vocodeur LPCNet (Valin & Skoglund, 2019), qui s'appuie sur une décomposition source-filtre du signal de parole et est donc un bon candidat pour la synthèse articulatoire. Nous présentons ce modèle, après un rapide tour d'horizon des vocodeurs neuronaux.

Les vocodeurs neuronaux sont des modèles capables de générer des sons de parole en ayant recours à des réseaux de neurones artificiels. Ils sont entraînés sur des volumes importants de parole et synthétisent des sons de haute qualité, dépassant souvent les technologies les ayant précédés. Néanmoins, ils sont très liés aux voix des locuteurs présents dans leur corpus d'entraînement et peuvent donner de mauvaises synthèses lorsqu'ils sont utilisés pour générer des voix éloignées de ce qu'ils ont appris (Perrotin et al., 2021).

WaveNet (van den Oord et al., 2016) est le premier vocodeur neuronal à voir le jour. Dans sa forme la plus épurée, le modèle génère du son échantillon par échantillon. Chaque génération d'échantillon est conditionnée par les valeurs de ceux qui le précèdent, et dont les valeurs sont lues grâce à des couches convolutionnelles dilatées. Le modèle peut être modifié pour ajouter à ce conditionnement des informations permettant de guider la génération des échantillons. Ces informations peuvent être le phonème courant, la syllabe ou bien directement des informations spectrales. WaveNet génère des sons de qualité, jugés supérieurs à ceux générés par la plupart des autres vocodeurs neuronaux apparus après lui lors de tests perceptifs (Govalkar et al., 2019). Cependant, le modèle nécessite une puissance et des temps de calcul importants pour être utilisé.

Pour proposer une synthèse de qualité comparable mais requérant une puissance et un temps de calcul plus raisonnables, Kalchbrenner et al., 2018 présentent WaveRNN. Plutôt que d'utiliser des couches de convolution, les auteurs ont recours à un réseau de neurones récurrent. Ce type de réseau présente à chaque pas de temps un état caché calculé à partir de l'entrée courante ainsi que de l'état caché précédent. Cet état caché encode de façon compacte les informations liées aux entrées précédentes. WaveRNN intègre l'état caché du pas de temps précédent dans le conditionnement de la génération de chaque échantillon. Ainsi, il n'est plus nécessaire de regarder pour chaque prédiction les échantillons précédents, sauvegardant du temps et de la puissance de calcul.

Bien d'autres vocodeurs neuronaux sont apparus suite à WaveNet, et parmi eux il convient de citer celui utilisé dans les travaux présentés dans ce manuscrit : LPCNet (Valin & Skoglund, 2019). LPCNet présente deux particularités qui le distinguent de la plupart des autres vocodeurs neuronaux. La première réside dans son principe de fonctionnement, qui combine le codage prédictif linéaire (en anglais *linear predictive coding*, LPC) avec un réseau de neurones. Le LPC est une méthode de codage et de synthèse de la parole basée sur la théorie source-filtre (Makhoul, 1975). La synthèse d'un échantillon par LPC requiert un signal d'excitation (la source) ainsi qu'une enveloppe spectrale (le filtre). L'enveloppe spectrale est traduite en une suite de coefficients, qui sont utilisés pour réaliser une somme pondérée des échantillons précédents. Cette somme, combinée au signal d'excitation, donne la valeur de l'échantillon suivant. LPCNet reprend cette méthode et y ajoute la génération du signal d'excitation par un réseau de neurones récurrent, conditionné par deux paramètres : la fréquence fondamentale du son et son degré d'harmonicité (définie comme la valeur de la fonction d'autocorrélation calculée sur une courte fenêtre temporelle). La synthèse par LPCNet est donc conditionnée par des paramètres de source et des paramètres de filtre. Cette séparation explicite entre ces deux types de paramètres est la seconde particularité de LPCNet, et c'est elle qui a motivé son choix pour les travaux présentés ici.

1.3 Apprentissage auto-supervisé de représentations

En apprentissage automatique, on a vu depuis une dizaine d'années un développement de plus en plus important de méthodes d'apprentissage auto-supervisé de représentations. Ces méthodes ont pour but d'extraire à partir de données brutes, et non étiquetées, des représentations de haut niveau. Elles consistent le plus souvent en l'entraînement d'un modèle sur une tâche prétexte. Une tâche prétexte est une tâche qui n'est pas celle pour laquelle sera utilisé le modèle après son entraînement. L'entraînement d'un modèle sur une tâche prétexte a pour but de le forcer à construire un espace, souvent de plus petite dimension que les données, dans lequel les représenter. Dans la suite de ce manuscrit, ces espaces seront qualifiés d'« espaces latents » et les représentations des données à l'intérieur de ceux-ci seront appelées « plongements ». Une fois construits, ces espaces de représentation peuvent fournir une meilleure base pour l'exécution d'autres tâches que l'espace des données original.

Un exemple d'apprentissage auto-supervisé de représentations est word2vec (Mikolov et al., 2013). Dans une de ses variantes, word2vec est un modèle dont la tâche prétexte d'entraînement est la prédiction d'un mot d'une phrase à partir de ceux qui l'entourent. L'espace latent obtenu après l'apprentissage du modèle présente des propriétés intéressantes telles que la proximité à l'intérieur de celui-ci des mots proches sémantiquement.

1.3.1 Auto-encodage

L'auto-encodage est une tâche prétexte dans laquelle un réseau de neurones est entraîné à compresser les données qui lui sont présentées puis à les reconstruire le plus fidèlement possible

à partir de leur représentation compressée. La compression est réalisée au niveau d'une couche cachée intermédiaire (*bottleneck layer*) dont la dimensionnalité est plus faible que celle des données à reconstruire. Ce « goulot d'étranglement » force les modèles à trouver une manière plus compacte de représenter les données, et peut ainsi être qualifié d'espace latent. Dans les modèles d'auto-encodage, la partie allant de l'entrée jusqu'à l'espace latent s'appelle l'encodeur et la partie allant de l'espace latent jusqu'à la sortie s'appelle le décodeur. La suite de cette section présente plusieurs variantes de cette technique.

1.3.1.1 Analyse en composantes principales

L'analyse en composantes principales (en anglais *principal component analysis*, PCA) est une méthode d'analyse statistique visant à transformer linéairement l'espace des données analysées en un nouvel espace où les dimensions sont décorrélées deux à deux. Ce nouvel espace est construit de façon à ce que chacune de ses dimensions retranscrive le mieux possible la variance des données non retranscrite par les dimensions précédentes.

Bien que cette méthode ne soit pas basée sur les réseaux de neurones, nous la présentons ici comme une « version linéaire » du processus d'auto-encodage. En effet, le sous-espace de dimension N, créé en ne retenant que les N premières composantes de plus forte variance de l'espace construit par PCA, est un espace latent des données. La transformation linéaire permettant de projeter les données dans ce sous-espace ainsi que celle permettant de reconstruire optimalement les données à partir de celui-ci peuvent également être vues comme un processus linéaire d'encodage et de décodage.

1.3.1.2 Auto-encodeur

Dans un auto-encodeur (AE) standard, les processus d'encodage et de décodage sont implémentés par des réseaux de neurones dont les paramètres sont estimés de façon conjointe et auto-supervisée, à partir d'un ensemble de données d'apprentissage, par rétropropagation du gradient de l'erreur de reconstruction (Hinton & Salakhutdinov, 2006).

Contrairement à la PCA, l'AE permet de modéliser des relations non-linéaires entre les données et leur représentation dans l'espace latent. Cependant, la construction des dimensions de ce dernier n'est pas contrainte lors de l'apprentissage, les rendant relativement difficiles à interpréter.

1.3.1.3 Auto-encodeur variationnel

L'auto-encodeur variationnel (en anglais *variational autoencoder*, VAE) peut être vu comme une version probabiliste de l'AE. Le VAE intègre un *prior* sur la structure de son espace latent, permettant notamment d'en faciliter l'interprétation et le contrôle (Kingma & Welling, 2014). Nous présentons ici les principes généraux du VAE, qui sera utilisé et spécifié plus en détail

à la section 4.1.1. Tout comme l'auto-encodeur, le VAE possède un encodeur et un décodeur implémentés à l'aide de réseaux de neurones. Cependant, contrairement à l'AE, les sorties de l'encodeur et du décodeur du VAE sont composées de paramètres de distributions et non de valeurs fournies de façon déterministe. De plus, le VAE fait une hypothèse forte sur l'espace latent : la distribution des plongements à l'intérieur de celui-ci doit suivre une distribution de probabilité définie *a priori* (typiquement une loi normale d'espérance 0 et d'écart-type 1, indépendante pour chaque composante).

Cette hypothèse conduit techniquement à deux différences par rapport à l'auto-encodeur classique. La première se situe au niveau de la traduction des données en plongements. La sortie de l'encodeur ne donne plus directement en sortie le plongement correspondant à l'entrée mais des paramètres pour la loi de probabilité choisie par le modélisateur. Les plongements sont le résultat d'un tirage à partir de cette distribution. La seconde concerne la fonction de coût utilisée pour l'apprentissage du modèle. En plus du critère de reconstruction, l'entraînement du VAE doit optimiser un terme de régularisation assurant que les lois de probabilité associées à chaque plongement doivent se rapprocher le plus possible du *prior*.

Le bénéfice de l'approche VAE est double par rapport à celle des auto-encodeurs. Premièrement, les espaces latents formés par les VAE ont tendance à être plus interprétables. Par exemple, lorsque les VAE sont utilisés sur des signaux de parole, ils sont susceptibles de représenter dans leur espace latent l'identité du locuteur ou les caractéristiques phonétiques du message encodé (Hsu et al., 2017). Deuxièmement, il est possible de générer de nouvelles données en fournissant au décodeur du VAE un plongement tiré aléatoirement en suivant le *prior* sur l'espace latent. Cependant, il faut noter que ces avantages sont obtenus au prix d'une moins bonne qualité de reconstruction de la part des VAEs par rapport aux auto-encodeurs, à cause de ce terme de régularisation qui vient contrecarrer jusqu'à un certain point le terme d'optimisation de la reconstruction des données.

1.3.1.4 Auto-encodeur variationnel quantifié vectoriel

L'auto-encodeur variationnel quantifié vectoriel (en anglais *vector quantized variational autoencoder*, VQ-VAE) est un auto-encodeur dont la conception s'inspire des méthodes de quantification vectorielle pour proposer un espace latent discrétisé (van den Oord et al., 2017). Tout comme le VAE, le VQ-VAE sera également utilisé dans ce travail et présenté plus en détail à la section 4.2.1. Son architecture comporte un encodeur et un décodeur implémentés à l'aide de réseaux de neurones ainsi qu'un mécanisme de discrétisation prenant place entre ceux-ci. Ce mécanisme fait intervenir un dictionnaire de plongements, dont la taille est fixe et définie *a priori*. Lorsqu'une donnée est présentée au VQ-VAE la sortie de son encodeur est comparée avec chacun des plongements présents au sein de ce dictionnaire, et le plongement le plus proche de la sortie de l'encodeur est alors sélectionné puis envoyé dans le décodeur. La définition des valeurs des plongements présents dans ce dictionnaire s'effectue lors de l'entraînement du réseau. Elle se fait grâce à une fonction de coût incluant, en plus de l'erreur de reconstruction, un terme s'assurant que la sortie de l'encodeur se rapproche du plongement sélectionné et un autre terme s'assurant que celui-ci se rapproche de la sortie de l'encodeur. Le résultat de la

combinaison de cette architecture avec cette méthode d'entraînement est un réseau capable de résumer des données provenant d'un espace continu à un nombre fini de classes, représentées par les plongements du dictionnaire.

1.3.2 Prédiction d'informations masquées

Une autre tâche prétexte utile à l'apprentissage auto-supervisé de représentations est la prédiction d'informations masquées. Le principe de cette tâche est, sur un signal découpé en plus petites parties, d'en masquer certaines, puis d'essayer de prédire les parties masquées à partir de celles restantes. Le recours à cette tâche repose sur l'hypothèse que les caractéristiques locales d'un signal sont en partie conditionnées par ses caractéristiques globales de plus haut niveau. Ainsi, afin de prédire correctement les parties masquées, un modèle entraîné sur cette tâche doit apprendre à extraire de chaque partie restante ses caractéristiques pertinentes au regard des attributs généraux du signal. Ces informations extraites forment des représentations qui peuvent ensuite être utilisées pour d'autres tâches. Dans la littérature, on peut trouver plusieurs manières d'implémenter cette tâche.

On peut d'abord distinguer une première famille d'implémentations dans lesquelles les parties de signal masquées sont celles situées après un instant de temps considéré (Chung et al., 2019 ; van den Oord et al., 2018). Leur but est d'entraîner un modèle au codage prédictif, qui est ici l'extraction de représentations du passé et du présent d'un signal contenant des informations utiles pour en prédire la suite. L'application de cette méthode à des signaux de parole permet d'obtenir des représentations porteuses d'informations phonémiques et d'informations caractérisant le locuteur.

On peut également distinguer une seconde famille d'implémentations dans lesquelles les parties de signal masquées sont celles à l'instant de temps considéré ainsi que d'autres sélectionnées aléatoirement parmi celles l'environnant, dans le passé comme dans le futur (Baevski et al., 2020 ; Hsu et al., 2021). Ces méthodes permettent de construire des représentations utilisables par exemple pour la reconnaissance automatique de la parole. En effet, avec seulement une dizaine de minutes de ces représentations couplées avec leur transcription, il est possible d'entraîner des modèles de reconnaissance de la parole aux performances équivalentes à celle de modèles entraînés sur des dizaines d'heures de parole brute (Baevski et al., 2020).

1.4 Architectures complètes d'apprentissage automatique de production de la parole

Nous avons passé en revue dans la première section de ce chapitre les grands enjeux théoriques sur lesquels s'appuient les théories du contrôle moteur de la production de la parole. Puis nous avons décrit les différentes composantes de modélisation permettant d'associer gestes articulatoires, représentations spectrales et sons, avant de tracer quelques grandes lignes des outils récents de modélisation et d'apprentissage statistique et notamment des modèles

neuronaux.

Sur cette base, de nombreux systèmes combinant des modèles articulatoires, des algorithmes de mise en correspondance statistique entre paramètres acoustiques et articulatoires et des principes de contrôle et d'apprentissage ont permis d'élaborer, dans des scénarios variés, des agents équipés d'un synthétiseur articulatoire capables d'explorer leur espace sensorimoteur et d'imiter des sons cibles de complexité variable (Bailly, 1997 ; Howard et Messum, 2014 ; Howard et Messum, 2007, 2011 ; Ishihara et al., 2008 ; Kröger et al., 2014 ; Miura et al., 2007 ; Moulin-Frier et al., 2014 ; Murakami et al., 2015 ; A. Philippsen, 2021 ; A. K. Philippsen et al., 2014 ; Pitti et al., 2021 ; Rasilo et Räsänen, 2017 ; Rasilo et al., 2013 ; voir aussi une revue dans Pagliarini et al., 2020). De manière frappante, tous ces modèles complets d'exploration/imitation se sont en général cantonnés à des expériences simples, à base d'apprentissage du contrôle de *stimuli* élémentaires, typiquement des voyelles tenues, des consonnes en contexte vocalique, et souvent des *stimuli* de synthèse (à l'exception de l'étude de Pitti et al., 2021 qui portait sur des phrases naturelles, mais en fait ne contenait pas de modèle articulatoire et constituait en réalité une étude purement acoustique).

À l'inverse, on a vu se développer pendant ces dernières années un ensemble de travaux portant sur l'apprentissage auto-supervisé de représentations de la parole à partir de très larges ensembles de données acoustiques, systématiquement multi-locuteurs et souvent multilingues, et, crucialement, non étiquetées (e.g. Chung et al., 2019). Ces représentations permettent d'entraîner des modèles génératifs du langage parlé qui peuvent être utilisés pour reproduire et étendre un ensemble de *stimuli* audio d'entrée (Polyak et al., 2021), ou pour connecter des systèmes de reconnaissance et de synthèse de la parole (Tjandra et al., 2020). Un certain nombre de développements récents dans ce domaine ont eu lieu dans le cadre du *Zero Resource Speech Challenge*, visant à construire des systèmes de traitement de la parole de façon auto-supervisée, sur la base de données acoustiques seules et sans ajout de connaissances supplémentaires, notamment textuelles (Dunbar et al., 2019 ; Morita & Koda, 2020). On citera également les travaux menés par Dupoux et coll., et notamment leur architecture de *Generative Spoken Language Modeling* (GSLM, Lakhotia et al., 2021) qui intègre (1) un encodage du signal de parole à l'aide de techniques d'apprentissage profond auto-supervisé de type CPC (van den Oord et al., 2018), wav2vec (Baevski et al., 2020), ou HuBERT (Hsu et al., 2021), (2) une découverte d'unités discrètes de la parole par quantification de ces représentations, et (3) la génération d'une voix de synthèse de très haute qualité en conditionnant un modèle *text-to-speech* neuronal de type Tacotron 2 (Shen et al., 2018) à partir de la séquence d'unités décodées. Il s'agit donc ici d'un système modélisant une boucle complète de perception et de production, qui permet d'une part des développements technologiques – comme par exemple le codage très bas débit (Polyak et al., 2021) – et d'autre part l'étude par le biais de la simulation de certains mécanismes qui sous-tendent l'acquisition de la parole et du langage. Cependant, il est important de noter que la quasi-totalité des systèmes développés n'intègrent pas de connaissances sur les processus de contrôle, de transformation articulatori-acoustique et de génération des sons (si ce n'est des connaissances élémentaires, par exemple sur les modèles de type source-conduit). Ainsi, ce champ de développements très spectaculaires, s'il a produit des avancées multiples sur les processus d'apprentissage de représentations à différents niveaux de la chaîne de traitement de la parole, ne s'est jusqu'à présent pas intéressé

de manière significative au problème de l'apprentissage du contrôle et de la production de la parole.

1.5 Objectifs de cette thèse

Le cadre de cette thèse s'inscrit précisément à la croisée des deux axes de recherche mentionnés dans la section précédente. Il adopte une approche similaire aux modèles et simulations de processus d'apprentissage, d'exploration et d'imitation présentés ci-dessus, mais il tente de leur faire franchir une étape qui semble majeure, en confrontant ces modèles au « vrai monde » des signaux de parole réelle, et si possible sans limiter leur complexité. Pour réaliser ces développements, nous proposerons dans tout ce qui suit des architectures entièrement « neuronales », c'est-à-dire de type réseaux de neurones profonds (*Deep Neural Networks*, DNN). Ces architectures nous permettront, par la puissance des algorithmes actuels d'apprentissage automatique, d'apprendre les correspondances entre données articulatoires et données acoustiques, de faire émerger des espaces de représentations latentes de divers types, et finalement de construire, brique après brique, les éléments d'un système complet d'apprentissage de certains mécanismes de base de la production de la parole. L'ensemble de ces développements s'inscrira dans un cadre d'apprentissage auto ou faiblement supervisé, dans lequel l'objectif ultime sera de développer des agents capables, sans avoir accès à des informations articulatoires ou phonologiques, de mettre en place des processus d'inférence permettant d'estimer les informations adéquates uniquement à partir d'entrées acoustiques fournies par leur environnement.

Chapitre 2

Jeux de données articulatoires et acoustiques

Sommaire

L'entièreté des travaux présentés dans ce manuscrit se base sur des corpus d'enregistrements acoustiques et articulatoires synchronisés de parole. Cette section commence par une présentation du type de données requis pour nos expériences et commun aux corpus utilisés, puis continue en présentant plus en détail le contenu et la méthode d'enregistrement de chacun d'eux. Ces corpus sont tous ouverts et disponibles au public.

2.1 Type de données utilisé

Les travaux présentés dans les sections suivantes se basent sur des enregistrements EMA (technique présentée en section 1.2.2.3) de la mâchoire, de la langue, des lèvres et du velum d'un locuteur produisant des sons de parole, synchronisés avec l'enregistrement acoustique correspondant. Dans chacun des corpus utilisés la position de la mâchoire est capturée à l'aide d'une bobine positionnée sur une incisive inférieure, celle de la langue par trois bobines positionnées sur l'apex, le milieu et le dos de celle-ci, celle des lèvres par deux bobines positionnées l'une sur la lèvre supérieure et l'autre sur l'inférieure, et, pour certains corpus, celle du velum à l'aide d'une bobine positionnée sur celui-ci. Le placement susmentionné des 7 bobines est illustré figure 2.1. Dans nos expériences, les coordonnées des bobines sont toutes exprimées en 2 dimensions dans le plan sagittal médian et échantillonnées à 100 Hz.

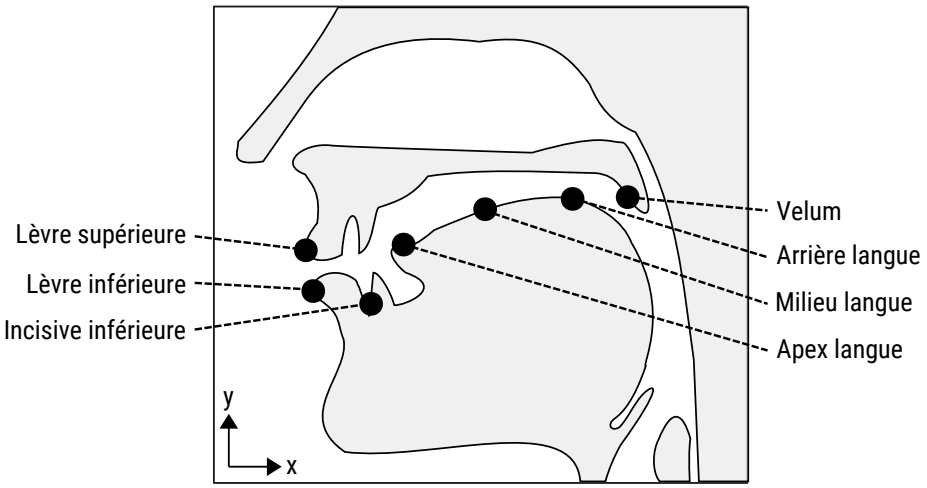

FIGURE 2.1 – **Placement des bobines EMA pour l'enregistrement des données articulatoires**

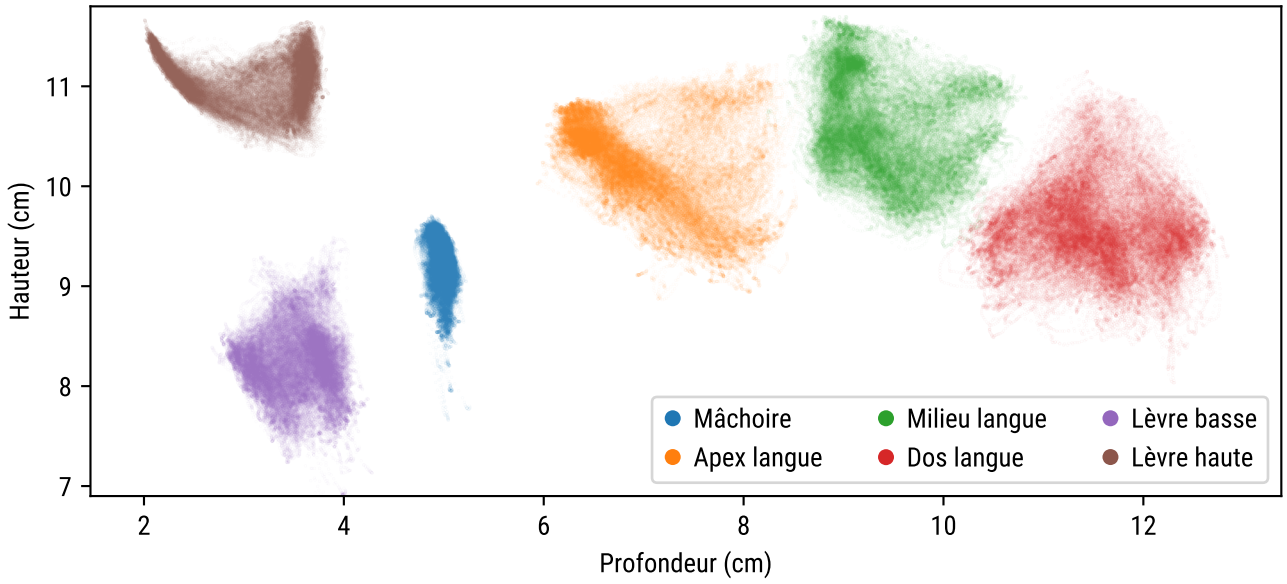

FIGURE 2.2 – **Distribution des positions des bobines EMA dans le plan sagittal médian pour le corpus PB2007** La référence $(0, 0)$ est arbitraire.

2.2 Corpus PB2007

Le corpus PB2007 (Badin et al., 2008) comprend l'enregistrement de 1 109 locutions produites par un locuteur français masculin. Cet ensemble de productions est constitué de 14 voyelles françaises isolées et de 224 séquences VCV voyelle-consonne-voyelle où C et V sont respectivement l'une des 16 consonnes et des 14 voyelles du français. Chacune de ces voyelles isolées et de ces séquences VCV a été enregistrée 2 fois et cette partie du corpus représente environ 54 % des productions. Le reste est constitué de mots isolés (36 % des productions) et de phrases (10 % des productions). Une fois les silences retirés, ce corpus représente 17 minutes de parole.

L'appareil d'enregistrement ayant servi à la capture des données EMA est le système Carstens 2D EMA (AG200). Contrairement aux autres corpus présentés dans ce manuscrit, celui-ci ne contient pas d'informations à propos du velum dont la position n'a pas été enregistrée. Les trajectoires des bobines ainsi acquises ont d'abord été filtrées à l'aide d'un filtre passe-bas à 20 Hz, puis ont été sous-échantillonnées de 200 Hz à 100 Hz, pour un total de 134 707 vecteurs d'observations articulatoires. La distribution des positions des bobines EMA pour ce corpus est illustrée figure 2.2.

Les données acoustiques ont été enregistrées à 22 050 Hz de façon synchronisée à l'enregistrement articulatoire. Ces enregistrements ont servi à l'étiquetage au niveau phonémique du corpus via une procédure d'alignement forcé basée sur les transcriptions manuelles du corpus et un modèle de Markov caché multi-locuteur. Ces alignements forcés ont ensuite été corrigés manuellement.

Ce corpus est présenté en détail dans Ben Youssef, 2011 et est disponible au téléchargement à l'adresse https://zenodo.org/record/6390598.

2.2.1 Corpus GB2016 et TH2016

Dans des expériences de généralisation à d'autres locuteurs ne nécessitant pas d'informations articulatoires, les corpus GB2016 et TH2016 ont été employés. Ceux-ci ne comportent que des enregistrements audio, dont le contenu est identique à celui de PB2007, mais produits par deux autres locuteurs masculins et français. La labellisation de ces deux corpus a été obtenue en alignant les transcriptions de PB2007 sur ces nouvelles versions en utilisant le *Montreal Forced Aligner* (McAuliffe et al., 2017).

2.3 Corpus BY2014

Le corpus BY2014 (Bocquelet et al., 2016) est composé de 925 productions d'un locuteur français masculin pour un total de 45 minutes de parole une fois les silences retirés. Ces productions incluent des voyelles françaises isolées, des séquences voyelle-consonne-voyelle en

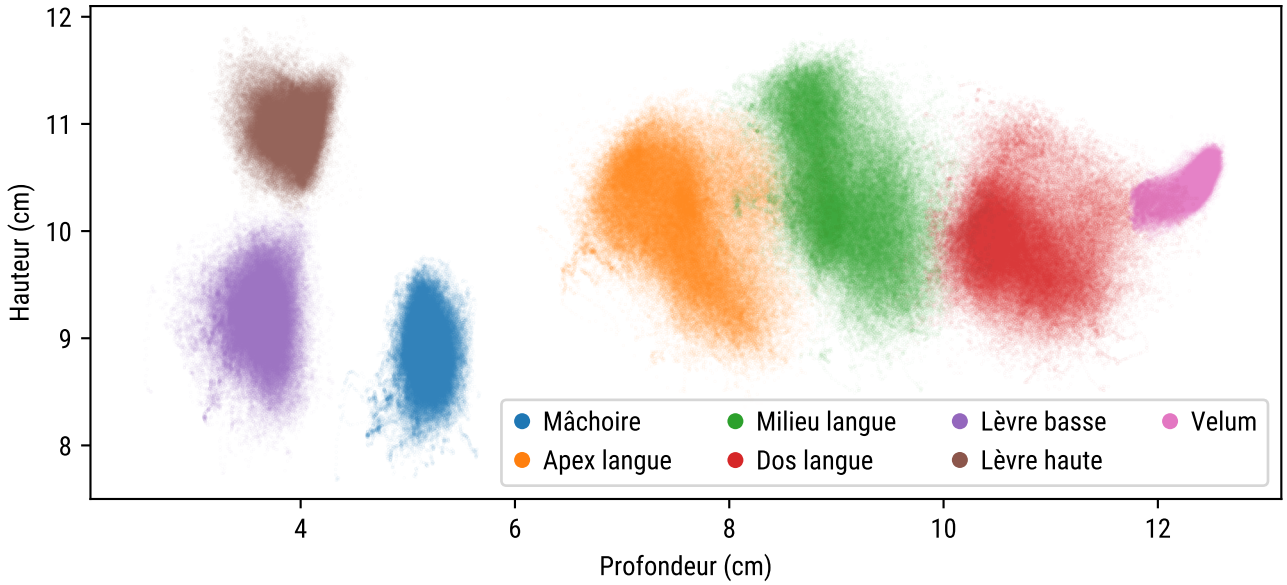

FIGURE 2.3 – **Distribution des positions des bobines EMA dans le plan sagittal médian pour le corpus BY2014** La référence $(0, 0)$ est arbitraire.

contexte, des phrases phonétiquement équilibrées et des phrases issues d'articles de presse.

Les données EMA ont été acquises grâce au système NDI Wave (NDI, Ontario, Canada), permettant l'enregistrement de la position de chaque bobine en 3 dimensions à 400 Hz. Comme nos travaux n'ont recours qu'à des coordonnées de bobine en 2 dimensions, les positions enregistrées ont été projetées sur un plan sagittal médian estimé *a posteriori* en prenant comme référence la position moyenne des 3 bobines de mâchoire, de dos de langue et de velum. La distribution des positions des bobines EMA pour ce corpus est illustrée figure 2.3.

À cause d'un défaut du système d'enregistrement, ce corpus présente des « trous » de durées variables dans les trajectoires capturées. Les sauts supérieurs à 50 ms, qui représentent 0,54 % de la durée totale du corpus et touchent environ 17 % des productions enregistrées, ont été retirés avant manipulation.

Les données acoustiques ont été enregistrées à 22 050 Hz de façon synchronisée à l'enregistrement articulatoire. Ces enregistrements ont d'abord été retranscrits manuellement sous forme de texte puis traduits en phonèmes grâce au phonétiseur Lia-PHON (Béchet, 2001). Ces phonèmes obtenus automatiquement ont ensuite été corrigés manuellement et alignés avec le son avec un système standard de reconnaissance de la parole afin d'obtenir la labellisation du corpus.

Ce corpus est décrit en détail dans Bocquelet, 2017 et est disponible au téléchargement à l'adresse https://zenodo.org/record/154083.

2.4 Corpus MOCHA

Le corpus MOCHA (Wrench, 1999) est constitué de productions de 2 locuteurs anglais, l'un masculin et l'autre féminin, respectivement désignés par « msak0 » et « fsew0 ». Lors de leur enregistrement, chacun des locuteurs a produit les mêmes 460 phrases courtes en anglais pour une durée respective de 17 et 20 minutes une fois les silences retirés.

Les trajectoires articulatoires ont été acquises avec un articulographe Carstens à un taux d'échantillonage de 500 Hz. Pour nos différentes expériences, nous les avons sous-échantillonné à 100 Hz. La distribution des positions des bobines EMA pour ce corpus est illustrée figure 2.4.

Les enregistrements acoustiques ont eux été acquis avec un microphone Audio-Technica ATM10a à 16 kHz. Ces enregistrements sont disponibles munis d'un étiquetage phonétique, sans spécification claire (d'après nos lectures aussi attentives que possibles) de la méthodologie avec laquelle ils ont été obtenus.

Ce corpus peut être téléchargé à l'adresse : https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

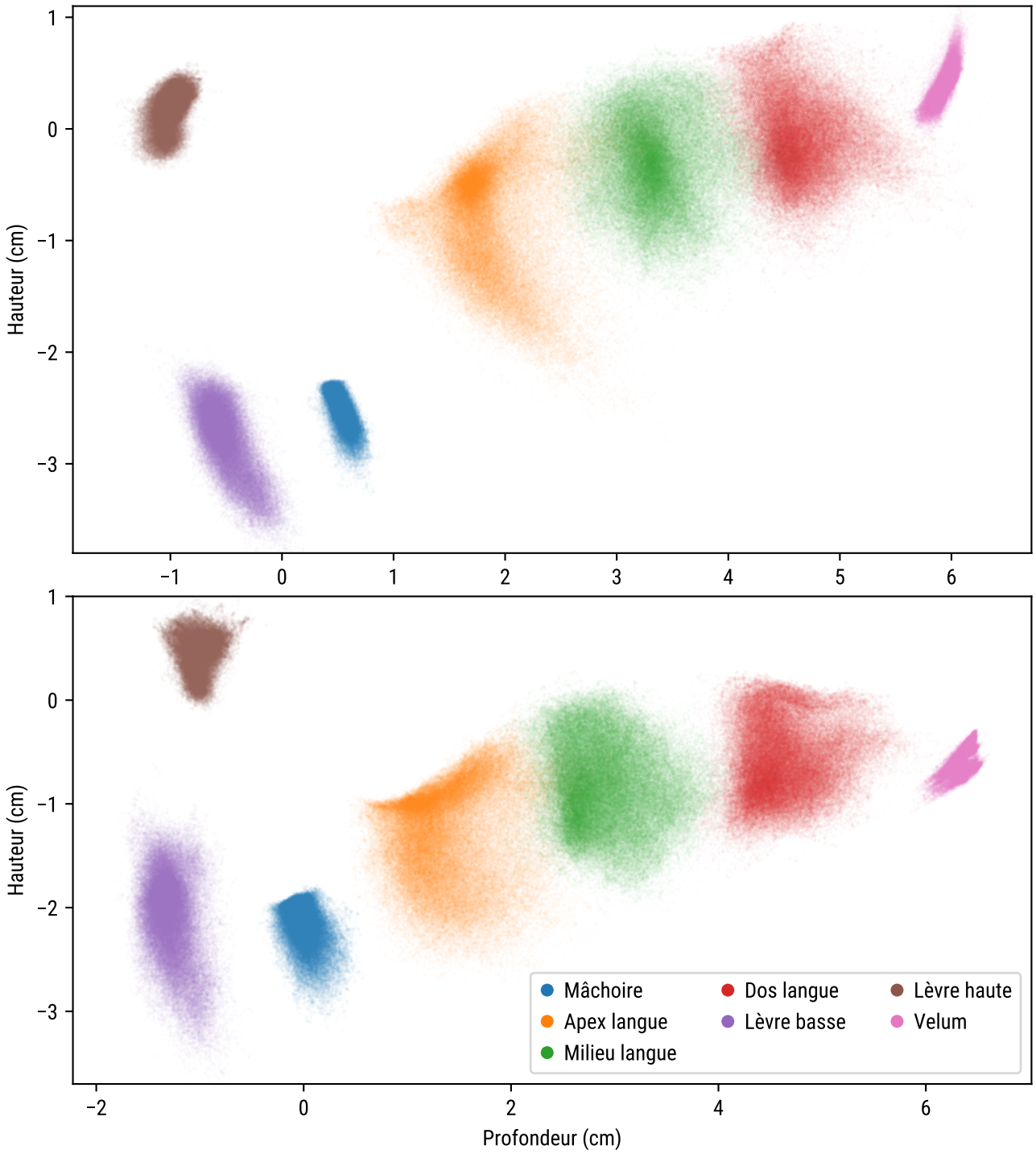

FIGURE 2.4 – **Distribution des positions des bobines EMA dans le plan sagittal médian pour le corpus MOCHA (en haut, locuteur msak0, en bas, locuteur fsew0)** La référence $(0, 0)$ est arbitraire.

Chapitre 3

Synthétiseur articulatoire neuronal

Sommaire

Le travail présenté dans cette section a fait l'objet de la publication :

— Georges, M.-A., Badin, P., Diard, J., Girin, L., Schwartz, J.-L. & Hueber, T. (2020). Towards an articulatory-driven neural vocoder for speech synthesis. *International Seminar on Speech Production (ISSP)*

Ce chapitre présente la construction du synthétiseur articulatoire qui sera piloté par l'agent. Le synthétiseur est construit en se basant sur les enregistrements articulatoires EMA et acoustiques d'un locuteur, appelé par la suite « locuteur de référence ». Ce processus de création, illustré figure 3.1, se déroule en trois étapes. La première est la création d'un modèle articulatoire pour traduire les enregistrements EMA en paramètres articulatoires interprétables représentant au mieux les degrés de liberté de l'appareil vocal. La seconde est la création d'un modèle articulatoire-vers-acoustique constitué d'un réseau de neurones entraîné à faire la correspondance entre les paramètres articulatoires d'une part et le contenu spectral du signal acoustique associé d'autre part. Enfin, la dernière étape est l'entraînement d'un vocodeur neuronal permettant la reconstruction d'un signal acoustique à partir du contenu spectral estimé.

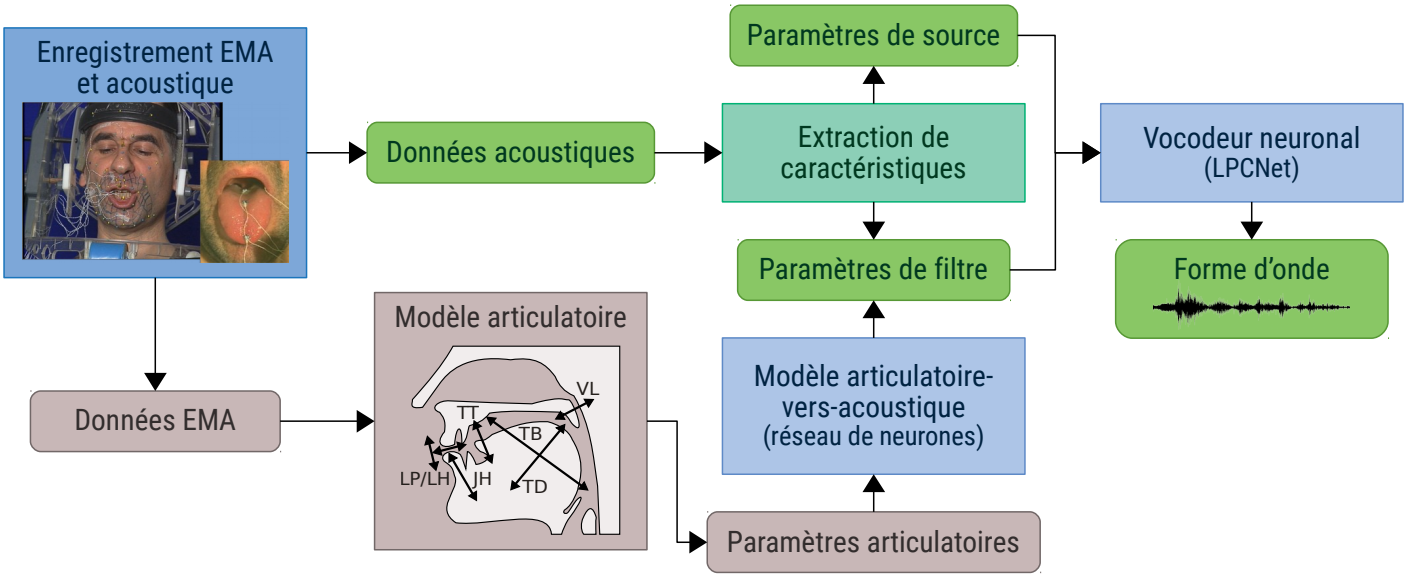

FIGURE 3.1 – **Vue d'ensemble du processus de création du synthétiseur articulatoire**

3.1 Modèle articulatoire

Cette première section se divise en trois sous-parties. La première présente le modèle articulatoire que nous utiliserons tout au long de ce manuscrit afin d'extraire des paramètres articulatoires interprétables à partir d'enregistrements EMA. Ce modèle est celui développé par Maeda, 1990, dont nous avons mentionné section 1.2.3.1 l'intérêt et la pertinence, toujours attestée plus de 30 ans après sa publication. Comme nous le verrons, ce modèle se base exclusivement sur des méthodes linéaires, susceptibles de limiter sa capacité à décrire finement les mouvements des articulateurs. C'est pourquoi la seconde et la troisième présentent chacune une tentative d'amélioration de ce premier modèle articulatoire en adaptant sa méthodologie aux réseaux de neurones.

3.1.1 Modèle articulatoire linéaire

Cette sous-partie présente la méthode de construction du modèle articulatoire qui sera utilisé tout au long de ce manuscrit. Le modèle articulatoire développé par Maeda, 1990, était au départ prévu pour fonctionner à partir d'enregistrements de formes de conduit vocal obtenues par cinéradiographie. La présente adaptation (Serrurier et al., 2012) est construite à partir d'un corpus d'enregistrements EMA et permet de traduire ce type de données en paramètres articulatoires interprétables, et en retour, de reconstruire les données à partir des paramètres.

3.1.1.1 Définition du modèle

Le modèle linéaire s'applique à des corpus de données EMA comprenant l'enregistrement des positions d'un nombre B_{EMA} de bobines, placées respectivement sur la mâchoire, l'apex, le milieu et l'arrière de la langue, la lèvre supérieure et inférieure, et potentiellement sur le velum.

Le modèle articulatoire est défini par :

$$\mathbf{e} \simeq \mathbf{aW} + \mathbf{m}_e,$$

avec :

- $\mathbf{e}$ vecteur de taille $(1, B_{EMA} \times D)$ contenant les coordonnées sur D dimensions des bobines EMA ;
- $\mathbf{a}$ vecteur de taille $(1, N_p)$ contenant la valeur des N_p paramètres articulatoires ;
- $\mathbf{W}$ matrice de taille $(N_p, B_{EMA} \times D)$ permettant de calculer les variations de position des bobines, autour de leur position moyenne, induites par l'action des paramètres articulatoires ;
- $\mathbf{m}_e$ vecteur de taille $(1, B_{EMA} \times D)$ décrivant la position moyenne [1] des bobines EMA.

Pour un vecteur de coordonnées EMA $\mathbf{e}$, le vecteur de paramètres articulatoires correspondant $\mathbf{a}$ est donné par le modèle inverse, qui est aussi une forme linéaire :

$$\mathbf{a} \simeq (\mathbf{e} - \mathbf{m}_e)\mathbf{W}_{inv},$$

avec $\mathbf{W}_{inv}$ matrice de taille $(B_{EMA} \times D, N_p)$.

Le corpus de données composé de M observations articulatoires est représenté par une matrice $\mathbf{E}$ de taille $(M, B_{EMA} \times D)$.

Les corpus utilisés dans cette thèse diffèrent par le nombre de bobines EMA utilisées lors de leur enregistrement. Ainsi, pour BY2014 et MOCHA, $B_{EMA} = 7$, et pour PB2007, $B_{EMA} = 6$ (ce corpus ne comporte pas de bobine de velum). Pour chaque corpus, chaque configuration articulatoire est ramenée dans le plan sagittal médian (donc $D = 2$). Dans la suite de cette section, nous détaillons la procédure d'estimation des paramètres dans le cas d'un corpus avec $B_{EMA} = 7$ bobines en $D = 2$ dimensions.

3.1.1.2 Estimation des paramètres

L'estimation des paramètres du modèle consiste à trouver les valeurs optimales de $\mathbf{m}_e$ et $\mathbf{W}$ pour le modèle direct, à partir d'un corpus de données EMA, puis à déterminer la valeur de $\mathbf{W}_{inv}$ pour le modèle inverse. Cette estimation se fait avec une méthode appelée « PCA guidée » et qui permet d'obtenir des paramètres de contrôle ayant un sens articulatoire – c'est à dire que chacun de ces paramètres n'influe potentiellement sur la position que d'un

1. La position moyenne du conduit n'a pas de véritable sens articulatoire ou physiologique, même si elle pourrait être proche d'une position de repos (difficile à définir et dépendante du locuteur), d'une position préphonatoire, ou d'un schwa qui se trouve approximativement au centre du triangle vocalique.

seul articulateur. Pour obtenir de tels paramètres, la PCA guidée repose sur une définition *a priori* de paramètres articulatoires de contrôle, et sur une analyse itérative des données basée sur une combinaison de PCA et régressions linéaires, permettant d'extraire itérativement la contribution de chacun des articulateurs. Cette méthode permet de conserver la propriété de la PCA classique de fournir des dimensions pour partie décorrélées, mais, au prix d'un abandon de l'orthogonalité des composantes, de converger vers des composantes interprétables.

Les paramètres du modèle articulatoire choisis sont au nombre de 7 : *JawHeight* (*JH*) décrit la hauteur de la mâchoire, *TongueBody* (*TB*) décrit la position antéro-postérieure de la langue, *TongueDorsum* (*TD*) décrit la hauteur de la langue, *TongueTip* (*TT*) décrit la hauteur de la pointe de la langue, *LipHeight* (*LH*) décrit l'écartement vertical des lèvres, *LipProtrusion* (*LP*) décrit la protrusion des lèvres, et enfin *Velum* (*VL*) décrit le degré de fermeture du velum.

Le reste de cette partie est consacré à l'estimation de ces paramètres. Les trajectoires induites par la variation de chacun des paramètres du modèle construit avec le corpus PB2007 sont visibles figure 3.2.

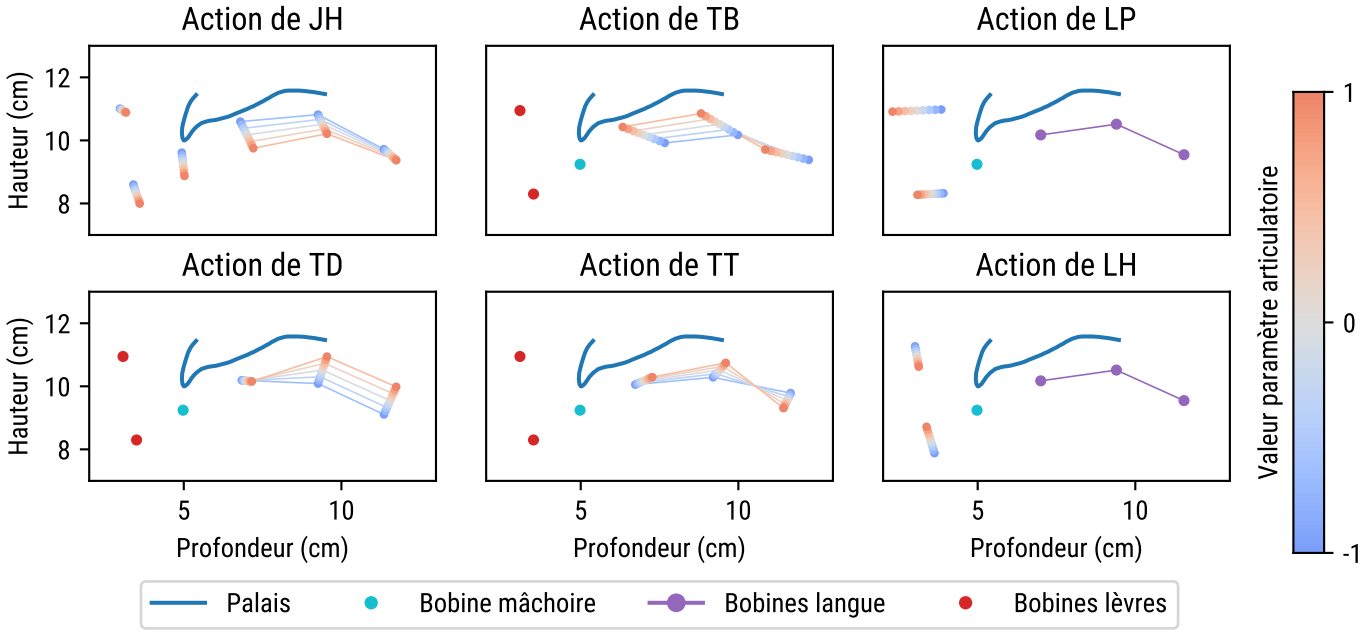

FIGURE 3.2 – **Trajectoires des bobines induites par la variation de chacun des paramètres articulatoires du modèle linéaire construit avec le corpus PB2007** Les points fixes représentent la position moyenne des bobines.

Mâchoire Le premier articulateur à être modélisé est la mâchoire, avec la relation suivante :

$$\mathbf{e}_{jaw} \simeq \mathbf{a}_{JH}\mathbf{W}_{JH} + \mathbf{m}_{e_{jaw}},$$

avec :

— $\mathbf{e}_{jaw}$ vecteur de taille $(1, 1 \times 2)$ contenant les coordonnées x et y de la bobine de mâchoire ;
— $\mathbf{m}_{e_{jaw}}$ vecteur de taille $(1, 1 \times 2)$ contenant les coordonnées en x et en y de la bobine de mâchoire en position moyenne ;
— $\mathbf{a}_{JH}$ vecteur de taille $(1, 1)$ contenant la valeur du paramètre *JawHeight* ;
— $\mathbf{W}_{JH}$ matrice de taille $(1, 1 \times 2)$ permettant de calculer les variations de la position de la bobine de mâchoire autour de sa position moyenne $\mathbf{m}_{e_{jaw}}$, induites par la valeur de $\mathbf{a}_{JH}$.

La recherche de $\mathbf{W}_{JH}$ et $\mathbf{m}_{e_{jaw}}$ se fait en travaillant sur la matrice $\mathbf{E}_{jaw}$, sous-ensemble de $\mathbf{E}$ ne contenant que les colonnes relatives aux positions de la bobine de mâchoire.

Le vecteur $\mathbf{m}_{e_{jaw}}$ correspond à la moyenne des positions de mâchoire enregistrées, et est estimé par :

$$\mathbf{m}_{e_{jaw}} = \frac{1}{M} \sum_{i=1}^{M} (\mathbf{E}_{jaw})_{i,:},$$

où le symbole « : » désigne l'ensemble de toutes les entrées sur une ligne ou une colonne donnée.

La matrice $\mathbf{W}_{JH}$ est obtenue en appliquant une PCA à $\mathbf{E}_{jaw}$. Les valeurs de $\mathbf{W}_{JH}$ sont alors définies comme celles permettant la projection d'une position de mâchoire $\mathbf{e}_{jaw}$ sur le premier axe de cette PCA.

Les projections des échantillons contenus dans $\mathbf{E}_{jaw}$ sur le premier axe de la PCA sont alors compilées dans une matrice $\mathbf{A}_{JH}$ de taille $(M, 1)$ pour le calcul des paramètres des articulateurs suivants.

