Modélisation de la coarticulation multimodale : vers l'animation d'une tête parlante intelligible

Théo Biasutto-Lervat

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Modélisation de la Coarticulation Multimodale

Vers l'animation d'une tête parlante intelligible

THÈSE

présentée et soutenue publiquement le 29 Janvier 2021

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Théo BIASUTTO--LERVAT

Composition du jury

Rapporteurs :	M. LABOISSIERE Rafaël	Chargé de Recherche, LPNC (UMR 5105)
	M. CHETOUANI Mohamed	Professeur, IRIS (UMR 722)
Examinateur :	Mme DEBLED-RENNESSON Isabelle	Professeur, Loria (UMR 7503)
Encadrant :	M. SLIM Ouni	Maître de Conférences, Loria (UMR 7503)

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Je tiens tout d'abord à remercier chaleureusement mon directeur de thèse, Slim Ouni, pour m'avoir accordé la confiance nécessaire au départ de cette aventure. Sa disponibilité, sa bonne humeur, ainsi que ses conseils avisés furent des moteurs indispensables de ces quatre dernières années.

Je profite de cette section pour remercier mes collègues de l'équipe MULTISPEECH, en particulier les occupants des salles C131 et C141, pour l'ambiance de travail et nos intéressants échanges, Vincent Colotte pour ses conseils de composition d'un corpus textuel, et Anne Bonneau pour son expertise précieuse en phonétique.

Je salue également mes anciens collègues de l'équipe MaIA et Larsen, en particulier Alain Dutech pour m'avoir donné goût à la recherche et aux réseaux de neurones récurrents deux ans avant le début de cette thèse, et Vincent Thomas pour les innombrables discussions sur les jeux de société, ainsi que mes camarades de l'équipe Kiwi, Amaury, Benjamin et Yacine, pour tous les bons moments partagés à la cantine et autour d'une tasse de thé entre 2014 et 2020.

Je dédie cette thèse à ma famille, en maigre remerciement pour leur soutien inconditionnel. Je pense particulièrement à mes parents pour la merveilleuse éducation dont j'ai pu bénéficier, à ma petite soeur et mes amis pour nos longues heures à pousser des cubes en bois, ainsi qu'à Morgane, dont la présence fut une incroyable source de réconfort pendant les périodes les plus difficiles de ces quatre dernières années.

Rien de ceci n'aurait été réalisable sans vous.

Résumé

Nous traitons dans cette thèse la modélisation de la coarticulation par les réseaux de neurones, dans l'objectif de synchroniser l'animation d'un visage virtuel 3D à de la parole. La prédiction de ces mouvements articulatoires n'est pas une tâche triviale, en effet, il est bien établi en production de parole que la réalisation d'un phonème est largement influencée par son contexte phonétique, phénomène appelé coarticulation. Nous proposons dans cette thèse un modèle de coarticulation, c'est-à-dire un modèle qui prédit les trajectoires spatiales des articulateurs à partir de la parole. Nous exploiterons pour cela un modèle séquentiel, les réseaux de neurones récurrents (RNN), et plus particulièrement les *Gated Recurrent Units*, capables de considérer la dynamique de l'articulation au cœur de leur modélisation. Malheureusement, la quantité de données classiquement disponible dans les corpus articulatoires et audiovisuels semblent de prime abord faibles pour une approche *deep learning*. Pour pallier cette difficulté, nous proposons une stratégie permettant de fournir au modèle des connaissances sur les gestes articulatoires du locuteur dès son initialisation. La robustesse des RNNs nous a permis d'implémenter notre modèle de coarticulation pour prédire les mouvements des lèvres pour le français et l'allemand, et de la langue pour l'anglais et l'allemand. L'évaluation du modèle fut réalisée par le biais de mesures objectives de la qualité des trajectoires et par des expériences permettant de valider la bonne réalisation des cibles articulatoires critiques. Nous avons également réalisé une évaluation perceptive de la qualité de l'animation des lèvres du visage parlant. Enfin, nous avons conduit une analyse permettant d'explorer les connaissances phonétiques acquises par le modèle après apprentissage.

Abstract

This thesis deals with neural network based coarticulation modeling, and aims to synchronize facial animation of a 3D talking head with speech. Predicting articulatory movements is not a trivial task, as it is well known that production of a phoneme is greatly affected by its phonetic context, a phoneme called coarticulation. We propose in this work a coarticulation model, i.e. a model able to predict spatial trajectories of articulators from speech. We rely on a sequential model, the recurrent neural networks, and more specifically the *Gated Recurrent Units*, which are able to consider the articulation dynamic as a central component of its modeling. Unfortunately, the typical amount of data in articulatory and audiovisual databases seems to be quite low for a *deep learning* approach. To overcome this difficulty, we propose to integrate articulatory knowledge into the networks during its initialization. The RNNs robustness allow uw to apply our coarticulation model to predict both face and tongue movements, in french and german for the face, and in english and german for the tongue. Evaluation has been conducted through objective measures of the trajectories, and through experiments to ensure a complete reach of critical articulatory targets. We also conducted a subjective evaluation to attest the perceptual quality of the predicted articulation once applied to our facial animation system. Finally, we analyzed the model after training to explore phonetic knowledges learned.

Sommaire

Introduction

La parole est certainement le moyen de communication le plus usité à travers le monde pour la vie quotidienne, ainsi que l'un des facteurs de l'incroyable développement de l'humanité à travers les âges. Cet outil formidable permet de transformer des pensées en concepts (conceptualisation), de manipuler des mots afin de former des phrases véhiculant ces idées (formalisation), et d'articuler ces phrases grâce à un système biomécanique complexe capable de créer un flux d'air, des vibrations, résonnances et modulations nécessaires à la réalisation acoustique de la parole (articulation) (Levelt, 1993). La parole ne peut cependant pas être réduite à un simple signal acoustique. En effet, de fortes interactions existent entre la modalité acoustique de la parole et sa contrepartie visuelle, c'est-à-dire entre le signal acoustique et les déformations du visage induit par les variations articulatoires. La dynamique du visage, et plus particulièrement des lèvres, des joues et de la mâchoire, est une conséquence directe de l'évolution temporelle du conduit vocal. Pour ces raisons, nous considérons la parole comme multimodale comprenant les modalités articulatoire, acoustique et visuelle, et nous désignons l'ensemble de ces trois modalités par parole multimodale. De la même manière, nous utiliserons la notion d'articulation multimodale pour référer simultanément aux mouvements du conduit vocal (modalité articulatoire) et aux déformations du visage lié à la parole (modalité visuelle).

De nombreuses études ont démontré l'importance de l'information visuelle pour la perception de la parole. En plus de permettre la communication d'informations de haut niveau comme les émotions ou la métacognition (Swerts and Krahmer, 2005; Granström and House, 2005), il a été établi que lorsque le signal acoustique est dégradé, la modalité visuelle apportée par le visage peut rétablir jusqu'à deux tiers de l'intelligibilité apportée par l'audio (Sumby and Pollack, 1954; Le Goff et al., 1994). Pour de nombreuses applications exploitant actuellement des technologies de synthèse de la parole, l'ajout de la modalité visuelle par le biais d'une tête parlante virtuelle permettrait donc d'augmenter l'intelligibilité de la parole synthétique, et ce même sans présence de modélisation interne (absence des mouvements de la langue) (Ouni et al., 2007). L'ajout de la modalité visuelle est cependant une tâche critique, car animer un visage virtuel peut se faire au détriment

de l'intelligibilité si le signal visuel n'est pas parfaitement congru au signal acoustique. En effet, lorsque nous observons un locuteur, nous utilisons un système neurologique de décodage multimodal, où la modalité visuelle influence notre compréhension de la modalité acoustique, et inversement (Skipper et al., 2007; Benoit et al., 2010). Ce mécanisme entraine une plus grande robustesse de la parole aux perturbations extérieures de par la redondance des informations au niveau visuel et acoustique, mais engendre néanmoins une grande sensibilité de l'humain à la moindre incohérence entre les deux modalités de la parole. Que ces incohérences soient dues à une mauvaise synchronisation entre le flux audio et visuel (Dixon and Spitz, 1980), ou à une distorsion phonétique (Green and Kuhl, 1989, 1991; Jiang et al., 2002b), celles-ci peuvent aboutir à d'importants effets sur la perception. L'exemple le plus notable est certainement l'effet McGurk (McGurk and MacDonald, 1976) : quand le stimulus audio 'ba' est couplé à un stimulus visuel 'ga', l'auditeur rapporte entendre prononcer 'da'.

Malgré ces difficultés, le développement de technologies de synthèse audiovisuelle de la parole pourrait être crucial pour les malentendantes qui exploitent bien plus le signal visuel qu'une vaste majorité de la population (MacSweeney et al., 2002; Campbell et al., 1998), mais également d'une utilité plus générale dans les lieux bruyants, comme les gares ou aéroports. De plus, l'intégration d'avatar virtuel doué de parole peut améliorer l'expérience de l'utilisateur dans nombreux cadres, comme les assistants virtuels, les sites internet ou les médias sociaux (Gibbs et al., 1993; Cosatto et al., 2003). Dans le secteur du divertissement, la synthèse audiovisuelle de la parole pourrait considérablement accélérer la réalisation de film d'animation en automatisant la production d'animation liée à la parole, et il en va de même pour l'industrie vidéoludique. Elle peut également être utilisée à des fins pédagogiques pour aider à capter l'attention de l'apprenant (Johnson et al., 2000), ou à l'apprentissage de la prononciation des langues étrangères (Hazan et al., 2005; Massaro, 2003). Cette utilisation de la synthèse audiovisuelle pour l'apprentissage des langues étrangères peut également être étendue au domaine médical, principalement comme outil de démonstration et de visualisation en orthophonie. En plus d'améliorer la capacité à transmettre des informations, qui est sans conteste l'objectif premier de la parole, l'utilisation d'un visage virtuel capable de parler rend l'interaction avec la machine plus naturelle (Pandzic et al., 1999; Sproull et al., 1996), ce qui renforce le confort de l'utilisateur (Dehn and Van Mulken, 2000) et sa confiance dans le système (Ostermann and Millen, 2000). Les utilisateurs interagissant avec un système informatique par le biais d'une tête parlante réagissent donc plus positivement (Pandzic et al., 1999) et sont plus engagés (Walker et al., 1994; Sproull et al., 1996).

Pour produire une modalité visuelle intelligible par le numérique, il semble donc pri-

mordial de parfaitement comprendre et modéliser la dynamique de nos articulateurs pendant la production de la parole. Nous pouvons pour cela nous pencher sur la phonétique articulatoire, domaine d'étude qui cherche à comprendre comment nous orchestrons les différents éléments de notre appareil phonatoire pour produire la parole. Le développement de nouvelles méthodes d'imagerie médicale dans les années 1950 a permis de grandes avancées en phonétique, nous permettant par exemple de caractériser chaque phonème par un lieu d'articulation, le point de constriction maximal du conduit vocal, ainsi qu'une manière d'articulation, la façon dont la constriction se forme et se relâche. Cependant, le processus nous permettant de produire une séquence de phonèmes dans son ensemble est encore mal compris. En effet, nous savons depuis les travaux de Scripture (1904) que les mouvements articulatoires cohabitants pendant la même période s'influencent les uns les autres par un mécanisme bien plus subtil qu'un simple phénomène de glissement entre une suite d'éléments statiques, hypothèse définitivement réfutée par l'étude acoustique de Joos (1948). Cette influence est connue sous le nom de coarticulation, du fait que les segments de la parole sont articulés conjointement et non pas grossièrement concaténés les uns aux autres (Menzerath and de Lacerda, 1933). Nous pouvons distinguer l'influence dite de la coarticulation rétentive (des phonèmes passés vers le phonème courant) et l'influence de la coarticulation anticipatrice (des phonèmes futurs vers le phonème courant). Par exemple, l'articulation labiale du /b/ et /t/ est différente dans les mots "butte" et "batte", /b/ subit une influence de la voyelle via les phénomènes d'anticipation, et /t/ subit une influence des phénomènes de rétention. La communauté scientifique tend aujourd'hui à attribuer les effets de la coarticulation anticipatrice à un complexe mécanisme de planification (Whalen, 1990), et les phénomènes rétentifs sont considérés comme essentiellement dus à la nature mécanique de l'appareil phonatoire humain (Daniloff and Hammarberg, 1973). Nous pouvons également noter que les effets de la coarticulation rétentive est bien moindre que ceux de la coarticulation anticipatrice (Gilbert, 1972), et que cette dernière a des effets attestés sur l'intelligibilité, par exemple dans le cas de la protrusion des lèvres (Cathiard et al., 1991). Pour synthétiser une parole multimodale, il est donc primordial de prendre en compte ces effets de coarticulation.

Dans notre travail de thèse, nous abordons le problème de la modélisation de la coarticulation en ayant comme objectif central la capacité à modéliser un nombre arbitraire d'articulateurs, et ce quelque soit la langue. D'une manière similaire à notre définition de l'articulation multimodale, nous proposons un modèle de coarticulation multimodale capable de prédire la dynamique du visage ou du conduit vocal en fonction de la modalité acoustique, réduite à sa représentation phonétique, d'une langue particulière. Motivés par notre étude bibliographique du chapitre 1, nous exploitons la puissance des méthodes

statistiques, et plus particulièrement de l'apprentissage profond (*deep learning*). Nous justifions au chapitre 2 notre choix d'une technique capable d'intégrer les notions de temporalité et de dynamique au coeur de sa modélisation, en l'occurrence les réseaux de neurones récurrents, et présentons également une procédure d'injection de connaissances articulatoires améliorant considérablement l'apprentissage du réseau. Cette méthode repose essentiellement sur une initialisation particulière des dernières couches de notre réseau de neurones. C'est une initialisation qui se base sur l'exploration préalable de nos corpus par des méthodes de réduction de la dimensionnalité, et nous permet de fournir au réseau de neurones une représentation efficace de l'espace articulatoire dès le début de la phase d'apprentissage. Les quatre corpus utilisés dans cette thèse, deux corpus pour la modalité visuelle (français et allemand) et deux corpus pour la modalité articulatoire (anglais et allemand), seront par ailleurs présentés au chapitre 3.

Le chapitre 4 est centré autour d'expériences permettant de valider l'adéquation de notre modèle au problème de modélisation de la coarticulation, et met plus particulièrement en lumière l'apport de notre procédure d'injection de connaissances articulatoires. Nous verrons en quoi notre modèle sans injection de connaissances permet la prédiction de trajectoires articulatoires de bonne qualité, mais souffre cependant de certaines difficultés lors de son apprentissage. Ces difficultés semblent complètement contournées lors de l'utilisation de notre procédure d'injection de connaissances, permettant d'améliorer les performances finales du modèle tout en réduisant considérablement le temps d'apprentissage. Nous expérimentons également avec les idées d'apprentissage par transfert, consistant en la réutilisation d'un modèle entrainé à la modélisation de la modalité visuelle vers la modélisation de la modalité articulatoire. Dans ce cas de figure, notre procédure semble permettre au modèle d'apprendre des caractéristiques plus génériques, plus facilement transférables à une autre modalité ou à un autre locuteur, en concentrant les informations propres au locuteur au niveau des dernières couches du réseau de neurones, qui ne seront pas réutilisées lors du transfert.

Nous évaluons plus finement les prédictions de notre modèle au chapitre 5. D'une part pour analyser en détail les erreurs de prédictions commises, que cela soit au niveau de l'erreur par région de l'espace articulatoire, ou au niveau de cibles articulatoires critiques de la modalité visuelle, comme la fermeture des lèvres pour la production d'une consonne bilabiale. Pour ces segments critiques, nous mettons en évidence certains manquements de notre modèle sans injection de connaissances, qui n'atteint pas une fermeture totale des lèvres lors des bilabiales. Ces manquements semblent être corrigés par l'apport d'une bonne représentation de l'espace articulatoire avant l'apprentissage. D'autre part, nous profitons de ces prédictions pour explorer ce que le réseau a modélisé de la coarticulation

anticipatrice, ce qui nous permet de mettre en évidence des parallèles intéressants avec les connaissances établies en phonétique articulatoire. Nous approfondissons ce parallèle en fin de chapitre par l'étude de la première couche récurrente du modèle, permettant d'obtenir une première idée de la représentation des phonèmes apprise par le modèle. En utilisant un outil de visualisation, nous mettons en avant de grandes similitudes entre cette représentation et l'organisation des phonèmes par lieu d'articulation. Nous terminons cette thèse par une application de notre modèle à une tête parlante virtuelle au chapitre 6, afin de former un système d'animation de la parole, permettant de synthétiser la parole visuelle depuis la modalité acoustique. Nous exploitons cette application pour conduire une évaluation subjective comparant les prédictions de notre modèle de coarticulation aux trajectoires articulatoires réelles. Cette expérience nous fournit des résultats encourageants quant à l'utilisation de notre modèle de coarticulation pour améliorer l'animation d'une tête parlante.

Chapitre 1

L'articulation multimodale : données, modélisation et prédiction

Nous pouvons représenter un modèle de production de la parole comme un processus à quatre niveaux : les fonctions linguistiques de haut niveau, la planification, le contrôle et la plateforme d'articulation. Les fonctions linguistiques dépassent amplement le cadre de cette thèse, et correspondent globalement à la prise en charge des contraintes sémantiques et syntaxiques de la parole nécessaire à faire passer son message. Nous pouvons également considérer certains aspects prosodiques comme faisant partie de cette étape, par exemple pour la préparation de phrase interrogative. Afin de concevoir notre système de prédiction de l'articulation, nous nous intéressons principalement aux niveaux suivants du processus : le planificateur, qui a pour charge d'élaborer le planning moteur nécessaire à la production de la parole en fonction des contraintes du processus linguistique, et le contrôleur, qui génère la séquence de commandes moteurs nécessaire au respect du planning moteur pour conduire la dynamique de la plateforme d'articulation. Cette dernière correspondant donc au modèle utilisé pour représenter l'appareil phonatoire. Il est bon de noter que très peu de consensus existent dans la littérature afin de définir la nature de ces différentes étapes, et ce même pour la terminologie.

Nous nous intéressons plus particulièrement dans ce chapitre aux modèles de contrôle moteur de la parole et les modèles de coarticulation. D'une part, le contrôle moteur de la parole correspond à la phase de contrôle du processus de production de la parole (section 1.2.1). Ces modèles cherchent à reproduire le mouvement de nos articulateurs à très bas niveau, avec une vision orientée mécanique et des approches inspirées par l'ingénierie. Ces derniers sont capables de générer des commandes moteurs depuis un planning, commandes permettant la manipulation d'un modèle de conduit vocal. D'autre part, la modélisation

de la coarticulation correspond à la phase de planification de la production de la parole (section 1.2.2). Elle fut historiquement étudiée et abordée avec une vision phonétique et/ou phonologique. Cette dernière cherche à expliciter le comportement des phénomènes de coarticulation, et à proposer un modèle capable de déterminer la trajectoire de certains articulateurs en fonction d'un contexte phonétique.

Ces deux types de modèles cherchent à modéliser comment fonctionnent certains aspects de la production de la parole, avec pour principal objectif une meilleure compréhension de ce mécanisme. Mais nous pouvons tenter de modéliser l'articulation avec pour principal objectif la reproduction fidèle de ces mouvements articulatoires par le numérique. Dans ce contexte, nous nous intéressons plus spécifiquement à deux tâches connexes : l'inversion acoustique, qui consiste à prédire la forme du conduit vocal en fonction du signal acoustique, et l'animation de la parole, dont l'objectif est l'animation automatique d'un visage en fonction d'un segment de parole. Nous proposons dans cette thèse une vision commune à ces deux problématiques sous le terme de prédiction de mouvement articulatoire, en argumentant que le coeur de ces deux tâches, indépendamment de la modalité d'entrée du système, est effectivement la prédiction de tels mouvements. Cette comparaison est de plus facilitée par le regroupement fait précédemment de l'articulation interne et externe au sein du terme d'articulation multimodale, permettant ainsi d'unifier leurs deux espaces de sortie. Nous avançons également que trois tendances bien distinctes peuvent se détacher pour l'inversion acoustique et l'animation de la parole. Premièrement, la prédiction par les connaissances, c'est-à-dire par un ensemble de règles, par l'utilisation d'un modèle numérique (par exemple l'utilisation d'un modèle de coarticulation pour piloter une tête virtuelle), ou par toutes méthodes se reposant essentiellement sur la connaissance de l'articulation humaine pour prédire les trajectoires des articulateurs (section 1.3.1). Deuxièmement, la prédiction par sélection de références, consistant en la sélection de données, qu'elles soient acquises ou simulées (e.g. à l'aide d'un modèle de conduit vocal), afin de les réutiliser lors d'une prédiction. Ce genre d'approche regroupe par exemple l'inversion acoustique par *codebook* ou la synthèse par concaténation (section 1.3.2). Dernièrement, nous présentons les approches par modélisation statistique, qui exploite d'importants volumes de données et des outils d'apprentissage machine tels que les modèles de Markov ou les réseaux de neurones (section 1.3.3).

Préalablement à ces deux sections, nous nous intéressons également principaux dispositifs permettant l'acquisition de ces données articulatoires multimodales (section 1.1.1), avec une attention particulière sur plusieurs points que nous jugeons critiques pour nos travaux. Premièrement, ces données multimodales doivent rendre compte de la dynamique et de la finesse de l'articulation, ce qui nécessite une résolution temporelle et spatiale de

qualité. Deuxièment, l'acquisition ne doit pas être invasive pour le locuteur, afin de limiter au maximum les désagréments, ce qui diminue sa fatigue durant l'acquisition, favorisant ainsi une articulation naturelle et augmentant sensiblement la quantité de données enregistrées par session d'acquisition. Troisièment, le matériel d'acquisition doit permettre l'enregistrement de données tridimensionnelles. Finalement, que cela soit pour le visuel ou pour l'articulation interne, nous nous intéressons aux principales méthodes de représentation et de paramétrisation de ces données (section 1.1.2) utilisées par les divers modèles de contrôle moteur de la parole et des systèmes de prédiction de mouvements articulatoires.

1.1 Données articulatoires multimodales

1.1.1 Matériels d'acquisition

De nombreuses méthodes d'acquisition de données articulatoires existent, ayant toutes avantages et défauts. Parmi elles, nous évoquerons les principales techniques utilisées pour l'acquisition de la dynamique du conduit vocal : l'échographie, la radiographie, l'imagerie par résonnance magnétique et l'articulographie électromagnétique pour l'acquisition de l'articulation interne du conduit vocal. Pour la modalité visuelle, nous nous intéresserons aux caméras RGB, RGB-D ainsi qu'aux systèmes de capture de mouvements. Dans tous ces cas de figure, nous considérons la modalité acoustique comme étant trivialement enregistrable à l'aide d'un microphone.

Acquisition de l'articulation interne

L'échographie par ultrason est une technique basée sur des ondes à très hautes fréquences. Lorsque celles-ci traversent les différents types de tissus, une partie de ces ondes sont partiellement réfléchies vers un dispositif de mesure, bien souvent situé proche de la source d'émission. En utilisant les lois de la physique sur le comportement des ondes, il est alors possible d'inférer la position des tissus traversés. Cette technique est facile d'accès et peu coûteuse, mais les images obtenues sont difficilement interprétables pour un utilisateur inexpérimenté. Il est par exemple possible de partiellement suivre les mouvements de la langue en apposant la sonde entre le menton et la proéminence laryngée.

Il y a quelques années, la filmographie par rayons X ou par microfaisceau de rayons X (Westbury et al., 1990) était utilisée pour obtenir des vidéos de la dynamique de l'entièreté du conduit vocal. Cette approche permettait l'acquisition de vidéo contenant le contour entier des différents articulateurs avec une fréquence allant jusqu'à 60Hz. Cependant, cette technique d'imagerie utilise un type de radiation ionisante dangereuse pour la santé

d'un sujet longuement exposé, soulevant d'importantes questions éthiques. L'imagerie par rayons X est donc strictement réglementée à l'heure actuelle, ne pouvant donc être utilisée dans notre contexte de recherche. De plus, les images ainsi obtenues sont très complexes à analyser, car l'ensemble des tissus traversés par les rayons seront présents sur le support d'impression.

Ces dernières années, l'imagerie par résonnance magnétique est devenue populaire pour l'étude de la production de la parole (Greenwood et al., 1992). En effet, cette dernière permet d'acquérir des images de grande précision de l'ensemble des articulateurs. Cependant un tel dispositif est très coûteux, et génère un bruit non négligeable venant parasiter les enregistrements acoustiques. De plus, l'acquisition d'image 3D ne peut se faire que si le patient reste immobile dans l'IRM, cette contrainte peut-être cependant levée si nous nous contentons d'une coupe sagittale du locuteur, cas où l'IRM est capable d'enregistrer des données temporelles.

L'idée principale de l'articulographie électromagnétique (Hixon, 1971) est de produire un champ magnétique dans lequel il est possible d'enregistrer la trajectoire spatiale de capteurs collés à l'intérieur comme à l'extérieur du conduit vocal. Directement câblés à l'articulographe, ces capteurs sont de très petites bobines qui vont créer de faible quantité de courant électrique lors de leurs déplacements dans le champ magnétique. En connaissance de ce courant et de la calibration du champ magnétique, il est possible de déduire la position des capteurs. Cette technique permet d'acquérir d'importants volumes de données avec une importante précision et une haute fréquence, jusqu'à 1250 échantillons par seconde et une utilisation typique à 250Hz, mais a quelques limites. Premièrement, il est impossible de récupérer les contours exacts des articulateurs, seuls quelques points d'intérêts peuvent donc être étudiés (24 sur les articulographes récents, avec une utilisation typique de 2 à 8 capteurs sur la langue). Deuxièment et certainement plus sujet à controverse, la présence de fils et de capteurs dans la bouche du sujet peut provoquer une gêne lors de l'élocution, et modifiant donc légèrement l'articulation. Il semble de plus très difficile de réussir à strictement disposer les capteurs dans la même position entre deux sessions d'enregistrement.

Acquisition de l'articulation externe

Pour l'acquisition de la modalité visuelle de la parole, une première solution évidente est l'utilisation d'une simple caméra permettant d'obtenir l'intégralité du signal visuel. Cependant, l'information tridimensionel est perdue lors de l'acquisition, et doit donc être estimée via des procédures complexes d'analyse de l'image. Ces dernières consistent en

la localisation de point de repère sur le visage, généralement au niveau des lèvres, des yeux, du nez et du menton. Bien que ces méthodes obtiennent aujourd'hui des résultats encourageants (Bulat and Tzimiropoulos, 2017), la quantité d'expertise nécessaire à la mise en place de ces méthodes peut être un frein non négligeable à la récolte des données.

Des alternatives matérielles existent cependant afin d'enregistrer également l'information de profondeur. Ces caméras de profondeur, parfois nommées caméras 3D, ou caméras RGB-D (Red Green Blue - Depth), permettent d'associer à chaque pixel trichromatique une information sur la distance de ce pixel par rapport à la caméra. Ceci peut être réalisé par la projection d'un motif infrarouge spécifique, dont les déformations nous renseignent sur le volume 3D filmé, ou par l'utilisation de lumières pulsées et de leurs réflexions. Malheureusement, une procédure de détection des points d'intérêts et/ou de segmentation de l'image doit toujours être effectuée afin de pouvoir étudier les articulateurs, et la précision des caméras n'est pas forcément adaptée à l'étude de la production de la parole, car ces dernières manquent grandement de précision au niveau de la profondeur (Bandini et al., 2015; Ouni and Dahmani, 2016).

Pour finir, les systèmes de capture de mouvement (MoCap pour *Motion Capture*) à base de marqueur exploitent le principe de la stéréovision. Des caméras suivent l'évolution de capteurs réfléchissants sur lesquels est projetée de la lumière infrarouge (bien souvent par les caméras elles-mêmes). Après calibration, il est possible de suivre avec grande précision spatiale (de l'ordre du dixième de millimètre) et temporelle (supérieur à 100Hz) l'évolution d'un marqueur dans l'espace tant que ce dernier est visible d'au moins deux caméras. De plus, cette technique permet de considérer un important nombre de capteurs, nous permettant d'approximer avec de nombreux points le contour réel des lèvres ou encore la forme des joues. Tout comme pour l'articulographie, le placement de ces capteurs doit cependant être minutieux, nécessite donc une longue période de préparation du sujet, et est difficilement reproductible à l'identique.

1.1.2 Représentation de l'espace articulatoire et visuel

Nous avons vu ci-dessus les principaux dispositifs d'acquisition de données d'articulation multimodale, permettant bien souvent d'obtenir des images (rayons X, IRM, caméras, échographie) ou des nuages de points (EMA, MoCap). Si le nuage de point est déjà une représentation simplifiée du visage ou du conduit vocal, il n'en est pas de même pour les images offrant l'intégralité du contour des articulateurs. Prédire les mouvements articulatoires avec le même niveau de détail que, par exemple, la coupe médio-sagittiale d'un IRM dynamique semble donc être un objectif très complexe. Pour pallier à cette problématique,

la communauté scientifique a proposé plusieurs représentations paramétriques du visage ou du conduit vocal permettant de simplifier la prédiction de leurs comportements tout en conservant leurs propriétés géométriques essentielles.

Modéliser le visage

Descendante directe du modèle proposé par Parke (1982), la norme MPEG-4 (Ostermann, 1998; Pakstas et al., 2002) défini un ensemble de points d'intérêts sur le visage (FDP : *Face Definition Parameters*) ainsi qu'un ensemble de 66 déplacements et rotations (FAP : *Face Animation Parameters*)) de ces points d'intérêts afin de contrôler l'animation du visage. Ces FAP furent le résultat d'une analyse de corpus audiovisuelle afin de modéliser les micro-expressions d'un visage.

Une autre paramétrisation des déformations du visage couramment usité est le FACS, *Facial Action Coding System* (Friesen and Ekman, 1978), un système d'inspiration anatomique qui définit un ensemble d'actions unitaires (*action units*, AU) représentant la contraction ou le relâchement d'un groupement musculaire. Une expression faciale peut donc être représentée comme une combinaison d'AU. Edwards et al. (2016) propose une surcouche aux FACs, JALI, introduisant un ensemble de visèmes (cf. paragraphe ci-dessous) et deux paramètres capables de quantifier l'impacte respectif de la mâchoire et des lèvres dans l'articulation. Il est alors possible pour les auteurs de représenter différents styles d'élocution par une simple manipulation de ces deux paramètres, JA et LI.

Bien que ne possédant pas de définition stricte dans la littérature, un visème peut-être envisagé comme le pendant visuel d'un phonème, ou comme un ensemble de phonèmes ayant la même apparence au niveau des lèvres (par exemple /p/ /b/ et /m/ ont tous le même visème d'une bouche complètement fermée). La relation phonème/visème n'est pas biunivoque, c'est-à-dire que plusieurs phonèmes possèdent la même résultante visuelle. Afin de représenter la dynamique d'un visage, il est possible de considérer une interpolation entre différents visèmes, méthode utilisée par de nombreux systèmes d'animation du visage. Cependant, un visème ne capture aucunement la dynamique intrinsèque à l'articulation, et l'évolution de la pondération associée à chaque visème se doit donc de tenir compte des phénomènes de coarticulation pour obtenir des animations de bonne qualité.

Pour pallier à ce défaut et être capable de considérer les phénomènes de coarticulation au niveau de ces "unités d'articulation", Taylor et al. (2012) introduisent l'utilisation de visème dynamique, afin de considérer le mouvement articulatoire dans son ensemble plutôt qu'une cible articulatoire précise (par exemple "ouverture de la bouche" plutôt

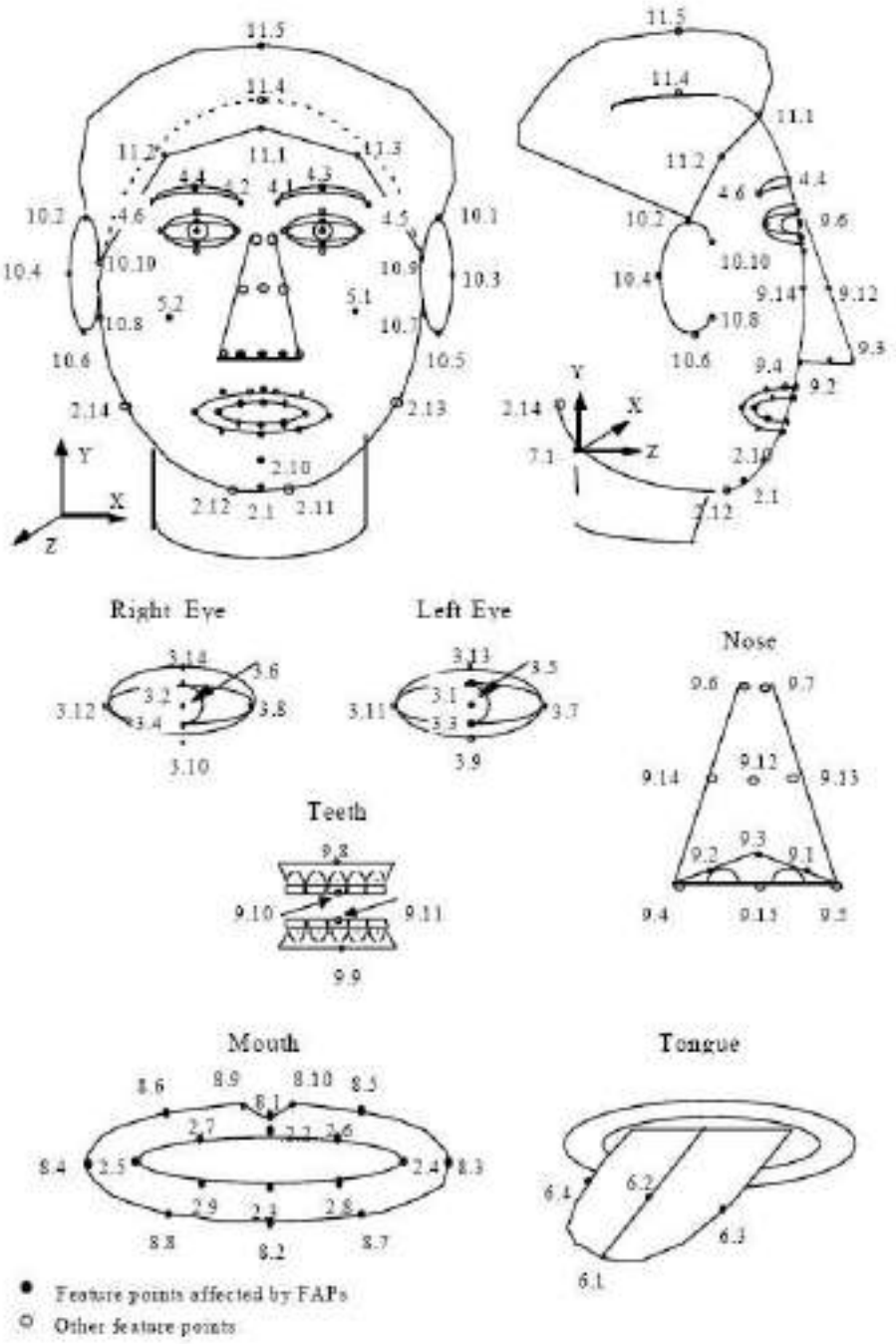


FIGURE 1.1 – Positionnement des paramètres de définition d'un visage (FDP) de la norme MPEG-4.

que "bouche ouverte"). Pour ce faire, les auteurs proposent une analyse des paramètres d'un modèle actif d'apparence (*Active Appearance Model*[1] (Cootes et al., 2001), AAM) d'un corpus audiovisuel acquis à l'aide d'une simple caméra, dont les points d'intérêts sont manuellement annotés, afin de segmenter chaque phrase en un ensemble de gestes articulatoires. Tous ces gestes seront par la suite analysés statistiquement afin de produire les caractéristiques utilisées par un algorithme de clustering pour regrouper les différents gestes articulatoires en visème dynamique.

Modéliser le conduit vocal

Deux grands types de modèles de conduit vocal peuvent être retrouvés dans la littérature, les modèles à fonction d'aire et les modèles articulatoires.

La fonction d'aire d'un conduit vocal se définit par l'aire de ce dernier sur le plan médio-sagittal, de la glotte aux lèvres. La forme géométrique du conduit n'est alors pas prise en compte, mais cette seule information est acoustiquement valide pour des fréquences inférieures à 4kHz, où le son se propage essentiellement le long du conduit vocal (Stevens and House, 1955). Stevens and House (1955) et Fant (1970) proposent ainsi des modèles à fonction d'aire à trois paramètres : la position de la constriction, l'aire à la constriction, et l'ouverture des lèvres. Le modèle de Fant (1970) est par exemple formé de quatre tubes : un tube pour les lèvres, une cavité avant, un tube pour la constriction de la langue, et une cavité arrière. La constriction de la langue est formalisée à l'aide d'une fonction parabolique ou hyperbolique afin de mieux représenter l'arrondie du corps de la langue. Schoentgen and Ciocca (1995) propose de remplacer ces tubes par des cônes afin d'obtenir des fonctions d'aire continue.

Les travaux sur les modèles articulatoires se sont essentiellement concentrés sur une représentation de la coupe médio-sagittale, car cette dernière est assez représentative du conduit vocal et permet d'obtenir relativement fidèlement l'acoustique à l'aide de modèle de passage de la coupe médio-sagittale à la fonction d'aire (Heinz, 1965). De plus, il s'agit d'une visualisation rendue possible par de nombreuses technologies (rayons X, IRM), ce qui permet de faire reposer les choix de conception du modèle sur des observations du conduit vocal. L'un des principaux modèles articulatoires, le modèle de Maeda (1990), repose justement sur une étude statistique de 1000 contours du conduit vocal obtenus

1. Les paramètres AAM sont une description statistique des déformations de la forme et texture d'un objet sur une image, qui a été utilisé avec succès pour représenter un visage (Edwards et al., 1998). Ce modèle repose sur l'utilisation de point d'intérêts sur l'objet à représenter, et d'une analyse par composantes principales des différentes positions de ces points d'intérêts ainsi que de la texture contenue par ces points.

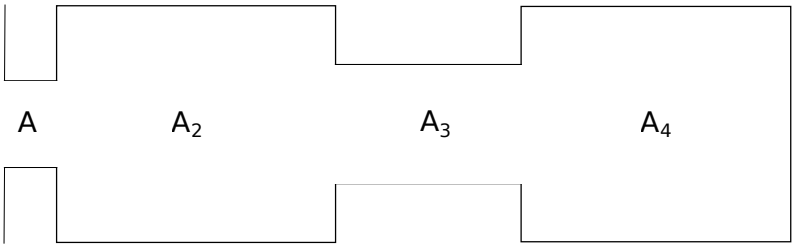


FIGURE 1.2 – Vue de coupe du modèle de Fant (1970), A_1 représente l'aire au niveau des lèvres, A_2 l'aire de la cavité avant, A_3 l'aire au niveau de la constriction de la langue, et A_4 la cavité arrière.

depuis des images cinéradiographiques. Pour cela, Maeda (1990) a étudié les intersections des contours de nombreux articulateurs (la langue, le larynx, les parois du pharynx, le pavillon labial, les incisives supérieures, le palais dur et le voile du palais) dans un système de coordonnées semi-polaires, et a conduit sur ces dernières une analyse factorielle à la recherche des composantes du modèle. Le modèle articulatoire en résultant est composé de sept paramètres : la position de la mâchoire, la forme et la position du corps de la langue, la position de l'apex de la langue, l'ouverture et la protrusion des lèvres et pour finir la hauteur du larynx.

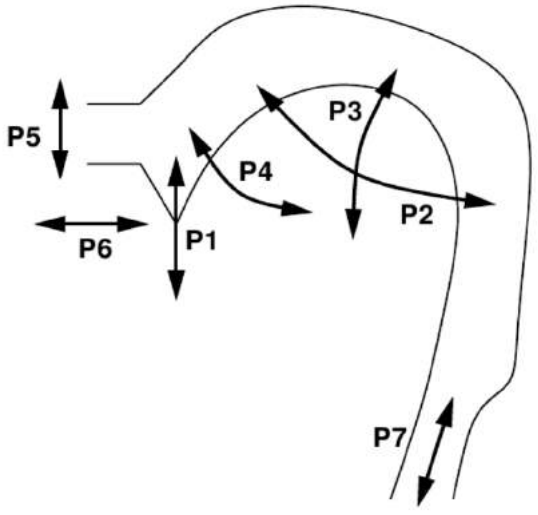


FIGURE 1.3 – Les sept paramètres du modèle de Maeda (1990).

1.2 Modèles de production de la parole

1.2.1 Contrôle moteur de la parole

Modèle DIVA

Le modèle DIVA, *Directions Into Velocities of Articulators*, fut développé pour répondre à plusieurs questions théoriques à propos du contrôle moteur de la parole, et se concentre principalement autour de la réplication de la production de la parole au niveau comportemental, neurologique et développemental (Tourville and Guenther, 2011; Guenther, 1994; Golfinopoulos et al., 2011; Guenther et al., 2006). Son unité de planning est le son de parole (*speech sound*), qui peut être un phonème, une syllabe ou encore un segment multisyllabique. Le planificateur est dans ce modèle une carte des sons de parole (*speech sound map*), qui associe pour chaque *speech sound* une trajectoire articulatoire, un signal acoustique et un signal somatosensoriel. La trajectoire articulatoire correspond à l'évolution temporelle des paramètres de la représentation utilisée pour l'espace articulatoire, le signal acoustique est représenté par les trois premiers formants, et le signal somatosensoriel englobe la position des articulateurs pour représenter la proprioception, ainsi que le degré de contacte entre les articulateurs pour représenter la sensation tactile.

DIVA est composé de trois sous-systèmes distincts. Le premier est un contrôle direct

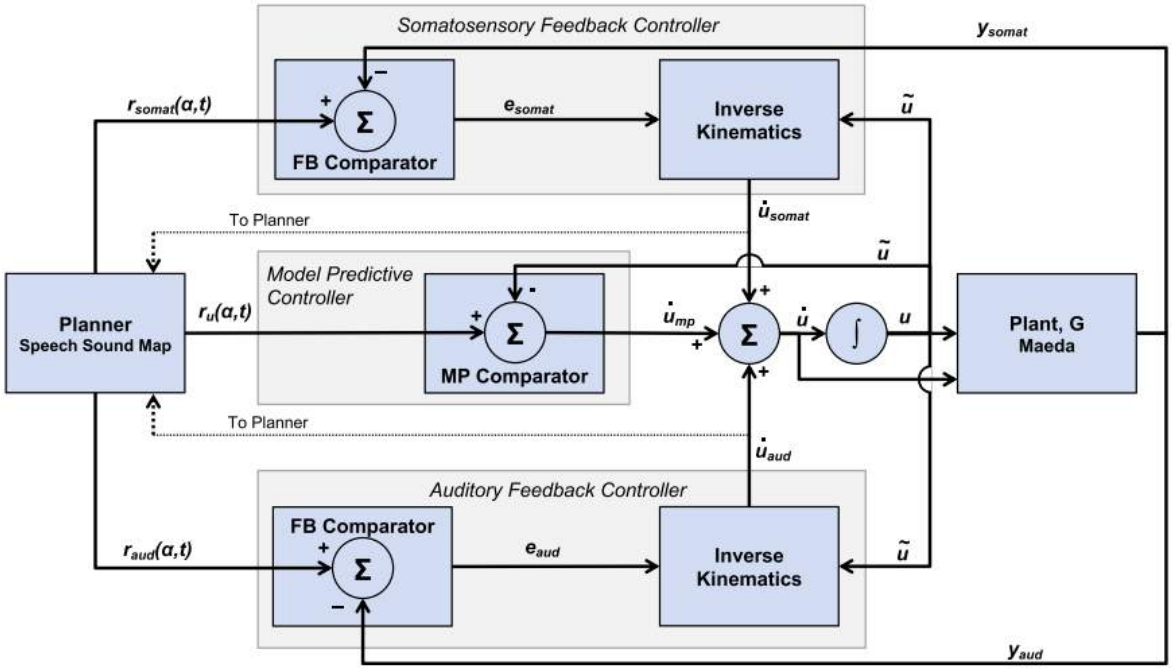


FIGURE 1.4 – Le modèle DIVA. Schéma extrait de (Parrell et al., 2019a)

de la position des articulateurs, qui compare la position des articulateurs avec la position cible donnée par le planificateur (*Model Predictive Controller*). Les deux autres sous-systèmes sont des boucles de rétroactions permettant de comparer le résultat acoustique et somatosensoriel de l'articulation avec les cibles présentes dans la carte des sons de parole, respectivement nommé le contrôleur de rétroaction auditif (*auditory feedback controller*) et le contrôleur de rétroaction somatosensoriel (*somatosensory feedback controller*). Ces trois erreurs sont individuellement transformées en commandes moteurs, qui seront finalement combinées en une unique commande moteur passée à la plateforme d'articulation. Cette dernière fut historiquement le modèle de Maeda, mais plusieurs autres représentations du conduit vocal furent utilisées le long du développement de DIVA.

Modèle TD

Formulé par Saltzman (1986, 1991), le modèle *Task Dynamics* repose sur l'hypothèse centrale que les mouvements articulatoires sont conduits par l'évolution d'un système dynamique dont les paramètres invariants sont déterminés par le contenu linguistique d'une phrase. Lors de sa création, son principal objectif fut la modélisation du passage d'une séquence de cibles linguistiques discrètes et invariantes à une suite continue de mouvements articulatoires dépendant de leurs contextes. Développé au sein du laboratoire de Haskins, la plateforme d'articulatoire utilisée par le modèle TD est principalement le synthétiseur articulatoire CASY (Rubin et al., 1996).

