Modélisation de la coarticulation labiale : mise en oeuvre sur une tête parlante

Vincent Robert

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

Modélisation de la coarticulation labiale : *Mise en œuvre sur une tête parlante*

THÈSE

présentée et soutenue publiquement le 12 novembre 2008

pour l'obtention du

Doctorat de l'Université Henri Poincaré – Nancy I

(Spécialité Informatique)

par

Vincent ROBERT

Composition du jury

Rapporteurs : L. Besacier, Maître de Conférence à l'Université Joseph Fourier de Grenoble.
O. Boëffard, Professeur à l'ENSSAT de Lannion.

Examinateurs : R. Sock, Professeur à l'Université Marc Bloch de Strasbourg.
Y. Laprie, Directeur de Recherche au CNRS - LORIA.
A. Bonneau, Chargée de Recherche au CNRS - LORIA.
J. P. Haton, Professeur à l'Université Henri Poincaré de Nancy.

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

Mis en page avec la classe thloria.

Remerciements

Je tiens tout d'abord à remercier Dominique MERY qui le premier m'a ouvert les portes de l'Ecole Doctorale ainsi que Nathalie PARLANGEAU qui, en co-encadrant mon stage de DEA, m'a permis de découvrir et d'apprécier l'univers de la recherche.

Je remercie vivement Yves LAPRIE et Anne BONNEAU d'avoir accepté d'encadrer ma thèse, même s'ils savaient que ce ne serait pas facile pour moi de conjuguer ma profession d'enseignant à temps plein avec celle de doctorant. Ils ont toujours su me motiver dans les moments difficiles et accepté que mes travaux de recherche prennent quelquefois davantage de temps que prévu. Leurs qualités d'écoute et la sympathie qu'ils m'ont témoignée ont conforté mon envie d'aller jusqu'au bout de cette aventure.

Je remercie aussi Brigitte WROBEL sans qui les corpus n'auraient peut-être jamais vu le jour. En passant beaucoup de temps au dépouillement des acquisitions vidéos, elle m'a permis de travailler sur des données d'excellente qualité. De plus, sa bonne humeur contagieuse a complètement occulté le côté fastidieux de l'enregistrement des corpus.

J'ai aussi une pensée toute particulière pour Jacques FELDMAR qui a consacré un temps considérable pour adapter à notre corpus une méthode spécifique de prédiction de la coarticulation. De plus, il est l'auteur du logiciel qui permet de synthétiser une tête parlante à partir de nos paramètres articulatoires, c'est-à-dire d'un élément indispensable aux tests de perception.

Bien évidemment, je remercie très sincèrement tous les membres du LORIA avec qui j'ai eu des contacts, particulièrement mes compagnons de bureau, Armelle et Christophe, qui m'ont beaucoup apporté sur le plan humain et scientifique ainsi que Marie Odile, Blaise, Slim et Martine C dont les conseils avisés m'ont aidé à résoudre certains problèmes épineux. Je n'oublie pas Martine K qui a toujours très gentiment répondu à mes sollicitations, même quand elle était débordée.

Je remercie enfin vivement mon épouse qui m'a soutenu pendant toutes ces années.

A ma fille Marion

Table des matières

Table des figures

Introduction générale

Pour comprendre la parole, l'entendre ne suffit pas toujours. Dans certains cas, le fait de de voir le visage de son interlocuteur peut grandement faciliter la compréhension. Des études perceptives ont montré que les informations visuelles apportées par le visage de l'interlocuteur dans le cadre de conditions de communication dégradées contribuent largement à l'intelligibilité de la parole. Cette thèse s'inscrit dans le cadre d'un projet qui consiste à développer une tête parlante destinée à améliorer la compréhension des sourds et des malentendants grâce à la lecture labiale. Au delà de cette application, le visage parlant peut aussi être utilisé pour rendre plus attractives les communications sur Internet ou d'autres supports, pour faciliter l'apprentissage des langues étrangères, etc. Mais c'est aussi dans le cadre de recherches fondamentales en parole que le visage parlant se révèle particulièrement intéressant car il peut permettre par "feed back" de vérifier certaines théories articulatoires.

Notre but n'est pas de créer une tête parlante réaliste capable de mystifier l'interlocuteur en lui faisant croire que le visage est réel. Ce type d'application, davantage réservé au monde du cinéma ou aux films d'animation ne permet pas à l'heure actuelle un codage de la parole en temps réel. Au contraire, la multitude d'informations à considérer nécessite un long traitement avec même quelquefois des retouches manuelles pour avoir un rendu parfait.

Notre but est d'obtenir un visage parlant dont les mouvements de la bouche sont en parfaite adéquation avec le signal sonore. A cette fin, il est nécessaire de prendre en considération la coarticulation, c'est-à-dire l'influence des sons adjacents sur le son en cours. Ce phénomène peut être anticipatoire (régressive) quand un son est influencé par le son suivant ou rétentif (progressive) dans le cas contraire.

Nous verrons dans le premier chapitre que la coarticulation est beaucoup plus complexe qu'un simple phénomène inertiel ou qu'une assimilation phonétique. L'assimilation, qui peut être progressive ou régressive désigne une modification pouvant affecter l'identité d'un son. Dans une séquence donnée, il y a assimilation quand le phonème le plus fort transmet partiellement ou totalement une ou plusieurs de ses particularités phonétiques à ses voisins. Par exemple, le mot "absurde" se prononce [apsyRdə]. Le son [s] sourd provoque le dévoisement de [b] qui peut être assimilé à [p]. La coarticulation est en revanche un phénomène plus général qui ne change pas forcément la nature des sons en présence, mais qui influence les mouvements des articulateurs.

Il existe un consensus quasi unanime pour affirmer que l'origine de la coarticulation rétentive est due à des effets inertiels. En revanche, plusieurs théories, que nous expliquerons au chapitre 2, rentrent en concurrence pour tenter d'expliquer le phénomène d'anticipation. Nous verrons que si certaines d'entre elles sont essentiellement basées sur des critères phonétiques et phonologiques, d'autres sont davantage liées à la dynamique des articulateurs.

Même si la coarticulation concerne l'ensemble des articulateurs, nous nous limiterons au chapitre 3 à la coarticulation labiale car les mouvements des lèvres sont les éléments essentiels de la tête parlante. L'état de l'art va nous permettre de faire le point sur la modélisation de ce phénomène par les techniques de synthèse d'images ou par les modèles de têtes de type géométriques ou biomécaniques. Certains modèles, comme les HMM (Hidden Markov Models) ou les réseaux de neurones sont proches des techniques utilisées en traitement de la parole alors que d'autres comme les modèles basés sur des règles ou les fonctions de dominance sont plus spécifiques à la coarticulation labiale.

Après avoir justifié le choix d'un modèle géométrique pour notre visage parlant, une étude monolocuteur va nous permettre dans le chapitre 4 d'étudier les principaux articulateurs qui contrôlent les mouvements de la bouche, leurs liens mutuels et leurs positions en fonction des sons. Parallèlement, une étude multilocuteur nous permettra d'estimer les degrés de variation des phonèmes et donc de cerner leurs caractéristiques intrinsèques.

L'étude des articulateurs et de leurs relations nous conduira au chapitre 5 à proposerr un algorithme de prédiction de la coarticulation en fonction de la séquence de phonèmes à articuler. Nous montrerons aussi comment cet algorithme formel peut être adapté à un locuteur précis.

La validation de cet algorithme se fera au chapitre 6 dans la phase de synthèse. Nous montrerons que la solution adoptée, bien que très légère en terme de volume de données à stocker est néanmoins très efficace. Les possibilités qu'offrent cet algorithme de pouvoir s'adapter à un locuteur précis, de gérer l'hyper ou l'hypo articulation, de pouvoir gérer n'importe quelle vitesse d'élocution constituent des atouts indéniables.

Malgré ces atouts, la tête parlante que nous pouvons générer est incomplète car seuls les mouvements des lèvres et de la mâchoire sont pris en compte. Dans le chapitre 7, nous allons estimer la qualité de notre synthèse, tout d'abord par une étude numérique en comparant les données réelles et estimées, puis par des tests de perception. Parallèlement, nous comparerons notre méthode de prédiction de la coarticulation avec une méthode basée sur les fonctions de dominance développée par Cohen et Massaro [19]. Cette dernière a été choisie car elle a servi de base à la construction de la tête artificielle nommée Baldi et a obtenu d'excellents scores comparée à d'autres modèles [10].

Nous terminerons cette thèse en proposant des perspectives pour prolonger ce travail.

Chapitre 1

Influence de la coarticulation dans la production de la parole

La parole naturelle ne correspond pas à une simple juxtaposition de sons isolés. Au contraire, l'articulation d'un son influence celle de ses voisins, c'est le phénomène de coarticulation. Dans le livre de William J. Hardcastle et Nigel Hewlet dédié à la coarticulation [39] dont nous avons extrait les références très anciennes, Barbara Kühnert et Francis Nolan indiquent que le principe même de la coarticulation est connu depuis plusieurs centaines d'années, mais que le terme date de 1933 quand Menzerath et Lacerda ont publié "Koartikulation, Steuerung und Lautabgrenzung" [61].

1.1 Les origines de la coarticulation

1.1.1 Les Premières théories

Brucke en 1856 [14] et Bell en 1867 [3] qui ont bâti les fondations de l'académie de Phonétique considéraient que les lettres alphabétiques avaient des réalisations physiques associées sous la forme de sons élémentaires. L'idée de base était que les sons se connectaient les uns les autres par un phénomène de glissement [78]. L'analogie peut être faite avec l'écriture liée. Chaque lettre a une forme bien définie, mais lors de l'écriture d'un mot, nous relions les lettres les unes aux autres.

Sievers, en 1876 [78] admet néanmoins la possibilité que dans certaines combinaisons de sons, les articulateurs qui ne sont pas impliqués peuvent anticiper la réalisation du son suivant tant qu'il n'y pas d'antagonisme. Paul, en 1898 [68] admet qu'un mot ne correspond pas à un nombre spécifique de sons indépendants, mais à une ligne continue constituée d'une infinité de sons.

1.1.2 Les premières mesures physiques

A partir de la fin du 19ème siècle, l'apparition de la " Kymography " a permis de quantifier la coarticulation de façon expérimentale. Ce procédé permet d'enregistrer des signaux en fonction du temps comme le signal acoustique, le débit d'air, les mouvements de langue. Rousselot en 1897 [75] a été l'un des premiers à utiliser ce procédé pour la parole. Néanmoins, Rousselot pensait que les données obtenues devaient et pouvaient être divisées en sons séparés. Quand il constatait que les courbes obtenues n'étaient pas jointives à la frontière entre sons, il en déduisait une erreur d'enregistrement. Il constata toutefois l'importance de l'anticipation. Dans une séquence Voyelle-Consonne-Voyelle (VCV) où la première voyelle est non arrondie et où la

seconde voyelle est arrondie, il remarqua que le geste d'arrondissement pouvait déjà commencer durant la réalisation de la première voyelle. De façon plus précise, Scripture, en 1902 [77] suppose que chaque mouvement articulatoire dépend des autres mouvements qui cohabitent pendant le même laps de temps. La langue n'est jamais immobile et n'occupe jamais exactement la même position d'une séquence à l'autre. Ainsi, Scripture réfute pour la première fois le fait que la parole correspond à une suite d'éléments statiques liés les uns aux autres. Scripture réfuta également la division de la parole en unités plus importantes que les phonèmes comme les syllabes par exemple.