Mâchoire et langue L'estimation des paramètres utilisés pour modéliser la langue se fait en 3 temps. D'abord, la contribution de la mâchoire, qui porte la langue, à sa position est estimée par régression linéaire. Ensuite, ce sont les paramètres représentant l'avancement (TB) et la hauteur de la langue (TD) qui sont extraits, puis enfin le paramètre représentant le mouvement de l'apex (TT).

La contribution de la mâchoire à la position de la langue est modélisée par l'équation :

$$\mathbf{e}_{tng} \simeq \mathbf{a}_{JH}\mathbf{W}_{JT} + \mathbf{m}_{e_{tng}} + \mathbf{r}_{tng},$$

avec :
— $\mathbf{e}_{tng}$ vecteur de taille $(1, 3 \times 2)$ contenant les coordonnées des 3 bobines de langue ;
— $\mathbf{m}_{e_{tng}}$ vecteur de taille $(1, 3 \times 2)$ contenant les coordonnées des 3 bobines de langue en position moyenne ;
— $\mathbf{a}_{JH}$ vecteur de taille $(1, 1)$ contenant la valeur du paramètre *JawHeight*, extrait à l'étape précédente ;
— $\mathbf{W}_{JT}$ matrice de taille $(1, 3 \times 2)$ permettant de calculer les variations de position des bobines de langue induites par JH ;

— $\mathbf{r}_{tng}$ vecteur de taille $(1, 3\times2)$ contenant les variations de position des bobines de langue supposées causées exclusivement par l'action de la langue elle-même. Ce vecteur sera qualifié par la suite de « mouvement résiduel de langue », c'est-à-dire les mouvements de celle-ci une fois pris en compte et éliminés ceux induits par la mâchoire.

Pour calculer $\mathbf{m}_{e_{tng}}$ et $\mathbf{W}_{JT}$, les calculs sont effectués sur la matrice $\mathbf{E}_{tng}$ contenant les positions des bobines de langue du corpus $\mathbf{E}$. Comme pour $\mathbf{m}_{e_{jaw}}$, la valeur de $\mathbf{m}_{e_{tng}}$ est définie comme la position moyenne des bobines de langue sur l'ensemble du corpus d'apprentissage.

La valeur de $\mathbf{W}_{JT}$ est obtenue par un processus classique de régression linéaire des variations des positions des 3 bobines de langues en fonction du paramètre articulatoire de mâchoire, en minimisant :

$$\|\mathbf{E}_{tng} - \mathbf{1}_{M\times1}\mathbf{m}_{e_{tng}} - \mathbf{A}_{JH}\mathbf{W}_{JT}\|^2,$$

avec $\mathbf{1}_{M\times1}$ matrice de taille $(M, 1)$ et dont chacune des entrées vaut 1.

Une fois la valeur de $\mathbf{W}_{JT}$ connue, il devient possible de calculer $\mathbf{R}_{tng}$, matrice de taille $(M, 3\times2)$ contenant pour chacun des échantillons du corpus la valeur des mouvements résiduels de langue après avoir éliminé les mouvements dus à l'action de la mâchoire :

$$\mathbf{R}_{tng} = \mathbf{E}_{tng} - \mathbf{1}_{M\times1}\mathbf{m}_{e_{tng}} - \mathbf{A}_{JH}\mathbf{W}_{JT}.$$

TongueBody* et *TongueDorsum Le mouvement résiduel de langue $\mathbf{r}_{tng}$ « nettoyé » de la contribution de la mâchoire *JH* est supposé être le résultat de l'action de *TongueBody*, *TongueDorsum* et *TongueTip*. Son modèle est :

$$\mathbf{r}_{tng} \simeq \mathbf{a}_{TBTD}\mathbf{W}_{TBTD} + \mathbf{r}_{tip},$$

avec les nouveaux termes :

— $\mathbf{a}_{TBTD}$, vecteur de taille $(1, 2)$ contenant les valeurs respectives de *TB* et *TD* ;

— $\mathbf{W}_{TBTD}$, matrice de taille $(2, 3 \times 2)$ permettant de calculer les variations de position des bobines de langue induites par *TB* et *TD* ;

— $\mathbf{r}_{tip}$, vecteur de taille $(1, 3\times2)$ contenant la partie du mouvement résiduel $\mathbf{r}_{tng}$ imputable à l'action de *TT*.

L'avancée et la montée de la langue sont des mouvements « globaux », qui ne renseignent pas sur l'activité de l'apex. Or, calculer la contribution de *TB* et *TD*, $\mathbf{W}_{TBTD}$, sur l'ensemble du mouvement résiduel de langue du corpus, $\mathbf{R}_{tng}$, conduirait à encoder également dans ces deux paramètres les variations en bout de langue, qui doivent en réalité être attribuées à *TT*. C'est pourquoi $\mathbf{W}_{TBTD}$ est estimée en deux étapes. D'abord, une PCA est effectuée sur la sous-partie de $\mathbf{R}_{tng}$ ne contenant que le mouvement résiduel des bobines milieu et arrière de langue. La matrice des 2 premières dimensions résultant de cette PCA donne une première matrice, $\mathbf{W}_{T1}$, et la projection des données sur ces deux dimensions donne les valeurs de *TB* et *TD* sur l'ensemble du corpus, $\mathbf{A}_{TBTD}$. Ensuite, une seconde matrice, $\mathbf{W}_{T2}$, est estimée par régression linéaire sur la sous-partie de $\mathbf{R}_{tng}$ ne contenant que le mouvement résiduel de la bobine avant de langue. Une fois $\mathbf{W}_{T1}$ et $\mathbf{W}_{T2}$ estimées, elles sont concaténées pour enfin former $\mathbf{W}_{TBTD} = [\mathbf{W}_{T2}\mathbf{W}_{T1}]$.

TongueTip $\mathbf{r}_{tip}$, le mouvement résiduel de langue « nettoyé » de toute contribution extérieure à *TongueTip*, est modélisé par :

$$\mathbf{r}_{tip} \simeq \mathbf{a}_{TT}\mathbf{W}_{TT},$$

avec les nouveaux termes :

- $\mathbf{a}_{TT}$, vecteur de taille $(1, 1)$ contenant la valeur TT ;
- $\mathbf{W}_{TT}$, matrice de taille $(1, 3 \times 2)$ permettant de calculer les variations de position des bobines de langue imputables à TT.

Pour estimer $\mathbf{W}_{TT}$, une PCA est appliquée au mouvement résiduel obtenu sur le corpus, $\mathbf{R}_{tip}$. $\mathbf{W}_{TT}$ est la matrice représentant la projection des données sur la première dimension de la PCA.

Mâchoire et lèvres La modélisation du mouvement des bobines de lèvres se fait en 3 étapes. D'abord, la contribution de la mâchoire à leur position est estimée. Ensuite, les valeurs de LP et LH sont estimées pour le corpus étudié, pour enfin permettre de calculer leur contribution.

Les positions des bobines des lèvres sont modélisées par l'équation :

$$\mathbf{e}_{lip} \simeq \mathbf{a}_{JH}\mathbf{W}_{JL} + \mathbf{m}_{e_{lip}} + \mathbf{r}_{lip},$$

avec :

- $\mathbf{e}_{lip}$ vecteur de taille $(1, 2 \times 2)$ contenant les coordonnées des 2 bobines des lèvres ;
- $\mathbf{m}_{e_{lip}}$ vecteur de taille $(1, 2 \times 2)$ contenant les coordonnées des bobines des lèvres en position moyenne ;
- $\mathbf{a}_{JH}$ vecteur de taille $(1, 1)$ contenant la valeur du paramètre *JawHeight* ;
- $\mathbf{W}_{JL}$ matrice de taille $(1, 2 \times 2)$ permettant de calculer les variations de position des bobines des lèvres imputables à JH ;
- $\mathbf{r}_{lip}$ vecteur de taille $(1, 2 \times 2)$ contenant les mouvements résiduels des variations de position des bobines des lèvres, une fois soustraite l'action de la mâchoire par régression linéaire, et supposées causées par la seule action des lèvres.

Pour calculer $\mathbf{m}_{e_{lip}}$ et $\mathbf{W}_{JL}$, les calculs sont effectués sur la matrice $\mathbf{E}_{lip}$ contenant les positions des bobines des lèvres du corpus $\mathbf{E}$.

Comme pour $\mathbf{m}_{e_{jaw}}$ et $\mathbf{m}_{e_{tng}}$, la valeur de $\mathbf{m}_{e_{lip}}$ est définie comme la position moyenne des bobines de lèvres sur l'ensemble du corpus d'apprentissage.

La valeur de $\mathbf{W}_{JL}$ est obtenue par régression linéaire entre l'articulateur mâchoire et les bobines des lèvres, en minimisant $\|\mathbf{E}_{lip} - \mathbf{1}_{M\times 1}\mathbf{m}_{e_{lip}} - \mathbf{A}_{JH}\mathbf{W}_{JL}\|^2$.

Une fois la valeur de $\mathbf{W}_{JL}$ connue, la matrice $\mathbf{R}_{lip}$ de taille $(M, 2 \times 2)$ contenant la valeur du mouvement résiduel de lèvres pour chacun des échantillons du corpus est calculée :

$$\mathbf{R}_{lip} = \mathbf{E}_{lip} - \mathbf{1}_{M\times 1}\mathbf{m}_{e_{lip}} - \mathbf{A}_{JH}\mathbf{W}_{JL}$$

Le mouvement résiduel de lèvres, $\mathbf{r}_{lip}$, suit la relation :

$$\mathbf{r}_{lip} \simeq \mathbf{a}_{LPLH}\mathbf{W}_{LPLH},$$

avec les nouveaux termes :

— $\mathbf{a}_{LPLH}$, vecteur de taille $(1, 2)$ contenant la valeur de LP et LH ;
— $\mathbf{W}_{LPLH}$, matrice de taille $(2, 2 \times 2)$ permettant de calculer les variations de position causées par LP et LH au mouvement résiduel $\mathbf{r}_{lip}$.

LipProtusion* et *LipHeight Si le calcul de $\mathbf{W}_{LPLH}$ était fait directement par une PCA appliquée à $\mathbf{R}_{lip}$, il n'est pas garanti que les deux premières composantes extraites reflètent des mouvement d'avancée et d'écartement des lèvres. C'est pourquoi $\mathbf{W}_{LPLH}$ est calculée en 2 étapes : d'abord des valeurs de LP et LH sont attribuées à chacun des échantillons du corpus, puis ensuite on cherche comment ces valeurs peuvent expliquer les variations de position observées.

Les valeurs de LP et LH sont estimées à partir des matrices $\mathbf{R}_{lip_x}$ et $\mathbf{R}_{lip_y}$, qui sont chacune une sous-matrice de $\mathbf{R}_{lip}$, ne contenant que les coordonnées x ou que les coordonnées y des deux bobines des lèvres.

Deux PCA sont appliquées respectivement à $\mathbf{R}_{lip_x}$ et $\mathbf{R}_{lip_y}$. La projection des données sur le premier axe de chacune de ces PCA donne $\mathbf{A}_{LP}$ et $\mathbf{A}_{LH}$, matrices de taille $(M, 1)$ contenant les valeurs de LP et LH pour chacun des échantillons du corpus. $\mathbf{A}_{LP}$ et $\mathbf{A}_{LH}$ sont concaténées pour former une unique matrice $\mathbf{A}_{LPLH}$ de taille $(M, 2)$.

Une fois les valeurs de LP et LH connues, leur contribution $\mathbf{W}_{LPLH}$ est calculée comme étant le terme minimisant :

$$\|\mathbf{R}_{lip} - \mathbf{A}_{LPLH}\mathbf{W}_{LPLH}\|^2.$$

Velum Le modèle fait l'hypothèse que la position du velum n'est pas influencée par celle des autres articulateurs. De ce fait, l'estimation des paramètres utilisés pour modéliser sa position se fait en une seule étape, et suit la même procédure que l'estimation des paramètres relatifs à la mâchoire. La position du velum est modélisée par la relation suivante :

$$\mathbf{e}_{vel} \simeq \mathbf{a}_{VL}\mathbf{W}_{VL} + \mathbf{m}_{e_{vel}},$$

avec :

— $\mathbf{e}_{vel}$ vecteur de taille $(1, 1 \times 2)$ contenant les coordonnées x et y de la bobine de velum ;
— $\mathbf{m}_{e_{vel}}$ vecteur de taille $(1, 1 \times 2)$ contenant les coordonnées en x et en y de la bobine de velum en position moyenne ;
— $\mathbf{a}_{VL}$ vecteur de taille $(1, 1)$ contenant la valeur du paramètre *Velum* ;
— $\mathbf{W}_{VL}$ matrice de taille $(1, 1 \times 2)$ permettant de calculer les variations de position de la bobine de velum autour de sa position moyenne $\mathbf{m}_{e_{vel}}$ induites par la valeur de $\mathbf{a}_{VL}$.

3.1.2 Adaptation non-linéaire littérale du modèle articulatoire

Le modèle linéaire que nous venons de présenter, bien que présentant un bon compromis entre adéquation aux données et interprétabilité des commandes, ne parvient pas à capturer

la totalité de l'information sur les positions des bobines. De ce fait, dans cette section, nous étudions si une adaptation du modèle articulatoire original au monde des réseaux de neurones permettrait d'en créer une nouvelle version, avec un meilleur pouvoir d'expression et par extension capable de fournir des paramètres articulatoires capturant davantage d'information tout en restant interprétables.

Pour ce qui concerne l'amélioration de l'adaptation aux données, on peut remarquer que la conception du modèle linéaire n'a recours qu'à deux outils mathématiques : la PCA et la régression linéaire. La PCA est utilisée à des fins de réduction de dimensionnalité et la régression linéaire est utilisée pour établir une relation entre deux espaces (celui d'un paramètre articulatoire associé à un articulateur et celui de la position des bobines supposées associées pour partie à un autre articulateur). Or, il existe parmi les réseaux de neurones deux outils fournissant les mêmes fonctionnalités : les auto-encodeurs pour la réduction de dimensionnalité, et les perceptrons multicouches pour la mise en correspondance d'espaces. Ainsi, il est possible de créer une version non-linéaire du modèle articulatoire décrit dans la partie précédente en suivant exactement les mêmes étapes mais en remplaçant l'utilisation de chacun des outils linéaires par leur équivalent non-linéaire. Concrètement, à chaque fois qu'une PCA dont on ne garde que n composantes est effectuée dans la version linéaire, un auto-encodeur lui est substitué avec n neurones dans son espace latent. De même, à chaque fois qu'une régression linéaire est employée, un perceptron multicouche lui est substitué.

En ce qui concerne l'interprétabilité, notre raisonnement a été le suivant. Dans la création du modèle non-linéaire, les réseaux de neurones utilisés pour l'extraction des paramètres articulatoires sont entraînés avec le même objectif que les méthodes linéaires employées pour le modèle original : reconstruire le mieux possible la position des bobines à partir d'un nombre limité de paramètres. Comme ces réseaux ont un très grand pouvoir d'expression, il leur est possible de s'ajuster plus efficacement que le modèle linéaire aux données d'entraînement. De ce fait, même si ces réseaux sont entraînés avec des méthodes classiques limitant leur surajustement aux données (*overfitting*), rien ne les empêche de converger vers des trajectoires très « chahutées » qui nous semblent difficilement interprétables, et de nature à faire perdre les intérêts du modèle initial. Nous avons donc décidé de limiter le pouvoir d'expression de ces réseaux à la découverte de paramètres fournissant des trajectoires « peu sinueuses », selon un critère de courbure de signe autant que possible constant : soit toujours convexes, soit toujours concaves (la figure 3.3 présente deux exemples de trajectoires, l'une respectant cette contrainte et l'autre ne la respectant pas). Nous formalisons cette contrainte dans la sous-section suivante.

3.1.2.1 Interprétabilité

Pour évaluer l'écart d'une trajectoire au critère proposé ci-dessus (courbes de sens de courbure unique), nous évaluons en chaque point de la trajectoire d'une bobine pilotée par un paramètre articulatoire donné, la courbure, comme :

$$c(a) = \frac{\ddot{x}(a)\dot{y}(a) - \dot{x}(a)\ddot{y}(a)}{\left(\dot{x}(a)^2 + \dot{y}(a)^2\right)^{3/2}},$$

avec :

— a, valeur du paramètre articulatoire ;
— $\dot{x}(a)$ et $\ddot{x}(a)$, dérivées première et seconde de la fonction $x(a)$ donnant la position en x de la bobine pour a ;
— $\dot{y}(a)$ et $\ddot{y}(a)$, dérivées première et seconde de la fonction $y(a)$ donnant la position en y de la bobine pour a.

Cette mesure donne une valeur nulle pour une trajectoire rectiligne, ou bien une valeur positive ou négative en fonction de l'orientation de la courbure. Cette valeur est d'autant plus grande que la courbure est importante. Cette mesure appliquée à deux exemples de trajectoire est disponible figure 3.3.

Suivant un seuil s donné positif et proche de zéro, une trajectoire en un point p est dite convexe si $c(p) \leq s$ et concave si $c(p) \geq -s$. Pour une trajectoire, la proportion de points se situant à des endroits dits concaves est notée P_v et la proportion de points se situant à des endroits dits convexes, P_x. La régularité de cette trajectoire est définie par $\max\{P_v, P_x\}$: plus cette valeur est grande, moins la courbe est sinueuse (voir figure 3.3).

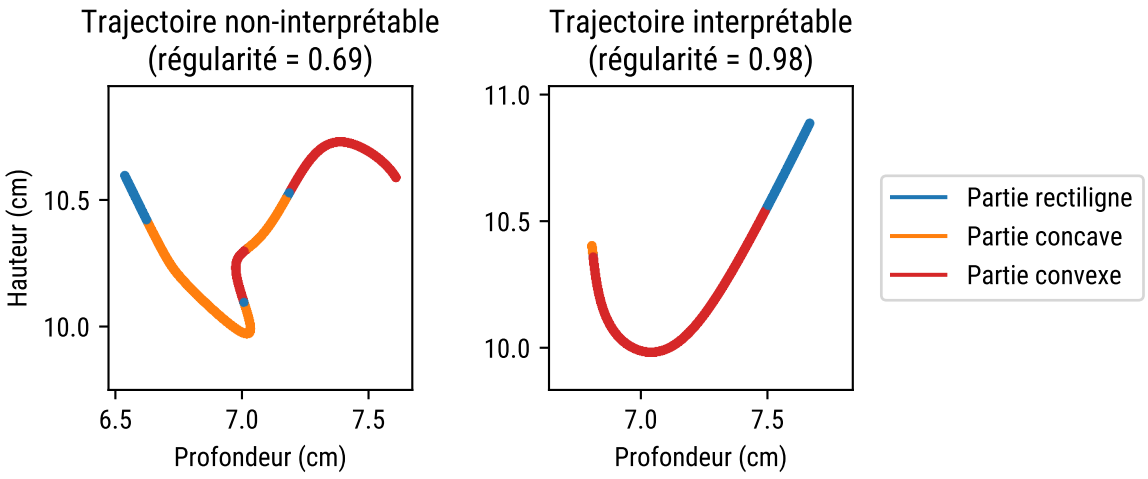

FIGURE 3.3 – **Exemple de trajectoire respectivement non-interprétable (à gauche) et interprétable (à droite)**

3.1.2.2 Évaluation

Méthode Pour évaluer la méthode de construction de modèles articulatoires non-linéaires, plusieurs modèles sont entraînés et évalués.

Afin d'obtenir à chaque étape de réduction de dimensionnalité ou de régression un réseau dont les trajectoires sont régulières, plusieurs réseaux avec seulement une couche cachée sont entraînés, et celui satisfaisant le critère de régularité présenté ci-dessous et présentant la plus faible erreur de reconstruction sur un ensemble de validation est retenu. Ainsi, pour chaque relation apprise entre un paramètre articulatoire et des positions de bobine, 3 réseaux sont entraînés pour chaque nombre de neurones cachés possible compris entre 2 et 7, soit un total de

18 réseaux. Le réseau retenu est celui présentant la plus faible erreur de reconstruction parmi ceux ayant une trajectoire dont la mesure de régularité est supérieure à 0,9 (avec $s = 0{,}1$).

Suivant cette procédure, 3 modèles articulatoires non-linéaires complets ont été construits. Chacun d'eux a été entraîné sur une partition aléatoire du corpus PB2007 comprenant 80 % des enregistrements, et testé sur les 20 % restant. 20 % des données d'entraînement ont servi pour l'*early stopping* et pour mesurer la régularité des trajectoires des paramètres articulatoires. La fonction de coût utilisée est l'erreur quadratique moyenne et la fonction d'activation est la tangente hyperbolique.

La prochaine sous-partie présente les résultats relatifs au processus de création d'un de ces modèles, tandis que la partie suivante compare les résultats de ces 3 modèles, chacun avec un modèle linéaire entraîné et évalué sur les mêmes données.

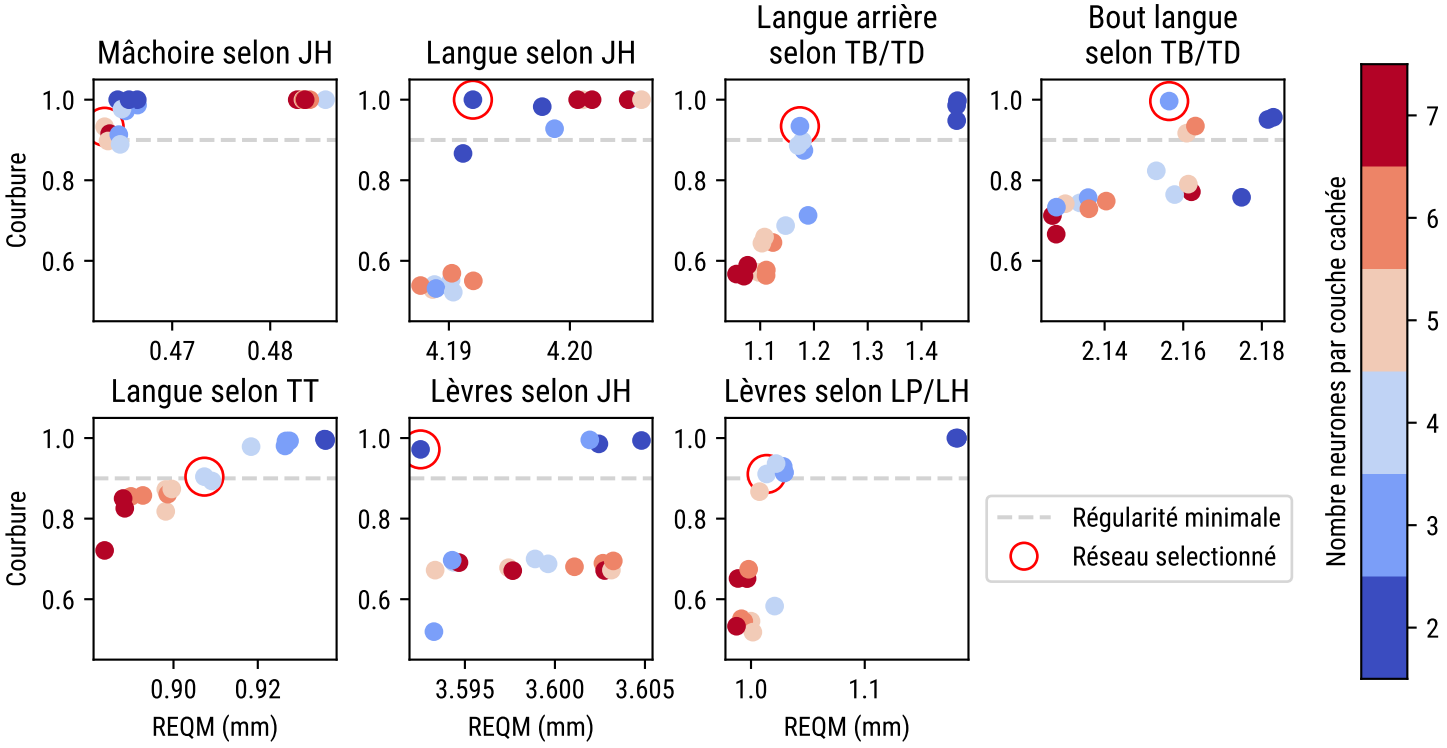

FIGURE 3.4 – **Erreur de reconstruction et courbure pour chacun des réseaux entraînés** Chaque point indique le résultat d'un des réseaux entraînés à établir la relation entre un paramètre articulatoire et la position des bobines EMA associées. La ligne pointillée (régularité minimale, fixée à 0,9) correspond à la proportion minimale de points de même courbure requise pour que les trajectoires extraites par un réseau soient considérées interprétables.

Évaluation qualitative Deux tendances semblent se dégager des différents architectures neuronales testées. Premièrement, et de façon attendue, plus le nombre de neurones dans la couche cachée est important, moins l'erreur de reconstruction est grande. Deuxièmement, l'augmentation de ce nombre permet au réseau de s'écarter davantage d'une trajectoire régu-

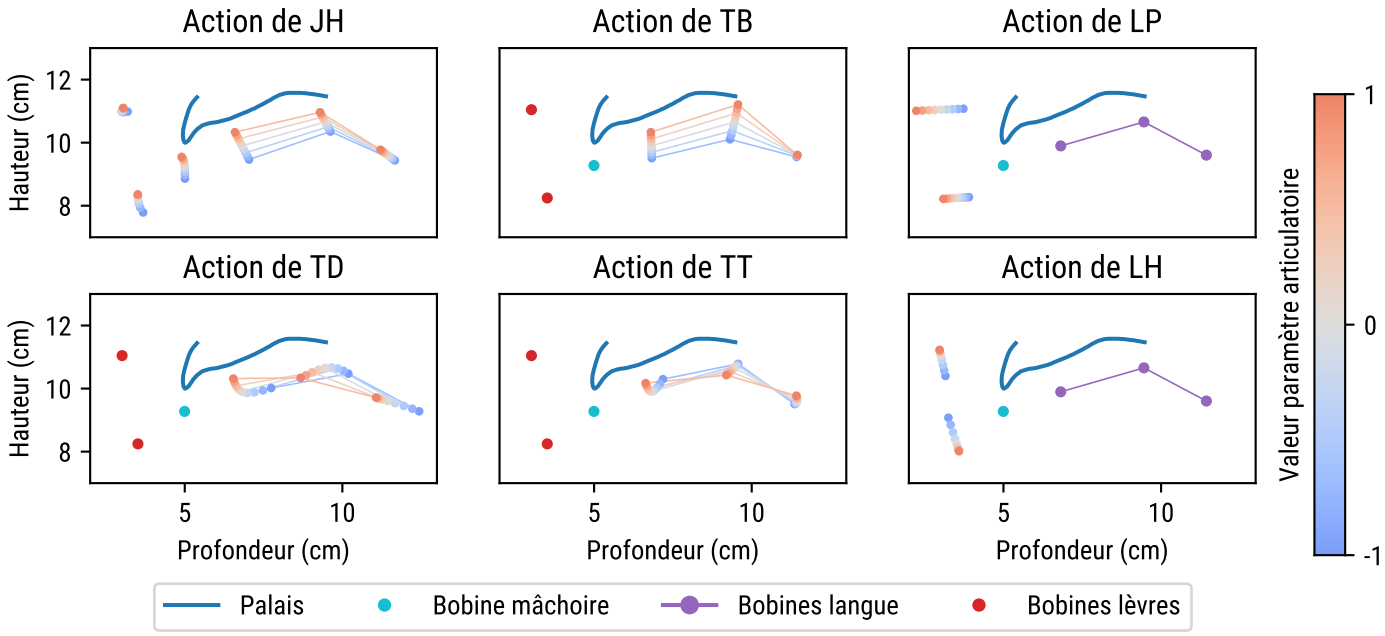

FIGURE 3.5 – **Trajectoires des bobines induites par la variation de chacun des paramètres articulatoires du modèle non-linéaire littéral construit avec le corpus PB2007** Les points fixes représentent la position moyenne des bobines.

lière. La figure 3.4 montre la valeur de ces mesures pour les différentes relations apprises lors de la création d'un modèle non-linéaire. La figure 3.5 présente les trajectoires induites par les paramètres articulatoires du modèle non-linéaire sélectionné.

Des pics de régularité avec une valeur 1 sont observés, ils se produisent lorsque le réseau concerné est « tombé » dans le minimum local d'une solution consistant en une trajectoire parfaitement rectiligne.

Comparaison au modèle linéaire Pour chacun des 3 modèles non-linéaires, un modèle linéaire a été entraîné sur les mêmes données d'entraînement et évalué sur les mêmes données de test à des fin de comparaison. Les résultats de cette comparaison sont visibles sur la figure 3.6.

À ce stade, le résultat de ce développement de modèle non linéaire est décevant. En effet, on peut remarquer que l'erreur de reconstruction globale est meilleure avec le modèle linéaire. Cependant, lorsqu'on regarde le détail de la reconstruction, on peut également remarquer que les performances de reconstruction de la part du modèle non-linéaire sont meilleures pour la reconstruction des coordonnées de la bobine de mâchoire ainsi que de celles de langue et de lèvres calculées seulement à partir de *JawHeight* (avant l'ajout de la contribution des paramètres *TongueBody*, *TongueDorsum* et *TongueTip* pour la langue, et *LipProtrusion* et *LipHeight* pour les lèvres). Le point commun de ces reconstructions est qu'elles ne font intervenir, dans le modèle non-linéaire, qu'un seul réseau de neurones pour leur calcul. En revanche, les autres reconstructions sont, elles, calculées à l'aide de plusieurs réseaux de neurones (par exemple, la bobine de langue arrière qui est calculée comme la somme de la sortie de deux réseaux de

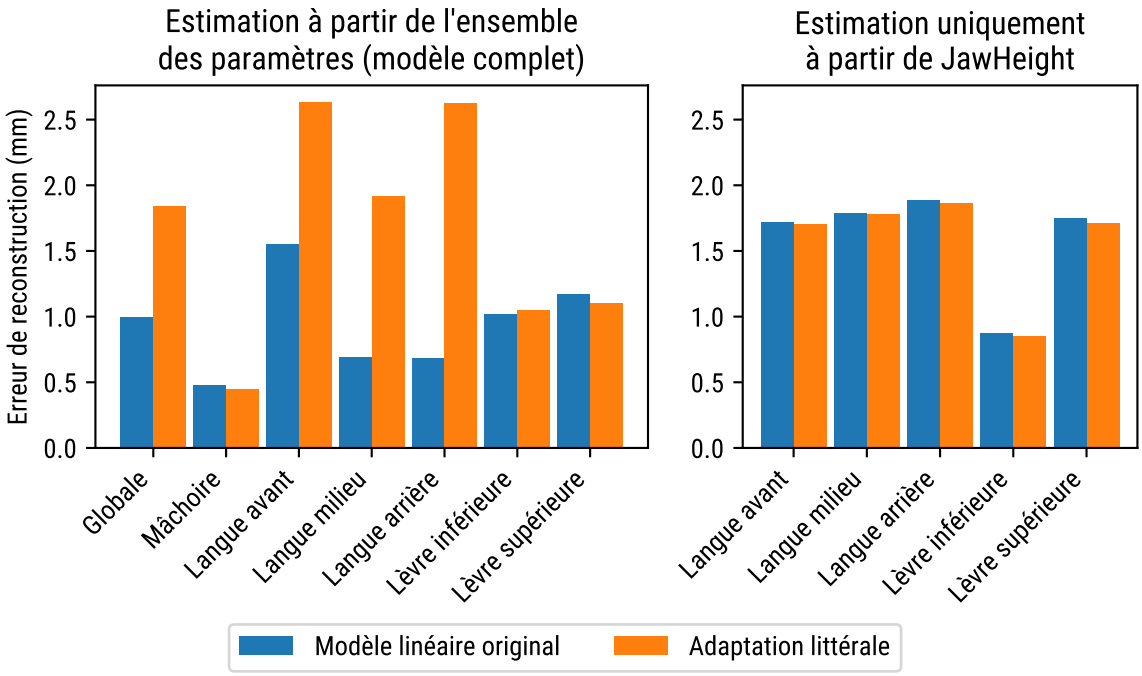

FIGURE 3.6 – **Erreur de reconstruction de la version linéaire vs. non-linéaire littérale pour le corpus PB2007** L'erreur globale correspond à l'erreur calculée sur l'ensemble des bobines. Chaque erreur est la moyenne des erreurs de 3 modèles entraînés indépendamment.

neurones, l'un se basant sur *JawHeight* et l'autre sur *TongueBody* et *TongueDorsum*). Cette différence de résultat suggère une mauvaise synergie entre les réseaux, qui pourrait provenir du fait qu'ils sont entraînés séparément. De ce fait, une approche bout à bout intégrant un entraînement conjoint de tous les réseaux œuvrant à l'extraction des paramètres articulatoires pourrait mener à une meilleur synergie et donc à une erreur de reconstruction globale plus faible. C'est ce que nous discutons dans la section suivante.

3.1.3 Adaptation non-linéaire *end-to-end* du modèle articulatoire

Nous proposons ici une architecture neuronale *end-to-end* permettant l'extraction des paramètres articulatoires par un unique réseau au cours d'un seul entraînement. Le travail présenté dans cette sous-section a une visée exploratoire et l'architecture proposée est focalisée sur la mâchoire et la langue, l'intégration des lèvres n'est pas traitée.

3.1.3.1 Définition du modèle

Le modèle articulatoire non-linéaire *end-to-end*, illustré figure 3.7, est constitué d'une succession d'auto-encodeurs et de régresseurs non-linéaires remplaçant la séquence de PCA et de régresseurs linéaires décrite dans la section 3.1.1.2. Ces sous-réseaux sont assemblés de façon à ne former qu'un seul réseau de neurones, prenant en entrée la position observée $\mathbf{e}$ des bobines

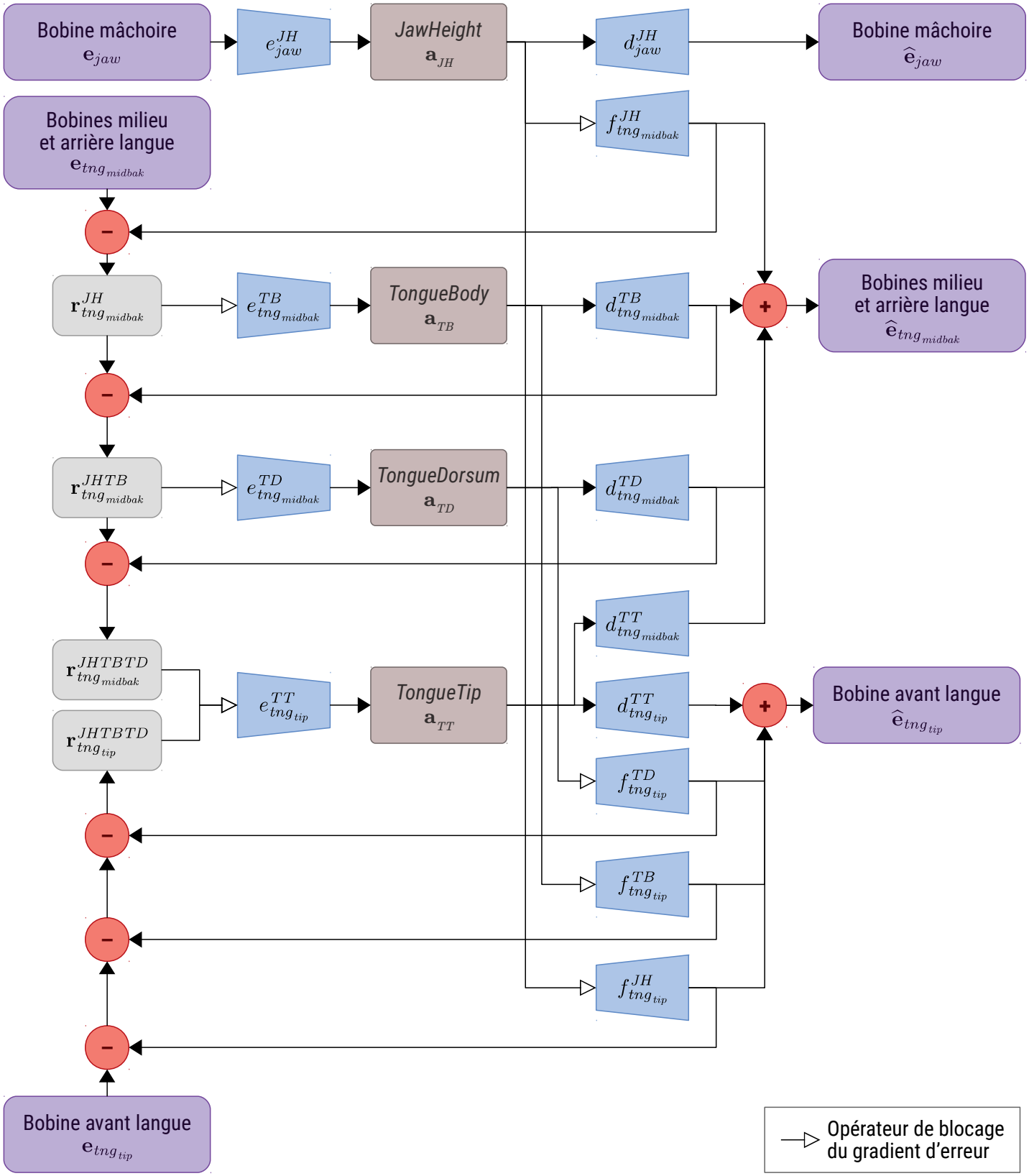

FIGURE 3.7 – **Architecture du modèle articulatoire *end-to-end*** Le modèle peut être vu comme un regroupement d'auto-encodeurs, chargé de reconstruire les coordonnées des bobines EMA et dont les espaces latents de chacun représentent les paramètres articulatoires.

EMA de mâchoire et de langue, et donnant en sortie sa reconstruction $\widehat{\mathbf{e}}$. Cet ensemble est entraîné conjointement afin de minimiser l'erreur de reconstruction $\|\mathbf{e}-\widehat{\mathbf{e}}\|^2$. Les espaces latents de chacun des auto-encodeurs n'ont qu'une seule dimension et représentent respectivement les paramètres articulatoires *JawHeight*, *TongueBody*, *TongueDorsum* et *TongueTip*. Comme nous le verrons, les connexions entre ces auto-encodeurs sont agencées de façon à proposer une extraction itérative des paramètres articulatoires reprenant directement la hiérarchie proposée dans la méthode linéaire originale.

Afin de pouvoir extraire de façon ciblée chacun des paramètres articulatoires, le vecteur de coordonnées EMA observées $\mathbf{e}$ est décomposé de la façon suivante :

$$\mathbf{e} = \mathbf{e}_{jaw} \oplus \mathbf{e}_{tng_{midbak}} \oplus \mathbf{e}_{tng_{tip}},$$

avec $\mathbf{e}_{jaw}$ vecteur contenant les coordonnées en x et y des bobines attachées à la mâchoire, $\mathbf{e}_{tng_{midbak}}$ vecteur contenant celles des bobines attachées sur le milieu (*middle*) et l'arrière (*back*) de la langue, $\mathbf{e}_{tng_{tip}}$ vecteur contenant celles des bobines attachées sur l'apex de la langue et $\oplus$ opérateur de concaténation de ces vecteurs. Pour les mêmes raisons, le vecteur contenant la reconstruction $\widehat{\mathbf{e}}$ de la position observée $\mathbf{e}$ est décomposé de la même manière. Nous allons maintenant présenter en détail comment chacune des parties de cette décomposition est reconstruite par le modèle articulatoire *end-to-end*, tout en veillant à faire le parallèle avec la méthode linéaire originale qui a guidé les choix de conception de ce nouveau modèle.

Reconstruction des bobines de mâchoire Comme pour le modèle articulatoire linéaire original, la reconstruction $\widehat{\mathbf{e}}_{jaw}$ de la position observée $\mathbf{e}_{jaw}$ des bobines de mâchoire se fait uniquement à partir du paramètre articulatoire *JawHeight*. Cette relation est représentée par un auto-encodeur :

$$e^{JH}_{jaw}(\mathbf{e}_{jaw}) = \mathbf{a}_{JH}, \qquad\qquad d^{JH}_{jaw}(\mathbf{a}_{JH}) = \widehat{\mathbf{e}}_{jaw},$$

avec e^{JH}_{jaw} encodeur permettant de projeter $\mathbf{e}_{jaw}$ vers l'espace latent $\mathbf{a}_{JH}$, qui représente le paramètre articulatoire *JawHeight*, et d^{JH}_{jaw} décodeur donnant la reconstruction $\widehat{\mathbf{e}}_{jaw}$.

Reconstruction des bobines de milieu et d'arrière de la langue Comme dans le modèle articulatoire linéaire original, la reconstruction $\widehat{\mathbf{e}}_{tng_{midbak}}$ de la position des bobines de milieu et d'arrière de langue $\mathbf{e}_{tng_{midbak}}$ se calcule comme la somme des contributions des paramètres articulatoires *JawHeight*, *TongueBody*, *TongueDorsum* et *TongueTip* :

$$\widehat{\mathbf{e}}_{tng_{midbak}} = f^{JH}_{tng_{midbak}}(\mathrm{sg}[\mathbf{a}_{JH}]) + d^{TB}_{tng_{midbak}}(\mathbf{a}_{TB}) + d^{TD}_{tng_{midbak}}(\mathbf{a}_{TD}) + d^{TT}_{tng_{midbak}}(\mathbf{a}_{TT}),$$

avec $f^{JH}_{tng_{midbak}}$ régresseur non-linéaire, et $d^{TB}_{tng_{midbak}}$, $d^{TD}_{tng_{midbak}}$ et $d^{TT}_{tng_{midbak}}$, décodeurs des auto-encodeurs respectivement en charge de l'estimation de $\mathbf{a}_{TB}$, $\mathbf{a}_{TD}$ et $\mathbf{a}_{TT}$. L'opérateur sg, pour *stop gradient*, est un opérateur de blocage du gradient d'erreur et est défini comme une fonction identité ayant une dérivée nulle, contraignant son opérande à rester inchangée lors de la rétropropagation du gradient d'erreur.

Avant de détailler l'implémentation des auto-encodeurs estimant les paramètres *TongueBody*, *TongueDorsum* et *TongueTip*, attardons-nous sur le terme $f^{JH}_{tng_{midbak}}(\text{sg}[\mathbf{a}_{JH}])$ de l'équation précédente. Selon le modèle articulatoire linéaire original, l'estimation de *JawHeight* doit se faire exclusivement à partir de la bobine de mâchoire. Ce paramètre est ensuite utilisé *a posteriori* pour estimer la position des bobines de langue via une régression linéaire. Cette façon de procéder est reproduite ici via l'utilisation de l'opérateur sg. Cet opérateur assure que, lors de la rétropropagation du gradient d'erreur, les poids de e^{JH}_{jaw}, encodeur en charge de l'estimation de $\mathbf{a}_{JH}$, ne soient pas ajustés pour donner une meilleure reconstruction $\widehat{\mathbf{e}}_{tng_{midbak}}$. De cette façon, le processus d'estimation de $\mathbf{a}_{JH}$, que nous avons vu dans la sous-section précédente, reste concentré uniquement sur la reconstruction de la bobine de mâchoire. En revanche, les poids du régresseur non-linéaire $f^{JH}_{tng_{midbak}}$ seront eux ajustés afin d'estimer le mieux possible les positions des bobines milieu et arrière de langue à partir de *JawHeight*, et donc la contribution de ce paramètre à ces positions.

Maintenant que nous avons défini les décodeurs des auto-encodeurs en charge d'estimer $\mathbf{a}_{TB}$ et $\mathbf{a}_{TD}$, concentrons-nous sur leurs encodeurs (l'estimation de $\mathbf{a}_{TT}$ sera traitée dans la sous-section suivante). Dans le modèle articulatoire linéaire, *TongueBody* est estimé à partir d'un mouvement résiduel calculé par soustraction de l'estimation par *JawHeight* à la position observée des bobines milieu et arrière de langue. Autrement dit, l'estimation de *TongueBody* se fait à partir de « la partie des positions des bobines de langue que *JawHeight* n'a pas permis d'estimer ». Dans cette implémentation, ce mouvement résiduel est calculé avec :

$$\mathbf{r}^{JH}_{tng_{midbak}} = \mathbf{e}_{tng_{midbak}} - f^{JH}_{tng_{midbak}}(\text{sg}[\mathbf{a}_{JH}])\,.$$

Ce mouvement résiduel est ensuite envoyé dans l'encodeur $e^{TB}_{tng_{midbak}}$ en charge de l'estimation de *TongueBody* :

$$e^{TB}_{tng_{midbak}}(\text{sg}[\mathbf{r}^{JH}_{tng_{midbak}}]) = \mathbf{a}_{TB}.$$

L'opérateur sg est employé ici afin que les paramètres de $d^{JH}_{tng_{midbak}}$ (qui intervient dans le calcul de $\mathbf{r}^{JH}_{tng_{midbak}}$) ne soient pas affectés par l'optimisation de $\mathbf{a}_{TB}$ lors de la rétropropagation du gradient de l'erreur de reconstruction globale. Le fait de fournir à l'auto-encodeur en charge de l'estimation de *TongueBody* un mouvement résiduel affranchi de la contribution de *JawHeight* constitue une façon de reproduire la hiérarchie entre ces deux paramètres présente dans le modèle articulatoire linéaire original.

De façon similaire, le calcul du paramètre articulatoire *TongueDorsum* se fait à partir du mouvement résiduel $\mathbf{r}^{JHTB}_{tng_{midbak}}$ issu de la soustraction de l'estimation de la position des bobines milieu et arrière de langue à partir de *TongueBody* et de *JawHeight* :

$$\mathbf{r}^{JHTB}_{tng_{midbak}} = \mathbf{r}^{JH}_{tng_{midbak}} - d^{TB}_{tng_{midbak}}(\mathbf{a}_{TB}).$$

Ce nouveau mouvement résiduel permet le calcul de *TongueDorsum*, via l'encodeur $e^{TD}_{tng_{midbak}}$ en charge de l'estimation de $\mathbf{a}_{TD}$:

$$e^{TD}_{tng_{midbak}}(\text{sg}[\mathbf{r}^{JHTB}_{tng_{midbak}}]) = \mathbf{a}_{TD}.$$

À noter que les appellations « *TongueBody* » et « *TongueDorsum* », respectivement pour le premier et le second paramètre extraits dans cette sous-section, sont à ce stade arbitraires.

En effet, *TongueBody* est censé représenter un mouvement plutôt horizontal de la langue et *TongueDorsum* un mouvement de celle-ci plutôt vertical. Or, il se peut qu'en fonction du corpus utilisé, ou de l'initialisation des poids du modèle avant son entraînement, le premier paramètre extrait encode plutôt un mouvement vertical et le second un mouvement plutôt horizontal. Dans ce cas, il convient d'inverser le nom de ces deux paramètres *a posteriori*.

Reconstruction des bobines de l'apex Comme pour la reconstruction de la position des bobines de milieu et d'arrière de langue, la reconstruction $\widehat{\mathbf{e}}_{tng_{tip}}$ de la position des bobines de l'apex se fait à partir des paramètres *JawHeight*, *TongueBody*, *TongueDorsum* et *TongueTip* :

$$\widehat{\mathbf{e}}_{tng_{tip}} = f^{JH}_{tng_{tip}}(\text{sg}[\mathbf{a}_{JH}]) + f^{TB}_{tng_{tip}}(\text{sg}[\mathbf{a}_{TB}]) + f^{TD}_{tng_{tip}}(\text{sg}[\mathbf{a}_{TD}]) + d^{TT}_{tng_{tip}}(\mathbf{a}_{TT}),$$

avec $f^{JH}_{tng_{tip}}$, $f^{TB}_{tng_{tip}}$, $f^{TD}_{tng_{tip}}$ régresseurs non-linéaires et $d^{TT}_{tng_{tip}}$ décodeur, reconstruisant chacun une partie de la position des bobines de milieu et d'arrière de langue, respectivement à partir de $\mathbf{a}_{JH}$, $\mathbf{a}_{TB}$, $\mathbf{a}_{TD}$ et $\mathbf{a}_{TT}$ dénotant la valeur des paramètres articulatoires *JawHeight*, *TongueBody*, *TongueDorsum* et *TongueTip*.

L'estimation du paramètre articulatoire *TongueTip* se fait à partir de deux mouvements résiduels décrivant la position des bobines de langue « nettoyée » de la contribution de *JawHeight*, *TongueBody* et *TongueDorsum*. Le premier mouvement résiduel décrit la position des bobines milieu et arrière :

$$\mathbf{r}^{JHTBTD}_{tng_{midbak}} = \mathbf{r}^{JHTB}_{tng_{midbak}} - d^{TD}_{tng_{midbak}}(\mathbf{a}_{TD}),$$

et le second décrit la position des bobines de l'apex :

$$\mathbf{r}^{JHTBTD}_{tng_{tip}} = \mathbf{e}_{tng_{tip}} - f^{JH}_{tng_{tip}}(\text{sg}[\mathbf{a}_{JH}]) - f^{TB}_{tng_{tip}}(\text{sg}[\mathbf{a}_{TB}]) - f^{TD}_{tng_{tip}}(\text{sg}[\mathbf{a}_{TD}]),$$

Ces nouveaux mouvements résiduels sont ensuite fournis à l'encodeur e^{TT}_{tng} en charge de l'estimation de $\mathbf{a}_{TT}$:

$$e^{TT}_{tng}(\text{sg}[\mathbf{r}^{JHTBTD}_{tng_{midbak}}] \oplus \text{sg}[\mathbf{r}^{JHTBTD}_{tng_{tip}}]) = \mathbf{a}_{TT}.$$

3.1.3.2 Implémentation

Nous implémentons ici le modèle en utilisant les données du corpus PB2007 ainsi que celles du corpus BY2014. Ainsi, pour chacun de ces corpus, le vecteur $\mathbf{e}_{jaw}$ contient les coordonnées 2D d'une bobine collée à l'incisive inférieure, le vecteur $\mathbf{e}_{tng_{midbak}}$ contient la position de deux bobines collées respectivement au milieu et à l'arrière de la langue et $\mathbf{e}_{tng_{tip}}$ la position d'un bobine collée sur l'apex. Pour chacun des corpus, 80 % des données sont aléatoirement choisies pour l'entraînement du modèle (20 % d'entre elles sont réservées pour le contrôle par du surapprentissage par *early-stopping*) et les 20 % restantes sont utilisées comme données de test.