Dans le modèle TD, le planificateur fournit au contrôleur des tâches articulatoires de

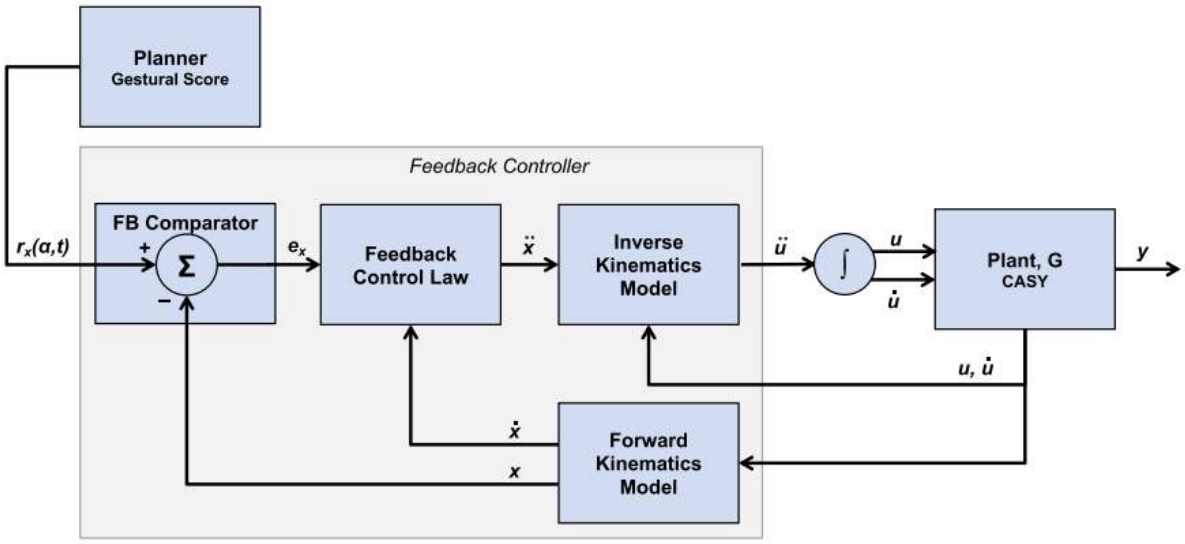


FIGURE 1.5 – Le modèle TD. Schéma extrait de (Parrell et al., 2019a)

haut niveau comme la fermeture des lèvres ou la constriction d'une zone du conduit vocal, dénommé *gestures*. En fonction de la tâche active et du temps, le planificateur fournit un score représentant l'état du geste articulatoire (*gestural score*). Cette définition permet au modèle TD d'être facilement mis en relation avec la phonologie articulatoire de Browman and Goldstein (1992). Le système de contrôle repose entièrement sur une comparaison dans l'espace des tâches de la cible imposée par le planificateur et de l'état de la plateforme d'articulation. L'erreur résultante permet de déterminer l'accélération nécessaire dans l'espace des tâches, elle-même utilisée pour calculer l'accélération correspondante dans l'espace des mouvements articulatoires.

Autres modèles

De nombreux autres modèles de production de la parole existent, parmi lesquelles nous pouvons retrouver SFC, FACTS ou encore GEPETTO.

SFC (Houde and Chang, 2015; Houde and Nagarajan, 2011; Houde et al., 2014) fut conçu avec pour objectif d'intégrer des progrès sur les architectures de *feedback* dans le contrôle moteur au problème spécifique de la production de la parole. Son avantage principal est de pouvoir prendre en compte la nature bruitée et retardée du *feedback* articulatoire lors de sa prédiction de l'état des articulateurs. Récemment, le modèle FACTs (Parrell et al., 2018, 2019b) fut proposé pour tenter de combiner les avantages du modèle TD et du modèle SFC, en particulier le concept de prédiction d'état par *feedback* de SFC ainsi que le modèle de planification et de contrôle du conduit vocale de TD.

Le modèle GEPPETO, proposé par Perrier et al. (1996); Patri (2018); Perrier et al. (2006) est un modèle de contrôleur basé sur l'hypothèse du point d'équilibre (Feldman et al., 1990), principalement utilisé dans le cadre du contrôle d'une langue biomécanique. GEPPETO fut formulé afin d'explorer trois hypothèses : (1) l'espace des tâches de la production de la parole est discrète et phonétique, (2) le rôle de la biomécanique n'est pas trivial dans le contrôle moteur de la parole, et (3) le contrôleur moteur de la parole utilise des principes de contrôle optimal. La sortie du planificateur est une séquence temporelle comprenant le phonème, sa durée, ainsi que l'effort nécessaire à son articulation catégorisé en trois niveaux (faible, moyen et fort), où l'effort est basé sur la force musculaire nécessaire à la production de la cible articulatoire. Cette séquence temporelle est transformée en une série de valeur λ pour les 6 muscles de la langue biomécanique, principalement via une procédure d'optimisation minimisant la modification de la longueur des muscles, dans un principe de moindre effort.

1.2.2 Modélisation de la coarticulation

Modèles *look-ahead*

Soutenu par des études expérimentales sur la coarticulation labiale (Daniloff and Moll, 1968) et vélaire (Moll and Daniloff, 1971), investiguant l'influence de la coarticulation anticipatrice, la théorie de la propagation des caractéristiques (*feature spreading*) proposé par Daniloff and Hammarberg (1973) puis Hammarberg (1976) est une rupture nette avec l'idée que la coarticulation a pour origine les mécanismes physiologiques universels de production de la parole. D'après Hammarberg (1976), une vision purement physiologique de la coarticulation crée une forte division entre l'intention et l'exécution, impliquant *de facto* que les processus mentaux ne prennent pas en compte les capacités de notre appareil phonatoire, ou que l'appareil phonatoire est incapable de répondre aux commandes spécifiées par le processus d'encodage phonétique. D'après ces auteurs, cette dichotomie contre-intuitive entre processus mentaux et capacités physiques des organes phonatoires peut être surmontée si nous considérons la coarticulation au niveau phonologique, de telle façon à ce que les articulateurs n'aient qu'à exécuter des instructions de haut niveau définies par des règles phonologiques.

Le modèle de coarticulation en résultant, connu sous le nom de modèle *look-ahead*, a été repris par Daniloff and Hammarberg (1973) depuis les travaux de Henke (1966). Dans cet algorithme, les segments de la parole sont associés à des caratéristiques phonologiques qui peuvent être présentes, absentes ou indéfinies. Pour un segment donné, toutes caractéristiques indéfinies au niveau phonologique le sera au niveau phonétique par un simple mécanisme de propagation *de droite à gauche*. Un autre modèle populaire de type *look-ahead* est le modèle de Öhman (1967), issue d'une étude des CVC en suédois, anglais et russe (Öhman, 1966). Dans ce modèle, la coarticulation existe de voyelle à voyelle, gestes par-dessus lesquels sont ajoutées les caractéristiques phonétiques nécessaires à la production de consonnes.

Une critique usuelle des modèles *look-ahead* est leurs incapacités à prendre en compte la nature graduelle de la coarticulation et son évolution temporelle. En témoignent les études par fibroscopie du vélum par Ushijima (1972) (langue japonaise) et Benguerel et al. (1977) (langue française) qui mettent en évidence que l'abaissement du vélum en anticipation d'un phonème nasal a une durée équivalente dans des séquences de 2 ou 3 segments précédents, remettant ainsi en cause la temporalité de la propagation des caractéristiques hypothétisée dans les modèles de type *look-ahead*. En plus de la façon de se propager au cours du temps, la nature binaire des caractéristiques phonétiques est fortement controversée par de nombreuses études, que cela soit par le fait que pour deux

caractéristiques contradictoires, la coarticulation débute pendant le premier segment et non pas après (Benguérel and Cowan, 1974; Sussman and Westbury, 1981), ou par le fait que dans une séquence V_1CV_2, V_1 et V_2 s'influence l'une l'autre même en cas de caractéristiques contradictoires (Butcher and Weiher, 1976; Farnetani et al., 1985; Magen, 1989). Pour surmonter une partie de ces limitations, Bladon and Al-Bamerni (1976) propose d'ajouter aux caractéristiques phonologiques binaires un coefficient de résistance à la coarticulation dans son modèle de résistance à la coarticulation (*coarticulation resistance model*) permettant de moduler la propagation de caractéristiques dans le temps.

Modèles *time-locked*

Fowler (1977) prend position contre les modèles supposant que les caractéristiques phonologiques sont utilisées comme paramètres d'entrée de notre système d'articulation, et contre une vision entièrement phonologique de la coarticulation. Le travail collaboratif entre psychologues et linguistes résultant de ces idées (Fowler, 1977; Bell-Berti and Harris, 1981) aboutit à la théorie de la coproduction et au domaine de la phonologie articulatoire. Dans la théorie de la coproduction, les gestes et cibles articulatoires ne sont pas modifiés au niveau phonologique : c'est leur nature temporelle et dynamique qui leur permet de se superposer, s'influençant les unes les autres parce que coproduite au sein du contexte phonétique. Cette hypothèse que les gestes articulatoires peuvent se *mélanger*, fut vérifiée par plusieurs études, que cela soit au niveau du larynx (Munhall and Löfqvist, 1992), du vélum (Bell-Berti and Krakow, 1991) ou encore sur les mouvements de la pointe de la langue (Farnetani, 1990; Farnetani and Busa, 1994).

L'influence temporelle de chaque geste articulatoire mis en lumière par cette théorie est à l'origine de la famille de modèle *time-locked*, parmi lesquelles nous retrouvons le modèle *task-dynamic* par Saltzman (1986, 1991), basé sur les *phonetic gestures*, une nouvelle unité dynamique de la phonologie articulatoire proposé par Ohala et al. (1986); Browman and Goldstein (1989, 1992). Nous retrouvons également dans cette catégorie le modèle à fonction de dominance de Löfqvist (1990), ayant lui-même grandement inspiré le modèle de Cohen and Massaro (1993). Dans ces deux derniers modèles, chaque phonème est associé à une cible articulatoire située en son milieu et à une fonction de dominance définissant l'influence du phonème au cours du temps. Pour connaitre la position d'un articulateur, il est alors possible de moyenner la position de toutes les cibles articulatoires du segment en fonction de leurs dominances à l'instant t. Une fonction de dominance typiquement usitée est celle proposée par Cohen and Massaro (1993), composée de deux exponentielles, l'une croissante jusqu'à la cible articulatoire, et l'autre décroissante à partir de cette cible (voir

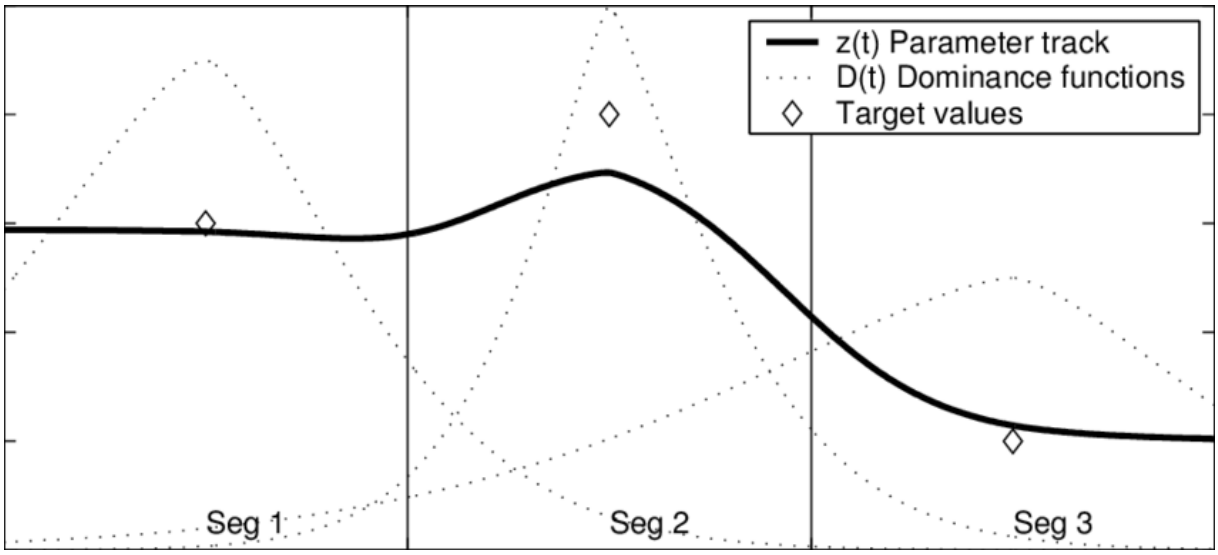


FIGURE 1.6 – Exemple de génération d'un paramètre articulatoire avec le modèle de Cohen and Massaro (1993), extrait de Beskow (2004). Les losanges représentent les valeurs cibles associées à chaque segment de parole, et les courbes en pointillés la fonction de dominance du paramètre pour ce segment. La trajectoire en noir est obtenue par une moyenne des cibles pondérées par leurs fonctions de dominance.

1.6). Ce type de modèle a cependant tendance à ne pas atteindre complètement certaines cibles articulatoires très résistantes à la coarticulation, comme la fermeture des lèvres lors de la production de bilabiale.

Autres modèles

Parmi les autres modèles de coarticulation difficilement classifiable entre les deux grandes familles *look-ahead* et *time-locked*, nous retrouvons le modèle hybride, expansioniste et "à fenêtre".

Le modèle hybride résulte d'une étude sur l'arrondissement des lèvres (Perkel, 1986), corroboré par une étude sur sur la coarticulation vélaire de Al-Bamerni and Bladon (1982), tend à démonter que les gestes articulatoires sont parfois scindés en deux étapes. Sur une séquence $/iC_nu/$, la protrusion est temporairement liée à l'influence de $/i/$, comme prédit par les modèles *look-ahead*, puis lié à $/u/$, comme prédit par le modèle *time-locked*. Cette observation aboutit au développement d'un modèle caractérisé par une prédiction des mouvements des articulateurs en deux phases, la première étant basée sur un algorithme proche du *look-ahead*, et la seconde sur une approche *time-locked*. Cette étude a cependant été faussée par l'influence des consonnes sur la protrusion, et a par conséquent trouvé de

nombreux détracteurs (Gelfer et al., 1989; Perkell and Matthies, 1992; Bell-Berti, 1980).

Une étude sur la coarticulation labiale française par Abry and Lallouache (1995) montre que la forme des lèvres est très stable lors de l'articulation de séquences $/iC_5y/$, mais devient très variable lors de l'accélération de l'articulation des mêmes séquences au sein d'une phrase. Dans ce second cas, les données semblent suivre des mouvements parfois en deux phases, et parfois en une unique phase. Ceci incite les auteurs à développer un nouveau modèle de coarticulation, le modèle d'expansion du mouvement, qui hypothétise que la durée du mouvement est fortement expansible et relativement peu compressible, d'où la tendance à anticiper quand le temps le permet. Cette propriété pourrait expliquer pourquoi le modèle *look-ahead* est observable jusqu'à ce que la suite de consonnes intervocaliques soit trop longue, c'est à dire jusqu'au point d'expansion maximal. Il est intéressant de noter que ce modèle est paramétrable, dans lequel l'expansion des mouvements est considérée comme une caractéristique propre au locuteur, alors que les contraintes de compression semblent être le même pour tous les sujets de l'étude.

Pour finir, le modèle proposé par Keating et al. (1988), *window model*, définit pour chaque phonème une fenêtre de variation de la position des articulateurs, ce qui permet donc une prise en compte mutuelle de la phonétique et de la phonologie. Cette fenêtre, calculée à partir de mesures effectuées en différents contextes, représente donc la "résistance" d'un phonème à la coarticulation : un phonème avec une large fenêtre sera très sensible aux phénomènes de coarticulation. La trajectoire de l'articulateur est ensuite définie comme le "chemin" passant par toutes ces fenêtres tout en minimisant l'effort articulatoire.

1.3 Prédiction des mouvements articulatoires

1.3.1 Prédire par les connaissances

Nous classifions ici les systèmes de prédictions de mouvements articulatoires basés sur un ensemble de règles ou sur l'utilisation d'un des modèles de coarticulation présentés à la section précédente. De notre point de vue, ces approches résultent toutes d'un haut niveau d'expertise dans la parole multimodale associée à des manipulations techniques du support visuel, et ont pour principale caractéristique de reconstruire les trajectoires articulatoires en se basant sur des connaissances expertes à propos de l'articulation et sur la capacité à modéliser son comportement.

Ces méthodes sont majoritairement retrouvées lors de la synthèse audiovisuelle en deux étapes, où le signal visuel est généré indépendamment du signal acoustique. La première

étape synthétise le signal acoustique, et la deuxième étape exploite des informations acoustiques et/ou phonétiques fournies par cette première synthèse acoustique de la parole pour guider les déformations d'une tête parlante virtuelle. Pour la prédiction de la modalité articulatoire, peu d'études peuvent être considérées comme étant basées sur les connaissances. Nous pouvons néanmoins retrouver des méthodes d'inversion dites "directe", cherchant à modéliser mathématiquement la relation entre un vecteur de paramètres acoustiques et un vecteur de paramètres articulatoires (Mermelstein, 1967; Schoentgen and Ciocea, 1997).

Pour les approches photographiques, une contribution notable est le système acteur (*Actors System*) de Scott et al. (1994) qui utilise des techniques de *morphing* pour générer une vidéo depuis une série d'images clefs. Ces points servant à l'interpolation sont sélectionnés par une simple table d'association entre phonème et image. Cependant, la table d'association ainsi que les points d'intérêts utilisés par l'interpolation sont tous deux manuellement défini, ce qui force Mattheyses and Verhelst (2015) à classer le système acteur comme étant basé sur un ensemble de règles, et nous incite donc à le qualifier de prédiction par les connaissances. Cette approche à base d'interpolation entre pseudo-visèmes fut utilisée dans de nombreuses études (Goyal et al., 2000; Ezzat and Poggio, 2000; Tiddeman and Perrett, 2002; Verma et al., 2003).

Dans le cas de la synthèse audiovisuelle en 3D, les modèles de coarticulation de type *look-ahead* et *time-locked* sont les plus fréquemment rencontrés dans la littérature. Par exemple, Pelachaud et al. (1991) exploite un modèle de type *look-ahead* pour contrôler les *actions units* d'une tête parlante reposant sur FACS, ou encore Elisei et al. (2001) qui utilise le modèle d'Öhman pour contrôler 4 paramètres articulatoires : la hauteur et largeur de la bouche, la protrusion des lèvres et l'ouverture de la mâchoire d'une tête parlante reposant sur la norme MPEG-4. Pour les algorithmes *time-locked*, Le Goff (1997) utilise le modèle de Cohen-Massaro légèrement modifié afin de définir les cibles articulatoires de chaque phonème depuis un corpus audiovisuel, et Cosi et al. (2002) propose une amélioration de ce modèle pour assurer l'atteinte de cible articulatoire critique (par exemple la fermeture des lèvres pour une bilabiale). Pour ce faire, l'auteur ajoute une fonction de résistance à la coarticulation pour chaque phonème, capable d'inhiber l'influence des fonctions de dominance du contexte phonétique.

Une intéressante étude de (Beskow, 2004) propose par ailleurs une comparaison du modèle d'Öhman (1967), de Cohen and Massaro (1993), de quelques méthodes statistiques (réseaux de neurones *feed-forward* et récurrents) ainsi que de son approche de type *look-ahead* à base de règles Beskow (1995). Ses résultats mettent en lumière quelques résultats critiques : (1) les réseaux de neurones récurrents obtiennent de meilleurs résultats que les

réseaux *feed-forward*, avec en particulier une trajectoire bien plus "lisse", (2) les meilleurs résultats objectifs sont atteints par le modèle de Cohen-Massaro, et (3) lors de l'application à sa tête parlante, son système à base de règles restaure une plus grande intelligibilité que les deux autres modèles de coarticulation.

1.3.2 Prédire par sélection de références

Initié par les travaux de Atal et al. (1978), l'inversion acoustique par *codebook* peut à première vue ressembler aux approches dérivées du *Actors System*. Ici aussi, ces méthodes se basent sur la construction d'une table de correspondance entre vecteurs acoustiques et vecteurs articulatoires. Cependant, les entrées de cette table ne sont pas ici conçues par l'humain, mais résultent d'un échantillonnage de l'espace articulatoire, espace obtenu par acquisition d'une faible quantité de données ou par l'utilisation d'un modèle de conduit vocal. Cet échantillonnage peut-être régulier (Atal et al., 1978), aléatoire (Schroeter and Sondhi, 1992; Boë et al., 1992), une interpolation depuis des vecteurs racines (Larar et al., 1988; Sorokin and Trushkin, 1996), ou encore adaptatif (Charpentier, 1984; Sorokin and Trushkin, 1996; Ouni and Laprie, 2005a; Potard and Laprie, 2007). De plus, là ou les systèmes "Acteurs" proposent une fonction d'interpolation pour effectuer la transition d'une référence à l'autre, en étant incapable de sortir des limites imposées par ces références, l'inversion acoustique par *codebook* propose elle de modéliser la relation acoustique/articulatoire aux voisinages d'une entrée de la table, afin de pouvoir modifier le vecteur articulatoire correspondant à la plus proche entrée de la table en fonction de la différence entre l'entrée de la table et le vecteur acoustique. Cette relation est définie bien différemment en fonction des auteurs : constante dans un petit voisinage (Atal et al., 1978; Larar et al., 1988; Schroeter and Sondhi, 1992), linéaire (Atal et al., 1978; Charpentier, 1984; Sorokin and Trushkin, 1996; Ouni and Laprie, 2005a), polynomiale (Potard and Laprie, 2007) ou encore stochastique (Laboissière, 1992; Hogden et al., 1996).

Nous pouvons noter qu'un grand nombre de ces méthodes sont des inversions dites *point à point*, cherchant à retrouver l'état du conduit vocal et/ou de certains articulateurs pour un vecteur acoustique donné, et non pas pour un segment de parole. L'utilisation naïve de telle méthode pour une inversion de segment entraîne donc une totale omission de la contrainte temporelle inhérente à la production de la parole, résultant dans des trajectoires impossibles à réaliser par l'être humain. Retrouver la trajectoire des articulateurs à partir des modèles à base de *codebook* se réalise donc principalement en deux étapes, une première recherche de trajectoire articulatoire initiale depuis le codebook, souvent à l'aide de la programmation dynamique (Ouni and Laprie, 2005b; Potard and Laprie,

2009), puis par une procédure de lissage afin de s'assurer d'un mouvement plausible des articulateurs en retirant les aspérités des trajectoires.

L'approche de synthèse par concaténation fut largement exploitée pour la synthèse 3D de la parole visuelle, suite à des travaux exploratoires de Hallgren and Lyberg (1998); Kuratate et al. (1998) sur la concaténation de polyphones, des configurations de *mesh* 3D correspondant aux phonèmes, équivalent des visèmes. Très rapidement, ces systèmes de concaténation se sont mis à prendre en compte le contexte lors de la sélection d'unité, afin d'englober et de considérer un maximum des effets de la coarticulation. Nous retrouvons par exemple une prise en compte du contexte phonétique (Minnis and Breen, 2000; Edge and Hilton, 2006), mais aussi du contexte visuel (Breen et al., 1996; Engwall, 2002). En particulier, Cao et al. (2004) prend en compte le contexte phonétique tout en minimisant le nombre de concaténations nécessaire à la synthèse. Dans son étude comparative, Bailly et al. (2002) montre que son système à base de concaténation aboutit à de meilleurs résultats qu'un ensemble de règles induit depuis le même corpus audiovisuel. Nous pouvons également citer des approches par concaténation étant de véritable synthèse audiovisuelle de la parole depuis le texte, à l'image des travaux de Ouni et al. (2013) qui réalisent leur synthèse concaténant des unités bimodales contenant le visuel et l'acoustique.

Du côté de la synthèse visuelle de la parole en 2D, un travail pionnier est celui de Bregler et al. (1997), qui propose le *Video Rewrite*, un système extractant les formes de bouches depuis une base de données et en les réorganisant par visèmes associés à chaque triphone. Ces visèmes pourront par la suite être selectionnés en fonction du flux audio, et ajoutés à une image de fond représentant le reste du visage. Bien entendu, une importante étape de traitement d'image est nécessaire afin que le rendu final soit perçu comme un tout, et non pas comme une simple superposition d'images.

Ces grands principes seront repris par de nombreuses études (Cosatto and Graf, 2000; Fagel, 2006; Thies et al., 2016), dont les plus récentes et performantes utilisent des modèles statistiques pour guider la sélection (Fan et al., 2016; Suwajanakorn et al., 2017), ce qui rapproche grandement ces modèles de ceux présentés à la section suivante. Par exemple, Suwajanakorn et al. (2017) utilise des réseaux LSTM pour prédire les coefficients ACP d'une représentation éparse du contour des lèvres servant de critère de sélection de l'unité à concaténer. Depuis le corpus vidéo, les auteurs ont extrait 18 points 3D formant le contour des lèvres, et on conduit une analyse en composante principale afin de réduire ces vecteurs de dimension 36. Nous pouvons également noter qu'un décalage temporel d est effectué pour fournir au réseau une quantité fixe d'information future par rapport à l'instant t (i.e. à l'instant $t + d$, le réseau doit prédire la représentation de la bouche à l'instant t), ceci dans le but de prendre en compte les effets de la coarticulation anticipative. L'étude de

Fan et al. (2015, 2016) quant à elle utilise des LSTM bidirectionnels afin de prédire les coefficients AAM correspondants à la partie basse du visage. Les auteurs ont ici profité des informations phonétiques (triphones) et acoustiques (MFCC, *Mel-Frequency Cepstral Coefficients*).

1.3.3 Prédire par modélisation statistique

Pour la prédiction de la modalité articulatoire, l'arrivée des technologies d'articulographie électromagnétique a permis l'acquisition d'important volume de données relativement facilement, en particulier en comparaison avec les technologies d'acquisitions existantes (cf. 1.1.1). Ces nouveaux corpus ont ainsi ouvert la voie à l'utilisation d'outils statistiques d'apprentissage automatique comme des réseaux de neurones, modèles de Markov cachés (HMM pour *Hidden Markov Model*), ou encore des mélanges de gaussiennes (GMM pour *Gaussian Mixture Model*) pour l'inversion acoustique. L'approche GMM (Toda et al., 2004, 2008) à l'avantage de ne pas avoir besoin d'informations autre que les données acoustiques et articulatoires, mais ne considère pas l'aspect temporel dans son ensemble, s'arrêtant à une modélisation *point à point* lors de l'apprentissage. A contrario, les HMM sont eux capables de tenir compte de la temporalité des enregistrements et sont donc un outil de choix pour résoudre ce problème d'inversion, comme en témoignent les nombreuses études ayant utilisé ces modèles (Hiroya and Honda, 2004; Zhang and Renals, 2008; Youssef et al., 2009; Hueber et al., 2012; Xie et al., 2015). En fonction des auteurs, la relation permettant de passer des paramètres articulatoires aux paramètres acoustiques en fonction de l'état du modèle peut grandement varier, et peuvent éventuellement nécessiter les séquences de phonèmes associés au signal acoustique lors de l'apprentissage, entrainant donc une nouvelle et coûteuse tâche de traitement des données, mais permettant l'exploitation d'informations phonétiques.

Débuté par les travaux préliminaires de Soquet et al. (1991) et Papcun et al. (1992) pour l'inversion point à point depuis des données acquises à l'aide de micro faisceaux de rayons X, les réseaux de neurones connaissent un vif succès depuis les années 2000 (Richmond, 2006, 2009; Uria et al., 2012; Wu et al., 2015; Xie et al., 2016). Il est a noté que les architectures neuronales dites *feed-forward* ne sont pas capable de traiter des séquences, une solution habituelle afin que ces derniers prennent partiellement en compte la temporalité est de leur fournir une fenêtre temporelle de vecteurs acoustiques à partir desquels le réseau doit inférer un vecteur articulatoire. Cette méthode à cependant certaines limites : l'ajout d'un hyperparamètre qu'est la taille de la fenêtre glissante, la supposition que les informations présentes en dehors de la fenêtre ne sont pas importantes

pour la prédiction et la nécessité de recourir à une étape de lissage des trajectoires, car ces dernières peuvent ici encore contenir d'importantes aspérités. Actuellement, les meilleurs résultats sur cette tâche sont obtenus à l'aide de réseaux de neurones récurrents (RNN) (Liu et al., 2015; Zhu et al., 2015), une architecture neuronales capable de surmonter ces limites en traitant une séquence dans son ensemble.

Il est également intéressant de noter que comme pour les modèles HMM, de nombreux auteurs ont exploité avec succès l'utilisation de l'information phonétique avec les réseaux de neurones. Nous retrouvons par exemple l'utilisation de la bimodalité acoustique/phonétique (Shahrebabaki et al., 2019) directement en entrée d'un réseau afin de prédire la trajectoire de données EMA, mais aussi des utilisations plus complexes proposées par Xie et al. (2016), comme la mise en place de *multi-task learning* pour apprendre à prédire le domaine articulatoire et phonétique depuis la modalité acoustique, ou l'utilisation des *features* apprises lors de la modélisation de la relation acoustique/phonétique pour aider à la modélisation de la relation acoustique/articulatoire.

L'utilisation de la modalité phonétique est elle aussi exploitée par Taylor et al. (2017) pour la synthèse visuelle de la parole. Dans cette étude, l'auteur entraîne un réseau de neurones profond à prédire les coefficients AAM correspondants à la partie basse du visage depuis une séquence de phonèmes. Ces coefficients sont par la suite lissés pour éliminer les aspérités inhérentes à l'utilisation d'un modèle *feed-forward* et d'une fenêtre temporelle, puis utilisés par une procédure de *retargeting* (Curio et al., 2006; Pighin and Lewis, 2006; Song et al., 2011; Zell et al., 2017; Ribera et al., 2017) afin de piloter les déformations d'un autre visage virtuel arbitraire. Cette approche à donc le grand avantage de pouvoir utiliser des données audiovisuelles faciles et peu couteuses à acquérir, les vidéographies, pour entraîner un modèle capable de piloter l'animation d'un modèle 3D.

Vers des approches *end-to-end*

Nous retrouvons également dans la littérature quelques études récentes sur la création de modèles *bout-en-bout*. Par exemple Karras et al. (2017), qui propose une architecture neuronale *feed-forward* prenant en entrée une fenêtre temporelle du signal acoustique brut. La réussite de système de base sur deux points principaux : premièrement l'utilisation de couches convolutionelles distinctes pour l'analyse du signal et pour la compression temporelle de la fenêtre, et deuxièmement la création d'une fonction de coût complexe, renforçant l'aspect temporel partiellement masqué par la fenêtre glissante par une pénalisation de trajectoires erratiques. Nous pouvons aussi citer Pham et al. (2018a) dont la méthode se repose également sur des couches convolutionelles analysant le signal acous-

tique sans aucune paramétrisation, mais aussi sur des réseaux récurrents afin de prendre en compte l'aspect temporel sans fonction de coût particulière. Dans cette dernière étude, le réseau est chargé de prédire les poids associés à chaque *blendshapes* et est donc intrinsèquement lié à la paramétrisation choisie pour un visage, ce qui n'en fait pas totalement une approche *end-to-end*.

Cette dernière remarque s'applique également à Zhou et al. (2018b), qui propose un système d'animation automatique depuis le signal audio à base de réseaux LSTM basé sur le modèle JALI. Dans cette étude, les auteurs proposent d'entraîner un réseau à reconnaitre les phonèmes depuis l'audio (une tâche de modélisation acoustique) ainsi qu'à prédire la position de certains points clefs du visage extrait depuis des données audiovisuelles (une tâche donc très semblable à l'inversion acoustique), ce réseau pré-entrainé est ensuite utilisé par d'autres réseaux LSTM pour en déduire l'ensemble des paramètres JALI. Les auteurs constatent une amélioration de la prédiction lors de l'utilisation de différentes modalités (acoustiques, phonétiques et visuelles) et de l'apprentissage par transfert (pré-entraînement du modèle sur des tâches connexes).

Cette tendance se dégage particulièrement pour les approches photographiques de la synthèse audiovisuelle, grâce aux récents progrès des modèles génératifs de réseaux de neurones, et plus particulièrement des réseaux *adversarials* proposés par Goodfellow et al. (2014) (Pham et al., 2018b; Zhou et al., 2018a; Song et al., 2018; Vougioukas et al., 2018; Zakharov et al., 2019; Sadoughi and Busso, 2019). Ces modèles sont globalement composés de deux réseaux, un réseau générant un échantillon, l'autre discriminant les échantillons générés de ceux issus du corpus, et sont entraînés dans une philosophie min-max, où le générateur doit maximiser l'erreur du discriminateur. Cette méthode d'entraînement couplée à la capacité des réseaux convolutionels à travail au niveau du pixel, permet l'élaboration de méthodes ne nécessitant en théorie pas phase de traitement de l'image pour corriger les artéfacts survenant lors d'une synthèse par concaténation (ajustement de la position de la tête, *glitch* au niveau de la mâchoire ou des dents, etc.).

1.4 Discussions

Dans ce chapitre, nous venons de constater une nette tendance des systèmes de prédiction de mouvements articulatoires vers la modélisation statistiques, que celà soit par des méthodes *end-to-end* ou par des systèmes hybrides ou le modèle statistique guide la sélection de références. Ces approches obtiennent actuellement des mouvements articulatoires bien plus fidèles à ceux observables chez l'humain que les anciennes approches utilisant explicitement un modèle de coarticulation. Cependant, de tels systèmes sont bien

souvent intrinsèquement liés à la représentation choisie pour l'espace articulatoire (visuel ou conduit vocal), et donc difficilement adaptable à de nouvelles modalités, contrairement aux modèles de coarticulation. Par exemple, le modèle de Öhman (1967), historiquement développé pour modéliser les mouvements de la langue, fut de nombreuses fois réutilisé pour inférer le mouvement des lèvres.

Malheureusement, ces différents modèles de coarticulation proviennent pour la plupart d'études expérimentales différentes, et ne considèrent que quelques paramètres influençant la production de la parole. Nature diverse des articulateurs, langue du locuteur, vitesse d'élocution, état émotionnel, anatomie... Cette multitude d'aspects à prendre en compte semble énormément complexifier le développement d'un modèle de coarticulation par un algorithme *ad hoc*, ce qui explique certainement la capacité de la communauté scientifique à contredire la prédiction de ces modèles de coarticulation lors de nouvelles conditions expérimentales. Par exemple, le cas du modèle *look-ahead*, dont l'étude de Abry and Lallouache (1995) montre que ce dernier n'est plus en accord avec les observations si nous augmentons la taille de la séquence de consonnes intervocaliques.

Dès lors, une première question vient à nous : est-il possible de combiner l'avantage de la modélisation statistique et de l'utilisation d'un modèle de coarticulation ? En d'autres termes, pouvons-nous exploiter la puissance de l'apprentissage automatique pour directement modéliser la coarticulation, afin d'aboutir à un modèle générique pouvant prendre en compte de nouvelles données lors d'acquisitions supplémentaires. Ainsi, nous pourrions réutiliser ce modèle avec de nombreuses technologies différentes pour la synthèse de l'articulation multimodale, tout comme il est le cas avec les modèles de coarticulation *look-ahead* et *time-locked*.

Une deuxième question, plus épineuse, est celle des données d'entrée nécessaires à une telle approche. Les modalités acoustiques et phonétiques semblent les plus couramment exploitées dans la littérature, que cela soit par les systèmes de prédiction ou par les modèles de contrôle moteur de la parole. L'acoustique étant une résultante directe de l'articulation, de nombreuses informations à propos de la dynamique des articulateurs sont contenues dans le signal audio. Cependant, exploiter ces indices acoustiques est loin d'être trivial, comme en témoigne la littérature sur l'inversion acoustique. De plus, la quantité de données articulatoires disponibles est toujours limitée, en particulier vis-à-vis du nombre de locuteurs. Une approche par apprentissage machine risque donc d'être difficilement transférable d'un locuteur à l'autre si nous nous contentons d'un corpus mono locuteur.

Si nous faisons abstraction de l'aspect temporel, la modalité phonétique a ce grand avantage d'être indépendante du locuteur, permettant une représentation de plus haut ni-

veau de la parole. Elle fut extensivement utilisée pour la modélisation de la coarticulation, mais aussi par le modèle de contrôle moteur GEPPETO. Une des plus grandes réussites de GEPPETO est par ailleurs la capacité de répliquer certaines caractéristiques cinématiques des mouvements articulatoires (Payan and Perrier, 1997; Perrier et al., 2003; Perrier and Fuchs, 2008). Ces travaux semblent indiquer qu'un contrôle explicite de ces phénomènes n'est pas nécessaire, car ces derniers sont des propriétés émergentes du modèle. Certaines informations contenues dans le signal acoustique peuvent donc être inférées depuis une représentation phonétique de la parole.

Nous pouvons également noter que l'utilisation de modalité visuelle fut elle aussi explorée pour la prédiction des mouvements articulatoires, que cela soit par le biais de simple modèle linéaire (Yehia et al., 1998; Jiang et al., 2002a; Bailly and Badin, 2002; Engwall and Beskow, 2003), des HMM et GMM (Ben Youssef et al., 2010), ou encore des machines à vecteurs de supports (Toutios et al., 2011). Bien que les améliorations portées au système d'inversion acoustique par le domaine visuel soient marginales, ces résultats attestent d'une corrélation exploitable entre le domaine visuel et articulatoire.

Pour finir, nous pouvons nous demander quel espace de sortie sélectionner pour un modèle de coarticulation basé sur l'apprentissage automatique. Si nous voulons conserver une grande indépendance vis-à-vis de la représentation choisie pour l'espace articulatoire, le nuage de point semble être un candidat de choix, et est d'ailleurs celui principalement utilisé dans les travaux sur l'inversion acoustique à base de modélisation statistique, ainsi que de plusieurs systèmes de prédictions de mouvements articulatoires. Un nuage de point ne contient cependant pas l'information complète quant aux contours des articulateurs ou aux légères déformations du visage, cette erreur d'approximation pouvant néanmoins être grandement réduite en augmentant le nombre de points. Il semble donc impératif de vérifier la capacité de notre modèle à pouvoir exploiter efficacement des nuages de points de petites comme de grandes tailles, afin de conserver une importante flexibilité vis-à-vis des données nécessaires à l'apprentissage de notre modèle.

Chapitre 2

Modéliser la coarticulation par les réseaux de neurones

Comme abordé durant la discussion du chapitre 1, la difficulté de modéliser la coarticulation par une approche algorithmique traditionelle nous a convaincu qu'une modélisation statistique de la coarticulation est une solution raisonnable. En effet, développer un modèle capable de prendre en compte de multiples paramètres comme le locuteur, les articulateurs considérés, le langage et de nombreux autres facteurs influançant la production de la parole, est une tâche extrêmement complexe. Un modèle statistique quand à lui, peut être entraîné sur d'important volume de donnée, et considérer un nombre de paramètres considérables. Cette approche par apprentissage machine est par ailleurs courament utilisée par les plus récents et performants systèmes de prédictions de mouvements articulatoires. Trois modèles semblent se démarquer dans la littérature par leurs capacité à modéliser les différents aspects de la parole multimodale, et en particulier par leurs capacités à gérer la dynamique de l'articulation : les modèles de Markov cachés (HMM), les réseaux de neurones convolutionels (CNN) et les réseaux de neurones récurrents (RNN).

L'utilisation des HMMs est sur le déclin, mais ces derniers possèdent l'avantage d'être une méthode bien établis avec des bases théoriques très solides, permettant une analyse du modèle bien plus aisée. Cependant, leurs tendances à "moyenner" les résultats de la prédiction par rapport aboutissent souvent à des phénomènes d'*undershooting* où la prédiction n'atteins pas les cibles articulatoires. Comme précédement évoqué dans l'ouvrage, cette incapacité est très critique pour certain phonème comme les bilabiales ou les labiodentales. Les réseaux de neurones se sont imposés ces dernières années comme un standard du traitement de la parole et des langages naturels, de part leurs capacités à modéliser des relations très complexe à partir de données brutes.

Les CNN se sont imposés comme un standard du traitement d'image de part leurs capacités à modéliser des caractéristiques indépendantes de la position. Grossièrement, quand un CNN analyse une image, un motif donné sera interpreté par le réseaux indépendament de sa position sur l'image. Cette capacité peut-être facilement utilisé pour rechercher des motifs dans le domaine temporelle, et de nombreuses études ont explorés l'utilisation des couches convolutionels pour analyser des séquences temporelles comme du texte ou un signal acoustique. Cependant, dans le cas la prédiction de l'articulation, l'utilisation d'une telle approche aboutis à la génération de trajectoire erratique. En effet, la fenêtre temporelle sur laquelle s'applique la convolution ne peut considérer le segment de parole dans son ensemble, et d'importante information peuvent être omise d'un pas de temps à l'autre, changeant radicalement la prédiction. Une procédure de lissage supplémentaire peut aider à corriger ces défauts, mais ces derniers trahissent d'une réelle incapacité de ces modèles à prendre en compte la dynamique de l'articulation, réalisant une modélisation que nous pouvons qualifier de "segment à point", ou un segment de parole résulte en la prédiction d'une unique configuration des articulateurs.

Les RNNs ne subissent pas cette limitation, ce qui leurs a certainement permis d'atteindre l'état de l'art dans l'inversion acoustique. Leurs capacités à traiter une séquence de taille arbitraire, et à réellement modéliser la relation entre les dynamiques des signaux d'entrée et de sortie, en font un outils de choix pour la prédiction de mouvements articulatoires. Les RNNs bidirectionnels semblent être particulièrement adaptés à notre tâche car ces derniers ont accès à l'intégralité de l'information future, permettant donc de considérer le segment de parole dans sa totalitée, ce qui nous affranchis de toute supposition quand à la durée maximale des phénomènes d'anticipation. De plus, nous ne nous attendons logiquement à ce que les phénomènes de coarticulation ne s'étendent pas sur de très grandes durées, là où les capacités des RNNs sont encore limitées. Nous présenterons en profondeur ce modèle à la section 2.1.

Pour finir, notre modèle reposera sur l'utilisation d'une représentation phonétique de la parole. En plus de la pertinance de cette représentation présentée en 1.4, et de sa nature indépendante du locuteur, cette simple représentation permet de prendres en compte de nombreux paramètres comme le langage, la vitesse ou le rythme d'élocution. La modalité d'entrée sera donc les phonèmes et leurs durées respectives composant le segment de parole dont nous voulons prédire l'articulation. Cette articulation sera représenté par un nuage de points, composé de points d'intérêts du visage et du conduit vocal, ces derniers étant stratégiquement positionner pour une approximation correcte du contours des articulateurs, comme il est d'usage dans la littérature. Ainsi, nous pouvons exploiter de nombreuses technologies d'acquisition différentes pour entrainer le modèle, mais pour

le cas des technologies d'imageries il sera néanmoins nécessaire de prendre en compte les erreurs résultant d'une procédure de détection automatique de points d'intérêts, ou d'investir une quantité non-négligeable de temps pour l'anotation manuelle des données.

2.1 Les réseaux de neurones récurrents

2.1.1 Les réseaux récurrents bidirectionels

Les réseaux de neurones *feed-forward* sont aujourd'hui connus pour être des approximateurs universels (Hornik et al., 1989), mais ces derniers ne peuvent traiter qu'un vecteur de taille fixe et ne possède aucune notion de temporalité. Les travaux de Rumelhart et al. (1986) ouvrent la voie à la création des réseaux de neurones récurrents, modèle capable de traiter une séquence temporelle de taille variable. Siegelmann and Sontag (1995) démontre par la suite que ceux-ci sont Turing-complet, et peuvent donc approximer n'importe quel algorithme. Pour réussir ce tour de force, les RNNs exploitent un état interne, leur servant de "mémoire". A l'instant t, l'état interne du RNN h_t est modifié en fonction de l'entrée x_t et de l'état interne précédent h_{t-1}, et devient donc un condensé de la séquence $(x_0, \ldots, x_t)$. Cet état interne permet au modèle d'apprendre les correlations temporelles présentes au sein du corpus.

Un réseau de neurones recurrents peut être formalisé comme suit :

$$\begin{aligned} h_t &= \mathcal{H}(x_t, h_{t-1}; \theta) \\ y_t &= W_{output}.h_t + b_{output} \end{aligned} \tag{2.1}$$

où h_t est l'état interne du réseau, x_t l'entrée à l'instant t et y_t la sortie correspondante. W_{output} sont les poids des connections de la couche de sortie (ici une simple couche linéaire), et b_{output} le vecteurs de biais associé. $\mathcal{H}$ est la fonction de transfer de la couche cachée recurrente, paramétrisé par l'ensemble θ.

Cependant, ces réseaux récurrents sont limités à l'utilisation de l'information *passé*, alors que l'accès à l'information *future* peut dans de nombreux cas aider à la prédiction de y_t. Ceci est particulièrement vrai dans le cas de la production de la parole et de la modélisation de la coarticulation, où la coarticulation anticipative joue un rôle majeur (cf. chapitre 1). Schuster and Paliwal (1997) propose dans ce contexte d'accès à l'information *future* un réseau récurrent bidirectionnel, simultanément entraîné dans les deux directions temporelles. La solution envisagée par Schuster et Paliwal est illustée à la figure 2.1, cette dernière consiste en la duplication de chaque couche récurrente du réseau : l'une des couches analysera la séquence de gauche à droite, comme un RNN classique, et

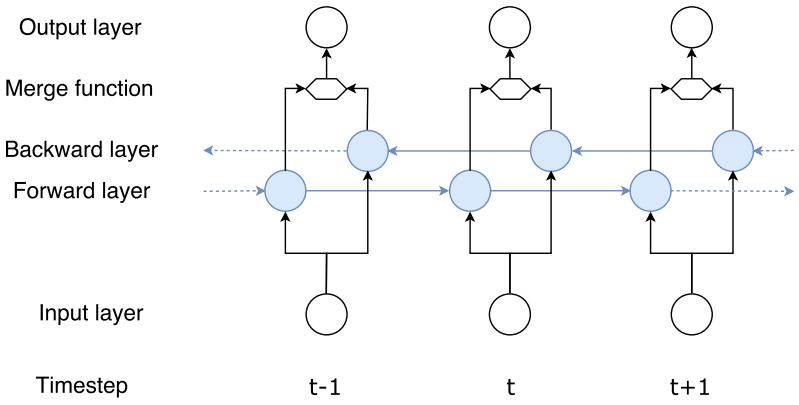


FIGURE 2.1 – Aperçu d'un réseau bidirectionnel.

l'autre traitera la séquence de droite à gauche. A l'instant t, deux états internes cohabiterons alors dans le réseau, l'un résumant la séquence $(x_0, \ldots, x_t)$ et l'autre résumant la séquence $(x_t, \ldots, x_n)$. Ces deux états peuvent alors être fusionné, usuellement par simple concaténation, avant d'être utilisé par la couche suivante.