Bien que le terme coarticulation n'existe pas encore à cette période, les phonéticiens avaient conscience de son existence. Cependant, les terminologies employées étaient vagues. Les termes assimilation, adaptation, similitude étaient souvent employés indifféremment. Jones en 1932 [43] a introduit une distinction entre similitude et assimilation. La similitude correspond au fait que deux sons deviennent voisins, mais ne changent pas leur identité (par exemple, dans le mot [pliz] en anglais, le phonème voisé /l/ est partiellement dévoisé. L'assimilation correspond au fait qu'un son puisse être remplacé par un autre son (/s/ changé en /S/ dans " jusque "). Cette distinction entre similitude et assimilation sera ensuite repris plus tard par Keating (1990) qui distingue la coarticulation phonologique et phonétique. L'assimilation ou la similitude peuvent être régressives quand un son est influencé par le son qui suit (par exemple, lors de la prononciation du mot "absent", le /b/ se transforme en /p/ et devient [apsã]). Elles peuvent être progressives quand un son est influencé par le son qui précède (par exemple, en Anglais, le pluriel de [dog] devient [dogz], c'est-à-dire que la consonne voisée [g] provoque le voisement de la consonne qui suit transformant [s] en [z]).

En 1933, Menzerath et Lacerda [61] ont introduit le terme de coarticulation en même temps que le terme de contrôle assisté (steering control). Selon eux, la coarticulation désigne le fait que les articulateurs essaient toujours d'anticiper le segment suivant et ils supposent que cette préparation commence aussi tôt que possible. Ils réservent le terme de contrôle assisté aux sons adjacents mettant en œuvre le même articulateur. Par exemple, dans la séquence [am], l'ouverture des lèvres du [a] est contrôlée par la consonne bilabiale qui suit.

En 1951, Stetson [80] a forgé les bases de la notion de coproduction. Il considère que c'est la coordination des mouvements articulatoires qui est primordiale et que cette coordination est modifiée par des changements tels l'augmentation de la vitesse d'élocution ou le stress. En considérant les syllabes comme unités fondamentales de la production de la parole, il a par exemple montré que la prononciation de plus en plus rapide de /tas tas tas / en continu conduit à /sta sta sta/. En outre, Stetson a observé que les mouvements articulatoires peuvent être présents bien que complètement cachés acoustiquement. Par exemple, dans la prononciation lente de /ispda/, les 3 consonnes intervocaliques sont clairement distinctes alors que lors de la prononciation rapide, les deux mouvements de la pointe de la langue au niveau des alvéoles se mélangent et le mouvement labial de /p/, même s'il est encore très visible sur le " kymogram " (enregistrement graphique des mouvements articulatoires) est recouvert par le mouvement global de la pointe de langue et ainsi, la phase de fermeture de cette bilabiale n'est plus nettement visible sur le signal acoustique.

Depuis 1960, on s'accorde à dire que l'intervalle d'influence de la coarticulation varie considérablement. En particulier certaines études (Benguerel and Cowan 1974 [7], Sussman and Westbury 1981 [82]) sur l'anticipation d'arrondissement du /u/ ont montré que l'anticipation pouvait

commencer 6 segments avant la voyelle cible. Par conséquent, la coarticulation est sans doute bien davantage que la conséquence de limitations physiologiques et d'effets inertiels. Par exemple, la première version du modèle de Lindblom [52] suggère qu'une séquence est réalisée par une chaîne de commandes. Comme les articulateurs ont des contraintes qui limitent leurs vitesses de déplacement, ils ne peuvent pas toujours atteindre leur cible avant la prochaine commande. Cette approche ne peut néanmoins pas prendre en compte le phénomène de forte anticipation.

Kozhevnikov et Chistovitch en 1965 [46] ont défini la syllabe (voyelle et toutes les consonnes qui précèdent celle-ci) comme unité de base pour la coarticulation. Leur recherche réalisée sur la protrusion dans une séquence /iC..Cu/ a montré que le début de la protrusion de la voyelle /u/ commence toujours à la première consonne même lorsque plusieurs consonnes précèdent la voyelle. Plus précisément, Öhman en 1967 [42] propose de séparer la coarticulation des voyelles de celle des consonnes. Selon lui, pour ce qui concerne les mouvements de la langue, les voyelles sont produites par de lents mouvements du corps de la langue sur lequel les gestes articulatoires des consonnes se superposent.

Wickelgren en 1969 et 1972 [85] [86] résout le problème de la coarticulation en considérant non plus le phonème comme constituant principal de la parole, mais l'allophone, c'est-à-dire le phonème dans son contexte (influence du phonème précédent et suivant). Néanmoins, cette théorie ne permet pas de prendre en compte l'effet coarticulatoire s'exerçant au delà du son adjacent. On pourrait bien sûr accroître le nombre d'allophones en tenant compte de plusieurs phonèmes précédents et suivants, mais ce nombre augmenterait alors de façon exponentielle. Cette approche semble trop systématique car le fait que certains phonèmes aient une plus grande influence que d'autres n'est pas pris en compte par cette théorie. Ceci nous amène au concept de résistance à la coarticulation proposé par Bladon et Al-Bamerni (1976) [12]. Ils attribuent à chaque allophone une valeur numérique quantifiant l'effet de la coarticulation anticipatrice et rétentive. Ces effets se propagent librement jusqu'à être inhibés par les effets des allophones voisins.

En ce qui concerne la coarticulation anticipatrice, la plupart des théories évoquées peuvent être classées en 3 ou 4 principales familles, associées à 3 ou 4 modèles théoriques : les modèles Look-Ahead, Time-Locked et hybrid. Récemment, Abry et Lallouache [2] ont présenté un 4ème modèle, le modèle expansionniste. Nous présenterons l'ensemble de ces modèles de façon détaillée dans le prochain chapitre.

1.2 Mesures de la coarticulation

Le concept de locus a été défini par Delattre, Liberman et Cooper [26] et Krull [47] a été le premier à utiliser les "équations du locus" de Lindlom [53] pour quantifier les effets de la coarticulation sur les syllabes CV. Pour une consonne donnée avec plusieurs contextes vocaliques, l'équation du locus représente les relations entre les valeurs des formants (essentiellement F2) au début de la transition CV et les valeurs des formants au milieu de la voyelle. Selon la théorie du locus, les consonnes et les voyelles ont des valeurs cibles pour les articulateurs et par conséquent des fréquences cibles pour les formants. Si la durée de la syllabe est suffisante, tous les phonèmes atteignent leurs cibles, mais si la durée de la syllabe est courte, la courbe représentant l'évolution du locus peut changer brusquement de sens avant d'avoir atteint les valeurs cibles. Bladon et Al-Bamerni [12], dans leur étude sur le degré de résistance à la coarticulation du phonème /l/

en Anglais ont repris cette idée en considérant comme mesure de la coarticulation la déviation de F2 en fonction du contexte.

En prenant en compte différentes voyelles, les loci permettent de dégager une droite de régression qui fournit un indicateur sur l'influence de la voyelle. Lorsque la pente est douce (proche de 0), ceci indique une absence d'effets coarticulatoires dus à la voyelle (i.e. une résistance maximale du geste consonantique à la coarticulation). Lorsque la pente est proche de 1, ceci indique une coarticulation maximale de l'articulation de la consonne avec celle de la voyelle (i.e. une résistance coarticulatoire minimale du geste consonantique). Quand les phonèmes ont une durée suffisante, les cibles des formants sont atteintes alors qu'elles peuvent ne pas l'être si la durée des phonèmes est faible (Fig. 1.1).

Néanmoins, certaines objections peuvent être formulées à l'égard de cette théorie du locus (Fant, 1973 [31] et Klatt en 1987 [45]) :

- La transition CV peut être une combinaison de mouvements rapides et lents. Par exemple, des mouvements rapides du bout de la langue peuvent être suivis par des mouvements lents du corps de la langue.
- Öhman [41] a montré que la voyelle précédente pouvait influencer le début de la transition CV et par conséquent la valeur cible du formant au niveau de la consonne ce qui tendrait à prouver que les loci des consonnes ne sont pas indépendants des voyelles voisines.

Klatt a émis l'hypothèse que les principaux effets de la voyelle sur l'articulation des consonnes sont dus à la position avant-arrière et à l'arrondissement des lèvres. Il a ainsi divisé les voyelles en 3 classes (avant, arrondies, en arrière et moins arrondies) et établi les droites de régression pour chacune des classes (Fig. 1.2) validant ainsi la théorie du locus pour chacune de ces classes.

En conclusion, les nombreuses études évoquées précédemment ont montré que la coarticulation est un phénomène beaucoup plus complexe que la simple similitude ou assimilation phonétique. Tout d'abord, la coarticulation s'intéresse à toutes les conséquences générées par la mise en fonctionnement simultanée de plusieurs articulateurs, même si les résultats sont difficiles voire impossibles à déceler à l'oreille. En outre, plusieurs études ont montré que les effets de la coarticulation s'étendent bien au delà des phonèmes immédiatement voisins prouvant ainsi que ce phénomène n'est pas complètement dû à des contraintes physiques ou des effets inertiels. Dans le chapitre qui va suivre, nous allons analyser les différentes théories tentant d'expliquer l'origine et l'ampleur de la coarticulation anticipatrice qui est prédominante en français.

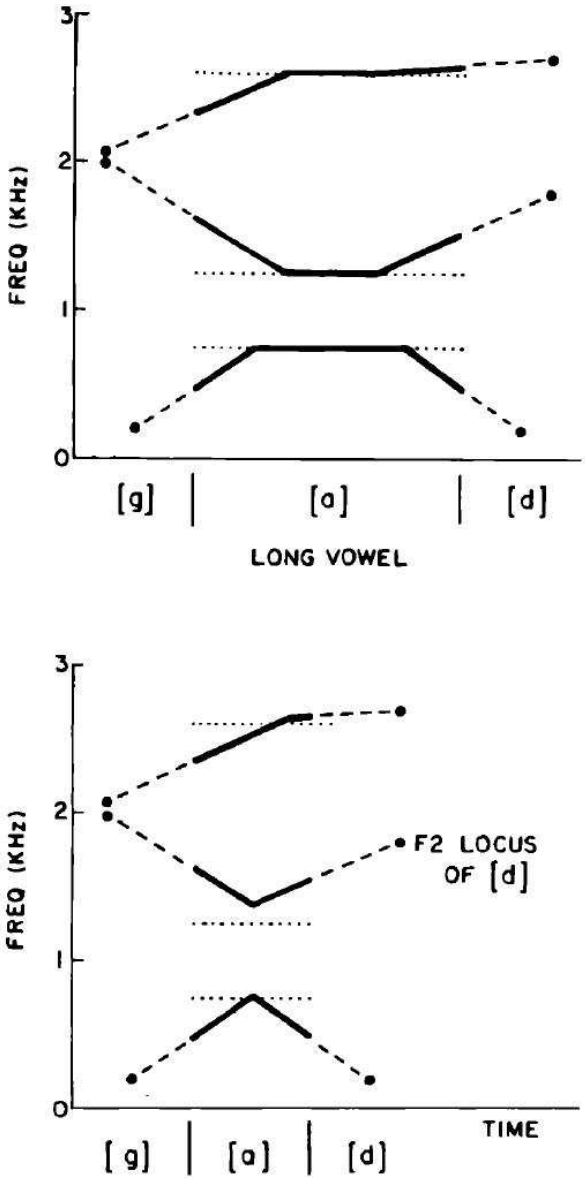

FIG. 1.1 – Variation du formant F2 en accord avec la théorie du locus. Si la voyelle est courte, la cible formantique n'est pas atteinte (figure du bas). Figure extraite de Klatt [45].