Chacun des sous-réseaux e, d et f est implémenté à l'aide de deux couches de 12 neurones pleinement connectées et dont la fonction d'activation est la tangente hyperbolique. Une procédure de *batch normalization* ainsi que de *dropout* ($p = 0{,}25$) sont appliquées à chacune de

ces couches cachées. Les paramètres du modèle complet sont ajustés par rétropropagation de l'erreur de reconstruction $\|\mathbf{e} - \widehat{\mathbf{e}}\|^2$ calculée sur des *mini-batches* de 32 observations issues du corpus d'entraînement.

3.1.3.3 Évaluation

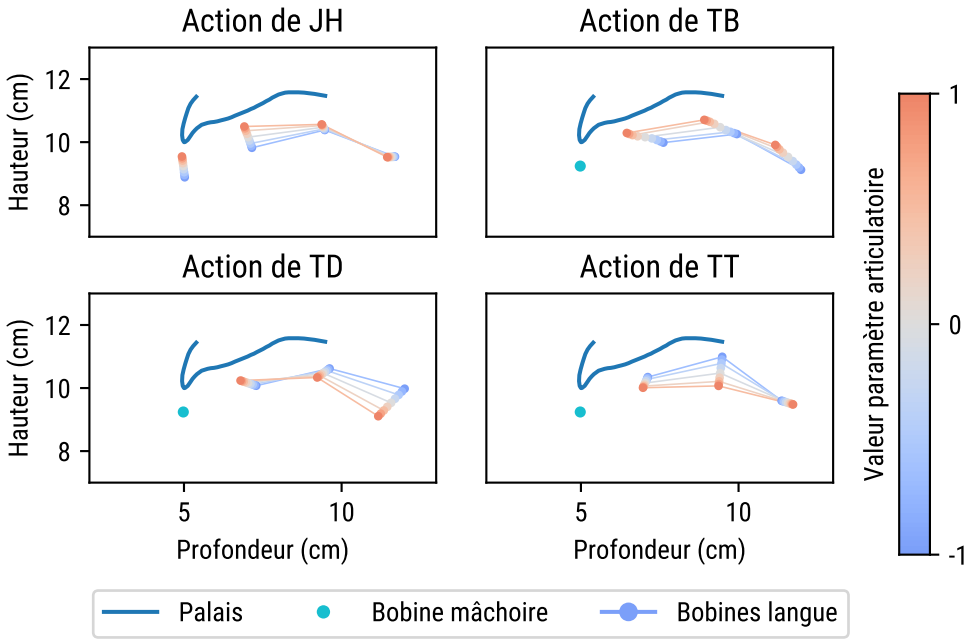

FIGURE 3.8 – **Trajectoires des bobines EMA induites par la variation de chacun des paramètres articulatoires du modèle non-linéaire *end-to-end* construit avec le corpus PB2007** Les points fixes représentent la position moyenne des bobines.

Évaluation qualitative La figure 3.8 montre les trajectoires tracées par le déplacement des bobines induit par la variation de la valeur des différents paramètres articulatoires. Tout d'abord, nous pouvons remarquer que, contrairement à l'adaptation non-linéaire littérale présentée dans la sous-section précédente, cette nouvelle adaptation non-linéaire *end-to-end* fournit des trajectoires interprétables directement et sans nécessiter de processus de sélection, probablement parce que l'estimation de l'ensemble des paramètres du modèle est ici simultanée et évite des sur-ajustements locaux. De plus, nous pouvons remarquer que les trajectoires présentées dans cet exemple sont très cohérentes par rapport à celles extraites par le modèle articulatoire linéaire (visibles sur la figure 3.2). En effet, la variation de *JawHeight* provoque un mouvement général vertical de chacune des bobines, la variation de *TongueBody* induit un mouvement d'avant en arrière des bobines de langue, la variation de *TongueDorsum* induit un mouvement vertical des bobines de langue et enfin la variation de *TongueTip* provoque un mouvement vertical des bobines de l'apex et du milieu de la langue ainsi qu'un mouvement plutôt horizontal de la bobine arrière. Ces mouvements se retrouvent également lors de la construction de modèles articulatoires *end-to-end* avec d'autres jeux de données (ces résultats ne sont pas présentés ici) mais ne sont parfois pas associés aux mêmes paramètres

articulatoires.

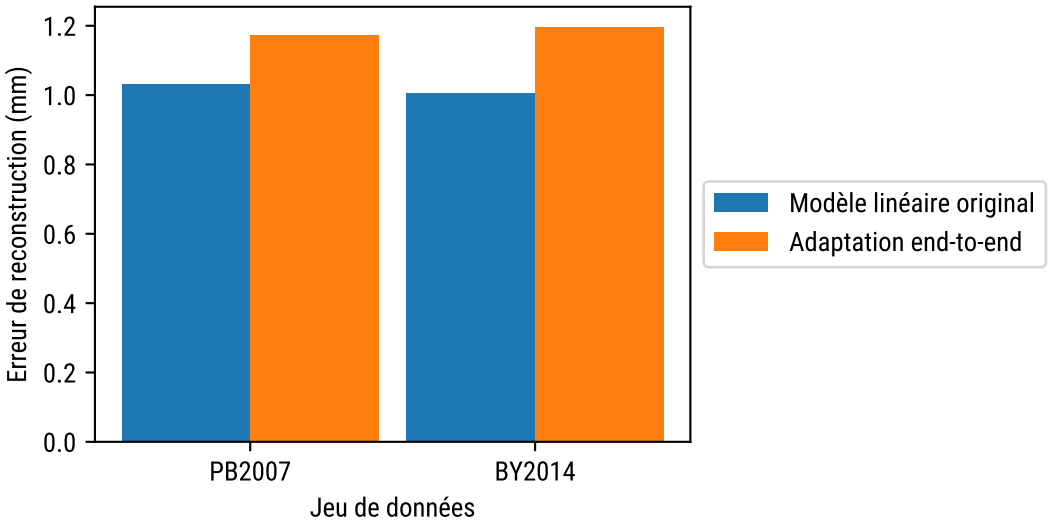

FIGURE 3.9 – **Erreur de reconstruction du modèle articulatoire *end-to-end* comparée à celle du modèle linéaire original** L'erreur de reconstruction correspond à la racine carrée de l'erreur quadratique moyenne (REQM) exprimée en millimètres entre la position de chacune des bobines reconstruites et leur position avant reconstruction.

Performance de reconstruction La figure 3.9 montre la racine de l'erreur quadratique moyenne (REQM) de reconstruction du modèle articulatoire non-linéaire *end-to-end* appliqué aux ensembles d'évaluation des corpus PB2007 et BY2014. Cette erreur est calculée en faisant la moyenne de l'erreur de 5 modèles initialisés et entraînés indépendamment, et sur des ensembles d'apprentissage et de tests différents. À des fins de comparaison, la figure montre également l'erreur obtenue avec le modèle linéaire (pour les mêmes configurations expérimentales).

Ces résultats s'avèrent décevants puisque l'erreur de reconstruction des modèles non-linéaires est supérieure pour les deux corpus (PB2007 et BY2014) à celle des modèles linéaires (autour de 1,2 mm contre environ 1 mm). Nous n'avons à ce jour pas d'explication pour interpréter cette relative mauvaise performance de l'approche non-linéaire *end-to-end* pour la construction d'un modèle articulatoire à partir de données EMA. D'autres expériences seront nécessaires pour comprendre l'origine précise de cette différence de performance. En effet, l'approche non-linéaire étant *a priori* capable d'aboutir à un ensemble de solutions contenant celle fournie par l'approche linéaire, nous pensons à ce stade à un problème non pas lié à l'approche elle-même, mais plutôt à son implémentation (choix de l'architecture de certains des sous-réseaux du modèle, sur-apprentissage ou encore mauvaise rétropropagation du gradient d'erreur).

3.1.3.4 Conclusions

Nous avons présenté une nouvelle méthode permettant une adaptation non-linéaire *end-to-end* du modèle articulatoire. Cette nouvelle méthode permet d'intégrer les principes introduits par la méthode originale au sein d'un seul et même réseau de neurones qui permet l'extraction simultanée de chacun des paramètres articulatoires.

Les résultats montrent que les trajectoires extraites par ce nouveau modèle sont interprétables et similaires à celles extraites par le modèle linéaire original. Néanmoins, les performances de reconstruction de ce nouveau modèle sont inférieures à celles du modèle linéaire. Malgré ce résultat décevant, cette nouvelle architecture nous semble tout de même prometteuse. En effet, à ce stade l'agencement des connexions au sein de celle-ci n'est sans doute pas optimal et une étude comparative les concernant pourrait potentiellement mener à une amélioration des performances.

3.1.4 Conclusion sur le modèle articulatoire

Nous avons présenté une méthode pour créer un modèle articulatoire à partir de données EMA. Ce modèle repose exclusivement sur des méthodes linéaires. De ce fait, nous avons présenté deux variantes non-linéaires, basées sur les réseaux neuronaux. Malheureusement, et malgré des développements prometteurs, ces deux tentatives ont donné des résultats mitigés. Nous utiliserons donc, dans la suite de ce manuscrit, la version linéaire originale du modèle articulatoire, présentée en section 3.1.1.

3.2 Modèle articulatoire-vers-acoustique

La seconde étape du processus de construction du synthétiseur articulatoire est la création d'un modèle articulatoire-vers-acoustique en charge d'estimer le contenu spectral associé à une configuration articulatoire. Nous implémentons ce modèle avec un réseau de neurones entraîné à faire la correspondance entre les enregistrements EMA, traduits sous forme de paramètres articulatoires (à l'aide du modèle articulatoire linéaire décrit précédemment), et les enregistrements acoustiques du locuteur de référence.

3.2.1 Implémentation

Similairement à Bocquelet et al., 2014, le modèle articulatoire-vers-acoustique est implémenté à l'aide d'un réseau de neurones *feedforward* contenant 4 couches cachées pleinement connectées de 256 neurones. La fonction d'activation tangente hyperbolique, une *batch normalization* ainsi qu'un *dropout* ($p = 0{,}25$) sont appliqués à chacune de ces couches. La sortie de la dernière couche cachée est dirigée vers la couche de sortie par une projection linéaire.

En entrée, le modèle reçoit un vecteur contenant 7 paramètres articulatoires (6 pour le corpus PB2007 qui, pour rappel, ne contient pas d'informations sur le velum), auxquels sont ajouté leurs dérivées première (vitesse) et seconde (accélération). Ces informations sur la dynamique visent à compenser le manque d'informations sur la géométrie du conduit vocal inhérent aux données EMA, qui ne décrivent la position que de quelques points de chair (contrairement à des données IRM, par exemple, qui fournissent le contour complet du conduit vocal). En sortie, le modèle prédit une observation acoustique contenant des caractéristiques spectrales sous la forme d'un vecteur de 18 coefficients cepstraux (en échelle Bark, calculés à l'aide d'une analyse par fenêtre glissante d'une taille de 20 ms et une *hop size* de 10 ms). Ce format est choisi pour être compatible avec LPCNet (Valin & Skoglund, 2019) qui est, comme nous le verrons dans la section suivante, le vocodeur neuronal qui se charge de la génération de la forme d'onde.

Pour l'entraînement du modèle, 80 % du corpus est utilisé pour l'apprentissage (avec un sous-ensemble de 20 % pour le contrôle du sur-apprentissage par *early-stopping*), les 20 % restants étant utilisés pour l'évaluation. La fonction de coût utilisée est l'erreur quadratique moyenne de reconstruction. Les poids du réseau sont mis à jour par descente de gradient stochastique à l'aide de l'algorithme d'optimisation Adam, sur des *mini-batches* de 32 trames.

3.2.2 Évaluation

Afin d'évaluer l'impact du format de représentation des données articulatoires sur les performances du modèle articulatoire-vers-acoustique, nous entraînons 3 variantes de celui-ci. La différence entre chacune de ces variantes du modèle est le format des données articulatoires fournies à son entrée. La première variante est celle décrite précédemment, prenant en entrée les paramètres articulatoires extraits des enregistrements EMA par le modèle articulatoire linéaire. La seconde prend en entrée les coordonnées EMA originales, et sert à évaluer la perte d'information induite par la traduction des coordonnées EMA en paramètres articulatoires. La troisième prend en entrée les valeurs des 7 premières composantes d'une PCA appliquée aux coordonnées EMA originales, et est utilisée pour évaluer la perte d'information, à nombre de paramètres équivalent, induite par le recours à des paramètres plus interprétables. Enfin, pour chacune de ces 3 variantes, nous entraînons des modèles sur les données articulatoires accompagnées ou non de leurs dérivées première et seconde, pour un total de 6 configurations expérimentales testées.

Pour chacune de ces configurations et pour chacun des corpus PB2007 et BY2014, nous entraînons 5 modèles articulatoire-vers-acoustique en utilisant une méthode de validation croisée à K blocs (*K-fold cross-validation*) avec $K = 5$. Cette répartition des données permet d'obtenir une resynthèse complète de chacun des corpus issue exclusivement d'ensembles d'évaluation.

La figure 3.10 présente la distribution de l'erreur de reconstruction des trames pour chacun des corpus et pour chacune des configurations considérées. Tout d'abord, nous pouvons remarquer que la distribution des erreurs est similaire pour chacun des corpus et pour chacune des conditions, avec une répartition principalement en dessous de 1,5. Nous évaluerons l'im-

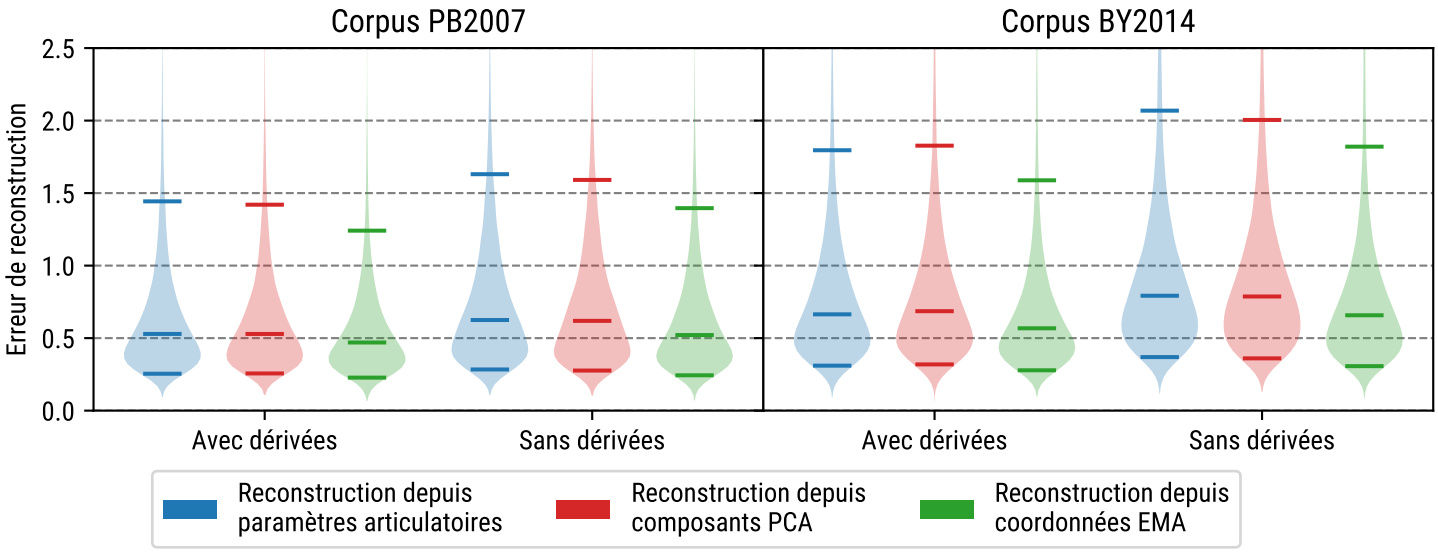

FIGURE 3.10 – **Distribution de l'erreur de reconstruction du modèle articulatoire-vers-acoustique pour les corpus PB2007 et BY2014 en fonction du type d'entrée articulatoire** Les types d'entrées considérés sont les paramètres articulatoires obtenus avec le modèle articulatoire, les données EMA brutes et les données EMA réduites par PCA. Pour une trame donnée l'erreur de reconstruction est la moyenne quadratique des erreurs sur toutes les dimensions du cepstre. Comme les dimensions du cepstre sont centrées réduites, cette erreur est sans unité. Les barres horizontales représentent de bas en haut le 5^{e} centile, la médiane et le 95^{e} centile de chacune des distributions.

pact perceptif de ces erreurs dans la section suivante. Nous pouvons également constater que les distributions des erreurs des modèles auxquels les dérivées des données articulatoires sont fournies sont davantage concentrées autour de valeurs plus faibles que celles des modèles qui en sont dénués, suggérant un apport bénéfique de celles-ci sur les performances du modèle. Ce constat se retrouve également au niveau des médianes des distributions, qui sont globalement plus basses pour les modèles avec dérivées. Ensuite, lorsque nous nous intéressons aux type de données articulatoires fournies au modèle, nous remarquons que les enregistrements EMA donnent lieu à une erreur de reconstruction globalement plus basse que celles obtenues lors de l'utilisation des paramètres articulatoires ou des composantes de la PCA classique. Ce résultat suggère que ces deux techniques de réduction de dimension n'ont pas permis de préserver toute l'information articulatoire contenue dans les enregistrements EMA originaux. Nous pouvons néanmoins remarquer que les distributions des erreurs de reconstruction à partir des paramètres articulatoires et des composantes de la PCA sont très similaires et que leurs médianes sont situées sensiblement aux mêmes niveaux. Ceci suggère que le gain en interprétabilité obtenu avec la méthode d'extraction des paramètres articulatoires ne s'est pas fait au prix d'une perte d'information pénalisante par rapport à une PCA classique.

3.3 Génération de la forme d'onde

Cette section présente la dernière étape de construction du synthétiseur articulatoire : l'ajout du vocodeur neuronal LPCNet pour lui permettre la génération de la forme d'onde à partir de la sortie du modèle articulatoire-vers-acoustique décrit précédemment. La suite de la section présente LPCNet et son entraînement, et se termine avec une évaluation objective et subjective du synthétiseur proposé.

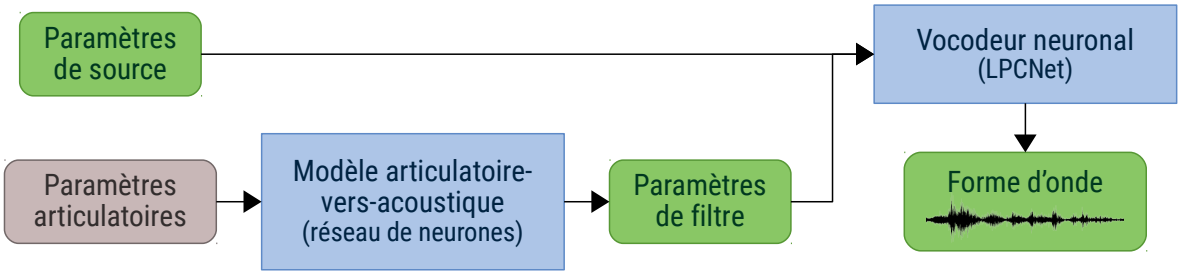

FIGURE 3.11 – **Vue d'ensemble du synthétiseur articulatoire**

3.3.1 Implémentation

3.3.1.1 Vocodeur neuronal LPCNet

LPCNet est un vocodeur neuronal que nous avons présenté en fin de section 1.2.5. Pour rappel, LPCNet prend en entrée deux types de paramètres, des paramètres décrivant le filtre et des paramètres décrivant la source, et est capable de générer à partir de ceux-ci une forme d'onde (signal audio de parole). Les paramètres de filtre sont 18 coefficients cepstraux exprimés en Bark (les mêmes que ceux fournis en sortie du modèle articulatoire-vers-acoustique décrit dans la section précédente) et les paramètres de source sont deux valeurs numériques décrivant respectivement la fréquence fondamentale du son et son degré de périodicité. Ce dernier peut être interprété dans une certaine mesure comme un niveau de voisement.

Cette distinction entre les paramètres de filtre et de source rend LPCNet particulièrement adapté pour être utilisé dans une tâche de synthèse articulatoire. En effet, les propriétés de filtrage du conduit vocal sont principalement conditionnées par la géométrie de celui-ci. Il est ainsi possible d'estimer les paramètres de filtre associés à une forme de conduit avec un modèle séparé, tel que le modèle articulatoire-vers-acoustique décrit dans la section précédente, et de brancher cet autre modèle directement sur l'entrée des paramètres de filtre de LPCNet. Le résultat de cette combinaison, illustré figure 3.11, donne un synthétiseur articulatoire, prenant en entrée des paramètres articulatoires et des paramètres de source, et donnant en sortie un signal audio.

3.3.1.2 Entraînement de LPCNet

Les jeux de données utilisés dans ce manuscrit ne contiennent pas une quantité d'enregistrements de parole suffisante pour entraîner depuis zéro et de façon convenable un vocodeur neuronal tel que LPCNet. De ce fait, nous utilisons une version de LPCNet fournie par ses auteurs, pré-entraînée sur un autre corpus de données et que nous adaptons aux corpus utilisés ici avec un *fine-tuning*. Cette version déjà entraînée l'a été sur le corpus *NTT Multi-Lingual Speech Database for Telephonometry*, contenant des productions en 21 langues différentes, durant 120 *epochs*. À partir de celle-ci, nous avons créé deux nouvelles versions adaptées respectivement aux corpus PB2007 et BY2014 en utilisant la procédure de *fine-tuning* fournie avec le code source de LPCNet [2] durant 50 *epochs*.

3.3.2 Évaluation

Cette sous-section s'intéresse à l'évaluation du synthétiseur articulatoire complet, comprenant le modèle articulatoire-vers-acoustique combiné au vocodeur neuronal LPCNet. Deux corpus sont considérés ici : PB2007 et BY2014. Pour chacun d'eux, nous réutilisons les modèles articulatoire-vers-acoustique entraînés dans la section 3.2. Les paramètres de source sont eux directement extraits des sons originaux et sont adjoints aux paramètres de filtre reconstruits. L'ensemble formé par ces deux types de paramètres est envoyé dans LPCNet afin d'obtenir une resynthèse complète de chacun des corpus.

3.3.2.1 Évaluation objective

L'évaluation objective du synthétiseur articulatoire est menée à l'aide de deux décodeurs phonétiques, l'un pour évaluer les resynthèses du corpus PB2007 et l'autre pour évaluer celles du corpus BY2014. Ces deux décodeurs partagent la même architecture, basée sur un modèle de Markov caché, et sont entraînés sur le contenu spectral de l'entièreté du corpus dont ils sont chargés d'évaluer les resynthèses. L'architecture employée ne comporte pas de modèle de langage (contrairement aux modèles plus traditionnels employés en reconnaissance de la parole), fournissant ainsi une évaluation « pure » (c'est-à-dire sans *a priori* linguistique) du contenu phonémique des resynthèses. Chacun de ces décodeur a été implémenté et entraîné à l'aide de l'outil *Hidden Markov Model Toolkit* (HTK, https://htk.eng.cam.ac.uk/).

La figure 3.12 présente les matrices de confusion fournies par ce décodeur acoustico-phonétique, en considérant une resynthèse pilotée par les paramètres articulatoires obtenus par le modèle articulatoire ainsi que leur version complétée de leurs dérivées première et seconde, pour le corpus PB2007. En observant les confusions sur les voyelles, nous remarquons que leur synthèse est globalement très bonne, avec un gain important apporté par l'utilisation des dérivées premières et secondes (précision moyenne de 78,79 % sans vs. 87,80 % avec). Nous pouvons cependant remarquer quelques confusions lors de la reconnaissance de /u/ qui est

2. Le code source de LPCNet est fourni par ses auteurs à l'adresse : https://github.com/xiph/LPCNet.

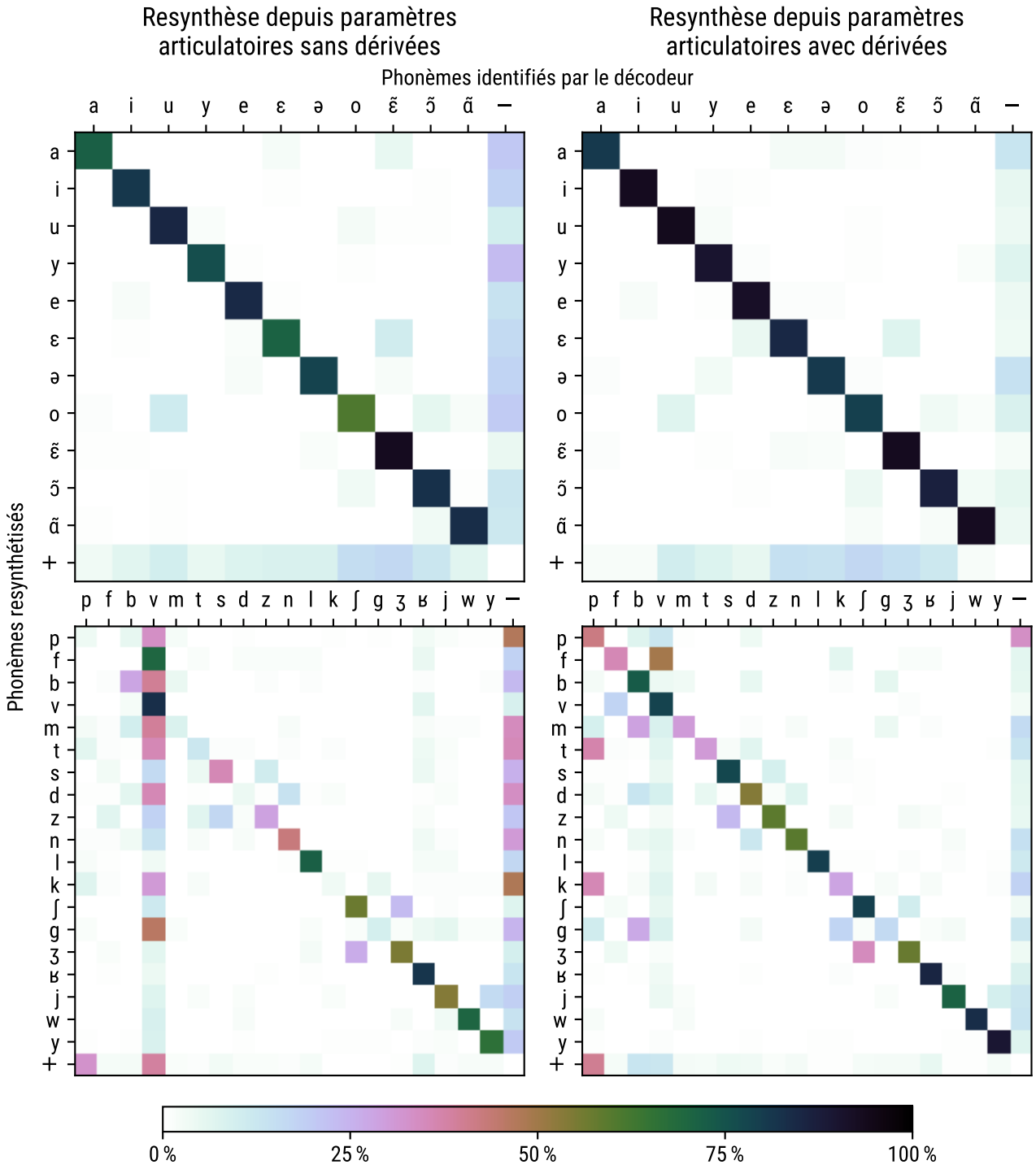

FIGURE 3.12 – **Résultats de la reconnaissance de phonèmes du décodeur HMM conduit sur les resynthèses du corpus PB2007** Les lignes « + » et les colonnes « – » représentent respectivement les insertions et les non-détections de phonèmes de la part du décodeur.

parfois pris pour un /o/ et lors de celle de /ɛ/, qui est parfois confondu avec son homologue nasalisé /ɛ̃/. Les matrices de confusion portant sur les consonnes montrent quant à elles un nombre plus important de confusions, principalement dans le cas du modèle sans dérivées. En effet, nous pouvons remarquer sur les resynthèses sans dérivées des erreurs presque systématiques sur les consonnes /p/, /f/, /m/, /t/, /d/, /k/ et /g/, qui sont majoritairement confondues avec la consonne /v/ ou simplement non détectées. Ces problèmes sont néanmoins limités par l'ajout des dérivées aux paramètres articulatoires (précision moyenne de 37,60 % sans les dérivées vs. 59,37 % avec). Cette différence peut s'expliquer par la nature dynamique du processus de production des consonnes qui requiert des mouvements rapides provoquant des constrictions en des points précis du conduit vocal et qui est probablement mieux décrit par les paramètres articulatoires complétés de leur vitesse et de leur accélération.

3.3.2.2 Évaluation subjective

En plus de l'évaluation objective, nous avons mené deux évaluations subjectives du synthétiseur à l'aide de tests perceptifs. La première de ces évaluations porte sur la qualité générale des resynthèses et la seconde porte sur leur contenu phonémique. Ces deux évaluations, qui sont décrites plus en détail dans les paragraphes suivants, ont d'abord été divisées en deux puis combinées entre elles de façon à former deux « parcours » de test composés de questions en provenance des deux évaluations. Ces deux parcours ont été construits avec l'outil *Web Audio Evaluation Tool* (Jillings et al., 2015) et ont été suivis respectivement par deux ensembles différents de 20 et 21 locuteurs francophones natifs recrutés à l'aide de la plate-forme *Prolific* (Palan & Schitter, 2018).

Qualité générale La qualité générale des resynthèses a été évaluée à l'aide de tests MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*) dont la méthodologie (ITU, 2015) permet d'évaluer la qualité perçue de sons reconstruits à l'aide de différentes méthodes par rapport à leur version originale. Un test MUSHRA est constitué d'une série de questions. Dans chaque question, un son est présenté au sujet comme étant un « son de référence ». Ce son est accompagné d'autres sons, qui sont des reconstructions de celui-ci, et le sujet doit fournir pour chacun de ces sons un score de similarité avec le son de référence (sur une échelle sans unité allant de « très similaire » à « très différent »). Afin d'assurer une calibration implicite de l'échelle de notation qu'emploie le sujet, la version originale du son de référence, appelée ultérieurement « ancre haute » ou « référence cachée », ainsi qu'une version volontairement très dégradée de celui-ci, qualifiée d'« ancre basse », sont introduites parmi les reconstructions du son de référence que le sujet doit noter. L'ajout du son original a pour but d'inciter le sujet à tenir compte des altérations mineures présentes dans les autres variations et l'ajout de l'ancre basse a pour but de limiter une notation trop basse des altérations modérées.

Suivant cette méthodologie, nous avons construit deux tests, le premier portant sur des resynthèses du corpus PB2007 et incorporé au premier parcours et le second portant sur le corpus BY2014 et incorporé au second parcours. Chaque test contient 30 ensembles de sons à évaluer. Pour chacun de ces ensembles, le son de référence est une phrase issue du corpus

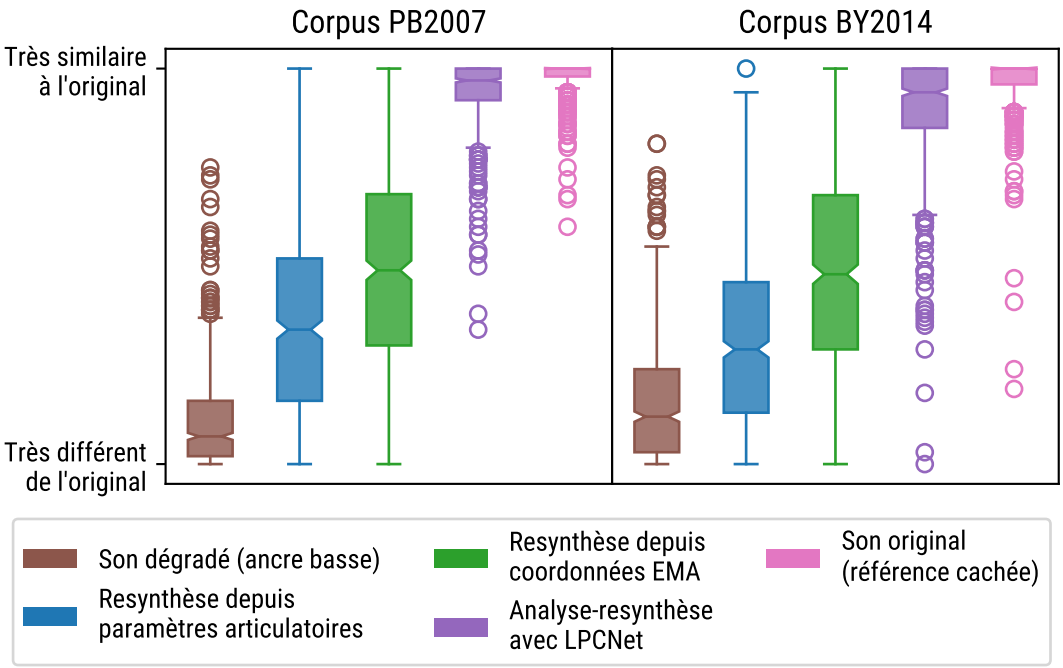

FIGURE 3.13 – **Résultats des tests MUSHRA conduits sur les resynthèses des corpus PB2007 et BY2014**

évalué. Le participant doit le comparer avec 5 autres sons, qui sont :

- S1 : resynthèse directe de la phrase par analyse-resynthèse avec LPCNet ;
- S2 : resynthèse à partir des données EMA accompagnées de leurs dérivées ;
- S3 : resynthèse à partir des paramètres articulatoires, eux aussi accompagnés de leurs dérivées ;
- S4 : resynthèse à partir des paramètres articulatoires auxquels a été ajouté un bruit (ancre basse) ;
- S5 : le son de référence (ancre haute).

L'ordre de présentation de ces sons est changé aléatoirement après la présentation d'un nouvel ensemble de *stimuli*. Les résultats de ce test MUSHRA, pour chaque corpus considéré, sont présentés à la figure 3.13.

Ces résultats sont en accord avec l'évaluation objective présentée précédemment. Tout d'abord, et de façon attendue, les meilleures performances sont bien obtenues pour les ancres hautes. De plus, les analyses-resynthèses des sons originaux avec LPCNet obtiennent des scores très proches de ceux des sons de référence, ce qui montre la « transparence » de ce vocodeur. Ensuite, les resynthèses à partir des paramètres articulatoires sont perçues comme inférieures en qualité à celles réalisées à partir des paramètres EMA (tout en restant bien supérieures aux ancres basses). Pour chaque corpus, un test de Kruskal-Wallis a montré une différence statistiquement significative entre les scores MUSHRA recueillis pour chaque condition de resynthèse ($p < 0{,}001$). Enfin, toujours pour chaque corpus, un test *post hoc* de Dunn a validé une différence significative pour chaque couple possible de conditions de resynthèse ($p < 0{,}001$).

Contenu phonémique L'évaluation du contenu phonémique des resynthèses a été réalisée à l'aide de deux tests de reconnaissance de phonèmes, le premier portant sur des resynthèses à partir des paramètres articulatoires et leurs dérivées, et le second sur des resynthèses à partir des données EMA et leurs dérivées. Chaque test ne comprend que des resynthèses du corpus PB2007 et est constitué de deux étapes. La première étape est la reconnaissance de consonnes au sein de resynthèses de sons de type « voyelle-consonne-voyelle » et la seconde est la reconnaissance de voyelles au sein de resynthèses de sons de voyelles isolées. Les consonnes resynthétisées pour la première étape sont /p/, /f/, /b/, /v/, /m/, /t/, /s/, /d/, /z/, /n/, /l/, /k/, /ʃ/, /g/, /ʒ/, /ʁ/, /j/, /w/ et /y/ en contexte vocalique /a/, /i/ ou /y/, et les voyelles resynthétisées pour la seconde étape sont /a/, /i/, /u/, /y/, /e/, /ɛ/, /ə/, /ø/, /o/, /ɛ̃/, /ɔ̃/ et /ɑ̃/. L'utilisateur doit écouter chacune de ces resynthèses et choisir parmi une liste le phonème reconnu. La liste de choix est constituée de l'intégralité des phonèmes associés à l'étape en cours ainsi qu'une option « pas de consonne » pour l'étape des consonnes.

La figure 3.14 montre les matrices de confusions obtenues en compilant les réponses des différents sujets pour les resynthèses à partir des paramètres articulatoires, et celles à partir des données EMA. Les matrices portant sur les voyelles montrent qu'une majorité d'entre elles ont très bien été reconnues par les sujets, suggérant une bonne resynthèse de celles-ci, avec une moyenne de leur diagonale de 82,32 % pour les resynthèses à partir des paramètres articulatoires et de 87,88 % pour celles à partir des données EMA. Cependant, on peut constater quelques faiblesses concernant la reconnaissance des voyelles nasalisées, probablement causées par l'absence d'informations sur la position du velum dans les données articulatoires du corpus PB2007. Les matrices portant sur les consonnes montrent quant à elles une proportion de reconnaissances correctes bien plus faible, avec une moyenne de leur diagonale de 28,36 % pour les resynthèses à partir des paramètres articulatoires et de 35,59 % pour celles à partir des données EMA. En effet, à part pour les consonnes /f/, /s/, /d/ et /w/ resynthétisées à partir des données EMA, le taux de bonnes réponses dépasse rarement les 50 %. Nous pouvons remarquer des confusions entre la plupart des consonnes et le /v/, *a fortiori* pour les resynthèses à partir des paramètres articulatoires, probablement dues à un son « étouffé » issu du synthétiseur. Nous pouvons également remarquer des confusions entre des consonnes telles que /k/ et /g/ ou /z/ et /s/ qui ne diffèrent que par leur trait de voisement. Ce résultat parait surprenant puisque le voisement est censé être codé lors des resynthèses par le paramètre « degré de périodicité du signal » qui est directement extrait des sons originaux et ne dépend donc pas du modèle articulatoire-vers-acoustique. Ceci suggère que la séparation entre paramètres de filtre et paramètres de source au sein de LPCNet n'est pas complète.

3.4 Conclusion

Dans cette section, nous avons présenté la méthode de construction d'un modèle articulatoire ainsi que celle d'un synthétiseur articulatoire complet. Le modèle articulatoire permet l'extraction de paramètres articulatoires interprétables à partir de données EMA et le synthétiseur permet de générer un signal de parole à partir de ceux-ci. L'évaluation objective du synthétiseur montre des résultats encourageants mais ceux de l'évaluation subjective sont moins

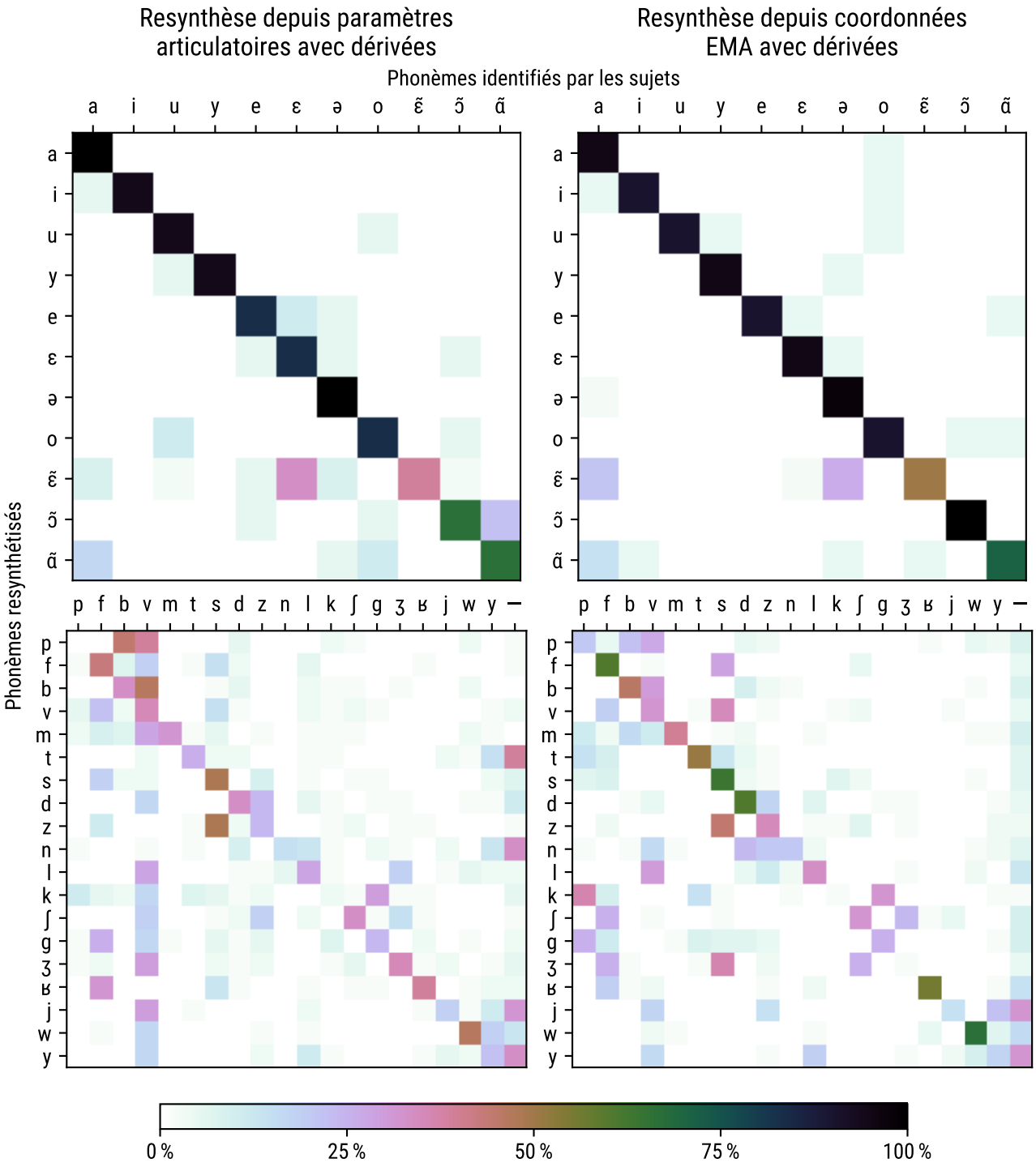

FIGURE 3.14 – **Résultats du test de reconnaissance de phonèmes conduit sur des resynthèses du corpus PB2007** La colonne « – » des matrices portant sur les consonnes symbolise l'absence de consonne.

bons. Cette différence pourrait néanmoins s'expliquer par de mauvaises conditions d'écoute lors des tests perceptifs et pourrait potentiellement être réduite avec un renouvellement de ceux-ci dans des conditions plus contrôlées. En conclusion, bien que présentant des résultats en demi-teinte, ce synthétiseur semble tout à fait exploitable pour la création d'un agent qui apprendrait à le piloter afin de produire de la parole.

Chapitre 4

Apprentissage de représentations de la parole intégrant des connaissances articulatoires

Sommaire

Nous allons maintenant, dans ce chapitre, explorer comment le fait pour un agent communicant de disposer à la fois de données acoustiques et de données articulatoires qui ont servi à produire ces données acoustiques peut contribuer à extraire des variables latentes plus robustes ou plus pertinentes pour représenter les sons de la parole dans l'objectif de les décoder (perception/reconnaissance) ou d'apprendre à les produire (production/synthèse). Nous avons vu dans les sections 1.3.1.3 et 1.3.1.4 que l'auto-encodeur variationnel (VAE) et l'auto-encodeur variationnel quantifié vectoriel (VQ-VAE) sont des outils de base pour l'apprentissage de représentations latentes. Nous explorons ici deux approches différentes, basées respectivement sur un développement autour du VAE et du VQ-VAE. Dans un premier temps (section 4.1), nous proposons d'utiliser des données articulatoires pour régulariser un VAE entraîné à reconstruire des données acoustiques, en contraignant un sous-ensemble des composantes de son espace latent à s'aligner sur la représentation articulatoire du signal encodé. Dans un second temps (section 4.2) nous utilisons directement les données articulatoires pour en extraire des représentations latentes, soit conjointement avec les données acoustiques soit de manière auto-

nome. Pour ce faire nous utilisons le VQ-VAE et nous évaluons l'apport spécifique des données articulatoires à la structuration phonétique de ces espaces latents.

4.1 Auto-encodeur variationnel régularisé articulatoirement

Le travail présenté dans cette section a fait l'objet de la publication :

— Georges, M.-A., Girin, L., Schwartz, J.-L. & Hueber, T. (2021). Learning robust speech representation with an articulatory-regularized variational autoencoder. *Conference of the International Speech Communication Association (Interspeech)*, 3345-3349

Cette section présente l'auto-encodeur variationnel régularisé articulatoirement (AR-VAE, pour *articulatory-regularized variational autoencoder*). Il a été montré que le VAE standard entraîné sur des données acoustiques est capable d'apprendre des représentations latentes pertinentes en dissociant des dimensions telles que l'identité du locuteur ou les caractéristiques phonétiques (Blaauw & Bonada, 2016 ; Hsu et al., 2017). L'AR-VAE est constitué d'un VAE également entraîné sur des données acoustiques, mais dont la construction d'une partie de l'espace latent est régularisée de façon à intégrer des informations articulatoires. Avec ce modèle, nous abordons deux questions de recherche à propos des connaissances articulatoires : (1) Peuvent-elles accélérer le processus d'apprentissage de représentations de la parole ? (2) Peuvent-elles rendre ces représentations plus robustes au bruit ?

4.1.1 Auto-encodeur variationnel

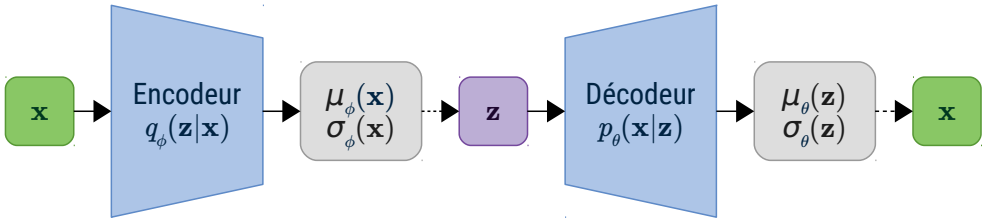

FIGURE 4.1 – **Représentation schématique de l'auto-encodeur variationnel** Les lignes en pointillé représentent un processus d'échantillonnage. La sortie est abusivement dénotée par $\mathbf{x}$ afin de rester consistant avec la formulation mathématique du modèle.

Cette sous-section présente plus en détail le VAE, représenté figure 4.1 et déjà introduit dans la sous-section 1.3.1.3. En notant $\mathbf{x}$ une donnée observée et $\mathbf{z}$ sa représentation dans un espace latent de dimension L, le VAE, introduit par Kingma et Welling, 2014 ; Rezende et al., 2014, modélise la densité de probabilité conjointe sur $\mathbf{x}$ et $\mathbf{z}$ par l'équation :

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}).$$

La distribution *a priori* $p(\mathbf{z})$ est implémentée à l'aide d'une distribution paramétrique. Le choix classique pour la représenter est une loi normale telle que $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_L)$ (avec $\mathbf{I}_L$ la matrice identité de taille L). Ce choix favorise l'orthogonalisation des dimensions latentes. La fonction de vraisemblance $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ joue le rôle de décodeur et décrit un processus de génération d'une observation $\mathbf{x}$ conditionnellement à sa représentation $\mathbf{z}$ dans l'espace latent. Cette fonction a également classiquement une forme Gaussienne, telle que :

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\theta(\mathbf{z}), \boldsymbol{\sigma}^2_\theta(\mathbf{z})),$$

avec $\boldsymbol{\mu}_\theta(\mathbf{z}) \in \mathbb{R}^F$ le vecteur de moyenne conditionnelle et $\boldsymbol{\sigma}^2_\theta(\mathbf{z}) \in \mathbb{R}^F_+$ vecteur contenant les valeurs de la diagonale de la matrice de covariance conditionnelle. Ces paramètres sont estimés à l'aide d'un réseau de neurones (le décodeur) paramétré par l'ensemble des poids $\boldsymbol{\theta}$.

Du fait des non-linéarités entre $\mathbf{z}$ et $\mathbf{x}$ dans $p_\theta(\mathbf{x}|\mathbf{z})$, $p(\mathbf{x})$ n'est pas calculable, et par conséquent $p(\mathbf{z}|\mathbf{x})$ ne l'est pas non plus avec la formule d'inversion de Bayes. De ce fait, cette dernière distribution est classiquement approximée par une distribution paramétrique $q_\phi(\mathbf{z}|\mathbf{x})$, définie telle que :

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})).$$

Les paramètres de cette distribution, également appelée modèle d'inférence, sont fournis par un autre réseau de neurones, appelé l'encodeur, dont l'entrée est $\mathbf{x}$ et les poids sont $\boldsymbol{\phi}$.

Les paramètres $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ sont estimés conjointement en maximisant, sur un corpus de données d'entraînement, la borne inférieure variationnelle (en anglais *variational lower bound*, VLB), définie par :

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - D_{\mathrm{KL}}\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big],$$

avec D_{KL} désignant la divergence de Kullback-Leibler. Le premier terme de la VLB représente la précision de reconstruction du processus d'encodage-décodage, et le second terme force $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ à s'approcher de la distribution *a priori* $p(\mathbf{z})$ de façon à forcer les composantes de $\mathbf{z}$ à être orthogonales.