Etendre l'équation 2.1 aux cas des réseaux bidirectionnels nous donnent les équations :

$$\begin{aligned} \overleftarrow{h}_t &= \overleftarrow{\mathcal{H}}(x_t, \overleftarrow{h}_{t+1}; \overleftarrow{\theta}) \\ \overrightarrow{h}_t &= \overrightarrow{\mathcal{H}}(x_t, \overrightarrow{h}_{t-1}; \overrightarrow{\theta}) \\ y_t &= W_{output}.\mathcal{M}(\overleftarrow{h}_t, \overrightarrow{h}_t) + b_{output} \end{aligned} \tag{2.2}$$

où $\overleftarrow{.}$ correspond aux éléments de la couche *backward* et $\overrightarrow{.}$ correspond aux éléments de la couche *forward*, et $\mathcal{M}$ est la fonction de fusion des informations passées et futures.

2.1.2 Les limites du *vanishing gradient*

Malgré les capacités théoriques des réseaux de neurones récurrents, leurs versions les plus simples sont difficilement utilisable en pratique, car ceci souffrent du problème du *vanishing/exploding gradient* (Bengio et al., 1994). En effet, si nous prêtons attention a l'implémentation la plus simple de la fonction de transfer $\mathcal{H}$ d'un RNN nous obtenons l'équation suivante :

$$h_t = f(W_r.h_{t-1} + W_i.x_t) \tag{2.3}$$

Dans cette équation W_i est la matrice des connexions entre l'entrée et l'état interne, W_r entre l'état interne et lui-même, et f est la fonction d'activation de la couche récurrente, usuellement la fonction sigmoïd ou tanh.

Nous pouvons nous rendre compte que le calcul du gradient des connexions récurrentes W_r nécessite une multiplication par W_r par élément de la séquence. Si nous prenons la comparaison avec les scalaires, nous pouvons nous rendre intuitivement compte que si W_r possède un petit rayon spectral (<1) le gradient va progressivement "disparaître", et progressivement "exploser" dans le cas contraire (>1). Ceci rend ces modèles peu enclin à l'apprentissage de correlation temporelle se déroulant sur une longue période.

Afin de contourner cette problématique, Hochreiter and Schmidhuber (1997) proposent un modèle de réseaux de neurones récurrents, les LSTM (*Long Short-term Memory*), où l'état interne est modifié dynamiquement en fonction de l'entrée par l'ajout d'incréments, encourageant ainsi l'information à rester bien plus longtemps au sein de ce dernier. Pour ce faire, chaque unité de la couche récurrente possède trois "portes" servant à ajuster les poids de ses connexions récurrentes en fonction de l'état interne et des données d'entrée, et deux états internes.

Pour l'architecture LSTM, $\mathcal{H}$ est implémentée par :

$$\begin{aligned} f_t &= \sigma(W_f.[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i.[h_{t-1}, x_t] + b_i) \\ o_t &= \sigma(W_o.[h_{t-1}, x_t] + b_o) \\ C_t &= f_t * C_{t-1} + i_t * \tanh(W_C.[h_{t-1}, x_t] + b_C) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{2.4}$$

où C est la cellule de mémoire, f, i et o sont respectivement les portes de *forget*, d'*input* et d'*output*. σ est la fonction sigmoïd, et les divers W et b correspondent aux matrices des poids du réseaux et les vecteurs de biais associés. La concaténation de vecteurs est dénotée $[.,.]$ et $*$ est un produit point-à-point.

Parmis les différentes variations des LSTM, les réseaux GRU (*Gated Recurrent Units*) proposé par Cho et al. (2014) sont particulièrement populaire. Cette architecture réduit la complexité des LSTM, principalement en supprimant une des portes et le double état interne, permettant ainsi de réduire le nombre de paramètre de l'architecture, et donc la quantitée théorique de données nécessaires à un bon apprentissage.

Pour l'architecture GRU, $\mathcal{H}$ est implémentée par :

$$\begin{aligned} z_t &= \sigma(W_z.[h_{t-1}, x_t] + b_z) \\ r_t &= \sigma(W_r.[h_{t-1}, x_t] + b_r) \\ h_t &= z_t * h_{t-1} + (1 - z_t) * \tanh(W_h.[h_{t-1} * r_t, x_t] + b_h) \end{aligned} \tag{2.5}$$

où z et r sont les portes d'*update* et de *reset*.

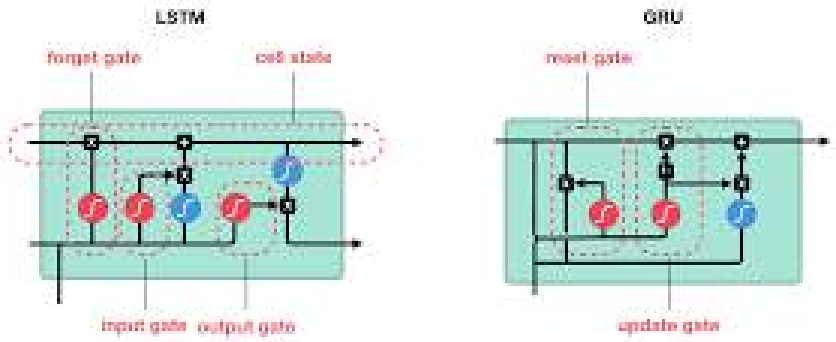


FIGURE 2.2 – Architecture de LSTM et GRU. (crédit : Michaël Nguyen)

2.2 Présentation de notre modèle

De part la faible quantité de données habituellement disponibles dans les corpus audiovisuels, nous avons minimiser le nombre de paramètres de nos réseaux de neurones, afin d'assurer un apprentissage efficace. De ce fait, nous avons privilégié l'utilisation de l'architecture GRU. Notre réseau est principalement composé de plusieurs couches consécutives de RNNs bidirectionnel, usuellement deux dans nos expérimentations, puis d'un réseau de neurones *feed-forward* permettant de projeter l'état interne de la dernière couche bidirectionnelle dans notre espace de sortie. Le passage de cet état interne à la position des articulateurs considérés peut s'effectuer directement, grâce à une unique couche linéaire, ou par le biais d'un espace latent pré-calculé, comme nous l'expliciterons à la section 2.3.

2.2.1 Représentation des données

Notre réseaux de neurones a pour objectif de prédire une séquence de positions spatiales $A = (a_0, \ldots, a_T)$, depuis une séquence de phonèmes $\Phi = (\phi_0, \ldots, \phi_T)$.

a_t représente ici un vecteur appartenant à $\Re^n$ composé des coordonées spatiales de tous les articulateurs considérés à l'instant t, ces derniers peuvent librement être choisis en fonction de la coarticulation que nous souhaitons modéliser. Dans cette étude, a_t est de nature variée en fonction des bases d'apprentissages utilisées. En particulier, nous pourrons retrouver dans cette étude l'utilisation d'un nuage de point 3D représentant le visage, acquéris avec un système de motion capture, d'un nuage de points 3D représentant la langue, acquéris avec un articulographe électro-magnétique, ou encore d'un nuage de points 2D représentant la langue dans l'axe médio-sagittal, acquéris aussi avec un articulographe électro-magnétique. n sera donc égale à $3n_{capteurs}$ ou $2n_{capteurs}$, avec $n_{capteurs}$ le nombre de points d'articulation considérés. Dans la suite de l'ouvrage, nous ferons la

distinction entre A, la séquence cible issue du corpus de données, et $\hat{A}$, la séquence telle que prédite par nos modèles.

ϕ_t quand à lui est un vecteur *one-hot* représentant le phonème articulé à l'instant t. Sa dimension est $n_{phonemes}$, le nombre de phonème du langage étudié. Cette encodage a l'avantage de préserver la durée de chaque phonème sans avoir à fournir explicitement cette information aux réseaux, et peut être considéré comme un signal binaire multidimensionnel synchronisé avec A. Nous pouvons noter que de part la faible dimension des vecteurs de la séquence d'entrée Φ, nous n'avons pas utilisé de couche d'*embedding* à l'entrée du réseaux afin de transformer ces vecteurs *one-hot* en vecteurs réels de plus faible dimension, comme il est courant lors de l'utilisation des réseaux de neurones pour le traitement de texte. Nous pouvons également assumer qu'un *embedding* implicite sera tout de même appris par le réseaux afin de transformer cette représentation, que nous pouvons presque qualifier de discrète, en une représentation continue, par exemple au niveau de W_i dans l'équation 2.3.

2.2.2 Procédure d'apprentissage

Tout réseau de neurones a besoin de trois éléments principaux pour son entraînement :

— Un corpus, contenant les données que nous voulons modéliser.

— Une fonction de coût, permettant de mesurer l'erreur entre une prédiction du réseaux et la valeur cible contenu dans le corpus.

— Un algorithme d'apprentissage, capable de modifier les paramètres du réseaux afin de minimiser l'erreur renvoyée par la fonction de coût.

Traditionnellement, la base de données est séparée en trois ensembles. L'ensemble principal, contenant généralement 80% des données, est nommé corpus d'apprentissage et est utilisé pour modifier les paramètres du réseau. Le reste des données est séparé entre le corpus de validation (ou de développement) et le corpus de test. Le corpus de validation est utilisé à plusieurs fin, comme le monitoring de l'entraînement afin d'éviter l'overfitting par des techniques dites d'*early stopping*, ou encore la sélection des hyper-paramètres. Le corpus de test, comme son nom l'indique, permet de mesurer les performances finales du modèle sur un ensemble inconnu, afin d'éviter la présence d'un biais maximisant les performances sur cet ensemble malgré une faible généralisation.

Parmis les différents usages possibles d'un RNN. Nous avons choisi l'approche *many-to-many* synchrone (à la toute droite de la figure 2.3), parfaitement adaptée à nos données. En effet, notre réseau va apprendre une fonction de transfert permettant de convertir un signal binaire multidimensionnel, notre séquence de données, en un signal continue

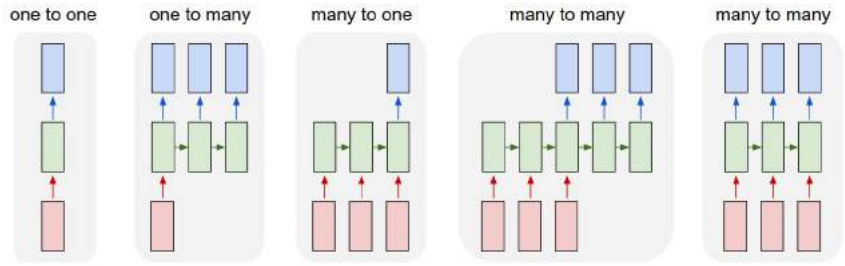


FIGURE 2.3 – Illustration des différentes utilisation d'un RNN. (crédit : Andrej Karpathy)

multidimensionnel, la trajectoire spatiale des articulateurs. Il s'agit donc d'une régression, pour laquelle la fonction de coût usuel est l'erreur quadratique moyenne. Nous avons défini comme erreur la distance euclidienne entre les vecteurs de coordonées a_t et $\hat{a}_t$.

Formellement, nous obtenons donc notre fonction de coût $\mathcal{L}$ via cette équation :

$$\mathcal{L}(A, \hat{A}) = \frac{1}{N} \sum_i \sum_j (a_{ij} - \hat{a}_{ij})^2 \tag{2.6}$$

où N est la taille de la séquence A, et $a_i j$ la dimension j de a_i, le vecteur des coordonées spatiales à l'instant i.

La méthode d'optimisation utilisée pour l'apprentissage du réseau est Adam Kingma and Ba (2015), une extension de la descente stochastique du gradient utilisant un pas d'apprentissage adaptif. Devenu un des standards de la communauté *Deep Learning*, Adam semble combiner les avantages de deux autres algorithmes d'apprentissage réputés, RMSprop Tieleman and Hinton (2012) et AdaGrad Duchi et al. (2011) : il est approprié pour des objectifs non-stationaire et des gradients *sparse*, la mise à jour des paramètres est invariante à l'échelle du gradient, et ses hyper-paramètres sont intuitifs.

$$\begin{aligned} m_t &= \beta_1 . m_{t-1} + (1 - \beta_1) . g_t \\ \hat{m}_t &= m_t / (1 - \beta_1^t) \\ v_t &= \beta_2 . v_{t-1} + (1 - \beta_2) . g_t^2 \\ \hat{v}_t &= v_t / (1 - \beta_2^t) \\ \theta_t &= \theta_{t-1} - \lambda . \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) \end{aligned} \tag{2.7}$$

où λ est le pas d'apprentissage, g_t le gradient de l'ensemble de paramètres θ_t à l'instant t, et ϵ est un très petit nombre nécessaire à la stabilité numérique. m, l'estimation de la dérivée de g, et v, l'estimation de sa dérivée seconde, sont calculés en utilisant deux moyennes mobiles exponentielles de paramètres respectifs β_1 et β_2. Une correction du biais est appliquée à ces moyennes mobiles afin de réduire l'importance des premiers échantillons, aboutissant à $\hat{m}$ et $\hat{v}$.

Pour finir, nous avons utilisé des *minibatch* pendant l'apprentissage, méthode consistant à présenter simultanément au modèle plusieurs exemples de la base d'apprentissage. La métode des *minibatch* est un compromis efficace entre la pure descente de gradient stochastique, où le gradient est calculé sur une unique instance du corpus, et la descent de gradient *batch*, où le gradient est calculé sur l'ensemble du corpus. Cette dernière méthode de calcule converge très rapidement, mais est très couteuse en mémoire, et dans le cas d'une optimisation non-convexe, converge vraisemblablement vers *saddle point* où un minima local peu apte à la généralisation. A contrario, la descente de gradient stochastique "pure" introduit une grande quantité de bruit due à la sélection aléatoire d'un élément du corpus d'apprentissage. La convergence en est donc grandement compliquée, mais le bruit aide à sortir des minima locaux et des *saddle point*, nous permettant d'obtenir des performances finales bien plus intéressantes. En utilisant des *minibatch*, nous diminuons le nombre d'étape nécessaire à l'apprentissage du modèle, tout en préservant les bonnes performances d'une descente de gradient stochastique. L'erreur utilisée pour le calcul des gradient sera, pour la suite de cette étude, la moyenne des erreurs au sein du *minibatch*.

Dans la majorité des expériences rapportés dans ce manuscript, nous utiliserons la procédure d'apprentissage présentée à la section 2.2.2, avec les hyper-paramètres d'Adam recommandés par l'auteur ($\lambda = 0,001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$, $\epsilon = 10^{-8}$) et une taille de minibatch de 2.

2.3 Stratégie d'injection de connaissances

2.3.1 Représentation latente de l'espace articulatoire

L'acquisition de données articulatoires est une tâche coûteuse en temps, mais aussi en ressources matérielles et humaines, et donc limite fortement la quantité de données habituellement disponible dans un corpus d'apprentissage. Afin de palier au mieux à ces faibles quantités de données, nous avons développé une procédure d'initialisation inspirée par l'apprentissage par transfert. Usuellement, lorsque nous évoquons l'apprentissage par transfer et les réseaux de neurones, celui-ci nous renvoie à l'entrainement d'un modèle pour une tâche A, et au remplacement de la couche de sortie du réseau avant l'apprentissage d'une nouvelle tâche B. L'idée principale d'une telle démarche est que si les deux tâches sont suffisament proches, alors l'apprentissage de caractéristiques nécessaires à résoudre la tâche A, calculées par les premières couches du réseau, peuvent aider à améliorer l'apprentissage et les performances vis-à-vis de la tâche B (Pan and Yang, 2009).

Dans cette thèse, nous utilisons également cette idée de réexploitation de caractéris-

tiques précédement apprises. Cependant, l'idée principale n'est plus la réutilisation des premières couches du réseau, servant à l'extraction de caractéristiques depuis les données d'entrées, mais l'injection de caractéristiques de haut niveau spécialement sélectionné pour aider à la bonne représentation et reconstruction des éléments de l'espace de sortie (i.e. le domaine articulatoire ou visuel). Ceci devrait forcer le réseau à utiliser cette réprésentation du domaine articulatoire ou visuel, afin que celui-ci n'ait pas cette tâche à effectuer et puisse se "concentrer" sur l'apprentissage de la coarticulation.

Pour obtenir cette représentation compacte de l'espace articulatoire, nous exploitons une procédure de réduction de la dimensionnalité. Cette dernière est sélectionnée pour nous permettre de calculer un espace latent, une fonction d'encodage et de décodage. Lors de l'application de ces méthodes à l'ensemble d'apprentissage à nos corpus audio-visuelle ou articulatoire, l'espace-latent représentera alors l'articulation du locuteur du corpus avec des attributs structurés de plus haut niveau, pouvant par exemple être relié à des mouvements articulatoires préçis comme l'ouverture ou l'étirement de la bouche. Notre procédure nécessite cependant l'existance d'une fonction de décodage dérivable. En effet, cette fonction sera directement intégrée au réseau afin de réaliser l'apprentissage depuis nos données d'acquisition et non pas depuis une représentation de ces données, afin d'exploiter toute la richesses des données brutes. De plus, ceci a le deuxième avantage de laisser l'opportunité au réseau de modifier à sa convenance l'espace latent lors de l'apprentissage de la coarticulation.

En étandant l'équation 2.2 des RNNs bidirectionnels à notre stratégie d'initialisation, nous obtenons :

$$\begin{aligned} \overleftarrow{h}_t &= \overleftarrow{\mathcal{H}}(x_t, \overleftarrow{h}_{t+1}; \overleftarrow{\theta}) \\ \overrightarrow{h}_t &= \overrightarrow{\mathcal{H}}(x_t, \overrightarrow{h}_{t-1}; \overrightarrow{\theta}) \\ z_t &= W_{output}.[\overleftarrow{h}_t, \overrightarrow{h}_t] + b_{output} \\ y_t &= f_{decoder}(z_t) \end{aligned} \tag{2.8}$$

où z_t est l'espace latent pré-calculé et $f_{decoder}$ la fonction de décodage.

2.3.2 Réduction de la dimensionnalité

Nous détaillons ci-dessous les trois principales méthodes de réduction de la dimensionnalité utilisé dans nos travaux : l'analyse en composante principale, la décomposition en *keyshape* et les réseaux *autoencodeurs*. Pour ces trois méthodes, nous expliciterons également leur intégration avec le réseau de neurones récurrents.

Analyse en composantes principales

Proposé par Pearson (1901) et formalisée par Hotelling (1933), l'analyse en composante principale consiste en la transformation de variables corrélées en un ensemble de variables décorrélées les unes des autres. Ces nouvelles variables, appelées composantes principales, sont une combinaison linéaire des variables originels dont l'objectif est la maximisation de la variance. La première composante principale est donc l'axe exprimmant la plus grande quantitée de la variance présente dans les données, et la dernière composante principale est celui exprimmant la plus petite quantitée de la variance.

Il est d'usage de centrer en zéro les données avant de calculer les composantes, et éventuellement de réduire leurs variances à 1. Dans nos travaux, nous n'effectuons que le centrage en zéro, car toutes nos données sont de même nature, et parce que nous considérons que la présence d'un articulateur ayant une plus grande déviation à sa moyenne (e.g. un point des lèvres) doit effectivement être plus important dans le calcul des composantes qu'un point ayant une faible déviation (e.g. un point peu mouvant des joues). Cette normalisation de la variance est également peu utile car nous exploitons des corpus monolocuteur (présenté au chapitre 3), il n'est donc pas nécessaire de projeter l'articulation des locuteurs sur une échelle commune.

L'analyse en composantes principales, en plus de réussir à décorréler les variables, permet de les débruiter si nous considérons que axes omis sont des axes bruités. Un choix usuel, que nous avons conservé pour la suite de cette thèse, est de sélectionner les premiers axes jusqu'à exprimmer plus de 98% de la variance totale.

La transformation des données originels d en variables décorrélées z s'effectue donc par une simple multiplication avec la matrice des composantes principales P, après soustraction des moyennes m. Il est donc possible de passer de l'espace des données à l'espace latent via ces expressions : $z = (d - m).P^\intercal$ et $d = P.z + m$, ce qui correspond à une couche linéaire dans un réseau de neurones. Appliqué à l'équation 2.8, nous obtenons :

$$\begin{aligned}
\overleftarrow{h}_t &= \overleftarrow{\mathcal{H}}(x_t, \overleftarrow{h}_{t+1}; \overleftarrow{\theta}) \\
\overrightarrow{h}_t &= \overrightarrow{\mathcal{H}}(x_t, \overrightarrow{h}_{t-1}; \overrightarrow{\theta}) \\
z_t &= W_{output}.[\overleftarrow{h}_t, \overrightarrow{h}_t] + b_{output} \\
y_t &= P.z_t + m
\end{aligned} \tag{2.9}$$

Décomposition en *keyshapes*

La décomposition en *keyshapes* provient du monde de l'animation, et est parfois évoquée sous le nom de *morph target animation*, *per-vertex animation*, *shape interpolation* or

encore *blendshapes*. Grande alternative à l'animation à base de squelette, cette dernière consiste en l'élaboration d'un ensemble de différentes versions "déformées" d'un mesh 3D, les *keyshapes*. L'animateur peut ensuite appliquer une intensité différentes pour chaque *keyshapes*, et les vertices finales du mesh 3D sont obtenus par interpolation linéaire.

Dans la plus directe des implémentations, nous pouvons considérer une matrice contenant l'ensemble des vertices des *keyshapes* K et un ensemble de poids p, le mesh résultant s est alors calculé une multiplication matricielle entre ces deux éléments. Dans cette version, chaque poids doit être positif, et la somme des poids de p doit être égale à 1 afin de conserver les proportions initiales du modèle 3D. Si nous voulons obtenir les poids associés à une forme finale s, il nous faut donc résoudre une équation matricielle de type $A = XB$. K étant vraisemblablement non-inversible, la résolution s'effectue usuellement par la méthode des moindres carrés non négatifs. Dans le cas de l'animation faciale, ces *keyshapes* sont dans de nombreuses études composées d'un ensemble de visèmes, parfois augmenté d'un ensemble de déformation lié à l'expressivité (e.g. sourire, haussement de sourcils).

Nous avons exploité une variante courante utilisé (e.g. par le logiciel d'animation propriétaire Maya), *delta blenshapes*, qui nécesitte que l'animateur désigne l'une des *keyshapes* comme une forme neutre k, la matrice K stockant dorénavant le vecteur de différence entre k et les autres *keyshapes*, et p n'a plus qu'une contrainte de positivité pour conserver. Dans ce cas, la forme finale s est obtenues par $s = K.p + k$. En appliquant cette méthode à l'équation 2.8, nous obtenons :

$$\begin{aligned} \overleftarrow{h}_t &= \overleftarrow{\mathcal{H}}(x_t, \overleftarrow{h}_{t+1}; \overleftarrow{\theta}) \\ \overrightarrow{h}_t &= \overrightarrow{\mathcal{H}}(x_t, \overrightarrow{h}_{t-1}; \overrightarrow{\theta}) \\ z_t &= ReLU(W_{output}.[\overleftarrow{h}_t, \overrightarrow{h}_t] + b_{output}) \\ y_t &= z_t.K + k \end{aligned} \tag{2.10}$$

où *ReLU* est la fonction d'activation *rectified linear unit* permettant d'assurer la positivité des poids associés à chaques *keyshapes*.

Pour finir, nous pouvons noter que l'utilisation des blendshapes comme espace latent permet de nous affranchir de l'étape de calcul des poids associés à chaque *keyshapes* lors d'une synthèse visuel de la parole, car ceux-ci peuvent être directement obtenus via z_t dans l'équation 2.8.

Réseaux autoencodeurs

Introduit par Kramer (1991) sous le nom de réseau auto-associatif, et maintenant connu sous le nom de réseau auto-encodeurs, ces réseaux ont la capacité d'apprendre

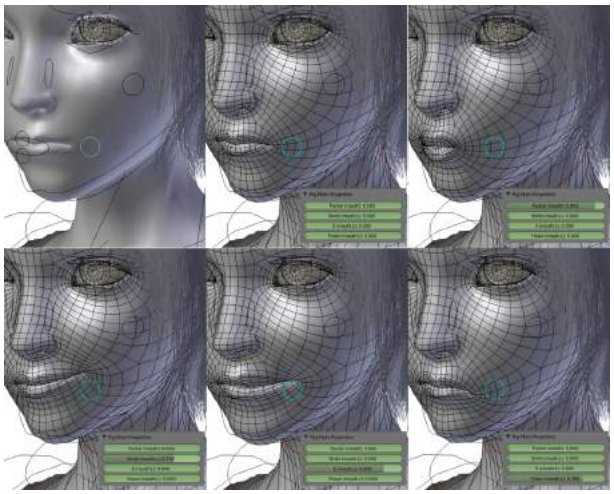


FIGURE 2.4 – Exemple de *keyshapes* du projet open-source Sintel.

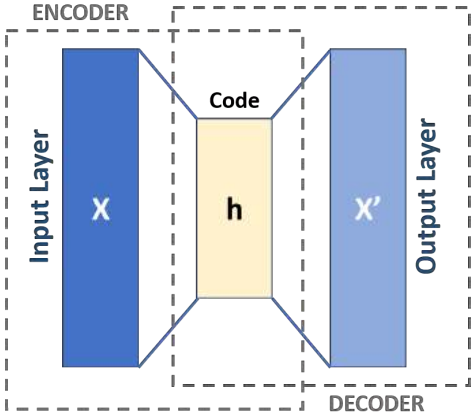


FIGURE 2.5 – Schéma d'un autoencodeur basique.

une représentation d'un ensemble de données de façon non-supervisé. Ces derniers sont capables d'apprendre des représentation plus compacte que l'ACP, principalement grâce à sa capacité d'exploiter des relations non-linéaires (Hinton and Salakhutdinov, 2006).

Dans sa version la plus généraliste, un réseau auto-encodeur se décompose en deux sections bien distinctes : un encodeur, allant de l'espace d'entrée à une représentation compacte, et un décodeur, allant de la représentation à l'espace originel. Entre ces deux, la jonction est effectué par un *bottleneck layer*, une couche dont la dimensionnalité est inférieur à la couche d'entrée (cf figure 2.5). Le réseau apprend alors à reconstruire les échantillons passés en entrée, forçant l'apprentissage d'une représentation compacte des données au niveau du goulot d'étranglement. De nombreuses variations existent autours des réseaux auto-encodeurs, que celà soit dans la procédure d'apprentissage ou dans l'architecture de l'encodeur et du décodeur.

Dans ce cas de figure, l'application à l'équation 2.8 devient trivial, avec $f_{decoder}$ le décodeur du réseau autoencodeur.

2.4 Discussions

Nous venons dans ce chapitre de présenter une architecture de réseau de neurones conçu pour l'apprentissage de la coarticulation, composée principalement de couches récurrentes capables d'apprendre la dynamique d'un signal, et de couches *feed-forward* initialisées dans le but de forcer le modèle à utiliser une représentation latente de l'espace articulatoire. Un schéma global de cette architecture est visible à la figure 2.6, dans le cas où la fonction de transfert $\mathcal{H}$ est composée de deux couches récurrentes, afin d'expliciter l'interconnexion des couches récurrentes. Pour simplifier la visualisation des connexions récurrentes, le modèle est dupliqué pour les pas de temps $t-1$ et $t+1$. Nous retrouvons les différentes couches contenant les paramètres du modèles (rectangles verts), en particulier la couche récurrente, la couche linéaire passant de l'état interne à l'espace articulatoire latent ($W_{output}.h_t + b_{output}$), ainsi que la fonction de décodage $f_{decoder}$. Nous retrouvons également dans cette figure les principales valeurs utilisées et produites par le réseau (cercles) : ϕ_t le phonème courant, h_t l'état interne de la couche récurrente, z_t les valeurs dans l'espace articulatoire latent, et finalement la prédiction $\hat{a}_t$. Nous pouvons également apercevoir le corpus dont provient les modalités phonétique et articulatoire utilisées pour l'apprentissage du modèle. Cet apprentissage est par ailleurs scindé en deux étapes, une phase d'injection de connaissances articulatoires, consistant en l'initialisation de $f_{decoder}$ depuis une méthode de réduction de la dimensionnalité appliquée aux données articulatoires, et l'apprentissage à proprement parler, reposant sur la comparaison de $\hat{a}_t$ et de la cible articulatoire a_t, afin d'obtenir le gradient de l'erreur nécessaire à l'optimisation des paramètres du modèle.

L'initialisation préalable de $f_{decoder}$ permet de découpler le nombre de paramètres du réseau initialisés aléatoirements du nombre de dimension de l'espace de sortie, rendant théoriquement la possibilité d'enrichir cet espace sans complexifier outre-mesure l'apprentissage, en augmentant par exemple le nombre de points du nuage pour améliorer la précision de la représentation de l'espace articulatoire). Cette affirmation semble être expérimentalement confirmée par notre chapitre 4, ainsi que par Karras et al. (2017) qui utilise lui-aussi une stratégie d'initialisation par l'ACP pour représenter l'entièreté d'un mesh 3D d'un visage.

Si nous tentons d'interpréter notre architecture par rapport aux deux principaux modèles de coarticulation, les modèles *look-ahead* et *time-locked*, nous pouvons noter quelques

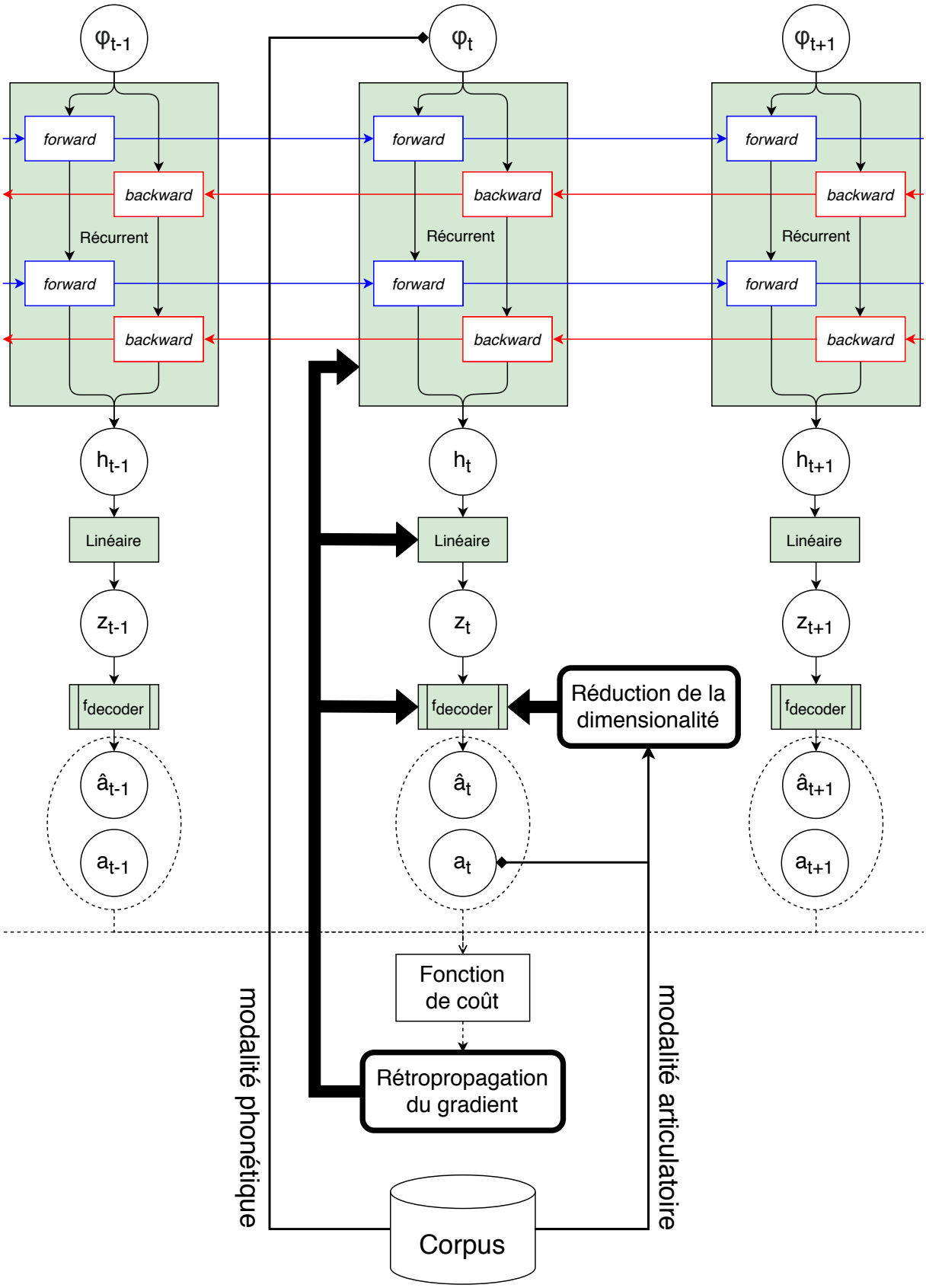


FIGURE 2.6 – Aperçu de notre modèle de coarticulation à base de réseau de neurones avec injection de connaissances articulatoires.

similitudes intéressantes. Premièrement, les couches récurrents bidirectionelles utilisent un mécanisme de lecture *de droite à gauche* semblable aux modèles *look-ahead*, permettant donc à des caractéristiques définies au niveau phonologique de s'étendre vers la gauche (cf éléments en rouge à la figure 2.6). Dans le réseau, ceci se traduit par une conservation de l'information future au sein de l'état interne de la couche récurrente. Néanmoins, la coexistance de cette lecture *de droite à gauche* avec une lecture *de gauche à droite* (cf éléments en bleu à la figure 2.6) rend plus pertinant le parallèle avec les modèles *time-locked*. Pour un instant t, la prédiction de notre modèle dépendra de l'entierté de la séquence, avec un accent important sur les données aux alentours de t. Dans un sens, l'articulation prédite sera donc le résultat des différents phonèmes de la série temporelle, tous *coproduient* à l'instant t, nous rapprochant ainsi fortement de la vision de la coarticulation proposé par les modèles *time-locked*.

Ce parallèle avec la vision de la phonétique articulatoire de Browman and Goldstein (1989), et donc avec les modèles *time-locked* et le modèle TD du contrôle moteur de la parole, est d'autant plus flagrant lorsque nous utilisons notre procédure d'injection de connaissances. Dans ce cas de figure nous pouvons considérer la recherche d'un espace latent par une procédure de réduction de la dimensionnalité comme une recherche statistique des mouvements articulatoires (*articulatory gesture*) présents dans nos données (cas de l'ACP et de l'autoencodeur), ou encore comme une sélection manuelle de ces mouvements (cas des *keyshapes* et visèmes). L'espace latent z_t de l'équation 2.8 correspondrait donc à des scores de gestes articulatoires (*gestural scores*) obtenues par combinaison linéaire de l'état interne h_t de la dernière couche récurrente. Cette vision de notre modèle ferait donc correspondre la majeur partie de notre réseau (de ϕ_t à z_t) au planificateur du modèle TD. Nous pouvons également noter une similitude majeur avec le modèle GEPETTO qui exploite la séquence phonétique comme unité de planning, tout comme notre modèle utilise cette séquence pour inférer les scores de gestes articulatoires.

Une critique pouvant être effectuée est l'absence totale de mécanisme de rétroaction (*feedback*), essentiel aux différents modèles de contrôle moteur de la parole. Dans les modèles DIVA et TD du contrôle moteur de la parole, les objectifs articulatoires sont une des entrées du modèle, servant à calculer les commandes moteurs nécessaire à leurs bonnes réalisations. Notre modèle quand à lui se doit d'inférer ces objectifs articulatoires depuis une unité de planning de plus haut niveau, le phonème, à l'instar du modèle GEPETTO. Cependant, Nous pouvons argumenter que notre modèle utilise bien une boucle de rétroaction somatosensoriel intervenant lors de l'apprentissage du modèle. Cette dernière consiste en la comparaison entre la position souhaitée des articulatoires a_t et la prédiction du modèle $\hat{a}_t$ au sein de la fonction de coût, l'objectif principale de l'algorithme

d'apprentissage étant de minimiser cette erreur. Une deuxième boucle, certainement plus discutable, nous semble localisée au niveau de la dernière couche récurrente. En effet, l'état interne h_t continent des informations partielles sur h_{t-1} (depuis la couche *forward*) et sur h_{t+1} (depuis la couche *backward*). Comme h_t est transformé par une couche linéaire en *gestural scores* z_t, eux-mêmes convertis en nuage de point à l'aide de $f_{decoder}$ (cf. équation 2.8), l'information nécessaire à la prédiction de l'articulation est donc vraisemblablement contenu dans cet état interne. Nous nous retrouvons donc avec une forme de boucle somatosensoriel où nous utilisons une information partielle de l'articulation à l'instant $t-1$ et $t+1$, une partie de h_{t-1} et h_{t+1}, afin de prédire $\hat{a}_t$.

Le modèle résultant de cette conception, bien qu'entièrement basé sur des techniques modernes de modélisation statistique, possède donc d'intéressant parrallèle avec les théories quand à la nature même de l'articulation chez l'humain, ce qui ouvre la voie à une analyse plus minutieuse du réseau de neurone après apprentissage comme nous l'aborderons au chapitre 5.

Chapitre 3

Corpus articulatoires multimodaux

Après avoir introduit les différents systèmes d'acquisition de données articulatoires à la section 1.1.1 ainsi que les besoins de notre modèle à la section 2.2.1, nous nous intéressons ici à la conception et l'acquisition des corpus audiovisuels et articulatoires nécessaire à l'apprentissage de notre modèle. Pour rappel, notre modèle nécessite d'une part la segmentation phonétique correspondant à un segment de parole, c.-à-d. les phonèmes et leurs durées respectives, et d'autre part les trajectoires des articulateurs considérés lors de la production du segment.

La modalité phonétique peut-être obtenue manuellement depuis le signal acoustique, au prix d'un laborieux travail d'annotation, ou automatiquement à l'aide d'une procédure d'alignement. Cette méthode permet un gain de temps considérable, mais nécessite néanmoins une transcription textuelle. Les réseaux de neurones étant réputés pour bénéficier de l'apport d'un faible bruit dans les données (Holmstrom and Koistinen, 1992; Neelakantan et al., 2015), ce dernier agissant comme un régularisateur aidant à lutter contre le surapprentissage (Bishop, 1995), nous estimons que l'apparition d'erreurs de segmentation inhérente à l'utilisation d'un modèle statistique n'est pas critique compte tenu des performances actuelles des modèles acoustiques.

Pour obtenir la dynamique du nuage de points représentant les articulateurs, deux systèmes d'acquisition nous semblent appropriés : l'articulographie électromagnétique pour obtenir la dynamique du conduit vocal (et plus particulièrement de la langue), et les systèmes de capture du mouvement pour le visage. Ces deux systèmes permettent d'enregistrer une quantité bien plus importante de données que les autres alternatives proposées à la section 1.1.1, tout en assurant l'acquisition de la trajectoire de points d'intérêts spécifiques, ce qui nous évite d'avoir recours à un algorithme de *tracking* supplémentaire et nous assure donc d'obtenir des trajectoires articulatoires très précises.

3.1 Corpus textuel

Malgré les progrès matériels et logiciels des dernières décennies, l'acquisition d'une base de données audiovisuelles ou articulatoires est encore une tâche coûteuse en temps ainsi qu'en ressources humaines et financières. De plus, il est important de noter que les modèles à base de réseaux de neurones utilisent dans la littérature une importante quantité de données, typiquement plusieurs centaines d'heures de parole pour l'apprentissage d'un modèle acoustique, ou plusieurs millions de phrases pour l'apprentissage d'un modèle de langage. Pour pallier ces contraintes, il est donc indispensable de préparer un corpus textuel minimisant la durée d'acquisition, mais maximisant la richesse phonétique, afin d'assurer un bon apprentissage du modèle.

Pour ce faire, nous adoptons une approche classique en deux temps. Premièrement, nous constituons un très grand corpus textuel, d'une taille suffisante à pouvoir hypothétiser sans risque que ce dernier contient un grand nombre de spécificités linguistiques, et que la richesse des contextes phonétiques soit suffisante pour couvrir un maximum des phénomènes de coarticulation existant. Dans un second temps, nous minimisons la taille de ce corpus tout en conservant la richesse phonétique. De nombreuses spécificités linguistiques peuvent alors être recherchées en fonction des besoins du corpus. Par exemple, dans le cas d'une synthèse de la parole par concaténation, il peut être intéressant de considérer une couverture maximale des di- ou triphones, en fonction de l'unité de concaténation souhaitée, ou encore de la position de cette unité dans le mot. Ce problème de minimisation est malheureusement NP-complet, entraînant *de facto* une impossibilité de trouver une solution optimale depuis un corpus suffisamment grand. Nous avons donc recours à un algorithme glouton afin d'obtenir une solution acceptable.

3.1.1 Français

Notre première étape fut la création d'un ensemble de 7000 phrases issu de la concaténation de plusieurs corpus de parole ouverts et d'anciens corpus internes à l'équipe de recherche. L'ensemble fut réduit à 2000 phrases en utilisant un algorithme glouton provenant de sojaTTS (Colotte and Lafosse, 2009) permettant de maximiser les spécificités linguistiques en minimisant le nombre de phrases retenues. L'algorithme prend en entrée un ensemble de phrases ainsi qu'une liste de critères (par exemple la couverture en diphone, positions dans le mot, dans la phrase, un groupe rythmique, etc.) et renvoie une liste ordonnée de phrases, où les premières phrases sont celles respectant un maximum de critères, et les dernières phrases sont celles en respectant le moins. En choisissant des critères axés par exemple sur la couverture phonétique, nos premières phrases seront donc

les plus riches phonétiquement.

Ces 2000 phrases présentent environ 200 occurrences pour les phonèmes les plus rares (`j`, `H`, `euf`) et plus de 6000 pour les phonèmes les plus courants (`l`, `a`, `R`). Nous couvrons également 92% des 1369 combinaisons diphoniques possibles depuis les 37 phonèmes du français, avec les 8% restants étant principalement des diphones impossibles ou inexistants dans la langue française (comme `HH`, `i@`, `a9`). Les diphones les plus fréquents, `la`, `d@` et `aR` apparaissent respectivement 206, 184 et 175 fois.

3.1.2 Allemand

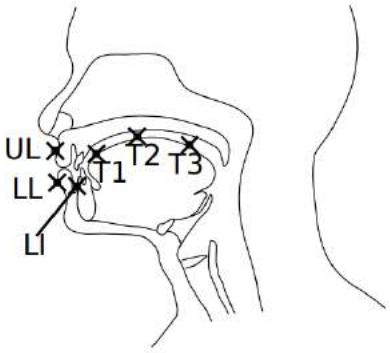
La préparation du corpus textuel allemand s'est déroulée sensiblement identiquement à celui du corpus français. Nous avons utilisé les corpus IFCASL (Trouvain et al., 2016) et VoxForge (MacLean, 2018) afin d'assembler un ensemble de phrases, dont la richesse devrait permettre un bon apprentissage de la coarticulation.

Cependant, nous avons une contrainte logistique forte pour l'acquisition du corpus allemand. En effet, le locuteur allemand n'étant disponible qu'une seule journée dans nos locaux, il nous a été donc impossible d'envisager un enregistrement sur plusieurs sessions. Nous avons donc limité le corpus à 1000 phrases, un nombre qui nous a semblé être la limite que nous pouvons acquérir en une journée d'acquisition. Pour ce faire, nous avons développé un simple algorithme glouton ad-hoc consistant à sélectionner les phrases une à une. À chaque itération, la phrase choisie est celle contenant le plus grand nombre des diphones les moins représentés par les phrases sélectionnées précédemment.

3.2 Corpus articulatoires

3.2.1 Corpus anglais

Nous avons utilisé MNGU0 (Richmond et al., 2011), un corpus articulatoire librement accessible, pour obtenir les trajectoires spatiales de la langue en anglais. Avec environ une heure de données acquises à l'aide d'un articulographe électromagnétique (EMA), MNGU0 est à notre connaissance la plus grande, qualitative et récente base de données articulatoire ouverte. MNGU0 possède 1354 phrases pour 67 minutes de parole, mais d'un unique orateur, et fut spécialement conçu pour assurer une grande richesse phonétique. De plus, la base de données disponible fut enregistrée en une seule journée, sans incident de décrochage d'un des capteurs, ce qui nous assure une unique distribution. Ce point est assez critique pour la qualité du corpus et de l'apprentissage de modèles statistiques,



label	location	label	location
UL	Upper lip	T1	Tongue tip
LL	Lower lip	T2	Tongue body
LI	Lower incisor	T3	Tongue dorsum

FIGURE 3.1 – Positions des capteurs de MNGU0 sur l'axe médio-sagittal. D'après Richmond et al. (2011)

comme démontré par Richmond (2009) lors d'une comparaison entre MNGU0 et MOCHA-TIMIT (Wrench and Hardcastle, 2000).