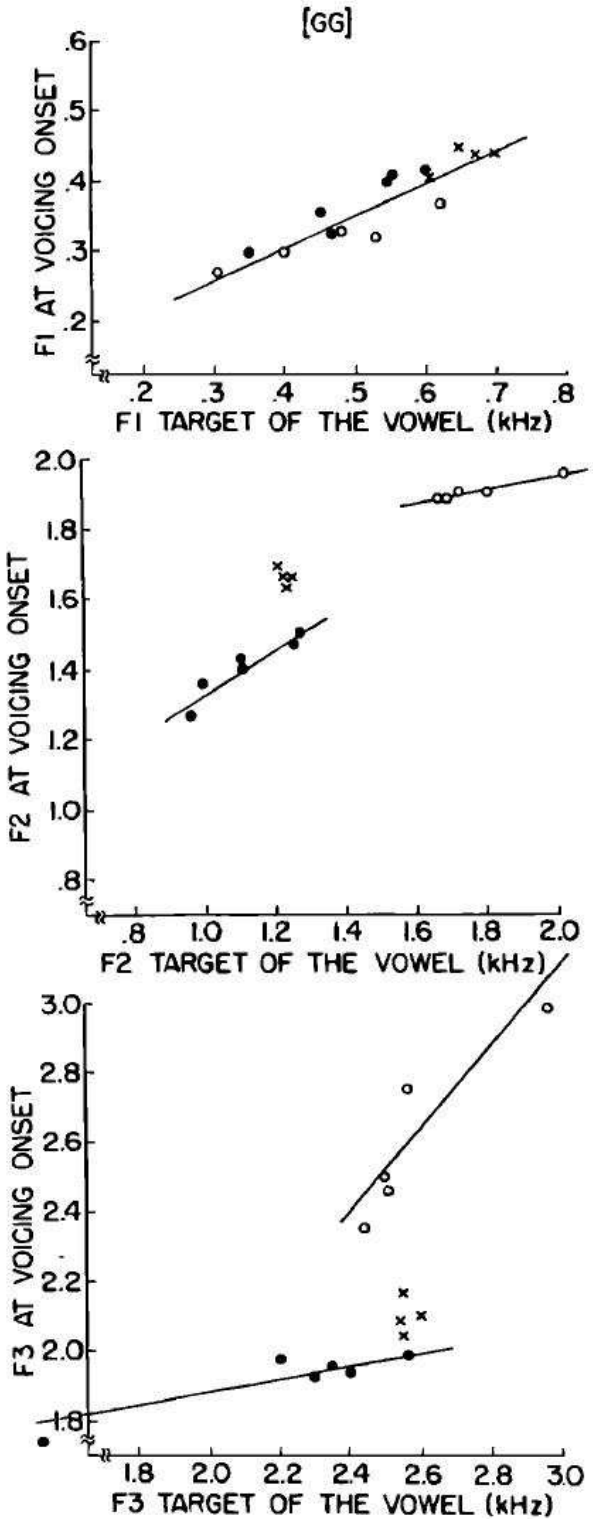

FIG. 1.2 – Séparation des voyelles en plusieurs classes et détermination des droites de régression pour chacune des classes (les cercles vides correspondent aux voyelles avant, les cercles pleins aux voyelles arrondies et et les croix correspondent aux voyelles postérieures moins arrondies) Figure extraite de Klatt (1987) [45].

Chapitre 2

Les principaux modèles théoriques caractérisant l'anticipation

Nous pouvons distinguer deux types de coarticulation selon la direction selon laquelle elle opère : la coarticulation anticipatrice (ou de droite à gauche) et la coarticulation rétentive (de gauche à droite). Mac Neilage [64] a trouvé que les effets de la coarticulation rétentive sont en principe plus réduits que les effets de la coarticulation anticipatrice. Ceci est en accord avec Daniloff et Hammarberg [24] qui affirment que la coarticulation rétentive est due essentiellement à l'inertie des articulateurs. La coarticulation anticipatrice ne fait l'objet d'aucun consensus clair et plusieurs modèles ont été élaborés à ce sujet.

2.1 Le modèle Look-Ahead

Ce modèle que l'on pourrait caractériser de "Tout Phonologique" a été présenté par Henke [40] et corroboré par l'étude expérimentale de Benguerel et Cowan [7]. Il suggère que l'anticipation peut commencer aussi tôt que possible tant qu'il n'existe pas d'antagonisme (Fig. 2.1). Les données d'entrée de ce modèle sont une suite de phonèmes qui sont étiquetés '+', '-' ou 'non spécifié' (noté 0) selon leur degré d'influence sur le paramètre articulatoire. Quand un phonème n'influence pas l'articulateur, la prochaine valeur spécifiée est anticipée. Les règles du modèle "Look Ahead" peuvent ainsi être formalisées par l'équation 2.1. La figure 2.2 montre un exemple de l'application de cette équation avec la phrase "une sinistre structure".

$$\begin{aligned} \pm\,(0)^n\,+\; &\longrightarrow \pm\,(+)^n\,+ \qquad n = 1, 2, 3, ... \\ and \quad \pm\,(0)^n\,-\; &\longrightarrow \pm\,(-)^n\,- \end{aligned} \tag{2.1}$$

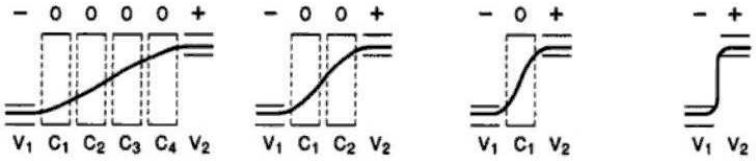

FIG. 2.1 – Evolution de la protrusion dans une séquence /iC..Cu/ selon le modèle Look-Ahead d'après E.Farnetani et D. Recasens [32].

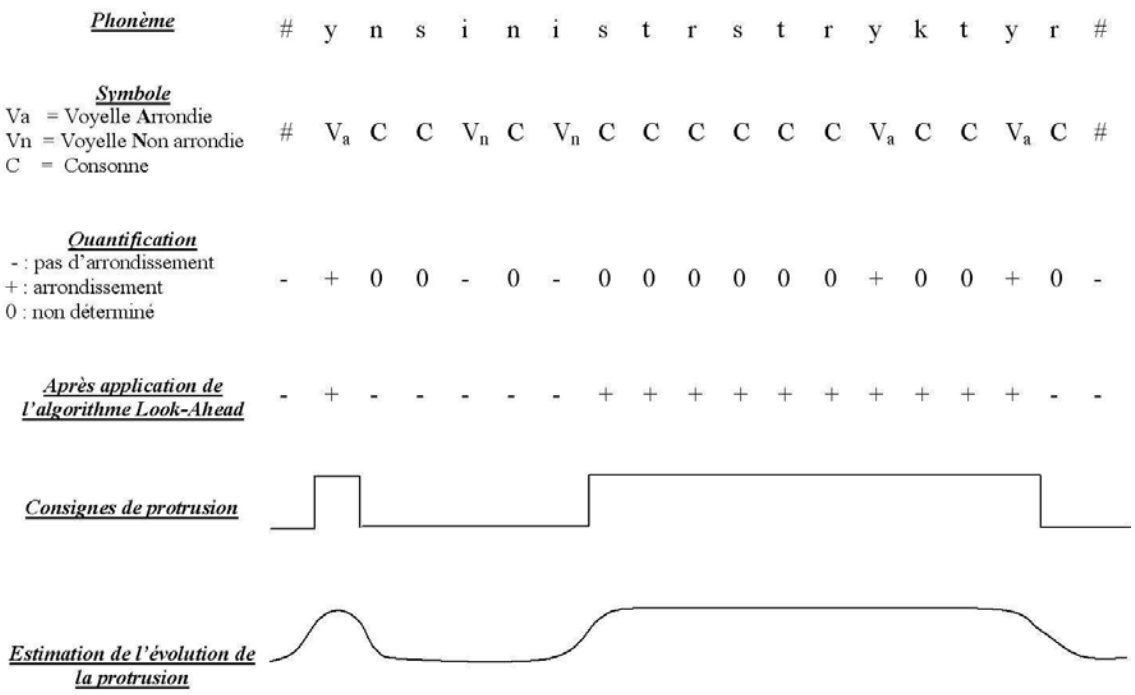

FIG. 2.2 – Evaluation du geste de protrusion en accord avec le modèle de Henke.

Plusieurs remarques peuvent être formulées sur ce modèle :

- L'affirmation selon laquelle le phénomème de coarticulation s'étend à tous les segments "neutres" qui précèdent un segment spécifié a été vérifiée par plusieurs études (Benguerel et Cowan[7] , Lubker[55], Sussman et Westbury[82]), mais aussi contredite par d'autres, notamment par Bell-Berti et Harris [5]. Ceci tend à prouver que le phénomème d'anticipation est plus complexe que ne le laisse penser cette théorie.
- L'affirmation selon laquelle l'anticipation est bloquée par les segments qui sont "spécifiés" ne semble pas non plus tout à fait exacte. En effet, de nombreuses études, notamment celle de Benguerel et Cowan [7] et celle de Sussman et Westbury [82] ont montré que le mouvement de coarticulation commence pendant l'exécution du segment contradictoire et non après, et spécialement quand plusieurs segments neutres interviennent.
- L'affirmation selon laquelle les segments neutres acquièrent la caractéristique du segment spécifié suivant est aussi remise en question par plusieurs études qui montrent que les segments neutres sont influencés, mais pas complètement modifiés. Par exemple, Engstrand [28] a montré que la protrusion diminue lors de la réalisation du /s/ dans la séquence /usu/. L'application de l'algorithme de Henke sur cette séquence aurait montré une stabilisation de la protrusion. De même, l'ouverture des lèvres dans la séquence /ata/ diminue au niveau du phonème /t/ alors que /t/ est quantifié comme neutre dans ce modèle.