4.1.2 VAE régularisé articulatoirement (AR-VAE)

Dans ce travail relatif à l'AR-VAE, qui est représenté figure 4.2, nous partons d'un jeu d'observations acoustiques et EMA conjointes. Chaque vecteur $\mathbf{x}$ est une observation acoustique décrivant le contenu spectral d'une trame du signal de parole (un vecteur de coefficients cepstraux exprimés en Bark) et $\mathbf{a}(\mathbf{x})$ est l'observation articulatoire correspondante, c'est-à-dire un vecteur de paramètres articulatoires, construit à partir des coordonnées EMA à l'aide de la technique décrite à la section 3.1.1. Pour forcer l'espace latent de l'AR-VAE à s'adapter à l'espace des paramètres articulatoires, nous ajoutons un troisième terme à la VLB en nous inspirant de la technique de régularisation présentée dans Roche et al., 2021 :

$$\mathcal{L}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log\, p_\theta(\mathbf{x}|\mathbf{z})\big] - D_{\mathrm{KL}}\big[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big] + \alpha\, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\mathcal{R}(\mathbf{z}, \mathbf{a}(\mathbf{x}))\big].$$

Ce nouveau terme de régularisation, $\mathcal{R}(\mathbf{z}, \mathbf{a}(\mathbf{x}))$, contraint pour chaque trame de parole les premières valeurs du vecteur de l'espace latent $\mathbf{z}$ à rester proches des valeurs correspondantes

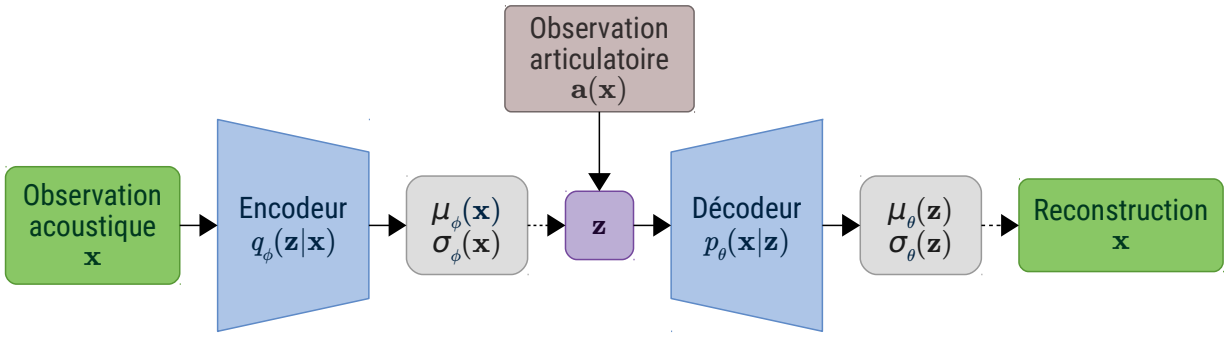

FIGURE 4.2 – **Représentation schématique du VAE régularisé articulatoirement (AR-VAE)** Lors de l'entraînement du modèle, les observations articulatoires sont utilisées pour régulariser une partie de l'espace latent du modèle (ses N premières dimensions). Les lignes en pointillé représentent un processus d'échantillonnage. La sortie est abusivement dénotée par $\mathbf{x}$ afin de rester consistant avec la formulation mathématique du modèle.

des paramètres articulatoires contenues dans $\mathbf{a}(\mathbf{x})$. Dans nos expériences, ce terme de régularisation est implémenté par l'erreur quadratique moyenne (EQM) :

$$\mathcal{R}\big(\mathbf{z}, \mathbf{a}(\mathbf{x})\big) = \|\mathbf{z}_{1:N} - \mathbf{a}(\mathbf{x})\|^2.$$

$\mathbf{z}_{1:N}$ désigne le sous-vecteur composé des N premières valeurs du vecteur latent, avec N correspondant au nombre de paramètres articulatoires. α est un facteur de pondération contrôlant le poids du terme de régularisation articulatoire, et qui sera modulé dans nos expériences.

4.1.3 Implémentation

Dans ce travail, l'AR-VAE est implémenté avec l'architecture suivante : l'encodeur avec 4 couches cachées entièrement connectées (256, 128, 64 et finalement 32 neurones) et le décodeur avec une architecture symétrique. L'espace latent possède autant de dimensions contraintes articulatoirement que de dimensions non contraintes. De ce fait, sa taille est de 12 pour les modèles entraînés sur PB2007 et de 14 pour ceux entraînés sur BY2014 (ces corpus contenant respectivement 6 et 7 paramètres articulatoires).

La tangente hyperbolique est utilisée comme fonction d'activation pour les neurones des couches cachées. L'algorithme d'optimisation Adam est utilisé pour l'apprentissage des paramètres de l'AR-VAE par descente de gradient stochastique sur des *mini-batches* de 32 observations (paires de vecteurs $\mathbf{x}$ et $\mathbf{a}(\mathbf{x})$). Pour chaque expérience, les ensembles de données ont été répartis de manière aléatoire, 80 % des données étant utilisées pour l'apprentissage et les 20 % restants pour l'évaluation.

4.1.4 Données d'apprentissage

Les données d'apprentissage utilisées pour la construction des AR-VAE présentés dans cette section proviennent des corpus PB2007 et BY2014. Les données acoustiques sont représentées par des vecteurs contenant 18 coefficients cepstraux exprimés en Bark et extraits par une fenêtre glissante de 20 ms avec une *hop size* de 10 ms. À partir des données EMA échantillonnées à 100 Hz, les données articulatoires synchrones sont représentées par des vecteurs contenant les paramètres articulatoires estimés à l'aide du modèle présenté section 3.1.1.

4.1.5 Expériences

4.1.5.1 Vitesse d'apprentissage et précision

Nous testons d'abord si l'introduction de la contrainte articulatoire peut accélérer le processus d'apprentissage. À cette fin, pour chacun des deux corpus de données, et pour chaque valeur de α prise dans $\{0\,;0{,}1\,;0{,}25\,;0{,}5\,;1\}$ nous avons entraîné 10 modèles AR-VAE (avec une initialisation différente à chaque fois) pendant 60 *epochs* (notons qu'un AR-VAE avec $\alpha = 0$ n'est pas contraint articulatoirement et est donc équivalent à un VAE conventionnel). À chaque *epoch*, nous avons calculé l'erreur de reconstruction, définie comme l'EQM entre les coefficients cepstraux reconstruits et originaux, sur l'ensemble de test. Pour chaque valeur de α (et pour chaque ensemble de données), nous avons finalement construit une version lisse de la courbe d'apprentissage en faisant la moyenne de l'erreur de reconstruction des 10 modèles sur leur ensemble de test. Ces courbes d'apprentissage sont présentées sur la figure 4.3a.

Tout d'abord, nous observons que presque tous les AR-VAE apprennent plus rapidement que les VAE classiques (c'est-à-dire que le tracé en pointillé bleu, correspondant au VAE conventionnel non régularisé, est presque toujours au-dessus des autres). Ensuite, comme le montre la figure 4.3b, la meilleure performance finale est obtenue avec l'AR-VAE sur les deux ensembles de données (avec $\alpha = 1$ pour PB2007 et avec $\alpha = 0{,}25$ pour BY2014). En conséquent, nous pouvons conclure que l'ajout de la contrainte articulatoire améliore l'apprentissage de représentations, tant en termes de vitesse d'apprentissage que de précision finale.

4.1.5.2 Robustesse face au bruit

Nous testons maintenant les performances de l'AR-VAE sur une tâche de débruitage de la parole. À cette fin, un bruit de type « cocktail-party » a été ajouté à chaque enregistrement audio. Nous avons testé différents rapports signal sur bruit (RSB) (aucun bruit, 10 dB, 5 dB et 0 dB). Des séquences d'observations acoustiques (c'est-à-dire des vecteurs contenant les 18 coefficients cepstraux en échelle Bark) ont été extraites de ces signaux audio bruités avec la même méthode que celle utilisée section 4.1.4. Pour chaque jeu de données, les VAE et AR-VAE ont été entraînés à reconstruire la version non bruitée de chaque observation acoustique à partir de son homologue bruité. Pour l'AR-VAE, nous avons d'abord comparé différentes

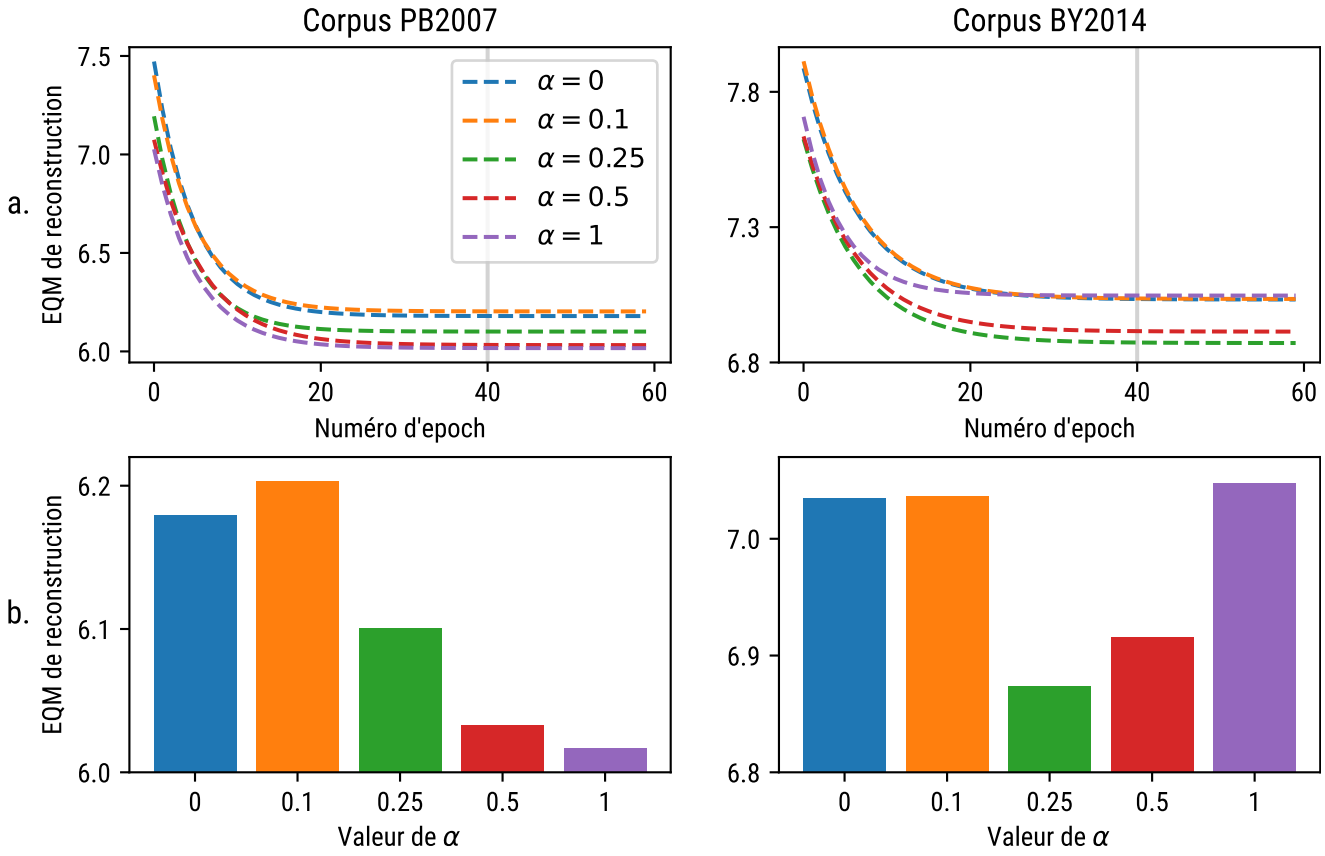

FIGURE 4.3 – **Effet de la contrainte articulatoire sur l'apprentissage** a. : Évolution de l'erreur de reconstruction sur l'ensemble de test au cours de l'apprentissage. Pour une meilleure visualisation, chaque courbe d'apprentissage est ajustée par une fonction exponentiellement décroissante. b. : Performance finale après convergence à l'*epoch* 40 sur l'ensemble de test.

valeurs du paramètre α pour cette tâche de débruitage. Pour des raisons de concision, nous ne présentons ici que les résultats avec $\alpha = 1$ qui ont fourni les meilleures performances parmi les valeurs testées. Comme pour la première expérience, nous entraînons 10 VAE et AR-VAE différents (avec une initialisation différente à chaque fois), et calculons la moyenne des résultats sur les 10 expériences. L'erreur de reconstruction sur les ensembles de données de test est illustrée figure 4.4. Ces résultats montrent que l'AR-VAE proposé surpasse le VAE conventionnel dans la tâche de débruitage pour tous les RSB considérés. La différence de performance est cependant plus prononcée pour les faibles niveaux de bruit.

Pour une analyse plus fine des performances, nous évaluons ensuite la qualité de parole débruitée au niveau segmental (phonétique). À cette fin, pour chaque phrase de l'ensemble d'évaluation, et pour chaque RSB considéré, nous avons resynthétisé un signal de parole synthétique en utilisant le vocodeur neuronal LPCNet, à partir des coefficients cepstraux reconstruits par le VAE ou l'AR-VAE, ainsi que les paramètres de source (hauteur et harmonicité) extraits des enregistrements originaux non-bruités. Le contenu phonétique du signal reconstruit est ensuite évalué à l'aide d'un décodeur acoustico-phonétique basé sur des HMM, déjà

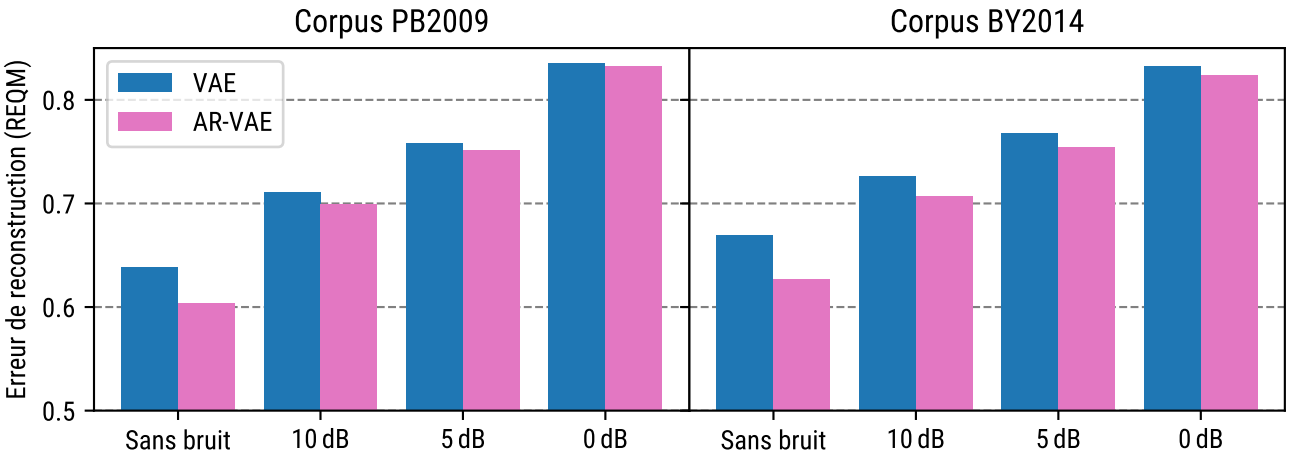

FIGURE 4.4 – **Erreur de reconstruction du VAE conventionnel et de l'AR-VAE sur l'ensemble de test pour la tâche de débruitage de la parole** Performance moyenne sur 10 expériences.

présenté section 3.3.2.1 et dont les paramètres ont été estimés à partir des signaux originaux (non-bruités) de l'ensemble d'apprentissage. La précision (*accuracy*) du décodage (qui prend en compte les erreurs d'insertion et d'omission) est présentée figure 4.5. De nouveau, l'AR-VAE surpasse le VAE, avec une marge importante (jusqu'à 10 %) lors du traitement de la parole propre, et avec une marge plus faible lors du traitement de la parole bruitée.

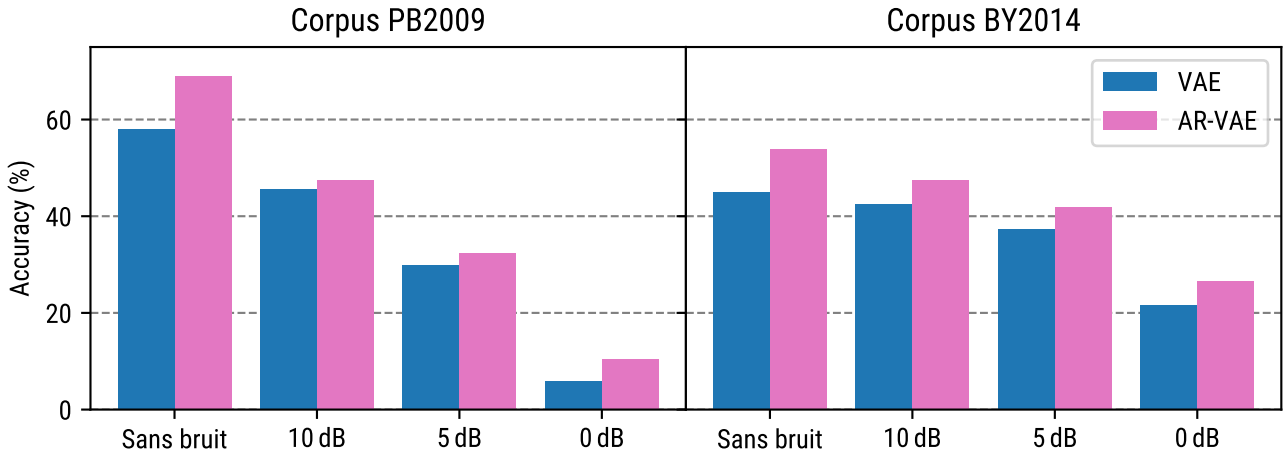

FIGURE 4.5 – **Précision d'un décodeur phonétique basé sur un HMM lors du traitement de signaux vocaux débruités par le VAE conventionnel et par l'AR-VAE**

Enfin, nous proposons une étude comparative des deux modèles VAE et AR-VAE basée sur un test perceptif de type MUSHRA, dont le paradigme a déjà été présenté section 3.3.2.2. Pour rappel, un test MUSHRA est un test dans lequel les participants doivent ordonner, sur une échelle sans unité allant de « très différent de l'original » à « très similaire à l'original », plusieurs *stimuli* audio en fonction de leur similarité avec un *stimulus* de référence (ITU, 2015). Pour construire ce test, nous avons d'abord sélectionné aléatoirement 20 phrases courtes du

corpus BY2014 (préféré au corpus PB2007 en raison de la présence de données articulatoires sur le velum). Ensuite, pour chaque phrase nous avons construit par analyse-resynthèse avec LPCNet le *stimulus* de référence à partir de la phrase originale, et l'ancre basse à partir d'une version bruitée de celle-ci avec un RSB de 0 dB. Afin de tester 4 niveaux de bruit différents sans que le test ne soit trop long pour les sujets, nous avons alterné successivement deux types de question. Dans chacun de ces deux types, seulement 2 niveaux de bruits A et B sont considérés ($A = 5$ dB et $B = 0$ dB pour le premier type et $A = 10$ dB et $B =$ pas de bruit pour le second). Indépendamment de l'ancre haute et de l'ancre basse, voici la liste des *stimuli* générés pour ces questions :

— S1 : analyse-synthèse d'un signal bruité avec un RSB de A, à partir des paramètres cepstraux inférés à l'aide du VAE, et des paramètres de source (f0, coefficient de périodicité) originaux, à l'aide du vocodeur LPCNet ;
— S2 : même méthode que S1 mais avec l'AR-VAE ;
— S3 et S4 : même méthode que S1 et S2 mais pour un RSB de B.

Ce test a été effectué en ligne par 23 locuteurs natifs français, recrutés via la plate-forme *Prolific* (Palan & Schitter, 2018).

Les résultats sont présentés dans la figure 4.6. Pour évaluer la significativité statistique de la différence entre les scores MUSHRA, nous avons d'abord effectué un test de Kruskal-Wallis qui a montré un effet significatif du facteur RSB ($p < 0{,}05$). Ensuite, un test post-hoc de Dunn a validé une augmentation statistiquement significative de la performance de VAE à AR-VAE pour un audio propre ($p < 0{,}001$) et pour un audio bruité avec RSB = 10 dB ($p < 0{,}05$), et n'a montré aucune différence significative entre les deux modèles pour RSB = 5 dB et RSB = 0 dB (c'est-à-dire des entrées très bruitées).

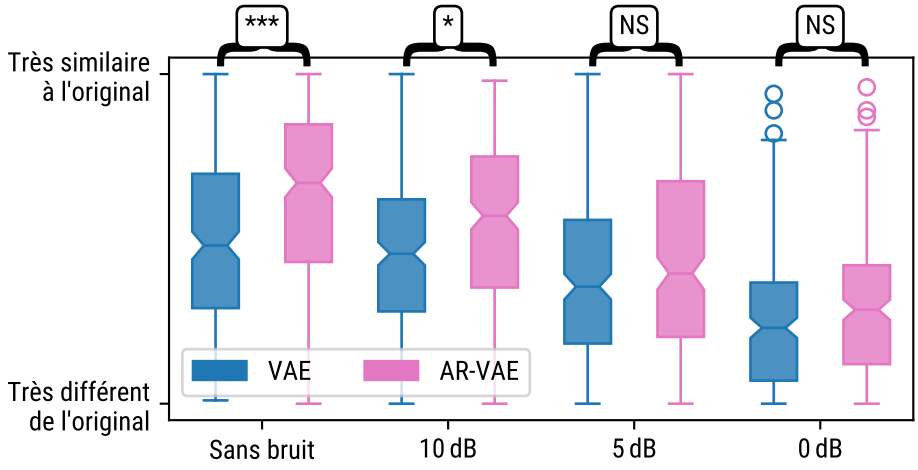

FIGURE 4.6 – **Scores MUSHRA de similarité par rapport au son original obtenus pour chaque niveau de bruit, pour le VAE conventionnel et l'AR-VAE** Par souci de clarté, nous omettons les scores obtenus sur l'ancre basse ainsi que sur la référence cachée. *** et * sont significatifs ($p < 0{,}001$ et $p < 0{,}05$), et NS sont des différences non significatives.

4.1.6 Conclusion

Dans cette section, nous avons montré comment l'injection de connaissances *a priori* de nature articulatoire permettrait d'apprendre des représentations plus robustes de la parole. Ces représentations ont été construites en appliquant une régularisation articulatoire sur une partie de l'espace latent d'un VAE auto-encodant des traits acoustiques de la parole, les obligeant à adopter des représentations mixtes acoustico-articulatoires. La régularisation articulatoire a été rendue possible grâce à l'ajout d'un terme supplémentaire dans la fonction de coût d'apprentissage. Cet ajout semble améliorer les performances du modèle, à la fois pour l'analyse-resynthèse des données non-bruitées, et pour la tâche de débruitage étudiée.

4.2 Découverte auto-supervisée de représentations acoustiques et articulatoire avec le VQ-VAE

Le travail présenté dans cette section a fait l'objet de la publication :

— Georges, M.-A., Schwartz, J.-L. & Hueber, T. (2022). Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE. *Conference of the International Speech Communication Association (Interspeech)*, 774-778

Dans cette seconde étude, nous n'utilisons plus les données articulatoires pour contraindre l'espace latent acoustique mais directement comme base d'apprentissage d'un espace latent dans un objectif de découverte d'unités de la parole à partir de données acoustiques et articulatoires. Pour ce faire, nous combinons les données acoustiques et articulatoires, pour en extraire soit un espace latent conjoint soit deux espaces latents indépendants. Ces espaces latents sont ici pré-catégorisés en utilisant des VQ-VAE, présentés dans la section 1.3.1.4. Nous étudions les possibles complémentarités des modalités acoustique et articulatoire dans la représentation des unités phonétiques, en nous appuyant sur les tests ABX qui constituent une mesure objective du contenu de ces représentations.

4.2.1 Auto-encodeur variationnel quantifié vectoriel

Le VQ-VAE (van den Oord et al., 2017), introduit en section 1.3.1.3 et représenté figure 4.7, peut être considéré comme une version du VAE classique avec un espace latent discret. Tout comme ce dernier, le VQ-VAE modélise la relation entre des données observées $\mathbf{x}$ et leur représentation latente $\mathbf{z}$. Il possède lui aussi un décodeur représentant la fonction de vraisemblance $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ et un encodeur représentant la probabilité *a posteriori* $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, tous deux implémentés avec des réseaux de neurones (dont les poids respectifs sont $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$).

Contrairement au VAE classique, l'espace latent $p(\mathbf{z})$ du VQ-VAE est discret, et l'*a priori*

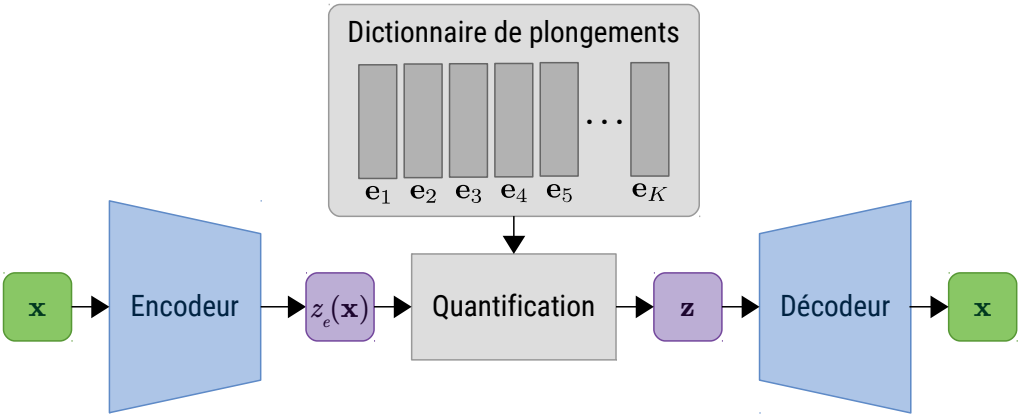

FIGURE 4.7 – **Représentation schématique de l'auto-encodeur variationnel quantifié vectoriel** La sortie est abusivement dénotée par $\mathbf{x}$ afin de rester consistant avec la formulation mathématique du modèle.

sur celui-ci est une loi uniforme discrète. Cette discrétisation est assurée par une étape de quantification vectorielle prenant place entre l'encodeur et le décodeur du modèle. Cette dernière s'appuie sur un dictionnaire de vecteurs de plongement $\mathbf{e} \in \mathbb{R}^{K \times D}$ contenant K vecteurs de dimension D. Une représentation latente $\mathbf{z}$ correspondant à une donnée observée $\mathbf{x}$ ne peut prendre comme valeur que celle de l'un des vecteurs contenus dans ce dictionnaire de vecteurs de plongement. Cette représentation est définie, à partir d'une observation d'entrée $\mathbf{x}$, comme le vecteur $\mathbf{e}_k$ du dictionnaire de plongements le plus proche de la sortie (continue) de l'encodeur $z_e(\mathbf{x})$. Ce mécanisme est traduit dans la définition du *posterior* du VQ-VAE, dont la distribution est catégorielle :

$$q_{\boldsymbol{\phi}}(\mathbf{z} = \mathbf{e}_k \mid \mathbf{x}) = \begin{cases} 1 & \text{si } k = \operatorname{argmin}_j \|z_e(\mathbf{x}) - \mathbf{e}_j\|, \\ 0 & \text{sinon.} \end{cases}$$

Les poids $\boldsymbol{\phi}$ de l'encodeur, les poids $\boldsymbol{\theta}$ du décodeur et la valeur des vecteurs $\mathbf{e}$ du dictionnaire de plongement sont entraînés par rétropropagation du gradient d'erreur, en considérant la fonction de coût suivante :

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x}, \mathbf{e}_k) = -\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z} = \mathbf{e}_k) + \|\text{sg}[z_e(\mathbf{x})] - \mathbf{e}_k\|^2 + \beta\|z_e(\mathbf{x}) - \text{sg}[\mathbf{e}_k]\|^2,$$

où $\mathbf{e}_k$ correspond au vecteur du dictionnaire de plongements le plus proche de la sortie de l'encodeur avant discrétisation $z_e(\mathbf{x})$, et où sg correspond à un opérateur de blocage du gradient d'erreur, défini comme une fonction identité lors de la *forward pass* et ayant une dérivée nulle, contraignant son opérande à rester inchangée lors de la *backward pass*. Le premier terme de cette fonction de coût assure que la sortie du VQ-VAE soit la plus proche possible de son entrée. Le second terme sert à mettre à jour la valeur du vecteur latent $\mathbf{e}_k$ afin que celle-ci se rapproche de la sortie de l'encodeur $z_e(\mathbf{x})$. Enfin, le troisième terme sert à mettre à jour l'encodeur de façon à ce que sa sortie $z_e(\mathbf{x})$ soit plus proche du vecteur latent $\mathbf{e}_k$. Le poids de ce dernier terme peut être modulé par l'hyperparamètre β. Les auteurs du VQ-VAE affirment que l'algorithme d'apprentissage est robuste face aux variations de β entre 0,1 et 2. Dans leurs

expériences, ils fixent la valeur de celui-ci à $\beta = 0{,}25$, valeur que nous utilisons également dans ce travail. Enfin, nous pouvons noter l'absence dans la fonction de coût du VQ-VAE, par rapport à celle du VAE classique, d'un terme faisant appel à la divergence de Kullback-Leibler pour assurer un tirage des représentations latentes en accord avec le *prior* uniforme $p(\mathbf{z})$. Cette absence est due au fait que la distribution $q(\mathbf{z}|\mathbf{x})$ prend toujours la même forme (une valeur de 1 sur le vecteur choisi du dictionnaire et 0 sur tous les autres), rendant l'ajout de ce terme inutile puisque celui-ci aurait une valeur constante.

4.2.2 Données d'apprentissage

Ce travail d'extraction d'unités de façon non supervisée est réalisé sur les données articulatoires et acoustiques des corpus PB2007 (un locuteur français) et MOCHA (deux locuteurs anglais), respectivement présentés en section 2.2 et en section 2.4. Les données articulatoires prennent la forme de vecteurs extraits toutes les 10 ms contenant chacun la valeur des 7 paramètres articulatoires (ou 6 pour PB2007, qui ne contient pas d'informations sur le velum) présentés en section 3.1.1 et extraits des enregistrements EMA des deux corpus. Un Mel-spectrogramme à 40 dimensions a été utilisé pour représenter le contenu acoustique du signal vocal (extrait de la forme d'onde vocale échantillonnée à 16 kHz et enregistrée en synchronisation avec les mouvements articulatoires, avec fenêtre glissante de 25 ms et une *hop size* de 10 ms en utilisant la boîte à outils Python *librosa*).

4.2.2.1 Implémentation

Pour chaque locuteur, nous avons entraîné trois VQ-VAE différents : l'un à partir des données articulatoires uniquement, appelé ensuite « VQ-VAE articulatoire », un second à partir des données acoustiques correspondantes, appelé « VQ-VAE acoustique », et enfin un troisième implémentant une stratégie de « fusion précoce » basée sur la concaténation des vecteurs de traits articulatoires et acoustiques, nommé « VQ-VAE acoustico-articulatoire ». Chaque VQ-VAE prend en entrée, et doit reconstruire en sortie, 5 trames concaténées des données encodées, amenant l'empan temporel d'une représentation latente à 50 ms (supposée correspondre à la durée typique d'événements acoustiques pertinents). Pour le reste, chaque VQ-VAE a été implémenté suivant la même architecture : l'encodeur a été construit avec 3 couches cachées entièrement connectées de 256 neurones chacune (des analyses préliminaires ont montré que des architectures plus complexes n'apportaient pas de gains significatifs de performance), avec la fonction tangente hyperbolique utilisée à l'entrée de ceux-ci (des couches de *dropout* et de *batch normalization* ont été insérées après chaque couche cachée avec un ratio de *dropout* de $p = 0{,}25$), et avec une couche linéaire finale de la taille de la dimension des vecteurs de plongement (i.e. D). Une structure duale a été utilisée pour le décodeur, mais avec une couche linéaire finale adaptée à la taille des données encodées.

L'apprentissage des modèles a été réalisé par rétropropagation avec l'optimiseur Adam, sur des *mini-batchs* de 8 séquences de 5 vecteurs d'observation. L'implémentation a été réalisée en utilisant la librairie *PyTorch*.

Pour chaque expérience, les ensembles de données ont été répartis de manière aléatoire, 80 % des données étant utilisées pour l'apprentissage et les 20 % restants pour les tests. 20 % de l'ensemble d'entraînement a été utilisé comme ensemble de validation (pour contrôler l'*early stopping*). Aussi, les données ont chaque fois été centrées et réduites en utilisant la moyenne et la variance calculées sur l'ensemble d'entraînement. Pour une meilleure robustesse des résultats expérimentaux, chaque modèle VQ-VAE a été entraîné et évalué 5 fois, avec à chaque fois un partitionnement aléatoire différent des ensembles de données. Les mesures de performance d'un VQ-VAE rapportées ici consistent en la performance moyenne sur ces 5 évaluations.

4.2.3 Test ABX

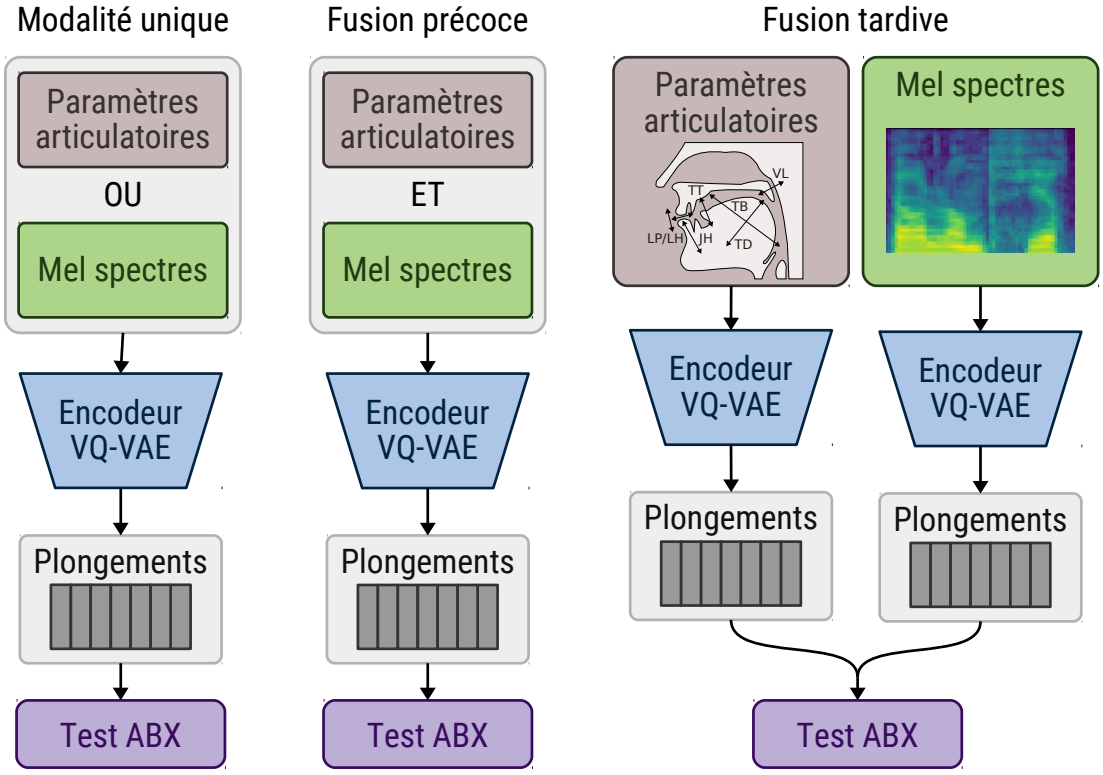

FIGURE 4.8 – **Cadre proposé pour l'apprentissage de représentations discrètes de la parole à partir de données articulatoires et acoustiques en utilisant le VQ-VAE**

Pour analyser les propriétés phonétiques des représentations latentes (c'est-à-dire les vecteurs des plongements) apprises par les différents VQ-VAE, nous avons mis en oeuvre un outil d'évaluation largement utilisé dans le cadre du *Zero Resource Speech Challenge* : les tests ABX. Cette méthode repose sur l'idée qu'une occurrence A d'une unité devrait être plus proche d'une autre occurrence X de la même unité, que de toute autre occurrence B d'une unité différente. Par exemple, si le type d'unités considéré est le phonème, le test ABX se déroule de la façon suivante :

1. Deux occurrences A et X du même phonème (par exemple /b/) et une occurrence B d'un autre phonème (par exemple /p/) sont choisies.

2. Les distances $d_{A,X}$ entre A et X et $d_{B,X}$ entre B et X sont évaluées (la distance est définie par le modélisateur).
3. Le test est considéré comme réussi si X est plus proche de A que de B, autrement dit si $d_{A,X} < d_{B,X}$.

En répétant ce test entre les occurrences de deux types d'unités et en mesurant leur taux de réussite moyen, on obtient alors un score de discriminabilité entre ces deux types d'unités. L'intérêt majeur de ce test est qu'il s'affranchit d'un modèle explicite de catégorisation des unités (par exemple un classifieur phonétique), utilisant seulement la discrimination basée sur la distance comme outil d'évaluation quantitative.

Le présent travail se concentre sur les consonnes dans des contextes vocaliques variés. Ainsi, les unités considérées lors des tests ABX sont les consonnes. Pour chaque locuteur, nous avons extrait de son ensemble de test correspondant toutes les séquences voyelle-consonne-voyelle (VCV, où les deux voyelles V avant et après la consonne C peuvent être différentes) et leurs représentations discrètes correspondantes dans les espaces latents de tous les VQ-VAE entraînés (VQ-VAE articulatoire, acoustique et articulatoire-acoustique). À partir de ces séquences, et pour toutes les consonnes, nous avons construit un ensemble de triplets A, B, X avec A et X les représentations associées à deux occurrences d'une même consonne, mais potentiellement dans un contexte vocalique différent, et B, les représentations associées à une occurrence d'une consonne différente, là encore dans un contexte vocalique possiblement différent. Nous avons ensuite comparé les distances entre A et X d'une part, et B et X d'autre part, et ce pour chaque type de VQ-VAE (VQ-VAE articulatoire, acoustique ou articulatoire-acoustique).

Dans ce travail, la distance entre deux représentations de consonne est définie comme suit :

1. Pour chacune des deux consonnes, nous extrayons de la séquence VCV la contenant la séquence des vecteurs de plongement correspondants en utilisant l'encodeur du VQ-VAE concerné.
2. Nous gardons uniquement les trames relatives à la consonne centrale (nous nous basons ici sur la segmentation disponible au niveau phonétique des jeux de données PB2007 et MOCHA).
3. Nous alignons temporellement les séquences de plongement des consonnes de A et X d'une part, et de B et X d'autre part, en utilisant l'algorithme DTW (*dynamic time warping*).
4. Nous calculons la similarité cosinus moyenne le long du chemin DTW (pour cette procédure, nous utilisons la boîte à outils fournie dans Schatz, 2016).

Pour limiter le coût de calcul, ce test ABX n'a pas été effectué pour chacun des triplets A, B, X possibles dans chaque ensemble de données, mais seulement à partir d'un sous-ensemble de 5 000 triplets (A, B, X) sélectionnés de manière aléatoire, en veillant à ce que chaque paire (A, B) possible soit représentée de manière égale. Pour chaque VQ-VAE et pour chaque locuteur, un score global de discriminabilité a été défini comme le taux de réussite moyen de tous les tests ABX individuels.

En plus de la stratégie de fusion précoce mentionnée ci-dessus, basée sur la concaténa-

tion des vecteurs de caractéristiques articulatoires et acoustiques et leur modélisation à l'aide d'un seul VQ-VAE (comme illustré sur la partie « fusion précoce » de la figure 4.8), nous avons étudié une autre approche pour combiner les deux modalités lors des tests ABX (partie « fusion tardive » de la figure 4.8). Lors du traitement des représentations A et X, nous calculons la distance entre elles avec $d^{fusion}_{A,X} = \omega.d^{ac}_{A,X} + d^{art}_{A,X}$, où $d^{ac}_{A,X}$ est la distance entre les représentations A et X obtenues avec le VQ-VAE acoustique, et $d^{art}_{A,X}$ est la distance entre les représentations A et X mais obtenues avec le VQ-VAE articulatoire, et avec ω un facteur de pondération entre ces deux modalités. La distance pour la paire (B, X) est calculée de la même manière. Un seul test ABX est alors considéré comme un « succès » lorsque $d^{fusion}_{A,X} < d^{fusion}_{B,X}$. Cette approche est appelée ici « fusion tardive ».

4.2.4 Expériences

4.2.4.1 Taille et dimension du dictionnaire de plongements

Afin de calibrer les espaces latents des différents modèles VQ-VAE utilisés dans la présente étude, nous avons d'abord réalisé une série d'expériences pour évaluer l'impact de leur dimension D et de la taille K du dictionnaire de plongements servant à leur quantification.

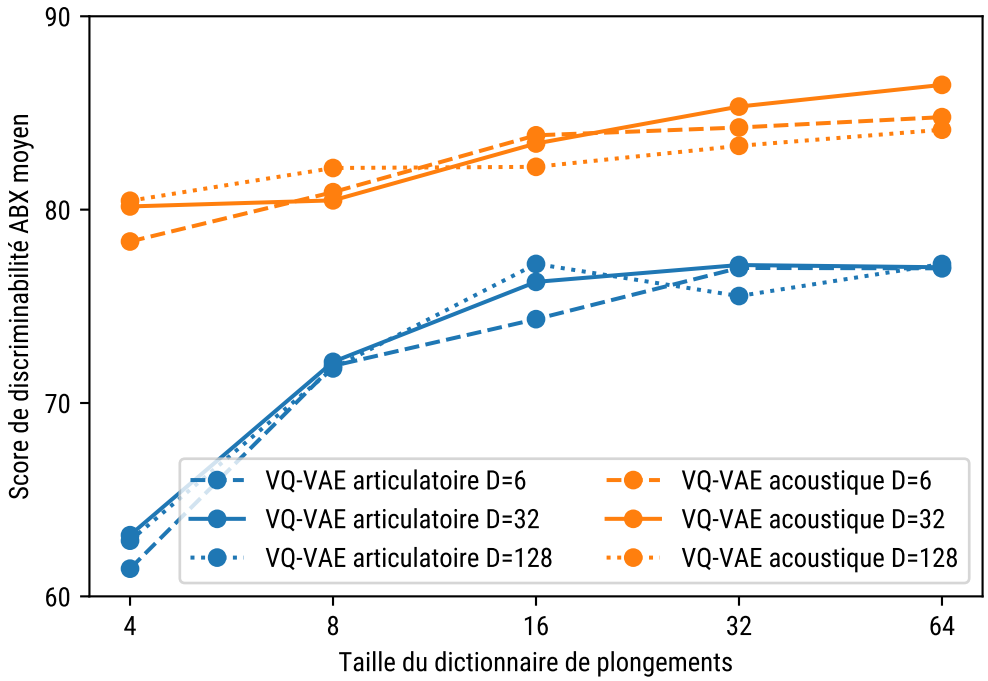

FIGURE 4.9 – **Score de discriminabilité ABX en fonction de la dimension D de l'espace latent et de la taille K du dictionnaire de plongements pour le locuteur PB** Chaque point représente le score moyen de 5 modèles entraînés à partir d'une initialisation différente.

Nous avons d'abord cherché à évaluer quel était l'impact de la dimension D de l'espace

latent des VQ-VAE sur leurs performances. Pour cela, nous avons entraîné 5 VQ-VAE articulatoires et 5 VQ-VAE acoustiques pour chacune des combinaisons possibles de dimension d'espace latent $D = \{6, 32, 128\}$ et de tailles de dictionnaire de plongements $K = \{4, 8, 16, 32, 64\}$. Ensuite, nous avons calculé, toujours pour chaque combinaison possible de ces paramètres, la moyenne des scores ABX des 5 VQ-VAE articulatoires et acoustiques. En observant la figure 4.9, sur laquelle sont reportés ces résultats, nous pouvons remarquer que la dimension D de l'espace latent ne semble avoir qu'un impact modéré sur les performances des VQ-VAE articulatoire et acoustique. En effet, les courbes associées à chacun de ces modèles et à chacune des dimensions D semblent se confondre. En revanche, la taille K du dictionnaire de plongements semble, elle, avoir un impact plus important sur les performances des modèles.

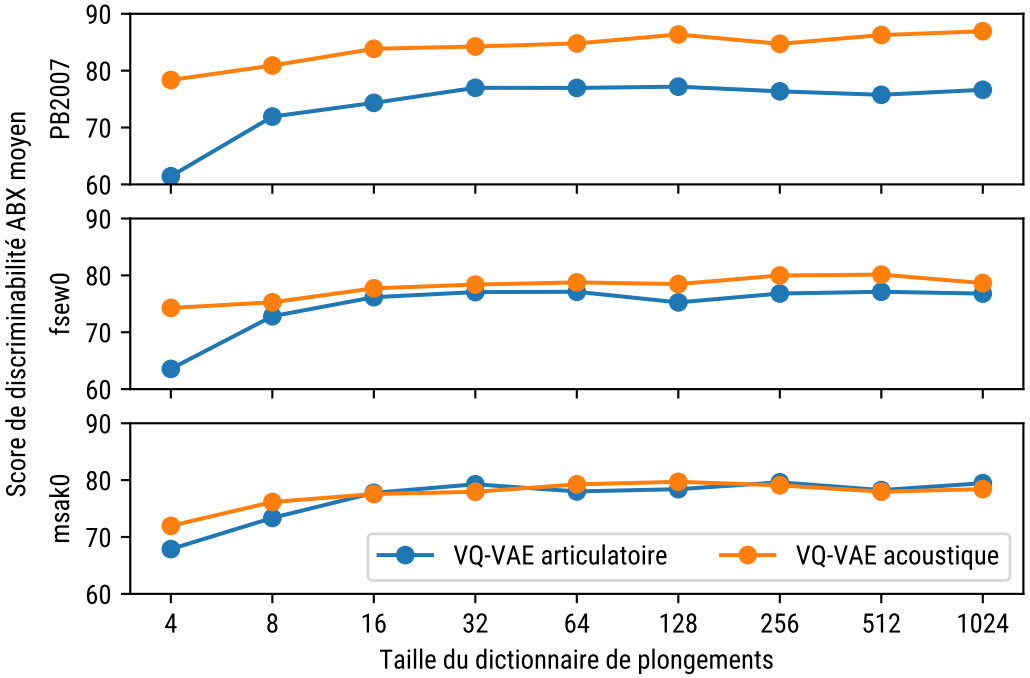

FIGURE 4.10 – **Score de discriminabilité ABX en fonction de la taille K du dictionnaire de plongements** La dimension de l'espace latent des VQ-VAE est de $D = 32$. Chaque point représente le score moyen de 5 modèles entraînés à partir d'une initialisation différente.

Nous avons par la suite étudié plus en profondeur l'impact du paramètre K. Pour ce faire, nous avons entraîné 5 VQ-VAE articulatoires et 5 VQ-VAE acoustiques avec $D = 32$ pour chacune des tailles de dictionnaire de plongements $K = \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ et pour chacun des locuteurs PB, fsew0 et msak0. Ensuite, nous avons calculé le score de discriminabilité ABX moyen de chaque groupe de 5 VQ-VAE articulatoires et acoustiques. En observant la figure 4.10, sur laquelle sont reportés ces résultats, nous observons une montée en performances des différents VQ-VAE en fonction de K. Cependant cette montée semble atteindre rapidement un plateau autour de $K = 32$ et $K = 64$.

À la lumière de ces résultats, nous décidons de fixer la dimension de l'espace latent des VQ-VAE que nous présenterons par la suite à $D = 32$ et la taille de leur dictionnaire de plongement à $K = 64$.

4.2.4.2 Comparaison des VQ-VAE articulatoire et acoustique

Pour les 3 locuteurs PB (issu du corpus PB2007), fsew0 et msak0 (tous deux issus du corpus MOCHA), les scores de discriminabilité ABX globaux étaient respectivement de 86,4 %, 78,4 %, 78,8 % pour les VQ-VAEs acoustiques, et de 77 %, 76,7 %, 78,3 % pour les VQ-VAEs articulatoires. Ainsi, les scores des modèles acoustique et articulatoire sont proches pour les deux locuteurs de MOCHA mais diffèrent plus fortement pour le locuteur PB, avec un score nettement plus faible pour le modèle articulatoire. Cela peut s'expliquer par le manque d'informations sur le vélum pour ce locuteur, rendant difficile l'encodage du trait de nasalité dans les représentations apprises.

Pour mieux comprendre les schémas d'erreurs pour chaque modalité, nous reportons sur la figure 4.11 le score moyen de discriminabilité ABX pour chaque paire de consonnes, pour le locuteur PB. Nous observons que pour le VQ-VAE articulatoire, les scores de discriminabilité sont plus faibles pour les paires de consonnes ayant le même lieu d'articulation (par exemple les palatales /s/ vs /d/ ou les labiales /f/ vs /b/) alors que pour le VQ-VAE acoustique, les scores sont plus faibles pour les paires de consonnes ayant le même mode d'articulation (par exemple les fricatives non voisées /f/ vs /s/).

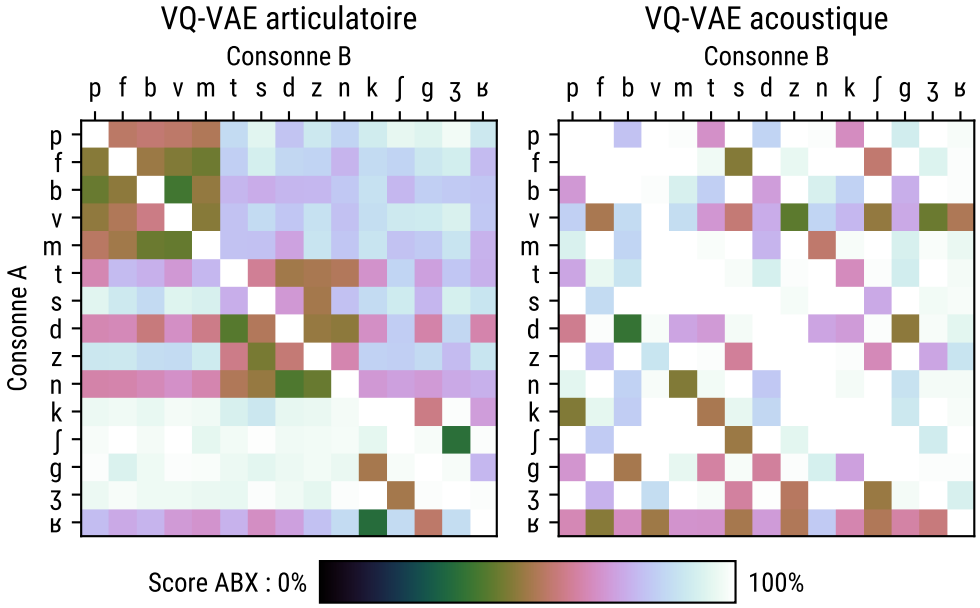

FIGURE 4.11 – **Scores de discriminabilité ABX pour chaque paire de consonnes calculés pour le locuteur PB** La diagonale est vide car une consonne n'est jamais testée contre elle-même.

En conséquence, nous avons effectué une analyse plus fine afin de mieux étudier comment les représentations apprises discriminent les consonnes en fonction de leur lieu et de leur mode d'articulation. À cette fin, nous avons défini deux scores distincts de discriminabilité ABX, respectivement appelés « score ABX de mode d'articulation » et « score ABX de lieu d'articulation ». Pour calculer le premier (centré sur le mode d'articulation), nous avons

regroupé les consonnes en trois sous-groupes ayant un lieu d'articulation similaire, à savoir les labiodentales, les palatales et les dorsales. Nous avons appliqué la méthodologie de test ABX au sein de chaque groupe (par exemple, pour le groupe labiodental, A est /abo/, X est /iba/ et B est /uvo/) et nous avons calculé le taux de réussite moyen pour les trois groupes. Pour calculer le second score (portant sur le lieu d'articulation), nous avons appliqué la même procédure mais après avoir regroupé les consonnes en fonction de leur mode d'articulation. Nous avons considéré les cinq sous-groupes suivants : les consonnes occlusives voisées, les consonnes occlusives non voisées, les fricatives/affriquées voisées, les fricatives/affriquées non voisées et les sonantes, c'est-à-dire les liquides et les nasales.