Un autre avantage de MNGU0 est la présence dans la base de données indispensables pour une bonne reproductibilité des recherches, et a une bonne comparaison des différents modèles. En particulier, nous y avons exploité le découpage préétabli des fichiers pour la base d'apprentissage, de validation et de test, mais aussi les données brutes enregistrées par le Cartsen AG500 à 200Hz et les segmentations phonétiques associées à la parole, obtenues à l'aide de l'outil d'alignement forcé `Multisyn` et du lexique `Combilex`.

Pour finir, nous avons effectué une stabilisation des mouvements de la tête en utilisant les capteurs supplémentaires prévus à cet effet. Cette étape est primordiale pour garantir que la tête ait la même position statique dans tous les fichiers. Les mouvements de la tête commencent par être retirés indépendamment pour chaque session, en utilisant au minimum 3 marqueurs pour calculer la rotation et translation entre la première frame et les suivantes. Il est impératif que ces marqueurs ne changent pas de place sur le sujet durant toute l'acquisition, nous privilégions donc des points au-dessus des oreilles, sur le crâne, le front, la mâchoire supérieure ou le nez. Pour MNGU0, ces points se situent sur les incisives supérieures et les oreilles.

3.2.2 Corpus allemand

Pour la langue allemande, nous avons utilisé un corpus articulatoire qui a été acquis dans le cadre d'une étude phonétique en cours au sein de l'équipe. Ce corpus inclut les 400

premières phrases de notre corpus textuel allemand. L'acquisition fut réalisée à l'aide d'un articulographe AG501 de la société Cartsen. La disposition des capteurs est semblable à MNGU0 sur le plan médio-sagittal, mais à cela s'additionne deux capteurs positionnés à la commissure des lèvres, ainsi que quatre capteurs supplémentaires positionnés sur les bords de la langue, afin d'obtenir des informations sur sa constriction difficilement observable si nous nous contentons de trajectoires en 2D sur le plan médio-sagittal. À ces capteurs visibles à la figure 3.2, s'ajoutent 3 capteurs positionnés sur les oreilles et la mâchoire supérieure, utilisés pour la stabilisation des mouvements de la tête avec la même procédure que pour MNGU0.

Une procédure d'alignement phonétique est finalement réalisée pour obtenir la segmentation phonétique des 400 phrases. Nous avons utilisé Kaldi et un modèle acoustique développé au sein de l'équipe de recherche pour réaliser cette tâche de segmentation automatiquement.

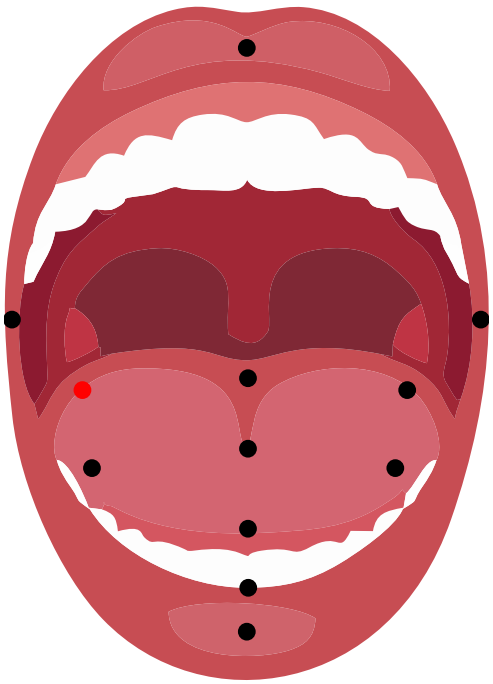


FIGURE 3.2 – Positions des capteurs pour le corpus articulatoire allemand. Le point rouge représente le capteur inutilisé pour la suite de l'étude, car étant tombé pendant l'acquisition.

3.3 Corpus audiovisuels

Nous avons utilisé 8 caméras Flex 13 de la société OptiTrack (fig. 3.3) comme système d'acquisition de données visuelles. Pour le signal acoustique, nous avons utilisé un microphone stéréo Blue Yeti à une fréquence d'échantillonnage de 48kHz, dans un lieu clos par des isolants phoniques sous forme de panneau en mousse. L'ensemble de l'acquisition est également filmé par une caméra numérique, afin de conserver une trace complète du déroulement de l'enregistrement.

La disposition des caméras du système de capture de mouvement est effectuée de façon à ce qu'au moins deux d'entre elles aient toujours une vue sur chaque marqueur, même lors des mouvements de tête des participants. La position du matériel est également contrôlée à l'aide d'un marquage au sol (visible au pied de la chaise à la figure 3.3) afin de conserver une même distance entre le sujet et dispositif d'acquisition et de s'assurer ainsi d'une certaine stabilité sur le plan acoustique.

Afin de rendre les données exploitables, une étape de post-traitement est nécessaire sur les données brutes, afin d'identifier chaque marqueur. Cette dernière est grandement facilitée par le logiciel propriétaire d'OptiTrack, Motive, afin de labelliser chaque capteur à chaque *frame*. Cette phase est rendue quasi automatique par une procédure de *tracking* du logiciel, mais certains cas de figure doivent être résolus manuellement (interpolation lors de la disparition du marqueur, suppression de faux marqueurs dus à des réflexions, résolution des confusions de l'algorithme de labellisation entre deux capteurs proches).

Pour les deux corpus suivants, la disposition des marqueurs visible à la figure 3.4 est grandement inspirée de la norme MPEG-4, ainsi que par une recherche sur les systèmes de *performance capture* du domaine de l'animation.

3.3.1 Corpus français

Ce corpus a été réalisé en collaboration avec Valérian Girard, Vincent Colotte et Sara Dahmani, la principale collaboratrice sur ce projet, et inclut quelques contraintes propres à ses travaux de thèses. En particulier, ce corpus contient de la parole expressive pour 6 émotions de bases (colère, dégoût, peur, tristesse, joie et surprise) et est exploité par Dahmani et al. (2019) dans un objectif de synthèse audiovisuelle de la parole prenant en compte l'état émotionnel. Dans la suite de ces travaux, nous n'exploiterons que la partie "neutre" du corpus.

Comme les besoins du corpus ne rendent pas envisageable l'acquisition sur une unique session d'enregistrement, nous avons développé une simple et peu couteuse stratégie afin de s'assurer d'une disposition quasi identique des marqueurs entre les différentes sessions.

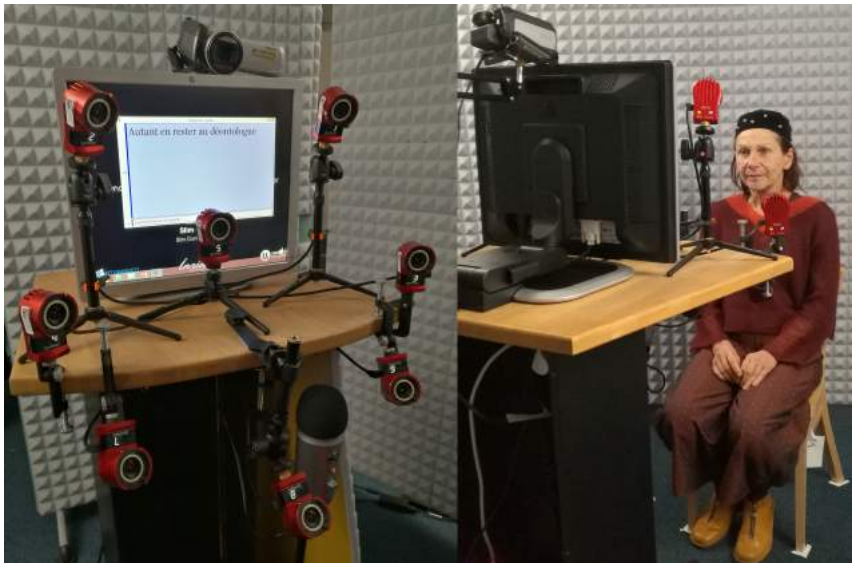


FIGURE 3.3 – Système de *MoCap* OptiTrack (caméra Flex 13) composé de 8 caméras. Ce système a été utilisé pour l'acquisition du corpus français et allemand.

(a) Corpus allemand (b) Corpus français

FIGURE 3.4 – Disposition des marqueurs de nos corpus audiovisuels Allemand (locuteur) et Français (locutrice).

Pour ce faire, une session préliminaire est nécessaire afin de réaliser un scan 3D du visage du participant, permettant une impression 3D relativement peu couteuse d'un masque sur mesure. Avant impression, le modèle de visage sera travaillé à l'aide d'un logiciel de modélisation 3D afin de placer des ouvertures aux positions souhaitées pour les marqueurs, ces trous servant de guide afin de placer avec précision les marqueurs d'une session d'enregistrement à l'autre. Alternativement, il est possible de réaliser le scan 3D après la disposition des marqueurs sur le participant, afin d'utiliser les irrégularités créées par ces derniers pour placer les perforations.

L'acquisition totale du corpus fut réalisée en 4 sessions. Les phrases furent enregistrées par groupe de 50 à 100, en marquant une pause de quelques secondes entre chaque phrase, afin de faciliter le travail de post-traitement et de minimiser la fatigue du sujet en incluant de courtes, mais fréquentes pauses. Le corpus entier fait un total d'environ 4h de parole. Nous avons sélectionné une femme adulte, francophone native, et actrice de théâtre. Nous espérons que l'expérience d'acteurs à avoir une bonne diction et à maintenir un niveau d'énergie constant durant une représentation, afin d'obtenir une prononciation claire et précise, et nous assureras une belle qualité d'articulation pour les données visuelles.

La figure 3.4b nous montre la disposition des différents marqueurs réfléchissants du système de MoCap. Nous avons utilisé 63 marqueurs de 3 mm de diamètre pour l'ensemble du visage, à l'exception du contour du visage où nous avons utilisés des marqueurs de 4mm. Finalement, 6 marqueurs de 9 mm ont été fixés à un bonnet afin de retirer les mouvements de tête par la même procédure que présentée pour les corpus articulatoires. Cependant, stabiliser la tête n'assure pas cette dernière d'avoir la même position d'une session d'enregistrement à l'autre. Pour résoudre cette dernière étape, nous réutilisons la procédure de stabilisation en exploitant les 5 points du front, car la position du bonnet n'est pas contrôlée d'une session à l'autre.

Pour finir, nous avons utilisé Kaldi et un modèle acoustique développé au sein de l'équipe de recherche pour obtenir l'alignement phonétique correspondant à chaque phrase. Le modèle acoustique a été entrainé sur le corpus ESTER (Gravier et al., 2004), et utilise un ensemble de 37 symboles, 36 phonèmes et un silence.

3.3.2 Corpus allemand

Ce dernier corpus se base sur l'ensemble de notre corpus textuel allemand, 1000 phrases, ce qui représente 98 minutes de parole enregistré en unique session d'une journée. D'une manière similaire au corpus français, nous avons utilisé 5 capteurs 9 mm pour le crâne, et 67 capteurs de 3 mm pour l'ensemble du visage.

La présence d'une unique session d'enregistrement simplifie également la procédure de stabilisation de la tête, pour laquelle nous avons utilisé les 5 marqueurs présents sur le crâne du locuteur, ainsi que la préparation du sujet, car il n'est plus nécessaire de créer un masque du locuteur pour assurer une pose reproductible des capteurs. Encore une fois, l'acquisition s'est déroulé par groupe de 50 à 100 phrases pour limiter au maximum la fatigue du locuteur, qui en l'occurrence est un phonéticien natif allemand.

L'alignement phonétique fut réalisé de façon identique à celui du corpus articulatoire allemand, c'est-à-dire en utilisant Kaldi et un modèle acoustique allemand de l'équipe MULTISPEECH.

3.4 Conclusion

Nous possédons maintenant un important volume de donnée permettant l'apprentissage de nos modèles. Ces dernières sont séparées en quatre corpus mono-locuteur distincts :

— un corpus articulatoire anglais ($\sim$ 67'),
— un corpus articulatoire allemand ($\sim$ 38').
— un corpus audiovisuel français ($\sim$ 236'),
— un corpus audiovisuel allemand ($\sim$ 98'),

Ces données nous permettent d'évaluer notre modèle vis-à-vis de deux modalités différentes, articulatoire et visuelle, dans un contexte multilingue (au minimum deux langages par modalité). De plus, la disparité de taille de ces corpus peut être exploitée pour appréhender si la quantité de données à un impacte flagrant sur les résultats. D'une manière similaire, la différence dans la nature des nuages de points entre le corpus articulatoire anglais (trajectoires 2D sur le plan médio-sagittal) et allemand (trajectoires 3D) pourra être exploitée pour s'assurer que notre procédure d'injection de connaissances permet un passage à l'échelle efficace.

La présence des deux modalités dans la langue allemande nous permettra d'expérimenter autour de l'utilisation de la multimodalité dans une tentative de profiter des corrélations entre domaine visuel et articulatoire, comme abordé à la section 1.4. Cependant, cette tâche est compliquée par deux points. Premièrement, l'acquisition n'ayant pas été réalisée en parrallèle, les durées des phonèmes sont différentes pour la même phrase : un apprentissage en simultané des deux modalités est donc difficilement envisageable. Deuxièmement, l'anatomie des deux locuteurs étant également différentes, la tâche en devient plus ardue, car le modèle se devra d'exploiter des relations entre visuel et articulatoire indépendantes du locuteur.

Chapitre 4

Apprentissage des modalités visuelle et articulatoire

Nous rapportons dans ce chapitre les résultats de l'apprentissage par nos modèles vis-à-vis des corpus audiovisuels français [2] et allemand à la section 4.2, puis vis-à-vis des corpus articulatoires en anglais [3] et en allemand à la section 4.3. Nous présentons d'abord à la section 4.1 deux métriques usuelles pour évaluer la qualité des prédictions de trajectoires articulatoires, la RMSE et la corrélation, ainsi qu'une mesure de la vitesse d'apprentissage permettant d'évaluer la facilité avec laquelle les réseaux de neurones sont capables de modéliser la coarticulation. Nous mettons également en lumière dans cette section la nature stochastique de l'apprentissage, ainsi qu'une simple stratégie pour prendre en compte cette variabilité lors de la mesure des performances.

Pour les deux familles de corpus, visuel et articulatoire, introduit au chapitre précédent, nous évaluerons avec soin l'influence de notre procédure d'initialisation dans le cadre d'une injection de connaissances articulatoires en exploitant l'analyse en composantes principales (cf. 2.3.2). Nous conduirons également des explorations spécifiques aux corpus français et anglais. Pour les données visuelles françaises, il s'agit d'une comparaison de trois espaces latents utilisables avec notre procédure d'initialisation. Pour les données articulatoires anglaises, nous présentons une étude de l'influence du nombre de couches récurrentes et de la taille de ces couches, ainsi qu'une comparaison entre les architectures LSTM et GRU (cf. 2.1.2).

Pour finir, nous terminerons par une étude autour de l'exploitation de la modalité vi-

2. Biasutto-Lervat, T., Dahmani, S., and Ouni, S. (2019). Modeling labial coarticulation with bidirectional gated recurrent networks and transfer learning. In *Interspeech 2019*, pages 2608–2612

3. Biasutto-Lervat, T. and Ouni, S. (2018). Phoneme-to-articulatory mapping using bidirectional gated rnn. In *Interspeech 2018*, pages 3112–3116

suelle pour la prédiction des données articulatoires. Pour cela, nous utilisons une procédure d'apprentissage par transfert permettant de réutiliser les couches récurrentes des modèles entraînées à prédire la modalité visuelle pour l'apprentissage de la modalité articulatoire. Cette section 4.4 explore également l'utilisation conjointe de la procédure d'initialisation et de la stratégie de transfert. Plus particulièrement, nous avons observé que les caractéristiques apprises au niveau des couches récurrentes pour la modalité visuelle sont plus utiles à l'apprentissage de la modalité articulatoire lors de l'utilisation de notre procédure d'injection de connaissances articulatoires. Ces expérimentations nous permettant de mieux comprendre et de confirmer la nature de l'apport de notre stratégie d'initialisation, ainsi que la présence d'un mécanisme sous-jacent commun à ces deux modalités.

4.1 Mesures objectives de l'apprentissage

Nous traitons dans cette section les performances des modèles et nous focalisons sur deux aspects de l'apprentissage : sa vitesse et ses performances finales. Comme l'entrainement est stochastique et les paramètres initiaux du modèle sont tirés aléatoirement, le déroulement de celui-ci peut grandement varier d'une instance à l'autre, et il en est de même pour la performance finale d'un modèle. Pour prendre en compte cet aspect, chaque réseau sera indépendamment entraîné plusieurs fois pour chacune de nos expériences. Dans la majeure partie de nos expériences, nous réaliserons un nombre arbitraire de 10 entrainements indépendants, afin d'obtenir une première idée de la distribution des performances tout en conservant un temps de calcul acceptable.

4.1.1 Vitesse d'apprentissage

Afin d'appréhender la vitesse d'apprentissage d'un modèle, nous avons simplement calculé durant l'apprentissage la valeur moyenne de la fonction de coût pour chaque époque. En plus de cette moyenne vis-à-vis de l'ensemble d'apprentissages, nous avons également surveillé l'évolution de l'erreur vis-à-vis de l'ensemble de validation, en calculant sa moyenne à la fin de chaque époque. Nous considérons la durée d'un apprentissage comme étant le nombre d'époques nécessaires pour minimiser l'erreur moyenne sur l'ensemble de validation.

En considérant tous les apprentissages indépendants, il est alors possible de calculer le nombre moyen d'époques d'apprentissage nécessaires pour obtenir les meilleures performances sur l'ensemble de validation. Il est cependant impossible, par souci de lisibilité, de rapporter l'intégralité des erreurs moyennes au cours de l'entrainement sur un

seul et même graphique. Pour pallier cette difficulté, nous présenterons principalement la courbe médiane des erreurs des différents apprentissages, c.-à-d. la médiane des erreurs moyennes pour chaque époque, afin de nous donner une vision du déroulement standard d'un entrainement.

4.1.2 Métriques de performances

Les paramètres ayant le plus faible taux d'erreurs (tel que définis par la fonction de coût) vis-à-vis de l'ensemble de validation sont ceux utilisés pour la mesure des performances. Nous utilisons pour cela deux métriques usuelles pour l'évaluation de trajectoires articulatoires. Premièrement, la racine carrée de l'erreur quadratique moyenne (RMSE, *Root Mean Squared Error*) permet de mesurer l'écart moyen entre la prédiction et la trajectoire originale tout en pénalisant les grandes erreurs et en minimisant les faibles déviations par le passage au carré. Deuxièmement, la corrélation de Pearson ρ nous permet de quantifier l'évolution conjointe de la prédiction et de la vérité terrain. Avec une corrélation de 1, les deux trajectoires sont linéairement corrélées, croissant et décroissant simultanément.

Nous calculons la moyenne des deux métriques pour chaque phrase de notre corpus de test, mais de manière indépendante pour chaque dimension du vecteur de coordonnées a_i. La performance finale sera donc la moyenne des RMSE moyennes de chaque dimension de l'espace articulatoire.

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{N}\sum_i (e_i - t_i)^2} \\ \rho &= \frac{\sum_i (e_i - \bar{e})(t_i - \bar{t})}{\sqrt{\sum_i (e_i - \bar{e}) \sum_i (t_i - \bar{t})}} \end{aligned} \tag{4.1}$$

où e_i est la prédiction à l'instant i, t_i la cible associée $\bar{e}$ est la moyenne des prédictions et $\bar{t}$ la moyenne des valeurs cibles.

Contrairement à la vitesse d'apprentissage, où rapporter l'intégralité des résultats est impossible par souci de clarté, il est tout à fait envisageable de visualiser l'ensemble des performances finales. Pour ce faire, nous aurons recours à l'utilisation de diagramme en violon. Ces derniers sont très similaires à une boite à moustache, et nous informe donc sur le minima, maxima, et la médiane par des traits horizontaux, mais leurs corps sont une estimation par noyaux de gaussienne de la distribution des échantillons, plutôt que d'être une "boîte" informant sur les quartiles.

4.2 Apprentissage de la modalité visuelle

4.2.1 Injection de connaissances

Nous avons débuté notre exploration de l'apprentissage de la modalité visuelle pour les langues française et allemande. La couche récurrente de notre modèle est composée de 2 couches de réseaux GRU bidirectionnel avec 128 unités par couche et direction, choix motivé par sa proximité avec les architectures de l'état de l'art dans l'inversion acoustique (Liu et al., 2015; Zhu et al., 2015). Nous réalisons 10 entrainements indépendants pour les deux corpus audiovisuels, et présentons les performances finales sous forme de diagramme en violon à la figure 4.1. Pour cette première approche, nous étudions particulièrement l'influence de notre procédure d'injection d'un espace articulatoire latent issu de l'analyse par composantes principales.

Nous pouvons observer des performances globales très encourageantes pour les deux langues, avec de très bonnes corrélations et RMSE. Sans procédure d'initialisation, les performances pour le corpus francophone oscillent entre une RMSE de 1,50 et 1,40 mm, avec l'essentiel de la distribution proche de la médiane à 1,48 mm, et la corrélation va

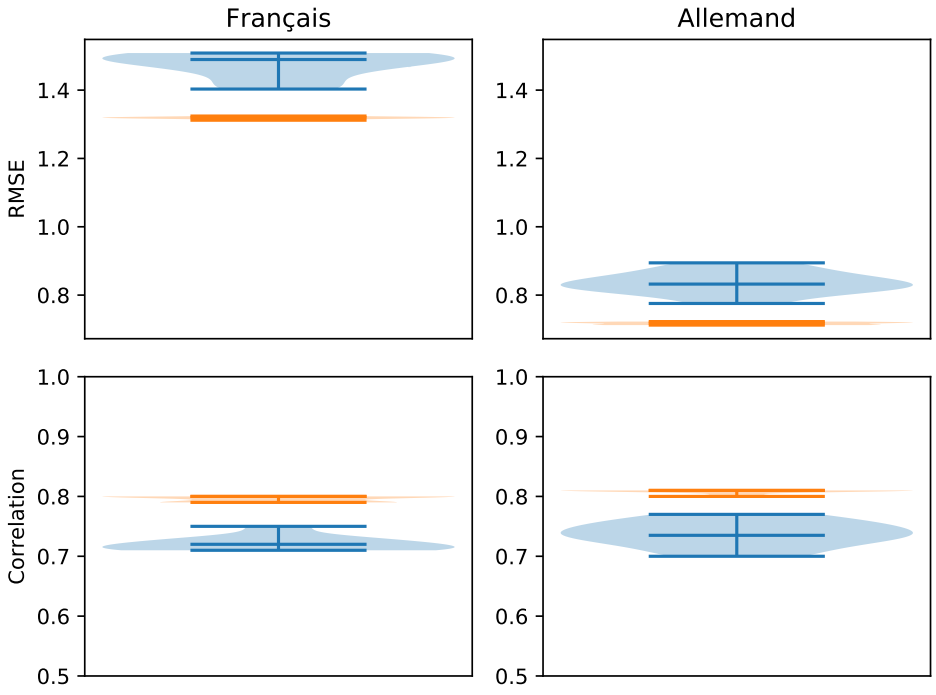
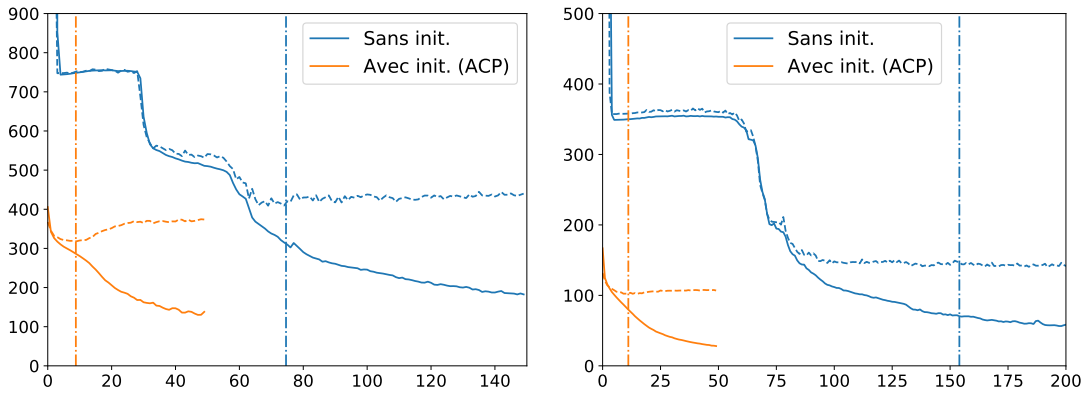


FIGURE 4.1 – Diagramme en violon des performances lors de l'apprentissage de la modalité visuelle. En bleu, sans stratégie d'initialisation, en orange, avec stratégie d'initialisation basée sur l'ACP.



(a) Modalité visuelle française (b) Modalité visuelle allemande

FIGURE 4.2 – Évolution de l'erreur moyenne durant 10 apprentissages indépendants. Les courbes pleines correspondent à l'ensemble d'apprentissages, et celles en pointillés à l'ensemble de validation. Les lignes verticales correspondent à l'époque moyenne où le réseau minimise l'erreur sur l'ensemble de validation.

de 0,71 à 0,75 avec une médiane à 0,72. Pour le corpus allemand, la RMSE oscille entre 0,77 et 0,89 mm avec une médiane à 0,83 mm, et la corrélation entre 0,7 et 0,77 avec une médiane à 0,73. Si nous exploitons la procédure d'injection de connaissances, la RMSE tombe entre 1,31 et 1,32 mm pour le français, avec une corrélation entre 0,79 et 0,8, et entre 0,71 et 0,72 mm pour l'allemand, avec une corrélation entre 0,8 et 0,81.

Étrangement, la RMSE est inférieure au millimètre dans le cas de la langue allemande, et ce malgré un corpus bien plus petit que pour la langue française. Nous expliquons cette différence par l'acquisition en une unique séance du corpus allemand, assurant une unique distribution des positions des capteurs. En effet, la distribution modélisée par le réseau de neurones sera en un certain sens une moyenne de ces multiples distributions, entrainant invariablement une légère erreur entre la trajectoire d'un échantillon du corpus, issue d'un positionnement spécifique des capteurs, et sa prédiction, issue d'une distribution modélisée par le réseau. Nous pouvons également souligner le plus grand nombre de phonèmes utilisés par le corpus allemand, rendant l'évaluation plus difficile en l'absence de grande quantité de données. La figure 4.1 nous permet également de noter l'apport de notre procédure d'initialisation, qui améliore les performances pour les deux langues, mais rend surtout le modèle bien moins sensible à la stochasticité de l'apprentissage, comme en témoigne la très faible variance des résultats sur 10 entrainements indépendants.

La figure 4.2 nous présente quant à elle l'évolution de l'erreur moyenne, telle que définie par la fonction de coût, sur l'ensemble d'apprentissage et de validation. Pour le français,

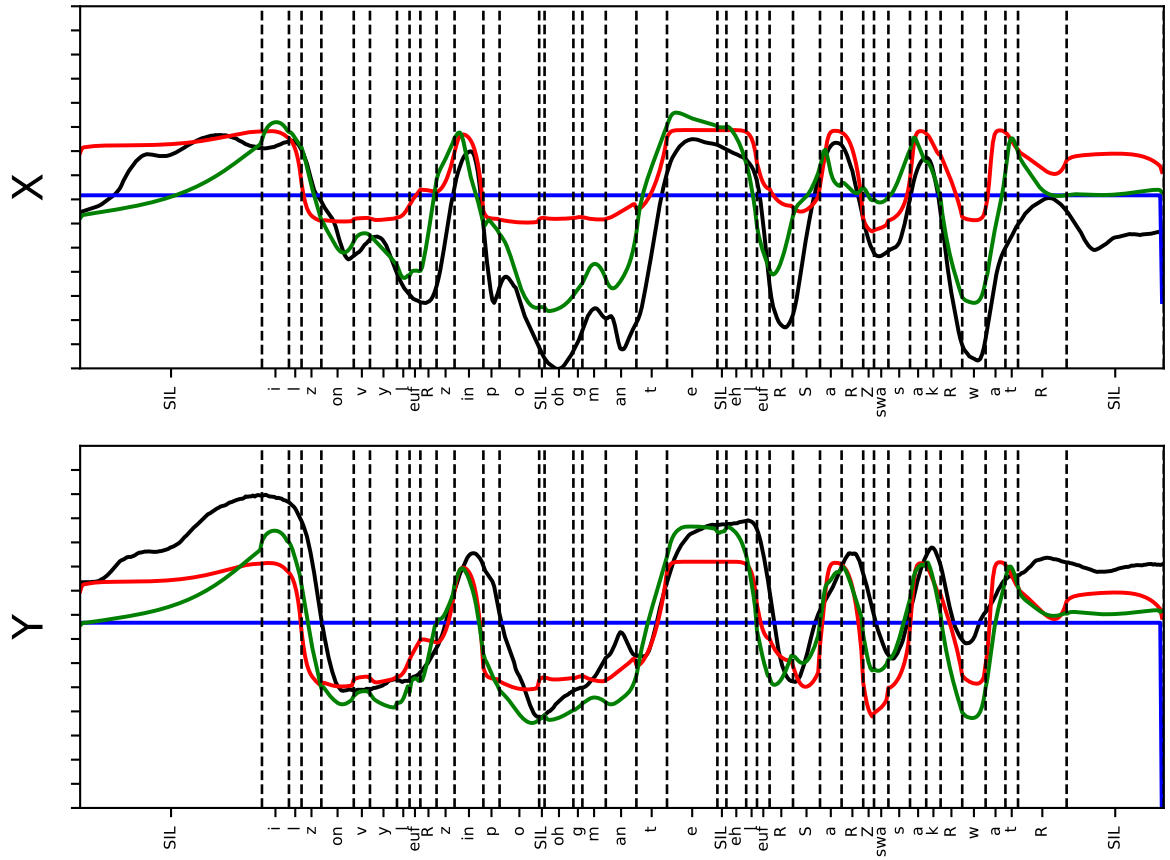


FIGURE 4.3 – Exemple de prédiction du capteur central de la lèvre inférieure pour le français, pour une valeur moyenne de la fonction de coût aux alentours de 750 (bleu), 550 (rouge) et 400 (vert). La trajectoire en noir correspond à la vérité terrain, et la segmentation phonétique est marquée en pointillé.

nous pouvons observer 3 paliers de stagnation durant l'apprentissage, situé approximativement aux alentours d'une erreur moyenne de 750, 550 et 400. Des exemples de prédictions associées à ces plateaux peuvent être observés sur la figure 4.3, où nous présentons un exemple de trajectoire du capteur centrale de la lèvre inférieure. Nous pouvons observer une prédiction stationnaire pour le plateau à 750 (en bleu), un début de mouvement avec un fort effet d'*undershooting* pour le plateau à 550 (en rouge), particulièrement visible sur l'axe X. Cet axe correspond à l'ouverture de la bouche, et nous pouvons constater jusqu'à 6 mm d'écart entre la courbe rouge et noire, ce qui correspond donc à une ouverture de la bouche bien plus légère à ce niveau d'apprentissage du modèle comparé à l'articulation du sujet. Nous obtenons finalement des résultats satisfaisants pour le dernier plateau aux alentours de 400 (en vert). Pour l'allemand, l'entrainement suit globalement la même tendance, avec un premier palier aux alentours d'une erreur moyenne de 350, un second aux alentours de 200, et un dernier légèrement au-dessus de 100. Il a cependant à noter que franchir ce second palier nécessite bien moins d'époques d'apprentissage que dans le cas de la base de données française. Dans les deux cas, nous pouvons apprécier l'impact de notre procédure d'initialisation sur la vitesse d'apprentissage du réseau qui réduit le nombre moyen d'époques nécessaires pour l'obtention du modèle le plus performant sur l'ensemble de validation. Pour le français, nous passons de 74,6 à 8,8 époques, et pour l'allemand de 154 à 11 époques, un ratio oscillant donc entre 8 et 14 fois plus rapidement.

4.2.2 Comparaison d'espaces latents injectables

Dans cette section, nous nous questionnons sur l'espace latent fourni au réseau via la procédure d'initialisation : existe-t-il des représentations plus efficaces que d'autres ? En particulier, un décodeur non linéaire est théoriquement capable de bien plus de richesse, serait-il meilleur qu'une simple transformation linéaire ? Pour répondre à ses interrogations, nous avons comparé les espaces latents et les décodeurs associés obtenus via trois méthodes de réduction de la dimensionnalité : l'analyse en composante principale, représentant des espaces latents linéaires calculés par des méthodes statistiques, les réseaux autoencodeurs, représentant les espaces latents non linéaires calculés par des méthodes statistiques, et la décomposition en blendshapes, représentant les espaces linéaires sélectionnés par rapport à des connaissances expertes. Nous effectuons ces expérimentations vis-à-vis de la base de données française.

Pour s'assurer une comparaison équitable des 3 méthodes, tous leurs espaces latents sont de dimension 13, ce qui correspond à plus de 98% de la variance lors de l'analyse en composante principale. L'autoencodeur quant à lui fut entrainé sur la même base

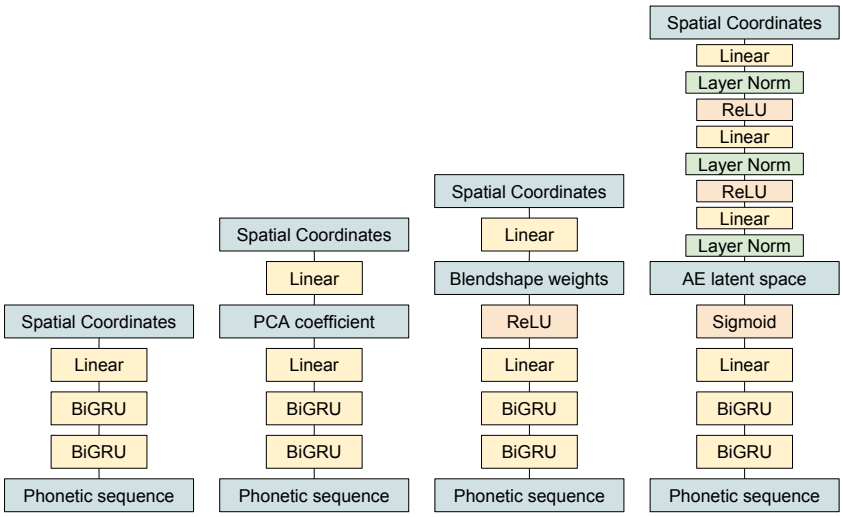


FIGURE 4.4 – Architecture des réseaux de neurones. De droite à gauche : baseline, initialisation avec les composantes principales, initialisation avec les keyshapes, initialisation avec l'autoencodeur.

TABLE 4.1 – Informations à propos de l'ACP, des blendshapes et de l'autoencodeur.

Méthode	ACP	Blendshape	Autoencodeur
Espace latent	$\Re^{13}$	$\Re_{+}^{13}$	$[0;1]^{13}$
Décodeur	Linéaire	Linéaire	Non-linéaire
RMSE (mm)	0,33	0,95	**0,32**

d'apprentissage avec une fonction de coût usuelle, l'erreur moyenne quadratique. Son encodeur et décodeur sont symétriques, composés chacun de 2 couches linéaires de 64 neurones avec une fonction d'activation ReLU. L'espace latent est compris entre 0 et 1 avec une fonction sigmoïde, et nous normalisons à chaque couche pour un apprentissage plus rapide et de meilleures performances (*Layer Normalization* (Ba et al., 2016)). Finalement, les 13 keyshapes utilisés pour la décomposition en blendshape seront présentés plus en détail à la section 6.1, ces derniers sont largement inspirés des travaux sur les visèmes de Benoit et al. (1992) et de Govokhina (2008). La table 4.1 récapitule ces informations et nous indique également l'erreur de reconstruction de chaque méthode. La figure 4.4 quant a elle nous présente l'architecture finale des quatre variantes de cette expérience : sans injection de paramètres, avec injection des composantes principales, avec injection des blenshapes, et avec injection de l'espace latent d'un autoencodeur.

Nous avons débuté cette comparaison par l'entrainement sur 150 époques des quatre modèles présentés sur la figure 4.4, sans notre procédure d'initialisation (c.-à-d. en utili-

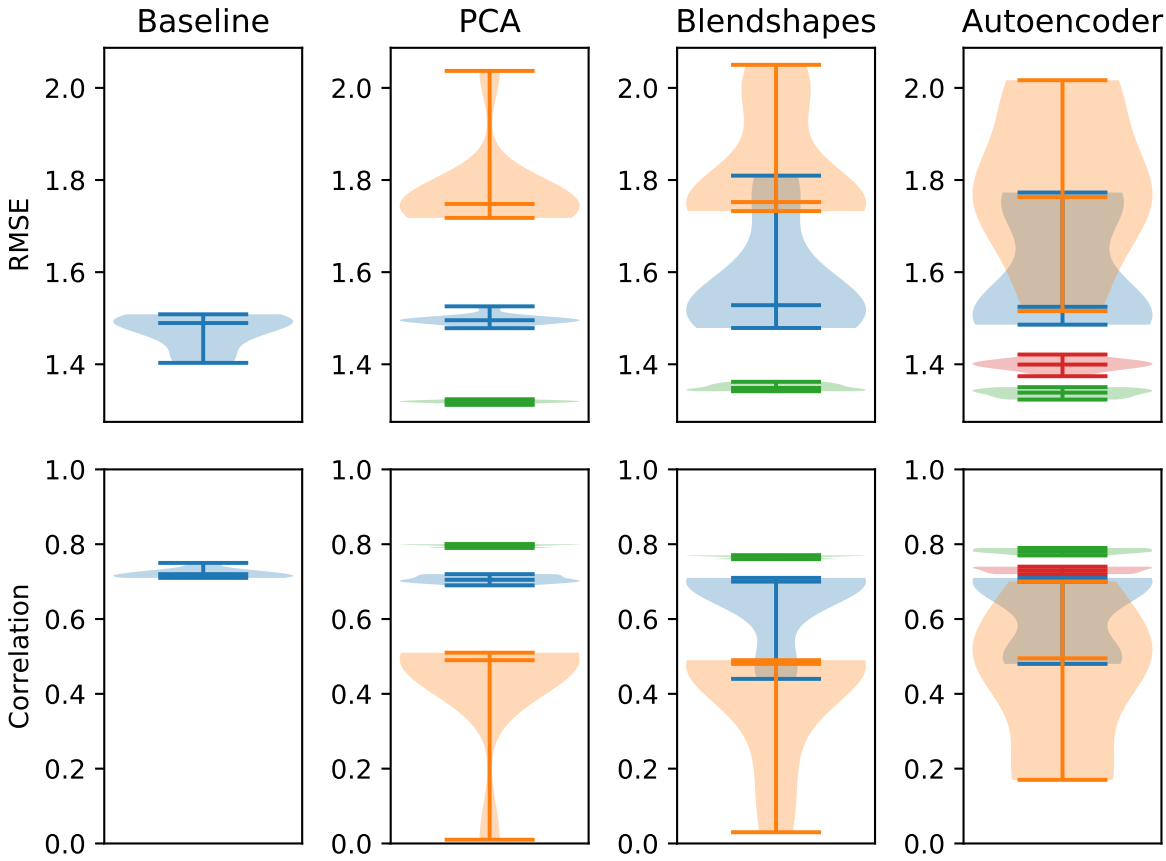


FIGURE 4.5 – Diagrammes en violons des performances pour 10 apprentissages indépendants. Les diagrammes en bleu et en orange correspondent aux modèles avec initialisation aléatoire, entrainés respectivement avec 150 et 50 époques. Les diagrammes en verts correspondent à notre stratégie d'initialisation (50 époques), et le diagramme en rouge correspond à une initialisation aléatoire avec utilisation de *Layer Normalization* (50 époques).

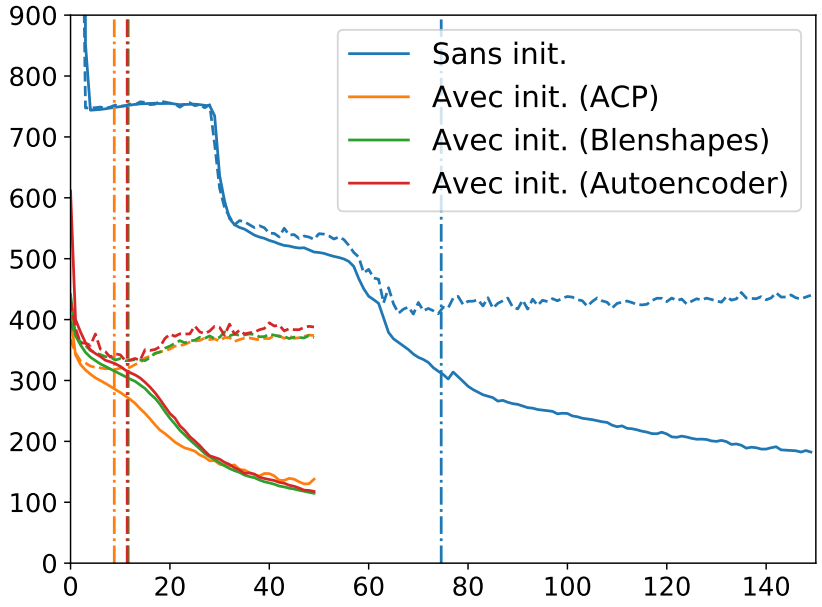


FIGURE 4.6 – Évolution de l'erreur moyenne durant 10 apprentissages indépendants. Les courbes pleines correspondent à l'ensemble d'apprentissages, et celles en pointillés à l'ensemble de validation. Les lignes verticales correspondent à l'époque moyenne où le réseau minimise l'erreur sur l'ensemble de validation.

sant une initialisation aléatoire). Ces résultats, reportés dans le diagramme en violon bleu de la figure 4.5, nous montrent en quoi l'architecture *baseline* (constituée seulement de deux couches GRU bidirectionnel) obtient les meilleures performances, avec une RMSE allant jusqu'à 1,4 mm et une performance médiane aux alentours des 1,5mm. Les autres architectures sont néanmoins fortement équivalentes, possédant toutes une performance médiane légèrement aux dessus de 1,5mm. Nous pouvons néanmoins noter la faible variance de l'architecture PCA, et les grandes variabilités des architectures blendshapes et autoencodeur, avec une performance minimale avoisinant les 1,8 mm de RMSE.

Maintenant que nous avons établi le niveau de performance de ces architectures avec une initialisation aléatoire, nous pouvons observer l'impact de notre procédure d'injection de connaissances, reporté dans les diagrammes en violon vert de la figure 4.5. Nous y remarquons immédiatement les performances similaires des trois modèles, tous meilleurs que notre *baseline* en termes de RMSE et de corrélation, malgré un entrainement de maximum 50 époques. Les trois architectures sont ici fortement équivalentes, avec une performance médiane proche de 1,35mm.

Pour approfondir la comparaison, nous avons également entrainé les architectures avec initialisation aléatoire sur une même durée (50 époques), et reporté ces résultats dans les

diagrammes orange. Nous pouvons alors observer un énorme écart de performance induit par notre stratégie d'initialisation. Pour terminer, dans le cas de l'autoencodeur, nous avons réalisé les précédentes expériences sans tenir compte des méthodes de normalisation utilisée lors de l'apprentissage de l'autoencodeur. Si nous réutilisons cette méthode lors d'une initialisation aléatoire (diagramme rouge), nous obtenons des performances supérieures à celle de la *baseline* (médiane à 1,4mm), mais toujours inférieures à notre méthode d'injection de connaissances articulatoires.

La figure 4.6 présente l'évolution de l'erreur moyenne durant l'apprentissage de notre *baseline* sans méthode d'initialisation, et l'utilisation des trois différentes méthodes de calcul d'un espace latent. En moyenne, notre *baseline* sans procédure d'initialisation met 74,6 époques pour minimiser son erreur sur l'ensemble de validation, contre 8,8 pour l'ACP, 11,6 pour les *blendshapes* et 11,4 pour l'autoencodeur, soit un apprentissage jusqu'à 8,5 fois plus rapide. De plus, les trois architectures exploitant notre procédure d'initialisation ne rencontrent aucun plateau dans l'évolution de l'erreur moyenne durant l'apprentissage, débutant l'apprentissage à des niveaux d'erreur très inférieurs au modèle sans initialisation.

4.3 Apprentissage de la modalité articulatoire

4.3.1 Injection de connaissances

Dans ces expériences, nous abordons les capacités des réseaux de neurones récurrents à prédire la dynamique de la modalité articulatoire, avec une application à deux langues : l'Anglais et l'Allemand. Nous utiliserons la procédure d'apprentissage présentée à la section 2.2.2 avec les paramètres habituels, et un réseau composé de deux couches de réseaux GRU bidirectionnel, avec 128 neurones par couche et direction, et finalement d'une couche linéaire allant vers l'espace de sortie. Pour la procédure d'initialisation, nous avons utilisé l'analyse en composante principale pour obtenir une représentation latente à injecter dans la dernière couche du réseau. Ici encore, nous reportons les résultats issus de 10 apprentissages indépendants pour prendre en compte la stochasticité de l'apprentissage.