Le modèle Look-Ahead proposé par Henke semble donc fournir une trame grossière de l'anticipation sans en considérer toutes les subtilités.

Une variante du modèle Look-Ahead est le modèle proposé par Kozhevnikov et Chistovich [46] qui met en évidence la notion de syllabe. Ce modèle, basé sur les relations de durée entre

les segments acoustiques et les données articulatoires labiales et linguales en Russe montre que la syllabe C_nV est à la fois une unité de rythme et une unité articulatoire. Selon les auteurs, un fort degré de coarticulation existe à l'intérieur des syllabes et un très faible degré entre les syllabes. Ce modèle entre en contradiction avec les données de Moll et Daniloff [63] qui ont montré que la coarticulation vélaire se faisait sentir deux voyelles avant une consonne nasale. De même, Benguerel et Cowan [7] ne sont pas d'accord avec le modèle syllabique de Kozhevnikov et Chistovich. Selon eux, il peut exister une coarticulation anticipatoire VC aussi bien que CV.

2.2 Le modèle Time-Locked

Ce modèle d'inspiration physique défendu par l'équipe de Haskins (Bell-Berti and Harris [5] puis par Browman et Goldstein [17] met en évidence la notion de gestes qui se superposent et par conséquent, dans cette théorie, la coarticulation est essentiellement due à la coordination des mouvements articulatoires. Un geste peut être considéré comme la formation et le relâchement au cours du temps d'une constriction par l'un des articulateurs. La durée d'un geste articulatoire étant relativement constante (pour des raisons de contraintes dynamiques sur les articulateurs de la parole), l'anticipation est donc relativement fixe par rapport au début du son acoustique à réaliser (d'où le nom de Time-Locked). Il se produit un chevauchement des gestes articulatoires du son qui précède et de celui à venir ; ainsi ce modèle est compatible avec la notion de coproduction définie par Fowler [33].

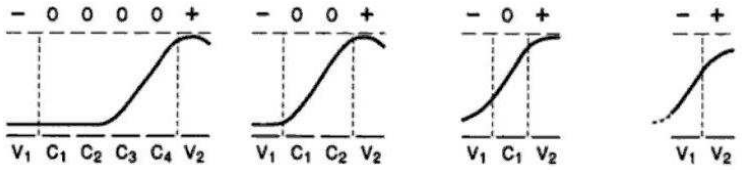

FIG. 2.3 – Evolution de la protrusion dans une séquence /iC..Cu/ selon le modèle Time-Locked d'après E.Farnetani et D. Recasens [32].

D'après ce modèle, la coarticulation est ainsi temporellement limitée et ne peut pas s'étendre de façon considérable. Contrairement au modèle Look-Ahead, la longueur de la séquence de phonèmes précédent la cible et le fait que certains phonèmes restent neutres n'est plus à prendre en considération. Leurs conclusions sont issues de mesures sur l'arrondissement des lèvres et d'analyses par ElectroMyoGrammes (EMG) - c'est-à-dire en mesurant le courant électrique généré par les contractions musculaires - de l'orbicularis oris (muscle circulaire situé autour de la bouche).

En conclusion, et à l'opposé du modèle Look-Ahead, basé sur des critères essentiellement phonétiques et phonologiques, le modèle Time-Locked met l'accent sur les contraintes dynamiques des articulateurs.

2.3 Le modèle Hybrid

Suite à une étude sur l'arrondissement des lèvres, Perkell et Chiang [71] sont davantage partisans d'un modèle hybride constitué de deux phases. La première phase relativement lente commence aussi tôt que possible et est essentiellement due à des critères phonologiques. La

deuxième phase, plus rapide, est déterminée par les caractéristiques dynamiques des articulateurs mis en jeu qui commence à un instant fixe avant la cible (Fig. 2.4). L'idée de cette théorie est que la coarticulation est une compétition entre des contraintes cinématiques et acoustiques qui peuvent varier d'un locuteur à un autre.

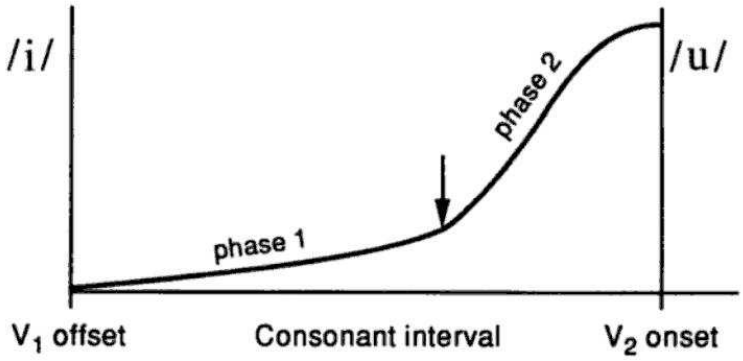

FIG. 2.4 – Evolution de la protrusion d'une séquence /iC...Cu/ selon le modèle hybride.

Ce modèle qui semble au premier abord mettre d'accord à la fois les partisans du modèle Time-Locked et Look-Ahead a au contraire trouvé de nombreux détracteurs [4][72][34]. Selon eux, même si des mesures sur l'arrondissement dans des séquences du type /ist#tu/ ont montré une évolution du mouvement de protrusion en deux phases cohérentes avec le modèle hybride, les résultats étaient biaisés par l'effet des consonnes intermédiaires et si cet effet était supprimé, le mouvement d'anticipation correspondait davantage au modèle Time-Locked. Pour supprimer l'effet des consonnes, Gelfer et al [34] ont comparé les résultats obtenus pour /iCi/ et /iCu/ ce qui a permis par soustraction d'identifier l'instant du début de protrusion.

2.4 Le modèle expansionniste

Après avoir mis en évidence les différences dues au contrôle de la jointure qui apparaît inévitablement dans les groupes de consonnes complexes habituellement manipulés dans les expériences sur l'anticipation et conscients de la variabilité inter et intra locuteurs du phénomène, Abry et Lallouache [2] ont émis l'hypothèse que la durée du mouvement est fortement expansible et relativement peu compressible d'où la tendance à anticiper quand le temps le permet. Selon eux, 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 où l'expansion n'est plus possible.

Le corpus analysé par Abry et Lallouache est composé de séquences /iC_ny/ avec un nombre variable de consonnes intercalées entre /i/ et /y/. Les conclusions peuvent se résumer en trois points :

- Premièrement, ils ont constaté que le mouvement de protrusion correspondait à une seule phase de mouvement.
- Le maximum de protrusion a lieu aux alentours de la voyelle arrondie.
- La relation entre la durée des consonnes et l'intervalle de protrusion n'est pas linéaire, mais davantage modélisé par une fonction hyperbolique. Ceci vient du fait que le mouvement de protrusion est largement expansible et peu compressible ; cette durée d'anticipation ne peut

par ailleurs pas être inférieure à un seuil minimum, en l'occurrence la durée d'anticipation qui est mesurée quand deux sons, pour lesquels le geste articulatoire considéré est pertinent, se succèdent immédiatement. La durée d'anticipation peut ainsi, pour certains locuteurs, atteindre celle que prédirait le modèle Look-Ahead, mais ce n'est pas systématiquement le cas.

Ce modèle est en accord avec les conclusions de Boyce et al [13] qui ont montré la forte influence de la vitesse d'élocution sur la forme du mouvement d'anticipation.

2.5 Pourquoi des modèles si différents ?

Il semble que les modèles purement phonologiques ne prennent pas suffisamment en compte les contraintes imposées par les articulateurs et classent de façon trop brutale les segments résistants à la coarticulation et ceux qui ne le sont pas. En effet plusieurs études [5][28] ont montré que des segments non spécifiés pouvaient avoir une influence. Par exemple, Engstrand a montré que le mouvement de protrusion pendant la séquence /usu/ diminuait au niveau du phonème /s/ alors que ce dernier n'est pas censé influencer ce phénomène. Il semble donc plus logique de quantifier graduellement les niveaux de résistance.

2.5.1 Quantification du degré de résistance à la coarticulation

La notion de résistance à la coarticulation a été introduite pas Bladon et Al Bamerni [12] dans une étude acoustique sur les allophones de /l/ en Anglais. L'étude analyse l'état des fréquences F1 et F2 dans les allophones cibles et dans un certain nombre de voyelles. Les résultats indiquent que les effets de la coarticulation CV décroît graduellement quand on passe de la version palatale de /l/ à la version vélaire. En outre, les effets coarticulatoires dépendent des éléments "frontières". Selon les auteurs, les effets de la coarticulation peuvent être modélisés en attribuant un degré de résistance à chaque allophone et à chaque élément "frontière". Les coefficients de résistance ne seraient donc pas universels, mais dépendant des spécificités de la langue ou des dialectes. Avec cette théorie, il est par exemple possible de prendre en considération les différences de nasalité entre l'Anglais et l'Américain. Cette caractéristique incite à penser que les degrés de coarticulation d'une langue dépendent à la fois des caractéristiques phonétiques et phonologiques.

2.5.2 Prise en compte mutuelle de la phonétique et de la phonologie

Le bon sens voudrait qu'il existe un intermédiaire entre le modèle Look-Ahead que l'on pourrait caractériser de tout phonologique et le modèle Time-Locked qui est purement articulatoire. Le modèle de Keating [44] a cette caractéristique et fait apparaître la notion de résistance à la coarticulation. A partir de mesures effectuées en différents contextes, P. Keating établit une fenêtre de variation possible pour les phonèmes au niveau d'un articulateur donné. Ainsi, une séquence donnée est associée à une suite de fenêtres. La variation de l'articulateur correspond à établir un "chemin" entre ces articulateurs en minimisant l'effort articulatoire comme le montre la figure 2.5. Plus la largeur de fenêtre est étroite, plus le phonème est "résistant" à la coarticulation.

Selon Keating , les différences de coarticulation entre langues peuvent être phonologiques quand des règles d'assimilation s'appliquent à une langue et pas à une autre ou phonétique quand les segments non quantifiés sont interprétés différemment d'une langue à l'autre. Le principe du "moindre effort" peut-être appliqué s'il n'y a pas d'antagonisme phonologique.