Nous avons reporté ces deux scores pour chaque locuteur dans la figure 4.12 (points bleus et orange). Les résultats confirment que le VQ-VAE articulatoire fournit des représentations latentes qui discriminent plutôt le lieu d'articulation (le score de lieu est plus élevé que le score de mode pour le VQ-VAE articulatoire) alors que le VQ-VAE acoustique structure l'espace latent principalement en termes de mode d'articulation (le score de mode est plus élevé que le score de lieu pour le VQ-VAE acoustique). On peut noter que la différence est plus prononcée pour le locuteur PB que pour les locuteurs fsew0 et msak0. Cet écart pourrait s'expliquer par la différence de contenu linguistique dans chaque ensemble de données, avec le corpus PB2007 contenant principalement des séquences VCV alors que le corpus MOCHA est principalement constitué de phrases.

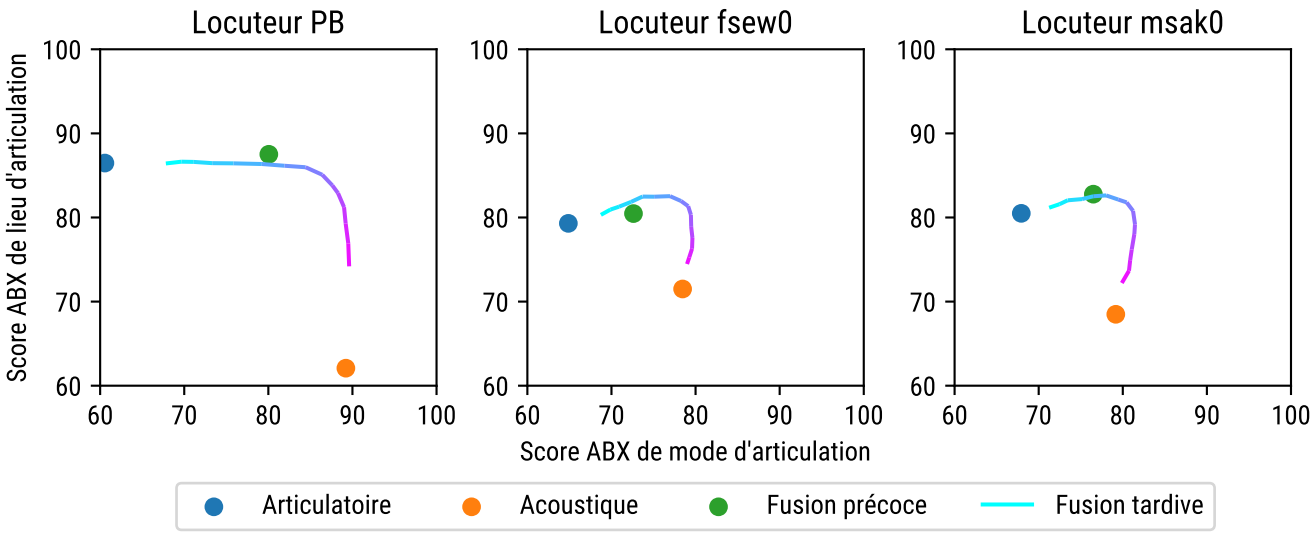

FIGURE 4.12 – **Scores ABX en fonction du lieu et du mode d'articulation pour les modalités articulatoire et acoustique, ainsi que pour les stratégies de fusion précoce et tardive** Les performances de la fusion tardive varient avec ω entre 10^{-1} (cyan) et 10^{1} (violet).

4.2.4.3 Fusion précoce vs. fusion tardive des modalités

Nous avons également reporté sur la figure 4.12 les scores de discrimination ABX du lieu et du mode d'articulation, à la fois pour la stratégie de fusion précoce (concaténation des vecteurs de données acoustiques et articulatoires, points verts) et la stratégie de fusion tardive (combinaison des distances ABX des VQ-VAE acoustiques et articulatoires, lignes colorées). Pour chaque locuteur, la stratégie de fusion précoce donne des résultats légèrement meilleurs que le VQ-VAE uniquement articulatoire en termes de lieu d'articulation et une performance modérément plus faible que le VQ-VAE uniquement acoustique en termes de mode d'articulation.

Quant à la stratégie de fusion tardive, en modulant la contribution des représentations issues des VQ-VAE acoustique et articulatoire (i.e. ω variant entre 10^{-1} et 10^{1}), les performances suivent une trajectoire avec une valeur optimale qui donne des performances presque aussi bonnes que chaque modalité prise indépendamment pour le jeu de données PB2007, et même de meilleures performances pour le jeu de données MOCHA. Par conséquent, ces expériences tendent à démontrer que ces stratégies de fusion peuvent tirer profit des modalités acoustique et articulatoire pour la découverte d'unités de parole.

4.2.5 Conclusion

Dans cette section, nous avons étudié l'utilisation de données articulatoires pour l'apprentissage de représentations discrètes de la parole à l'aide de VQ-VAE. Nos simulations effectuées sur 3 locuteurs (2 langues différentes) montrent le bénéfice potentiel d'exploiter des connaissances articulatoires pour découvrir des unités de la parole de façon auto-supervisée. Dans un certain sens, ces résultats ne sont pas vraiment surprenants et correspondent bien aux attentes concernant le rôle des caractéristiques articulatoires dans la représentation du lieu d'articulation des consonnes (Liberman et al., 1967). La proposition de stratégies de fusion, précoce ou tardive, apporte un éclairage supplémentaire, suggérant que la complémentarité des représentations acoustique et articulatoire pourrait en effet être cruciale pour fournir des représentations phonétiques robustes et complètes. Il est néanmoins important de confirmer ces attentes sur des corpus plus étendus, et sur un plus grand nombre de locuteurs et de langues.

Chapitre 5

Modélisation d'un agent apprenant la parole

Sommaire

Nous sommes maintenant dotés des différentes composantes nécessaires à notre projet : un synthétiseur vocal pilotable à partir de paramètres articulatoires interprétables, ainsi que différentes techniques d'apprentissage auto-supervisé de représentations de la parole, exploitant conjointement des données acoustiques et articulatoires. Nous abordons dans ce dernier chapitre le développement d'un agent communicant, apprenant à parler en répétant des sons fournis par son environnement, et construisant ses connaissances des relations acoustico-articulatoires de façon auto-supervisée (c'est-à-dire sans jamais avoir accès, pour un *stimulus* acoustique perçu, au geste articulatoire qui doit lui être associé). Cette dernière phase de notre étude s'est déroulée en deux temps, qui constituent les deux sections de ce chapitre.

D'abord, nous avons développé une première version d'agent, dite « agent à but imitatif », dont le but est d'apprendre à répéter les sons qu'il perçoit en visant à s'approcher le plus possible de leur contenu spectro-temporel, sans chercher à identifier puis à reproduire un quelconque contenu linguistique (phonétique, syllabique, lexical, etc.).

Ensuite, nous proposons un agent dit « agent à but communicatif » dont l'objectif n'est pas de s'approcher du contenu spectral du *stimulus* de parole qu'il perçoit mais plutôt d'en extraire une séquence d'unités linguistiques discrètes, puis de la reproduire. Cette approche

nous apparaît plus plausible d'un point de vue développemental car, même si des phénomènes de convergence phonétique restent possibles, l'apprentissage de la parole par l'enfant tient probablement plus à l'apprentissage d'un « code » qu'à une imitation parfaite du timbre des voix qu'il entend.

5.1 Agent à but imitatif

Le travail présenté dans cette section a fait l'objet de la publication :

— Georges, M.-A., Diard, J., Girin, L., Schwartz, J.-L. & Hueber, T. (2022). Repeat after Me : Self-Supervised Learning of Acoustic-to-Articulatory Mapping by Vocal Imitation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8252-8256

5.1.1 Vue d'ensemble

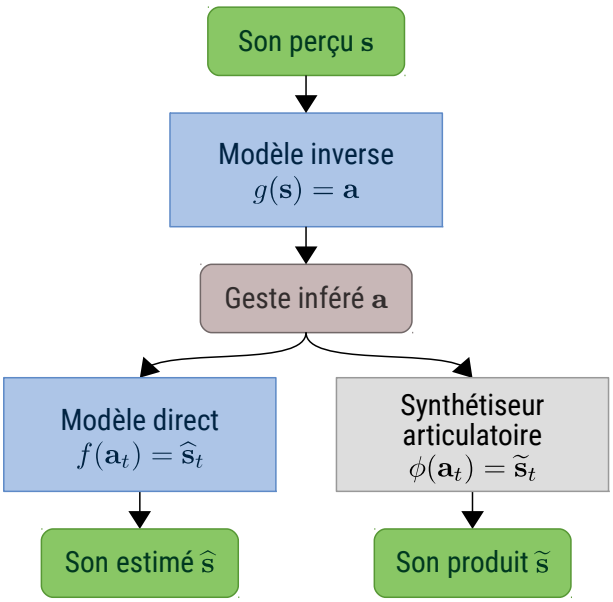

FIGURE 5.1 – **Architecture de l'agent à but imitatif**

L'architecture de l'agent à but imitatif proposé est présentée sur la figure 5.1. Nous décrivons ici les 3 modules qui composent cet agent, ainsi que les notations mathématiques associées.

Le synthétiseur articulatoire Le synthétiseur articulatoire ϕ permet à l'agent de produire, à partir d'une séquence de vecteurs de paramètres articulatoires $\mathbf{a} = [\mathbf{a}_1, \cdots, \mathbf{a}_T]$ (où T est la longueur de la séquence), la séquence de vecteurs de paramètres acoustiques correspondants $\widetilde{\mathbf{s}}$. Ce synthétiseur joue le rôle de *plant* que l'agent va devoir apprendre à piloter pour pouvoir répéter les *stimuli* acoustiques de parole qu'il perçoit. Lors de nos simulations, le synthétiseur articulatoire, qui vise à approximer la physique de l'appareil vocal, a déjà été pré-entraîné sur un locuteur de référence et n'est pas mis à jour lors de l'apprentissage de l'agent.

Le modèle direct Le modèle direct f de l'agent est un modèle interne du *plant*, tel que décrit dans la section 1.1.2.2. Il lui permet de prédire la conséquence acoustique $\widehat{\mathbf{s}}$ de l'exécution d'une séquence de commandes articulatoires $\mathbf{a}$. À l'initialisation de l'agent, le modèle direct ne contient aucune connaissances et devra être entraîné afin de donner de bonnes estimations du résultat acoustique de commandes articulatoires.

Le modèle inverse Le modèle inverse g est lui aussi un modèle interne, et permet d'estimer la commande articulatoire $\mathbf{a}$ à envoyer au synthétiseur pour s'approcher du *stimulus* auditif perçu $\mathbf{s}$. Comme le modèle direct, ce modèle inverse est totalement naïf à l'initialisation de l'agent et doit être entraîné pour donner de bonnes estimations.

5.1.2 Données d'apprentissage

Dans ce travail de création d'un agent apprenant la parole, nous utilisons les jeux de données PB2007, GB2016 et TH2016 présentés à la section 2.2. Pour rappel, ces corpus contiennent les mêmes 1 109 locutions (principalement des séquences voyelle-consonne-voyelle mais également des mots et des phrases en français), prononcées par trois locuteurs adultes masculins différents (un locuteur par corpus). En retirant les silences, chacun de ces corpus contient environ 16 minutes de parole. Seul le premier corpus, PB2007, contient, en plus des productions acoustiques, les enregistrements des mouvements articulatoires associés. Bien que l'entraînement de l'agent n'ait pas directement recours à ces données articulatoires (au contraire, le but de l'agent est de trouver des gestes seulement à partir des sons qu'il perçoit), elles ont été utilisées pour la création du synthétiseur articulatoire qu'il pilote, selon la procédure décrite au chapitre 3. De ce fait, nous désignerons dans la suite de ce travail le locuteur du corpus PB2007 par le nom de LR pour « locuteur de référence » (puisque l'agent aura « la même voix » que celui-ci) et GB2016 et TH2016 seront désignés par L1 et L2 pour « locuteur 1 » et « locuteur 2 ».

Pour ces trois locuteurs, les enregistrements audio ont été transformés en autant de séquences de vecteurs de paramètres acoustiques. Chaque vecteur contient 18 coefficients cepstraux exprimés sur une échelle en Bark, extraits à l'aide d'une analyse en fenêtre glissante, en considérant une taille de fenêtre de 25 ms, et une *hop size* de 10 ms. Les représentations articulatoires estimées par le modèle inverse et transmises à la fois au modèle direct ainsi qu'au synthétiseur, sont des séquences de vecteurs de 6 paramètres articulatoires, tels que décrits à

la section 3.1.1 et spécifiant le degré d'activation de la langue, de la mâchoire et des lèvres.

5.1.3 Implémentation

Le synthétiseur articulatoire Le synthétiseur articulatoire ϕ suit la même implémentation que la partie du synthétiseur articulatoire présenté au chapitre 3 modélisant la relation articulatoire-vers-acoustique, elle-même présentée section 3.2. Pour rappel, il est donc constitué d'un réseau de neurones *feedforward* (perceptron multicouche) de 4 couches de 256 neurones pleinement connectées. La fonction d'activation pour les couches cachées est la tangente hyperbolique. Une *batch normalization* ainsi qu'un mécanisme de type *dropout* (avec $p = 0{,}25$) sont appliqués à chaque couche cachée. La sortie de la dernière couche cachée est ensuite dirigée vers la couche de sortie par une projection linéaire. Le synthétiseur prend en entrée un seul vecteur de paramètres articulatoires et donne en sortie le vecteur de coefficients cepstraux correspondants. Ainsi, pour traiter une séquence entière de vecteurs de paramètres articulatoires $\mathbf{a} = [\mathbf{a}_1, \cdots, \mathbf{a}_T]$, il doit traduire indépendamment chacun des vecteurs $\mathbf{a}_t$ qui la composent.

Il est important de noter que dans cette architecture d'agent proposée, le synthétiseur articulatoire ne doit pas nécessairement être implémenté à l'aide d'un réseau de neurones et pourrait très bien l'être par exemple à l'aide d'un modèle physique tel que ceux présentés en section 1.2.4.1.

Le modèle direct Le modèle direct f est implémenté également sous la forme d'un réseau de neurones *feedforward*. Bien que cela ne soit pas un pré-requis de notre approche, nous avons utilisé la même architecture que celle du synthétiseur articulatoire. Cependant, contrairement à ce dernier, les paramètres du modèle direct sont initialisés aléatoirement au début de l'apprentissage de l'agent. Ce modèle ne contient de ce fait aucune connaissance préalable sur les propriétés de l'« appareil vocal de l'agent » (son synthétiseur). L'agent doit donc les découvrir au fur et à mesure qu'il apprend à reproduire les *stimuli* auditifs qu'il perçoit pour entraîner son modèle direct.

Le modèle inverse Le modèle inverse g, contrairement aux deux autres parties du modèle, est implémenté à l'aide d'un réseau de neurones récurrent (RNN, *recurrent neural network*). Ce choix est motivé par la nature *one-to-many* de la relation acoustique-vers-articulatoire qu'il est chargé d'approximer (cf. section 1.2.4.2). Cette relation présente des ambiguïtés lors de l'inversion de certains sons en gestes, qui peuvent être partiellement résolues en considérant le contexte de ces sons. Par conséquent, le modèle inverse est implémenté à l'aide de 2 couches unidirectionnelles passé-vers-futur de LSTM (*long short-term memory*, Hochreiter et Schmidhuber, 1997). Après des expériences préliminaires (menées sur des corpus de validation), la dimension des états cachés de chaque couche de LSTM a été fixée à 32. Un mécanisme de *dropout* ($p = 0{,}25$) est également utilisé dans le but d'améliorer les capacités de généralisation du modèle. Chaque état caché à un temps t de la dernière couche LSTM est ramené dans

la dimensionnalité de l'espace des paramètres articulatoires à l'aide d'une projection linéaire. Le modèle inverse permet donc de traduire d'un bloc une séquence de T vecteurs de traits acoustiques perçue par l'agent $\mathbf{s} = [\mathbf{s}_1, \cdots, \mathbf{s}_T]$ en une séquence de T vecteurs de paramètres articulatoires $\mathbf{a} = [\mathbf{a}_1, \cdots, \mathbf{a}_T]$.

5.1.4 Apprentissage

L'apprentissage du modèle inverse et du modèle direct de l'agent se fait en suivant un algorithme basé sur le principe d'apprentissage par accommodation présenté en section 1.1.2.3. Cet algorithme d'apprentissage consiste en la répétition en boucle de cette suite d'étapes :

1. L'agent reçoit un son $\mathbf{s}$ tiré du jeu de données d'apprentissage, qu'il va tenter d'imiter.
2. Le modèle inverse est utilisé pour estimer une séquence de paramètres articulatoires $\mathbf{a}$ dont l'exécution produirait un son proche de celui perçu, avec $\mathbf{a} = g(\mathbf{s})$.
3. Le modèle direct est utilisé pour estimer le résultat acoustique $\widehat{\mathbf{s}}$ si la séquence de paramètres articulatoires inférés était exécutée, avec $\widehat{\mathbf{s}} = f(\mathbf{a})$.
4. L'agent utilise son synthétiseur articulatoire pour exécuter la séquence de paramètres articulatoires inférés, donnant le son produit $\widetilde{\mathbf{s}} = \phi(\mathbf{a})$.
5. Le modèle direct est mis à jour en rétropropageant l'erreur $\|\widetilde{\mathbf{s}} - \widehat{\mathbf{s}}\|^2$ entre l'estimation $\widehat{\mathbf{s}}$ du résultat acoustique de l'exécution du geste inféré $\mathbf{a}$ et le résultat acoustique réel $\widetilde{\mathbf{s}}$ donné par le synthétiseur pour le même geste articulatoire $\mathbf{a}$.
6. Le modèle inverse est mis à jour en rétropropageant l'erreur $\|\mathbf{s} - \widehat{\mathbf{s}}\|^2$ entre le son perçu $\mathbf{s}$ et le résultat $\hat{s}$ du chaînage du processus d'inversion (par le modèle inverse) et d'estimation acoustique (par le modèle direct, dont les poids sont ici gelés).

À noter qu'en pratique plusieurs sons $\mathbf{s}$ sont envoyés à l'agent sous forme de *mini-batchs* (de taille 8 dans le présent travail), et l'algorithme est exécuté parallèlement pour chacun de ces sons. De plus, nous faisons le choix d'arrêter la procédure d'apprentissage lorsque l'erreur d'imitation $\|\mathbf{s} - \widetilde{\mathbf{s}}\|^2$ entre le son perçu $\mathbf{s}$ et le résultat de l'imitation $\widetilde{\mathbf{s}}$ calculée sur l'ensemble de validation cesse de décroître pendant 10 *epochs* (stratégie dite d'*early stopping*).

Avant de décrire la dynamique d'apprentissage de cet algorithme d'accommodation, concentrons-nous sur deux choix de modélisation importants pris lors de l'élaboration de l'étape 6. Le premier est le choix de geler les poids du modèle direct lors de la rétropropagation du gradient de l'erreur $\|\mathbf{s} - \widehat{\mathbf{s}}\|^2$. Si le modèle direct n'était pas gelé, la suite modèle inverse-modèle direct serait alors équivalente à la suite encodeur-décodeur d'un auto-encodeur classique. De ce fait, l'espace de gestes $\mathbf{a}$ serait un espace latent non contraint et rien ne garantirait qu'il ait un sens articulatoire. En gelant les poids du modèle direct et en supposant que celui-ci soit bien entraîné, l'agent se voit contraint, pour réduire l'erreur entre $\mathbf{s}$ et $\widehat{\mathbf{s}}$, de mettre à jour son modèle inverse pour trouver des gestes articulatoires $\mathbf{a}$ adaptés pour obtenir une bonne imitation. Cette mise à jour du modèle inverse est donc guidée par la rétropropagation passant d'abord au travers du modèle direct, fournissant la dérivée $\partial\widehat{\mathbf{s}}/\partial\mathbf{a}$. Le second choix important est celui de réduire l'erreur entre le son perçu $\mathbf{s}$ et la sortie du modèle direct $\widehat{\mathbf{s}}$ plutôt que l'erreur entre $\mathbf{s}$ et la sortie du synthétiseur articulatoire $\widetilde{\mathbf{s}}$. Bien que techniquement possible et menant

probablement à de meilleurs résultats, le choix de ne pas utiliser le synthétiseur à l'étape 6 est motivé par trois raisons. (1) La rétropropagation de l'erreur au travers du synthétiseur articulatoire donnerait au modèle inverse de l'information sur la relation entre les sons et les gestes au travers du calcul de la dérivée $\partial\widetilde{\mathbf{s}}/\partial\mathbf{a}$. Or, le synthétiseur approxime la physique de la production de la parole et il est impossible d'imaginer chez l'enfant apprenant à parler un transfert direct de connaissances entre un phénomène physique et un processus cognitif tel que son modèle interne inverse. (2) Le synthétiseur articulatoire contient des connaissances sur le lien articulatoire-vers-acoustique avant même le début du processus d'apprentissage. Ainsi, son utilisation pour la mise à jour du modèle inverse supposerait que l'enfant apprenant à parler aurait déjà des connaissances sur le comportement de son conduit vocal avant même de l'avoir utilisé. Enfin, la raison (3), qui est plutôt un avantage technique qu'un choix de modélisation, est la possibilité d'utiliser n'importe quel type de synthétiseur articulatoire (par exemple basé sur des réseaux de neurones ou sur une simulation des phénomènes acoustiques physiques). En effet, si le gradient devait être calculé au travers de celui-ci, il faudrait que la relation entre sa sortie et son entrée soit différentiable, ce qui n'est pas toujours le cas pour certaines approches de synthèse articulatoire.

Discutons à présent de la dynamique observée pour cet apprentissage par accommodation des relations acoustico-articulatoires. Au début de l'apprentissage, les modèles direct et inverse sont initialisés aléatoirement et ne contiennent aucune connaissance préalable sur la relation entre les modalités acoustique et articulatoire. De ce fait, les séquences de gestes $\mathbf{a}$ données par le modèle inverse à l'étape 2 sont inférées « au hasard », résultant en une exploration aléatoire de l'espace articulatoire. Le résultat acoustique $\widetilde{\mathbf{s}}$ de l'exécution par le synthétiseur articulatoire à l'étape 4 de ces gestes $\mathbf{a}$ inférés aléatoirement donnent à l'agent de véritables exemples de la relation articulatoire-vers-acoustique. Grâce à ces exemples, cette relation peut être apprise localement par le modèle direct avec la mise à jour de ses poids à l'étape 5. En conséquence, les connaissances emmagasinées par le modèle direct au fil de l'apprentissage vont permettre au modèle inverse, à l'étape 6, de bénéficier d'un gradient d'erreur de la part du modèle direct le guidant vers l'inférence de gestes articulatoires de plus en plus adaptés à l'imitation du son perçu. Cette amélioration du modèle inverse va rendre ses inférences à l'étape 2 de moins en moins hasardeuses, faisant évoluer l'exploration de l'espace articulatoire, auparavant aléatoire, vers une exploration centrée sur des régions pertinentes pour l'imitation des sons perçus.

5.1.5 Expériences

5.1.5.1 Dynamique d'apprentissage

Évolution de l'erreur La figure 5.2 montre, pour trois instances de l'agent entraînées chacune sur un locuteur différent, l'évolution de l'erreur du modèle inverse et du modèle direct au cours de l'apprentissage (à noter que, tout au long de ce chapitre, les erreurs sont calculées à la trame, même si les processus d'apprentissage sont, eux, définis sur des séquences de trames, c'est-à-dire des sons). On remarque que l'erreur du modèle inverse g (c'est-à-dire

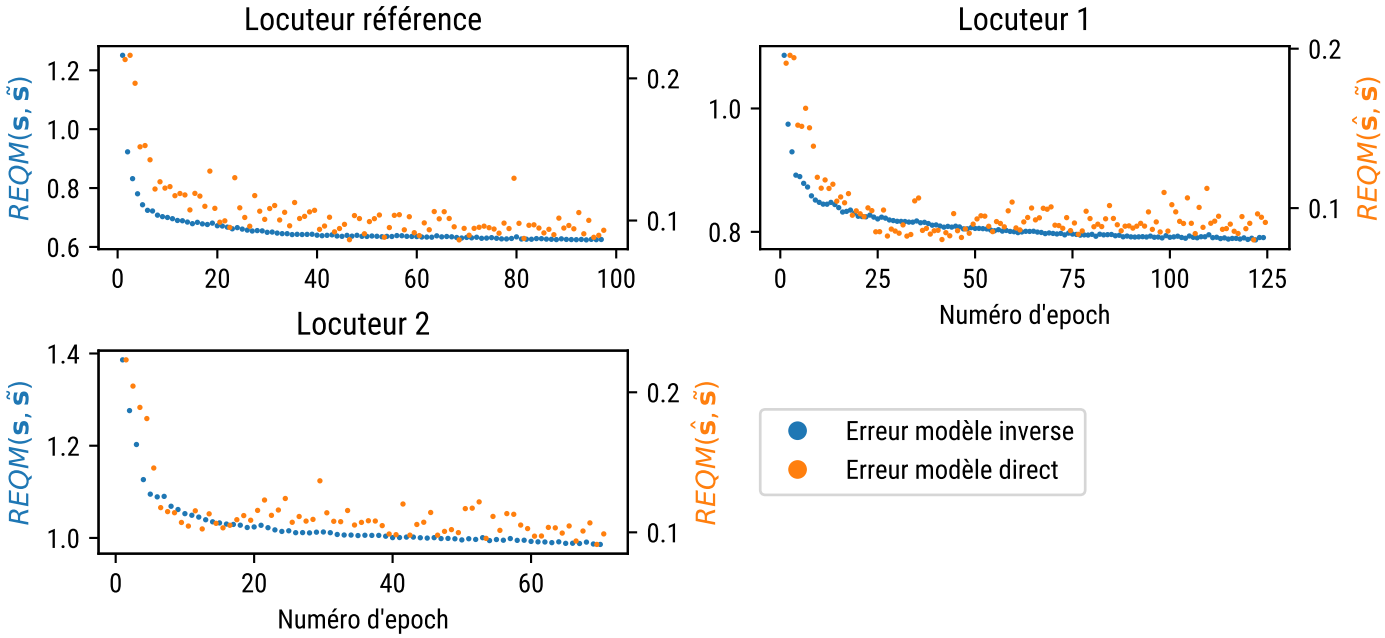

FIGURE 5.2 – **Exemples typiques de l'évolution de l'erreur des modèles inverse et direct de l'agent au cours de l'apprentissage** La racine de l'erreur quadratique moyenne (REQM) du modèle inverse est affichée en bleu sur l'axe y de gauche, celle du modèle direct en orange sur l'axe y de droite. L'axe x représente les itérations successives (*epochs*) de l'algorithme d'apprentissage.

la différence entre le son perçu par l'agent $\mathbf{s}$ et son imitation $\widetilde{\mathbf{s}}$) ainsi que l'erreur du modèle direct f (c'est-à-dire la différence entre l'estimation du résultat acoustique $\widehat{\mathbf{s}}$ de l'exécution du geste inféré $\mathbf{a}$ et le résultat acoustique réel $\widetilde{\mathbf{s}}$) diminuent toutes deux rapidement au cours de l'apprentissage, suggérant que la méthode d'apprentissage de l'agent se comporte comme attendu.

Nous pouvons également remarquer que l'évolution de l'erreur du modèle direct semble moins régulière que celle du modèle inverse. Le modèle direct est entraîné à prédire le résultat acoustique des gestes inférés $\mathbf{a}$. Or, ces gestes $\mathbf{a}$ sont inférés par le modèle inverse, qui est en constante évolution du fait de son apprentissage. De ce fait, le modèle direct est, à chaque *epoch*, entraîné à prédire le résultat acoustique de gestes différents. Cette différence dans les gestes fournis au modèle direct pourrait expliquer l'irrégularité locale de l'évolution de son erreur.

En dernier lieu, nous pouvons finalement remarquer que l'ordre de grandeur de l'erreur du modèle inverse (par exemple autour de 0,6 à la fin de l'apprentissage pour LR) et celui de l'erreur du modèle direct (autour de 0,1 pour LR) ne sont pas les mêmes. L'erreur du modèle inverse est intrinsèquement liée à la performance du synthétiseur (en effet, si par exemple le synthétiseur était très mauvais l'agent ne pourrait pas imiter convenablement les sons perçus). À l'inverse, l'erreur du modèle direct n'est, elle, pas limitée par la performance du synthétiseur. Cette différence de dépendance entre les deux erreurs pourrait ainsi expliquer l'écart entre leur

ordre de grandeur.

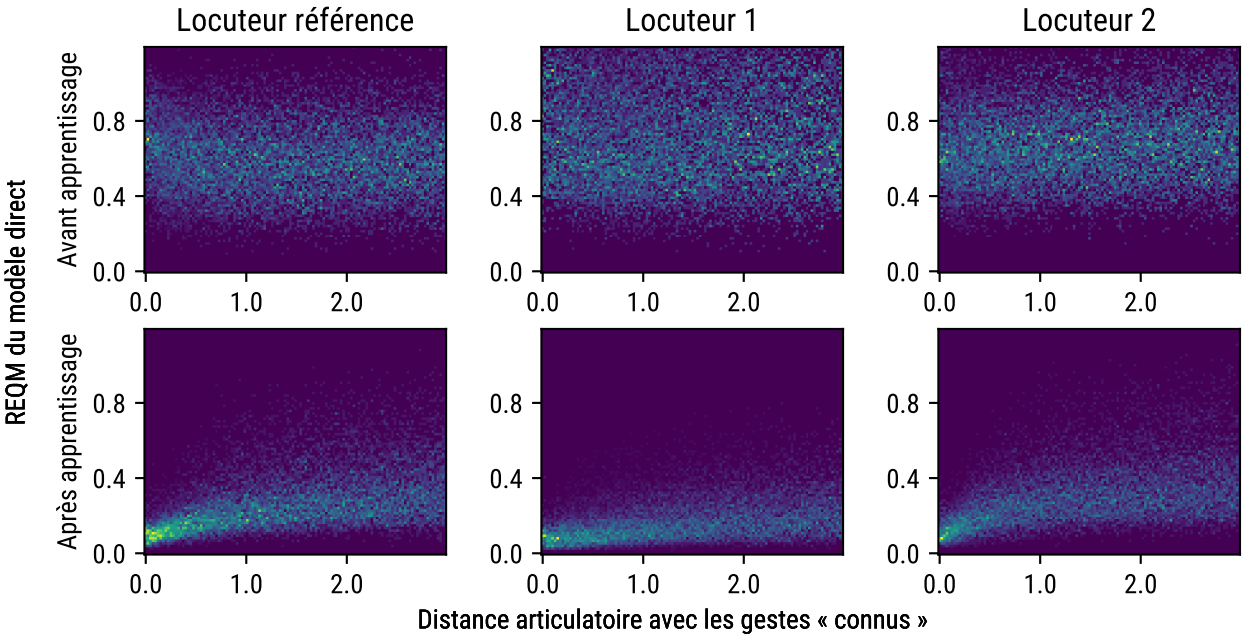

FIGURE 5.3 – **Erreur du modèle direct sur des gestes aléatoires, en fonction de leur proximité avec les gestes « connus » de l'agent, avant (en haut) et après (en bas) son apprentissage** Les gestes « connus » de l'agent sont définis comme ceux qu'il utilise pour répéter les *stimuli* auditif du corpus de test, après entraînement.

Évaluation du modèle direct La figure 5.3 permet d'évaluer plus précisément la finesse du modèle direct en début et fin d'apprentissage. À cette fin, nous avons cherché à mesurer l'erreur de prédiction du modèle direct (différence entre sa sortie $\widehat{\mathbf{s}}$ et celle du synthétiseur $\widetilde{\mathbf{s}}$ pour le même geste $\mathbf{a}$) sur des gestes sélectionnés aléatoirement et avec lesquels l'agent est plus ou moins « familier ». Pour le tirage aléatoire des gestes, nous avons fourni à l'agent, après son entraînement complet, les sons appartenant au corpus de test. Le modèle inverse a estimé pour chacun de ces sons, une séquence de vecteurs de paramètres articulatoires. Nous avons alors ajouté à chacun de ces gestes un décalage aléatoire, dont la magnitude sert à définir le « degré de familiarité » de l'agent avec ces gestes modifiés (l'agent est ainsi dit familier avec les gestes proches de ceux qu'il a inférés et moins familier avec les gestes éloignés). Nous avons ensuite fourni ces gestes aléatoirement modifiés au synthétiseur ainsi qu'au modèle direct afin de mesurer l'erreur de prédiction de ce dernier. Les erreurs de prédiction pour trois agents entraînés respectivement sur chacun des locuteurs sont présentés en bas de la figure 5.3. La partie haute affiche ces mêmes erreurs de prédiction, calculées à partir des mêmes gestes aléatoires, mais en utilisant le modèle direct des agents au début de leur apprentissage (*epoch* 1). Les résultats montrent qu'au début de l'apprentissage, le modèle direct est uniformément mauvais pour déduire les résultats acoustiques de l'exécution des gestes. En revanche, après l'apprentissage, le modèle direct est précis pour les vecteurs de paramètres articulatoires dont le décalage est faible par rapport à ceux auxquels l'agent est « familier », tandis que sa précision diminue lorsque les paramètres articulatoires s'éloignent des gestes qu'il a originellement sélectionnés.

Ce résultat suggère que l'algorithme d'apprentissage par accommodation entraîne le modèle direct en se focalisant sur des régions de l'espace articulatoire utiles pour répéter les sons qu'il a perçu.

5.1.5.2 Évaluation acoustique

L'agent a ensuite été testé sur la qualité du contenu phonémique du signal de parole produit. Pour cela nous avons utilisé dans un premier temps une évaluation objective réalisée à l'aide d'un décodeur acoustico-phonétique basé sur des HMM, puis dans un second temps une évaluation subjective par des auditeurs humains.

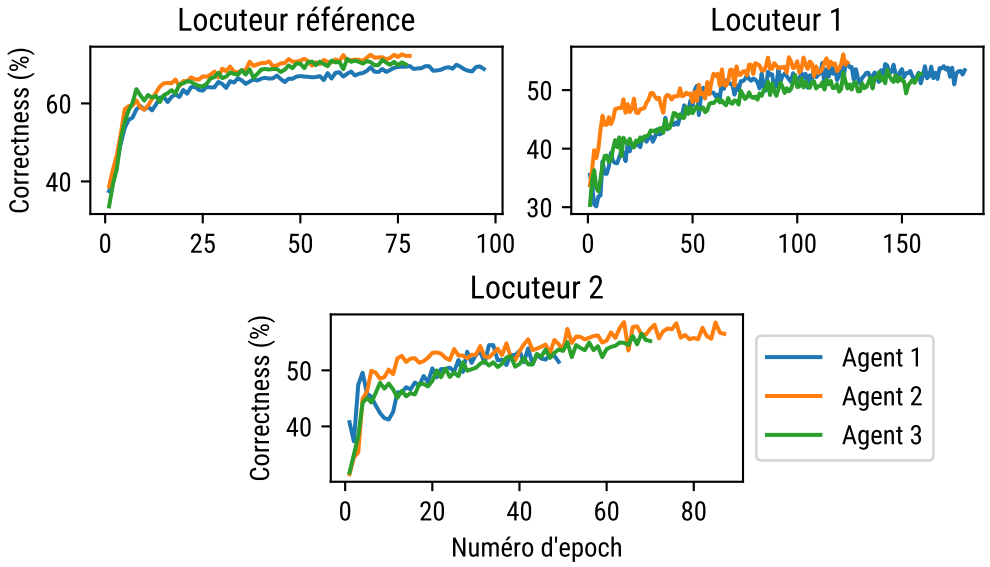

FIGURE 5.4 – **Évolution de la *correctness* du décodage au niveau phonétique des productions acoustiques réalisées par l'agent répétant des *stimuli* auditifs du locuteur de référence ou de locuteurs tierces, au fur et à mesure de son apprentissage** Chaque courbe représente l'évolution d'un agent différent.

Évaluation objective Un décodeur acoustico-phonétique basé sur un ensemble de 39 HMMs de type gauche-droite, à 3 états (identique à celui décrit en section 3.3.2.1) a d'abord été entraîné sur l'ensemble de données audio du locuteur de référence (LR), en utilisant les mêmes éléments et caractéristiques acoustiques que ceux utilisés pour entraîner les modèles direct et inverse (l'entraînement a été effectué à l'aide de la boîte à outils HTK et d'une procédure standard). Ensuite, à chaque *epoch* d'apprentissage, les imitations des sons de l'ensemble de test par l'agent, pour 9 simulations (3 simulations pour chacun des 3 locuteurs), ont été décodées au niveau phonétique. La précision du décodage est évaluée à l'aide de la métrique *correctness*[1]. Les résultats sont présentés à la figure 5.4. De façon attendue, l'évolution de la qualité phonétique est fortement corrélée à celle de l'erreur de reconstruction dans le domaine spectral, tel que présenté sur la figure 5.2. La *correctness* atteint des valeurs autour de 70 % pour

l'imitation du LR, et des valeurs plus faibles, autour de 60 %, pour les deux autres locuteurs. Cela n'est pas surprenant si l'on considère que « la voix » de l'agent (la voix avec laquelle a été construit le synthétiseur articulatoire piloté par l'agent) est celle de LR et est différente de L1 et L2, rendant plus difficile l'imitation de ces derniers. Néanmoins, l'intelligibilité globale des énoncés produits à la fin de l'entraînement est plutôt satisfaisante.

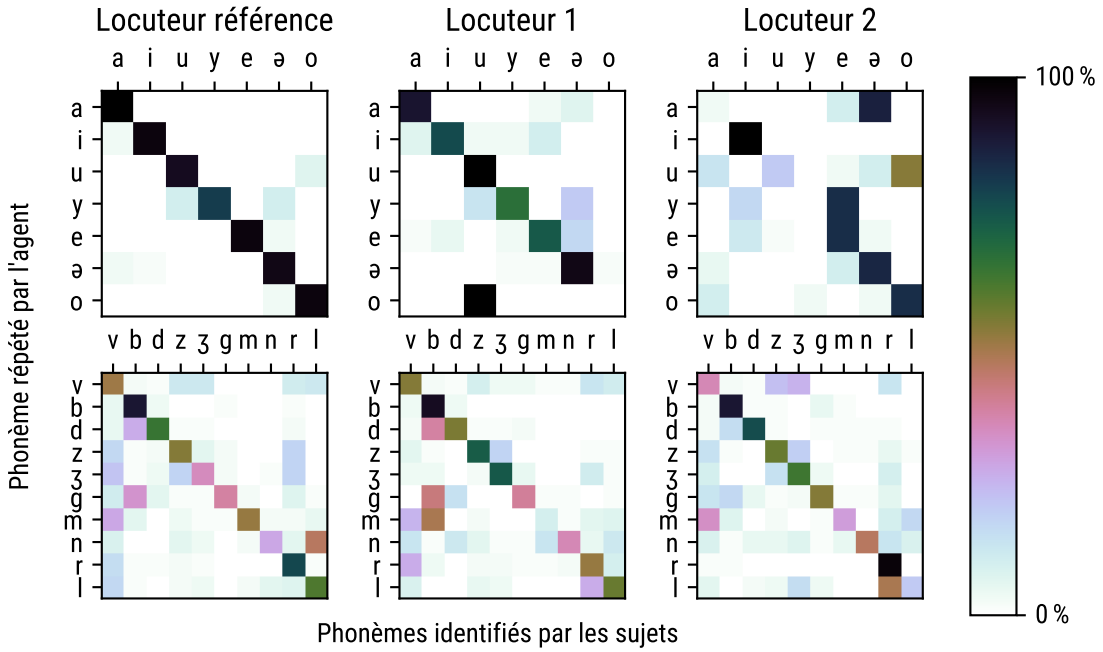

FIGURE 5.5 – **Évaluation perceptive du contenu phonétique des productions acoustiques de l'agent sur une tâche d'identification** Matrices de confusion, la vérité terrain est la séquence phonétique des *stimuli* que l'agent perçoit et doit imiter. Les voyelles /ɛ/ et /ø/ n'apparaissent pas car fusionnées respectivement avec /e/ et /ə/ *a posteriori*, du fait de leur proximité acoustique introduisant des confusions jugées comme non causées par la qualité de répétition de l'agent.

Évaluation subjective Pour mieux évaluer la qualité de l'imitation, par l'agent, de chacun des 3 locuteurs considérés dans ce travail, nous avons extrait de leur corpus respectif un ensemble de voyelles isolées /a i u y e ɛ ø ə o/ ainsi qu'un ensemble de consonnes voisées /v b d z ʒ g m n r l/ dans un contexte voyelle-consonne-voyelle VCV, avec V une des trois voyelles /a i u/. Ensuite, nous avons présenté chaque *stimulus* auditif à notre agent (en veillant à ce que chaque production ne soit pas inclue dans l'ensemble d'entraînement). La séquence de vecteurs de paramètres acoustiques résultante (vecteurs de 18 coefficients cepstraux exprimés en Bark) est convertie en une forme d'onde à l'aide du vocodeur neuronal LPCNet (présenté

1. La mesure de *correctness* a été choisie car l'*accuracy* habituellement utilisée était trop pénalisée par l'insertion de phonèmes par le décodeur lors du décodage. Nous pensons que ces insertions sont causées par une trop grande différence de stabilité entre les trajectoires acoustiques sur lesquelles a été entraîné le décodeur et celles produites par l'agent. Comme ces phonèmes insérés n'étaient pas perceptibles lors de l'écoute des répétitions (nous décrivons dans la suite comment ces répétitions sont rendues audibles) et comme la *correctness* permet de mesurer le pourcentage de phonèmes restitués, nous avons opté pour celle-ci.

en fin de section 1.2.5), en la combinant avec les paramètres de source (f0 et harmonicité) extraits du son original. L'intelligibilité des *stimuli* ainsi synthétisés, pour chacun des trois locuteurs (LR, L1, L2), a été évaluée par 20 auditeurs, à l'aide d'un test d'écoute en ligne (le recrutement des auditeurs a été réalisé à l'aide de la plate-forme *Prolific*, Palan et Schitter, 2018). Lors de ce test, les sujets devaient identifier la voyelle ou la consonne dans chaque séquence de type VCV présentée (questionnaire à choix forcé parmi la liste des voyelles ou celle des consonnes énoncées précédemment, complété par une option additionnelle « absence de consonne »). Les résultats de ce test sont présentés sur la figure 5.5 sous forme de matrices de confusions (entre d'une part, le contenu phonétique du *stimulus* auditif à répéter, et d'autre part, celui du *stimulus* produit par l'agent, identifié par les sujets du test perceptif). Pour les voyelles, la précision (moyenne des valeurs de la diagonale des matrices de confusion), pour chaque locuteur, varie de 94 % pour LR à des scores plus faibles, respectivement 72 % et 54 % pour L1 et L2. Ces scores plus faibles pour les locuteurs L1 et L2 sont probablement dus à un problème de normalisation inter-locuteur, c'est-à-dire qu'une réalisation acoustique d'un phonème par ces locuteurs peut avoir un contenu spectral proche de celui d'une réalisation acoustique d'un autre phonème par le locuteur de référence qui « donne sa voix » à l'agent (par exemple, les productions de /o/ par L1 sont proches de celles de /u/ par LR).

Les scores de reconnaissance des consonnes sont plus faibles, avec notamment des confusions entre des consonnes partageant un mode de production similaire (par exemple /g/ avec /b/, /ʒ/ avec /z/ ou /r/ avec /l/), et des performances beaucoup plus semblables entre locuteurs imités que pour les voyelles (précision de 58,48 % pour LR, 61,90 % pour L1 et 61,61 % pour L2).

5.1.5.3 Évaluation des trajectoires articulatoires

Évaluation qualitative L'observation des trajectoires articulatoires inférées par l'agent – qui prennent la forme de séquences de paramètres articulatoires –, ainsi que de leurs projections dans l'espace des bobines EMA grâce au modèle articulatoire, laisse apparaître plusieurs phénomènes. Lors de la production des voyelles, les positions adoptées sont stables et semblent correspondre à des gestes « naturels ». En revanche, les trajectoires employées pour la production des consonnes présentent quelques anomalies. Premièrement, les constrictions ne sont pas toujours réalisées aux lieux caractéristiques associés aux consonnes imitées, et ce principalement lors des instants de silence. Deuxièmement, des mouvements brusques peuvent apparaître.

Ces phénomènes apparaissent sur la figure 5.6, qui montre l'évolution des paramètres articulatoires inférés, ainsi que leur projection dans l'espace des bobines EMA à trois instants clés, lors de l'imitation d'un /øgø/ produit par le locuteur L2. Le premier instant mis en évidence (à 230 ms) montre la configuration articulatoire adoptée lors de la production de la voyelle /ø/. La configuration adoptée semble pertinente, avec une langue vers l'avant et des lèvres protruses (*LipProtrusion* proche de 1). Le second instant mis en évidence (à 360 ms) correspond au moment précédant l'occlusion complète du conduit lors de la production du /g/. À cet instant, l'arrière de la langue est presque au contact du palais grâce à l'action

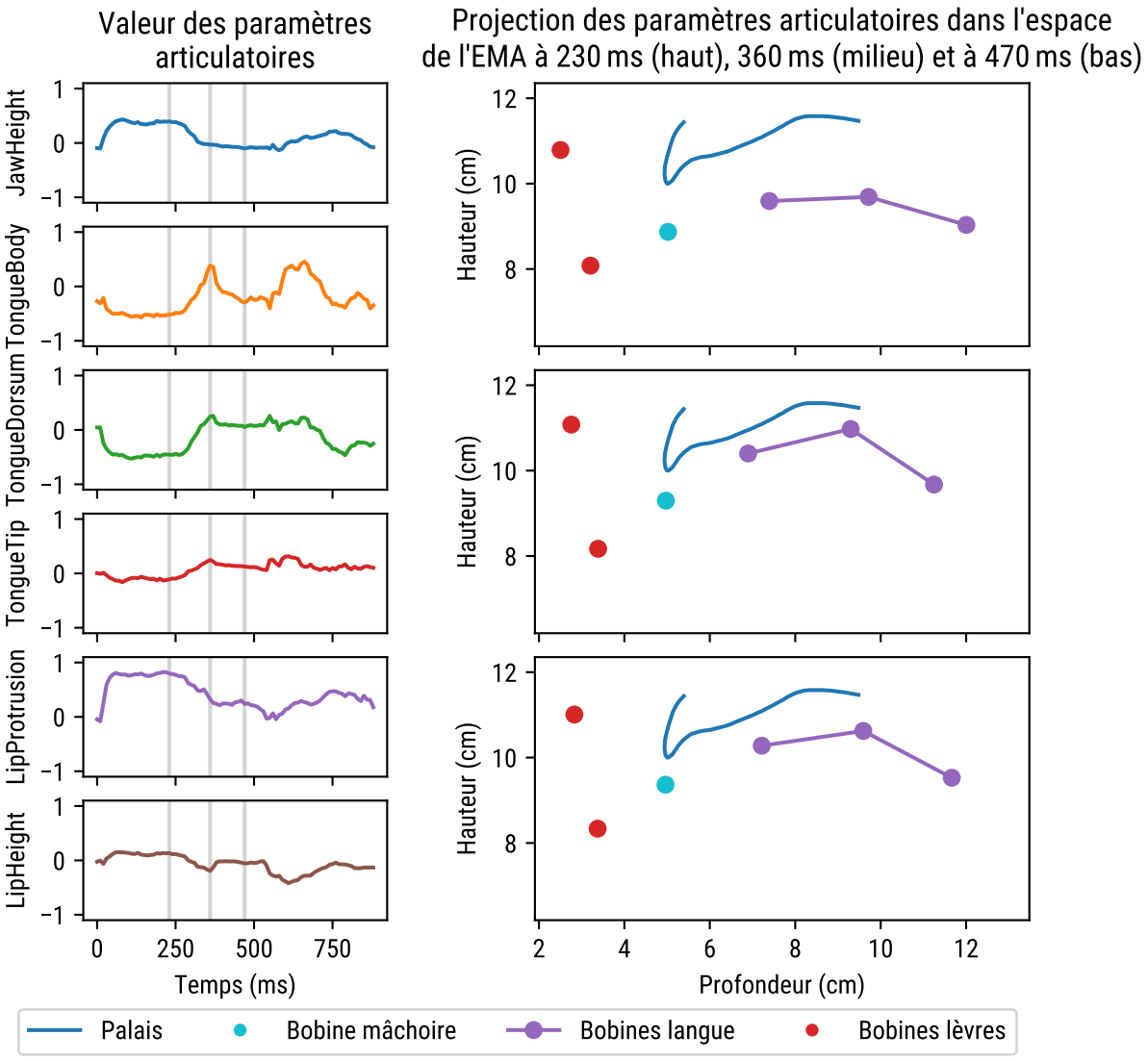

FIGURE 5.6 – **Évolution des paramètres articulatoires inférés par l'agent lors de la répétition d'un øgø prononcé par L2 ainsi que leur projection dans l'espace des bobines EMA à trois instants clés** À 230 ms : production du ø. À 360 ms : instant précédant l'occlusion lors de la production du g. À 470 ms : instant correspondant au « silence voisé » de la production du g.

conjointe des paramètres *TongueBody* et *TongueDorsum*, ce qui concorde bien avec le geste caractéristique associé à la production de cette consonne. En revanche, au troisième instant de temps (à 470 ms), qui correspond à l'instant où la langue est contre le palais lors de la production classique du /g/, nous pouvons remarquer que l'agent n'a pas adopté cette stratégie. Au contraire, il semblerait que l'occlusion ait plutôt été réalisée au niveau des incisives, avec un mouvement de l'avant de la langue (correspondant à la diminution nette du paramètre *TongueBody*), qui ne correspond pas à ce qui est attendu ici. Enfin, lorsque l'on regarde la valeur des paramètres articulatoires autour de 660 ms – ce qui correspond à l'instant de la production du /g/ où l'occlusion est relâchée –, l'agent a adopté une position semblable à celle précédant l'occlusion (à 360 ms), où l'arrière de la langue est proche du palais (remontée du paramètre *TongueBody*). Ainsi, le geste employé par l'agent pour la réalisation du /g/ semble bien correspondre à la stratégie classique associée à la production de cette consonne, sauf lors de l'occlusion, qui est réalisée à un autre endroit. Ce résultat peut être expliqué par la nature uniquement acoustique de l'objectif poursuivit par l'agent lors de son apprentissage, et qui a pu le pousser à employer des occlusions satisfaisantes sur le plan acoustique mais incohérentes avec des stratégies articulatoires « normales ».