Les performances finales des réseaux ainsi que l'erreur moyenne au cours de l'apprentissage sont respectivement rapportées aux figures 4.7 et 4.8. Sans procédure d'injection de connaissances articulatoires, les performances pour le corpus anglais oscillent entre 0,94 et 0,99 millimètre de RMSE pour une corrélation entre 0,88 et 0,99 (médiane respective à 0,95 mm et 0,89). Pour le corpus allemand, la RMSE va de 1,23 à 1,34 mm et la corrélation de 0,59 à 0,66 (médiane respective à 1,29 mm et 0,62). Lors de l'utilisation de

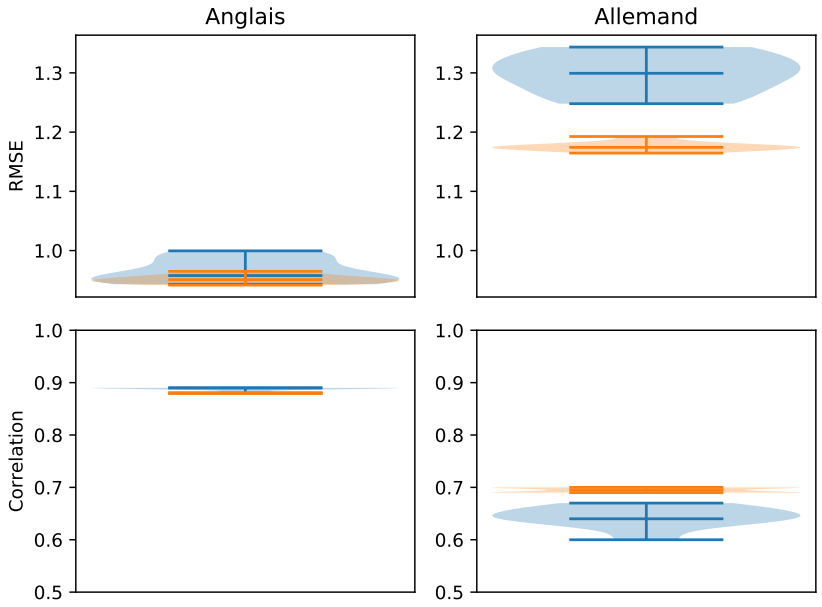


FIGURE 4.7 – Diagramme en violon des performances des architectures lors de l'apprentissage de la modalité articulatoire. En bleu, sans stratégie d'initialisation, en orange, avec stratégie d'initialisation basée sur l'ACP.

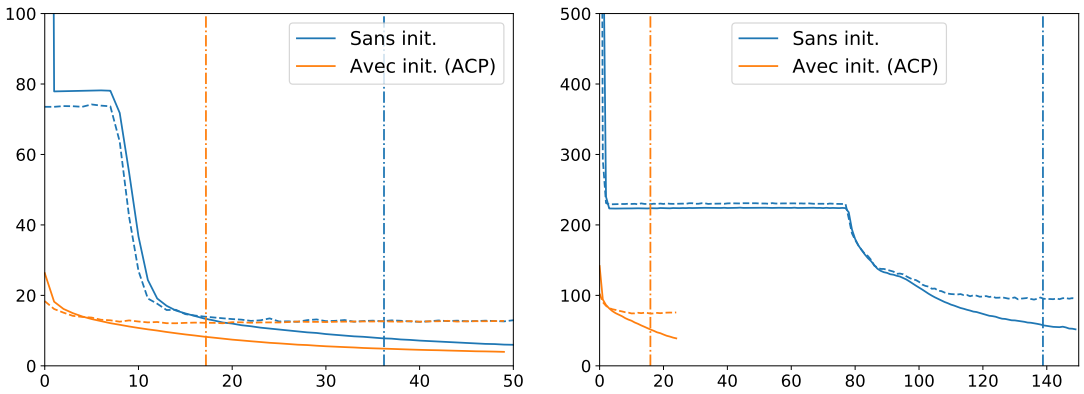
notre procédure d'injection de connaissances articulatoires, les performances ne sont pas améliorées pour le corpus anglais MNGU0, mais leur variance est légèrement diminuée : la RMSE va alors de 0,94 à 0,96 mm (médiane à 0,95) et la corrélation minimum, maximum et médiane vaut 0,88. L'impacte de cette procédure est plus notable pour le corpus allemand : la RMSE varie entre 1,12 et 1,16 mm (médiane à 1,14) et la corrélation entre 0,69 et 0,70 (médiane à 0,69).

Ces résultats nous amènent à formuler les remarques suivantes :

— Les performances sur la base de données MNGU0 sont très bonnes, légèrement inférieures à l'état de l'art dans l'inversion acoustique (Liu et al., 2015; Zhu et al., 2015) et ce malgré l'utilisation de séquences phonétiques bien moins riche en information que le signal acoustique.

— Malgré une taille minimale, les performances sur nos données articulatoires allemandes sont bonnes. Nous attribuons ce succès à notre procédure de création du corpus, optimisant la richesse phonétique de ces 400 phrases. Il est cependant probable que le corpus de test soit trop petit pour que ces résultats soient complètement révélateurs.

— Sur les deux corpus, le modèle passe par une phase de stagnation lors de l'apprentissage, comme pour la modalité visuelle, avec là aussi des prédictions stationnaires



(a) Modalité articulatoire anglaise (b) Modalité articulatoire allemande

FIGURE 4.8 – Évolution de l'erreur moyenne durant 10 apprentissages indépendants. En bleu, sans stratégie d'initialisation, en orange, avec initialisation basée sur l'ACP. Les courbes pleines correspondent à l'ensemble d'apprentissages, les courbes en pointillés à l'ensemble de validation, et les lignes verticales correspondent à l'époque moyenne à laquelle la meilleure performance vis-à-vis de l'ensemble de validation.

lors de ce plateau (cf figure 4.9). Notre procédure d'initialisation aide à contourner cette difficulté, ce qui entraine une nette accélération de l'entrainement : le temps moyen pour atteindre les meilleures performances est approximativement divisé par deux pour la base anglaise et par 10 pour le corpus allemand.

— La procédure d'initialisation permet d'atteindre un nouveau niveau de qualité des prédictions pour le corpus allemand, avec une RMSE plus faible, une corrélation plus haute, et des performances moins variables d'un apprentissage à l'autre. Cette tendance n'est par contre pas retrouvée pour le corpus anglais, à l'exception de la baisse des variabilités. Nous considérons ces résultats comme étant le fruit de l'espace articulatoire de MNGU0, beaucoup moins complexe que l'espace articulatoire de notre corpus allemand (2D contre 3D, capteurs supplémentaires). Plus l'espace articulatoire est simple à modéliser, moins notre procédure d'initialisation améliorera l'apprentissage.

4.3.2 Influence de la taille, la profondeur et l'architecture du réseau

Dans cette section, nous présentons des expérimentations qui ont pour objectif l'amélioration de la procédure d'apprentissage tout en explorant l'influence de la taille et de la profondeur des couches récurrentes. La plus notable différence dans la procédure d'appren-

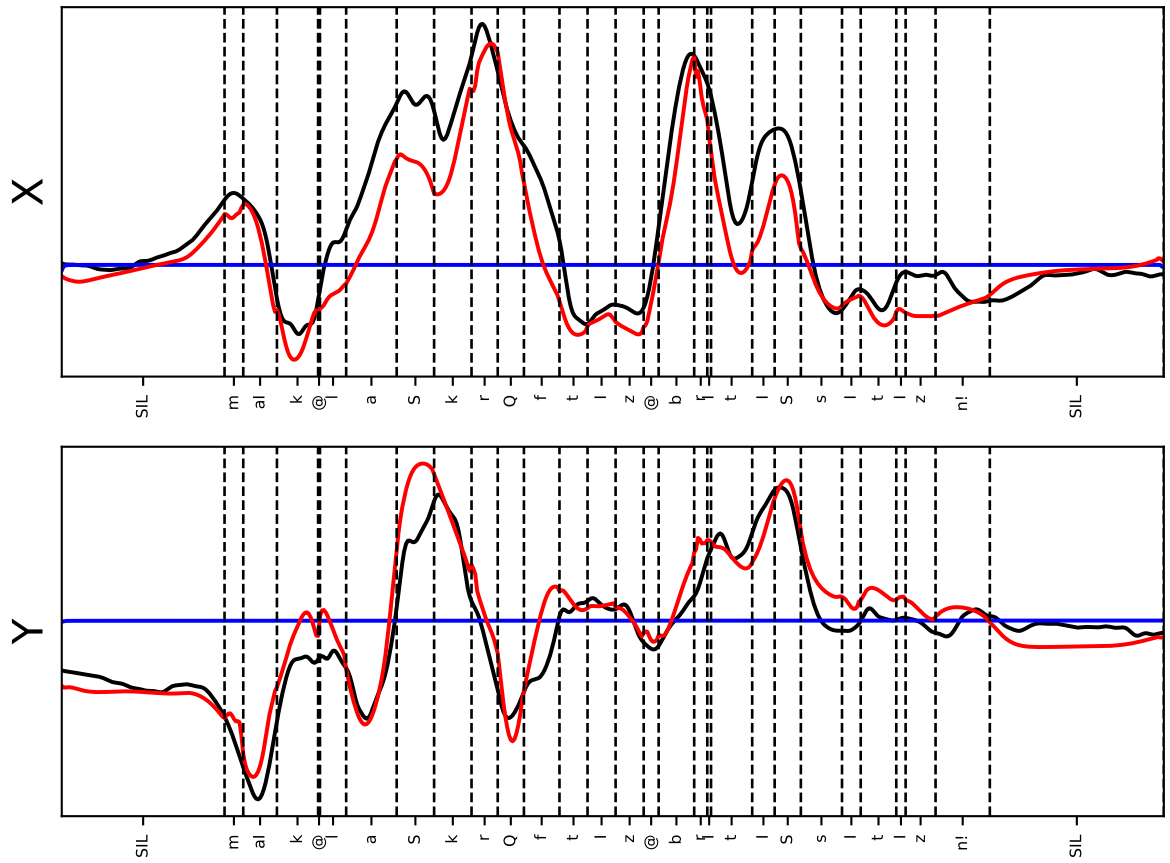


FIGURE 4.9 – Exemple de prédiction du capteur à la pointe de la langue pour le corpus anglais. La trajectoire en noir correspond à la vérité terrain, et la segmentation phonétique est marquée en pointillé. En rouge, les trajectoires prédites par le réseau le plus performant vis-à-vis de l'ensemble de validation (époque 44), et en bleu, les trajectoires prédites lors de la stagnation de la fonction de coût (époque 5)

tissage est l'absence d'utilisation de minibatch, que nous justifions par la faible quantité de données disponible dans MNGU0. En effet, plusieurs études pointent vers le fait que de petites tailles de minibatch aboutissent à une meilleure performance des réseaux (Wilson and Martinez, 2003; Masters and Luschi, 2018) (cf. sections 2.2.2 pour une présentation des minibatch). Dans un second temps, nous avons également intégré à notre procédure d'apprentissage une stratégie de *learning decay*, qui consiste à diminuer la valeur du pas d'apprentissage λ en fonction de l'évolution de l'erreur moyenne. Cette méthode permet à l'algorithme d'optimisation de plus finement ajuster les paramètres du modèle lorsque ce dernier est proche d'une solution acceptable, en effectuant de plus petits déplacements dans l'espace de recherche. En pratique, nous divisons par deux le pas d'apprentissage λ si la performance du réseau vis-à-vis de l'ensemble de validation ne connaît pas d'amélioration pendant 5 époques consécutives (*learning decay*), et nous arrêtons l'apprentissage si nous ne constatons pas d'amélioration pendant 10 époques consécutives (*early stopping*) afin de sauvegarder du temps de calcul.

Notre première expérience dans ce contexte fut une comparaison entre les architectures LSTM et GRU. Ici encore, nous avons choisi un réseau possédant 2 couches récurrentes bidirectionnelles de 128 unités dans chaque direction. La figure 4.10 nous indique clairement de très bonnes RMSE et corrélation pour chaque capteur. Avec une RMSE moyenne de 0,6mm, les dynamiques de la mâchoire (LI) et les lèvres (UL, LL) sont particulièrement réalistes, comparable à l'état de l'art dans l'inversion acoustique. La lèvre supérieure (UL) aux dynamiques plus stables a même atteint l'excellente RMSE moyenne de $0,4$mm. La langue a quant à elle une RMSE médiane aux alentours de $1,2$mm. Les performances de GRU sont légèrement supérieures à celle de LSTM : la différence de RMSE entre les deux architectures est de $0,007$mm pour la médiane, et de $0,01$mm pour les extrêmes.

Notre seconde expérience pour cette procédure d'apprentissage spécifique fut une exploration de l'influence du nombre de couches et d'unités par couche du réseau. Nous avons testé une profondeur allant d'une à quatre couches, pour un nombre d'unités par couche allant de 32 à 256 dans chaque direction. Une première observation en regardant la figure 4.11 est la bonne performance du couple de paramètres précédents (2 couches, 128 unités par couches et direction) en comparaison avec les autres jeux de paramètres. Ce résultat renforce notre conviction en la similarité de la modélisation de la coarticulation et de l'inversion acoustique, et en effet la complexité de ces deux tâches semble être relativement équivalente pour un BRNN. Deuxièment, le modèle ne possédant qu'une couche obtient des performances mitigées à côté des autres réseaux, certainement car le faible nombre de paramètres ainsi que le manque de profondeur ne fournit pas aux réseaux une capacité de modélisation suffisante à l'apprentissage de la relation complexe entre une séquence

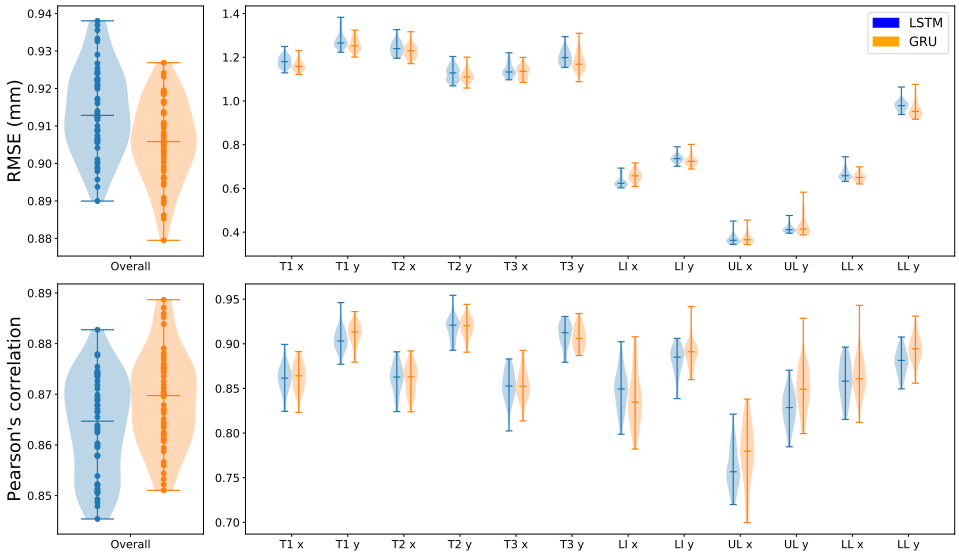


FIGURE 4.10 – Diagramme en violon de la RMSE (en mm) et de la corrélation de 50 réseaux de neurones entrainés indépendamment. Les diagrammes de gauche correspondent aux performances globales, alors que ceux de droite nous donnent les détails des performances par dimension. Chaque point représenté au sein des diagrammes en violon correspond à un apprentissage spécifique.

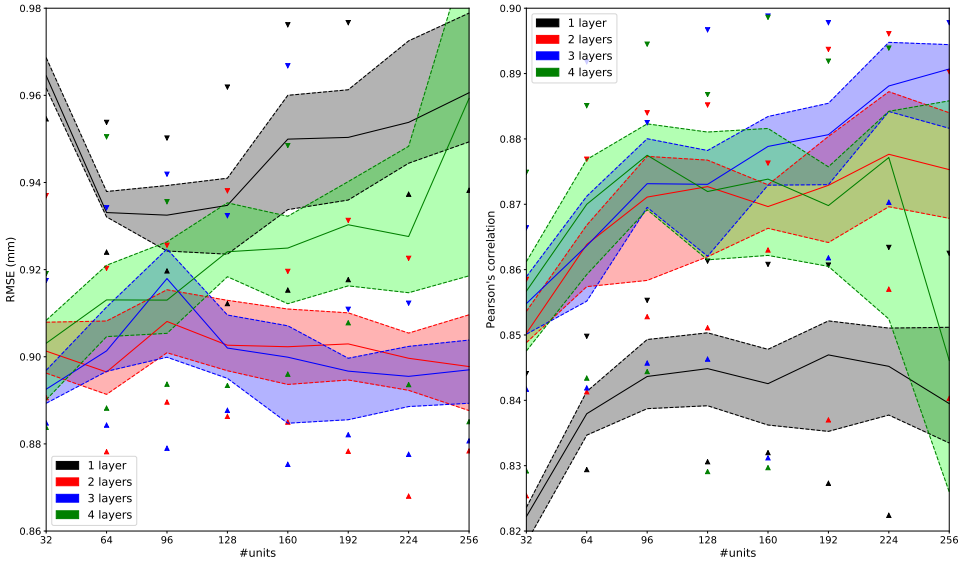


FIGURE 4.11 – Performances globales des réseaux GRU en fonction du nombre de couches et de la taille des couches. Chaque réseau a été entrainé 20 fois indépendamment, la performance médiane pour chaque configuration est une ligne pleine, les lignes en pointillés indiquent le premier et troisième quartile, et les triangles indiquent quant à eux les extrêmes. Une semi-transparence de la même couleur que les lignes aide à la visualisation de la distribution des performances de chaque configuration.

phonétique et la dynamique de ces articulateurs. Le modèle possédant 4 couches est lui aussi légèrement en dessous des modèles à 2 et 3 couches, particulièrement lorsque nous augmentons le nombre d'unités par couche. Ceci s'explique par l'augmentation importante du nombre de paramètres du modèle, rendant l'entraînement du réseau très compliqué par rapport à la quantité de données présentes dans le corpus d'apprentissage. Il est aussi à envisager que notre procédure d'apprentissage ne soit pas adaptée à un modèle si profond, en particulier en présence de si peu de données.

4.4 De la modalité visuel à la modalité articulatoire

Nous exploitons ici les techniques de l'apprentissage par transfert pour tenter d'améliorer nos performances vis-à-vis de la prédiction de mouvements articulatoires en allemand. Cette approche est motivée par la nature identique des données d'entrées du réseau pour les deux corpus (phonèmes allemands et leurs durées respectives), par la présence d'informations communes dans les données audiovisuelles et articulatoires (la dynamique de la mâchoire et des lèvres), ainsi que par le fait que ces modalités ne sont que deux représentations partielles du même phénomène (l'articulation). Nous estimons donc pouvoir réutiliser les caractéristiques apprises quant à la dynamique de la modalité visuelle, ainsi que des caractéristiques phonologiques apprises depuis les séquences phonétiques du corpus audiovisuel, pour un apprentissage plus efficace de la modalité articulatoire. Plus particulièrement, nous exploiterons les caractéristiques acquises au niveau des couches récurrentes.

Concrètement, notre procédure d'apprentissage par transfert consiste à conserver les couches récurrentes du réseau aux meilleures performances sur la base audiovisuelle, et de remplacer les couches de sorties par de nouvelles. Si nous reprenons les équations 2.2 où 2.8, notre procédure consiste donc à conserver les fonctions de transfert des biRNNs ($\overleftarrow{\mathcal{H}}$ et $\overrightarrow{\mathcal{H}}$) depuis l'apprentissage sur la base audiovisuelle, à remplacer nos couches de sorties par de nouvelles couches adaptées à la base articulatoire (W_{output}, b_{output} et $f_{decoder}$ si notre procédure d'injection de connaissance est utilisée). Nous avons expérimenté autour de l'utilisation des fonctions $\mathcal{H}$ issue des meilleurs réseaux *baseline* (sans d'initialisation) et *PCA* (avec initialisation) de la figure 4.5. Cette procédure est récapitulée pour nos quatre variantes à la figure 4.12, où les rectangles jaunes correspondent aux couches initialisées aléatoirement, les verts aux couches initialisées avec notre procédure d'injection de connaissance, et les bleus aux couches transférées depuis l'apprentissage de la modalité visuelle. Les variantes 1 et 2 profitent d'un transfert depuis les couches récurrentes du modèle sans injection de connaissances articulatoires, et les variantes 3 et 4 réutilisent les

couches récurrentes du modèle avec injection de connaissances.

La figure 4.13 nous récapitule les différences de performance pour 10 apprentissages indépendants de chaque architecture : sans transfert, avec transfert depuis un modèle sans procédure d'initialisation, et avec transfert depuis un modèle avec procédure d'initialisation. Pour ces trois architectures, nous avons essayé deux variantes : avec et sans procédure d'initialisation depuis les composantes principales. Sans procédure d'initialisation, les RMSE finales varient de 1,35 à 1,24 mm pour le modèle sans transfert, de 1,31 à 1,25 mm pour un transfert depuis la baseline audiovisuelle, et de 1,23 à 1,20 mm pour le transfert depuis le modèle audiovisuel avec initialisation. Si nous utilisons conjointement notre procédure d'initialisation basée sur l'ACP, nous obtenons des RMSE oscillants entre 1,16 et 1,12 mm pour le modèle sans transfert, entre 1,20 et 1,16 mm pour le transfert depuis la baseline audiovisuelle, et entre 1,16 et 1,13 mm pour le transfert depuis le modèle audiovisuel avec initialisation.

Une première remarque est la très faible modification des performances finales lors d'un transfert des couches récurrentes depuis un modèle sans procédure d'initialisation. Par rapport à notre baseline sans transfert, le gain est quasi-inexistant lors de ce transfert, avec deux médianes aux alentours de 1,3 mm de RMSE. Les performances tendent à stagner si nous utilisons ce transfert en conjonction avec notre procédure d'initialisation avec une RMSE médiane augmentant d'environ 0,05 mm. Cependant, cette tendance change totalement lors d'un transfert de couches récurrentes depuis un modèle entraîné avec notre procédure d'initialisation. Dans ce cas de figure, les performances s'améliorent sans procédure d'initialisation (1,22 contre 1,3 mm de résultat médian) et restent stables avec une procédure d'initialisation.

À côté des résultats vis-à-vis de la performance, nous pouvons aussi étudier la vitesse d'apprentissage des modèles. La figure 4.14 nous indique l'évolution de l'erreur moyenne sur l'ensemble d'apprentissage et de validation. Sans procédure d'initialisation spécifique au corpus lingual allemand, nous passons de 138 époques d'apprentissage pour obtenir la plus faible erreur sur l'ensemble de validation, à 64 époques pour le cas d'un transfert depuis la *baseline* visuelle sans initialisation, et à 46 époques pour le cas d'un transfert depuis le modèle avec initialisation. Si nous utilisons notre procédure d'initialisation, nous n'avons besoin que de 16 époques d'apprentissage pour notre *baseline*, de 12 pour le modèle avec transfert depuis la *baseline* visuelle, et tombons jusqu'à 7 époques pour le modèle avec transfert depuis le modèle initialisé avec les composantes principales du domaine visuel allemand.

Que cela soit en termes de performances ou de vitesse d'apprentissage, un transfert depuis un modèle utilisant notre procédure d'initialisation semble plus efficace qu'un trans-

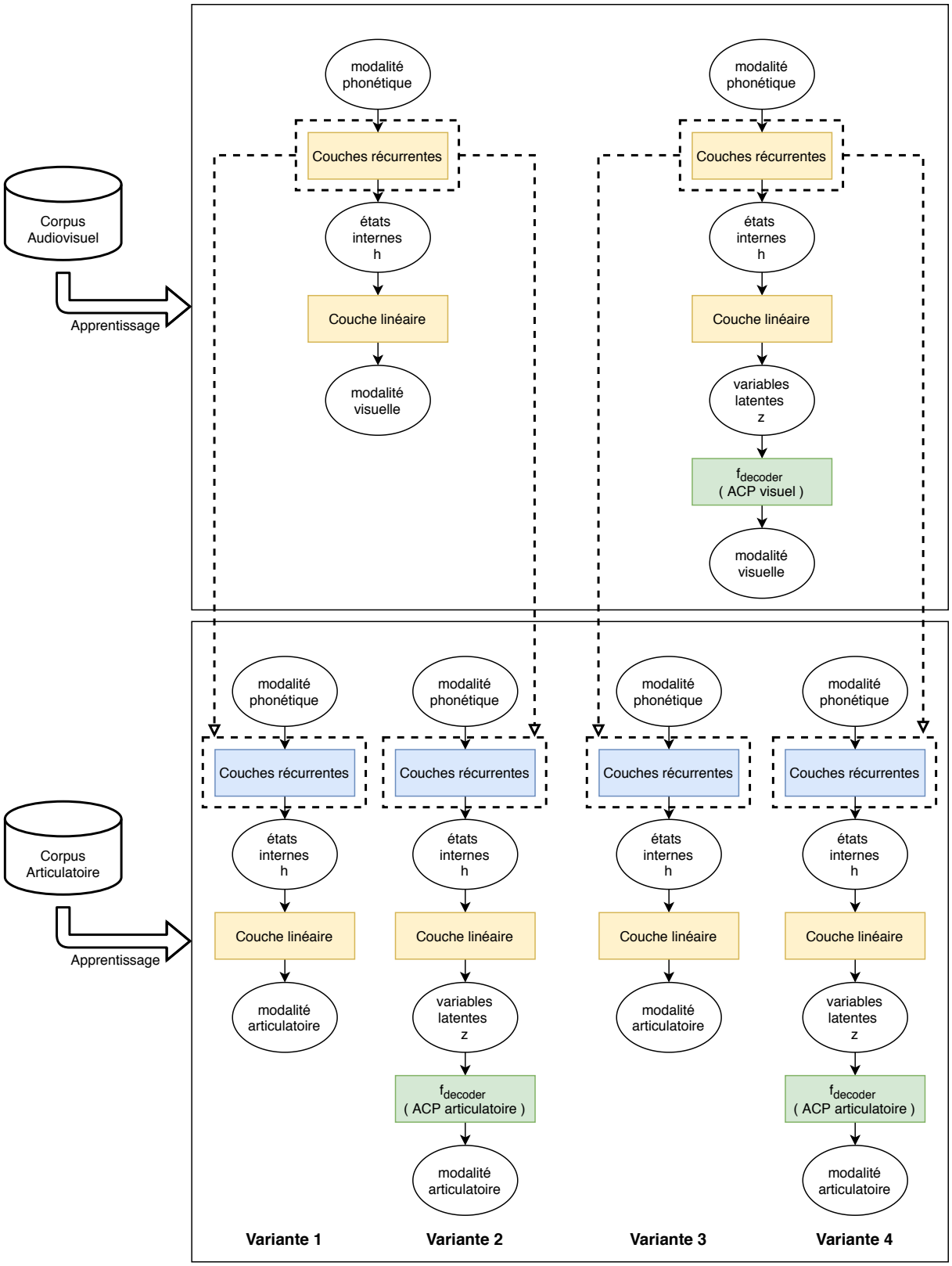


FIGURE 4.12 – Les quatre variations utilisées pour notre expérimentation sur l'apprentissage par transfert. Nous retrouvons en jaune les couches du réseau initialisées aléatoirement, en vert celles initialisées par notre méthode d'injection de connaissances, et en bleu celles initialisée par un transfert depuis la modalité visuelle.

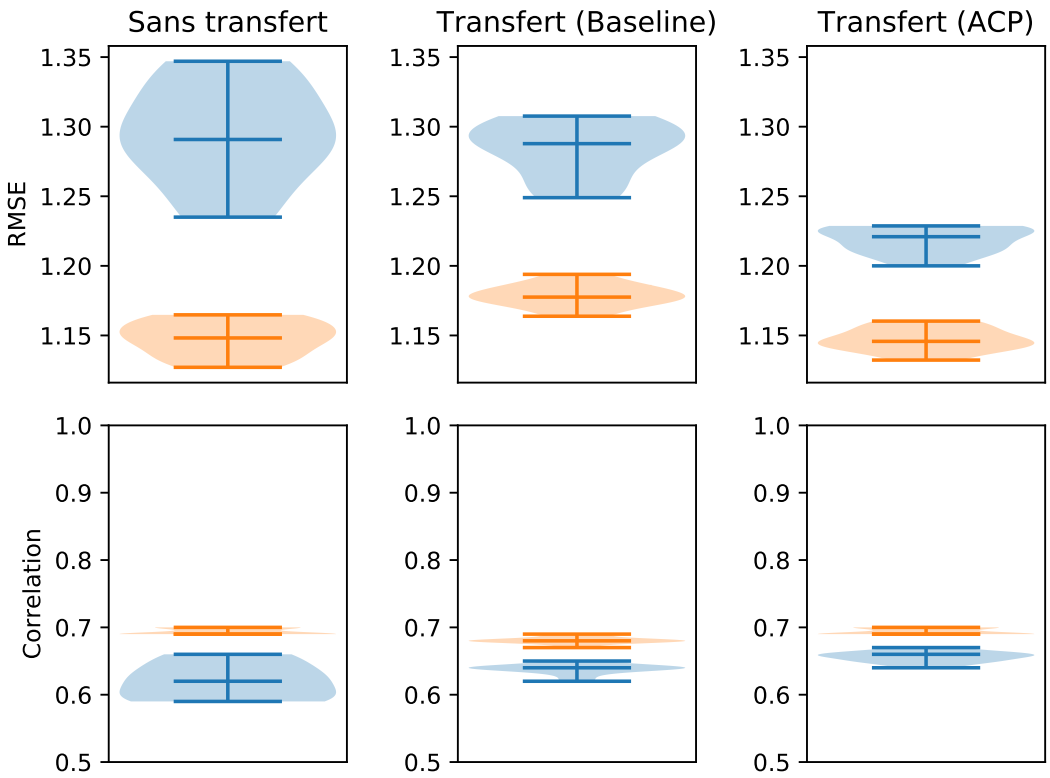
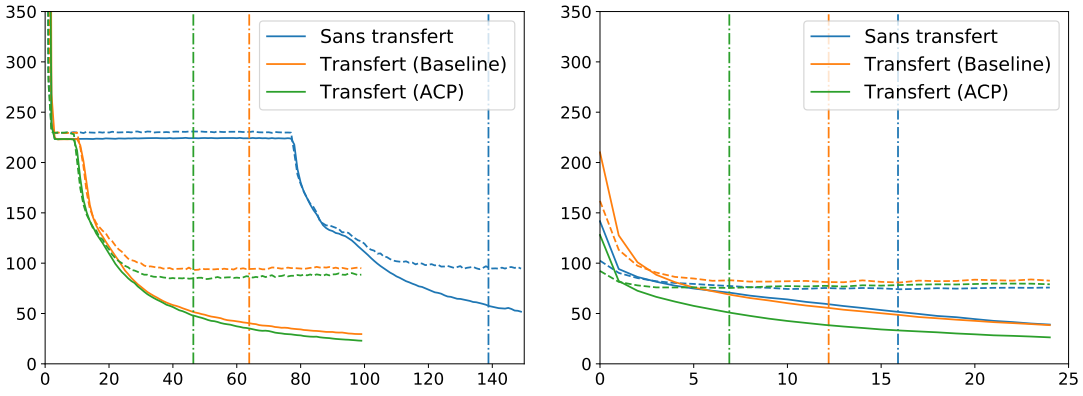


FIGURE 4.13 – Diagramme en violon des performances des architectures lors de l'apprentissage de la coarticulation linguale allemande. En bleu, sans stratégie d'initialisation, en orange, avec stratégie d'initialisation basée sur l'ACP.



(a) Sans initialisation (b) Avec initialisation basée sur l'ACP

FIGURE 4.14 – Erreur moyenne de 10 apprentissages indépendants durant un apprentissage avec transfert depuis la modalité visuelle.

fert depuis un modèle *baseline* sans initialisation. Dans ce dernier cas de figure, la fonction $\mathcal{H}$ du réseau doit apprendre des *features* capables de représenter efficacement l'espace articulatoire, car la prédiction est directement obtenue depuis l'état interne grâce à une transformation linéaire. A contrario, si nous utilisons notre procédure, l'espace articulatoire est "condensé" dans $f_{decoder}$, permettant à $\mathcal{H}$ de se concentrer sur l'apprentissage de caractéristiques propres à la coarticulation, permettant la modélisation du *mapping* entre espace phonétique et espace articulatoire. Les *features* ainsi apprises au sein de $\mathcal{H}$ semblent plus transférables d'un locuteur à l'autre, et d'un domaine articulatoire à l'autre (du labial au lingual dans notre cas). Ceci semble par ailleurs confirmé lors de l'utilisation de notre procédure d'initialisation avec un espace latent propre aux données linguales en allemand. Dans cette configuration, l'utilisation du *transfer learning* a de nouveau un impacte plus important sur la vitesse d'apprentissage du modèle dans le cas où le transfert s'effectue depuis un réseau initialisé avec l'ACP. La réutilisation de $\mathcal{H}$ n'est pas plus utile qu'une initialisation aléatoire si cette fonction contient des informations liées à un espace articulatoire spécifique, et peut même aboutir à une très légère diminution des performances, lorsqu'utilisée en conjonction avec notre procédure d'initialisation.

4.5 Discussions

La modélisation de la coarticulation est loin d'être triviale, même lors de l'utilisation de modèle aussi puissant que les réseaux de neurones. Durant l'apprentissage, notre modèle est confronté à une (voire deux) phase de stagnation de l'erreur vis-à-vis de l'ensemble d'apprentissage et de validation. Durant ces phases de stagnation, la prédiction correspondante est sujette à de forts phénomènes d'*undershooting* se traduisant par des mouvements articulatoires limités, comme une très faible ouverture de la bouche. Cette difficulté semble être contournée par notre procédure d'initialisation, qui consiste à préentrainer la dernière couche du réseau via une méthode de réduction de la dimensionnalité. D'un certain point de vue, notre procédure consiste à injecter dans le réseau une connaissance préalable de l'articulation, en lui fournissant un espace latent capable de représenter efficacement les différents mouvements des articulateurs considérés.

Un parallèle intéressant peut-être fait avec l'apprentissage de la parole chez l'humain. En effet, nos premiers vocalisations et babillages ont certainement pour objectif la découverte de nos articulateurs, ainsi que les conséquences acoustiques résultant de leurs utilisations (Davis and MacNeilage, 1995). Cette étape du développement de l'enfant semble donc être les fondations de l'acquisition de la parole. D'une manière similaire, le réseau de neurones peut "apprendre" l'articulation après un certain nombre d'époques,

qui peut être comparé au temps nécessaire à l'enfant pour apprendre à prononcer les sons de la parole et à correctement les lier entre eux. Ce temps, ou plutôt ces époques, peuvent être vus comme une phase d'exploration de la relation entre l'espace articulatoire et la représentation phonétique de la parole. De toute évidence, les mécanismes d'apprentissage chez l'humain ne sont pas du tout similaires. Toutefois, nous pouvons voir notre procédure comme une injection de connaissances préalables sur l'espace articulatoire, l'équivalent de la phase de découverte de ses articulateurs par l'enfant, nécessaire au bon apprentissage de la relation entre phonétique et articulatoire par notre modèle. Notons que dans le cas où nous injectons des connaissances par des techniques statistiques de réduction de dimensionnalité, nous réduisons le temps d'apprentissage nécessaires pour obtenir les mêmes performances sans ces connaissances. Cependant, nous ne pouvons pas supposer que ces connaissances sont de la même nature que celles apprises par le réseau.

D'autre part, l'influence positive de l'apprentissage par transfert sur la vitesse d'apprentissage, voire sur les performances, permet de confirmer l'existence d'un mécanisme commun aux deux modalités, articulatoire et visuelle, qui semble tout à fait exploitable par une modélisation par les réseaux de neurones. En revanche, nous ne savons pas si cette amélioration provient de l'utilisation de mécanisme articulatoire commun (p. ex. la mâchoire influence à la fois la langue et les lèvres) ou celle des caractéristiques phonologiques (p. ex. les phonèmes s'influencent d'une manière similaire pour les deux modalités).

4.6 Conclusion

Dans ce chapitre, nous avons démontré en quoi notre modèle à base de réseau de neurones récurrents est capable d'apprendre avec une grande précision la coarticulation labiale, linguale et mandibulaire, et ce pour trois langues différentes. En utilisant une simple représentation de la séquence phonétique, nous avons obtenu des performances similaires à l'état de l'art dans l'inversion acoustique (pour MNGU0 : 0,868 mm de RMSE moyenne pour notre meilleur réseau, contre 0,816 mm pour l'inversion articulatoire Liu et al. (2015)), et ce sans la richesse des informations contenues dans le signal acoustique. Depuis des labels phonétiques arbitraires, notre modèle semble capable de définir la place et le mode d'articulation de chaque phonème, ainsi que de prendre en compte les phénomènes de coarticulation.

L'utilisation de notre procédure d'initialisation fournie au réseau des connaissances articulatoires l'aidant grandement à apprendre la relation entre phonème et trajectoire articulatoire, lui permet de modéliser beaucoup plus rapidement les phénomènes de coarticulation, d'avoir une plus petite variabilité des résultats par rapport à l'apprentissage,

ainsi que d'atteindre de nouvelles performances. Nous pouvons cependant nuancer grandement ces résultats : l'amélioration des performances reste relativement faible, avec un gain moyen sur la RMSE globale avoisinant les 0,1 mm (une amélioration relative aux alentours des 10%), et semble avoir un impact proportionnel à la complexité de l'espace articulatoire.

Une remarque importante peut également être effectuée quant aux données. En effet, l'utilisation de matériels basés sur la présence de capteurs collés au locuteur entraine de grandes difficultés à enregistrer des données sur plus d'une session, limitant de facto le volume de données disponibles. Malgré une procédure permettant de minimiser l'erreur de placement (cf. section 3.3.1), l'existence de multiples distributions perturbe nos mesures objectives des performances, aboutissant à des RMSE globales bien plus élevées pour la base audiovisuelle française que pour son homologue allemande, et ce malgré un temps de parole considérablement plus grand (approximativement 4h contre 1h30). Ces résultats tendent à nous faire penser que la richesse phonétique et la qualité d'un corpus est toutes aussi, et peut-être plus, important que la taille de ce dernier pour l'apprentissage de la coarticulation. Cette richesse nécessite tout de même un certain volume de donnée pour pouvoir s'exprimer, certainement en fonction du langage, comme en témoigne les résultats mitigés sur le corpus articulatoire allemand, notre plus courte base de données.

Finalement, nos expériences autour de l'apprentissage par transfert nous montrent que notre procédure d'initialisation permet l'apprentissage de caractéristiques de plus haut niveau dans les couches récurrentes, plus indépendantes du locuteur et de l'espace articulatoire, et donc plus efficacement transférable. En effet, les spécificités de l'espace articulatoire d'un locuteur sont majoritairement capturées par la méthode de réduction de la dimensionnalité, permettant aux couches récurrentes de se focaliser sur la modélisation de la relation entre phonèmes et des gestes articulatoires communs à tous les locuteurs.

Chapitre 5

Analyse du modèle de coarticulation

Si le chapitre 4 nous a démontré la capacité de notre modèle à base de réseaux récurrents à apprendre à modéliser la coarticulation, la qualité finale de nos prédictions est loin d'être convenablement évaluée. En effet, l'utilisation d'une métrique "globale" comme la RMSE ou la corrélation ne peut rendre compte avec finesse de la trajectoire prédite pour les différents articulateurs. Pour combler ces lacunes, nous proposons dans ce chapitre une étude avancée de notre modèle en deux grands axes.

Le premier axe d'étude est une évaluation fine des performances de nos modèles, afin de mieux appréhender la nature des erreurs de prédiction commises. En particulier, il semble intéressant d'analyser où se situe l'erreur d'un point de vue spatial, c'est-à-dire en analysant la RMSE en fonction des différentes régions du visage et de la langue. Comme nous représentons ces modalités sous forme d'un nuage de points, il est trivial de calculer cette RMSE pour chaque point du nuage. Un second point d'intérêts est d'assurer que le modèle est capable d'atteindre certaines cibles critiques pour l'intelligibilité de la parole, comme la fermeture des lèvres lors de la production de bilabiales (/p/, /b/, /m/).

Un deuxième axe d'étude est l'analyse de notre réseau après apprentissage, l'objectif principal de cette exploration étant de comprendre ce que notre modèle a appris de la coarticulation. Deux directions nous semblent envisageables pour cela : une analyse des poids du réseau après-apprentissage, et une analyse des trajectoires prédites en fonction de la séquence phonétique. Pour les poids du réseau, nous nous attarderons principalement sur les poids de la couche initialisée à l'aide de notre méthode afin de mesurer les modifications faites à $f_{decoder}$, mais également sur les poids de la première couche du réseau, qui transforme les vecteurs *one-hot* représentant les phonèmes en un vecteur réel de plus haute dimension, et servant donc implicitement de couche d'*embedding*. Nous hypothétisons que l'analyse de cet *embedding* puisse révéler des similitudes avec les connaissances

phonétiques, comme le regroupement en cluster de phonèmes ayant un impact similaire sur l'articulation. Finalement, nous proposerons une simple méthode pour interpréter l'impact de la coarticulation anticipative telle que modélisée par le réseau de neurones. Là encore, nous pensons pouvoir dégager des tendances du modèle pouvant être mis en relation avec ce que nous savons de la coarticulation chez l'humain.

5.1 Évaluation objective

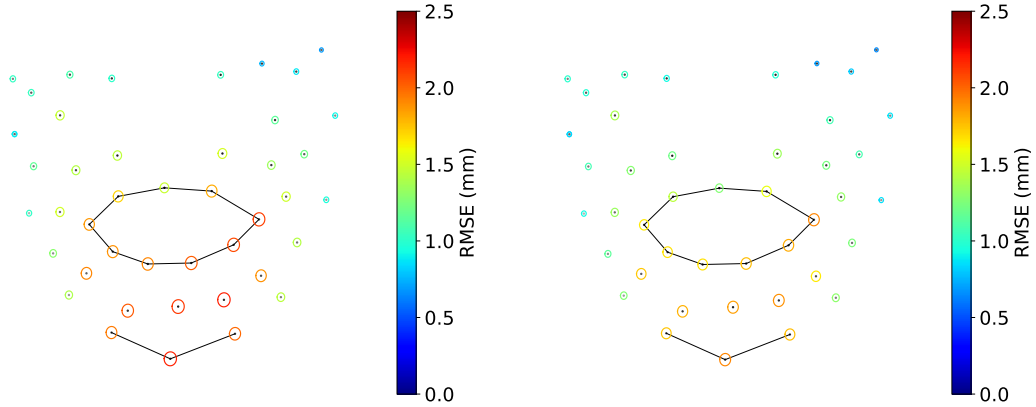
5.1.1 Localisation de l'erreur

Nous détaillons dans cette section la RMSE moyenne par capteurs pour nos corpus, en considérant la moyenne des 10 apprentissages indépendants. Nous effectuons une comparaison entre cette erreur pour un modèle ne profitant pas de la procédure d'initialisation, et un modèle initialisé avec les composantes principales. Ce travail n'est pas réalisé pour MNGU0, car la procédure d'injection de connaissances ne semble pas avoir d'impact sur ce corpus (cf section 4.3), et l'erreur par capteur a déjà été présentée à la figure 4.10.

Les figures 5.1 et 5.2 représente donc le visage de nos locuteurs français et allemand sous forme d'un nuage de points, et 5.3 présente la langue et les lèvres du deuxième locuteur allemand. Dans ces figures, la position moyenne de tous les points du nuage est représentée en trois dimensions (avec une légère transparence pour la profondeur), et un cercle est tracé autour de chaque point avec pour rayon la RMSE moyenne sur ce point. Un code couleur est aussi utilisé pour faciliter la lecture. Ce nuage de points représente donc le visage pour 5.1 et 5.2, et la langue et les lèvres pour 5.3.

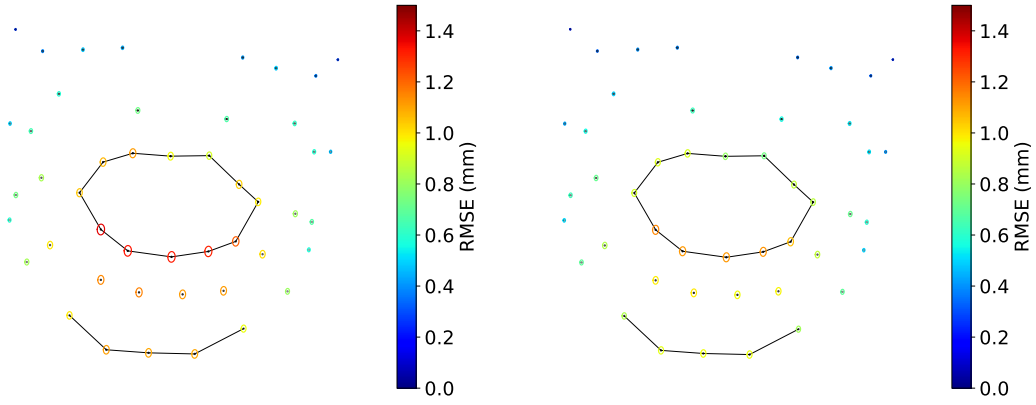
Comme attendu, l'erreur n'est pas répartie uniformément sur l'ensemble des capteurs, mais au contraire localisée sur les points ayant le plus de variances. Pour la modalité visuelle, il s'agit principalement de la zone couvrant les lèvres et le menton, et plus particulièrement la lèvre inférieure, avec une RMSE moyenne variant de 0,70 à 2,19 mm pour le français, et de 0,26 à 1,37 mm pour l'allemand. Ceci s'explique en grande partie par la nature "neutre" de la parole enregistrée : en l'absence d'émotion, très peu de mouvement est observable au niveau des joues et des pommettes. Pour la modalité articulatoire, l'erreur est la plus importante le long du dos de la langue, avec une erreur la plus importante au niveau du centre de celle-ci (de 0,18 à 2,11 mm). Nous considérons ces erreurs comme raisonnables, en particulier car 2 mm ne peut suffire à distordre le visage dans un mouvement qui ne semble pas naturel.