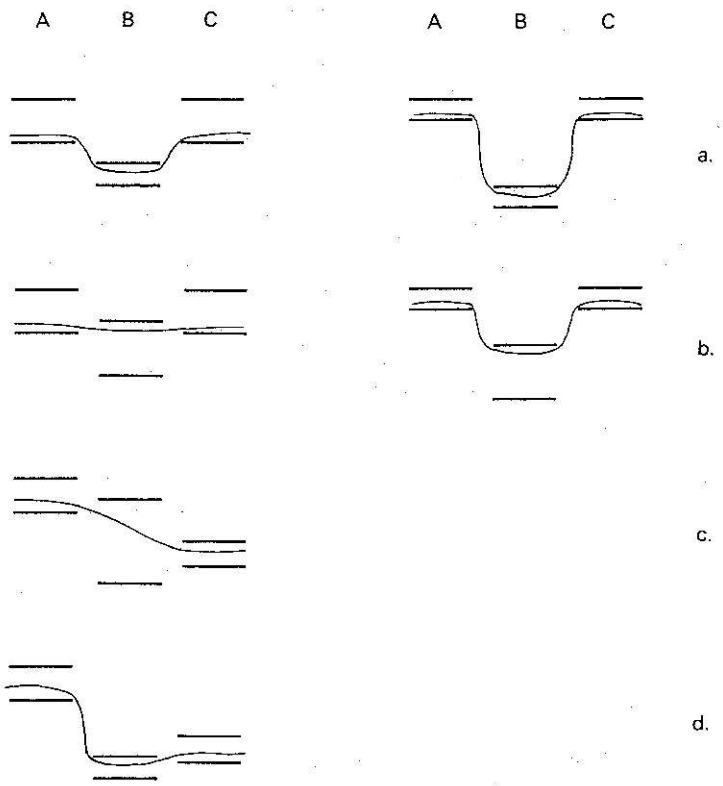

FIG. 2.5 – Illustration d'une séquence de 3 phonèmes associée à des fenêtres de largeurs variables (selon Keating [44]).

Une critique sérieuse de la théorie de Keating vient de Browman et Goldstein [18] qui considèrent que la représentation de la parole en deux entités distinctes que sont la phonétique et la phonologie ne permet pas de prendre en compte les interactions entre ces deux domaines. En particulier, Bell-Berti ([6]) considère difficile la mise en relation des caractères phonologiques non spécifiés avec les positions articulatoires spécifiques. Une autre critique de la théorie de Keating est formulée par F. H. Guenther ([37]) qui reproche à Keating de ne pas avoir tenu compte de l'influence et la compensation mutuelle des articulateurs. Par exemple, il fait remarquer que le degré de liberté assez grand de la lèvre supérieure et de la mâchoire lors de la prononciation de la voyelle /a/ se réduit considérablement si on étudie la déformation combinée de ces deux articulateurs. Blackburn[11] est en accord avec la théorie de Keating mais propose de remplacer la notion de fenêtre par une probabilité.

Contrairement au modèle de Henke, les modèles de Keating et Blackburn ne classent pas de façon si brutale les segments. Néanmoins, ces modèles ne tiennent pas compte de l'influence mutuelle des articulateurs et de leurs contraintes physiques.

2.6 Conclusion

Ces principaux modèles théoriques insistent tous sur l'importance de la coarticulation anticipatrice ; certains considèrent que ce phénomène est essentiellement du à des contraintes liées aux

articulateurs alors que d'autres y voient davantage un caractère phonologique ou phonétique. En tout cas, il est évident qu'il est indispensable de tenir compte de ce phénomène pour obtenir une tête parlante réaliste. Dans le chapitre suivant, nous allons étudier les différentes voies qui ont été choisies pour construire les visages parlants. Si certaines d'entre elles semblent davantage liées au monde de l'animation, du cinéma et n'accordent pas une importance prioritaire à l'articulation, d'autres en revanche considèrent l'intelligibilité comme le critère essentiel et s'inspirent en partie des modèles théoriques que nous venons d'aborder dans ce chapitre.

Chapitre 3

Mise en œuvre de la coarticulation labiale pour la modélisation de têtes parlantes : état de l'art

L'intelligibilité d'une tête parlante est essentiellement due aux mouvements labiaux, Massaro [58] et Ouni [66] l'ont notamment montré dans leurs études. C'est pour cela que la prédiction de ces mouvements constitue l'élément essentiel de notre travail de recherche. Dans l'avenir, il pourra être envisagé de modéliser l'intérieur de la bouche, les dents, les mouvements des yeux, les battements des paupières, etc. mais pour l'instant, nous allons nous concentrer sur les mouvements des lèvres et de la mâchoire.

3.1 Importance de la composante visuelle pour l'intelligibilité de la parole

Sumby et Pollack [81] ont montré que l'intelligibilité décroît quand le ratio Parole sur bruit diminue et quand la taille du vocabulaire augmente. Parallèlement, ils ont montré que la visualisation du locuteur augmentait considérablement l'intelligibilité en présence de bruit et que cette information bimodale était très résistante au bruit. En résumé, la contribution visuelle est d'autant plus grande que le bruit est important (Fig. 3.1). Les conclusions de Sumby ont été reprises par Benoit [8] et par D Massaro [58].

L'apport de la contribution visuelle doit être la plus précise possible, sinon, elle risque de fausser l'information acoustique et l'apport d'intelligibilité apporté serait nul voire négatif. Un exemple très connu est l'effet Mc Gurk [60]. Typiquement, si on diffuse la séquence audio /baba/ et que les mouvements labiaux représentent la séquence /gaga/ alors on aura l'impression d'entendre le son /dada/.

3.2 Principe de modélisation des visages parlants

Deux grandes approches cohabitent dans ce domaine : les approches basées images qui essaient de reconstruire une vidéo la plus réaliste possible en associant des bribes de vidéos enregistrées et les approches basées modèle (3D) dans lesquelles on modélise la tête sous la forme d'un objet 3D.

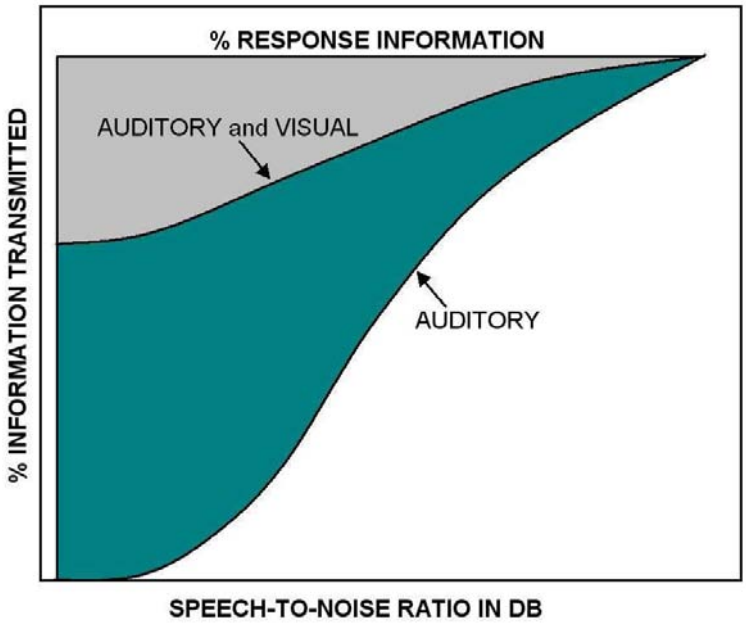

FIG. 3.1 – Contributions du message oral et visuel à l'intelligibilité en fonction du ratio Signal/Bruit (De Sumby et Pollack [81]).

3.2.1 Par synthèse d'images

Ce type d'approche privilégie le rendu et l'esthétisme du visage parlant et est principalement destiné à créer des animations les plus réalistes possibles. Les informations à manipuler étant très importantes, la plupart des méthodes existantes utilisent la superposition d'images associées à des zones fixes du visage avec des vidéos représentant les zones en mouvement. Le raccordement des différentes parties est un point délicat afin que la vidéo finale soit perçue comme un tout. L'un des premiers systèmes fonctionnant sur ce principe est "Video Rewrite" de Bregler et al[16]. Il superpose des formes de bouche plausibles sur une image de fond. La synthèse est effectuée en deux étapes (Fig. 3.2) :

- La première étape extrait les phonèmes du flux audio, puis cherche les visèmes(équivalents visuels des phonèmes) de la base d'apprentissage associés à chaque triphone.
- La seconde étape mélange et réalise un morphing entre les visèmes successifs. Des ajustements sont aussi effectués pour corriger les mouvements de tête.

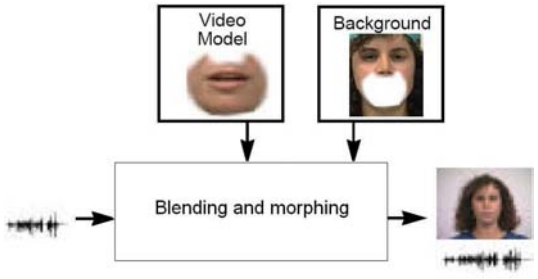

FIG. 3.2 – Modélisation d'un visage parlant par synthèse (Video Rewrite).

Le modèle de Cosatto [21] effectue une synthèse similaire à celle de Bregler en utilisant 6 portions du visage au lieu de 2 : les yeux, la bouche, les dents (supérieures et inférieures), le menton et le front). Cette décomposition multiple a pour but de limiter le nombre d'images de référence et le nombre de paramètres nécessaires pour contrôler une partie. L'avantage de ce modèle, contrairement à celui de Bregler est de pouvoir générer plus facilement des émotions particulières. Ce système utilise 12 visèmes associés à des monophones et utilise la technique de coarticulation de Cohen et Massaro [19] pour tenir compte du contexte. En fait, chaque phonème est associé aux paramètres définissant la forme de bouche associée. A chaque intervalle de temps, les paramètres qui contrôlent la forme de la bouche sont calculés en tenant compte des phonèmes précédents et suivants.

Plutôt que de diviser le visage en portions, Scott et al [76] ont choisi de contrôler la déformation de chaque pixel de l'image. Des images clés (visèmes) ainsi que des points de contrôles sont judicieusement choisis, puis un phénomène de morphing est appliqué pour reconstruire la forme estimée. MikeTalk [30] du MIT est un autre système basé sur le morphing utilisant des algorithmes de type flux optique pour mesurer les déformations entre images. Ces dernières techniques présentées sont toutes basées sur des visèmes associés à des monophones et considèrent la coarticulation comme une interpolation linéaire entre formes adjacentes.

Ces méthodes sont définies pour des images 2D (même si des artifices existent pour reconstruire du pseudo 3D) et restent associées à un locuteur précis. Même si dans certaines conditions, elles peuvent produire des résultats ultra-réalistes, elles nécessitent une base de données dense et un temps de calcul non négligeable incompatible avec le temps réel. Les techniques qui utilisent le visème comme forme élémentaire ne permettent pas de prendre en compte toute la complexité du phénomène de coarticulation qui peut s'étendre bien au delà des phonèmes voisins. De ce fait, les combinaisons de type VC...CV relativement complexes ne peuvent pas être correctement synthétisées.

3.2.2 Modélisation en 3D

Alors que les solutions exposées précédemment visent à obtenir un visage parlant le plus réaliste possible, les modèles 3D constituent une solution plus légère, plus rapide, donc mieux adaptée au domaine temps réel et cherchent à synthétiser avec une grande précision les mouvements de la bouche et de la mâchoire. Dans ce type de modèle, la tête virtuelle est un objet 3D que l'on cherche à animer. Certains modèles contrôlent les déformations avec des paramètres purement géométriques alors que d'autres utilisent des paramètres articulatoires ou biomécaniques.