Évaluation quantitative Afin d'évaluer si les trajectoires articulatoires inférées par l'agent peuvent être une bonne base pour la découverte d'unités de la parole, nous avons appliqué le même protocole que celui décrit dans la section 4.2, mais en remplaçant les données articulatoires enregistrées sur des sujets humains par celles inférées par l'agent à partir des données acoustiques associées. Pour rappel, cette méthode consiste à entraîner un VQ-VAE soit sur des données acoustiques soit sur des données articulatoires. Pendant cet entraînement, le VQ-VAE va apprendre de façon auto-supervisée des représentations latentes discrètes de ces modalités. Ces représentations sont ensuite évaluées grâce à une méthodologie basée sur des tests ABX afin de mesurer à quel point elles permettent de discriminer les consonnes les unes des autres. Le résultat de cette évaluation est traduit en deux scores de discriminabilité : le premier décrivant à quel point ces représentations permettent de discriminer entre elles les consonnes partageant le même mode d'articulation (e.g. occlusives, fricatives, affriquées) et le second de discriminer celles partageant le même lieu d'articulation (e.g. labiales, dentales, palatales).

Les résultats de cette évaluation sont affichés figure 5.7. Les scores de discriminabilité montrent que les représentations issues des données articulatoires inférées par l'agent par inversion acoustico-articulatoire (affichées en bleu sur la figure) ont des propriétés bien différentes de celles de représentations obtenues à partir de données articulatoires réelles (affichées en rouge pour LR). En effet, comme nous l'avons vu dans la section 4.2, les représentations articulatoires « réelles » fournissent un score de discriminabilité en terme de lieu d'articulation supérieur aux représentations issues de la modalité acoustique (autour de 80 % vs 60 %) et inversement, fournissent un score inférieur en terme de mode d'articulation (autour de 60 % vs 90 %). Cette tendance n'est cependant pas visible sur les données articulatoires obtenues par inversion du signal acoustique, ou alors dans une bien moindre mesure. En effet, par rapport aux représentations acoustiques des locuteurs LR et L1, les représentations articulatoires « de l'agent » n'apportent qu'un gain très faible en terme de discriminabilité sur le lieu d'articulation (inférieur à 5 %) et une perte du même ordre en terme de mode d'articulation. Cette

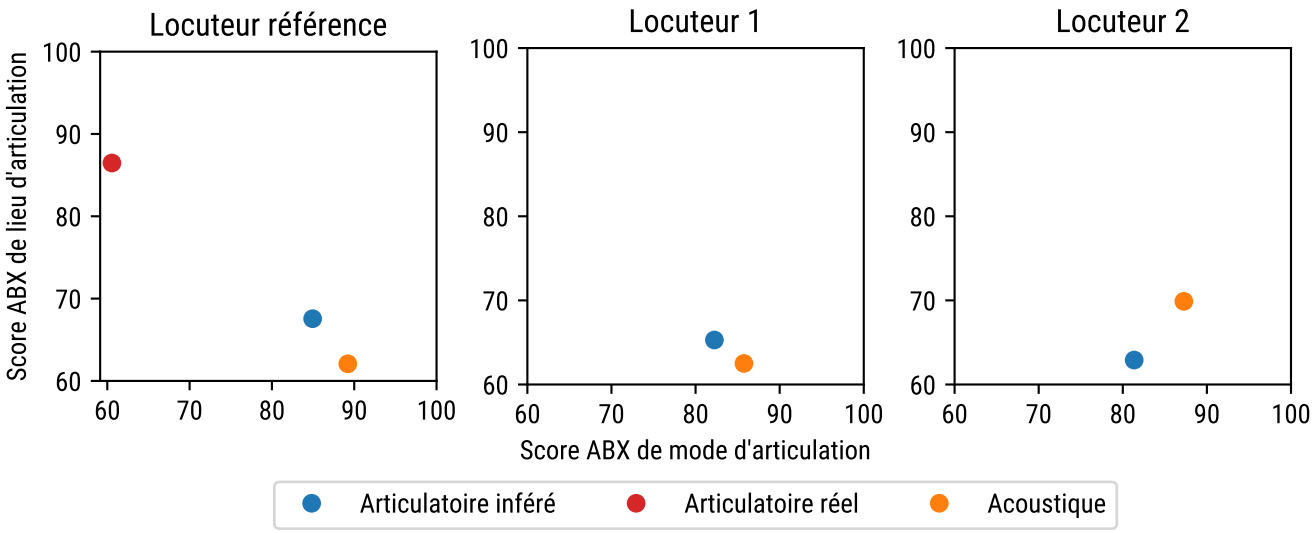

FIGURE 5.7 – **Scores ABX en fonction du lieu et du mode d'articulation des représentations issues des données articulatoires inférées par l'agent et des données acoustiques réelles** À des fins de comparaison, les scores associés à des représentations extraites des données articulatoires réelles du LR sont affichés (en rouge).

différence de discriminabilité n'apparaît cependant pas pour le locuteur L2. Ceci pourrait s'expliquer par le fait que les représentations acoustiques de ce locuteur fournissent déjà un score élevé de discriminabilité en terme de lieu d'articulation (par rapport à celles des locuteurs LR et L1), modérant de fait le gain potentiel apporté par les représentations « de l'agent ». Cette explication est également corroborée par la similarité des scores des représentations articulatoires inférées du locuteur L2 avec ceux des locuteurs LR et L1 (points bleus sur la figure), suggérant que l'apport des représentations articulatoires inférées est similaire entre locuteurs, et que ce serait plutôt le score intrinsèque des représentations acoustiques pour L2 qui serait différent des scores pour LR et L1 (points oranges sur la figure).

Ces résultats indiquent que les trajectoires articulatoires obtenues par inversion du signal acoustique restent éloignés de mouvements « réels », et n'apportent donc pas le même type d'informations. Bien que décevant, ce phénomène peut s'expliquer par deux raisons. La première est que l'inférence des trajectoires par l'agent n'est pas contrainte de façon à donner des gestes « réalistes » (par exemple avec un critère de minimisation d'un coût énergétique, comme décrit section 1.1.2.1), mais a pour seul objectif de permettre *in fine* la meilleure synthèse sonore. Autrement dit, rien n'empêche l'agent de changer très rapidement de configuration articulatoire sans aucun respect de la biomécanique de son appareil vocal, et créer ainsi des trajectoires peu plausibles (mais restant efficaces d'un point de vue acoustique), comme nous l'avons illustré sur la figure 5.6. La seconde provient de l'implémentation même du synthétiseur articulatoire. En effet, ce dernier est un réseau de neurones entraîné sur un nombre limité de données de paroles (de l'ordre d'une quinzaine de minutes). À cause de cela, le synthétiseur est forcé d'extrapoler lorsque des configurations articulatoires éloignées de celles contenues dans son corpus d'apprentissage lui sont présentées, sans garantie que ces extrapolations soient cohérentes. Ainsi, l'agent peut potentiellement trouver et exploiter ces

zones afin d'améliorer ses productions au niveau acoustique, encore une fois au détriment de la plausibilité des gestes exécutés. Ce phénomène, s'il existe, devrait cependant apparaître dans une moindre mesure pour le locuteur LR, dont la voix a été utilisée pour construire le synthétiseur, et donc pour lequel l'agent aurait potentiellement moins besoin d'extrapoler afin de mieux l'imiter.

5.1.6 Conclusion générale sur l'agent à but imitatif

Nous avons présenté une nouvelle architecture d'agent apprenant de façon auto-supervisée à imiter des sons de parole de plusieurs locuteurs, en pilotant un synthétiseur articulatoire. L'architecture de cet agent comprend, en plus du synthétiseur, un modèle inverse chargé de retrouver les gestes qui ont pu être à l'origine d'un son de parole perçu, ainsi qu'un modèle direct chargé de prédire quelles seraient les conséquences acoustiques de l'exécution d'un geste de parole. Ceux-ci sont implémentés à l'aide de réseaux de neurones afin de pouvoir gérer la complexité de données issues du monde réel. Un algorithme d'apprentissage tirant parti du mode d'entraînement des réseaux de neurones a également été proposé, et permet l'apprentissage simultané des deux modèles internes inverse et direct, seulement à partir d'enregistrements acoustiques de parole.

Les résultats montrent que l'algorithme d'apprentissage est efficace et qu'il permet aux modèles inverse et direct d'acquérir des connaissances sur la relation entre les modalités articulatoire et acoustique. Suite à cet apprentissage, les imitations de l'agent restituent une partie satisfaisante du contenu phonémique du signal de parole original. Cependant, les gestes inférés n'apportent que peu d'informations sur les lieux d'articulations des consonnes par rapport à la modalité acoustique, contrairement aux gestes originaux enregistrés.

Enfin, la façon de communiquer de cet agent – en imitant les sons de parole qu'il perçoit – reste probablement très éloignée de la nôtre. Une première étape pour la rendre plus réaliste pourrait être l'introduction d'un mécanisme permettant à l'agent de construire, en même temps qu'il apprend ses modèles internes inverse et direct, des unités de la parole. Ce faisant, il pourrait répéter les sons de parole non plus en cherchant à s'en rapprocher le plus possible acoustiquement mais plutôt en cherchant à produire un son associé aux mêmes unités que celui perçu. C'est cet objectif qui motive le développement d'un nouvel agent, appelé « agent à but communicatif », que nous décrivons dans la section suivante.

5.2 Agent à but communicatif

Dans cette partie, nous proposons une mise à jour de l'agent à but imitatif, baptisée « agent à but communicatif ». L'objectif poursuivi par ce nouvel agent au cours de son apprentissage est triple : il doit (1) se construire un répertoire d'unités discrètes représentant la parole de façon compacte, (2) apprendre à reconnaître ces unités dans les sons qu'il perçoit et (3) apprendre à produire des sons contenant les mêmes unités que celles perçues, toujours en

pilotant un synthétiseur articulatoire. Pour permettre à ce nouvel agent d'apprendre de façon auto-supervisée à accomplir cet objectif, cette mise à jour introduit un mécanisme basé sur les VQ-VAE et permettant à l'agent de construire deux répertoires d'unités, à partir des modalités acoustique et articulatoire.

Comme nous le verrons, malgré l'introduction de ce mécanisme de découverte d'unités et de ce nouvel objectif d'apprentissage, cette mise à jour de l'agent souffre encore de problèmes similaires à sa version précédente concernant l'inférence de trajectoires articulatoires par son modèle inverse, qui ne présentent pas les mêmes propriétés que des trajectoires articulatoires réelles. Ainsi, suite à la présentation de ce nouvel agent, nous ne chercherons pas dans cette section à l'évaluer en profondeur, mais plutôt à tenter de résoudre ces problèmes d'inversion en proposant de nouveaux ajustements. De ce fait, cette partie est constituée d'une série de travaux exploratoires dont le but est d'inspecter des pistes d'amélioration de l'agent à but communicatif plutôt que d'en proposer une version aboutie.

5.2.1 Architecture

5.2.1.1 Vue d'ensemble

L'architecture de l'agent à but communicatif, présentée figure 5.8, se base sur celle de l'agent à but imitatif et y ajoute deux éléments : un VQ-VAE chargé de la découverte d'unités à partir des sons de parole perçus par l'agent (que nous appellerons par la suite « VQ-VAE acoustique ») et un autre VQ-VAE chargé, lui, de la découverte d'unités à partir des trajectoires articulatoires inférées par l'agent (que nous appellerons ensuite « VQ-VAE articulatoire »). Pour rappel, le VQ-VAE a été présenté en section 1.3.1.4 et détaillé en section 4.2.1.

Le rôle des modèles inverse et direct ainsi que sur celui du synthétiseur articulatoire sont les mêmes que pour l'agent à but imitatif, décrits en section 5.1.1 – mais la procédure d'apprentissage est bien évidemment différente pour le modèle inverse, nous le verrons.

VQ-VAE acoustique Le VQ-VAE acoustique permet d'associer à chacun des instants de temps t d'un son perçu $\mathbf{s}$ une représentation latente discrète o_t^s, que nous qualifions d'« unité acoustique ». Pour trouver l'unité acoustique à un instant t de la séquence de trames acoustiques $\mathbf{s}$, le VQ-VAE encode la sous-séquence de $2\tau + 1$ trames $\mathbf{s}_{t\pm\tau} = [\mathbf{s}_{t-\tau}, \cdots, \mathbf{s}_{t+\tau}]$ centrée sur la trame d'intérêt $\mathbf{s}_t$. La sortie de l'encodeur $z_e^s(\mathbf{s}_{t\pm\tau})$ est ensuite comparée à chacun des vecteurs du dictionnaire de plongements $\mathbf{e}^s \in \mathbb{R}^{K\times D}$ du VQ-VAE acoustique (qui contient K vecteurs de dimension D). Le vecteur du dictionnaire de plongements dont la distance euclidienne avec la sortie de l'encodeur $z_e^s(\mathbf{s}_{t\pm\tau})$ est la plus petite fournit l'unité acoustique décodée $\mathbf{o}_t^s$. Dans la suite de ce chapitre, à des fins de clarté de présentation, nous identifions l'unité acoustique décodée $\mathbf{o}_t^s$ avec le vecteur de plongement représentant cette unité. Pour traiter une séquence de trames acoustiques $\mathbf{s}$ entière, cette procédure est répétée pour chacun de ses instants t, donnant la séquence d'unités acoustiques $\mathbf{o}^s$ de même longueur que $\mathbf{s}$.

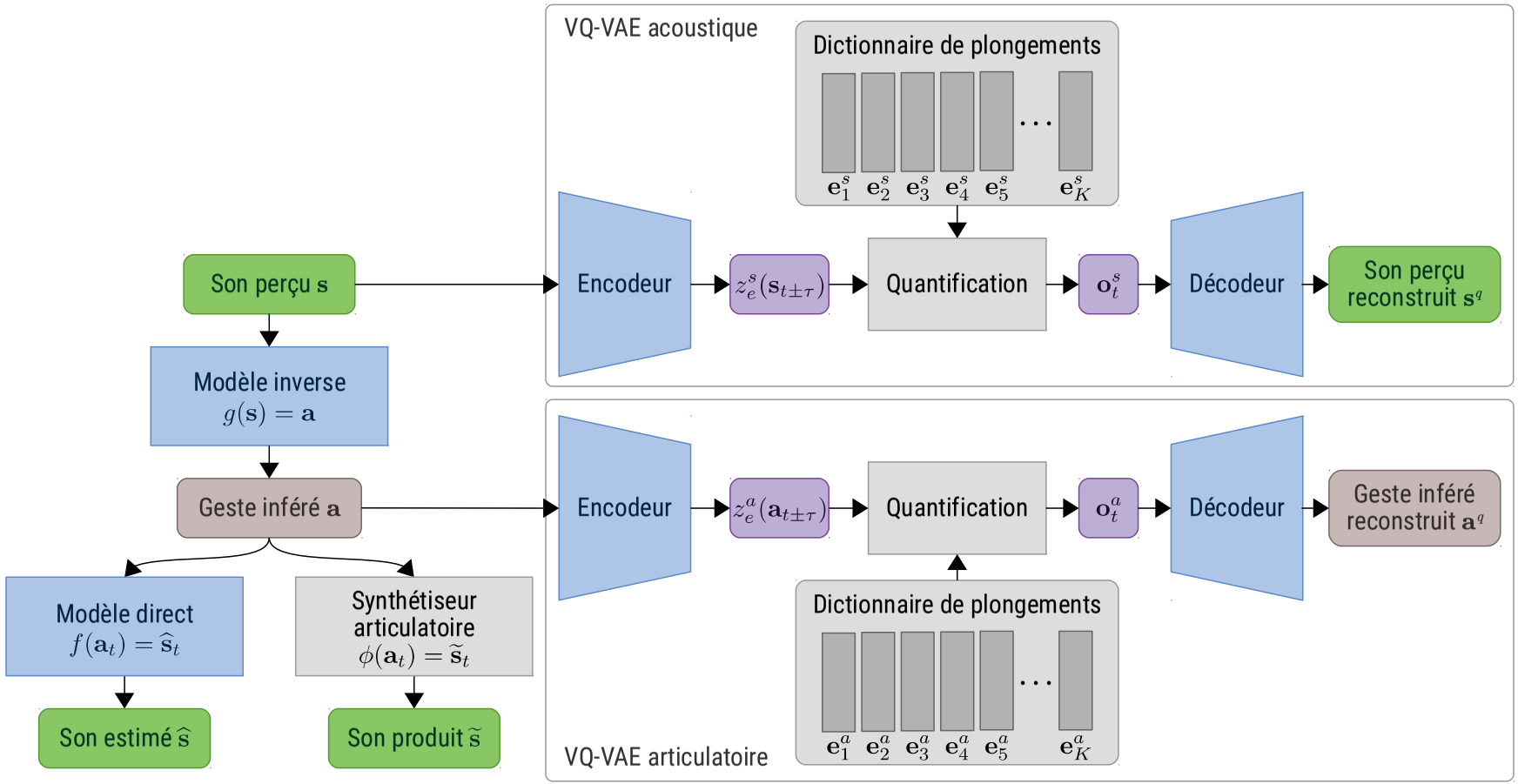

FIGURE 5.8 – **Architecture de l'agent à but communicatif** Par souci de lisibilité, l'envoi de la sortie du modèle direct dans le VQ-VAE acoustique, nécessaire pour l'entraînement du modèle inverse, n'est pas affichée.

Au début de l'apprentissage, le VQ-VAE acoustique est vierge de toute connaissance. Son répertoire d'unités acoustiques (son dictionnaire de plongements) est d'abord initialisé aléatoirement et les valeurs des vecteurs qu'il contient – ainsi que celles des poids de son encodeur et de son décodeur – sont ajustées au cours de l'apprentissage de l'agent. Ces ajustements se font en suivant la procédure d'apprentissage classique des VQ-VAE, qui consiste ici à entraîner le VQ-VAE acoustique à fournir en sortie la reconstruction $\mathbf{s}^q_{t\pm\tau}$ la plus fidèle possible de chaque paquet $\mathbf{s}_{t\pm\tau}$ extrait pour chacun des instants t des sons $\mathbf{s}$ perçus par l'agent.

VQ-VAE articulatoire Le VQ-VAE articulatoire permet d'associer à chacun des instants t d'une trajectoire articulatoire $\mathbf{a}$ une représentation latente discrète $\mathbf{o}^a_t$, que nous qualifions d'« unité articulatoire ». Son fonctionnement est le même que celui du VQ-VAE acoustique. Il permet ainsi d'associer une unité $\mathbf{o}^a_t$ à un instant t d'une séquence de paramètres articulatoires $\mathbf{a}$ en traitant la sous-séquence $\mathbf{a}_{t\pm\tau}$. La sortie de son encodeur $z^a_e(\mathbf{a}_{t\pm\tau})$ pour ce paquet est ensuite comparée à chacun des vecteurs de son dictionnaire de plongements $\mathbf{e}^a \in \mathbb{R}^{K\times D}$, et le plongement le plus proche de cette sortie au sens de la distance euclidienne est sélectionné (notons que, comme au chapitre précédent, nous utilisons les mêmes valeurs de D et K pour les deux VQ-VAE, articulatoire et acoustique). Ce vecteur de plongement fournit l'unité articulatoire $\mathbf{o}^a_t$.

Tout comme le VQ-VAE acoustique, le répertoire d'unités du VQ-VAE articulatoire est d'abord initialisé aléatoirement et est mis à jour au cours de l'entraînement de l'agent. L'objectif d'entraînement du VQ-VAE articulatoire est de fournir en sortie la reconstruction $\mathbf{a}^q_{t\pm\tau}$ la plus fidèle possible de chaque sous-séquence $\mathbf{a}_{t\pm\tau}$ extraite pour chacun des instants t des trajectoires articulatoires $\mathbf{a}$ inférées par l'agent.

5.2.1.2 Données d'apprentissage

Les corpus de données et les noms adoptés pour les désigner sont les mêmes que ceux employés pour l'agent à but imitatif, comme décrit en section 5.1.2. Pour rappel, ces jeux de données sont PB2007, GB2016 et TH2016, et sont ici respectivement désignés par les noms LR (pour « locuteur référence »), L1 et L2 (pour « locuteur 1 » et « locuteur 2 »). Le corpus PB2007 est appelé « locuteur référence » car, comme pour le précédent travail portant sur l'agent à but imitatif, c'est sur celui-ci qu'est construit le synthétiseur articulatoire.

Le type de représentation utilisé pour décrire les modalités acoustique et articulatoire est cependant différent. En effet, les signaux acoustiques sont ici représentés sous la forme de Mel-spectogrammes (l'analyse acoustique s'effectue en utilisant la librairie Python *librosa*, en considérant des fenêtres de tailles 25 ms, une *hop size* de 10 ms, et une quantification du spectre d'amplitude sur 40 filtres triangulaires répartis sur l'échelle de Mel). Les trajectoires articulatoires sont, elles, décrites par des séquences de vecteurs contenant les 6 mêmes paramètres articulatoires (détaillés en section 3.1.1 et décrivant le degré d'activation de la langue, de la mâchoire et des lèvres) auxquels sont adjoints 2 paramètres décrivant la hauteur et le degré d'harmonicité (extraits avec la même procédure que celle utilisée dans le cadre du vocodeur

LPCNet). À nouveau, il est important de noter que ces trajectoires articulatoires ne sont utilisées que pour la création du synthétiseur articulatoire et que l'agent n'y aura jamais accès directement. Ce changement de types de représentation par rapport à ceux employés pour l'agent à but imitatif a été décidé afin de donner à ce nouvel agent un accès plus explicite aux informations liées à la source glottique des sons perçus $\mathbf{s}$, inférés $\widehat{\mathbf{s}}$ et produits $\widetilde{\mathbf{s}}$, et un contrôle sur celle des sons produits (via l'ajout des 2 paramètres de source à l'espace articulatoire $\mathbf{a}$, qui est l'espace de contrôle du synthétiseur).

5.2.1.3 Implémentation

L'implémentation interne des modèles inverse et direct ainsi que celle du synthétiseur articulatoire de l'agent à but communicatif est identique à celle utilisée pour l'agent à but imitatif et présentée en section 5.1.3. Ainsi, seule la dimension des couches d'entrée et de sortie des modèles est adaptée à la nouvelle représentation des données (i.e. vecteurs de Mel-spectre d'amplitude de dimension 40 pour la modalité acoustique, vecteurs contenant 6 paramètres de filtre et 2 paramètres de source pour la modalité articulatoire).

L'architecture du VQ-VAE acoustique est directement inspirée de celle implémentée dans la section 4.2.2.1. L'encodeur et le décodeur du VQ-VAE acoustique sont tous deux implémentés à l'aide de 3 couches de 256 neurones pleinement connectées. La fonction d'activation utilisée pour les couches cachées est la tangente hyperbolique. Une *batch normalization* ainsi qu'un mécanisme de *dropout* (avec $p = 0{,}25$) sont utilisés pour chaque couche cachée. Le VQ-VAE acoustique auto-encode des sous-séquences de 5 trames consécutives ($\tau = 2$, représentant un empan temporel de 50 ms). Les sous-séquences centrées sur les τ premières et dernières trames des sons sont formées à l'aide d'un *padding* en miroir. Le dictionnaire de plongements comporte 64 vecteurs de dimension 32. La valeur du terme β de pondération d'un des termes de la fonction de coût spécifique au VQ-VAE est fixée à 0,25 (de façon similaire à van den Oord et al., 2017).

L'implémentation technique du VQ-VAE articulatoire est, elle, identique à celle du VQ-VAE acoustique à la différence près des dimensions de son entrée et de sa sortie, qui sont adaptées afin de correspondre à la dimensionnalité de l'espace des paramètres articulatoires.

5.2.1.4 Apprentissage

L'apprentissage complet de l'agent à but communicatif se déroule en deux temps. Le premier temps est une procédure d'entraînement pour les modèles inverse et direct ainsi que pour le VQ-VAE acoustique et le second temps est une autre procédure d'entraînement, cette fois pour le VQ-VAE articulatoire. Afin d'être plus en accord avec l'apprentissage réel de la parole, nous pourrions combiner ces deux phases de découverte d'unités acoustiques et articulatoires en une seule. En effet, il est peu probable que la découverte d'unités chez l'enfant se fasse en deux phases aussi nettement séparées. Cependant, comme nous le verrons, la découverte des unités articulatoires par l'agent (l'entraînement de son VQ-VAE articulatoire) n'impacte pas,

dans la version actuelle, l'entraînement des autres parties du modèle. De ce fait, nous pouvons nous permettre de l'entraîner une fois le reste du modèle stabilisé, et obtenir un résultat final similaire à celui d'un entraînement simultané tout en économisant des ressources de calcul.

Apprentissage du VQ-VAE acoustique et des modèles inverse et direct L'apprentissage conjoint des modèles inverse et direct, ainsi que du VQ-VAE acoustique, s'effectue à l'aide d'une procédure proche de celle déjà décrite pour l'agent à but imitatif (voir section 5.1.4). Elle consiste à itérer la succession d'étapes suivante :

1. L'agent reçoit un son $\mathbf{s}$ tiré du jeu de données d'apprentissage, qu'il va tenter de répéter.
2. Pour chacun des instants t du son perçu $\mathbf{s}$, la sous-séquence de trames $\mathbf{s}_{t\pm\tau}$ est présentée au VQ-VAE acoustique dont l'encodeur et le module de quantification infèrent l'unité acoustique $\mathbf{o}^s_t$ et dont le décodeur infère la reconstruction $\mathbf{s}^q_{t\pm\tau}$ de $\mathbf{s}_{t\pm\tau}$.
3. Les poids du VQ-VAE acoustique ainsi que son dictionnaire de plongements (le dictionnaire d'unités acoustiques) sont mis à jour en minimisant la fonction de coût spécifiée section 4.2.1.
4. Le modèle inverse est utilisé pour estimer quelle a pu être la séquence de paramètres articulatoires $\mathbf{a}$ à l'origine du son perçu, avec $\mathbf{a} = g(\mathbf{s})$ (inversion acoustico-articulatoire dans l'espace du locuteur de référence).
5. Le modèle direct est utilisé pour estimer le résultat acoustique $\widehat{\mathbf{s}}$ si la séquence de paramètres articulatoires inférés était exécutée, avec $\widehat{\mathbf{s}} = f(\mathbf{a})$ (internalisation du processus physique de production de la parole).
6. L'agent utilise son synthétiseur articulatoire (le *plant*) pour exécuter la séquence de paramètres articulatoires inférés par le modèle inverse, produisant le signal de parole synthétique $\widetilde{\mathbf{s}} = f(\mathbf{a})$.
7. Le modèle direct est mis à jour en rétropropageant l'erreur $\|\widetilde{\mathbf{s}} - \widehat{\mathbf{s}}\|^2$ entre l'estimation $\widehat{\mathbf{s}}$ du résultat acoustique de l'exécution du geste inféré $\mathbf{a}$ et le résultat acoustique réel $\widetilde{\mathbf{s}}$ donné par le synthétiseur.
8. Pour chacun des instants t de l'estimation du résultat acoustique $\widehat{\mathbf{s}}$, la sous-séquence de trames $\widehat{\mathbf{s}}_{t\pm\tau}$ est envoyée dans l'encodeur du VQ-VAE acoustique, qui donne en sortie $z^s_e(\widehat{\mathbf{s}}_{t\pm\tau})$.
9. Le modèle inverse est mis à jour en rétropropageant l'erreur $\|\mathbf{o}^s - z^s_e(\widehat{\mathbf{s}})\|^2$ cumulant sur l'ensemble des trames du son présenté l'erreur entre l'unité acoustique $\mathbf{o}^s_t$ extraite du son perçu $\mathbf{s}$ par le VQ-VAE à l'instant t, et le résultat $z^s_e(\widehat{\mathbf{s}}_{t\pm\tau})$ du chaînage des processus d'inversion (par le modèle inverse), d'estimation acoustique (par le modèle direct, dont les poids sont ici gelés) et d'encodage par l'encodeur du VQ-VAE acoustique (dont les poids sont ici aussi gelés) au même instant. Ce processus est commenté plus en détail ci-dessous.

À noter qu'en pratique plusieurs sons $\mathbf{s}$ sont envoyés simultanément à l'agent sous forme de *mini-batches* (de taille 8 dans le présent travail), et la procédure est exécutée parallèlement pour chacun de ces sons. Cette procédure est répétée jusqu'à ce que l'entraînement du VQ-VAE acoustique et des modèles inverse et direct soit jugé complet. Dans ce travail, un critère d'*early*

stopping est utilisé pour arrêter l'apprentissage lorsque la moyenne des erreurs $\|\mathbf{o}^s - z_e^s(\widehat{\mathbf{s}})\|^2$ calculée sur un ensemble de validation cesse de décroître pendant 25 *epochs* après un minimum de 50 *epochs*.

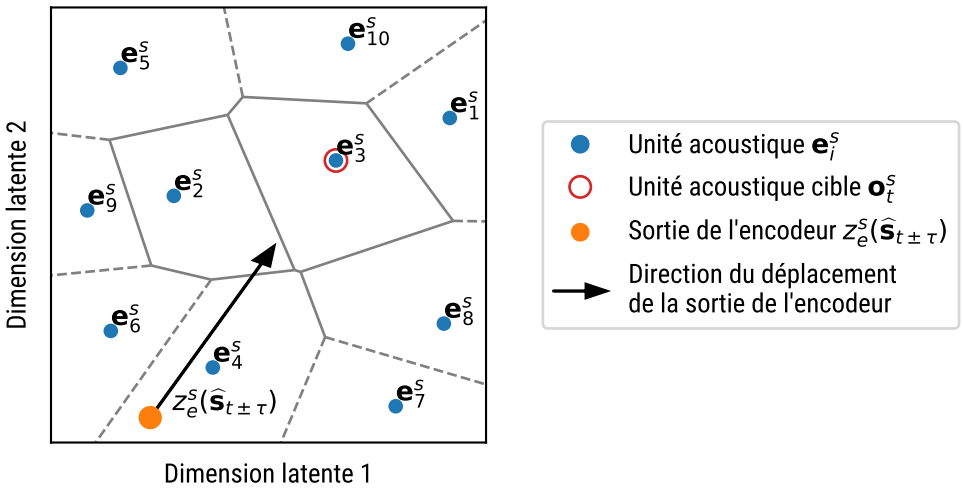

FIGURE 5.9 – **Illustration en 2 dimensions de l'effet de la rétropropagation de l'erreur $\|\mathbf{o}_t^s - z_e^s(\widehat{\mathbf{s}}_{t\pm\tau})\|^2$ dans l'espace de sortie de l'encodeur du VQ-VAE acoustique** Les points bleus représentent les vecteurs de plongements $\mathbf{e}_i^s$ découverts par l'agent, qui correspondent aux différentes unités de la parole. Les lignes autour de chacun de ces vecteurs de plongement délimitent les portions de l'espace de sortie de l'encodeur qui leur sont associées. La rétropropagation de l'erreur $\|\mathbf{o}_t^s - z_e^s(\widehat{\mathbf{s}}_{t\pm\tau})\|^2$ va déplacer la sortie de l'encodeur dans la direction indiquée par la flèche. Cette sortie finira par atteindre la cellule associée à l'unité cible, même si au départ de l'algorithme elle se trouve plus proche d'une autre unité.

Avant de décrire l'apprentissage du VQ-VAE articulatoire, prenons le temps de décrire le mécanisme d'apprentissage du modèle inverse en jeu à l'étape 9. Dans ce travail, les vecteurs du dictionnaire de plongements du VQ-VAE acoustique sont définis comme les unités acoustiques de la parole découvertes par l'agent. Ainsi, le but de l'agent, « produire un son comportant les mêmes unités que le son qu'il perçoit », peut être formalisé en « produire un son dont le passage dans le VQ-VAE acoustique doit donner la même séquence d'unités que la séquence d'unités $\mathbf{o}^s$ extraite du son perçu $\mathbf{s}$ ». Pour comprendre comment cet objectif est traduit en une fonction de coût différentiable, il est utile de rappeler que l'association d'une sous-séquence de trames de signal acoustique $\mathbf{s}_{t\pm\tau}$ à une unité par le VQ-VAE acoustique se déroule en deux étapes. La première est le passage de cette sous-séquence dans son encodeur qui fournit $z_e^s(\mathbf{s}_{t\pm\tau})$, et la seconde est la substitution par son module de quantification de cette sortie par le vecteur du dictionnaire de plongements le plus proche de celle-ci, $\mathbf{o}_t^s$. Si l'on veut qu'une sous-séquence acoustique présentée au VQ-VAE acoustique soit associée à une unité en particulier, il suffit alors de rapprocher la sortie de son encodeur, pour cette sous-séquence, du vecteur de plongement représentant l'unité ciblée. Ainsi, tel qu'illustré sur la figure 5.9, en rapprochant suffisamment la sortie de l'encodeur du vecteur représentant l'unité cible, ce dernier finira par être le plus proche de celle-ci parmi tous les vecteurs du dictionnaire de plongement et deviendra, de fait, l'unité associée à cette sous-séquence. De ce fait, en minimisant la distance

$\|\mathbf{o}_t^s - z_e^s(\widehat{\mathbf{s}}_{t\pm\tau})\|^2$ entre chacune des unités $\mathbf{o}_t^s$ extraites pour chacun des instants t du son perçu $\mathbf{s}$ et la sortie de l'encodeur du VQ-VAE $z_e^s(\widehat{\mathbf{s}}_{t\pm\tau})$ calculée au même instant t à partir du résultat acoustique estimé par le modèle direct $\widehat{\mathbf{s}}$, on encourage l'agent à produire un signal de parole comportant la même séquence d'unités que la séquence $\mathbf{o}^s$ extraite du son perçu $\mathbf{s}$. Comme les poids du VQ-VAE acoustique et du modèle direct sont gelés lors de la rétropropagation de ces erreurs, le seul degré de liberté laissé à l'agent pour les réduire dans le processus de minimisation à l'étape 9 est la mise à jour de son modèle inverse. Ce dernier se retrouve donc contraint de s'adapter pour trouver un geste $\mathbf{a}$ dont la conséquence acoustique anticipée $\widehat{\mathbf{s}}$ par le modèle direct donnera en sortie de l'encodeur du VQ-VAE acoustique à chaque instant $z_e^s(\widehat{\mathbf{s}}_{t\pm\tau})$ un vecteur plus susceptible d'être associé à l'unité $\mathbf{o}_t^s$ extraite du son perçu $\mathbf{s}$ au même instant.

Apprentissage du VQ-VAE articulatoire La procédure d'apprentissage du VQ-VAE articulatoire consiste elle aussi en la répétition en boucle d'une suite d'étapes. Ces étapes sont les suivantes :

1. L'agent reçoit un son $\mathbf{s}$ tiré du jeu de données d'apprentissage.
2. La séquence de paramètres articulatoires correspondante $\mathbf{a}$ est estimée par le modèle inverse g, tel que $\mathbf{a} = g(\mathbf{s})$.
3. Pour chaque instant t, une unité (discrète) articulatoire $\mathbf{o}_t^a$ est estimée par l'encodeur et le module de quantification du VQ-VAE articulatoire à partir de la sous-séquence $\mathbf{a}_{t\pm\tau}$. Une version reconstruite de cette dernière, notée $\mathbf{a}_{t\pm\tau}^q$, est obtenue en sortie de son décodeur.
4. Les paramètres du VQ-VAE articulatoire ainsi que le dictionnaire d'unités associé sont mis à jour par rétropropagation de la fonction de coût.

Comme pour la première procédure, celle-ci est stoppée avec un critère d'*early-stopping*, arrêtant l'entraînement lorsque l'erreur de reconstruction du VQ-VAE articulatoire ne baisse plus durant 25 *epochs*, après un minimum de 50 *epochs*.

5.2.1.5 Évaluation

Évolution des erreurs La figure 5.10 montre l'évolution des différentes fonctions de coût minimisées par l'agent et présentées section 5.2.1.4 pour l'optimisation des modèles direct et inverse, ainsi que des VQ-VAE acoustique et articulatoire.

Nous pouvons d'abord remarquer que, globalement, la valeur des erreurs mesurées baisse au cours de l'apprentissage, suggérant une acquisition de connaissances par les différentes parties de l'agent. Nous devons cependant constater que cette baisse n'est pas régulière pour chacun des modèles. En effet, l'évolution de l'erreur du VQ-VAE acoustique, l'erreur d'estimation du modèle direct ainsi que l'erreur du modèle inverse présentent toutes trois des irrégularités, principalement au début de leur entraînement. Pour le modèle inverse et le modèle direct, nous supposons que ces irrégularités sont causées par des dépendances entre leurs performances et celles du VQ-VAE acoustique. En effet, tant que le VQ-VAE acoustique et le modèle direct

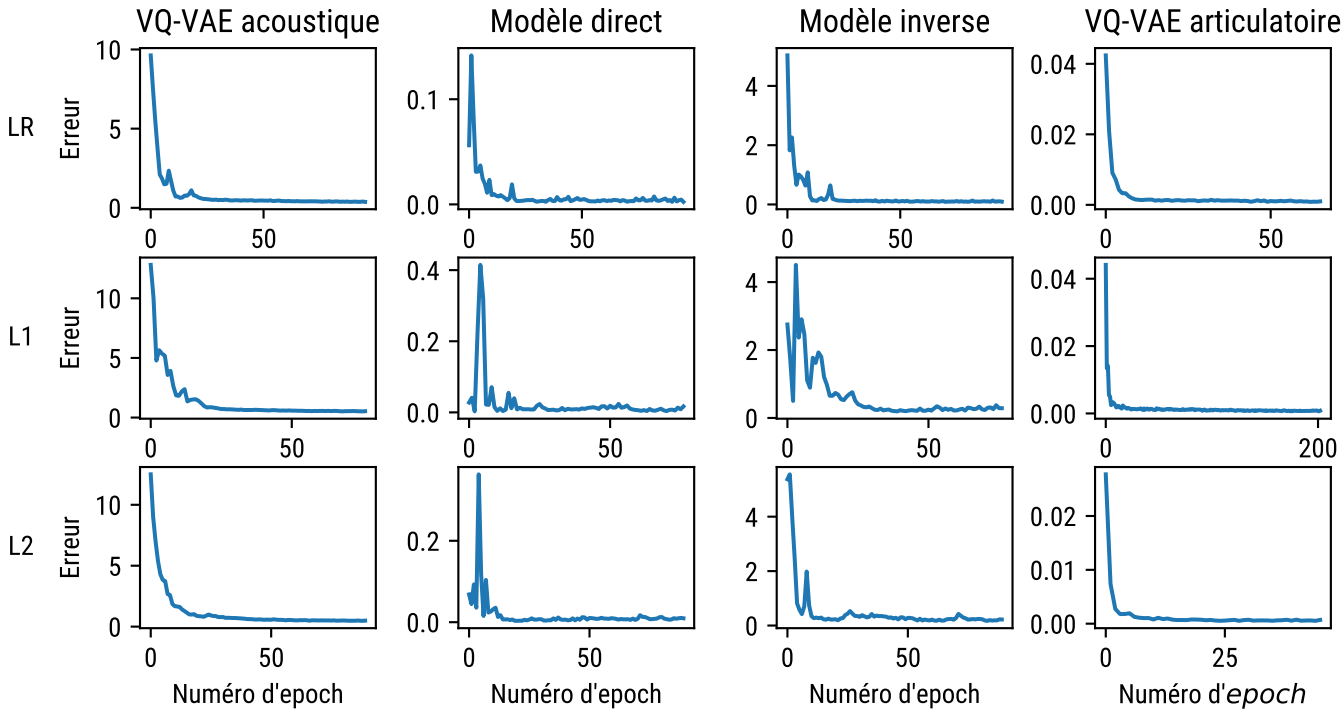

FIGURE 5.10 – **Exemples de l'évolution de l'erreur sur l'ensemble de test des différents modules de l'agent à but communicatif au cours de l'apprentissage (VQ-VAE acoustique et articulatoire, modèles internes direct et inverse)** L'erreur du modèle direct correspond à la différence entre sa sortie $\widehat{\mathbf{s}}$ et celle du synthétiseur $\widetilde{\mathbf{s}}$. L'erreur du modèle inverse correspond à la distance entre les unités décodées à partir du son perçu d'une part, et celles décodées à partir du son estimé par le modèle direct d'autre part (les unités étant représentées par leur plongement respectif, fourni par l'encodeur et le dictionnaire de plongements du VQ-VAE).

n'acquièrent pas suffisamment de connaissances, l'évolution du modèle inverse ne peut pas être guidée vers l'inversion de gestes articulatoires corrects. De plus, tant que les unités ne sont pas découvertes par le VQ-VAE acoustique et tant que le modèle inverse n'infère pas de gestes suffisamment adéquats pour s'en approcher, les régions de l'espace articulatoire explorées par le modèle inverse sont susceptibles de varier grandement, impactant de fait l'erreur du modèle direct, qui n'est entraîné et évalué que sur celles-ci. Le VQ-VAE acoustique quant à lui est entraîné directement sur les sons perçus et ne dépend donc pas des autres modules de l'agent. De ce fait, nous attribuons les irrégularités de l'évolution de sa fonction de coût à des « conflits momentanés » entre son objectif de reconstruction et son objectif de régularisation de son espace latent.

La figure 5.10 montre également l'évolution de la fonction de coût du VQ-VAE articulatoire au cours de la seconde phase d'apprentissage des agents entraînés. Nous remarquons que cette évolution est plus régulière que celles observées précédemment. Ceci n'est pas surprenant si l'on tient compte du fait que cette seconde phase n'inclut que l'entraînement du VQ-VAE articulatoire, qui dépend seulement du modèle inverse qui a déjà été stabilisé au cours de la

phase précédente.

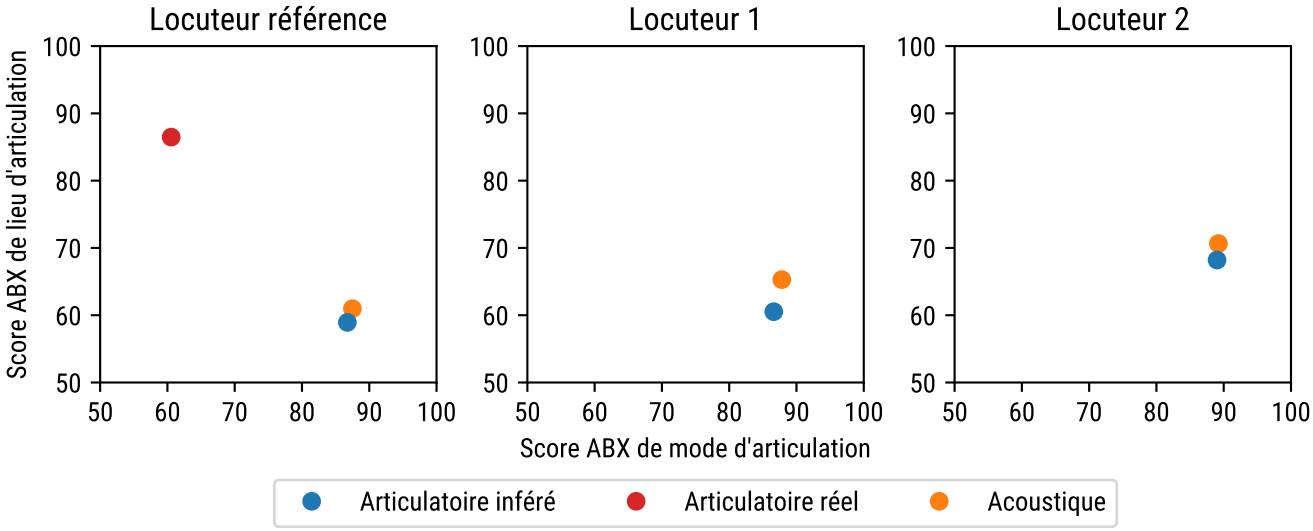

FIGURE 5.11 – **Scores ABX en fonction du lieu et du mode d'articulation des représentations issues des données articulatoires inférées par l'agent à but communicatif et des données acoustiques réelles** À des fins de comparaison, les scores associés à des représentations extraites des données articulatoires réelles du locuteur de référence (LR, sur lequel est construit le synthétiseur articulatoire et qui définit une sorte de borne supérieure de performance) sont affichés (en rouge).

Évaluation ABX Afin d'évaluer les unités articulatoires et acoustiques de la parole extraites par les deux VQ-VAE de l'agent à but communicatif, nous avons appliqué sur celles-ci le protocole d'évaluation basé sur les tests ABX, déjà présenté en section 4.2. Les résultats de cette évaluation, calculés en moyennant les scores de 5 agents entraînés indépendamment pour chaque locuteur, sont affichés sur la figure 5.11. Pour rappel, le « score ABX de lieu d'articulation » et le « score ABX de mode d'articulation » reflètent respectivement à quel point les unités découvertes permettent de discriminer entre elles des consonnes ayant un lieu ou un mode d'articulation différent.

Les unités extraites de la modalité acoustique (en orange) permettent de mieux discriminer deux consonnes en fonction de leur mode d'articulation, plutôt qu'en fonction de leur lieu d'articulation. Comme l'entraînement du VQ-VAE acoustique est indépendant des modèles interne et direct, ce résultat est en réalité similaire à celui présenté à la section 4.2, à savoir l'évaluation d'unités de la parole découvertes par un VQ-VAE entraîné seul sur des données acoustiques.

Les unités extraites des gestes inférés par le modèle inverse présentent le même problème que celles extraites par l'agent à but imitatif et présentées en section 5.1.5.3. En effet, lorsque l'on compare leurs scores (en bleu) aux scores obtenus en considérant les trajectoires articulatoires réelles (enregistrées sur le locuteur LR), extraites par un VQ-VAE indépendant et ajoutées à titre indicatif (points rouges), on remarque une nette différence. Alors que les

unités extraites des trajectoires réelles affichent une meilleure discrimination du lieu d'articulation (par rapport au mode d'articulation), les unités extraites à partir des trajectoires articulatoires inférées affichent des scores similaires à ceux des unités acoustiques. Ce résultat suggère que les trajectoires articulatoires inférées par l'agent sont encore trop éloignées de trajectoires réelles, et n'apportent pas autant d'informations sur le lieu d'articulation. Cette hypothèse est corroborée par l'observation qualitative des trajectoires inférées par l'agent, qui présentent globalement les mêmes caractéristiques que celles inférées par l'agent à but imitatif (voir section 5.1.5.3).

5.2.1.6 Conclusion

Nous avons proposé une mise à jour de l'agent à but imitatif, appelée « agent à but communicatif ». Cette mise à jour a pour but la découverte d'unités acoustiques et articulatoires de la parole, permettant de répéter des sons de parole en produisant des trajectoires acoustiques contenant les mêmes unités que celles perçues dans les signaux répétés. Une première évaluation de cette version de notre nouvel agent montre que son apprentissage se déroule bien, c'est-à-dire qu'il est capable d'apprendre conjointement, et de façon auto-supervisée, deux dictionnaires d'unités de la parole (acoustiques et articulatoires), ainsi que deux modèles, direct et inverse, des relations acoustico-articulatoire. Cependant, comme pour l'agent à but imitatif, l'évaluation des contenus articulatoires par le paradigme ABX suggère que les trajectoires articulatoires inférées par ce nouvel agent manquent toujours de réalisme et semblent avoir une dynamique trop proche de celle du signal acoustique répété (c'est-à-dire qu'elles semblent apporter le même type d'informations concernant le lieu et le mode d'articulation des consonnes). Dans les sections suivantes, nous proposons quelques pistes pour tenter de résoudre ce problème.

5.2.2 Ajout d'un mécanisme de babillage forcé

Nous proposons ici d'introduire un mécanisme de babillage dans la procédure d'apprentissage de l'agent. Notre hypothèse est que le babillage fournit aux enfants des exemples de trajectoires articulatoires qui contiennent et informent sur certaines propriétés de la parole réelle. En conséquence, si l'agent a accès à de telles trajectoires durant son apprentissage, les inférences réalisées par son modèle inverse seraient possiblement guidées vers des trajectoires articulatoires plus réalistes, notamment pour la réalisation des consonnes.

5.2.2.1 Implémentation

L'implémentation d'un mécanisme de babillage au sein de la procédure d'apprentissage de l'agent est réalisée en donnant périodiquement des trajectoires articulatoires directement à l'agent, qui les exécute et qui se sert du résultat pour mettre à jour son modèle inverse. Si nous souhaitions être le plus possible en phase avec les observations du comportement de l'enfant, nous pourrions par exemple fournir à l'agent des trajectoires inspirées de celles décrites par

MacNeilage et Davis, 2001, à savoir des variations rythmiques du degré d'ouverture de la mâchoire produisant des occlusions à différents endroits du conduit et servant d'exemples de proto-consonnes. Ça n'est pas le choix que nous avons fait. En effet, nous désirons ici savoir si cette voie d'évolution du modèle est prometteuse et nous faisons le choix de nous placer dans le scénario le plus favorable possible afin de voir si les performances de l'agent peuvent réellement être améliorées par l'introduction d'un mécanisme de babillage. Ainsi, les gestes donnés à l'agent pour le babillage sont ceux qui ont été enregistrés sur le locuteur de référence LR (c'est-à-dire le locuteur ayant servi pour la création du synthétiseur articulatoire).

En pratique, seule la première phase de l'algorithme d'apprentissage présenté en section 5.2.1.4 est modifiée. La nouvelle procédure d'apprentissage consiste en la répétition des étapes suivantes :

1. Entraînement de l'agent en suivant la procédure d'apprentissage du VQ-VAE acoustique et des modèles inverse et direct de l'agent. Cet entraînement se fait sur le corpus du locuteur cible (LR, L1 ou L2) durant une *epoch* complète.
2. Entraînement supervisé du modèle inverse de l'agent en lui fournissant directement les sons et les trajectoires articulatoires du corpus LR durant une *epoch* complète.

Cet apprentissage s'arrête suivant le même critère que celui utilisé dans la section précédente (lorsque l'erreur du modèle inverse sur un corpus de validation cesse de décroître durant 25 *epochs* après un minimum de 50 *epochs*). Une fois cet apprentissage terminé, le VQ-VAE articulatoire est entraîné en suivant la procédure décrite en section 5.2.1.4.

5.2.2.2 Évaluation

L'évolution des valeurs des fonctions de coût pour cette version de l'agent incluant le mécanisme de babillage étant très similaire aux évolutions observées sans introduire ce mécanisme, nous nous concentrerons directement sur l'évaluation ABX de ses unités articulatoires (la découverte des unités acoustiques, ne dépendant que des signaux acoustiques perçus, n'est pas différente de celle de l'agent sans babillage). Ces résultats sont affichés figure 5.12 (points cyans), aux côtés de ceux observés sur l'agent sans mécanisme de babillage (points bleus). Les scores ABX des unités acoustiques de cet agent sont également illustrés (points orange) ainsi que, à titre comparatif, ceux d'unités extraites directement à partir des trajectoires articulatoires réelles de LR (point rouge) par un VQ-VAE indépendant. Chacun des scores affichés est calculé en faisant la moyenne des résultats de 5 simulations, chacune avec une initialisation différente.