Nous pouvons remarquer que la procédure d'initialisation ne réduit pas non plus cette erreur uniformément, ayant au contraire une tendance à diminuer l'erreur commise sur



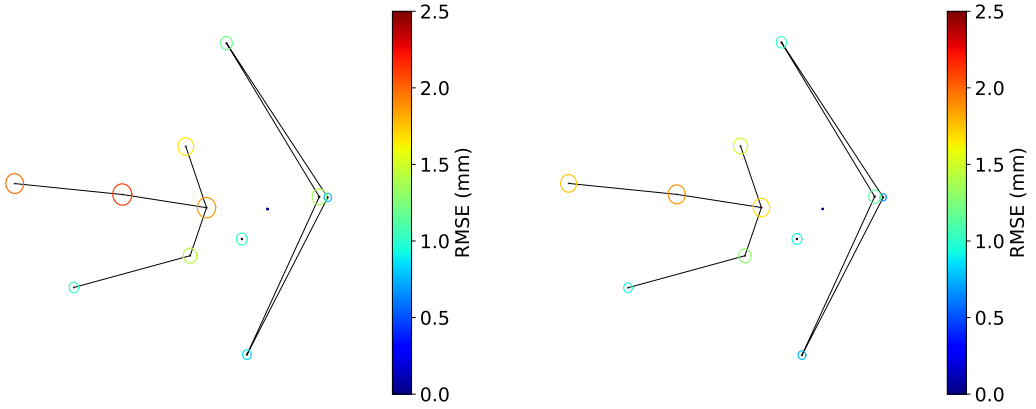
(a) Sans initialisation (b) Avec initialisation basée sur l'ACP

FIGURE 5.1 – Positions des capteurs sur le visage du locuteur du corpus Français. La couleur et la taille du cercle autour d'un marqueur indiquent l'erreur en RMSE. Les marqueurs des lèvres et du menton sont reliés pour aider à la visualisation.



(a) Sans initialisation (b) Avec initialisation basée sur l'ACP

FIGURE 5.2 – Positions des capteurs sur le visage du locuteur du corpus Allemand. La couleur et la taille du cercle autour d'un marqueur indiquent l'erreur en RMSE. Les marqueurs des lèvres et du menton sont reliés pour aider à la visualisation.



(a) Sans initialisation (b) Avec initialisation basée sur l'ACP

FIGURE 5.3 – Positions des capteurs sur la langue et les lèvres du locuteur du corpus articulatoire allemand. La couleur et la taille du cercle autour d'un marqueur indiquent l'erreur en RMSE. Vue du dessus, les marqueurs des lèvres (à droite) et de la langue (à gauche) sont reliés pour aider à la visualisation.

ces zones les plus mouvantes. Ainsi, l'erreur maximale sur un capteur passe de 2,19 à 1,92 mm pour la base française, alors que l'erreur minimale passe de 0,70 à 0,66 mm. Pour les deux corpus allemands, le même phénomène est observé : l'erreur maximale passe de 1,37 à 1,15 mm pour le visuel et de 2,11 à 1,88 mm pour l'articulatoire, alors que l'erreur minimale passe de 0,27 à 0,23 mm pour le visuel et de 0,18 à 0,16 mm pour l'articulatoire.

Une dernière note peut être effectuée vis-à-vis du corpus français, où l'erreur semble supérieure sur un des deux côtés du visage, contrairement aux résultats sur le corpus allemand, où l'erreur est sensiblement répartie symétriquement sur le visage. Nous pensons que ces résultats sont un autre témoignage direct de l'impact des différentes distributions présentes dans nos données, issues des multiples sessions d'acquisition nécessaires, comme rapporté à la section 3.3.1.

5.1.2 Cibles visuelles critiques

Dans cette section, nous examinons deux gestes articulatoires qui peuvent être considérés comme critiques pour l'intelligibilité de la parole, car étant les deux gestes les plus visibles : la fermeture de la bouche lors de la production de phonèmes bilabiaux (/p/, /b/, /m/), et la protrusion (pour /u/ et /y/ par exemple). En effet, de faibles déviations dans les trajectoires des articulateurs peuvent fortement impacter l'intelligibilité de la parole. Il a été par exemple démontré que cette fermeture est nécessaire à la bonne intelligibilité des phonèmes bilabiaux, en particulier McGurk and MacDonald (1976) nous montre en

quoi le son /ba/ associé au visuel /ga/ engendre la perception du son /da/.

Pour ce faire, nous proposons les deux tests suivants :

Test 1 Calculer le minimum de l'ouverture de la bouche pour certains phonèmes.

Test 2 Calculer le maximum de la protrusion pour certains phonèmes.

Le test 1 nous permet de détecter si le modèle a capturé la fermeture complète de la bouche durant la production de bilabiales (/b/, /p/, /m/), mais peut aussi être utilisé dans le cas des labiodentales (/f/, /v/), et le test 2 nous permet de vérifier si le réseau de neurones a correctement appris la protrusion plus ou moins prononcé pour certains phonèmes (/O/, /o/, /oh/, /y/, /u/, /eu/, /euf/, /swa/, /S/ et /Z/ pour le français, /o :/, /O/, /U/, /y/, /y :/, /S/, /C/ et /Z/ pour l'allemand). Nous avons défini l'ouverture de la bouche comme la distance euclidienne entre le capteur central de la lèvre inférieure et celui de la lèvre supérieure, et la protrusion comme la profondeur du capteur central de la lèvre supérieure. Une fois l'ouverture minimale et la protrusion maximale calculés pour tous les phonèmes considérés au sein de l'ensemble de test du corpus, nous pouvons afficher la distribution de ces valeurs sous forme de diagramme en violon (figures 5.4 et 5.5), et comparer cette distribution à celle obtenue depuis les prédictions du modèle sur l'ensemble de test. Pour la figure 5.4, plus les valeurs sont basses, et plus la bouche est fermée. Pour la figure 5.5, plus les valeurs sont hautes, et plus la protrusion des lèvres est marquée.

La figure 5.4 nous montre ainsi la distribution de l'ouverture minimale pour les bilabiales et les labiodentales françaises et allemandes. Les résultats pour le français particulièrement intéressant, dans le sens où nous retrouvons trois distributions quasiment identiques pour les consonnes bilabiales, avec des moyennes très proches, et deux distributions aux moyennes légèrement plus élevées pour les labiodentales (avec bien plus de variances pour /v/). Nous pouvons par ailleurs observer l'impact important de la procédure d'initialisation pour ce corpus. Sans cette dernière, la valeur moyenne des ouvertures (diagrammes orange) est supérieure à la moyenne des valeurs cibles, et la production de /b/, /p/ et /m/ ressemble très fortement à celle de /f/. Les phénomènes d'*undershooting* déjà abordés au chapitre 4 sont donc ici mis en lumière pour des segments critiques de la parole. Nous pouvons également apprécier la bonne correction de cette distribution lors de l'intégration de l'espace latent pré-calculé (diagramme vert). Ces résultats sont nettement moins visibles pour notre locuteur allemand, où l'apport de notre procédure d'initialisation est moindre. Le bénéfice de notre initialisation est plus mitigé du côté de l'analyse de la protrusion, où aucune tendance ne se dégage réellement, que cela soit pour l'allemand comme le français. Nous pouvons néanmoins affirmer que la variabilité de cette protrusion semble moindre pour le corpus allemand, où les distributions provenant des

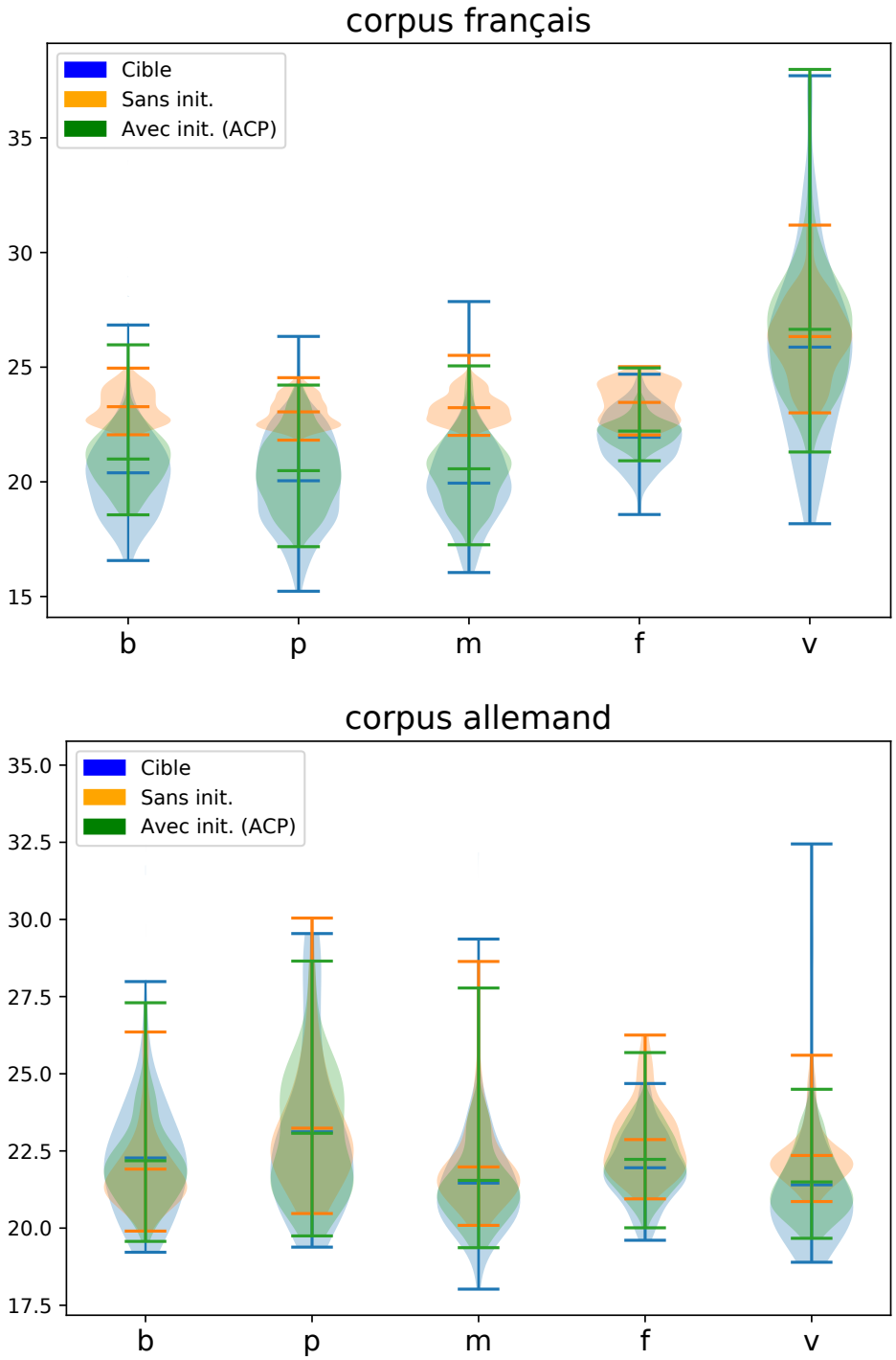


FIGURE 5.4 – Distribution de l'ouverture minimum en millimètre de la bouche pour les corpus audiovisuels lors de la production de bilabiales et de labiodentales. Cette ouverture est définie comme la distance entre les deux capteurs centraux des lèvres.

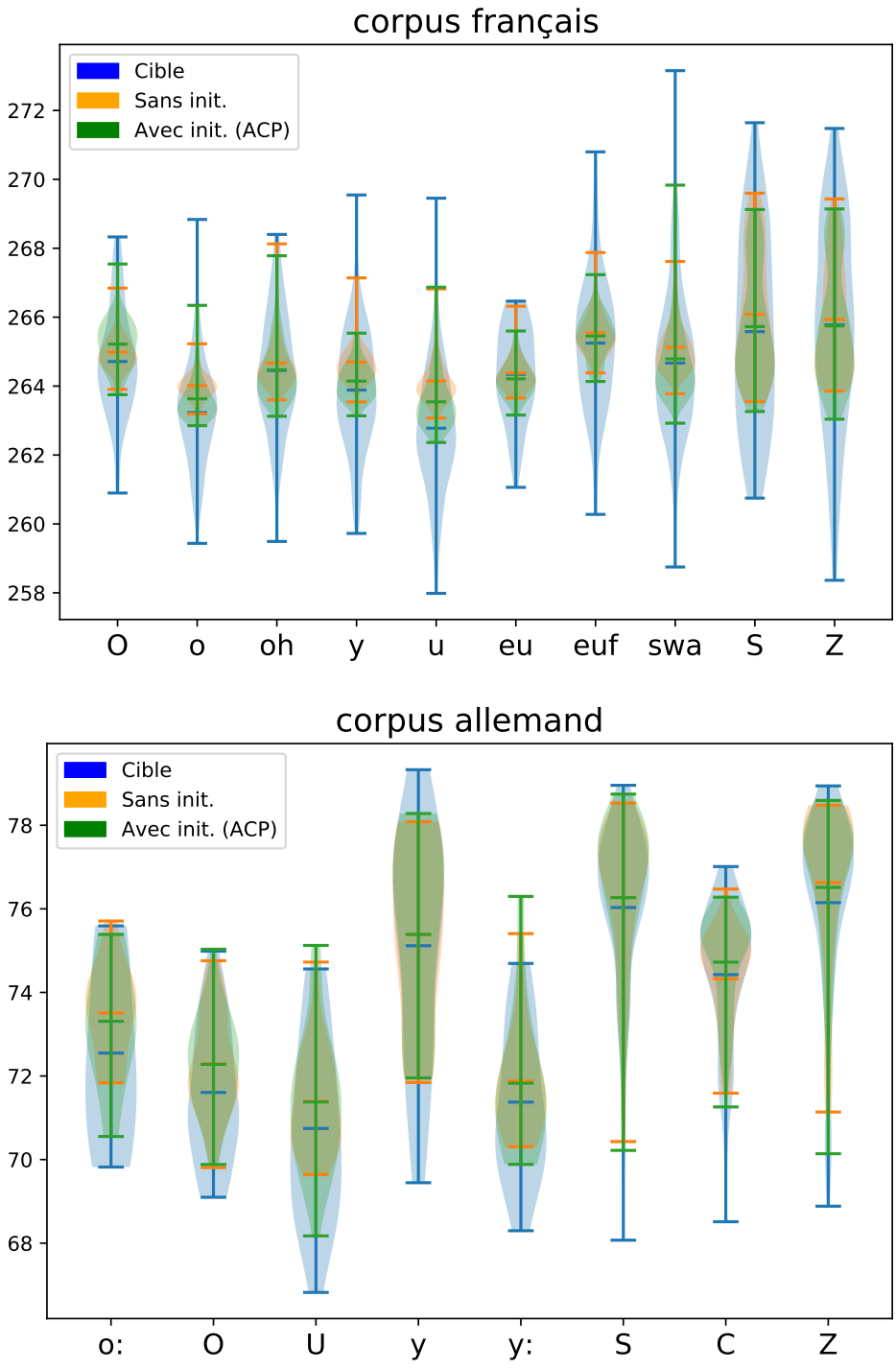


FIGURE 5.5 – Distribution de la protrusion maximale pour les corpus audiovisuels.

prédictions semblent plus proches des distributions réelles que pour le corpus français.

5.2 Étude de la coarticulation anticipative modélisée

Dans cette section, nous proposons une analyse de l'influence de la coarticulation anticipative modélisée par les réseaux de neurones récurrents, avec pour objectif la validation de cette modélisation par rapport aux connaissances phonétiques et phonologiques de ces langues. Pour ce faire, nous exploitons l'ensemble de test de nos corpus, et analysons la prédiction obtenue à partir de phrases tronquées, comme illustrée par la figure 5.6. En comparant la prédiction de référence, obtenue à l'aide de la phrase complète (en vert sur la figure), avec la prédiction des phrases tronquées (en bleu), nous pouvons calculer une différence par phonème dépendant du nombre d'unités phonétiques composant son contexte futur. Ainsi, nous obtenons un tableau qui pour un phonème et une taille de contexte futur, nous donne l'erreur commise sur ce phonème. Nous tronquons chaque phrase, phonème par phonème, en conservant un minimum de 4 unités (silence inclus), et considérons une taille de contexte futur variant de 0 à 6 phonèmes.

Ces résultats sont récapitulés à la figure 5.7 pour nos deux corpus audiovisuels. Pour les données français, les phonèmes les plus résistants à la coarticulation (à droite sur la figure) sont ceux dont l'articulation labiale joue un rôle majeur pour leur production, comme l'ensemble des voyelles nécessitant un arrondissement des lèvres (/u/, /H/, /o/, /O/, /y/ et /oh/), puis par les bilabiales et les labiodentales. A contrario, les phonèmes

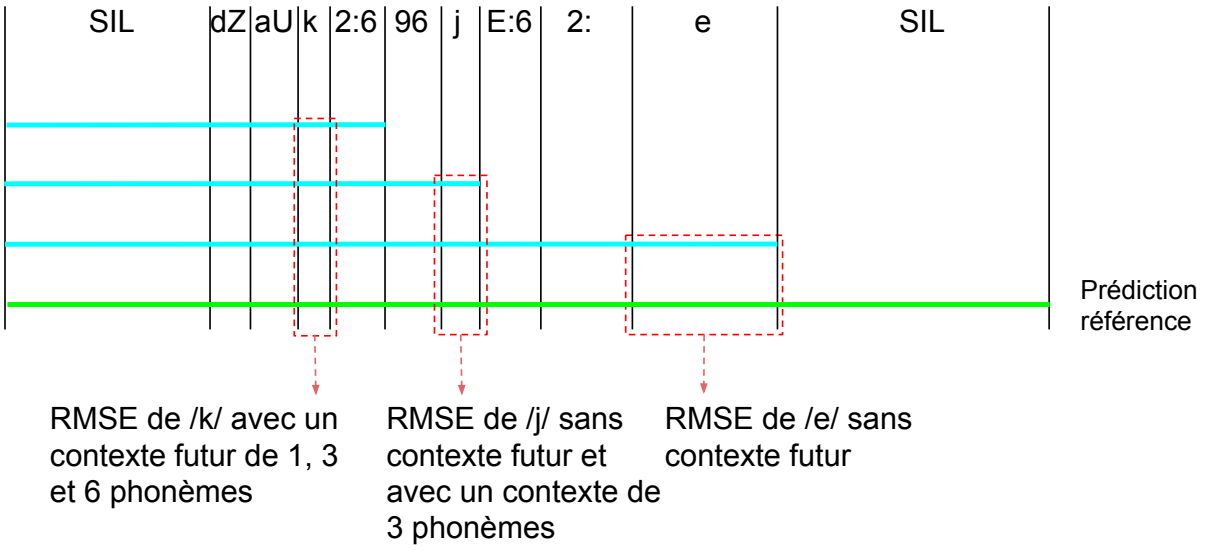


FIGURE 5.6 – Exemple de calcul de la RMSE pour différentes tailles de contexte futur.

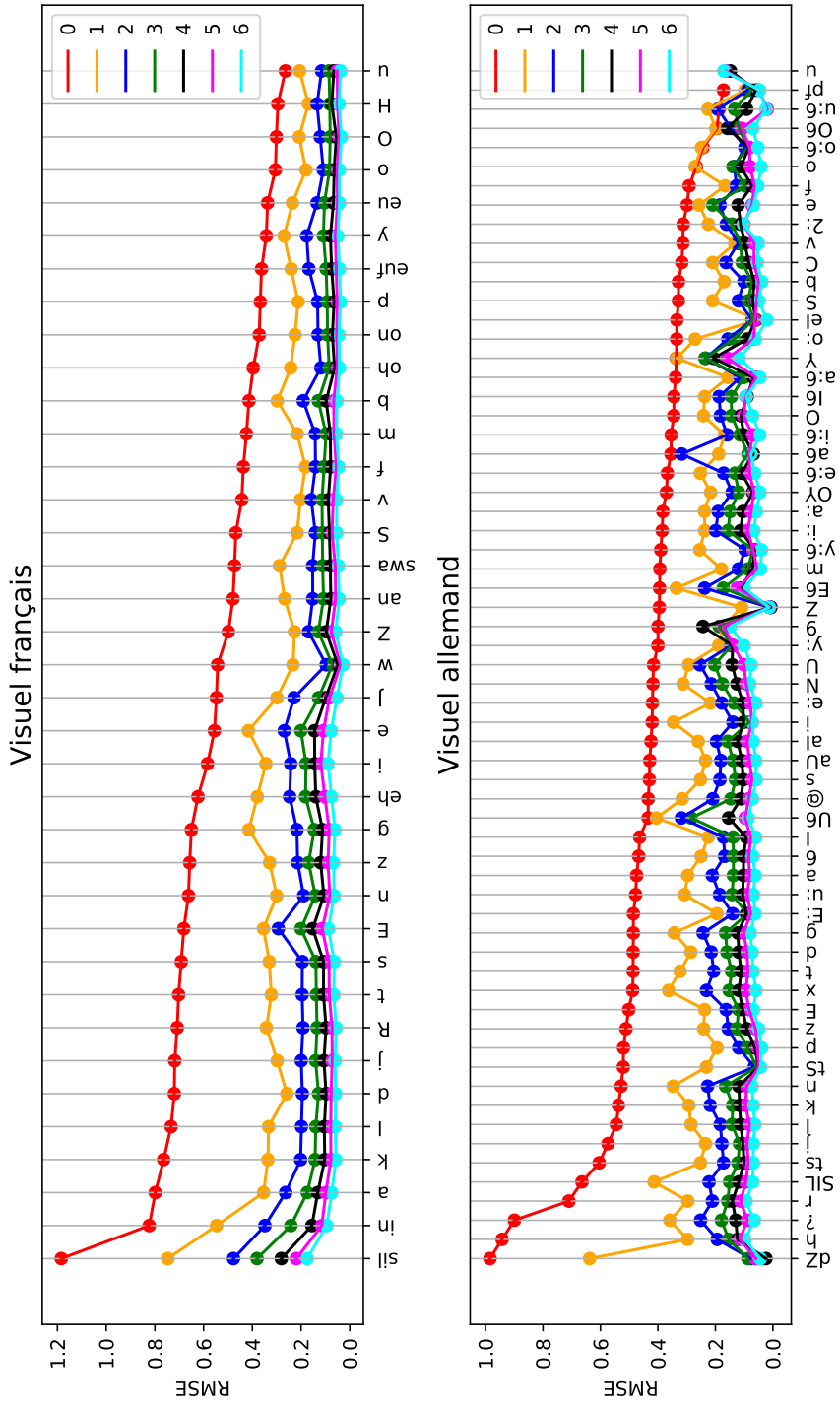


FIGURE 5.7 – Résistance des phonèmes au contexte futur. La RMSE est exprimée en millimètre, et permet de mesurer l'écart entre la prédiction avec un contexte futur complet, et la prédiction avec un contexte futur tronqué. Le nombre de phonèmes disponible dans le contexte futur est indiqué par la légende.

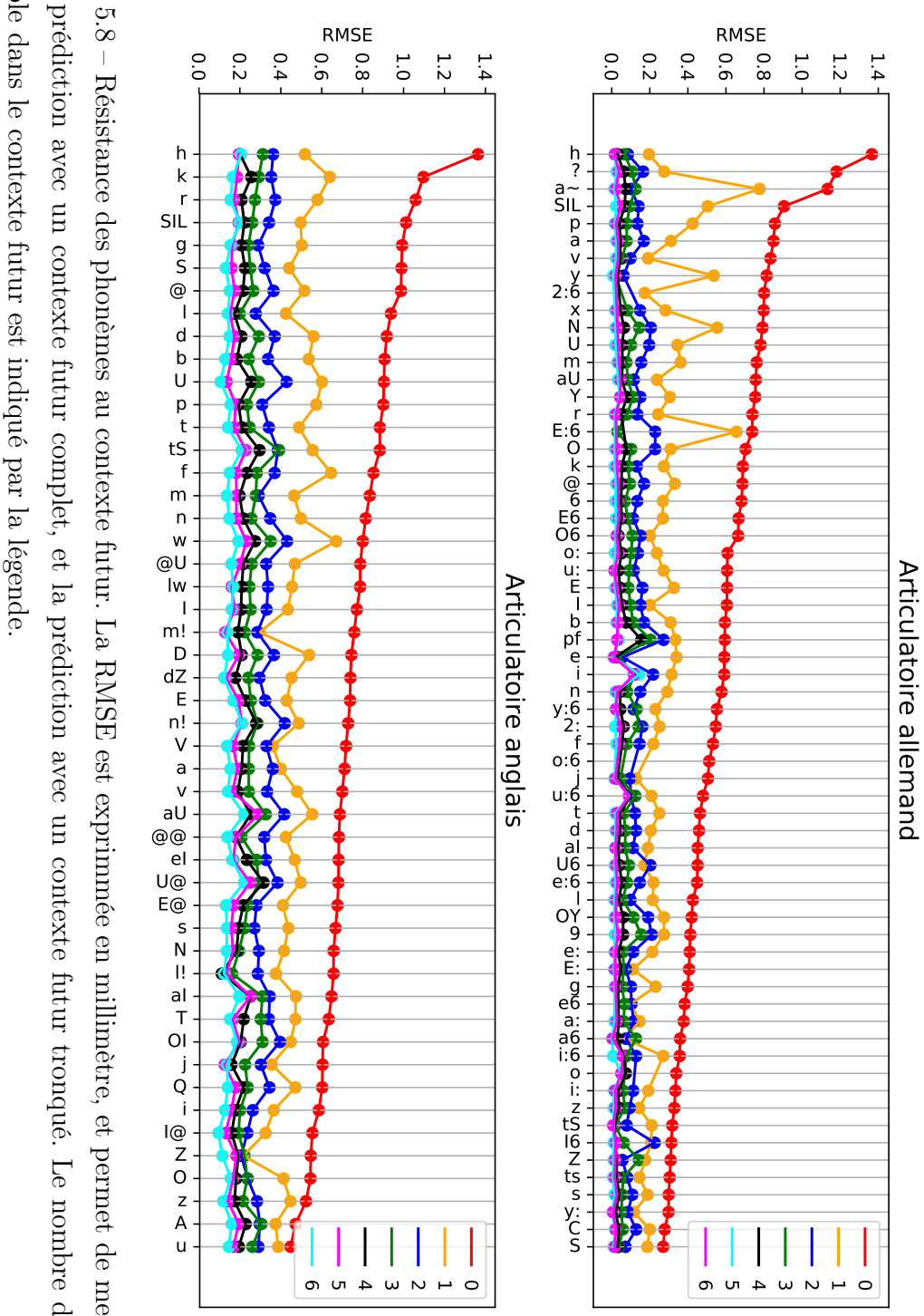


FIGURE 5.8 – Résistance des phonèmes au contexte futur. La RMSE est exprimmée en millimètre, et permet de mesurer l'écart entre la prédiction avec un contexte futur complet, et la prédiction avec un contexte futur tronqué. Le nombre de phonèmes disponible dans le contexte futur est indiqué par la légende.

les plus sensibles à la coarticulation anticipative sont ceux dont le point d'articulation est défini par la langue, principalement des consonnes. Notons le cas particulier de la voyelle /a/, dont l'ouverture de la bouche correspondante semble fortement influencé par le contexte phonétique. Des résultats similaires peuvent être observés pour le corpus visuel allemand, dans lequel nous retrouvons beaucoup de diphtongue et de voyelle nécessitant de l'arrondissement parmi les moins sensibles à l'anticipation, ainsi que les labiodentale. Similairement au français, de nombreuses consonnes dont l'articulation est principalement réalisée par la langue se retrouvent parmi les plus sensibles au contexte futur, comme /dZ/, /ts/, /l/, /k/, /n/ ou encore /z/. L'absence d'influence au niveau labiale permet donc au réseau d'obtenir implicitement Une différence notable est cependant pour le phonème /p/, qui semble bien plus sensible au contexte phonétique futur pour l'allemand que le français. Ce phénomène s'explique très certainement de par la production des plosives sourdes allemandes, présentant une expiration caractéristique, et donc une réalisation plus longue que pour le français (Jessen, 1998). Cette taille plus importante peut donc permettre une meilleure anticipation du phonème suivant au niveau labial, ce qui pourrait être une explication à ce phénomène. Bien entendu, les contextes phonétiques présents dans notre corpus de test de taille très modeste peuvent également influencer ce résultat.

Quelques similitudes peuvent aussi s'établir au niveau des résultats pour les corpus articulatoires allemand et anglais à la figure 5.8, cependant, il est bon de rappeler que notre corpus articulatoire allemand contient peu de données, et ces résultats ne sont peut-être pas représentatifs. Le /h/ est pour les deux langues le phonème le plus sensible à la coarticulation, très proche du SIL, ce qui indique que notre modèle assimile le silence à une aspiration. Les prédictions pour les occlusives /k/ et /g/ sont bien plus sensibles au contexte futur dans le cas de l'anglais, mais les bilabiales sont à peu près sur le même niveau pour les deux langues. Parmi les phonèmes les moins sensibles à la coarticulation anticipative, nous retrouvons principalement les fricatives dont la position de la constriction nécessaire à leurs articulations ne peut être grandement modifiée, sous peine d'en changer totalement la réalisation. Nous retrouvons ainsi les fricatives alvéolaires /s/ /z/ et palato-alvéolaires /Z/ /S/ pour les deux langues (ainsi que /C/ pour la langue allemande). Une différence se remarque également entre les occlusives alvéolaires /t/ /d/, bien plus sensible pour l'anglais que pour l'allemand.

5.3 Étude des paramètres du réseau après apprentissage

Malgré leurs puissantes capacités de modélisation, les réseaux de neurones souffrent d'un profond problème d'interprétabilité, et sont souvent associés à des boites noires dont les motifs de décision ne peuvent pas être facilement explicités. Dans cette section, nous tentons d'ouvrir cette boîte noire, non pas pour expliquer pourquoi le réseau aboutit à telle prédiction, mais pour tenter d'apercevoir une fraction de cette modélisation interne. Pour ce faire, nous étudions la valeur de certains paramètres du réseau après apprentissage, en nous intéressant particulièrement à deux questions. Premièrement, l'espace articulatoire latent injecté par la procédure d'injection de connaissances est-il grandement modifié durant l'apprentissage ? Et deuxièmement, quelle est la représentation que notre modèle a faite des phonèmes ?

5.3.1 Modification de l'espace articulatoire latent injecté

Afin de nous assurer pleinement de l'apport d'un espace latent pré-calculé lors de l'initialisation du réseau de neurones, nous avons étudié l'évolution de $f_{decoder}$ (cf. équation 2.8 à la section 2.3) après l'apprentissage du modèle.

Dans le cas d'une initialisation via l'ACP, $f_{decoder}(x)$ est définie comme une simple multiplication matricielle, $W.x + b$, où W est la matrice des composantes principales et b la moyenne de chaque élément de x. Afin de mesurer les changements sur cette transformation linéaire, nous pouvons prêter attention à la distance cosine moyenne entre les composantes telles que calculées par l'ACP et ces mêmes composantes après l'apprentissage de la coarticulation. À une distance cosine de 0, deux droites sont colinéaires, et à contrario deux droites orthogonales ont une distance cosine de 1. Pour l'évolution des moyennes b, nous pouvons simplement calculer la RMSE du vecteur avant et après apprentissage. La table 5.1 récapitule cette distance cosine et RMSE pour nos quatre corpus différents. Nous pouvons y remarquer la très faible RMSE entre les moyennes calculées statiquement avant l'ACP, et celle post-apprentissage, inférieur au dixième de millimètre. Pour la distance cosine des composantes, deux groupes se distinguent clairement en fonction de la modalité. Pour la modalité articulatoire, ces composantes ne sont pratiquement pas modifiées, avec une distance cosine moyenne de 0,01 pour le corpus lingual en anglais et de 0,06 pour le corpus lingual en allemand. En revanche, dans le cas de la modalité visuelle, nous obtenons une distance cosine moyenne de 0,13 pour l'allemand et de 0,25 pour le français. L'espace latent pré-calculé semble donc être utilisé tel quel par le ré-

TABLE 5.1 – Distance cosinus des composantes principales et RMSE des moyennes, avant et après apprentissage.

Corpus	Modalité	Distance cosinus des composantes W	RMSE (mm) des moyennes b
Français	visuelle	0,25	0,07
Allemand	visuelle	0,13	0,05
Allemand	articulatoire	0,06	0,05
Anglais	articulatoire	0,01	0,08

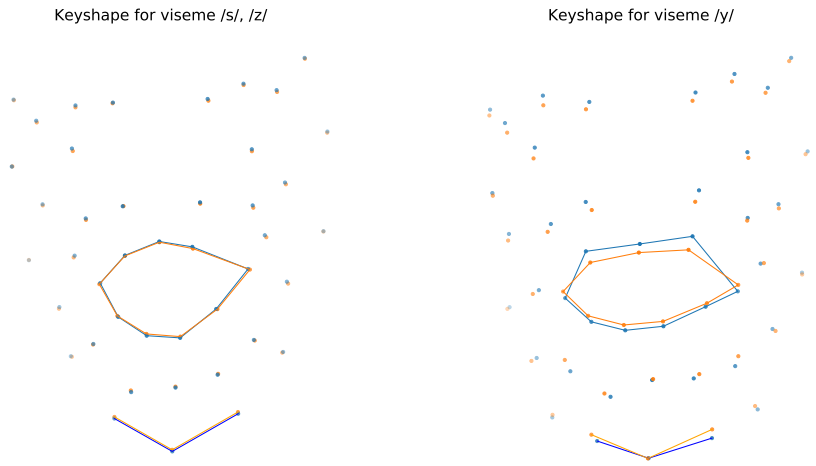


FIGURE 5.9 – Différence de *keyshapes* avant et après fine-tuning. Les points en bleu correspondent au keyshape originale, et les points en orange au keyshape après apprentissage. Les points correspondant aux lèvres et au menton ont été reliés pour faciliter la lecture.

seau de neurones dans le cas de la modalité articulatoire, et légèrement modifié pour la modalité visuelle.

Analyser la raison de cette plus grande modification de $f_{decoder}$ pour la modalité visuelle que la modalité articulatoire. Nous proposons néanmoins deux hypothèses pour ce phénomène :

— La plus grande quantité de données présente dans les corpus audiovisuel facilite la modification de l'espace latent injecté pour l'adapter à notre tâche,

— La définition du visage est bien plus détaillée (plus grand nombre de points) que la définition de la langue, entrainant une augmentation de la non-linéarité des mouvements, et donc une plus mauvaise représentation par un espace latent linéaire.

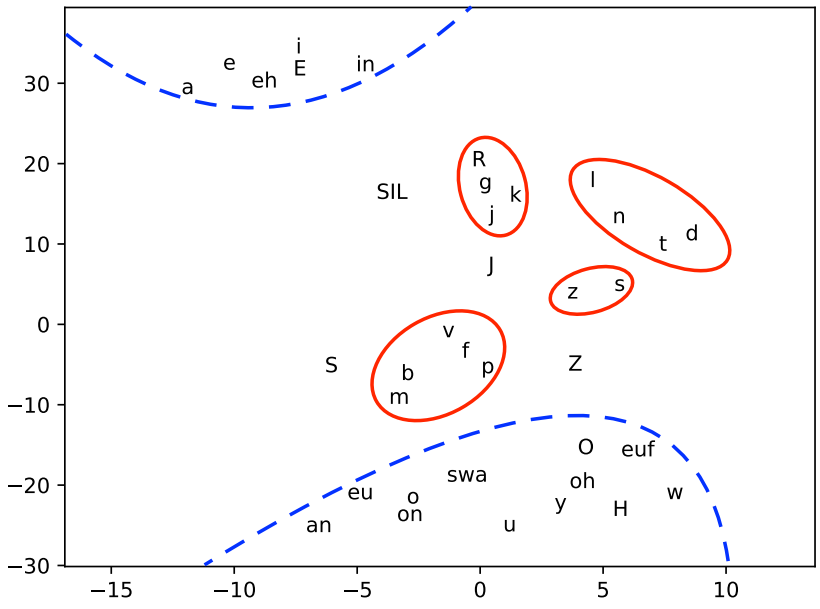
Nous pouvons également étudier la transformation des *keyshapes* par apprentissage, pour le cas de la coarticulation en français. Si nous calculons la RMSE entre les deux matrices de *keyshapes*, nous obtenons un résultat moyen de $0,76$mm pour les 10 apprentissages indépendants. La figure 5.9 met en lumière ces très faibles modifications en présentant le *keyshape* ayant subi le moins de modifications (correspondant au viseme /s/ et /z/) côte à côte avec le *keyshape* ayant subi le plus de modification (correspondant au viseme /y/). Nous pensons que ces résultats renforcent à la fois la validité de nos *keyshapes*, car ceux-ci ne sont pas modifiés outre mesure par la descente de gradient, ainsi que la bonne utilisation de ces derniers durant l'apprentissage de l'espace latent. De plus, ces légères modifications peuvent probablement améliorer le côté applicatif de la synthèse audiovisuelle en adaptant la technologie à notre modèle de coarticulation. En effet, notre modèle nous donne ici accès à des *keyshapes* légèrement raffinés, plus adapté à nos modèles de coarticulation, aboutissant à de meilleurs résultats pour l'animation de la parole, et permettant donc de guider l'animateur pour la création de ces modèles 3D.

5.3.2 Analyse de la couche d'*embedding*

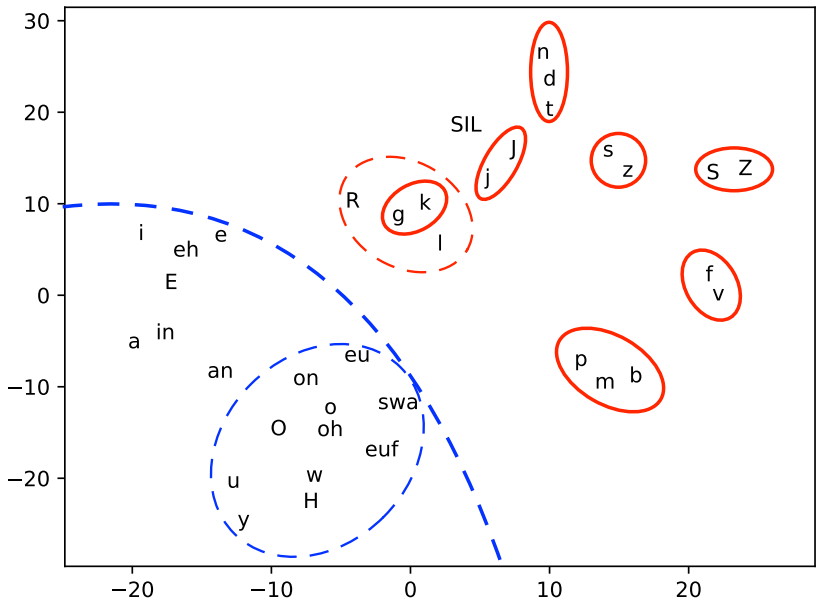
Pour les réseaux GRU (équation 2.5, cf. section 2.1), nous considérons la couche d'embedding comme étant la concaténation de toutes les matrices de poids connectant le vecteur d'entrée x_t aux différentes portes, c'est-à-dire une section de W_z, W_r et W_h. Chaque phonème étant représenté par un vecteur *one-hot*, la multiplication matricielle entre ce vecteur dont l'élément i est "chaud" et la matrice de poids va nous donner un vecteur égal à la $i$ème ligne de matrice. La concaténation des différentes matrices contient donc l'ensemble des embeddings de dimension 768 (128 neurones par couche, 3 matrices, 2 couches récurrentes pour la bidirectionnalité) pour tous les phonèmes.

Nous avons utilisé t-SNE (Maaten and Hinton, 2008) pour visualiser une projection en deux dimensions de cet espace de très haute dimension, avec comme paramètres une perplexité de 5, un learning rate de 1, et 1000 itérations. La figure 5.10 nous montre le résultat de cette projection pour un modèle entraîné sur le corpus audiovisuel en français, avec ou sans notre procédure d'injection de connaissances. Bien que t-SNE reste un outil de projection, ne pouvant refléter avec exactitude un espace de très haute dimension, ce dernier peut nous donner un aperçu des clusters formés par les phonèmes dans l'espace original. Sur ce graphique, nous pouvons observer la projection des *embeddings* des phonèmes sans utilisation de la procédure d'injections de connaissances (figure 5.10a) et avec utilisation de la procédure basée sur l'ACP (figure 5.10b).

Une première remarque pouvant être formulée par rapport aux deux graphiques est

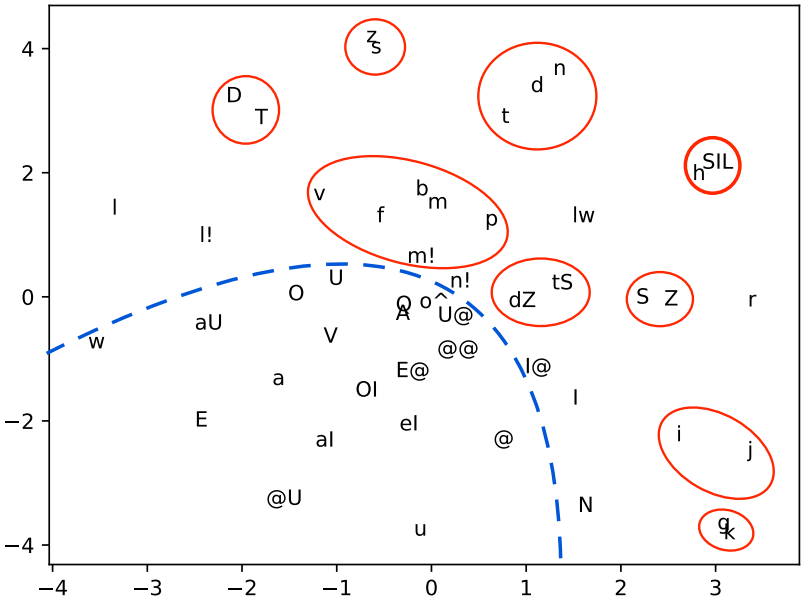


(a) Sans injection de connaissances

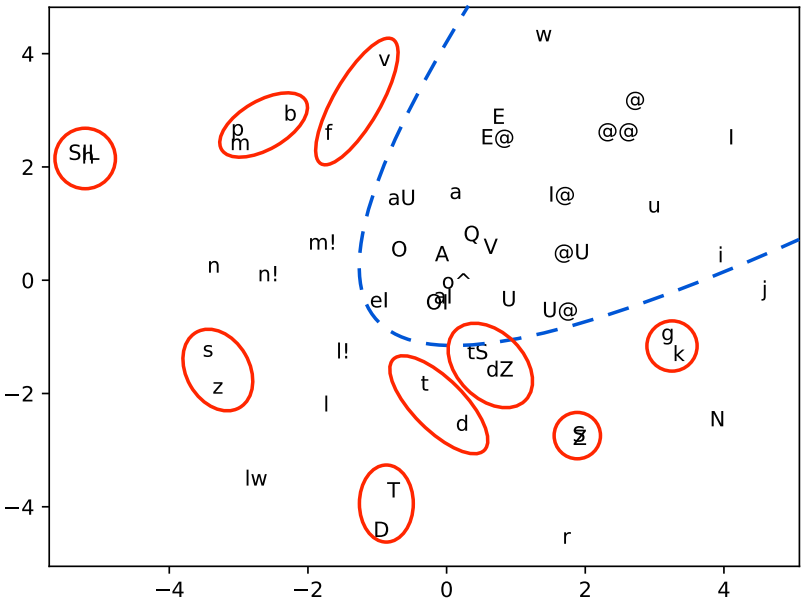


(b) Avec injection de connaissances (ACP)

FIGURE 5.10 – Projection 2D avec t-SNE de la première couche du modèle après apprentissage de la base audiovisuelle française.



(a) Sans injection de connaissances



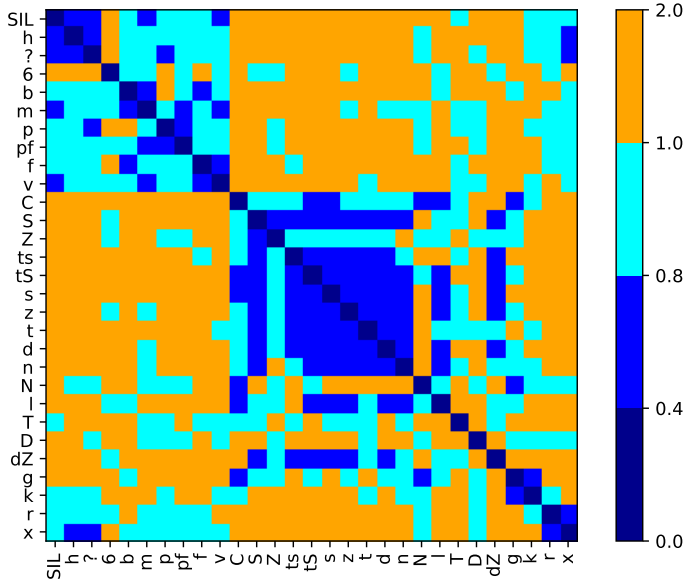
(b) Avec injection de connaissances (ACP)

FIGURE 5.11 – Projection 2D avec t-SNE de la première couche du modèle après apprentissage de la base articulatoire anglaise.