Dans le cas des déformations contrôlées par des paramètres géométriques, les modèles utilisent des transformations géométriques simples (rotation, translation) pour déformer les points modélisant le visage 3D. Baldi, la tête parlante développée au Perceptual Science Laboratory (Université de Californie - Santa Cruz) et Swen, développée au "Royal Institute of Technology" (KTH) font partie de ce type de modèles (Fig. 3.3). Ces modèles sont tous des descendants du modèle proposé par Parke [67] qui proposait le maillage de la figure 3.4 comme modèle de base. Ces méthodes par interpolation quoique simples à mettre en œuvre exigent un maillage très fin et énormément de réglages pour pouvoir modéliser les mouvements des articulateurs.

Le principal problème avec les modèles géométriques est de pouvoir identifier avec précision les points caractéristiques du visage qui doivent servir de base au maillage tridimensionnel. La norme MPEG4 fournit des informations précises à ce sujet. Cette norme de codage d'objets audiovisuels [65] modélise un visage par 84 points 3D (FP : Feature Points), lesquels sont pilotés

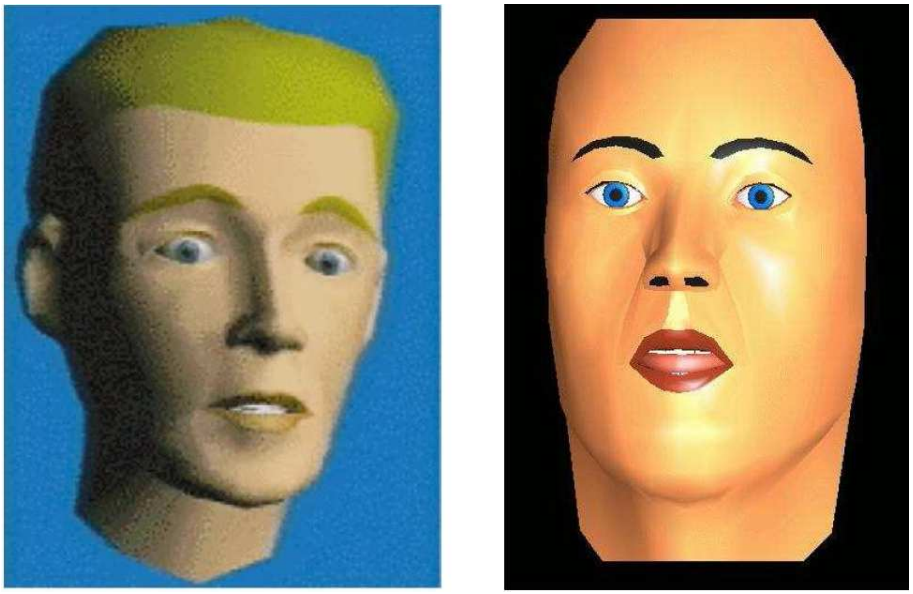

FIG. 3.3 – Exemple de visages parlants utilisant des déformations géométriques.

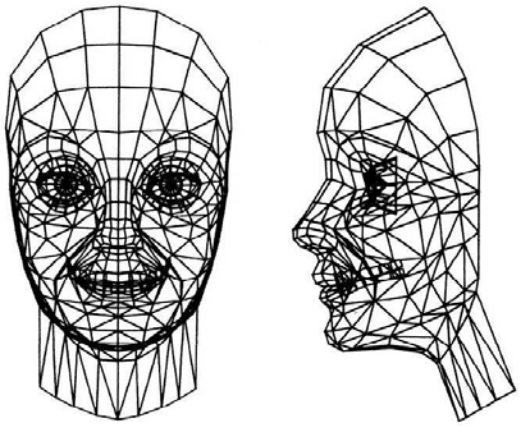

FIG. 3.4 – Modèle de Parke.

par 68 paramètres, les FAPs(Facial Action Parameters) qui sont soit purement géométriques, soit articulatoires. Le passage de paramètres purement géométriques à des paramètres articulatoires est une optimisation visant à produire des mouvements cohérents. Les paramètres articulatoires de la norme MPEG4 sont ceux contrôlant la hauteur, largeur de la bouche, la protrusion des lèvres, le mouvement de la mâchoire. Les chercheurs de L'Institut de Communication Parlée (ICP) ont développé un modèle [27] basé sur ces paramètres articulatoires et contrôlés par le modèle de coarticulation d'Öhman [42].

Bien que la norme MPEG4 fournisse des points de repères précis servant de base à la construction du maillage, ce dernier peut encore être affiné s'il est contrôlé par les paramètres biomécaniques, c'est-à-dire les muscles et les articulateurs du visage. Les actions sur les muscles se propagent au niveau des articulateurs et engendrent des déformations faciales. Les tissus de la

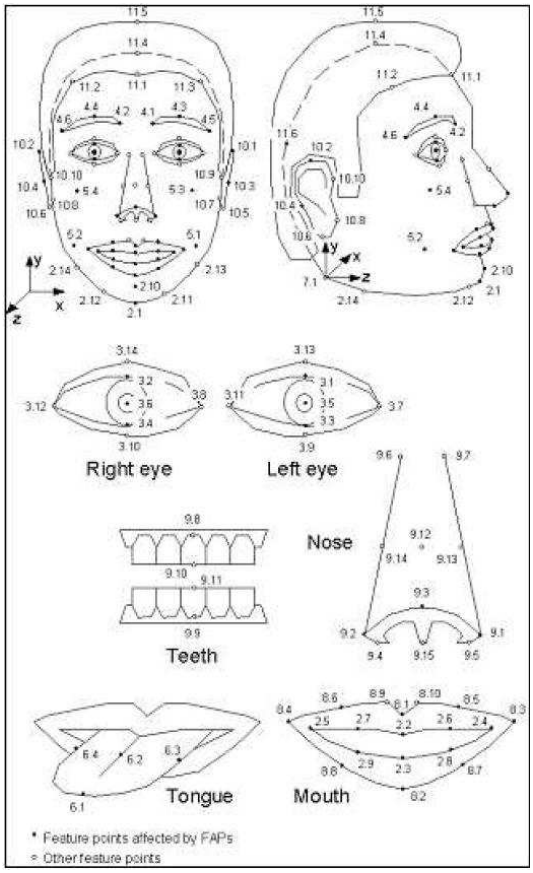

FIG. 3.5 – Points de contrôle (FP) de la norme MPEG4.

peau sont souvent modélisés par des ressorts inter connectés [73]. Le modèle de visage de Lucero et al [56] (Fig. 3.6) est construit sur ce principe. Les contractions musculaires sont le résultat de la déformation de zones ellipsoïdales autour des points de contrôle. Ce modèle très prometteur est également très complexe à mettre en œuvre car le visage comporte près de 250 muscles auxquels s'ajoutent les interactions os/tissus très difficiles à modéliser.

En conclusion, les modèles par synthèse d'image ou par maillage 3D ne rentrent pas réellement en concurrence et des études continuent à exister dans ces deux domaines. Si les premiers semblent davantage destinés au domaine des films d'animation, les seconds sont mieux adaptés aux solutions légères visant à produire une animation à la volée. Dans cette thèse, inspirée d'un projet qui vise, grâce à la lecture labiale à aider les sourds et les malentendants, il est important de concentrer la synthèse sur les éléments du visage essentiels à la compréhension ; en outre, le but final est de pouvoir générer une synthèse temps réel ; en conséquence, notre choix s'est porté sur la création d'un modèle géométrique.

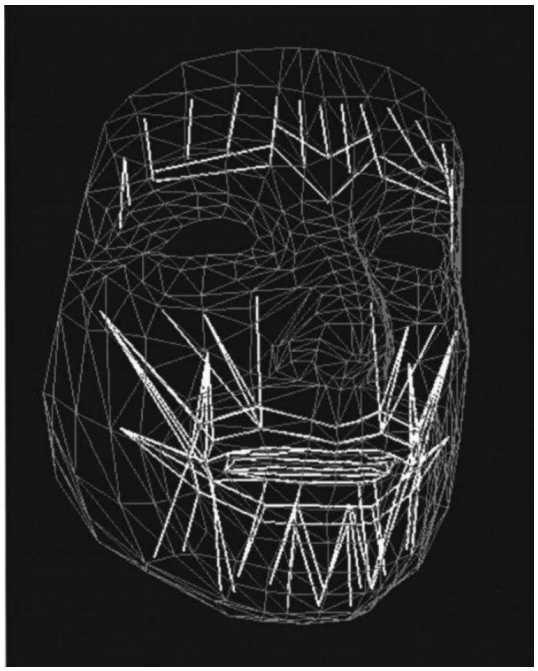

FIG. 3.6 – Lignes d'actions des muscles faciaux du modèle de Lucero et al [56].

3.3 Les modèles expérimentaux de modélisation de la coarticulation

Les modèles expérimentaux de la coarticulation sont variés ; certains ont leur propre modèle théorique associé alors que d'autres se rapprochent beaucoup en ce qui concerne la coarticulation anticipatrice de modèles présentés au chapitre 2. Par exemple, la définition de visèmes peut être associée au concept d'assimilation car chaque visème correspond à un état précis des différents paramètres labiaux alors que la notion de vecteurs cibles fait davantage référence au concept de coproduction car dans ce cas, les paramètres labiaux restent bien distincts les uns des autres.

Différentes techniques ont été utilisées pour définir les paramètres du modèle de coarticulation : certaines font appel à des règles (phonétiques par exemple) et permettent un contrôle local de chaque paramètre alors que d'autres, basées sur des critères statistiques, offrent davantage un contrôle global du mécanisme de coarticulation.

3.3.1 Construction d'un modèle à base de règles

Dans ce type de modèle, des règles phonétiques ou articulatoires servent de base à la prédiction de l'anticipation ou à la rétention des mouvements. Pelachaud [69] a imaginé un modèle basé sur une série de visèmes. Un visème a été associé à chaque phonème quantifié comme "non déformable". Dans le cas des phonèmes "déformables", ceux-ci sont modélisés visuellement en accord avec le modèle théorique de type Look-Ahead (cf. 2.1). Ce modèle prédit que chaque ajustement articulatoire commence juste après une position clé et se poursuit jusqu'à la prochaine. Deux règles sont prises en compte : l'effet anticipatoire et rétentif. L'ajustement articulatoire est effectué sur une suite de consonnes (qui n'appartiennent pas à la série des phonèmes "peu déformable") qui suivent ou précèdent une voyelle. Un algorithme en trois étapes est appliqué.

Dans la 1ère étape, les règles de coarticulation anticipatrices et rétentives sont appliquées à tous les clusters qui sont définis comme dépendants du contexte. L'étape suivante prend en compte les temps de contraction et de relaxation des muscles. La troisième étape étudie la façon dont les étapes 1 et 2 peuvent être combinées. Le modèle de Pelachaud peut donc être associé au modèle théorique de type hybride.

Beskow a lui aussi développé un modèle à base de règles et de visèmes ([9]). Un ensemble de paramètres ont été définis pour quantifier les mouvements articulatoires. En ce qui concerne les lèvres, Beskow a choisi l'arrondissement, l'occlusion bilabiale et l'occlusion labiodentale.