Premièrement, nous pouvons remarquer que le mécanisme de babillage a eu un effet sur le score ABX des unités articulatoires de l'agent entraîné sur le corpus LR. Celles-ci permettent de mieux discriminer les consonnes en fonction de leur lieu d'articulation que celles de l'agent entraîné sans babillage. Ceci suggère que les trajectoires inférées par cet agent sont plus proches des trajectoires réelles, qui elles aussi – et avec une propension encore plus grande – fournissent des unités reflétant mieux le lieu que le mode d'articulation. Cependant, le mécanisme de babillage ne semble pas avoir eu d'effets sur les unités articulatoires découvertes par les agents entraînés respectivement sur L1 et L2. Cette différence pour ces locuteurs par rapport à LR

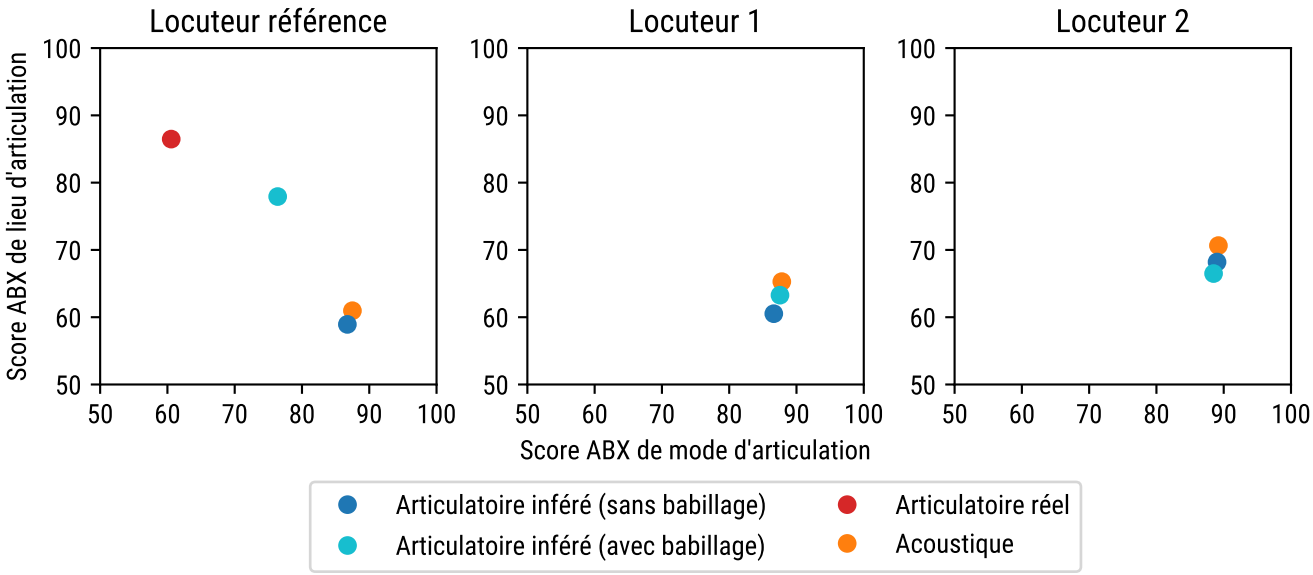

FIGURE 5.12 – **Scores ABX en fonction du lieu et du mode d'articulation des représentations issues des données articulatoires inférées par l'agent à but communicatif avec babillage forcé** Les points cyans montrent les scores d'unités articulatoires extraites par des agents avec babillage, les points bleus sans babillage. À des fins de comparaison, les scores associés à partir des sons (en orange), et à partir des données articulatoires réelles du LR (en rouge) sont affichés.

est probablement due au fait que ceux-ci ne partagent pas « la même voix » que le synthétiseur articulatoire piloté par l'agent. Ce résultat est doublement décevant. Premièrement parce qu'il indique qu'un mécanisme de babillage, même dans une forme idéale (et irréaliste), ne semble pas permettre à l'agent d'inférer des trajectoires articulatoires présentant les mêmes propriétés que des trajectoires articulatoires réelles. Deuxièmement, nous aurions pu espérer que l'utilisation d'unités acoustiques comme cibles pour les répétitions de l'agent introduirait une forme de normalisation de la voix des locuteurs cibles. En effet, ces unités acoustiques sont supposées être des représentations du signal de parole de plus haut niveau qu'une représentation spectrale, et leur utilisation comme cible aurait pu gommer, au moins en partie, les effets indésirables dûs aux différences entre « la voix de l'agent » (celle du synthétiseur) et celles des locuteurs L1 et L2.

5.2.2.3 Conclusion

Nous avons introduit un mécanisme de babillage dans l'agent à but communicatif en lui faisant exécuter directement les trajectoires articulatoires réelles enregistrées sur LR. Ce mécanisme semble avoir permis à l'agent entraîné sur les productions de LR d'inférer des gestes plus réalistes, mais cet effet ne se réplique pas aux autres agents entraînés sur les productions des locuteurs L1 et L2. Cette différence est probablement due au fait que l'agent partage « la même voix » que LR et en ait une différente de celles de L1 et L2. Ce résultat est décevant

car il suggère que le processus d'inversion articulatoire (le modèle inverse) de l'agent n'est pas généralisable à d'autres locuteurs dont le timbre de voix est différent du sien.

5.2.3 Ajout d'un mécanisme de normalisation acoustique

Comme nous l'avons vu dans la section précédente, l'agent à but communicatif, même avec un mécanisme de babillage très favorable, ne semble pas être en capacité d'inférer des trajectoires articulatoires « réalistes » lorsqu'il est entraîné à répéter les productions de locuteurs autres que celui ayant servi à la création du synthétiseur articulatoire qu'il pilote. Dans cette section, nous étudions l'apport d'un mécanisme de normalisation acoustique des données issues des locuteurs L1 et L2, par rapport au locuteur de référence LR, sur la qualité des trajectoires articulatoires inférées. La méthode que nous proposons ici ne cherche pas à être plausible d'un point de vue anatomique ou neuro-physiologique, mais plutôt à trouver quelles sont les limites de notre agent à but communicatif (on notera néanmoins que des mécanismes de normalisation acoustique sont déjà présents dès la petite enfance et ont été décrits par exemple dans Kuhl, 1979, 1983).

5.2.3.1 Normalisation

Nous distinguons ici deux caractéristiques des productions des locuteurs que nous cherchons à normaliser : la « dynamique » et le timbre. La dynamique est relative au rythme de la parole et le timbre est relatif à son contenu spectral. Chacune de ces caractéristiques présente un intérêt différent à être normalisée. En effet, la dynamique affecte directement la temporalité des trajectoires articulatoires que l'agent va inférer et le timbre va, lui, affecter les cibles articulatoires à atteindre.

Nous pouvons produire 3 versions normalisées d'un corpus : une version où le timbre est normalisé, une version où la dynamique est normalisée et une version où les deux le sont. Comme le but est d'étudier à quel point les productions utilisées pour entraîner l'agent doivent être proches de celles utilisées pour construire le synthétiseur (celles de LR), nous créons pour les corpus L1 et L2 trois nouvelles versions, pour chacune des trois déclinaisons susmentionnées, et nous créons pour le corpus LR deux nouvelles versions, normalisées pour copier respectivement la dynamique des corpus L1 et L2.

Techniquement, pour effectuer la normalisation des jeux de données, nous tirons partie du fait que ces derniers partagent le même contenu linguistique. Ainsi, nous pouvons estimer les paramètres d'une fonction de normalisation (conversion), de façon supervisée, à partir d'un corpus associant une production d'un locuteur source (par exemple L1 ou L2) et celle d'un locuteur de référence (LR), pour un énoncé donné.

Les deux sous-sections suivantes expliquent comment s'effectue cette conversion, d'abord pour la dynamique, puis pour le timbre.

Normalisation de la dynamique Afin de produire une nouvelle version des productions d'un corpus en reprenant la dynamique de celles d'un autre corpus, nous avons recours à l'algorithme de déformation temporelle dynamique (ou DTW pour *Dynamic Time Warping*). Pour rappel, les sons des productions sont représentés comme des séquences de vecteurs contenant chacun 40 coefficients Mel-spectraux (chaque vecteur décrivant le contenu spectral du signal de parole sur une fenêtre de 10 ms). L'algorithme DTW fournit un chemin (non-linéaire) d'alignement entre une séquence de vecteur issue d'une production « source » – dans notre cas produite par le locuteur à normaliser – et une production « cible », ici produite par locuteur de référence. En interpolant la séquence source le long du chemin DTW, tel qu'illustré figure 5.13, il est alors possible de lui appliquer la même dynamique que celle de la séquence cible. En répétant cette procédure « d'alignement puis d'interpolation le long du chemin DTW » pour chacune des productions du corpus à normaliser, on obtient une nouvelle version de celui-ci, contenant des productions alignées temporellement sur celles du corpus servant de référence. Dans ce travail, nous avons utilisé l'implémentation de l'algorithme DTW de la librairie Python *librosa*, et nous avons utilisé la distance euclidienne comme critère pour mesurer la similarité entre deux vecteurs.

Normalisation du timbre Pour produire une nouvelle version d'un corpus reprenant le timbre d'un autre locuteur, nous procédons à une conversion de voix[2] trame à trame de chacune de ses productions. Pour ce faire, nous créons d'abord un réseau de neurones de type *feedforward*, comprenant 3 couches cachées de 256 neurones pleinement connectées, dont la fonction d'activation est la tangente hyperbolique, et dont l'entrée et la sortie sont de même dimensionnalité que les vecteurs représentant les productions. Nous entraînons ensuite ce réseau en lui fournissant en entrée chacun des vecteurs de Mel-spectre du corpus à normaliser et en rétropropageant l'erreur quadratique moyenne entre sa sortie et les vecteurs du corpus dont on cherche à copier le timbre et qui sont alignés avec ceux du corpus à normaliser selon la méthode décrite dans la sous-section précédente (80 % des vecteurs sont utilisées pour l'entraînement et 20 % pour la procédure dite d'*early stopping*). Une fois le réseau entraîné, nous l'utilisons pour convertir les vecteurs du corpus à normaliser soit dans son état original soit dans sa version dont la dynamique a été normalisée (afin d'obtenir soit une nouvelle version dont seul le timbre a été normalisé, soit une nouvelle version dont à la fois le timbre et la dynamique ont été normalisés).

5.2.3.2 Évaluation

Muni de ces nouvelles versions des corpus dont le timbre et la dynamique ont été normalisés, nous les utilisons maintenant pour conduire l'entraînement d'agents en suivant exactement la même procédure que celle décrite pour l'agent à but communicatif avec babillage forcé. Comme les résultats concernant l'évolution des erreurs d'apprentissage sont très similaires à ceux de l'agent à but communicatif original, nous nous concentrerons ici uniquement sur les

2. La conversion de voix est un thème de recherche en soi qui a donné lieu à de multiples travaux. Voir Sisman et al., 2020 pour une revue récente.

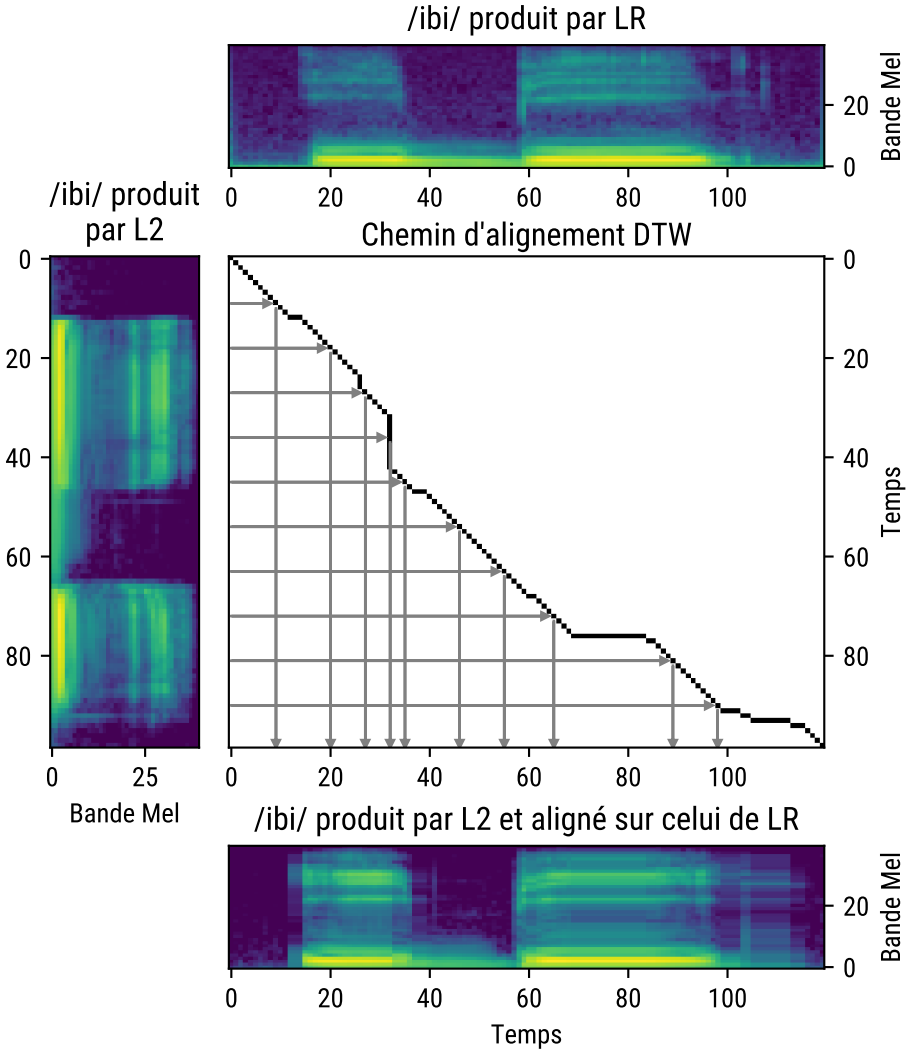

FIGURE 5.13 – **Alignement temporel de deux productions** Un chemin d'alignement DTW est d'abord calculé entre les deux productions. Les flèches illustrent comment ce chemin est ensuite utilisé pour aligner temporellement la production de L2 sur celle de LR.

résultats des tests ABX, qui fournissent, nous l'avons dit, une évaluation objective du contenu articulatoire des trajectoires inférées. Chacun des scores affichés dans cette section a été calculé en faisant la moyenne de ceux de 5 agents entraînés dans les mêmes conditions, mais à partir d'une initialisation différente.

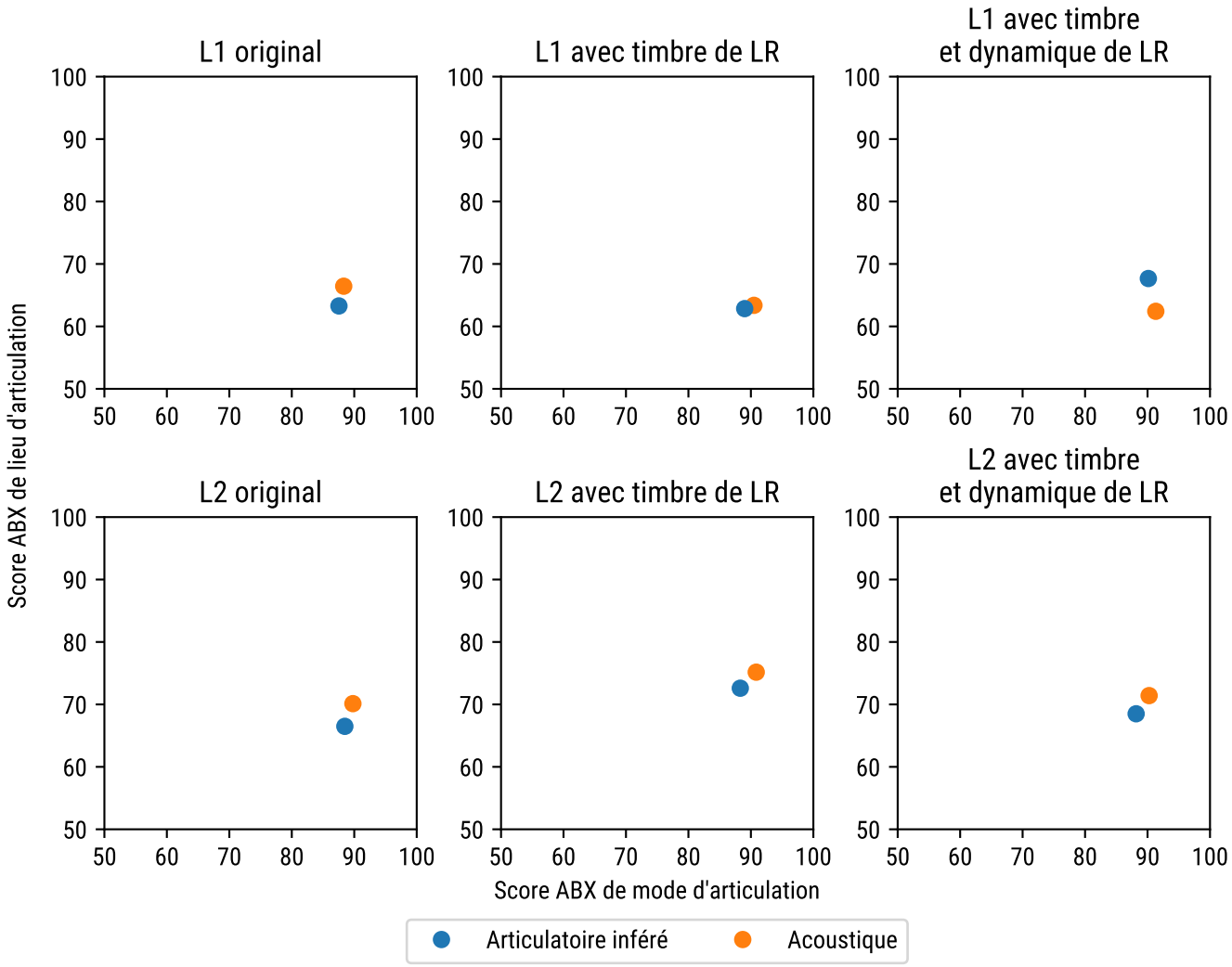

FIGURE 5.14 – **Scores ABX en fonction du lieu et du mode d'articulation des représentations extraites par l'agent à but communicatif, avec babillage forcé, et entraîné sur les corpus L1 et L2 originaux et normalisés** La procédure de normalisation permet la conversion du timbre et de la dynamique d'un locuteur tiers vers ceux du locuteur de référence.

Normalisation des caractéristiques de L1 et L2 vers celles de LR La figure 5.14 montre les résultats des tests ABX conduits sur les unités extraites par les agents entraînés à répéter les versions des corpus L1 et L2, dont soit le timbre seul, soit le timbre et la dynamique ont été normalisés vers ceux de LR. Les résultats de tests conduits sur les unités d'agents entraînés sur les corpus L1 et L2 originaux sont également affichés à des fins compara-

tives. Tout d'abord, nous pouvons remarquer qu'aucune des normalisations ne semble avoir eu d'effet significatif sur les scores ABX pour les unités extraites à partir des observations acoustiques. Aussi, nous pouvons remarquer qu'aucune des versions normalisées des jeux de données n'a permis à l'agent d'extraire des unités articulatoires permettant de mieux discriminer les consonnes en fonction de leur lieu que de leur mode d'articulation. Ce résultat suggère que les techniques de normalisation employées ici n'ont pas permis à l'agent d'apprendre à inférer des trajectoires « réalistes », qui elles fourniraient une meilleure discrimination en terme de lieu qu'en terme de mode d'articulation. Ce résultat est surprenant puisque ces nouvelles versions des corpus L1 et L2 sont supposées être proches du corpus LR, dont les trajectoires articulatoires réelles ont été utilisées pour forcer l'agent à babiller en parallèle de son apprentissage par répétition.

Il semble également que l'ajout de la normalisation de la dynamique, en plus de celle du timbre, n'ait pas eu d'impact significatif sur les scores ABX. Ceci suggère que les corpus dont le timbre a été normalisé sont encore trop éloignés de LR pour que l'ajout de la normalisation de la dynamique ait un effet significatif.

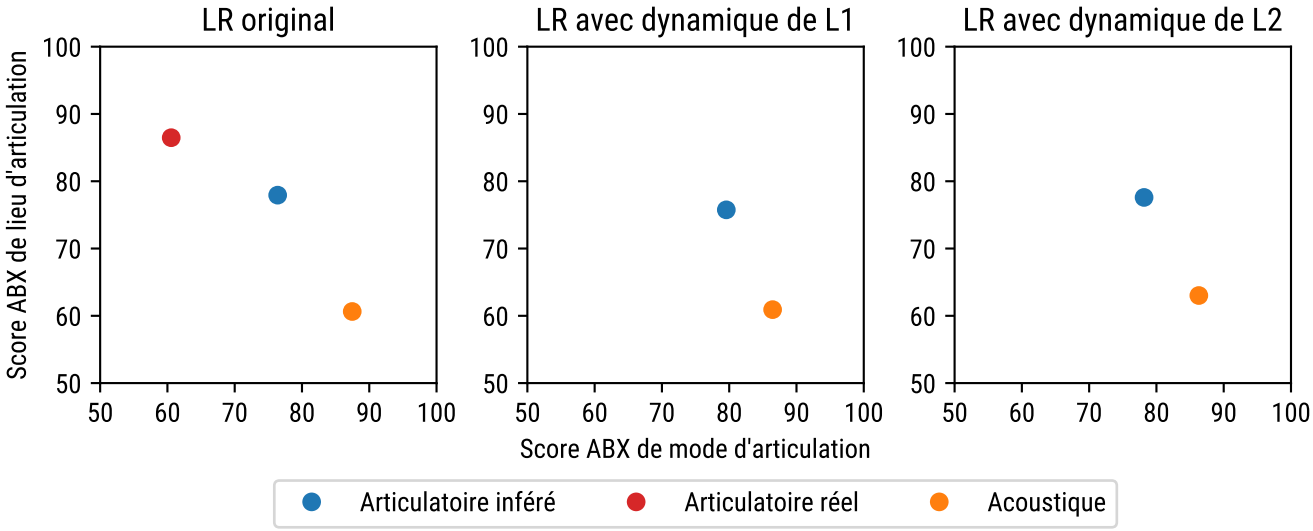

FIGURE 5.15 – **Scores ABX en fonction du lieu et du mode d'articulation des représentations extraites par l'agent à but communicatif, avec babillage forcé, et entraîné sur le corpus LR original ou normalisé** La procédure de normalisation permet la conversion du timbre et de la dynamique d'un locuteur tiers vers ceux du locuteur de référence. À des fins de comparaison, les scores associés à partir des données articulatoires réelles de LR (en rouge) sont affichés.

Normalisation de la dynamique de LR vers celle de L1 et L2 La figure 5.15 montre les résultats des tests ABX conduits sur les unités extraites par les agents entraînés à répéter les versions du corpus LR dont la dynamique a été normalisée pour reproduire celle de L1 et L2 ainsi que, à des fins comparatives, par des agents entraînés à répéter le corpus LR original. Nous pouvons remarquer que les unités articulatoires extraites par les agents permettent lors des tests ABX une discrimination meilleure en terme de lieu d'articulation et moins bonne en

terme de mode que les unités acoustiques, se rapprochant du schéma de performance d'unités extraites à partir de trajectoires articulatoires réelles (c.f. section 4.2.4.2). Lorsque l'on compare entre eux les scores des trois agents, on remarque qu'ils sont tous très similaires. Ceci suggère qu'une différence entre la dynamique des trajectoires articulatoires réelles utilisées pour forcer l'agent à babiller, et celle des productions à répéter, n'est pas pénalisante pour l'inférence de trajectoires articulatoires « réalistes ». Ceci suggère également que c'est la différence de timbre de voix qui est l'élément ayant le plus d'impact sur les performances des unités articulatoires extraites par l'agent à but communicatif avec babillage.

5.2.3.3 Conclusion

L'agent à but communicatif comprenant un mécanisme de babillage n'extrait des unités articulatoires « réalistes » qu'à partir du corpus LR, qui a servi à construire le synthétiseur articulatoire qu'il pilote et qui contient les trajectoires articulatoires utilisées pour le forcer à babiller. Nous avons, dans cette section, normalisé le timbre et la dynamique des corpus L1 ou L2 pour qu'ils s'approchent de ceux de LR, et normalisé la dynamique du corpus LR pour qu'elle reprenne celle de L1 et L2. Le résultat de ces normalisations est un nouvel ensemble de jeux de données, chacun plus ou moins similaire à LR – qui est le corpus sur lequel l'agent avec babillage présente les meilleures performances. Plusieurs agents à but communicatif avec babillage ont été entraînés sur chacun de ces nouveaux corpus. Les résultats montrent qu'une normalisation du timbre et de la dynamique des corpus L1 et L2 ne semble pas avoir d'effet sur le « réalisme » des trajectoires articulatoires inférées par l'agent, mais qu'une normalisation de la dynamique du corpus LR n'empêche aucunement l'agent d'inférer des trajectoires articulatoires « réalistes ».

5.2.4 Conclusion générale sur l'agent à but communicatif

Nous avons présenté dans cette section une mise à jour de l'agent à but imitatif, baptisée « agent à but communicatif ». Cette mise à jour introduit un mécanisme de découverte autosupervisée d'unités de la parole ainsi qu'un nouvel objectif d'apprentissage. Le mécanisme de découverte d'unités repose sur deux VQ-VAE entraînés respectivement sur les sons perçus et les trajectoires articulatoires inférées, permettant la découverte d'unités acoustiques et d'unités articulatoires. L'objectif d'apprentissage de ce nouvel agent est de décoder les unités acoustiques présentes dans les sons qu'il perçoit et de produire, en pilotant son synthétiseur articulatoire, un nouveau son comportant les mêmes unités que celles perçues. Ce nouvel objectif est intégré dans une mise à jour de la procédure d'apprentissage, dont l'efficacité pour entraîner les différentes parties composant l'agent est confirmée par l'évolution des valeurs des fonctions de coût que cette nouvelle procédure met en jeu. Cependant, alors que l'agent à but communicatif semble atteindre correctement son objectif (produire des sons contenant les mêmes unités que celles identifiées dans les sons perçus), les unités extraites à partir des données articulatoires inférées présentent des propriétés différentes (et d'une certaine façon décevantes) de celles obtenues à partir de données articulatoires réelles. En effet, celles-ci

représentent les consonnes davantage en terme de mode d'articulation qu'en terme de lieu, ce qui les rend assez similaires aux unités extraites directement à partir du signal acoustique.

Pour tenter de remédier à ce problème, nous avons successivement exploré deux pistes. Toutes deux n'avaient pas pour but d'être plausibles d'un point de vue du développement de la parole mais plutôt de nous permettre d'investiguer si, en se plaçant dans un scénario favorable, l'agent était capable d'inférer des trajectoires articulatoires plus réalistes. La première piste explorée a été l'ajout d'un mécanisme de babillage à la procédure d'entraînement de l'agent. L'introduction de ce mécanisme, qui consiste à forcer l'agent à exécuter les trajectoires articulatoires enregistrées sur le locuteur de référence (LR) et à se servir du résultat pour mettre à jour son modèle inverse, ne s'est montrée efficace que pour l'inférence de trajectoires articulatoires à partir des sons produits par ce même locuteur. Ainsi, la seconde piste explorée à été la normalisation acoustique des différents jeux de données LR, L1 et L2 afin d'en produire de nouvelles versions, chacune plus ou moins proche du locuteur de référence LR. Les résultats montrent que cette procédure de normalisation n'est utile que si elle est quasi-parfaite. Aussi, une erreur même faible dans la conversion des *stimuli* perçus – pour les ramener dans l'espace acoustique du synthétiseur de l'agent en terme de timbre – aboutit systématiquement à une forte dégradation de la qualité des trajectoires articulatoires inférées.

Ainsi, les résultats observés suite à l'évaluation des trajectoires articulatoires de l'agent s'avèrent décevants. Malgré la combinaison de processus de quantification au niveau du VQ-VAE, supposée réduire la variabilité acoustique, de processus de normalisation acoustique inter-locuteurs, et l'introduction d'un processus de guidage du modèle inverse par « babillage forcé », nous n'avons pas réussi à générer des trajectoires articulatoires vraiment réalistes et incorporant dans l'agent les connaissances articulatoires attendues pour améliorer le processus de découverte d'unités phonétiques pertinentes. Nous discuterons dans le prochain chapitre de perspectives qui pourraient permettre d'espérer des gains significatifs de performance pour tenter de dépasser ces limites actuelles de nos simulations.

Chapitre 6

Discussion

Sommaire

6.1 État des lieux

Pour démarrer cette discussion nous allons dans cet état des lieux rappeler les principales contributions de ce travail mais également les mettre en perspective avec quelques autres travaux qui se sont développés dans la même temporalité en proposant des architectures qui par certains aspects sont proches de celles que nous avons conçues.

6.1.1 Contributions

La principale contribution de ce travail de thèse est la création d'un agent capable d'apprendre « à parler » de façon auto-supervisée, uniquement à partir de *stimuli* acoustiques issus de son environnement. Cet agent est basé sur un module de découverte d'unités phonétiques, un synthétiseur articulatoire qui joue le rôle d'appareil vocal virtuel, et deux modèles internes respectivement direct et inverse qui représentent la façon dont le cerveau internalise les relations complexes entre le contenu spectral des signaux de parole d'une part, et les trajectoires articulatoires associées d'autre part. La conception de cet agent est basée sur les outils du *machine learning* et plus précisément les réseaux de neurones profonds qui permettent un entraînement conjoint de tous ses modules, un ajout simple de nouveaux modules ou contraintes et la capacité de travailler avec un volume potentiellement très important de données issues du monde réel. Cet agent fournit ainsi un cadre général de modélisation et peut, comme nous

le verrons dans la section 6.2, être modifié afin d'étudier différents aspects du développement de la parole.

Afin de rendre l'agent capable de produire des sons de parole de bonne qualité à partir de commandes articulatoires interprétables, nous avons commencé par définir le processus de création d'un nouveau synthétiseur articulatoire. Ce processus permet, à partir d'un corpus d'enregistrements EMA et acoustiques d'un locuteur, de créer un synthétiseur en trois étapes. La première est la construction d'un modèle articulatoire à partir des données EMA permettant de traduire ces dernières en paramètres articulatoires décrivant la position des différents articulateurs du conduit vocal. La seconde est l'entraînement d'un réseau de neurones jouant le rôle de modèle articulatoire-vers-acoustique chargé d'estimer, à partir des paramètres articulatoires précédemment extraits, le contenu spectral du signal à générer. Enfin, la troisième étape est l'entraînement d'un vocodeur neuronal capable de transformer le contenu spectral estimé par le modèle articulatoire-vers-acoustique en une forme d'onde. Suite à ce processus, l'assemblage du modèle articulatoire-vers-acoustique avec le vocodeur neuronal forme le synthétiseur articulatoire complet, prenant en entrée des paramètres articulatoires et fournissant en sortie un signal de parole réaliste.

Ensuite, nous avons proposé deux études visant à quantifier l'apport d'informations articulatoires sur la création de représentations de la parole. Dans la première de ces études, les paramètres articulatoires ont été utilisés lors de l'entraînement d'un auto-encodeur variationnel (VAE) chargé de reconstruire un signal acoustique de parole afin de régulariser la construction d'une partie de son espace latent. Ce VAE régularisé articulatoirement (AR-VAE) montre de meilleures performances de reconstruction qu'un VAE classique, suggérant l'intérêt d'exploiter des connaissances articulatoires explicites pour le décodage de la parole. La seconde étude a porté sur la découverte auto-supervisée d'unités discrètes de la parole à partir d'informations acoustiques, articulatoires ou bien du mélange de ces deux modalités. Pour cela, des auto-encodeurs variationnels quantifiés vectoriels (VQ-VAE) ont été entraînés à reconstruire le contenu spectral d'un signal de parole, ses paramètres articulatoires ou bien la fusion de ces deux informations, et, par là même, à construire des représentations discrètes des modalités acoustique et articulatoire dans leur espace latent. Une évaluation de ces représentations, via une méthodologie ABX focalisée sur les consonnes, révèle que les représentations issues des données acoustiques et articulatoires prises séparément encodent respectivement mieux le mode et le lieu d'articulation, et que la fusion de ces deux modalités permet d'obtenir des représentations efficaces à la fois pour le mode et le lieu.

Enfin, nous avons conçu une première architecture de l'agent, qualifiée d'« agent à but imitatif » et proposé une évolution de celle-ci, qualifiée d'« agent à but communicatif ». L'architecture de l'agent à but imitatif possède, en plus du synthétiseur articulatoire, un modèle direct chargé de représenter la relation articulatoire-vers-acoustique et un modèle inverse chargé de représenter la relation acoustique-vers-articulatoire. L'utilisation de ces deux modèles est directement inspirée des modèles théoriques du contrôle moteur de la parole, tels que ceux présentés au chapitre 1. Cet agent tente d'apprendre la relation inverse acoustique-vers-articulatoire de façon auto-supervisée, en répétant, à l'aide de son « conduit vocal » – ici représenté par son synthétiseur articulatoire – les *stimuli* auditifs qu'il perçoit. Cet agent est

dit purement imitatif car la distance entre le son perçu et le son produit est évaluée au niveau du spectre d'amplitude, l'agent cherche donc d'une certaine manière à copier le timbre du locuteur qu'il perçoit. Pour le second agent proposé, nous cherchons à définir une fonction objective plus réaliste et faisons l'hypothèse qu'il suffit à l'agent de chercher à préserver un « code linguistique », par exemple phonétique, qu'il doit également découvrir de façon auto-supervisée. L'architecture de l'agent à but communicatif intègre donc deux nouveaux modèles permettant la découverte auto-supervisée d'unités de la parole, respectivement à partir des modalités acoustique et articulatoire. La procédure d'apprentissage de cette deuxième version a également été modifiée pour permettre l'entraînement de ces deux modèles de découverte d'unités ainsi que pour intégrer l'utilisation des unités acoustiques découvertes par l'agent comme cible de ses répétitions des sons perçus.

6.1.2 Positionnement par rapport à des développements récents

Au cours de l'avancement de cette thèse, d'autres travaux d'apprentissage automatique auto-supervisé de commandes articulatoires pour de la synthèse acoustique ont été développés parallèlement. Dans cette sous-section, nous présentons trois d'entre eux en montrant comment nos travaux s'en démarquent.

ArticulationGAN Le premier de ces travaux porte sur le modèle ArticulationGAN (Beguš et al., 2022). L'architecture de ce modèle est une version modifiée d'un réseau antagoniste génératif (*generative adversarial network*, GAN). Pour rappel, un GAN est un modèle composé de deux réseaux de neurones qui sont mis en concurrence lors de leur entraînement : un générateur et un discriminateur. Le générateur produit une observation (par exemple une image ou un son), à partir d'une observation tirée d'une distribution aléatoire. Le discriminateur est (dans sa formulation originale) un classifieur binaire chargé de déterminer si les observations qui lui sont présentées sont issues du générateur ou bien des données d'apprentissage. Dans ArticulationGAN, le générateur, appelé « générateur articulatoire », produit des trajectoires articulatoires sous la forme de coordonnées x et y de 6 bobines EMA (les mêmes que celles présentées sur la figure 2.1, à l'exception de celle attachée au velum) et un paramètre de voisement. Ces trajectoires sont ensuite envoyées dans un synthétiseur articulatoire qui est un réseau de neurones pré-entraîné et qui les transforme en formes d'ondes. Le discriminateur est, lui, entraîné à identifier si les formes d'ondes qui lui sont présentées sont des données réelles ou sont issues du générateur.

Les résultats montrent que ce modèle est capable d'apprendre à produire des sons de mots de bonne qualité. Cependant, en l'absence de contraintes, les trajectoires articulatoires employées pour les générer semblent éloignées de trajectoires réelles.

La principale différence avec l'agent conçu dans ce travail de thèse est que ArticulationGAN, du fait de son architecture de type GAN, est un modèle génératif. De ce fait, ArticulationGAN modélise plutôt un processus de génération spontanée de parole conditionnée par les sons perçus dans l'environnement langagier plutôt qu'un processus de répétition directe de

sons perçus. De plus, lors de l'entraînement du modèle le gradient d'erreur est rétropropagé au travers du synthétiseur articulatoire – et non au travers d'un modèle direct appris. Comme énoncé en section 5.1.4, ce choix est critiquable car il donne au modèle, avant même son entraînement, l'accès à la connaissance complète de la relation articulatoire-vers-acoustique. Notons enfin que l'espace de contrôle du synthétiseur articulatoire est défini par des coordonnées de bobines EMA plutôt que par des paramètres articulatoires interprétables – nous reviendrons sur ce choix en section 6.2.1.

MirrorNet MirrorNet (Siriwardena et al., 2022) est un modèle dont le but est d'apprendre à répéter les sons qu'il perçoit le plus fidèlement possible (au sens de la proximité du contenu spectro-temporel), et dont l'architecture est très similaire à celle de l'agent à but imitatif présenté en section 5.1. En effet, le modèle est un auto-encodeur entraîné à reproduire une entrée acoustique et dont l'espace latent est directement un ensemble de paramètres servant à piloter un synthétiseur articulatoire, également afin de reproduire le son d'entrée. De ce fait, l'encodeur et le décodeur du modèle peuvent être respectivement assimilés au modèle inverse et au modèle direct de l'agent à but imitatif. Les sons sont modélisés par des spectrogrammes de longueur fixe et les paramètres articulatoires sont des variables représentant la position des lèvres ainsi que les constrictions entre la langue et le palais.

Les résultats montrent que le modèle est capable d'apprendre à répéter les sons qu'il perçoit de façon satisfaisante sur le plan acoustique. Cependant, au niveau des trajectoires articulatoires, un pré-entraînement supervisé de l'encodeur et du décodeur de MirrorNet est nécessaire pour que le modèle apprenne, lors de son entraînement principal, à inférer des trajectoires articulatoires « réalistes » à partir des sons qu'il perçoit.

Simulating vocal learning of spoken language Dans leur travail, van Niekerk et al., 2023 proposent une procédure de recherche de paramètres pour piloter un synthétiseur articulatoire dans le but de produire des sons de type « consonne-voyelle » (CV). Pour cela, les auteurs commencent par créer un « encodeur syllabique », qui consiste en un réseau de neurones entraîné de façon supervisée à reconnaître la consonne et la voyelle de sons de type CV. Ensuite, ils utilisent une approche de type TPE (*Tree-structured Parzen Estimator*, Bergstra et al., 2011) pour rechercher des paramètres d'un synthétiseur articulatoire prédéfini (Birkholz et al., 2006 ; Birkholz et al., 2010) permettant la synthèse de sons qui sont identifiés par l'encodeur syllabique comme correspondant à des séquences CV en particulier. Ils comparent également cette approche à une recherche de paramètres de synthétiseur guidée par la proximité spectro-temporelle avec des sons CV perçus. Chacune de ces recherche est également guidée par divers types d'objectifs articulatoires conjoints (tels qu'imposer une occlusion lors de la production de la consonne).

Les résultats montrent que la recherche guidée par l'encodeur syllabique permet de produire des sons davantage identifiés comme portant la syllabe cible que la recherche guidée par la proximité du contenu spectro-temporel. Malgré la présence d'un objectif articulatoire, les auteurs ne proposent pas d'évaluation des gestes issus du processus de recherche.

Ces deux types d'objectifs perceptifs peuvent être reliés respectivement à ceux de notre agent à but imitatif (pour la recherche guidée par le contenu spectro-temporel, voir section 5.1) et de notre agent à but communicatif (pour la recherche guidée par la syllabe cible, voir section 5.2). Ainsi, le décodeur syllabique peut être vu comme un VQ-VAE acoustique qui aurait été pré-entraîné pour identifier les consonnes et les voyelles, et le processus de recherche des paramètres de synthétiseur comme une sorte de modèle inverse. Cependant, contrairement à l'agent à but communicatif, la procédure de recherche proposée ne comporte pas de modèle direct de la relation articulatoire-vers-acoustique. Concrètement, cela signifie que pour guider la recherche vers des gestes adaptés, la procédure utilise un processus d'échantillonage. Le risque est de ne pas permettre à l'agent d'accumuler des connaissances sur les liens acoustico-articulatoires réutilisables d'une situation à une autre, et donc de le mettre en situation de ne pouvoir exploiter aucune connaissance préalable sur son conduit vocal dans la production d'une syllabe jusqu'alors inconnue.

6.2 Perspectives

Nos expériences sur l'agent ont montré que ce dernier était capable d'apprendre à répéter les sons qu'il perçoit en inférant des gestes dont le résultat acoustique est soit proche du son perçu pour l'agent à but imitatif, soit porteur des mêmes unités pour l'agent à but communicatif. Cependant, des problèmes et limitations demeurent, à différents niveaux. D'abord, la qualité de restitution sonore semble encore insuffisante (voir notamment les résultats des évaluations perceptives présentées aux sections 3.3.2 et 5.1.5.2). Par ailleurs, le problème de la normalisation entre locuteurs, introduit section 1.2.4.2, n'a été que peu abordé dans ce travail (voir section 5.2.3). Enfin, les unités extraites à partir de trajectoires articulatoires inférées par l'agent n'ont pas les mêmes propriétés que celles d'unités extraites à partir de données réelles (section 5.1.5.3). Ceci suggère que les gestes inférés par l'agent sont encore trop éloignés de gestes « naturels » et sont probablement sélectionnés de façon *ad hoc* pour atteindre au mieux l'objectif acoustique, au détriment de la cohérence articulatoire – ce qui est d'ailleurs probablement aussi le cas des autres travaux récents dans le domaine, comme nous venons de le voir en section 6.1.2. Cette section présente plusieurs pistes d'amélioration de l'agent que nous avons imaginées et qui pourraient potentiellement permettre de pallier ces problèmes. De manière essentielle, nous verrons que chacune de ces perspectives, quelles que soient la difficulté ou l'ampleur des questions à résoudre, peut renvoyer à une implémentation neuronale compatible avec l'architecture globale de l'agent développé.

6.2.1 Amélioration du synthétiseur

Une première perspective d'amélioration de l'agent serait le perfectionnement de son synthétiseur articulatoire. En effet, comme nous l'avons vu dans la section 3.3.2, le synthétiseur articulatoire dans sa forme actuelle ne permet pas une restitution parfaite des phonèmes, et en particulier des consonnes, lors de resynthèses à partir de trajectoires articulatoires réelles. De ce fait, si l'agent essaie de produire ces phonèmes à l'aide du synthétiseur, il risque de devoir

employer des trajectoires articulatoires conçues pour contourner les limitations du synthétiseur, et, ainsi, probablement éloignées de trajectoires « naturelles ».

Amélioration du modèle articulatoire Une première piste d'amélioration du synthétiseur se situe au niveau du modèle articulatoire sur lequel il repose. Pour rappel, le modèle articulatoire actuellement utilisé est un modèle linéaire construit à partir d'enregistrements EMA, et permet de traduire ces derniers en une série de paramètres articulatoires interprétables. Ces paramètres articulatoires sont ensuite utilisés par le modèle articulatoire-vers-acoustique qui essaie de déduire le contenu spectral du signal qui résulterait de leur exécution. Or, lorsque l'on compare les performances du modèle articulatoire-vers-acoustique prenant en entrée les paramètres articulatoires à celles d'un homologue prenant en entrée les coordonnées EMA originales, on remarque que ce dernier affiche de meilleures estimations du signal acoustique (section 3.2.2). Cette différence peut s'expliquer par une perte d'informations induite par le processus de transformation des coordonnées EMA en paramètres articulatoires (à l'aide du modèle articulatoire présenté en section 3.1.1). Une première solution pour améliorer les performances du synthétiseur consisterait alors à abandonner purement et simplement cette étape d'extraction de composantes interprétables, et à effectuer la synthèse directement à partir des données EMA brutes. C'est par exemple la solution choisie par Beguš et al., 2022, nous l'avons vu. Mais le passage par un modèle, extrayant des variables plus directement interprétables et potentiellement moins dépendantes du corpus lui-même, nous semble intéressant pour la suite de ces travaux. C'est pourquoi nous conservons la perspective d'une amélioration du modèle articulatoire, et donc de ce processus de transformation, susceptible de réduire cette perte d'information et de potentiellement mener à de meilleures synthèses. C'est ce raisonnement qui nous a menés à tenter de porter le modèle articulatoire linéaire original vers le monde des réseaux de neurones, d'abord de façon littérale en section 3.1.2, puis de façon *end-to-end* en section 3.1.3. Ces tentatives ouvrent la voie vers un modèle articulatoire non-linéaire plus performant et ainsi vers un meilleur synthétiseur articulatoire. Elles se sont révélées à ce stade peu fructueuses, conduisant dans nos premières simulations à des performances globales inférieures à celles du modèle linéaire (voir section 3.1.3.3). Mais la mise en œuvre d'une architecture globale *end-to-end* qui inclut dans son espace de recherche la solution linéaire garantit que, si les contraintes techniques de solution initiale et de poids des contraintes dans l'espace de recherche sont résolues, nous devrions obtenir des gains de performance, possiblement intéressants pour la suite du travail.

Inclusion de la source dans le modèle articulatoire-vers-acoustique Une autre piste d'amélioration du synthétiseur est l'inclusion de la source dans le modèle articulatoire-vers-acoustique. Pour le moment, celui-ci n'est chargé d'estimer que la « partie filtre » du signal à générer sous la forme d'un contenu spectral de basse résolution (18 coefficients cepstraux sur une échelle Bark). C'est ensuite LPCNet, le vocodeur neuronal utilisé en aval afin de générer la forme d'onde, qui prend en entrée ce contenu spectral estimé ainsi que deux paramètres de source décrivant respectivement la fréquence fondamentale du signal à générer et son degré de périodicité.

Cette manière de procéder est problématique car elle ne permet pas, comme nous l'avons vu en section 3.3.2, de contrôler finement la présence de voisement dans les resynthèses par l'intermédiaire du paramètre indiquant le degré de périodicité du signal à générer. Ce problème indique que, malgré leur faible résolution spectrale, les paramètres de filtre portent encore trop d'informations sur la source et que ces informations prennent le pas sur les deux autres paramètres dédiés à celle-ci lors de la génération de la forme d'onde par LPCNet. Ce problème indique également que les paramètres de filtre associés à deux configurations articulatoires très similaires sont différents en fonction de la présence de voisement ou non (par exemple dans le cas de /k/ vs. /g/). De ce fait, le modèle articulatoire-vers-acoustique, qui ne prend en entrée que les paramètres articulatoires, contrôle et choisit arbitrairement si le signal de parole généré est voisé ou non.

Ces dysfonctionnements pourraient potentiellement être réglés par l'ajout des deux paramètres de source en entrée du modèle articulatoire-vers-acoustique. En effet, ceci permettrait d'indiquer à ce dernier s'il doit produire le contenu spectral d'un signal voisé ou non, ce qui réduirait potentiellement l'erreur de LPCNet lors de la synthèse de la forme d'onde finale. Cette modification aurait également un second avantage, qui est celui d'ajouter les paramètres de source à l'espace de contrôle de l'agent. En effet, ce dernier n'a pour l'instant accès qu'au modèle articulatoire-vers-acoustique pour produire des signaux et n'a donc pas de contrôle, dans sa version à but imitatif, ni sur la hauteur, ni sur le voisement du signal qu'il produit. Avec cet ajout, l'agent pourra plus explicitement décider s'il veut produire un son voisé ou non. Un premier pas dans cette direction a été réalisé dans la version à but communicatif de l'agent mais son effet n'a pas eu à ce stade les effets escomptés.

Évolution vers une approche *end-to-end* Enfin, une dernière piste d'amélioration du synthétiseur pourrait être le basculement vers une approche *end-to-end*, telle que celles présentées dans Y.-W. Chen et al., 2021 ; Wu et al., 2022. Dans cette approche, le synthétiseur serait constitué d'un unique modèle reconstruisant une forme d'onde directement à partir des paramètres articulatoires, sans passer par une représentation spectrale intermédiaire.

6.2.2 Inférence des trajectoires articulatoires

Une seconde perspective d'amélioration de l'agent se situe au niveau de son modèle inverse et de la façon dont ce dernier infère les trajectoires articulatoires adéquates pour reproduire les sons perçus. Pour le moment, ce processus d'inférence de gestes n'est nullement contraint et aucun mécanisme ne l'empêche de proposer une séquence de gestes défiant les lois de la bio-mécanique si cette séquence de gestes permet de mieux atteindre l'objectif d'apprentissage. En effet, le modèle inverse pourrait par exemple inférer un geste où la langue basculerait en un seul pas de temps (soit 10 ms) d'une position très avancée à une position très reculée si cela permet d'obtenir un résultat acoustique permettant de réduire l'erreur d'apprentissage. Les prochains paragraphes de cette sous-section présentent plusieurs propositions qui permettraient l'introduction de contraintes au sein du modèle inverse, susceptibles de limiter ou d'interdire l'inférence de séquences gestuelles « aberrantes ».

Interdiction des gestes impossibles Du fait de l'absence de contraintes sur l'inférence des paramètres articulatoires par le modèle inverse, ce dernier peut donner une combinaison de paramètres qui, une fois reprojetés dans l'espace des coordonnées des bobines EMA à l'aide du modèle articulatoire, donnent des configurations physiquement impossibles comme par exemple des bobines de langue traversant le palais. Un mécanisme de détection de ces gestes impossibles permettrait d'ajouter un terme à la fonction d'erreur minimisée par le modèle inverse lors de l'entraînement de l'agent. Ce terme permettrait de pénaliser fortement le choix de ces gestes afin de tenter d'empêcher leur usage.

La mise en œuvre de cette technique présente cependant un écueil : comment détecter ces gestes impossibles ? Dans les jeux de données PB2007 et BY2014 nous disposons du contour du palais sous la forme d'un ensemble de coordonnées dans l'espace des bobines EMA. Ce contour peut permettre de détecter une partie des gestes impossibles dans lesquels les bobines de la langue traversent le palais. L'ajout d'une telle information de contour présenterait également l'intérêt de permettre au modèle d'évaluer la taille des constrictions, qui fournissent potentiellement une information clé pour mieux évaluer les résonances, la génération de frictions, et globalement la sortie acoustique.

Néanmoins, ces corpus ne disposent pas de suffisamment d'informations pour détecter d'autres types de gestes impossibles tels que ceux où la position des articulateurs est trop extrême (par exemple un avancement de la langue ou un degré d'ouverture de la mâchoire trop importants). L'enregistrement d'un nouveau corpus de données EMA dans lequel le locuteur enregistré adopterait des positions extrêmes avec chacun de ses articulateurs aiderait à proposer des limites dans l'espace des bobines EMA. Ces limites pourraient permettre de définir comme impossibles les paramètres articulatoires induisant leur traversée par l'une des bobines.