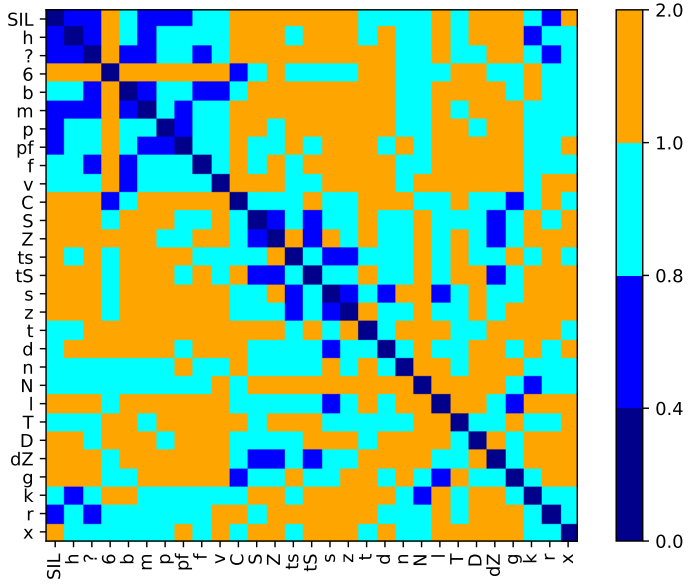
la nette séparation entre les consonnes et les voyelles faites par le modèle. Cependant, les *embeddings* appris lors de l'utilisation de l'espace articulatoire latent pré-calculé semblent bien mieux regrouper les phonèmes en fonction de leurs places et manières d'articulation. Par exemple, le cluster des bilabiales (/b/, /p/ et /m/) semble regrouper avec le cluster des labiodentales (/f/ et /v/) en comme phonèmes intervenant les lèvres lors de l'articulation, sans la procédure d'injection de connaissances, alors que ces 5 phonèmes forment deux clusters bien identifiables avec la procédure d'injection de connaissances. D'une manière similaire, nous pouvons identifier de nombreux clusters intéressants à la figure 5.10b, dont les frontières sont par ailleurs nettement plus marquées que la figure 5.10a. Par exemple le cluster des consonnes fricatives alvéolaires (/s/ et /z/) et postalvéolaires (/S/ et /Z/), ou encore celui des occlusives alvéolaires (/t/, /d/ et /n/) ou vélaires (/k/ et /g/). Pour les consommes, nous pouvons remarquer un regroupement des voyelles nécessitant de la protrusion, en particulier /u/ et /y/, ou des phonèmes proches les uns des autres dans leurs réalisations (/e/, /E/ et /eh/, ou encore /o/, /O/ et /oh/).

Pour le corpus articulatoire en anglais, des résultats similaires peuvent être observés à la figure 5.11. Nous retrouvons par exemple les clusters (/b/, /p/ et /m/), (/f/ et /v/), (/t/, /d/ et /n/), (/s/ et /z/), (/S/ et /Z/), mais aussi des clusters de phonèmes spécifiques à l'anglais comme les affriqués palato-alvéolaires sourdes (/tS/ et /dZ/) ou les fricatives lingodentale (/T/ et /D/). Nous pouvons aussi remarquer la proximité du silence, SIL, et de la fricative glottale sourde /h/. Ce phénomène s'explique certainement par l'absence d'un symbole particulier pour les moments d'inspiration du locuteur, alors assimilé à un silence, dont la réalisation est très proche de celle d'un h aspiré. Pour finir, la procédure d'injection de connaissances a un impact minime sur la différence entre ces espaces d'*embedding*, un résultat que nous pouvons éventuellement mettre en corrélation avec l'impact minime de notre procédure d'injection de connaissances sur les métriques de performances globales du modèle (cf. section 4.3, figure 4.7).

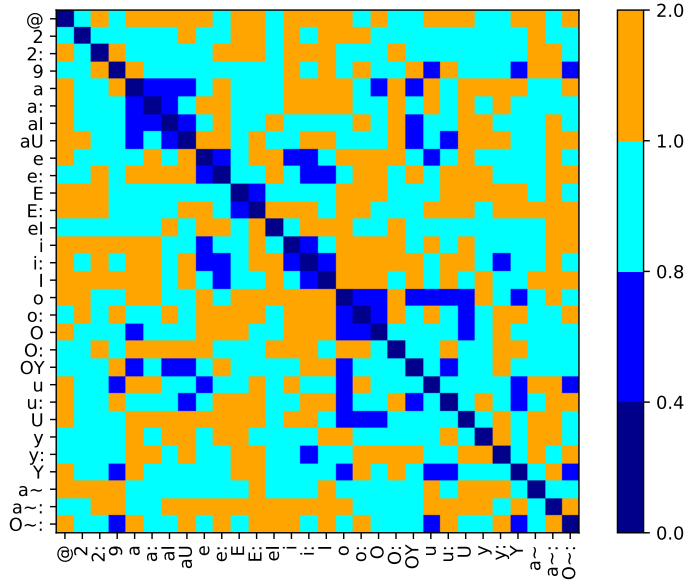
Nous n'avons malheureusement pas été en mesure de produire une projection lisible des *embedding* correspondant à nos deux corpus allemands à l'aide de t-SNE. Pour combler cette lacune, nous proposons une analyse semblable à celle réalisée à la section précédente, basée sur la distance cosine entre deux phonèmes dans l'espace d'embedding. Pour tous les phonèmes considérés par notre modèle, nous récapitulons la distance cosine entre les vecteurs d'embedding suite à l'apprentissage du corpus articulatoire à la figure 5.12, et à la figure 5.13 pour le corpus audiovisuel. Pour simplifier le travail d'analyse et augmenter la lisibilité, nous avons séparé les voyelles des consonnes. La visualisation résultante est moins facilement interprétable qu'une projection 2D, mais quelques remarques peuvent être formulées, en particulier :



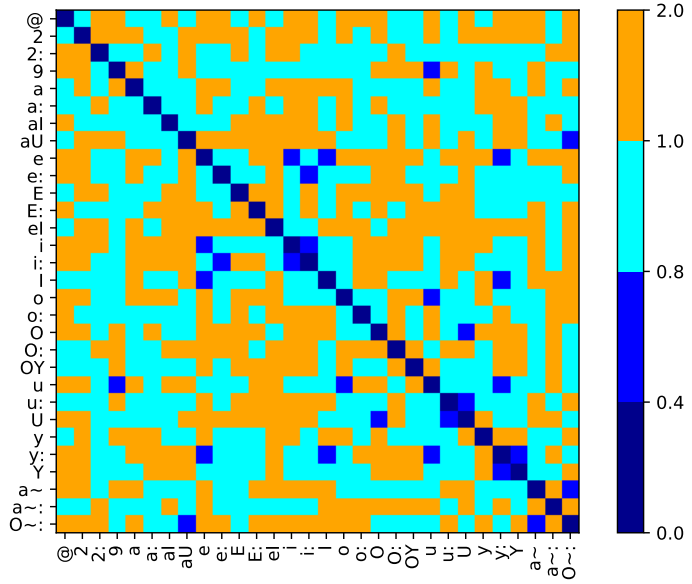
(a) Consonnes - Sans injection de connaissances



(b) Consonnes - Avec injection de connaissances (ACP)

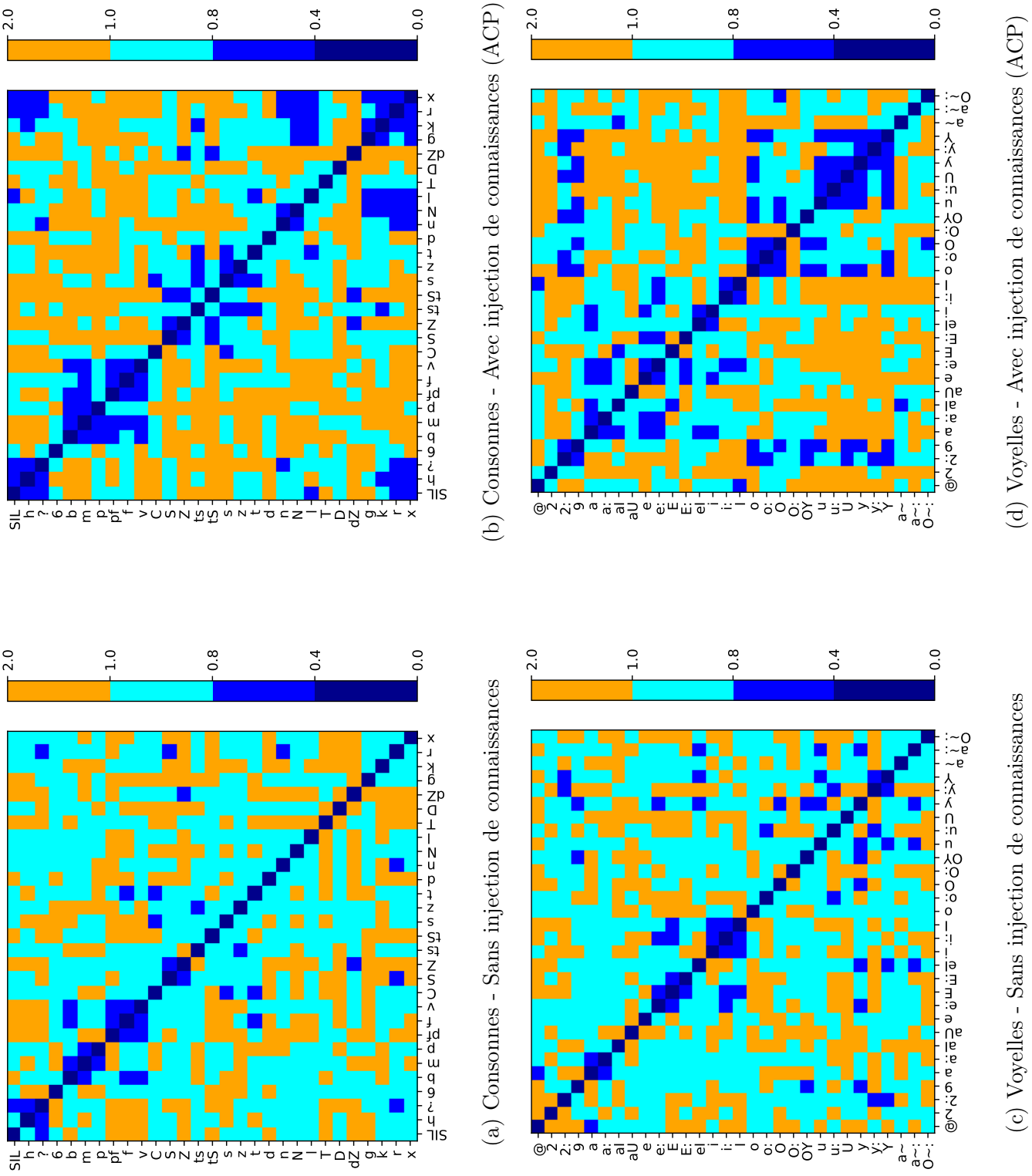


(c) Voyelles - Sans injection de connaissances



(d) Voyelles - Avec injection de connaissances (ACP)

FIGURE 5.12 – Distance cosine entre les vecteurs d'*embedding* des phonèmes allemand après apprentissage du corpus articulatoire



(a) Consonnes - Sans injection de connaissances

(b) Consonnes - Avec injection de connaissances (ACP)

(c) Voyelles - Sans injection de connaissances

(d) Voyelles - Avec injection de connaissances (ACP)

FIGURE 5.13 – Distance cosine entre les vecteurs d'*embedding* des phonèmes allemand après apprentissage du corpus audiovisuel

— Le cluster des bilabiales /b/ /p/ /m/ semble presque confondu avec celui des fricatives labiodentales /f/ et /v/ et avec l'affriqué /pf/.
— Nous retrouvons le cluster des fricatives /s/ et /z/, proche des affriqués /tS/, /ts/ et /dZ/.
— Les voyelles longues sont souvent proches de leurs voyelles courtes correspondantes (/a/ et /a :/, /E/ et /E :/, /o/ et /o :/).
— Le silence est comme pour l'anglais très proche du h aspiré /h/, mais aussi du coup de glotte allemand / ?/.
— Les clusters sont plus petits lors de l'utilisation de notre procédure d'injection de connaissances articulatoires, le réseau semble capable d'une meilleure différentiation.

5.4 Discussions

Ce chapitre nous a permis de valider avec plus de certitude la qualité des prédictions de notre modèle, en particulier la qualité par capteur (section 5.1.1) ainsi que sur des sections critiques du signal visuel (section 5.1.2). Cette première partie de chapitre qui focalise sur l'évaluation de nos modèles nous permet de confirmer la très bonne capacité des réseaux de neurones récurrents à modéliser la coarticulation, permettant à la fois d'obtenir une bonne dynamique des articulateurs tout en atteignant les cibles articulatoires critiques. Cette particularité nous permet d'affirmer que notre modèle est capable de lever une des limites de modèles basés sur les fonctions de dominance comme le modèle de Cohen-Massaro, limite également partagée par les modèles basés sur des HMMs qui souffrent d'une tendance à moyenner les résultats. La deuxième partie du chapitre fut consacré à l'étude de notre modèle, et a permis de mettre en lumière plusieurs résultats intéressants, ainsi que quelques limites de notre modèle et de nos corpus. Les expérimentations réalisées sur la coarticulation anticipative (section 5.2) nous permettent d'appréhender ce que le réseau a modélisé de la coarticulation anticipative, et permettent également de comprendre quel est le contexte minimal permettant une bonne prédiction de l'articulation par notre algorithme, qui semble être aux alentours de 3 à 4 phonèmes d'avance pour des résultats acceptables.

L'étude des poids du réseau (section 5.3) nous apporte deux conclusions. Premièrement, l'espace latent injecté dans le réseau de neurones par notre procédure d'injection de connaissances n'est pas beaucoup modifié par l'apprentissage (section 5.3.1), ce qui est un indice fort quant à la validité de notre procédure pour la modélisation de la coarticulation, cet espace latent étant exploité pratiquement tel quel par le modèle. Deuxièmement,

l'étude de la couche implicite d'*embedding* du réseau nous révèle que le modèle est capable de catégoriser correctement les labels arbitraires des phonèmes en fonction de leurs places et leurs manières d'articulation. Plus surprenant, cette catégorisation est capable de se réaliser d'une manière indépendante de la modalité utilisée, par exemple le regroupement des consonnes occlusives /k/ et /g/, dont l'articulation est conditionné par la langue, s'effectue bien lors de l'utilisation de la seule modalité visuelle. Ce phénomène s'explique certainement par l'un de ces deux points : il est soit effectivement possible pour notre réseau d'inférer partiellement les mouvements de la langue depuis la modalité visuelle, soit l'utilisation de /k/ et /g/ est conditionné par des règles phonologiques de plus haut niveau que le réseau est capable de prendre en compte. Même si nous ne pouvons pas trancher entre ces deux hypothèses en l'état actuel de nos travaux, il est très encourageant de constater de nombreuses cohérences entre les résultats obtenues dans nos expérimentations et les connaissances en phonétique articulatoire. Il est cependant important de souligner que ce que nous considérons comme la couche d'*embedding* de notre modèle n'est qu'une vue très partielle, négligeant complètement l'aspect dynamique, et pouvant donc omettre des informations phonétiques présentes dans les couches récurrentes.

Chapitre 6

Application à l'animation de la parole

Dans ce chapitre, nous abordons l'intégration de notre modèle de coarticulation à un système d'animation afin de développer une tête parlante virtuelle, nous permettant ainsi de générer la synchronisation labiale d'un modèle 3D arbitraire en fonction d'un segment de parole.

Comme évoqué durant l'introduction, l'intégration de la modalité visuelle à un système de synthèse à de nombreux bénéfices. Si nous nous attardons sur les capacités de ces systèmes à augmenter l'intelligibilité de la parole, à accroitre l'intérêt de l'utilisateur pour le système, et à permettre une visualisation fine de l'articulation, il semble alors indéniable qu'une telle technologie peut avoir un impact positif sur l'apprentissage des langues vivantes, intuitions confirmés par plusieurs études (Hazan et al., 2005; Massaro, 2003). En effet, un professeur virtuel muni de capacité d'articulation réaliste et de bonne qualité peut être très utile pour l'apprentissage de la prononciation. C'est par ailleurs l'un des objectifs de METAL (Modèles et Tracers au service de l'Apprentissage des Langues), l'un des lauréats de l'appel à projets e-FRAN (Formation, Recherche et Animation Numériques dans l'éducation). Les projets e-FRAN ont pour objectifs l'étude et la conception d'outils numériques pour l'éducation. Le projet METAL, qui finance ces travaux de thèse, s'intéresse à la création d'outils pédagogique pour l'apprenant et l'enseignant, afin d'aider à l'enseignement des langues vivantes, et plus particulièrement du Français et de l'Allemand. Pour les besoins du projet, nous appliquons donc ici notre modèle de coarticulation à la langue allemande, mais cette utilisation peut être facilement transférée à toutes autres langues vivantes.

Dans ce contexte d'apprentissage de la prononciation, la qualité de l'articulation doit être le critère principal d'évaluation d'un tel système. Ce critère doit être évalué vis-à-vis de la perception des humains pour valider que l'intelligibilité de l'articulation est

convaincante. Bien que l'étude de mesures objectives comme la RMSE ou le coefficient de corrélation soit des indices précieux sur la qualité de prédiction de notre modèle, de même que nos expériences sur les cibles articulatoires critiques (cf. section 5.1.2), ces derniers ne peuvent pas nous confirmer avec certitudes la perception qu'un humain peut avoir des trajectoires des articulateurs une fois intégrées à un système d'animation de la parole. En effet, nous avons présenté en début de thèse la grande sensibilité de l'humain à la moindre incohérence entre le signal visuel et acoustique de la parole. Ces perturbations, parfois minime et donc difficilement identifiable à l'aide de métriques de performances globales comme nous en avons fait l'usage au chapitre 4, peuvent aller jusqu'à une forte dégradation de l'intelligibilité du discours. Pour attester de la qualitée perçue de nos prédictions de la coarticulation, une fois appliquées à une tête parlante virutelle, nous avons conduit une expérimentation perceptive aurpès de locuteurs natifs allemand.

Plutôt que de comparer deux résultats issus de nos prédiction pour chercher par exemple un éventuel apport de notre procédure d'injection de connaissances, nous avons choisi de comparer ces prédictions aux trajectoires naturelles, telles que contenues dans notre corpus audiovisuel allemand. Notre expérience consiste donc à faire noter par les participants leurs niveaux de préférences entre deux animations de la parole, l'une d'entre elles étant une resynthèse depuis les trajectoires originales acquises par le système de capture de mouvement, et l'autre étant prédite par notre modèle de coarticulation basé sur le réseau de neurones ayant obtenu la plus basse RMSE (version avec injection de connaissances articulatoires). Cette tâche semble de prime abord très complexe, il semble à première vue impossible de pouvoir faire "mieux" que les mouvements articulatoires originaux. Cependant, nous espérons aboutir à un résultat où les participants sont en incapacité de faire la différence entre mouvements articulatoires prédits et mouvements articulatoires naturels, ce qui validerait l'efficacité de notre méthode de modélisation de la coarticulation pour la synthèse visuelle de la parole.

6.1 Système d'animation de la parole

Nous présentons à cette section le système d'animation de la parole développé au sein l'équipe Multispeech au Loria. Ce dernier s'articule autour de trois modules principaux, récapitulés à la figure 6.1 :

— Une procédure d'alignement phonétique, capable d'obtenir la segmentation phonétique nécessaire depuis le signal acoustique de la parole.
— Un modèle capable de prédire la dynamique du visage en fonction de la segmentation phonétique, dans ce cas le modèle de coarticulation présenté dans cette thèse.

— Un système d'animation facial couplé à une procédure de *retargeting*, afin d'animer l'avatar virtuel selon la dynamique prédite par le modèle de coarticulation.

La figure 6.1 nous donne un aperçu de notre système d'animation de la parole. Ce dernier prend en entrée le signal acoustique correspondant à la parole, sur lequel est appliqué une procédure d'alignement afin d'obtenir une représentation phonétique du discours. Cette procédure d'alignement est réalisée avec les mêmes modèles acoustiques que ceux utilisés pour la préparation de nos corpus (cf. chapitre 3). Comme introduit au le chapitre 2, cet alignement phonétique sera exploité par notre modèle, lui permettant de prédire la dynamique du visage du locuteur de notre corpus. Nous présentons dans cette section la procédure permettant le *retargeting* de l'animation faciale vers notre modèle 3D, cette dernière repose essentiellement sur l'utilisation d'un espace de *blendshape* commun aux modèles 3D et au locuteur. Cette technique nous permets d'étendre très simplement notre système d'animation vers de nouveaux modèles 3D et de nouvelles langues. Par

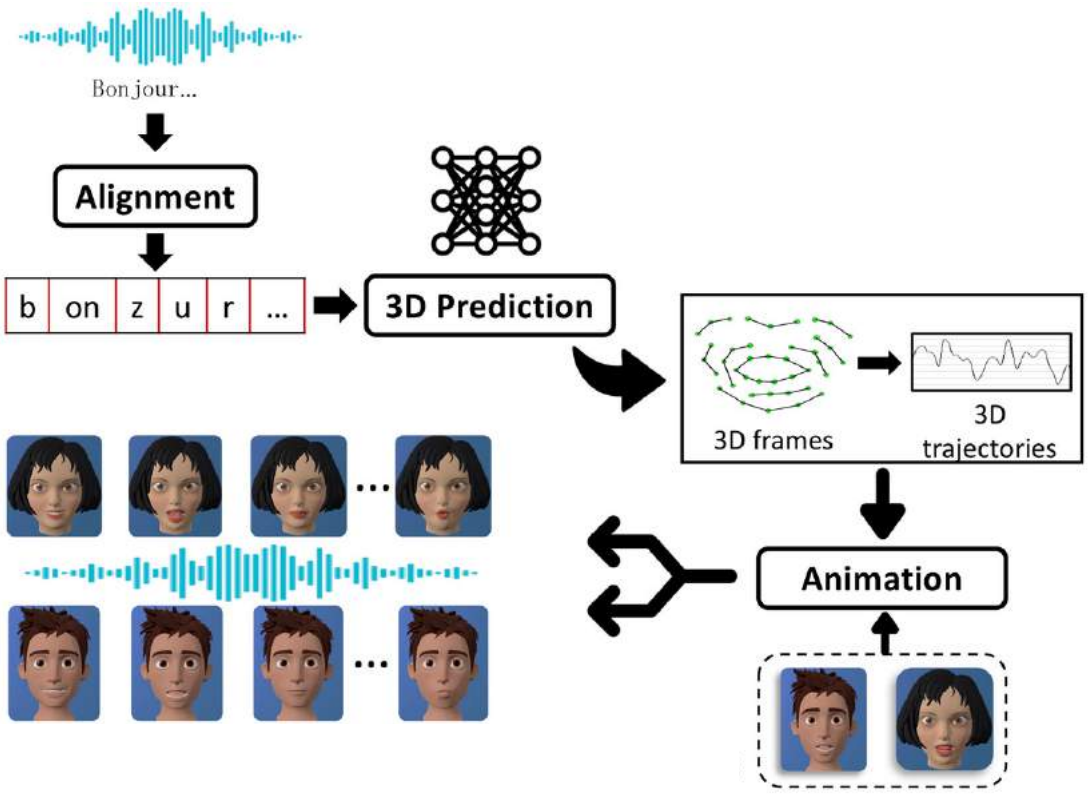


FIGURE 6.1 – Aperçu général de notre procédure d'animation faciale guidée par la parole. Le système prend en entrée un signal acoustique dont il extrait l'alignement phonétique utilisé par notre modèle de coarticulation. Ce dernier prédit finalement la dynamique du visage permettant de guider l'animation en fonction du segment de parole.

exemple, nous avons débuté par la langue française et le modèle 3D féminin, puis ajouté le modèle 3D masculin (ces deux modèles sont visibles à la figure 6.1). Le passage d'une langue à l'autre se fait en remplaçant le modèle acoustique de la procédure d'alignement et le modèle de coarticulation du module de prédiction 3D par des modèles adaptés à la nouvelle langue, ce que nous avons réalisé pour la langue allemande.

6.1.1 Animation par *blendshape*

Notre plateforme repose sur l'animation par *blenshape*, ou *morph target animation*, qui est devenue aujourd'hui un des standards des méthodes d'animation au côté de l'animation par squelette par exemple. Cette dernière est particulièrement prisée, car donne à l'animateur et au graphiste un grand contrôle sur la position de chaque sommet du modèle 3D, et permet d'animer une surface dont les mouvements sont difficilement représentables par un squelette, comme du tissu, la peau, ou dans notre cas les expressions faciales. Comme abordé à la section 2.3.2, cette méthode consiste en la création de plusieurs versions cible d'un même modèle 3D. Ces déformations cibles, les *keyshapes*, seront interpolés linéairement afin d'obtenir l'apparence souhaitée sur le modèle 3D. L'évolution du poids associé à chaque *keyshapes* va donc permettre de contrôler l'animation du modèle.

Dans le cadre de l'animation de la parole, il semble assez direct que nous puissions faire un intéressant parallèle entre ces *keyshapes* et les visèmes, présentés à la section 1.1.2. Pour rappel, un visème est ce que nous voyons lors de la production d'un phonème. En considérant ces derniers comme l'ensemble des *keyshapes*, le système nous laisse alors apercevoir d'intéressant lien avec la théorie de la coproduction Löfqvist (1990), en particulier avec les modèles de coarticulation par fonction de dominance comme le modèle de Cohen and Massaro (1993). Avec cette vision, la variation de la forme neutre du modèle 3D vers l'un des visèmes représente un geste articulatoire, et la coproduction de ces gestes s'exprime par la superposition des différents gestes articulatoires induits par l'interpolation des différents visèmes.

6.1.2 *Retargeting* du nuage de points vers le modèle 3D

Ce système de *retargeting* de la dynamique d'un nuage de points vers l'animation d'un modèle 3D est grandement inspiré par les travaux de Chuang and Bregler (2002). Pour chaque visème, le système possède un modèle 3D correspondant (*keyshapes*), dont l'ensemble des sommets est stocké dans un vecteur B_i formant la matrice des *keyshapes* $B = [B_1, ..., B_k]$. Pour chacun de ces visèmes, nous cherchons également l'équivalent depuis le corpus de capture de mouvement utilisé pour l'apprentissage du modèle, ce qui nous

donne une matrice de nuage de points $F = [F_1, ..., F_k]$ très similaire à la matrice des visèmes du modèle 3D B. Pour obtenir les poids associés à chaque *keyshapes* depuis un nuage de point du corpus, nous résolvons l'équation suivante :

$$argmin_x(\|F.x - y\|_2) \tag{6.1}$$

Dans laquelle y est le nuage de point représentant le visage, tel qu'obtenu depuis la prédiction du réseau de neurones (3D frames sur la figure 6.1), et compte tenu des *keyshapes* sous forme de nuage de point stocké dans la matrice F. Le vecteur des poids x peut ensuite être réutilisé avec la matrice B pour animer le modèle 3D.

Une solution peut-être facilement obtenue par une régression positive par minimisation des moindres carrés, mais cette procédure de régression est appliquée indépendamment à chaque frame, aucune notion de temporalité n'est utilisée lors du *retargeting*. Un filtre passe-bas est donc appliqué sur l'évolution temporelle de ces poids afin d'assurer des trajectoires lisses, les trajectoires erratiques pouvant aboutir à l'apparition de léger tremblement du modèle lors de l'animation. Pour conclure, nous pouvons donc animer la tête parlante depuis un segment de parole, en utilisant le modèle de coarticulation présenté dans ces travaux pour estimer y.

6.1.3 Sélection des visèmes

Le système d'animation exploite un ensemble de 13 visèmes et d'une position neutre, ensemble inspiré par les travaux de Benoit et al. (1992) qui propose une liste de 17 visèmes comme adaptation pour le Français de la liste de 16 visèmes proposé par la norme MPEG4 (Pakstas et al., 2002). Une autre liste de visème fut proposée par Govokhina (2008), qui réduis cet ensemble à seulement 8 visèmes. Govokhina (2008) utilise la mesure de Bhattacharyya (Mak and Barnard, 1996) pour calculer la distance entre les cibles articulatoires associées à chaque phonème. Les plus proches phonèmes furent regroupés ensemble pour définir des catégories visuelles. Une grande différence entre ces deux listes est l'utilisation d'information interne par Benoit et al. (1992), qui sépare ainsi le visème [t, d, n, ɳ] de [l] et de [ʁ].

Notre système d'animation ne modélisant pas les mouvements de la langue, l'équipe à d'abord exploité la classification de Govokhina (2008). Cependant, lors de l'utilisation de seulement 8 visèmes, la reconstruction du nuage de point après projection dans l'espace des blenshapes, c'est-à-dire la différence entre y et $F.x$ dans l'équation 6.1, était très importante (une RMSE supérieur à 2 mm et une corrélation très critique, pouvant tomber jusqu'à $\rho = 0,1$). La liste fut donc étendue à 22 visèmes, afin de pouvoir distinguer les plus

Phonème	Visème / Keyshape	Phonème	Visème / Keyshape
#, ə		i	
p, b, m		j	
f, v		o, ɔ,ɔ̃, ɑ̃	
ʃ , ʒ		y	
s, z		u	
t, d, n, ɲ, k, g, l, ʁ		w,ɥ	
a, e, ɛ, ɛ :, ɛ̃		ø, oe	

TABLE 6.1 – Liste des visèmes de l'application, avec les correspondances phonétiques et le modèle 3D associés.

fins détails lors de l'articulation, ce qui permis de descendre la RMSE à la reconstruction en dessous du milimètre et avec une corrélation supérieur à 0,8. La dernière étape fut de diminuer cette liste à 13 visèmes en supprimant les *keyshapes* avec le plus faible impacte sur la précision des reconstructions. La liste complète de nos 13 visèmes et de la forme neutre est disponible au tableau 6.1.

6.2 Évaluation subjective

6.2.1 Protocole

Chaque participant a 50 comparaisons à effectuer, toutes évidemment sélectionnées depuis l'ensemble de tests de nos corpus audiovisuels, et 6 niveaux de préférence pour répondre à la question "Lequel de ces avatars semble avoir la meilleure articulation ?". L'ordre d'apparition des 50 paires de vidéos, ainsi que la position (gauche ou droite) de chaque vidéo de la paire est tirée aléatoirement pour chaque participant. L'expérimentation a été conduite à l'aide d'une plateforme web auprès de 10 locuteurs allemands natifs. Comme le corpus allemand a été prononcé par un homme, nous utiliserons la version masculine du modèle 3D pour l'évaluation subjective suivante afin de ne pas perturber les participants en proposant l'animation d'un visage féminin avec la voix du sujet du corpus audiovisuel allemand.

Après avoir décodé les résultats et associer chaque label (A ou B) des vidéos avec leurs véritables natures (prédiction ou originale), nous pouvons obtenir le pourcentage de sélection des 6 choix proposés aux participants. Ces 6 valeurs de l'échelle entre A et B ont pour signification "la prédiction", "plutôt la prédiction", "un peu la prédiction", "un peu l'orignal", "plutôt l'original" et "l'original". Ce choix d'un nombre pair niveaux est délibéré afin de forcer l'utilisateur à donner une préférence pour l'un des deux échantillons, afin de ne pas favoriser l'apparition de choix médians. En effet, nous pensons que donner aux participants la possibilité de sélectionner ce choix "neutre" peut favoriser grandement l'apparition du résultat espéré, et donc légèrement biaiser nos conclusions. En forçant l'utilisateur à établir sa préférence, nous nous assurons que cette égalité dans la préférence entre resynthèse et prédiction émerge au niveau global, et ne soit pas le résultat d'une sélection médiane locale, plus "confortable" pour un participant indécis.

Pour finir, nous transformons également ces choix catégoriques en un taux d'appréciation, note exprimée en pourcentage, afin d'obtenir une métrique facilement interprétable nous permettant d'évaluer la préférence globale d'un participant entre les trajectoires articulatoires synthétisées et les trajectoires naturelles. Les deux choix extrêmes ("la prédic-

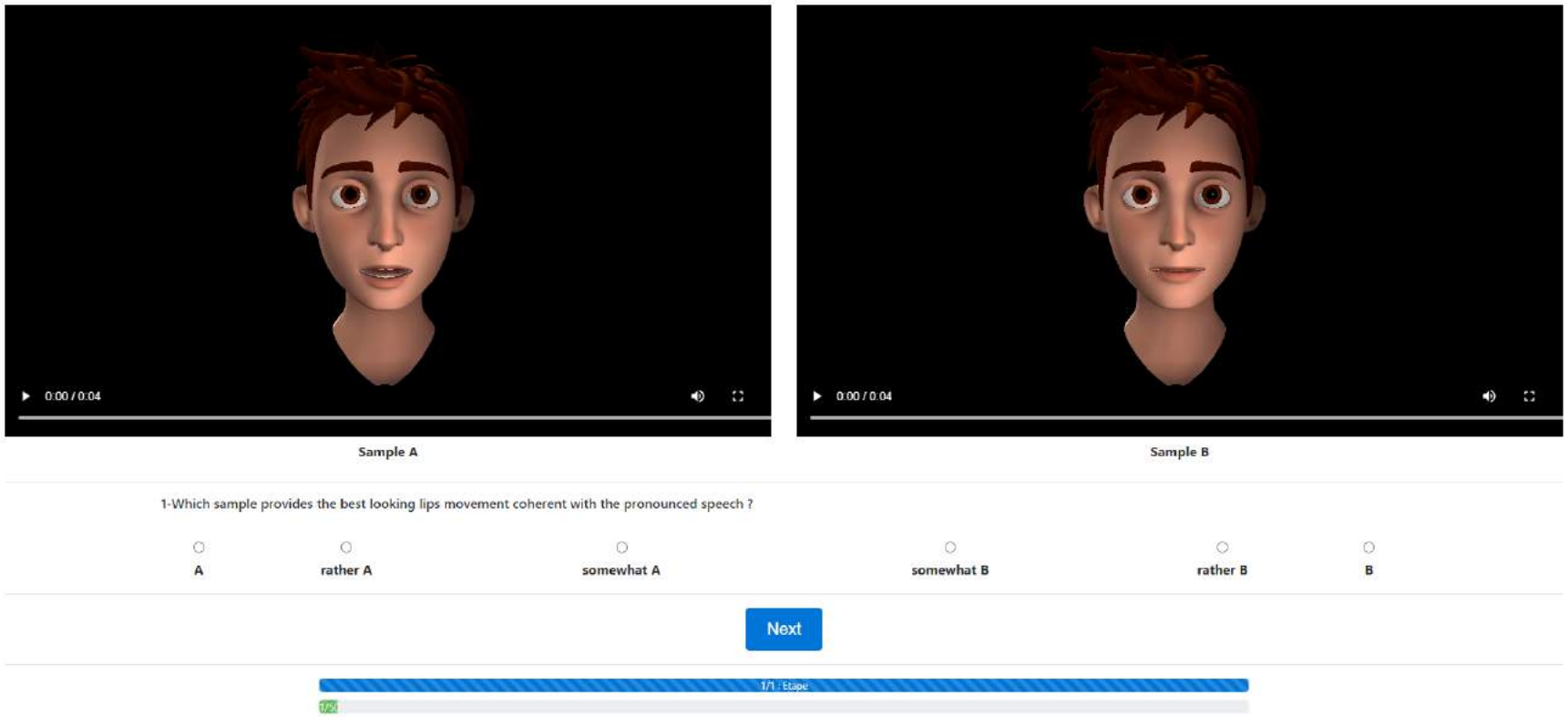


FIGURE 6.2 – Example d'écran de l'application web utilisé pour l'expérience perceptive. Nous pouvons voir côte à côte les deux animations du même segment de parole, ainsi que les 6 choix proposés aux participants pour évaluer leurs préférences.

tion" et "l'original") sont traduits par une approbation de 100% pour la vidéo sélectionnée et de 0% pour l'autre. Les taux d'approbations sont par la suite répartis uniformément pour les 4 autres choix, à savoir 80%/20% pour les choix "plutôt la prédiction" et "plutôt l'original", et 60%/40% pour les choix "un peu la prédiction" et "un peu l'original".

6.2.2 Résultats

La figure 6.3 nous présente le pourcentage de sélection des 6 catégories, avec un code couleur que nous réexploiterons dans les figures suivantes. Dans ce graphique, un premier élément marquant est le résultat a priori mitigé de notre modèle de coarticulation, ayant la faveur de l'utilisateur dans 36,8% des cas contre 63,2% de préférence pour les vidéos synthétisés depuis les trajectoires originales des articulateurs. Cependant, nous pouvons considérer les deux choix médians (un peu la prédiction et un peu l'original) comme un marqueur de l'hésitation forte des participants, de leur difficulté à choisir entre les deux vidéos, et donc entre les trajectoires issues de notre modèle et celle acquise par *motion capture*. Avec ce point de vue, nous pouvons considérer que ces deux choix, représentant 39,2% des sélections, sont actuellement des résultats positifs pour notre expérience, car mettent en lumière une réelle difficulté de l'humain à différencier les deux vidéos. Ce

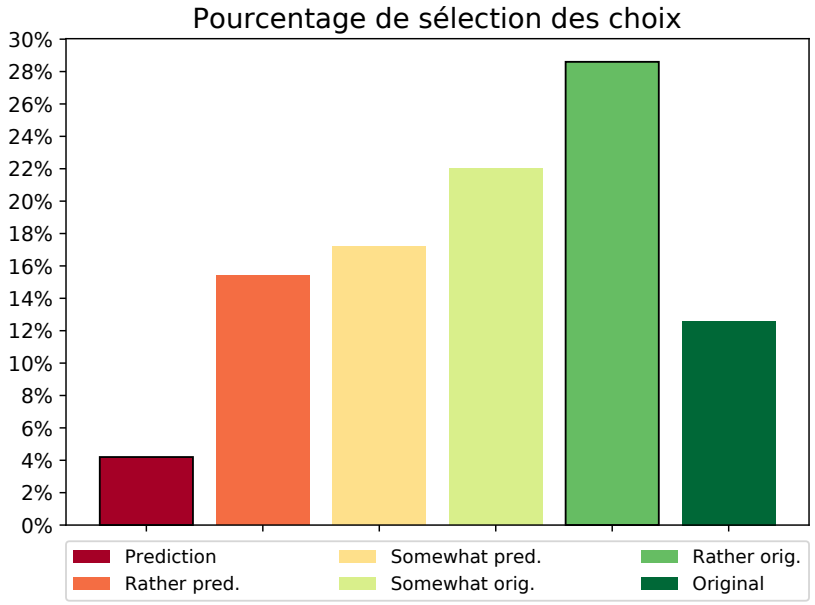


FIGURE 6.3 – Résultat de l'expérience perceptive pour la langue allemande : pourcentage de sélection de chaque catégorie. Les catégories avec un contour noir sont significatives ($p < 0,01$).

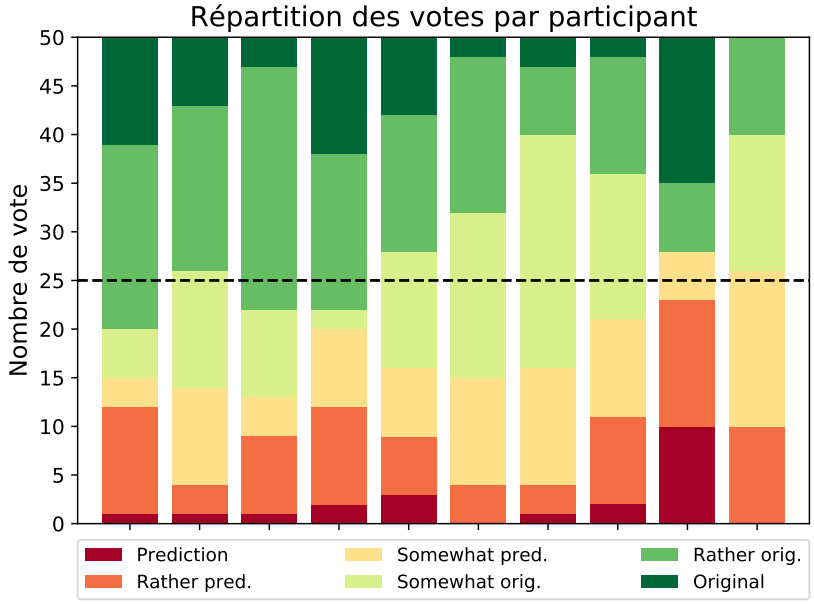


FIGURE 6.4 – Résultat de l'expérience perceptive pour la langue allemande : taux d'appréciation de la prédiction pour les 10 participants.

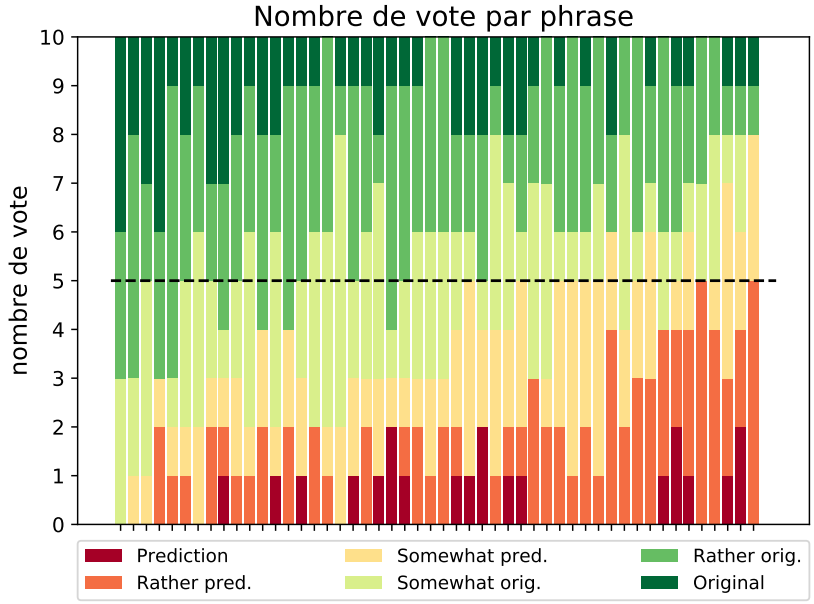


FIGURE 6.5 – Résultat de l'expérience perceptive pour la langue allemande : taux d'appréciation de la prédiction pour les 50 phrases.

choix nous amène donc à interpréter la figure 6.3 de la manière suivante : Dans 58,8% des cas le participant à une préférence pour la coarticulation issue de notre modèle ou a de grandes difficultés à différencier les deux vidéos, et dans 41,2% des cas le participant à une préférence pour les trajectoires originales des articulateurs, dont 12,6% de préférence forte et sans hésitations.

Les résultats ci-dessus sont cependant à considérer avec une prudence particulière. En effet, si nous appliquons un test de Student indépendamment sur ces 6 catégories pour nos 10 participants, avec comme hypothèse nulle la moyenne issue d'un hasard total ($\mu_0 = \frac{nombre\ d'echantillons}{nombre\ de\ choix} = 50/6$), seule la sélection des catégories "prédiction" et "plutôt original" est significative ($p < 0,01$). Nous pouvons donc qu'affirmer avec certitudes ces deux points :

— Le choix "prédiction" n'est quasiment jamais sélectionné, ce qui représente clairement la difficulté à prédire des trajectoires articulatoires plus réalistes que ceux acquis par la capture de mouvement.

— Le choix "plutôt original" est celui le plus souvent sélectionné, ce qui dénote d'un doute très souvent présent pour le participant lors de l'expression de sa préférence, mais d'une tendance globale vers l'animation générée depuis la *motion capture*.

6.2.3 Taux d'appréciation

Cette tendance pour les animations issues des trajectoires originales se nuance cependant si nous interprétons ces résultats en termes de taux d'appréciation, c'est-à-dire en convertissant les votes catégoriques en pourcentage d'appréciation, comme présenté à la section 6.2.1. Nous pouvons observer l'appréciation moyenne par participant à la figure 6.6. Globalement, cette moyenne vaut 41,36%, une approbation correspondant à la catégorie "un peu l'originale", et relativement proche de notre limite théorique de 50%. De plus, la différence entre l'approbation moyenne pour les prédictions et pour les trajectoires originales est statistiquement significative ($p < 0,001$).

Dans cette figure, ainsi qu'à la figure 6.7, nous rappelons le niveau d'appréciation des différents choix par une ligne horizontale en pointillé reprenant le code couleur de la figure 6.3, ainsi que deux lignes noires : l'une en pointillée représentant l'objectif idéal de 50%, et l'autre continue qui représente le taux d'appréciation moyen des vidéos issues de la prédiction.

Nous pouvons remarquer que trois des participants ont eu de grandes difficultés à différencier les paires de vidéos, avec un taux d'appréciation entre 47,2% et 50,4%. Quatre participants sont aux alentours du choix "un peu l'original" avec une approbation allant de

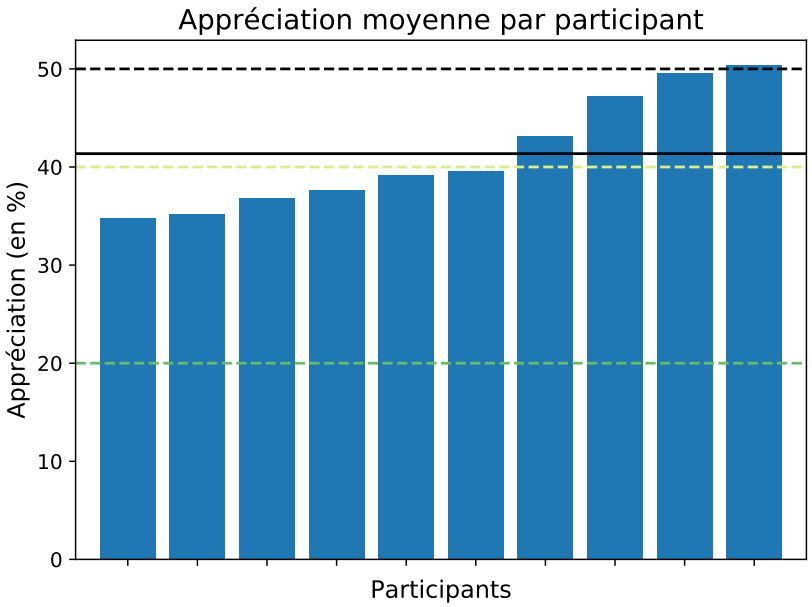


FIGURE 6.6 – Résultat de l'expérience perceptive pour la langue allemande : taux d'appréciation de la prédiction pour les 10 participants.