Ces modèles sont intéressants car ils permettent de contrôler très précisément les mouvements coarticulatoires. Des réglages ponctuels sont possibles en fonction par exemple de certaines combinaisons de phonèmes. Beskow [10] a montré que les têtes parlantes pilotées par de tels modèles obtenaient d'excellents résultats au niveau de la perception.

3.3.2 Modèle basé sur les fonctions de dominance

3.3.2.1 Modèle de base de Löfqvist

Le modèle expérimental proposé par Löfqvist [50] repose sur la théorie des gestes présentée précédemment à la section 2.2 et peut donc être associé au modèle théorique de type Time-Locked. Il fait référence au modèle de coproduction où un ensemble de gestes se "mélangent" pour former le geste final. Löfqvist a représenté l'influence de chaque articulateur par une fonction de dominance (Fig. 3.7). Les fonctions de dominance sont propres à chaque articulateur (fig. 3.8) et diffèrent selon les locuteurs. En fonction de la vitesse d'élocution, les segments se chevauchent plus ou moins pour donner le geste final. Löfqvist et Yoshioka ont constaté dans une étude plus ancienne [51] qu'une séquence /s#k/ pouvait être produite avec un ou deux gestes laryngaux. Lors d'une prononciation rapide, seul un geste apparaît alors que deux gestes distincts sont observables à vitesse lente. A vitesse intermédiaire, on observe un mélange des deux gestes.

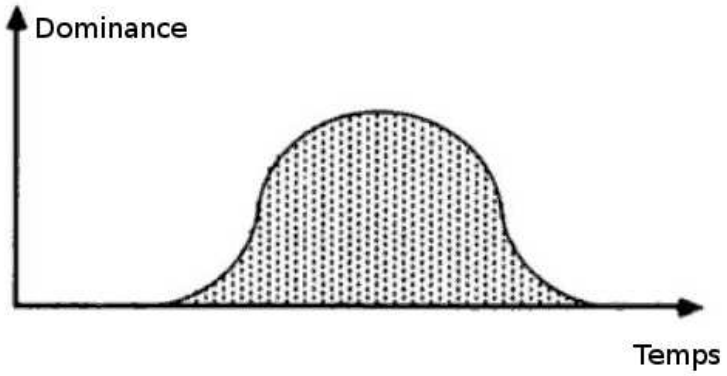

FIG. 3.7 – Représentation de l'influence d'un segment (dominance) en fonction du temps. Schéma extrait de Lôfqvist [51].

3.3.2.2 Modèle de Cohen et Massaro

Le modèle de Cohen et Massaro [19] s'inspire très nettement du modèle de Löqvist présenté précédemment. Dans ce modèle, chaque segment est associé à un vecteur cible. Des fonctions de

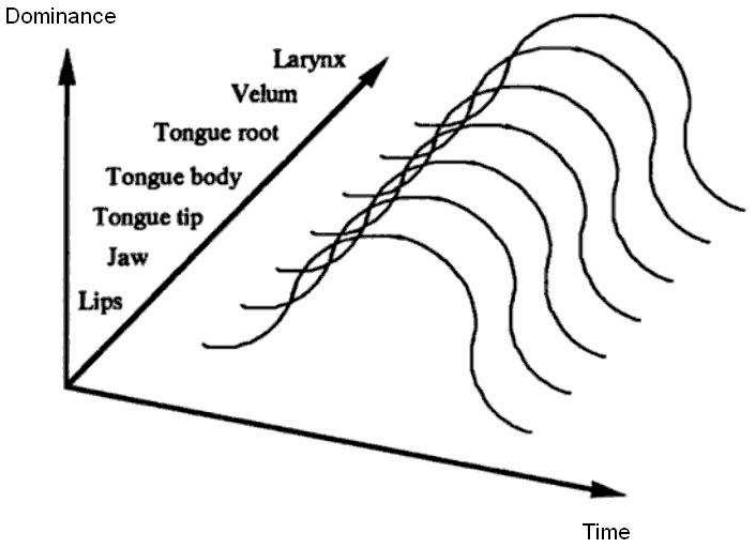

FIG. 3.8 – Influence simultanée des différents articulateurs. Schéma extrait de Lôfqvist [51].

dominances sont utilisées afin de pondérer les valeurs cibles en fonction du temps. Les fonctions de dominance sont constituées d'une paire de fonctions exponentielles, l'une prenant en compte la coarticulation anticipatrice et l'autre prenant en compte la coarticulation rétentive. Les durées de ces exponentielles étant indépendantes du contexte, le modèle de Cohen et Massaro peut être associé au type " Time-Locked ". La forme générale de la fonction de dominance est donnée par l'équation 3.1. Dans cette équation τ correspond à l'écart temporel avec le centre du segment. c permet de contrôler le degré de transition et θ contrôle l'importance de la coarticulation anticipatrice ou rétentive. Afin de tenir compte de ces deux types de coarticulation, Cohen et Massaro ont décomposé l'équation 3.1 en 2 parties, les équations 3.2 et 3.3. Dans ces équations, $\theta_{\leftarrow sp}$ quantifie l'importance de l'anticipation alors que $\theta_{\rightarrow sp}$ caractérise l'influence de la coarticulation rétentive.

$$D = e^{-\theta\tau^c} \tag{3.1}$$

$$D_{sp} = \alpha_{sp} e^{-\theta_{\leftarrow sp}|\tau|^c} \quad \texttt{si} \quad \tau \geq 0 \tag{3.2}$$

$$D_{sp} = \alpha_{sp} e^{-\theta_{\rightarrow sp}|\tau|^c} \quad \texttt{si} \quad \tau < 0 \tag{3.3}$$

Dans les deux cas, la distance temporelle τ au pic de dominance est donnée par l'équation 3.4 où t est l'instant recherché, $t_{0\ sp}$ représente le décalage entre le centre du segment de parole et le pic de dominance et $t_{c\ sp} = t_{start\ s} + duration_s/2$ représente l'instant associé au centre du segment, c'est-à-dire l'instant de début incrémenté de la moitié de la durée du segment.

$$\tau = t_{c\ sp} + t_{0\ sp} - t \tag{3.4}$$

La combinaison de ces fonctions de dominance pour chaque segment conduit à l'équation 3.5 où N est le nombre de segments dans la séquence et T_{sp} correspond à la valeur cible du segment considéré.

$$F_p(t) = \frac{\sum_{s=1}^{N}(D_{sp}(t) \times T_{sp})}{\sum_{s=1}^{N} D_{sp}(t)} \tag{3.5}$$

La figure 3.9 représente l'évolution au cours du temps des fonctions de dominance et de la fonction résultante.

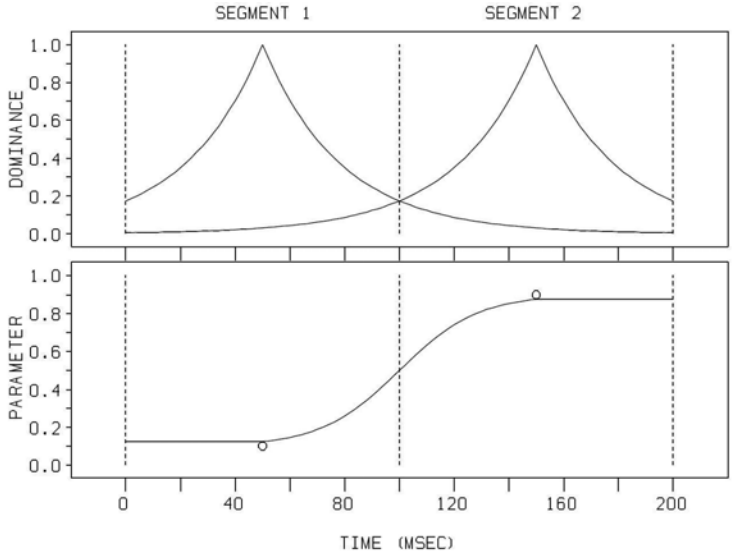

FIG. 3.9 – Représentation des fonctions de dominance de deux segments en fonction du temps et de la fonction résultante (en bas). Les cercles dans la figure du bas indiquent les valeurs des paramètres de contrôle. Figure extraite de Cohen et Massaro [19].

L'avantage de la technique proposée par Cohen et Massaro est :
- de pouvoir contrôler aisément l'influence d'un segment de parole grâce au paramètre α (Fig. 3.10)
- de pouvoir contrôler aisément l'importance de la coarticulation anticipatrice ou rétentive (Fig. 3.11)
- de pouvoir contrôler l'étendue de la coarticulation (Fig. 3.12)
- de pouvoir s'adapter par le contrôle des durées des segments à différentes vitesses d'élocution (Fig. 3.13)

Ce modèle offre donc une souplesse importante permettant de s'adapter au locuteur et à sa vitesse d'élocution. Néanmoins, la figure 3.13 montre que plus la vitesse d'élocution croît, plus le mouvement résultant s'éloigne de la cible au niveau du phonème /t/ lors de la prononciation de /utu/. En ce qui concerne la protrusion et la séquence /utu/, ceci paraît parfaitement logique. En revanche, si l'on considère par exemple la prononciation de /apa/, le modèle de Cohen et Massaro ne peut pas garantir la complète fermeture des lèvres au niveau de /p/ surtout quand la vitesse d'élocution est rapide. Or, ceci est un critère indispensable pour les bilabiales.

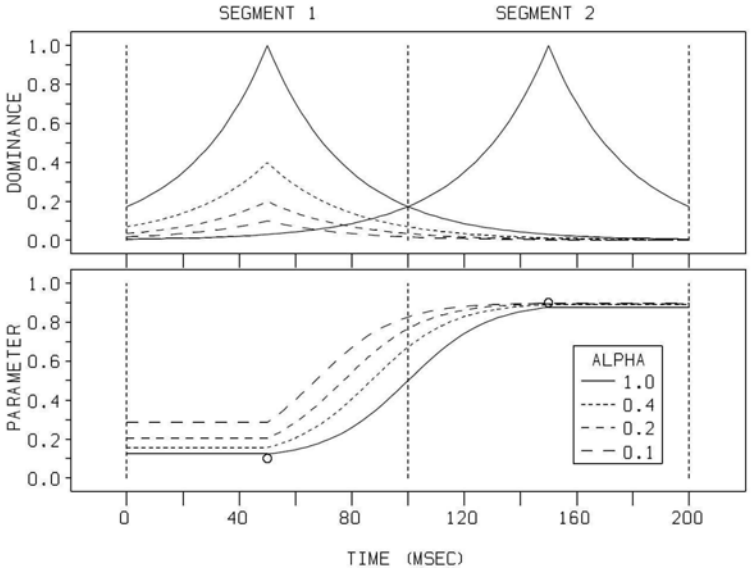

FIG. 3.10 – Influence du paramètre α. Figure extraite de Cohen et Massaro [19].