Minimisation du *jerk* Comme nous l'avons vu en section 1.2.4.2, certains systèmes d'inversion acoustique-vers-articulatoire proposés dans la littérature intègrent des principes de régularisation des trajectoires articulatoires. Ces systèmes s'appuient notamment sur la proposition de Hogan, 1984 selon laquelle les mouvements humains sont organisés de façon à emprunter des trajectoires lisses qui minimisent l'intégrale sur la trajectoire du *jerk* (qui est, rappelons-le, la dérivée de l'accélération).

Ce principe pourrait être intégré au sein de l'agent au travers d'un terme supplémentaire ajouté à la fonction d'erreur minimisée par son modèle inverse lors de son entraînement. Ce terme intégrerait la somme des changements d'accélération des paramètres articulatoires le long des trajectoires inférées par le modèle inverse et sa minimisation entraînerait un lissage de celles-ci. Cette solution a déjà été utilisée sous des formes diverses dans des travaux similaires (Rasilo et al., 2013 ; van Niekerk et al., 2023). De plus, elle est relativement simple à implémenter sous réserve de prendre en compte comme hyperparamètre la pondération de ce nouveau terme par rapport au terme de cible acoustique afin d'éviter d'inférer des trajectoires trop lisses n'atteignant pas les objectifs acoustiques escomptés.

Inférence de vitesse Une autre façon d'obtenir un lissage des trajectoires articulatoires inférées par le modèle inverse serait de lui retirer toute possibilité de produire des changements trop brusques des paramètres articulatoires d'un pas de temps à l'autre. Pour le moment, lorsque l'on demande au modèle inverse d'inférer une trajectoire articulatoire pour produire un signal acoustique, celui-ci fournit une série de paramètres articulatoires en définissant directement leur valeur à chaque pas de temps. Au lieu de cela, le modèle inverse pourrait plutôt contrôler, toujours à chaque pas de temps, la vitesse de ces paramètres (à noter que c'est le principe de contrôle implémenté dans le modèle DIVA, présenté en section 1.1.3.1).

Concrètement, ce nouveau modèle inverse g_v inférant la vitesse des paramètres articulatoires pourrait être implémenté par un réseau de neurones défini par :

$$g_v(\mathbf{s}_t, \mathbf{a}_{t-1}) = \Delta_{\mathbf{a}_t},$$

avec $\mathbf{s}_t$ vecteur décrivant les caractéristiques du signal acoustique que l'on cherche à inverser à l'instant t, $\mathbf{a}_{t-1}$ la valeur des paramètres articulatoires à l'instant précédent et $\Delta_{\mathbf{a}_t}$ le degré de changement à appliquer à ceux-ci (leur vitesse) pour approcher le résultat acoustique de leur exécution de $\mathbf{s}_t$. Grâce à $\Delta_{\mathbf{a}_t}$, la valeur des paramètres articulatoire peut être connue à chaque instant de temps t avec $\mathbf{a}_t = \mathbf{a}_{t-1} + \Delta_{\mathbf{a}_t}$. Pour empêcher des mouvements trop rapides, une contrainte pourrait alors être imposée en définissant $\Delta_{\mathbf{a}_t}$ en fonction de la valeur des neurones de la dernière couche $\mathbf{y}$ du réseau g_v par l'équation :

$$\Delta_{\mathbf{a}_t} = \tanh(\mathbf{y})\Lambda_{\mathbf{a}},$$

dans laquelle la fonction tangente hyperbolique assure que les valeurs de $\Delta_{\mathbf{a}_t}$ restent bornées dans $]-\Lambda_{\mathbf{a}}, \Lambda_{\mathbf{a}}[$. Les valeurs du vecteur $\Lambda_{\mathbf{a}}$ pourraient être définies à partir des enregistrements articulatoires ayant servi à la construction du synthétiseur piloté par l'agent afin de déterminer une vitesse maximale pour chacun des paramètres articulatoires en phase avec la vérité terrain.

6.2.3 Normalisation des locuteurs

Un troisième enjeu concerne la question de la normalisation des sons produits par différents locuteurs pour un même contenu phonétique. Cette question n'a été que peu travaillée dans cette thèse – voir section 5.2.3 – et elle est pourtant essentielle, tant pour le développement de l'agent à but imitatif que de l'agent à but communicatif. La question de la normalisation peut s'envisager à deux niveaux différents, celui de la prise en compte des différences physiques/physiologiques entre locuteurs et de mécanismes permettant de recadrer les sons produits par différents locuteurs dans un espace acoustique commun, et celui de la prise en compte des variations inter-individuelles (accents, idiosyncrasies), et de la définition d'un espace de représentation phonétique cohérent permettant l'accès aux unités phonologiques.

Normalisation acoustique Ce premier niveau est essentiel à considérer par l'agent à but imitatif. En effet, une première façon de permettre à l'agent de compenser les différences entre les locuteurs serait de l'équiper d'un module de normalisation chargé de transformer les signaux

acoustiques perçus. Ce module serait chargé de convertir le timbre de voix des locuteurs perçus pour rapprocher les sons de ceux que peut produire le synthétiseur articulatoire piloté par l'agent. L'agent à but imitatif, dans la version actuelle, mesure sa performance d'imitation entre le son perçu et le son qu'il produit en évaluant la proximité de leurs contenus spectraux. De ce fait, si le son perçu est produit avec une voix trop différente de celle du synthétiseur qu'il pilote, l'agent risque de devoir recourir à des configurations articulatoires non « naturelles » pour compenser cet écart. Il est pourtant admis que les bébés sont très tôt capables d'imiter certaines propriétés des sons qu'ils entendent : voir par exemple les données de X. Chen et al., 2004 montrant que les bébés de 1 à 7 jours semblent capables d'associer des sons /m/ vs. /a/ à des gestes de fermeture vs. ouverture de la bouche et des lèvres ; la célèbre étude de Mampe et al., 2009 montrant que les pleurs de bébés allemands vs. français de 2 à 5 jours sont déjà façonnés par la prosodie de leur langue ; ou les données d'imitation articulatoire de *stimuli* audiovisuels de voyelles par des bébés de 4 mois (Kuhl & Meltzoff, 1996).

La littérature sur les méthodes de normalisation acoustique est ancienne et bien connue (on en trouvera une présentation récente dans Johnson et Sjerps, 2021). On peut les séparer en méthodes « intrinsèques » vs. « extrinsèques ». Dans les méthodes intrinsèques, le signal acoustique à un moment donné est supposé contenir en lui-même les éléments de la normalisation, par exemple par l'appel à la fréquence fondamentale ou l'introduction de rapports entre formants (ou distances dans un espace pseudo logarithmique comme celui des Bark), pour retirer autant que faire se peut l'influence de variations de taille du conduit vocal. Dans les méthodes extrinsèques, on cherche à estimer la taille du conduit vocal d'un locuteur ou à estimer quelques propriétés de la distribution acoustique de ses productions, par un apprentissage préalable sur un corpus de données acoustiques de ce locuteur. Depuis les premières propositions de Lobanov, 1971 ou Nearey, 1978, de nombreuses techniques ont été appliquées d'abord à la reconnaissance automatique de la parole, et plus récemment à l'inversion articulatori-acoustique (Sivaraman et al., 2016). Ces différentes techniques peuvent être aisément implémentées dans l'agent, à but imitatif ou communicatif, pour améliorer ses performances. Néanmoins, il est probable que ce type de normalisation acoustique ne soit pas suffisant pour résoudre complètement les problèmes rencontrés. C'est pourquoi, pour l'agent à but communicatif, un autre niveau de normalisation est probablement requis.

Normalisation phonétique au niveau des unités Dans l'agent à but communicatif, un autre niveau de normalisation doit être introduit, qui concerne, indépendamment des différences physiques ou physiologiques, les variations inter-individuelles de production associées au style, aux accents, aux idiosyncrasies de tous ordres. C'est alors au niveau du système de découverte des unités acoustiques de la parole qu'il faut opérer, et on peut envisager (au moins) trois types de développement.

D'abord, l'utilisation d'architectures capables d'introduire des connaissances structurelles plus larges, de type wave2vec ou HuBERT (Baevski et al., 2020 ; Hsu et al., 2021) en lieu et place du VQ-VAE acoustique actuellement utilisé, pourrait permettre d'extraire de façon autosupervisée des représentations de la parole encodant des informations articulatori-acoustiques assez efficacement reliées aux unités phonémiques (Wells et al., 2022). Ensuite, un certain

nombre de travaux récents dans le domaine de l'apprentissage auto-supervisé de représentations de la parole (Mohamed et al., 2022) introduisent des variables indexicales ou stylistiques (identité, genre, état émotionnel, contexte d'interaction) qui permettent de séparer jusqu'à un certain point les contributions phonétiquement pertinentes de celles qui ne le sont pas, améliorant ainsi la représentation des unités indépendamment des variations inter-individuelles.

Enfin, une piste importante dans le contexte de la réalisation d'un agent autonome consiste à introduire des interactions avec un tuteur, et/ou à réinsérer les informations acoustiques dans le contexte visuel de l'interaction, afin de permettre à l'agent d'associer les productions de différents agents à un même contexte d'interaction, et donc potentiellement à en extraire des informations sémantiques susceptibles de contraindre l'extraction d'informations sur les unités phonologiques.

6.2.4 Temporalités multiples

Les solutions présentées dans les sous-sections précédentes ne remettent pas directement en cause l'hypothèse simplificatrice sur la gestion du temps tout au long des différentes simulations que nous avons mises en œuvre. Cette hypothèse stipule que les représentations manipulées par l'agent possèdent toutes la même temporalité. Autrement dit, lorsque l'agent doit répéter un son représenté par T trames de caractéristiques acoustiques, celui-ci associe à chacune de ces T trames une unité acoustique, un vecteur de paramètres articulatoires et une unité articulatoire. Ce principe entraîne des limitations telles que l'obligation pour l'agent de toujours produire un son de longueur strictement identique au son qu'il cherche à répéter.

Cette sous-section présente une première proposition en vue d'une nouvelle architecture d'agent pour laquelle le son produit peut avoir une dynamique/temporalité différente de celle du son perçu. Cette architecture est toujours en développement et la façon dont elle pourrait être entraînée n'est pas encore totalement définie. La suite de cette sous-section présente d'abord son fonctionnement général et ce qu'elle permettrait avant de discuter des conditions dans lesquelles elle pourrait être entraînée.

Architecture et fonctionnement Le but de ce nouvel agent, dont l'architecture est présentée figure 6.1, reste de « répéter » les sons qu'il perçoit. On peut rester, à ce stade, flou sur ce qu'il s'agit de « répéter », contenu acoustique ou phonétique dans un sens large, pour l'agent à but imitatif, ou contenu informationnel et donc linguistique, pour l'agent à but communicatif – nous y reviendrons. La répétition d'un son se ferait en suivant cette série d'étapes :

1. L'agent perçoit un son en provenance de son environnement.
2. L'agent infère avec son modèle inverse une trajectoire articulatoire de même longueur que la séquence d'observations acoustiques correspondant au son perçu.
3. Ces deux séquences d'observations sont envoyés dans le décodeur d'unités. La séquence des unités est typiquement de longueur plus petite que celle des séquences d'observations acoustiques et articulatoires. Là encore, nous restons à ce stade imprécis sur le contenu de ces unités, nous y reviendrons plus loin.

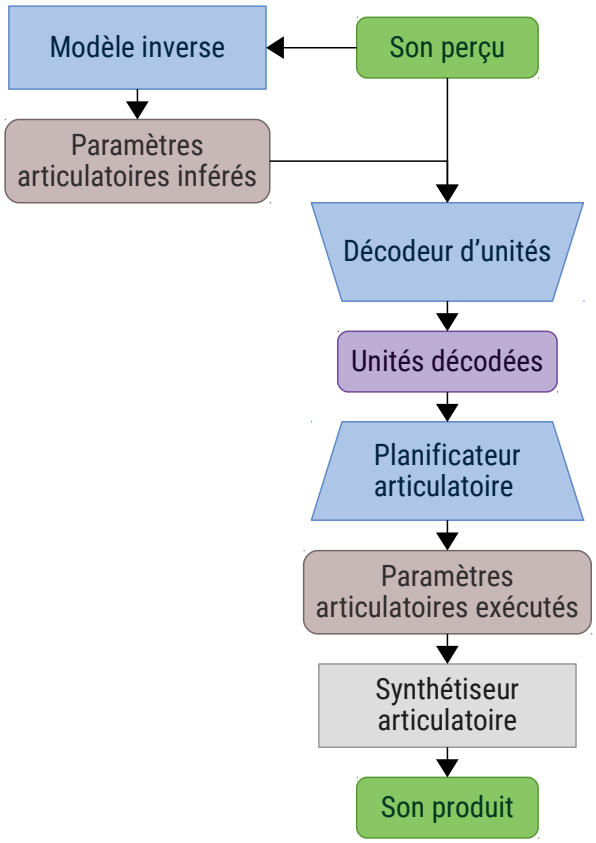

FIGURE 6.1 – **Proposition pour une nouvelle architecture d'agent pour laquelle le son produit peut avoir une dynamique/temporalité différente de celle du son perçu**

4. La série d'unités décodées est ensuite envoyée dans le planificateur articulatoire dont la tâche est de trouver une séquence de paramètres articulatoires dont l'exécution résulterait en un signal acoustique compatible avec ces unités, c'est-à-dire qui générerait la même séquence d'unités s'il était donné en entrée au décodeur d'unités. Cette nouvelle séquence de paramètres articulatoires est typiquement de longueur plus grande que la séquence d'unités.

5. La série de paramètres articulatoires est utilisée pour piloter le synthétiseur articulatoire qui donne en sortie le son produit par l'agent.

Le cœur de cette architecture est la concaténation de deux processus, respectivement le passage d'une séquence d'observations acoustiques à une séquence d'unités, puis le passage de cette séquence d'unité à une trajectoire articulatoire (de taille arbitraire), à destination du synthétiseur. Ces deux processus pourraient typiquement être implémentés par des architectures de type *seq2seq* (séquence à séquence). Ce mode de fonctionnement pourrait être apparenté à la mise en cascade d'un modèle de reconnaissance automatique de parole avec un modèle de synthèse vocale où le texte serait remplacé par des unités auto-découvertes par l'agent et le son de sortie par une trajectoire articulatoire permettant de produire le son désiré.

À ce stade, il est important de revenir sur la nature du processus de répétition, et, en lien, le contenu des unités du décodeur. On peut en effet proposer différents types d'unités, qui tracent un continuum entre répétition « acoustique » ou « acoustico-phonétique » dans un agent à but imitatif, pour lequel les unités restent proches du contenu d'entrée, correspondant par exemple à des « états acoustiques », silence, prévoisement, transition de formant, climax formantique, bruit de friction... ; et répétition du contenu informationnel, dans un agent à but communicatif pour lequel les unités sont proches des unités linguistiques, typiquement des phones.

Entraînement Dans cette nouvelle architecture, trois modèles doivent être entraînés pour permettre à l'agent d'accomplir sa tâche de répétition des sons perçus : le modèle inverse, le décodeur d'unités et le planificateur articulatoire.

La tâche du modèle inverse est de trouver une séquence de paramètres articulatoires dont l'exécution produirait un son proche du son perçu. Cette séquence de paramètres articulatoires inférés étant de longueur identique au son perçu, le fonctionnement du modèle inverse est donc le même que celui de l'agent à but imitatif et de l'agent à but communicatif présentés dans le chapitre précédent. De fait, nous pourrions utiliser ces précédentes architectures pour entraîner un modèle inverse puis l'importer dans cette nouvelle architecture – avec l'objectif d'exploiter ensuite cette entrée articulatoire inférée pour contribuer au bon fonctionnement du décodeur d'unités.

La réelle difficulté se situe au niveau de l'entraînement du décodeur d'unités et du planificateur articulatoire, du fait de la variation de la longueur des séquences présentes en entrée et en sortie de ces modèles. Cette difficulté pourrait être résolue en mettant en place un apprentissage par renforcement dans lequel un score serait calculé en comparant le contenu phonétique du son produit par l'agent à celui du son perçu. Ce score pourrait être calculé par exemple à l'aide d'un modèle de reconnaissance automatique de parole externe jouant le rôle d'un tuteur donnant à l'agent un *feedback* sur la qualité de ses répétitions. En cherchant à maximiser ce score, l'agent pourrait trouver des gestes adaptés pour produire des sons dont le contenu phonétique – dans un sens à définir, dépendant de la nature des unités, voir discussion précédente – soit suffisamment proche de celui du son perçu. Un second objectif devrait être ajouté à cet entraînement afin de régulariser cet espace des unités auto-découvert. Par exemple, nous pouvons imaginer qu'une contrainte poussant l'agent à inférer des séquences d'unités les plus courtes possibles l'inciterait à trouver une façon compacte de représenter l'information pertinente pour produire un son au contenu phonétique proche de celui du son perçu.

6.3 Conclusion générale

Dans ce travail de thèse, nous avons mené différents travaux de modélisation portant sur l'apprentissage auto ou faiblement supervisé des représentations acoustiques, articulatoires et phonétiques de la parole. L'aboutissement de ces travaux est une architecture d'agent

apprenant « à parler », en répétant les *stimuli* auditifs qu'il perçoit. Cet agent s'appuie sur une architecture complète, intégrant un synthétiseur articulatoire, un modèle interne direct et un modèle interne inverse, ainsi qu'un ou plusieurs espaces de représentation phonétique, au sein d'un processus d'apprentissage global *end-to-end* permettant potentiellement d'en apprendre toutes les composantes conjointement à partir d'ensembles de données acoustiques réelles. Cette architecture a, comme discuté précédemment, des limites et des manques, qui suggèrent de nombreuses perspectives que nous avons esquissées. Mais son intérêt majeur, à notre sens, est de s'inscrire au confluent de deux approches, le plus souvent disjointes : celles des théories du contrôle et de la production de la parole et celles des méthodes d'apprentissage automatique, en l'occurrence ici d'apprentissage profond auto-supervisé. À ce titre, on peut, à partir de ce travail, proposer des recherches dans au moins deux types de directions. D'une part, développer chacune des composantes de cette architecture dans une perspective d'apprentissage massif, en intégrant des modules d'apprentissage neuronal de plus en plus sophistiqués et des corpus acoustiques multi-locuteurs de plus en plus fournis. Mais aussi, d'autre part, interroger, à la lumière de ces simulations, des données et hypothèses sur les processus de contrôle, et particulièrement sur le calendrier développemental de l'apprentissage de la parole et du langage chez l'enfant. Ainsi, le lien tracé entre modèles cognitifs et algorithmes d'apprentissage massif nous semble tracer des perspectives prometteuses, à la fois fondamentales et applicatives.

Bibliographie

Alexander, R., Sorensen, T., Toutios, A. & Narayanan, S. (2019). A modular architecture for articulatory synthesis from gestural specification. *The Journal of the Acoustical Society of America*, *146*(6), 4458-4471 (cité en page 20).

Atal, B. S., Chang, J. J., Mathews, M. V. & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, *63*(5), 1535-1555 (cité en page 23).

Avni, R. & Patil, H. (2016). Jerk Minimization for Acoustic-To-Articulatory Inversion. *ISCA Speech Synthesis Workshop (SSW)*, 13-15 (cité en page 25).

Badin, P., Elisei, F., Bailly, G. et al. (2008). Can you "read tongue movements" ? *Conference of the International Speech Communication Association (Interspeech)*, 2635-2637 (cité en page 37).

Badin, P. & Fant, G. (1984). Notes on vocal tract computation. *Speech Transmission Laboratory - Quarterly Progress Status Report*, *25*(2-3), 53-108 (cité en page 22).

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 12449-12460 (cité en pages 31, 32, 132).

Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, *22*(2-3), 251-267 (cité en page 32).

Barnaud, M.-L. (2018). *Modélisation bayésienne du développement conjoint de la perception, l'action et la phonologie* (thèse de doct.). Université Grenoble Alpes. (Cité en page 11).

Barnaud, M.-L., Schwartz, J.-L., Bessière, P. & Diard, J. (2019). Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. *PLOS ONE*, *14*(1), 1-34 (cité en page 15).

Beautemps, D., Badin, P. & Bailly, G. (2001). Linear degrees of freedom in speech production : Analysis of cineradio-and labio-film data and articulatory-acoustic modeling. *The Journal of the Acoustical Society of America*, *109*(5), 2165-2180 (cité en page 21).

Béchet, F. (2001). LIA PHON : un systeme complet de phonétisation de textes. *Traitement Automatique des Langues*, *42*(1), 47-67 (cité en page 38).

Beguš, G., Zhou, A., Wu, P. & Anumanchipalli, G. K. (2022). Articulation GAN : Unsupervised modeling of articulatory learning. *arXiv preprint arXiv :2210.15173* (cité en pages 125, 128).

Ben Youssef, A. (2011). *Contrôle de têtes parlantes par inversion acoustico-articulatoire pour l'apprentissage et la réhabilitation du langage* (thèse de doct.). Université de Grenoble. (Cité en page 37).

Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, *24* (cité en page 126).

Birkholz, P., Jackèl, D. & Kroger, B. J. (2006). Construction and control of a three-dimensional vocal tract model. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *1* (cité en pages 20, 126).

Birkholz, P., Kroger, B. J. & Neuschaefer-Rube, C. (2010). Model-based reproduction of articulatory trajectories for consonant–vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(5), 1422-1433 (cité en pages 20, 126).

Blaauw, M. & Bonada, J. (2016). Modeling and Transforming Speech Using Variational Autoencoders. *Conference of the International Speech Communication Association (Interspeech)*, 1770-1774 (cité en page 72).

Blandin, R., Arnela, M., Félix, S., Doc, J.-B. & Birkholz, P. (2021). Comparison of the Finite Element Method, the Multimodal Method and the Transmission-Line Model for the Computation of Vocal Tract Transfer Functions. *Conference of the International Speech Communication Association (Interspeech)*, 3330-3334 (cité en page 22).

Bocquelet, F. (2017). *Toward a brain-computer interface for speech restoration* (thèse de doct.). Université Grenoble Alpes. (Cité en page 38).

Bocquelet, F., Hueber, T., Girin, L., Badin, P. & Yvert, B. (2014). Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications. *Conference of the International Speech Communication Association (Interspeech)* (cité en page 60).

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C. & Yvert, B. (2016). Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces. *PLOS Computational Biology*, *12*(11), 1-28 (cité en pages 24, 37).

Browman, C. P. & Goldstein, L. (1992). Articulatory phonology : An overview. *Phonetica*, *49*(3-4), 155-180 (cité en page 7).

Buchaillard, S., Perrier, P. & Payan, Y. (2009). A biomechanical model of cardinal vowel production : Muscle activations and the impact of gravity on tongue positioning. *The Journal of the Acoustical Society of America*, *126*(4), 2033-2051 (cité en page 21).

Bullock, D., Grossberg, S. & Guenther, F. H. (1993). A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. *Journal of Cognitive Neuroscience*, *5*(4), 408-435 (cité en page 11).

Calka, M., Perrier, P., Ohayon, J., Grivot-Boichon, C., Rochette, M. & Payan, Y. (2021). Machine-Learning based model order reduction of a biomechanical model of the human tongue. *Computer Methods and Programs in Biomedicine*, *198*, 105786 (cité en page 21).

Callan, D. E., Kent, R. D., Guenther, F. H. & Vorperian, H. K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research*, *43*(3), 721-736 (cité en page 21).

Chen, X., Striano, T. & Rakoczy, H. (2004). Auditory–oral matching behavior in newborns. *Developmental Science*, *7*(1), 42-47 (cité en page 132).

Chen, Y.-W., Hung, K.-H., Chuang, S.-Y., Sherman, J., Huang, W.-C., Lu, X. & Tsao, Y. (2021). EMA2S : An end-to-end multimodal articulatory-to-speech system. *IEEE International Symposium on Circuits and Systems (ISCAS)*, 1-5 (cité en page 129).

Chung, Y.-A., Hsu, W.-N., Tang, H. & Glass, J. (2019). An unsupervised autoregressive model for speech representation learning. *Conference of the International Speech Communication Association (Interspeech)*, 146-150 (cité en pages 31, 32).

Dang, J. & Honda, K. (2004). Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America*, *115*(2), 853-870 (cité en page 21).

Demolin, D. & Metens, T. (2009). L'imagerie par résonance magnétique en temps réel pour l'étude de la parole. In A. Marchal & C. Cavé (Éd.), *L'imagerie Médicale pour L'étude de la Parole* (p. 257-274). Lavoisier. (Cité en page 19).

Diehl, R. L. & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121-144 (cité en page 6).

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., Besacier, L., Sakti, S. & Dupoux, E. (2019). The Zero Resource Speech Challenge 2019 : TTS Without T. *Conference of the International Speech Communication Association (Interspeech)*, 1088-1092 (cité en page 32).

Engwall, O. (2002). *Tongue talking : studies in intraoral speech synthesis* (thèse de doct.). KTH Royal Institute of Technology. (Cité en page 20).

Fant, G. (1970). *Acoustic theory of speech production*. Walter de Gruyter. (Cité en pages 18, 22).

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *Journal of Motor Behavior*, *18*(1), 17-54 (cité en page 17).

Galantucci, B., Fowler, C. A. & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361-377 (cité en page 7).

Georges, M.-A., Badin, P., Diard, J., Girin, L., Schwartz, J.-L. & Hueber, T. (2020). Towards an articulatory-driven neural vocoder for speech synthesis. *International Seminar on Speech Production (ISSP)* (cité en pages 2, 3, 41).

Georges, M.-A., Diard, J., Girin, L., Schwartz, J.-L. & Hueber, T. (2022). Repeat after Me : Self-Supervised Learning of Acoustic-to-Articulatory Mapping by Vocal Imitation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8252-8256 (cité en pages 3, 90).

Georges, M.-A., Girin, L., Schwartz, J.-L. & Hueber, T. (2021). Learning robust speech representation with an articulatory-regularized variational autoencoder. *Conference of the International Speech Communication Association (Interspeech)*, 3345-3349 (cité en pages 2, 3, 72).

Georges, M.-A., Schwartz, J.-L. & Hueber, T. (2022). Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE. *Conference of the International Speech Communication Association (Interspeech)*, 774-778 (cité en pages 3, 79).

Ghosh, P. K. & Narayanan, S. (2010). A generalized smoothness criterion for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, *128*(4), 2162-2172 (cité en page 24).

Ghosh, P. K. & Narayanan, S. S. (2011). A subject-independent acoustic-to-articulatory inversion. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4624-4627 (cité en page 25).

Girin, L., Hueber, T. & Alameda-Pineda, X. (2017). Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *25*(3), 662-673 (cité en page 25).

Gosztolya, G., Pintér, Á., Tóth, L., Grósz, T., Markó, A. & Csapó, T. G. (2019). Autoencoder-based articulatory-to-acoustic mapping for ultrasound silent speech interfaces. *International Joint Conference on Neural Networks (IJCNN)*, 1-8 (cité en pages 23, 24).

Govalkar, P., Fischer, J., Zalkow, F. & Dittmar, C. (2019). A comparison of recent neural vocoders for speech signal reconstruction. *ISCA Speech Synthesis Workshop (SSW)*, 7-12 (cité en page 27).

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*(5), 350-365 (cité en page 8).

Guenther, F. H., Hampson, M. & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*(4), 611-633 (cité en page 8).

Guenther, F. H. & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, *25*(5), 408-422 (cité en page 13).

Harris, C. M. & Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature*, *394*(6695), 780-784 (cité en page 9).

Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504-507 (cité en page 29).

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780 (cité en page 92).

Hogan, N. (1984). An organizing principle for a class of voluntary movements. *Journal of Neuroscience*, *4*(11), 2745-2754 (cité en pages 9, 25, 130).

Houde, J. F. & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, *5*, 82 (cité en page 14).

Howard, I. S. & Messum, P. (2014). Learning to Pronounce First Words in Three Languages : An Investigation of Caregiver and Infant Behavior Using a Computational Model of an Infant. *PLOS ONE*, *9*(10), 1-21 (cité en page 32).

Howard, I. S. & Messum, P. R. (2007). A computational model of infant speech development. *International Conference on Speech and Computer (SPECOM)*, 756-765 (cité en page 32).

Howard, I. S. & Messum, P. R. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, *15*(1), 85-117 (cité en page 32).

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R. & Mohamed, A. (2021). HuBERT : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460 (cité en pages 31, 32, 132).

Hsu, W.-N., Zhang, Y. & Glass, J. (2017). Learning latent representations for speech generation and transformation. *Conference of the International Speech Communication Association (Interspeech)*, 1273-1277 (cité en pages 30, 72).

Hueber, T. & Denby, B. (2009). Analyse du conduit vocal par imagerie ultrasonore. In A. Marchal & C. Cavé (Éd.), *L'imagerie Médicale pour L'étude de la Parole* (p. 147-174). Lavoisier. (Cité en page 19).

Hueber, T., Girin, L., Alameda-Pineda, X. & Bailly, G. (2015). Speaker-Adaptive Acoustic-Articulatory Inversion using Cascaded Gaussian Mixture Regression. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *23*(12), 2246-2259 (cité en page 25).

Imai, S. (1983). Cepstral analysis synthesis on the Mel frequency scale. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *8*, 93-96 (cité en page 27).

Ishihara, H., Yoshikawa, Y., Miura, K. & Asada, M. (2008). Caregiver's sensorimotor magnets lead infant's vowel acquisition through auto mirroring. *International Conference on Development and Learning (ICDL)*, 49-54 (cité en page 32).

ITU. (2015). *Method for the subjective assessment of intermediate quality level of audio systems* (rapp. tech. ITU-R BS.1534-3). International Telecommunication Union. (Cité en pages 66, 77).

Jillings, N., Moffat, D., De Man, B. & Reiss, J. D. Web Audio Evaluation Tool : A browser-based listening test environment. In : *Sound and Music Computing Conference (SMC)*. 2015, juillet (cité en page 66).

Johnson, K. & Sjerps, M. J. (2021). Speaker normalization in speech perception. *The Handbook of Speech Perception*, 145-176 (cité en page 132).

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S. & Kavukcuoglu, K. (2018). Efficient neural audio synthesis. *International Conference on Machine Learning (ICML)*, 2410-2419 (cité en page 27).

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *9*(6), 718-727 (cité en page 9).

Kello, C. T. & Plaut, D. C. (2004). A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, *116*(4), 2354-2364 (cité en page 23).

Kingma, D. P. & Welling, M. (2014). Auto-Encoding Variational Bayes. *International Conference on Learning Representations (ICLR)* (cité en pages 29, 72).

Kröger, B. J., Kannampuzha, J. & Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics*, *2*(1), 1-28 (cité en page 32).

Kuhl, P. K. (1979). Speech perception in early infancy : Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, *66*(6), 1668-1679 (cité en page 116).

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*(2-3), 263-285 (cité en page 116).

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850-11857 (cité en page 6).

Kuhl, P. K. & Meltzoff, A. N. (1996). Infant vocalizations in response to speech : Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, *100*(4), 2425-2438 (cité en page 132).

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A. et al. (2021). On Generative Spoken Language Modeling from Raw Audio. *Transactions of the Association for Computational Linguistics*, *9*, 1336-1354 (cité en page 32).

Laprie, Y., Elie, B., Tsukanova, A. & Vuissoz, P.-A. (2018). Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech. *European Signal Processing Conference (EUSIPCO)*, 2110-2114 (cité en pages 20, 21).

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431-461 (cité en page 88).

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36 (cité en page 6).

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606-608 (cité en page 132).

Lu, C., Nakai, T. & Suzuki, H. (1993). Finite element simulation of sound transmission in vocal tract. *Journal of the Acoustical Society of Japan*, *14*(2), 63-72 (cité en page 22).

MacNeilage, P. F. & Davis, B. L. (2001). Motor mechanisms in speech ontogeny : Phylogenetic, neurobiological and linguistic implications. *Current Opinion in Neurobiology*, *11*(6), 696-700 (cité en page 114).

Maeda, S. (1990). Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. *Speech Production and Speech Modelling* (p. 131-149). Springer. (Cité en pages 20, 21, 42).

Maeda, S. & Honda, K. (1994). From EMG to formant patterns of vowels : The implication of vowel spaces. *Phonetica*, *51*(1-3), 17-29 (cité en page 21).

Makhoul, J. (1975). Linear prediction : A tutorial review. *Proceedings of the IEEE*, *63*(4), 561-580 (cité en page 28).

Mampe, B., Friederici, A. D., Christophe, A. & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology*, *19*(23), 1994-1997 (cité en page 132).

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner : Trainable Text-Speech Alignment Using Kaldi. *Conference of the International Speech Communication Association (Interspeech)*, 498-502 (cité en page 37).

Ménard, L., Schwartz, J.-L. & Boë, L.-J. (2004). Role of Vocal Tract Morphology in Speech Development. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1059-1080 (cité en page 21).

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, *53*(4), 1070-1082 (cité en page 20).

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781* (cité en page 28).

Miura, K., Yoshikawa, Y. & Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. *Advanced Robotics*, *21*(13), 1583-1600 (cité en page 32).

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L. et al. (2022). Self-supervised speech representation learning : A review. *IEEE Journal of Selected Topics in Signal Processing* (cité en page 133).

Morita, T. & Koda, H. (2020). Exploring TTS Without T Using Biologically/Psychologically Motivated Neural Network Modules (ZeroSpeech 2020). *Conference of the International Speech Communication Association (Interspeech)*, 4856-4860 (cité en page 32).

Moulin-Frier, C., Brochard, J., Stulp, F. & Oudeyer, P.-Y. (2017). Emergent Jaw Predominance in Vocal Development through Stochastic Optimization. *IEEE Transactions on Cognitive and Developmental Systems*, *12*(3), 378-389 (cité en page 12).

Moulin-Frier, C., Diard, J., Schwartz, J.-L. & Bessière, P. (2015). COSMO ("Communicating about Objects using Sensory–Motor Operations") : A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics*, *53*, 5-41 (cité en page 16).

Moulin-Frier, C., Nguyen, S. M. & Oudeyer, P.-Y. (2014). Self-organization of early vocal development in infants and machines : the role of intrinsic motivation. *Frontiers in Psychology*, *4* (cité en page 32).

Moulin-Frier, C. & Oudeyer, P.-Y. (2013). Exploration strategies in developmental robotics : a unified probabilistic framework. *International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 1-6 (cité en pages 11, 12).

Mullen, J., Howard, D. M. & Murphy, D. T. (2007). Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(2), 577-585 (cité en page 22).

Murakami, M., Kröger, B., Birkholz, P. & Triesch, J. (2015). Seeing [u] aids vocal learning : Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing. *International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 208-213 (cité en page 32).

Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club. (Cité en page 132).

Neiberg, D. & Ananthakrishnan, G. (2008). On the Non-uniqueness of Acoustic-to-Articulatory Mapping. *FONETIK 2008*, 9-13 (cité en page 23).

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, *46*(2), 135-147 (cité en pages 8, 9).

Nguyen, S. M. & Oudeyer, P.-Y. (2012). Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner. *Paladyn*, *3*(3), 136-146 (cité en page 12).

Niikawa, T., Matsumura, M., Tachimura, T. & Wada, T. (2000). Modeling of a speech production system based on MRI measurement of three-dimensional vocal tract shapes during fricative consonant phonation. *International Conference on Spoken Language Processing (ICSLP)*, *2*, 174-177 (cité en page 22).

Ouni, S. & Laprie, Y. (2005). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, *118 (1)*, 444-460 (cité en page 24).

Pagliarini, S., Leblois, A. & Hinaut, X. (2020). Vocal imitation in sensorimotor learning models : a comparative review. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(2), 326-342 (cité en pages 11, 32).

Palan, S. & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27 (cité en pages 66, 78, 99).

Parrell, B., Ramanarayanan, V., Nagarajan, S. & Houde, J. (2018). FACTS : A Hierarchical Task-based Control Model of Speech Incorporating Sensory Feedback. *Conference of*

the International Speech Communication Association (Interspeech), 1497-1501 (cité en page 15).

Parrell, B., Ramanarayanan, V., Nagarajan, S. & Houde, J. (2019). The FACTS model of speech motor control : Fusing state estimation and task-based control. *PLOS Computational Biology*, *15*(9), 1-26 (cité en page 15).

Parrot, M., Millet, J. & Dunbar, E. (2020). Independent and automatic evaluation of speaker-independent acoustic-to-articulatory reconstruction. *Conference of the International Speech Communication Association (Interspeech)* (cité en page 26).

Patri, J.-F., Diard, J. & Perrier, P. (2015). Optimal speech motor control and token-to-token variability : a Bayesian modeling approach. *Biological Cybernetics (Modeling)*, *109*(6), 611-626 (cité en page 17).

Patri, J.-F., Diard, J. & Perrier, P. (2019). Modeling sensory preference in speech motor planning : a Bayesian modeling framework. *Frontiers in Psychology*, *10*, 2339 (cité en page 17).

Payan, Y. & Perrier, P. (1997). Synthesis of VV sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. *Speech Communication*, *22*(2-3), 185-205 (cité en page 17).

Perrier, P. (2005). Control and representations in speech production. *ZAS Papers in Lingustics*, *40*, 109-132 (cité en page 8).

Perrier, P. (2006). About speech motor control complexity. In J. Harrington & M. Tabain (Éd.), *Speech Production : Models, Phonetic Processes, and Techniques* (p. 13-26). Psychology Press : New-York, USA. (Cité en page 9).

Perrier, P., Payan, Y., Zandipour, M. & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants : A modeling study. *The Journal of the Acoustical Society of America*, *114*(3), 1582-1599 (cité en pages 17, 21).

Perrotin, O., El Amouri, H., Bailly, G. & Hueber, T. (2021). Evaluating the extrapolation capabilities of neural vocoders to extreme pitch values. *Conference of the International Speech Communication Association (Interspeech)*, 11-15 (cité en page 27).

Philippsen, A. (2021). Goal-directed exploration for learning vowels and syllables : a computational model of speech acquisition. *KI-Künstliche Intelligenz*, *35*(1), 53-70 (cité en page 32).

Philippsen, A. K., Reinhart, R. F. & Wrede, B. (2014). Learning how to speak : Imitation-based refinement of syllable production in an articulatory-acoustic model. *International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 195-200 (cité en page 32).

Pitti, A., Quoy, M., Boucenna, S. & Lavandier, C. (2021). Brain-inspired model for early vocal learning and correspondence matching using free-energy optimization. *PLOS Computational Biology*, *17*(2), 1-27 (cité en page 32).

Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A. & Dupoux, E. (2021). Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. *Conference of the International Speech Communication Association (Interspeech)*, 3615-3619 (cité en page 32).

Qin, C. & Carreira-Perpinán, M. A. (2007). An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. *Speech Communication* (cité en page 23).

Rasilo, H. & Räsänen, O. (2017). An online model for vowel imitation learning. *Speech Communication*, *86*, 1-23 (cité en page 32).

Rasilo, H., Räsänen, O. & Laine, U. K. (2013). Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, *55*(9), 909-931 (cité en pages 32, 130).

Rezende, D. J., Mohamed, S. & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning (ICML)*, 1278-1286 (cité en page 72).

Richmond, K., King, S. & Taylor, P. (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech & Language*, *17*(2-3), 153-172 (cité en page 23).

Roche, F., Hueber, T., Garnier, M., Limier, S. & Girin, L. (2021). Make That Sound More Metallic : Towards a Perceptually Relevant Control of the Timbre of Synthesizer Sounds Using a Variational Autoencoder. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *4*, 52-66 (cité en page 73).

Rolf, M., Steil, J. J. & Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. *IEEE Transactions on Autonomous Mental Development*, *2*(3), 216-229 (cité en page 11).

Saha, P. & Fels, S. (2020). Learning Joint Articulatory-Acoustic Representations with Normalizing Flows. *Conference of the International Speech Communication Association (Interspeech)*, 3196-3200 (cité en page 23).

Savariaux, C., Perrier, P. & Orliaguet, J. P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube : A study of the control space in speech production. *The Journal of the Acoustical Society of America*, *98*(5), 2428-2442 (cité en page 8).

Schatz, T. (2016). *ABX-Discriminability Measures and Applications* (thèse de doct.). Université Paris 6 (UPMC). (Cité en page 83).

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J. & Conrad, B. (1987). Electromagnetic articulography : Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, *31*(1), 26-35 (cité en page 19).

Schwartz, J.-L., Basirat, A., Ménard, L. & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT) : A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336-354 (cité en page 7).

Serrurier, A., Badin, P., Barney, A., Boë, L.-J. & Savariaux, C. (2012). The tongue in speech and feeding : Comparative articulatory modelling. *Journal of Phonetics*, *40*(6), 745-763 (cité en page 42).

Serrurier, A., Badin, P., Lamalle, L. & Neuschaefer-Rube, C. (2019). Characterization of inter-speaker articulatory variability : a two-level multi-speaker modelling approach based on MRI data. *The Journal of the Acoustical Society of America*, *145*(4), 2149-2170 (cité en page 20).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R. et al. (2018). Natural TTS Synthesis by Conditioning WaveNet

on Mel Spectrogram Predictions. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779-4783 (cité en page 32).

Shiga, Y. & King, S. (2004). Accurate spectral envelope estimation for articulation-to-speech synthesis. *ISCA Speech Synthesis Workshop (SSW)*, 19-24 (cité en page 23).

Siriwardena, Y. M., Espy-Wilson, C. & Shamma, S. (2022). Learning to Compute the Articulatory Representations of Speech with the MIRRORNET. *arXiv preprint arXiv :2210.16454* (cité en page 126).

Sisman, B., Yamagishi, J., King, S. & Li, H. (2020). An overview of voice conversion and its challenges : From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 132-157 (cité en page 117).

Sivaraman, G., Mitra, V., Nam, H., Tiede, M. & Espy-Wilson, C. (2019). Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion. *The Journal of the Acoustical Society of America*, *146*(1), 316-329 (cité en page 26).

Sivaraman, G., Mitra, V., Nam, H., Tiede, M. K. & Espy-Wilson, C. Y. (2016). Vocal Tract Length Normalization for Speaker Independent Acoustic-to-Articulatory Speech Inversion. *Conference of the International Speech Communication Association (Interspeech)*, 455-459 (cité en page 132).

Story, B. H., Vorperian, H. K., Bunton, K. & Durtschi, R. B. (2018). An age-dependent vocal tract model for males and females based on anatomic measurements. *The Journal of the Acoustical Society of America*, *143*(5), 3079-3102 (cité en page 20).

Tang, Q., Wang, W. & Livescu, K. (2018). Acoustic feature learning using cross-domain articulatory measurements. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4849-4853 (cité en page 26).

Thompson, M., Houde, J. F. & Nagarajan, S. S. (2019). Learning and adaptation in speech production without a vocal tract. *Scientific Reports*, *9*(1), 1-11 (cité en page 10).

Tjandra, A., Sakti, S. & Nakamura, S. (2020). Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 976-989 (cité en page 32).

Toda, T., Black, A. & Tokuda, K. (2004). Acoustic-to-articulatory inversion mapping with Gaussian mixture model. *International Conference on Spoken Language Processing (ICSLP)* (cité en page 24).

Toda, T., Black, A. W. & Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, *50*(3), 215-227 (cité en page 24).

Tourville, J. A. & Guenther, F. H. (2011). The DIVA model : A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952-981 (cité en page 13).

Turrisi, R., Tavarone, R. & Badino, L. (2018). Improving generalization of vocal tract feature reconstruction : from augmented acoustic inversion to articulatory feature reconstruction without articulatory data. *IEEE Spoken Language Technology Workshop (SLT)*, 159-166 (cité en page 26).

Valin, J.-M. & Skoglund, J. (2019). LPCNet : Improving neural speech synthesis through linear prediction. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5891-5895 (cité en pages 27, 28, 61).

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016). Wavenet : A generative model for raw audio. *arXiv preprint arXiv :1609.03499* (cité en page 27).

van den Oord, A., Li, Y. & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv :1807.03748* (cité en pages 31, 32).

van den Oord, A., Vinyals, O. & Kavukcuoglu, K. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, *30* (cité en pages 30, 79, 107).

van Niekerk, D. R., Xu, A., Gerazov, B., Krug, P. K., Birkholz, P., Halliday, L., Prom-on, S. & Xu, Y. (2023). Simulating vocal learning of spoken language : Beyond imitation. *Speech Communication* (cité en pages 126, 130).

van Vugt, F. T. & Ostry, D. J. (2018). From known to unknown : moving to unvisited locations in a novel sensorimotor map. *Annals of the New York Academy of Sciences*, *1423*(1), 368-377 (cité en pages 9, 10).

van Vugt, F. T. & Ostry, D. J. (2019). Early stages of sensorimotor map acquisition : learning with free exploration, without active movement or global structure. *Journal of Neurophysiology*, *122*(4), 1708-1720 (cité en page 10).

Wells, D., Tang, H. & Richmond, K. (2022). Phonetic Analysis of Self-supervised Representations of English Speech. *Conference of the International Speech Communication Association (Interspeech)*, 3583-3587 (cité en page 132).

Westbury, J. R., Turner, G. & Dembowski, J. (1994). X-ray microbeam speech production database user's handbook. *University of Wisconsin* (cité en page 26).

Wolpert, D. M. & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*(11), 1212-1217 (cité en pages 9, 11, 15).

Wrench, A. (1999). The MOCHA-TIMIT articulatory database. (Cité en page 39).

Wu, P., Watanabe, S., Goldstein, L., Black, A. W. & Anumanchipalli, G. K. (2022). Deep Speech Synthesis from Articulatory Representations. *Conference of the International Speech Communication Association (Interspeech)*, 779-783 (cité en page 129).

Zen, H., Nankaku, Y. & Tokuda, K. (2010). Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(2), 417-430 (cité en page 24).

Self-supervised learning of the relationships between sounds, gestures and units for the control of speech production: towards an agent learning to speak

Abstract — This thesis aims to study, through modeling and simulation, the learning mechanisms of the relationships between the speech sounds, the underlying articulatory gestures and the phonetic units. The employed methodology is based on deep learning, with an emphasis on self or weakly supervised learning, a paradigm that approaches (to some extent) human learning. To this end, we propose a computational agent capable of learning "to speak" in a self-supervised manner, solely from speech sounds from its environment. First, in order to make the agent capable of producing good quality speech sounds, we develop an articulatory synthesizer, exploiting articulatory and acoustic recordings of a reference speaker and driven by a limited number of parameters describing the main degrees of freedom of the vocal apparatus. Then, we propose two studies aiming to quantify the contribution of articulatory information on the learning of speech representations. In the first study, we simulate the access to articulatory representations during speech perception by evaluating, on a denoising task, the addition of articulatory constraints on the latent space of a variational autoencoder (VAE). In a second study, we investigate the self-supervised discovery of discrete phonetic units using vector quantized variational autoencoders (VQ-VAEs). We show the complementarity of acoustic and articulatory information for structuring the units dictionary. Finally, we propose two versions of the full computational agent, the first referred to as an "imitative agent" and the second as a "communicative agent". These two types of agents must learn to speak in a self-supervised way, by repeating the speech sounds they perceive, by driving the articulatory synthesizer previously developed. To do so, they are provided with two internal models, respectively direct and inverse, which represent the way the brain internalizes the complex relations between the spectral content of the speech signals on the one hand, and the associated articulatory trajectories on the other hand. The imitative agent aims to produce repetitions whose spectral content is as close as possible to that of the perceived sounds. The architecture of the communicative agent adds to the previous one two phonetic unit discovery modules, one based on acoustic information and the other on articulatory information inferred by the agent. These different modules are trained jointly from acoustic *stimuli* provided by different speakers. Both types of agents appear to be able to learn to speak, but present a certain number of limitations which open up many perspectives for future developments.

Keywords: Speech production, computational models, articulatory synthesis, representation learning.

Apprentissage auto-supervisé des relations entre sons, gestes articulatoires et unités de la parole pour le contrôle de la production : vers un agent apprenant à parler

Résumé — Ce travail de thèse vise à étudier, par le biais de la modélisation et de la simulation, les mécanismes d'apprentissage des relations entre les sons de la parole, les gestes articulatoires sous-jacents et les unités phonétiques. La méthodologie employée est basée sur l'apprentissage automatique profond (*deep learning*), avec un accent sur l'apprentissage auto ou faiblement supervisé (*self-supervised learning*), paradigme qui s'approche (dans une certaine mesure) de l'apprentissage humain. Pour ce faire, nous proposons un agent computationnel capable d'apprendre « à parler » de façon auto-supervisée, uniquement à partir de sons de parole issus de son environnement. D'abord, afin de rendre l'agent capable de produire des sons de parole de bonne qualité, nous élaborons un synthétiseur articulatoire, exploitant des enregistrements articulatoires et acoustiques d'un locuteur de référence et piloté par un nombre restreint de paramètres décrivant les degrés de liberté principaux de l'appareil vocal. Ensuite, nous proposons deux études visant à quantifier l'apport d'informations articulatoires sur l'apprentissage de représentations de la parole. Dans la première étude, nous simulons l'accès à des représentations articulatoires lors de la perception de la parole en évaluant, sur une tâche de débruitage, l'ajout de contraintes articulatoires sur l'espace latent d'un auto-encodeur variationnel (VAE). Dans une seconde étude, nous nous intéressons à la découverte auto-supervisée d'unités phonétiques discrètes, grâce à des auto-encodeurs variationnels quantifiés vectoriels (VQ-VAE). Nous montrons une complémentarité des informations acoustiques et articulatoires pour la structuration du dictionnaire d'unités. Enfin, nous proposons deux versions de l'agent computationnel complet, la première qualifiée d'« agent à but imitatif » et la seconde d'« agent à but communicatif ». Ces deux types d'agents doivent apprendre à parler de façon auto-supervisée, en répétant les sons de parole qu'ils perçoivent, au moyen du synthétiseur articulatoire développé préalablement. Pour ce faire, ils sont dotés de deux modèles internes respectivement direct et inverse qui représentent la façon dont le cerveau internalise les relations complexes entre le contenu spectral des signaux de parole d'une part, et les trajectoires articulatoires associées d'autre part. L'agent imitatif cherche à produire des répétitions dont le contenu spectral est le plus proche possible de celui des sons perçus. L'architecture de l'agent à but communicatif ajoute à la précédente deux modules de découverte d'unités phonétiques, l'un basé sur les informations acoustiques et l'autre sur les informations articulatoires inférées par l'agent. Ces différents modules sont entraînés conjointement à partir de *stimuli* acoustiques fournis par différents locuteurs. Ces deux types d'agents apparaissent effectivement capables d'apprendre à parler, mais présentent un certain nombre de limitations qui ouvrent sur de nombreuses perspectives pour des développements futurs.

Mots clés : Production de la parole, modèles computationels, synthèse articulatoire, apprentissage de représentations.