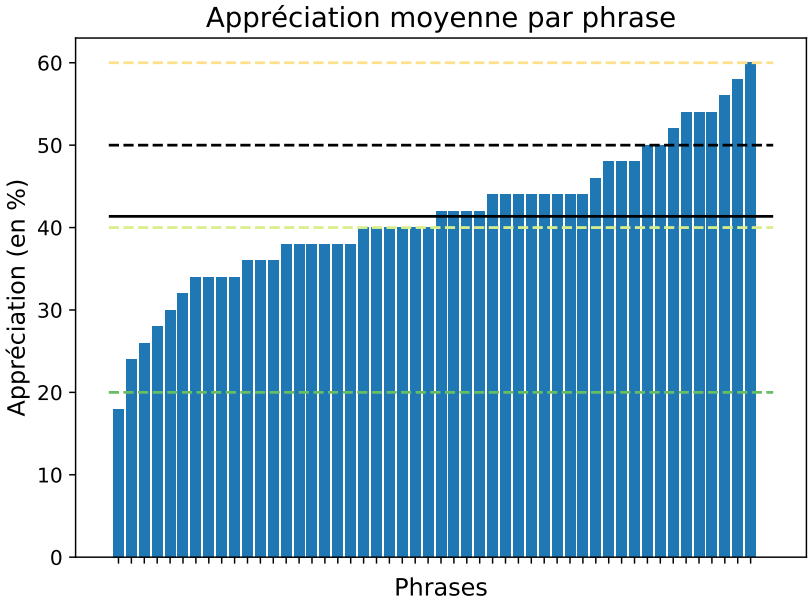


FIGURE 6.7 – Résultat de l'expérience perceptive pour la langue allemande : taux d'appréciation de la prédiction pour les 50 phrases.

37,6% à 43,2%, et trois participants sont à mi-chemin entre "un peu l'original" et "plutôt l'original" avec une approbation variant de 34,8% à 36,8%. Si nous portons attention au taux d'appréciation moyen par phrase (figure 6.7), nous distinguons 13 phrases que nous considérons avoir réussi notre test avec plus de 46% de taux d'appréciation (dont 9 ayant atteint ou même dépassé le score optimal de 50%), 10 phrases dont les prédictions n'ont pas totalement convaincu les participants (5 variant de 18% à 30% d'approbation, et 5 variant de 32% à 34%), et 27 phrases aux alentours de notre moyenne avec une approbation allant de 36% à 44%.

6.3 Discussion

Ce chapitre propose une méthode permettant d'animer le visage d'un avatar virtuel arbitraire en fonction d'un segment de parole, ayant aboutis à la demande de brevet. Notre système repose essentiellement sur l'utilisation conjointe du modèle de coarticulation présenté dans cette thèse et d'une procédure de *retargeting* permettant de transférer l'articulation d'un locuteur à un visage virtuel 3D arbitraire. Ce mappage entre la dynamique du nuage de points représentant le visage et l'animation du modèle 3D repose sur la définition d'un espace commun, celui des *blenshapes*, où chaque dimension représente la pondération d'un *keyshape* spécifique. Cette méthode requiert néanmoins la définition de chaque keyshape pour chaque locuteur (sous forme de nuage de points) et modèle 3D, une tâche réalisée manuellement. De plus, notre ensemble de *keyshapes* est optimisé pour le français, mais ces 13 visèmes couvrent un très grand espace articulatoire, qui semble suffisamment riche pour une application à la langue allemande. Par conséquence, le nuage de point tel que prédit par notre modèle de coarticulation contient vraisemblablement des informations plus fine quand aux déformations du visage, et il peut donc sembler dommage de perdre une partie de cette information lors du calcul des poids nécessaires aux *blendshapes*. En contrepartie de cette perte d'informations, nous ne somme plus limité à la physionomie du locuteur, et pouvons ainsi animer des avatars à l'apparence bien différente. Cette considération est primordiale dans l'optique d'une utilisation de ce système par des professionels de l'industrie du divertissement.

Lors de notre évaluation subjective, nous avons demandé à des locuteurs natifs allemands de faire un choix entre une animation originale du locuteur, resynthétisé depuis les données de capture de mouvement, et une animation guidée par notre modèle de coarticulation. Nous estimons que si ce choix est trop difficile pour les participants, si ces

3. Deep Lypsinc Dynamic, n°PCT/EP2020/072272 du 07/08/2020

derniers ne peuvent différencier significativement les deux animations, alors le modèle de coarticulation pilotant la tête parlante peut être considéré comme de très bonne qualité. Nous avons mis en place une simple métrique du taux d'appréciation permettant de mesurer la préférence entre les mouvements articulatoires originaux et nos prédictions, avec un objectif théorique de 50% correspondant à une appréciation égale des deux animations. Nos modèles obtiennent un score honorable de 41,36%. Plus intéressant, notre analyse de l'appréciation par phrase met en lumière une forte dissonance au sein de notre corpus de test, l'appréciation moyenne variant de 18% à 60%. Une analyse approfondie des phrases dont les prédictions sont les moins appréciées pourrait permettre de mettre en évidence des erreurs de notre modèle, que nous pouvons tenter de corriger en introduisant plus de phrases avec ce contexte phonétique problématique dans le corpus d'apprentissage, ou en pondérant la fonction d'erreur pour accorder une plus grande importance à ces contextes sous-représentés.

Conclusion

Plusieurs contributions se dégagent de cette thèse, la plus importante étant certainement la confirmation que les réseaux de neurones récurrents bidirectionels sont bien adaptés au problème de modélisation de la coarticulation. En effet, les sections 4.2, 4.3 et 5.1 nous montrent en quoi notre modèle est capable d'obtenir de très bonnes performances en terme de RMSE et de corrélation, et nous confirment également la capacité de cette approche à atteindre des cibles articulatoires critiques contrairement à certains modèles *time-locked* (par exemple le modèle de Cohen and Massaro (1993), cf. section 1.2.2). De plus, notre modèle fut également appliqué à un système d'animation de la parole, et obtient de très honorables résultats lorsque confrontés à l'appréciation subjective de locuteurs natifs (section 6.2).

Notre modèle de coarticulation peut s'appliquer avec succès à différentes langues, différentes modalités articulatoires, et différents locuteurs, mais de nombreuses pistes restent encore ouvertes afin de prendre l'intégralité de ces données en considération. En effet, nous pouvons considérer que notre réseau de neurones apprend à "cloner" l'articulation d'un locuteur spécifique, et n'apprend pas réellement à articuler d'une manière indépendante à l'anatomie du locuteur. Nous pouvons donc retrouver dans nos prédictions des spécificités d'articulation propres au locuteur du corpus d'apprentissage, qui ne sont pas imposées par des contraintes articulatoires, mais des choix personnels d'articulation. Une piste de recherche qui nous semble donc particulièrement intéressante est d'adresser l'aspect monolocuteur de notre modèle à l'aide de notre procédure d'injection de connaissances articulatoires. Nous pensons que cette dernière peut tout à fait exploiter l'utilisation d'une fonction $f_{decoder}$ plus expressive, capable d'exploiter un espace latent représentant l'espace articulatoire de nombreux locuteurs. Pour être compatible avec notre approche, cette fonction devrait permettre d'obtenir le nuage de point correspondant à un locuteur depuis deux principales informations : l'une permettant d'isoler l'identité du locuteur et ses caractéristiques anatomiques, et l'autre permettant de retrouver l'état de l'espace articulatoire indépendamment de tous locuteurs. L'utilisation de formes plus avancées de réseau autoencodeur, dont la version la plus simple fut utilisée avec succès par notre mo-

dèle, nous semble particulièrement prometteuse, en particulier les réseaux autoencodeurs variationnels (voir l'article de Kingma and Welling (2019) pour une introduction de ce modèle). L'effort d'acquisition nécessaire à une telle tâche nous semble cependant très important, alors qu'il existe d'ores et déjà une importante quantité de ressources vidéo alliant parole et signal visuel (la plateforme Youtube par exemple). Explorer l'utilisation de ces ressources et d'un système de *tracking* afin d'extraire la position spatiale de points d'intérêts pourrait donc être un point d'entrée intéressant à ce problème de généralisation de l'espace articulatoire, du moins pour la modalité visuelle.

Contrairement à de nombreuses approches sur l'inversion acoustique, mais d'une manière similaire aux modèles de coarticulation "traditionnels", notre modèle est capable de prédire cette articulation depuis une représentation de haut niveau de la parole : l'alignement phonétique, c'est-à-dire les phonèmes et leurs durées respectives. Si l'inversion acoustique permet d'exploiter de nombreux indices acoustiques quant à la forme du conduit vocal, ces modèles n'expliquent pas comment l'articulation est planifiée chez l'humain. Le succès de notre modèle semble être un indice supplémentaire quant à la nature phonétique de cette unité de planning, ou tout du moins une confirmation que cette représentation de la parole, bien que compacte et symbolique, contient une quantité non-négligeable d'informations quand à la dynamique des articulateurs. De plus, l'analyse de la couche d'*embeddings* de notre modèle (sections 5.3.2 et 5.2), nous confirme que le réseau de neurones parvient à capturer des informations cohérentes avec les connaissances en phonétique articulatoire. En effet, l'utilisation de l'algorithme t-SNE nous permet de mettre en évidence certains clusters de phonèmes dont les caractéristiques phonétiques sont semblables. D'une manière très étonnante, certains de ces clusters (comme le cluster /t/ /d/ /n/ et le cluster /k/ /g/) apparaissent lors de l'apprentissage depuis la modalité visuelle seule, alors que le lieu et la manière d'articulation les caractérisant sont définis par la langue. Plusieurs explications s'offrent à nous pour ce phénomène, il est soit possible de partiellement inférer les mouvements de la langue depuis la seule modalité visuelle, ou ces phonèmes on un effet semblable sur la coarticulation des phonèmes voisins, et peuvent donc être regroupés par rapport à leurs effets, ou encore l'utilisation même de certains phonèmes au sein d'un mot répond à des règles phonologiques dépendant du langage, que le réseau parvient à saisir.

L'utilisation de l'alignement phonétique peut éventuellement servir de première approche afin d'étuder les capacités multilingues de notre modèle. Plutôt que d'utiliser un ensemble de phonèmes spécifiques à chaque langue, et donc des entrées différentes pour chaque langue, nous pourrions utiliser un ensemble international de phonèmes, par exemple celui définit par le corpus GlobalPhone (Schultz et al., 2013). Il serait alors

primordial de vérifier si les différentes stratégies de coarticulation observées chez l'humain, par exemple les différences de production du segment /utu/ en turque et en anglais (Boyce, 1990), sont bien capturées et reproduites par notre modèle de coarticulation. Une piste de réalisation intéressante pour de tels travaux nous semble être de travailler sur les *embeddings* des phonèmes. Plutôt que de laisser le réseau inférer ces derniers lors de l'apprentissage, nous pouvons certainement apporter aux réseaux une représentation efficiente du domaine phonétique. Nous pourrions pour ce faire appliquer les techniques provenant du traitement automatique des langues, et plus particulièrement les méthodes comme word2vec (Mikolov et al., 2013), Elmo (Peters et al., 2018) ou encore BERT (Devlin et al., 2018). A la place d'appliquer ces méthodes au niveau du mot, nous pourrions les utiliser au niveau du phonème. Ceci nous permettrait de définir un *embedding* propre à chaque phonèmes dans chaque langues, que nous utiliserions en entrée de notre modèle de coarticulation. Pour reprendre l'exemple du segment /utu/, cette méthode nous permettrait de fournir des informations différentes au réseau lors de l'utilisation d'un /t/ turque ou anglais, lui permettant éventuellement de différencier les deux stratégies de coarticulation.

Un point grandement mis en avant durant cette thèse est l'utilisation de notre procédure d'injection de connaissances, permettant de fournir au modèle une représentation des gestes articulatoires compatibles avec le réseau de neurones. Cette approche décompose le réseau en deux grands modules, un premier composé de couches récurrentes, et un second module initialisé grâce à une méthode de réduction de la dimensionnalité. Cette méthode peut-être interprétée comme une architecture de réseaux de neurones basée sur les principes de la phonétique articulatoire, comme abordée dans la discussion du chapitre 2. L'apport de cette méthode semble se concrétiser en trois points particuliers soulevés au chapitre 4 :

— Il diminue drastiquement le temps d'apprentissage nécessaire au modèle,
— Plus le nuage de points représentant l'espace articulatoire est précis (c'est-à-dire plus le nombre de points du nuage augmente), et plus l'impact de notre procédure sur les performances est importante,
— Les *features* apprises au sein des couches récurrentes sont plus indépendantes du locuteur, car les caractéristiques de ce dernier sont essentiellement capturées dans les gestes articulatoires.

Ce troisième point est particulièrement intéressant, car les couches récurrentes entrainées en conjonction avec l'espace articulatoire latent pré-calculé semblent être plus facilement transférables d'un locuteur à l'autre, mais également d'une modalité articulatoire à une autre. Dans nos expériences, cela se traduit par un passage du domaine visuel au domaine

articulatoire plus aisé (cf. section 4.4. À ces trois points, nous pouvons également ajouter la certitude que cet espace articulatoire latent n'est que très légèrement modifié par l'apprentissage (section 5.3.1), ce qui nous confirme que le réseau exploite bel et bien les gestes articulatoires injectés avant l'apprentissage. Cette particularité nous semble être un avantage de choix à exploiter pour l'extension de notre modèle à un contexte multilocuteurs.

Malheureusement, les réseaux de neurones souffrent encore d'un problème d'interprétabilité, et notre modèle ne fait pas exception. La majorité des méthodes permettant d'expliquer le résultat obtenu par un réseau de neurones furent développées pour des problèmes de classification, et permettent par exemple de fournir une *heatmap* précisant quelles sont les entrées ayant le plus contribué à sa classification. Cela peut donc être une analyse au niveau du pixel pour des images (Bach et al., 2015; Landecker et al., 2013; Simonyan et al., 2013), ou des mots pour du texte (Li et al., 2015; Arras et al., 2017). Malheureusement, l'application de ces méthodes à des problèmes de régression et à des réseaux de neurones récurrents est à notre connaissance un domaine encore embryonnaire. Le problème d'explainabilité donc semble être un candidat de choix pour de futures études. De plus, la capacité d'expliquer l'influence de l'entrée sur les résultats de la régression pourrait s'appliquer à d'autres problèmes impliquant des séquences temporelles, comme la prédiction du marché boursier (Selvin et al., 2017; Nelson et al., 2017). Les méthodes évoquées au paragraphe précédent semblent être particulièrement prometteuses, mais de légers changements dans l'architecture du réseau de neurones pourraient également aider à augmenter l'explainabilité du modèle. Nous pensons plus particulièrement à l'insertion d'un mécanisme d'attention au niveau des couches récurrentes de notre réseau, d'une manière similaire aux travaux de Bahdanau et al. (2015) sur la traduction automatique. Nous espérons que cette méthode permettrait de visualiser simplement où se situe l'information nécessaire à la prédiction, par le biais des scores d'attention, en mettant ainsi en lumière l'influence respective de chaque phonème sur la position des articulateurs.

En plus de ces problématiques d'interprétabilité, deux perspectives supplémentaires de travaux à plus long terme nous semblent envisageables. Premièrement, nous pouvons souligner l'aspect déterministe de notre modèle, se traduisant par l'impossibilité de générer deux articulations différentes depuis une unique séquence phonétique. Chez l'humain, de subtiles variations peuvent se produire lors de la production de la même séquence phonétique, variations que notre modèle n'est pas en capacité de modéliser. Aller d'un modèle prédictif (modélisant $P(y|x)$) à un modèle génératif (modélisant $P(x \wedge y)$) nous semble être une approche envisageable, en exploitant par exemple les récents progrès sur les réseaux adversariaux. En plus de gérer ces variations articulatoires stochastiques, il nous semble

être une extension logique de notre modèle que de prendre en compte les variations articulatoires induites par l'état émotionel du locteur. Secondement, nous pouvons définir une nouvelle façon de faire de l'inversion acoustique, en couplant notre modèle de coarticulation à un modèle acoustique, et en utilisant la modalité phonétique comme point de pivot entre les deux modèles. A la place de considérer le phonème d'entrée comme un vecteur *one-hot*, il nous semble intéressant d'expérimenter autour de l'utilisation d'un vecteur de probabilités tel que fournit par le modèle acoustique. Contrairement aux approches actuelles en inversion acoustique, cette modification de notre modèle de coarticulation permettrait donc de prédire l'articulation depuis la parole de n'importe quel locuteur, tout en conservant un corpus articulatoire monolocuteur.

Table des figures

Liste des tableaux

Bibliographie

Abry, C. and Lallouache, T. (1995). Le mem : un modèle d'anticipation paramétrable par locuteur : Données sur l'arrondissement en français. *Les Cahiers de l'ICP. Bulletin de la communication parlée*, (3) :85–99.

Al-Bamerni, A. and Bladon, A. (1982). One-stage and two-stage temporal patterns of velar coarticulation. *The Journal of the Acoustical Society of America*, 72(S1) :S104–S104.

Arras, L., Horn, F., Montavon, G., Müller, K.-R., and Samek, W. (2017). " what is relevant in a text document ?" : An interpretable machine learning approach. *PloS one*, 12(8) :e0181142.

Atal, B. S., Chang, J. J., Mathews, M. V., and 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.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv :1607.06450*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7) :e0130140.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bailly, G. and Badin, P. (2002). Seeing tongue movements from outside. In *Seventh International Conference on Spoken Language Processing*.

Bailly, G., Gibert, G., and Odisio, M. (2002). Evaluation of movement generation systems using the point-light technique. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 27–30. IEEE.

Bandini, A., Ouni, S., Cosi, P., Orlandi, S., and Manfredi, C. (2015). Accuracy of a markerless acquisition technique for studying speech articulators. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Bell-Berti, F. (1980). Velopharyngeal function : A spatial–temporal model. In *Speech and language*, volume 4, pages 291–316. Elsevier.

Bell-Berti, F. and Harris, K. S. (1981). A temporal model of speech production. *Phonetica*, 38(1-3) :9–20.

Bell-Berti, F. and Krakow, R. A. (1991). Anticipatory velar lowering : A coproduction account. *The Journal of the Acoustical Society of America*, 90(1) :112–123.

Ben Youssef, A., Badin, P., and Bailly, G. (2010). Can tongue be recovered from face ? The answer of data-driven statistical models. In *Interspeech*, Makuhari, Japan.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5 :157–166.

Benguérel, A.-P. and Cowan, H. A. (1974). Coarticulation of upper lip protrusion in french. *Phonetica*, 30(1) :41–55.

Benguerel, A.-P., Hirose, H., Sawashima, M., and Ushijima, T. (1977). Velar coarticulation in french : a fiberscopic study. *Journal of phonetics*, 5(2) :149–158.

Benoit, C., Lallouache, T., Mohamadi, T., and Abry, C. (1992). A set of french visemes for visual speech synthesis.

Benoit, M. M., Raij, T., Lin, F.-H., Jääskeläinen, I. P., and Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human brain mapping*, 31(4) :526–538.

Beskow, J. (1995). Rule-based visual speech synthesis. In *Fourth European Conference on Speech Communication and Technology*.

Beskow, J. (2004). Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology*, 7(4) :335–349.

Biasutto-Lervat, T., Dahmani, S., and Ouni, S. (2019). Modeling labial coarticulation with bidirectional gated recurrent networks and transfer learning. In *Interspeech 2019*, pages 2608–2612.

Biasutto-Lervat, T. and Ouni, S. (2018). Phoneme-to-articulatory mapping using bidirectional gated rnn. In *Interspeech 2018*, pages 3112–3116.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1) :108–116.

Bladon, R. A. W. and Al-Bamerni, A. (1976). Coarticulation resistance in english/l. *Journal of Phonetics*, 4(2) :137–150.

Boë, L.-J., Perrier, P., and Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production : proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20(1) :27–38.

Boyce, S. E. (1990). Coarticulatory organization for lip rounding in turkish and english. *The Journal of the Acoustical Society of America*, 88(6) :2584–2595.

Breen, A., Bowers, E., and Welsh, W. (1996). An investigation into the generation of mouth shapes for a talking head. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2159–2162. IEEE.

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite : Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360.

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6(2) :201–251.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology : An overview. *Phonetica*, 49(3-4) :155–180.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem ?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030.

Butcher, A. and Weiher, E. (1976). An electropalatographic investigation of coarticulation in vcv sequences. *Journal of Phonetics*, 4(1) :59–74.

Campbell, R., Burnham, D., Dodd, B., Campbell, R., Away, G., and Burnham, D. K. (1998). *Hearing by eye II : Advances in the psychology of speechreading and auditory-visual speech*, volume 2. Psychology Press.

Cao, Y., Faloutsos, P., Kohler, E., and Pighin, F. (2004). Real-time speech motion synthesis from recorded motions. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 345–353.

Cathiard, M., Tiberghien, G., Cirot-Tseva, A., Lallouache, M., and Escudier, P. (1991). Visual perception of anticipatory rounding during acoustic pauses : A cross-language study. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, volume 4, pages 50–53. Aix-en-Provence.

Charpentier, F. (1984). Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic nonlinearities. *Speech Communication*, 3(4) :291–308.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chuang, E. and Bregler, C. (2002). Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report, Stanford University*, 2(2) :3.

Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation*, pages 139–156. Springer.

Colotte, V. and Lafosse, A. (2009). Soja : French text-to-speech synthesis system.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6) :681–685.

Cosatto, E. and Graf, H. P. (2000). Photo-realistic talking-heads from image samples. *IEEE Transactions on multimedia*, 2(3) :152–163.

Cosatto, E., Ostermann, J., Graf, H. P., and Schroeter, J. (2003). Lifelike talking faces for interactive services. *Proceedings of the IEEE*, 91(9) :1406–1429.

Cosi, P., Caldognetto, E. M., Perin, G., and Zmarich, C. (2002). Labial coarticulation modeling for realistic facial animation. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 505–510. IEEE.

Curio, C., Breidt, M., Kleiner, M., Vuong, Q. C., Giese, M. A., and Bülthoff, H. H. (2006). Semantic 3d motion retargeting for facial animation. In *Proceedings of the 3rd symposium on Applied perception in graphics and visualization*, pages 77–84.

Dahmani, S., Colotte, V., Girard, V., and Ouni, S. (2019). Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. In *INTERSPEECH*, pages 2598–2602.

Daniloff, R. and Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research*, 11(4) :707–721.

Daniloff, R. G. and Hammarberg, R. E. (1973). On defining coarticulation. *Journal of Phonetics*, 1(3) :239–248.

Davis, B. L. and MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research*, 38(6) :1199–1211.

Dehn, D. M. and Van Mulken, S. (2000). The impact of animated interface agents : a review of empirical research. *International journal of human-computer studies*, 52(1) :1–22.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Dixon, N. F. and Spitz, L. (1980). The detection of audiovisual desynchrony. *Perception*, 9 :719–721.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.*, 12 :2121–2159.

Edge, J. and Hilton, A. (2006). Visual speech synthesis from 3d video. In *IET European Conference on Visual Media Production*, pages 174–174.

Edwards, G. J., Taylor, C. J., and Cootes, T. F. (1998). Interpreting face images using active appearance models. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 300–305. IEEE.

Edwards, P., Landreth, C., Fiume, E., and Singh, K. (2016). Jali : an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)*, 35(4) :1–11.

Elisei, F., Odisio, M., Bailly, G., and Badin, P. (2001). Creating and controlling video-realistic talking heads. In *AVSP*, pages 90–97. Citeseer.

Engwall, O. (2002). Evaluation of a system for concatenative articulatory visual speech synthesis. In *Seventh International Conference on Spoken Language Processing*.

Engwall, O. and Beskow, J. (2003). Resynthesis of 3d tongue movements from facial data. In *Eighth European Conference on Speech Communication and Technology*.

Ezzat, T. and Poggio, T. (2000). Visual speech synthesis by morphing visemes. *International Journal of Computer Vision*, 38(1) :45–57.

Fagel, S. (2006). Joint audio-visual units selection — the JAVUS speech synthesizer. In *International Conference on Speech and Computer*, St. Petersburg, Russia.

Fan, B., Wang, L., Soong, F. K., and Xie, L. (2015). Photo-real talking head with deep bidirectional lstm. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4884–4888. IEEE.

Fan, B., Xie, L., Yang, S., Wang, L., and Soong, F. K. (2016). A deep bidirectional LSTM approach for video-realistic talking head. *Multimedia Tools and Applications*, 75 :5287–5309.

Fant, G. (1970). Analytical constraints on the composition of speech spectra. *Acoustic Theory of Speech Production, Second Printing*, pages 48–62.

Farnetani, E. (1990). Vcv lingual coarticulation and its spatiotemporal domain. In *Speech production and speech modelling*, pages 93–130. Springer.

Farnetani, E. and Busa, M. G. (1994). Italian clusters in continuous speech. In *Third International Conference on Spoken Language Processing*.

Farnetani, E., Vagges, K., and Magno-Caldognetto, E. (1985). Coarticulation in italian/vtv/sequences : a palatographic study. *Phonetica*, 42(2-3) :78–99.

Feldman, A., Adamovich, S., Ostry, D., and Flanagan, J. (1990). The origin of electromyograms—explanations based on the equilibrium point hypothesis. In *Multiple muscle systems*, pages 195–213. Springer.

Fowler, C. A. (1977). *Timing control in speech production*, volume 134. Indiana University Linguistics Club.

Friesen, E. and Ekman, P. (1978). Facial action coding system : a technique for the measurement of facial movement. *Palo Alto*, 3.

Gelfer, C. E., Bell-Berti, F., and Harris, K. S. (1989). Determining the extent of coarticulation : Effects of experimental design. *The Journal of the Acoustical Society of America*, 86(6) :2443–2445.

Gibbs, S., Breiteneder, C., De Mey, V., and Papathomas, M. (1993). Video widgets and video actors. In *Proceedings of the 6th annual ACM symposium on User interface software and technology*, pages 179–185.

Gilbert, J. H. (1972). *Speech and cortical functioning*. Academic Press New York.

Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., Nieto-Castanon, A., and Guenther, F. H. (2011). fmri investigation of unexpected somatosensory feedback perturbation during speech. *Neuroimage*, 55(3) :1324–1338.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Govokhina, O. (2008). *Modèles de trajectoires pour l'animation de visages parlants*. PhD thesis, Thèse de l'Institut National Polytechnique de Grenoble.

Goyal, U. K., Kapoor, A., and Kalra, P. (2000). Text-to-audiovisual speech synthesizer. In *International Conference on Virtual Worlds*, pages 256–269. Springer.

Granström, B. and House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech communication*, 46(3-4) :473–484.

Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., McTait, K., and Choukri, K. (2004). The ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*.

Green, K. P. and Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics*, 45 :34–42.

Green, K. P. and Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology : Human Perception and Performance*, 17 :278–288.

Greenwood, A., Goodyear, C., and Martin, P. (1992). Measurements of vocal tract shapes using magnetic resonance imaging. *IEE Proceedings I (Communications, Speech and Vision)*, 139(6) :553–560.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological cybernetics*, 72(1) :43–53.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, 96(3) :280–301.

Hallgren, A. and Lyberg, B. (1998). Visual speech synthesis with concatenative speech. In *AVSP*, Terrigal-Sydney, Australia.

Hammarberg, R. (1976). The metaphysics of coarticulation. *Journal of Phonetics*, 4(4) :353–363.

Hazan, V., Sennema, A., Iba, M., and Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by japanese learners of english. *Speech communication*, 47(3) :360–378.

Heinz, J. M. (1965). On the relations between lateral cineradiographs area functions and acoustic spectra of speech. *Proc. 5th Int. Congr. Acoust. Liege, 1965*.

Henke, W. L. (1966). Dynamic articulatory model of speech production using computer simulation.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786) :504–507.

Hiroya, S. and Honda, M. (2004). Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Speech and Signal Processing*, 12(2) :175–185.

Hixon, T. J. (1971). An electromagnetic method for transducing jaw movements during speech. *The Journal of the Acoustical Society of America*, 49(2B) :603–606.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9 :1735–1780.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., and Saltzman, E. (1996). Accurate recovery of articulator positions from acoustics : New conclusions based on human data. *The Journal of the Acoustical Society of America*, 100(3) :1819–1834.

Holmstrom, L. and Koistinen, P. (1992). Using additive noise in back-propagation training. *IEEE transactions on neural networks*, 3(1) :24–38.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2 :359–366.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6) :417.

Houde, J. F. and Chang, E. F. (2015). The cortical computations underlying feedback control in vocal production. *Current opinion in neurobiology*, 33 :174–181.

Houde, J. F. and Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in human neuroscience*, 5 :82.

Houde, J. F., Niziolek, C., Kort, N., Agnew, Z., and Nagarajan, S. S. (2014). Simulating a state feedback model of speaking. In *10th International Seminar on Speech Production*, volume 202, page 205.

Hueber, T., Ben-Youssef, A., Bailly, G., Badin, P., and Elisei, F. (2012). Cross-speaker acoustic-to-articulatory inversion using phone-based trajectory hmm for pronunciation training. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Jessen, M. (1998). *Phonetics and phonology of tense and lax obstruents in German*. Benjamins Amsterdam.

Jiang, J., Alwan, A., Keating, P., Auer, E., and Bernstein, L. (2002a). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing*, 11 :1174–1188.

Jiang, J., Alwan, A., Keating, P. A., Auer, E. T., and Bernstein, L. E. (2002b). On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Applied Signal Processing*, 11 :1174–1188.

Johnson, W. L., Rickel, J. W., Lester, J. C., et al. (2000). Animated pedagogical agents : Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in education*, 11(1) :47–78.

Joos, M. (1948). Acoustic phonetics.(linguistic society of america, language monograph 23.) baltimore.

Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36(4) :94 :1–94 :12.

Keating, P. A. et al. (1988). The window model of coarticulation : articulatory evidence. *UCLA Working papers in Phonetics*, 69 :3–29.

Kingma, D. P. and Ba, J. (2015). Adam : A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv :1906.02691*.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2) :233–243.

Kuratate, T., Yehia, H., and Vatikiotis-Bateson, E. (1998). Kinematics-based synthesis of realistic talking faces. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*.

Laboissière, R. (1992). *Préliminaires pour une robotique de la communication parlée : inversion et contrôle d'un modèle articulatoire du conduit vocal*. PhD thesis, Grenoble INPG.

Landecker, W., Thomure, M. D., Bettencourt, L. M., Mitchell, M., Kenyon, G. T., and Brumby, S. P. (2013). Interpreting individual classifications of hierarchical networks. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 32–38. IEEE.

Larar, J. N., Schroeter, J., and Sondhi, M. M. (1988). Vector quantization of the articulatory space. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(12) :1812–1818.

Le Goff, Guiard-marigny, T., Cohen, M., and Benoit, C. (1994). Real-time analysis-synthesis and intelligibility of talking faces. In *In 2nd International conference on Speech Synthesis*, pages 53–56.

Le Goff, B. (1997). Automatic modeling of coarticulation in text-to-visual speech synthesis. In *Fifth European Conference on Speech Communication and Technology*.

Levelt, W. J. (1993). *Speaking : From intention to articulation*, volume 1. MIT press.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv :1506.01066*.

Liu, P., Yu, Q., Wu, Z., Kang, S., Meng, H., and Cai, L. (2015). A deep recurrent approach for acoustic-to-articulatory inversion. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4450–4454. IEEE.

Löfqvist, A. (1990). *Speech as Audible Gestures*, pages 289–322. Springer Netherlands, Dordrecht.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov) :2579–2605.

MacLean, K. (2018). Voxforge. *Ken MacLean.[Online]. Available : http ://www. voxforge. org/home.[Acedido em 2012]*.

MacSweeney, M., Calvert, G. A., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., Woll, B., and Brammer, M. J. (2002). Speechreading circuits in people born deaf. *Neuropsychologia*, 40(7) :801–807.

Maeda, S. (1990). Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. J. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic.

Magen, H. S. (1989). *An acoustic study of vowel-to-vowel coarticulation in English*. na.

Mak, B. and Barnard, E. (1996). Phone clustering using the bhattacharyya distance. In *Fourth International Conference on Spoken Language Processing*.

Massaro, D. W. (2003). A computer-animated tutor for spoken and written language learning. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 172–175.

Masters, D. and Luschi, C. (2018). Revisiting small batch training for deep neural networks. *arXiv preprint arXiv :1804.07612*.

Mattheyses, W. and Verhelst, W. (2015). Audiovisual speech synthesis : An overview of the state-of-the-art. *Speech Communication*, 66 :182–217.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264 :746–748.

Menzerath, P. and de Lacerda, A. (1933). *Koartikulation, stenerung and lautabgrenzung*, volume 19. F Dummler.

Mermelstein, P. (1967). Determination of the vocal-tract shape from measured formant frequencies. *The Journal of the Acoustical Society of America*, 41(5) :1283–1294.

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

Minnis, S. and Breen, A. (2000). Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *Interspeech*, Bejing, China.

Moll, K. L. and Daniloff, R. G. (1971). Investigation of the timing of velar movements during speech. *The Journal of the Acoustical Society of America*, 50(2B) :678–684.

Munhall, K. and Löfqvist, A. (1992). Gestural aggregation in speech : Laryngeal gestures. *Journal of Phonetics*, 20(1) :111–126.

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. (2015). Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv :1511.06807*.

Nelson, D. M., Pereira, A. C., and de Oliveira, R. A. (2017). Stock market's price movement prediction with lstm neural networks. In *2017 International joint conference on neural networks (IJCNN)*, pages 1419–1426. IEEE.

Ohala, J. J., Browman, C. P., and Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, 3 :219–252.

Öhman, S. E. (1966). Coarticulation in vcv utterances : Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1) :151–168.

Öhman, S. E. (1967). Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, 41(2) :310–320.

Ostermann, J. (1998). Animation of synthetic faces in mpeg-4. In *Proceedings Computer Animation'98 (Cat. No. 98EX169)*, pages 49–55. IEEE.

Ostermann, J. and Millen, D. (2000). Talking heads and synthetic speech : An architecture for supporting electronic commerce. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 1, pages 71–74. IEEE.

Ouni, S., Cohen, M. M., Ishak, H., and Massaro, D. W. (2007). Visual contribution to speech perception : measuring the intelligibility of animated talking heads. *EURASIP J. Audio Speech Music Process.*, 2007(1) :3–3.

Ouni, S., Colotte, V., Musti, U., Toutios, A., Wrobel-Dautcourt, B., Berger, M.-O., and Lavecchia, C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1) :16.

Ouni, S. and Dahmani, S. (2016). Is markerless acquisition technique adequate for speech production ? *The Journal of the Acoustical Society of America*, 139(6) :EL234–EL239.

Ouni, S. and Laprie, Y. (2005a). 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.

Ouni, S. and Laprie, Y. (2005b). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of Acoustic Society of America*, 118(1) :444–460.

Pakstas, A., Forchheimer, R., and Pandzic, I. S. (2002). *MPEG-4 Facial Animation : The Standard, Implementation and Applications*. John Wiley & Sons, Inc.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10) :1345–1359.

Pandzic, I. S., Ostermann, J., and Millen, D. (1999). User evaluation : Synthetic talking faces for interactive services. *The Visual Computer*, 15(7) :330–340.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., and Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *The Journal of the Acoustical Society of America*, 92(2) :688–700.

Parke, F. I. (1982). Parameterized models for facial animation. *IEEE computer graphics and applications*, (9) :61–68.

Parrell, B., Lammert, A. C., Ciccarelli, G., and Quatieri, T. F. (2019a). Current models of speech motor control : A control-theoretic overview of architectures and properties. *The Journal of the Acoustical Society of America*, 145(3) :1456–1481.

Parrell, B., Ramanarayanan, V., Nagarajan, S., and Houde, J. (2019b). The facts model of speech motor control : Fusing state estimation and task-based control. *PLoS computational biology*, 15(9) :e1007321.

Parrell, B., Ramanarayanan, V., Nagarajan, S. S., and Houde, J. F. (2018). Facts : A hierarchical task-based control model of speech incorporating sensory feedback. In *Interspeech*, pages 1497–1501.

Patri, J.-F. (2018). *Bayesian modeling of speech motor planning : variability, multisensory goals and perceptuo-motor interactions*. PhD thesis.

Payan, Y. and 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.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) :559–572.

Pelachaud, C., Badler, N. I., and Steedman, M. (1991). Linguistic issues in facial animation. In *Computer Animation'91*, pages 15–30. Springer.

Perkel, J. (1986). Preliminary support for a hybrid model of anticipatory coarticulation. In *Proceedings of the 12th International Conference of Acoustics*.

Perkell, J. S. and Matthies, M. L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel/u : Within-and cross-subject variability. *The Journal of the Acoustical Society of America*, 91(5) :2911–2925.

Perrier, P. and Fuchs, S. (2008). Speed–curvature relations in speech production challenge the 1/3 power law. *Journal of neurophysiology*, 100(3) :1171–1183.

Perrier, P., Ma, L., and Payan, Y. (2006). Modeling the production of vcv sequences via the inversion of a biomechanical model of the tongue. *arXiv preprint physics/0610170*.

Perrier, P., Ostry, D. J., and Laboissière, R. (1996). The equilibrium point hypothesis and its application to speech motor control. *Journal of Speech, Language, and Hearing Research*, 39(2) :365–378.

Perrier, P., Payan, Y., Zandipour, M., and 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.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv :1802.05365*.

Pham, H. X., Wang, Y., and Pavlovic, V. (2018a). End-to-end learning for 3d facial animation from speech. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 361–365, New York, NY, USA. ACM.

Pham, H. X., Wang, Y., and Pavlovic, V. (2018b). Generative adversarial talking head : Bringing portraits to life with a weakly supervised neural network. *CoRR*, abs/1803.07716.

Pighin, F. and Lewis, J. P. (2006). Facial motion retargeting. In *ACM SIGGRAPH 2006 Courses*, pages 2–es.

Potard, B. and Laprie, Y. (2007). Compact representations of the articulatory-to-acoustic mapping. In *Eighth Annual Conference of the International Speech Communication Association*.

Potard, B. and Laprie, Y. (2009). A robust variational method for the acoustic-to-articulatory problem. In *Tenth Annual Conference of the International Speech Communication Association*.

Ribera, R. B. i., Zell, E., Lewis, J., Noh, J., and Botsch, M. (2017). Facial retargeting with automatic range of motion alignment. *ACM Transactions on Graphics (TOG)*, 36(4) :1–12.

Richmond, K. (2006). A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Ninth International Conference on Spoken Language Processing*.

Richmond, K. (2009). Preliminary inversion mapping results with a new ema corpus. In *Tenth Annual Conference of the International Speech Communication Association*.

Richmond, K., Hoole, P., and King, S. (2011). Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Twelfth Annual Conference of the International Speech Communication Association.*

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., and Browman, C. (1996). Casy and extensions to the task-dynamic model. In *1st ETRW on Speech Production Modeling : From Control Strategies to Acoustics ; 4th Speech Production Seminar : Models and Data.*

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. pages 318–362. MIT Press.

Sadoughi, N. and Busso, C. (2019). Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing.*

Saltzman, E. (1986). Task dynamic coordination of the speech articulators : A preliminary model. *Experimental brain research series*, 15 :129–144.

Saltzman, E. (1991). The task dynamic model in speech production. *Speech motor control and stuttering*, pages 37–52.

Schoentgen, J. and Ciocca, S. (1995). Direct calculation of the vocal tract area function from measured formant frequencies. In *Fourth European Conference on Speech Communication and Technology.*

Schoentgen, J. and Ciocea, S. (1997). Kinematic formant-to-area mapping. *Speech communication*, 21(4) :227–244.

Schroeter, J. and Sondhi, M. M. (1992). Speech coding based on physiological models of speech production. *Advances in Speech Signal Processing*, pages 231–267.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone : A multilingual text & speech database in 20 languages. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8126–8130. IEEE.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45 :2673–2681.

Scott, K. C., Kagels, D., Watson, S., Rom, H., Wright, J., Lee, M., and Hussey, K. (1994). Synthesis of speaker facial movement to match selected speech sequences.

Scripture, E. W. (1904). *The elements of experimental phonetics.* C. Scribner's Sons.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K., and Soman, K. (2017). Stock price prediction using lstm, rnn and cnn-sliding window model. In *2017 international conference on advances in computing, communications and informatics (icacci)*, pages 1643–1647. IEEE.

Shahrebabaki, A. S., Olfati, N., Imran, A. S., Siniscalchi, S. M., and Svendsen, T. (2019). A phonetic-level analysis of different input features for articulatory inversion. *Proc. Interspeech 2019*, pages 3775–3779.

Siegelmann, H. and Sontag, E. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, 50 :132–150.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks : Visualising image classification models and saliency maps. *arXiv preprint arXiv :1312.6034*.

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices : how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17(10) :2387–2399.

Song, J., Choi, B., Seol, Y., and Noh, J. (2011). Characteristic facial retargeting. *Computer Animation and Virtual Worlds*, 22(2-3) :187–194.

Song, Y., Zhu, J., Wang, X., and Qi, H. (2018). Talking face generation by conditional recurrent adversarial network. *CoRR*, abs/1804.04786.

Soquet, A., Saerens, M., and Jospa, P. (1991). Acoustic-articulatory inversion based on a neural controller of a vocal tract model : further results. *Artificial Neural Networks*, pages 371–376.

Sorokin, V. N. and Trushkin, A. V. (1996). Articulatory-to-acoustic mapping for inverse problem. *Speech Communication*, 19(2) :105–118.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., and Waters, K. (1996). When the interface is a face. *Human-computer interaction*, 11(2) :97–124.

Stevens, K. N. and House, A. S. (1955). Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America*, 27(3) :484–493.

Sumby, W. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26 :212.

Sussman, H. M. and Westbury, J. R. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech, Language, and Hearing Research*, 24(1) :16–24.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama : Learning lip sync from audio. *ACM Trans. Graph.*, 36 :95 :1–95 :13.

Swerts, M. and Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1) :81–94.

Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., Hodgins, J., and Matthews, I. (2017). A deep learning approach for generalized speech animation. *ACM Trans. Graph.*, 36(4) :93 :1–93 :11.

Taylor, S. L., Mahler, M., Theobald, B.-J., and Matthews, I. (2012). Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face : Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395.

Tiddeman, B. and Perrett, D. (2002). Prototyping and transforming visemes for animated speech. In *Proceedings of Computer Animation 2002 (CA 2002)*, pages 248–251. IEEE.

Tieleman, T. and Hinton, G. (2012). RMSprop Gradient Optimization.

Toda, T., Black, A., and Tokuda, K. (2004). Acoustic-to-articulatory inversion mapping with gaussian mixture model. In *Eighth International Conference on Spoken Language Processing*.

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

Tourville, J. A. and Guenther, F. H. (2011). The diva model : A neural theory of speech acquisition and production. *Language and cognitive processes*, 26(7) :952–981.

Toutios, A., Ouni, S., et al. (2011). Predicting tongue positions from acoustics and facial features. In *12th Annual Conference of the International Speech Communication Association-Interspeech 2011*.

Trouvain, J., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., Möbius, B., et al. (2016). The ifcasl corpus of french and german non-native and native read speech.

Uria, B., Murray, I., Renals, S., and Richmond, K. (2012). Deep architectures for articulatory inversion. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Ushijima, T. (1972). Fiberscopic observation of velar movements during speech. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 6 :25–38.

Verma, A., Rajput, N., and Subramaniam, L. V. (2003). Using viseme based acoustic models for speech driven lip synthesis. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 5, pages V–720. IEEE.

Vougioukas, K., Petridis, S., and Pantic, M. (2018). End-to-end speech-driven facial animation with temporal gans. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 133.

Walker, J. H., Sproull, L., and Subramani, R. (1994). Using a human face in an interface. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 85–91.

Westbury, J., Milenkovic, P., Weismer, G., and Kent, R. (1990). X-ray microbeam speech production database. *The Journal of the Acoustical Society of America*, 88(S1) :S56–S56.

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, 18(1) :3–35.

Wilson, D. R. and Martinez, T. R. (2003). The general inefficiency of batch training for gradient descent learning. *Neural networks*, 16(10) :1429–1451.

Wrench, A. and Hardcastle, W. (2000). A multichannel articulatory database and its application for automatic speech recognition. In *Proc. 5th International Seminar on Speech Production*, pages 205–308, Kloster Seeon, Bavaria.

Wu, Z., Zhao, K., Wu, X., Lan, X., and Meng, H. (2015). Acoustic to articulatory mapping with deep neural network. *Multimedia Tools and Applications*, 74(22) :9889–9907.

Xie, X., Liu, X., and Wang, L. (2016). Deep neural network based acoustic-to-articulatory inversion using phone sequence information. In *Interspeech*, pages 1497–1501.

Xie, X., Liu, X., Wang, L., and Su, R. (2015). Generalized variable parameter hmms based acoustic-to-articulatory inversion. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2) :23–43.

Youssef, A. B., Badin, P., Bailly, G., and Heracleous, P. (2009). Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden markov models. In *Tenth Annual Conference of the International Speech Communication Association.*

Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468.

Zell, E., Lewis, J., Noh, J., Botsch, M., et al. (2017). Facial retargeting with automatic range of motion alignment. *ACM Transactions on Graphics (TOG)*, 36(4) :154.

Zhang, L. and Renals, S. (2008). Acoustic-articulatory modeling with the trajectory HMM. *Signal Processing Letters*, 15 :245–248.

Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. (2018a). Talking face generation by adversarially disentangled audio-visual representation. *CoRR*, abs/1807.07860.

Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., and Singh, K. (2018b). Visemenet : Audio-driven animator-centric speech animation. *ACM Trans. Graph.*, 37(4) :161 :1–161 :10.

Zhu, P., Xie, L., and Chen, Y. (2015). Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings. In *Sixteenth Annual Conference of the International Speech Communication Association.*