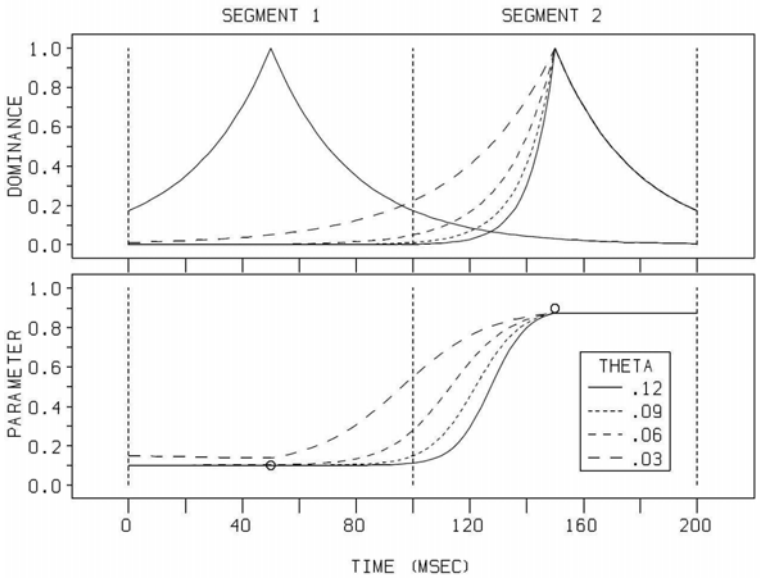

FIG. 3.11 – Influence du paramètre θ. Figure extraite de Cohen et Massaro [19].

Le modèle proposé par Cosi [22] propose plusieurs améliorations à ce modèle dont l'une résout ce problème . L'auteur propose de remplacer le paramètre c de la fonction de dominance par un facteur différent pour chaque phonème et ayant une valeur différente selon qu'il s'agisse de l'anticipation ou de la rétention (3.6). En outre, Cosi a ajouté un terme de résistance $R(\tau)$ à la coarticulation qui permet comme dans le cas des bilabiales de forcer la fonction résultante à atteindre la cible. Enfin, un terme $s(\tau)$ a été ajouté afin de contrôler les mouvements articulatoires au voisinage de la cible. L'équation résultante de la combinaison de fonctions de dominance (l'équation 3.5) devient l'équation 3.7.

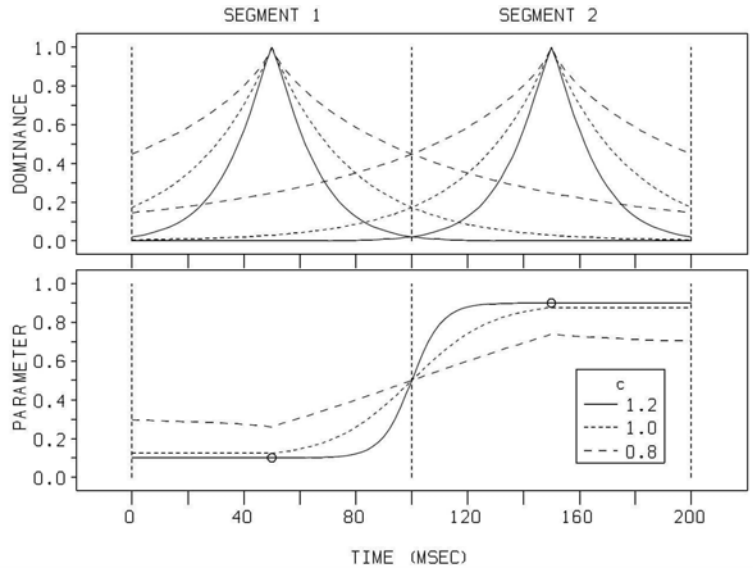

FIG. 3.12 – Influence du paramètre c. Figure extraite de Cohen et Massaro [19].

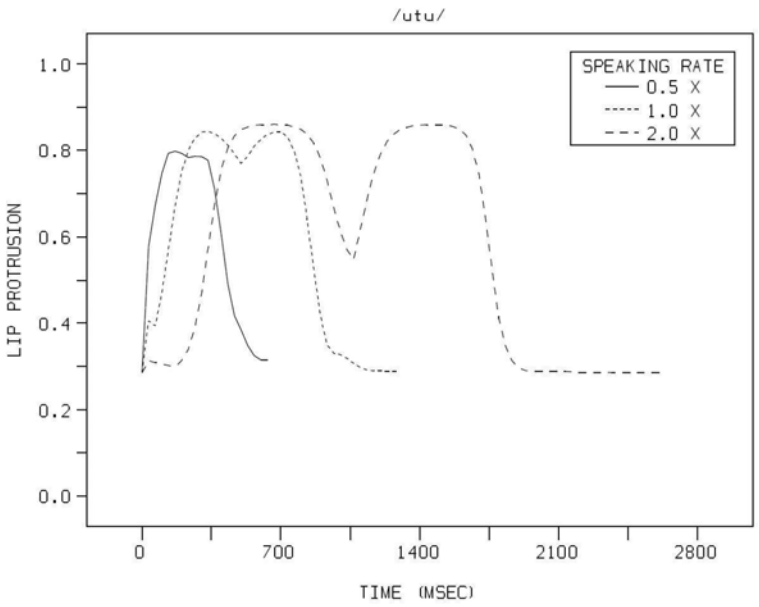

FIG. 3.13 – Evolution de la protrusion des lèvres prédite par le modèle de Cohen et Massaro pour la séquence /utu/ à différentes vitesses d'élocution. Figure extraite de [19].

$$D(\tau) = \begin{cases} \alpha\ e^{-\theta_{bw}|\tau|^{c_{bw}}} & \texttt{si}\ \ \tau \leq 0 \\ \alpha\ e^{-\theta_{fw}|\tau|^{c_{fw}}} & \texttt{si}\ \ \tau > 0 \end{cases} \tag{3.6}$$

$$F_p(t) = \frac{\sum_{i=1}^{N}(T_i.S_i(t-t_i).R_i(t-t_i).D_i(t-t_i)}{\sum_{i=1}^{N} R_i(t-t_i).D_i(t-t_i)} \tag{3.7}$$

Cosi semble avoir corrigé les principaux défauts du modèle de Cohen et Massaro, mais les nombreux paramètres obtenus rendent les fonctions de dominance très complexes et le modèle devient difficilement inversible, ce qui oblige lors de la phase d'apprentissage à définir une multitude de paramètres manuellement, rendant ainsi l'opération délicate et longue.

En conclusion, la technique proposée par Cohen et Massaro ou l'amélioration suggérée par Cosi doit donc s'appliquer à chacun des articulateurs. Le fait de ne pas tenir compte de la dépendance entre articulateurs risque néanmoins de provoquer des erreurs dans la synthèse des mouvements, notamment à cause d'erreurs de synchronisation. Le Goff [35] a choisi de tenir compte de l'influence mutuelle des articulateurs pour la synthèse visuelle de la parole en considérant non plus chacun des paramètres articulatoires pris isolément, mais en choisissant le visème comme paramètre de base. Le problème posé ici est l'attribution de visèmes aux phonèmes qui n'ont pas de contrainte labiale particulière. Même si on choisit les consonnes dans un contexte VCV, cela signifie que l'on ne prend en compte que le phonème précédent et le phonème suivant. Les autres phonèmes vont quand même être pris en compte lors du moyennage final des fonctions de dominance, mais de manière différente. En effet, une partie du phénomène de coarticulation est prise en compte dans le visème lui même et une autre partie est prise en compte lors du calcul final de la moyenne et par conséquent, rien ne garantit la cohérence globale de l'approche.

3.3.3 Modèle d'Öhman

Öhman [42] a défini un modèle particulier pour évaluer l'ampleur de la coarticulation des mouvements de la langue. Selon lui, les mouvements sont pilotés par les voyelles auxquelles se superpose le geste consonantique. Une analogie peut être faite avec le procédé de modulation qui correspond à la superposition d'un ensemble de gestes rapides (mouvement consonantal) sur des gestes plus lents (celui des voyelles). La famille de ce modèle est de type Look-Ahead car le début du geste commence aussi tôt que possible tant qu'il n'y a pas d'antagonisme. Reveret et al . [74] ont adopté ce modèle pour la modélisation de la coarticulation d'une tête parlante en français.

3.3.3.1 Description de la méthode utilisée

Öhman échantillonne le conduit vocal en cinquante segments numérotés de 0 à 49 comme le montre la figure 3.14. Pour caractériser la forme du conduit vocal, il définit $s(x,t)$ qui est la distance entre l'origine du segment x ($0 \leq x \leq 49$) et le point d'intersection avec le contour de la langue en fonction du temps. Une première approximation de la coarticulation pour une séquence VCV est d'écrire :

$$s(x,t) = (1 - k(t)).s(x,v) + k(t).s(x,c) \tag{3.8}$$

Dans cette formule,v représente l'influence de la voyelle et c celle de la consonne ; le paramètre k varie en fonction du temps de 0 à 1 ; il vaut 1 au niveau de la consonne (à la constriction) et 0 au niveau de la voyelle.

L'inconvénient de cette modélisation est qu'elle privilégie la consonne. En effet, quand k=1, $s(x) = s(x;c)$, c'est-à-dire que seule la consonne intervient, ce qui ne correspond pas à la réalité. Dans le cas par exemple de la prononciation des séquences /ada/ et /udu/, la voyelle intervient évidemment sur la forme de la langue pendant l'occlusion.

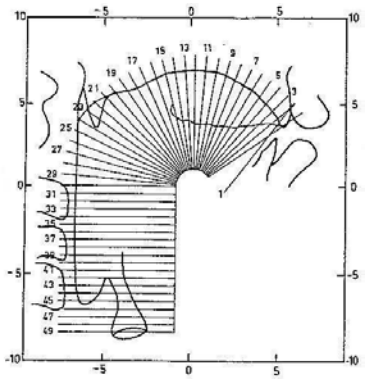

FIG. 3.14 – Echantillonnage du conduit vocal (selon Öhman).

Pour résoudre ce problème, l'idée consiste à séparer la consonne en deux objets invariants $c(x)$ et $w_c(x)$. $c(x)$ représente la cible si la consonne était appliquée seule alors que $w_c(x)$ correspond à des valeurs entre 0 et 1 et représente l'influence d'une voyelle arbitraire. L'équation 3.8 devient :

$$s(x,t) = v(x) + k(t).[c(x) - v(x)]w_c(x) \tag{3.9}$$

Ici, $c(x)$ et $w_c(x)$ ne dépendent pas du temps. Comme la séquence VCV est symétrique, la fonction $v(x)$ est aussi indépendante du temps.

Une approximation au premier ordre de $v(x)$ donne :

$$v(x) = \alpha a(x) + \beta u(x) + \gamma i(x) \tag{3.10}$$

avec $\alpha + \beta + \gamma = 1$ et $0 \leq \alpha, \beta, \gamma \leq 1$

$a(x)$, $u(x)$ et $i(x)$ sont les cibles associées aux voyelles /a/, /u/ et /i/. Il suffit de fixer 2 des coefficients α, β ou γ pour trouver le troisième. On peut exprimer les coefficients en fonction de 2 variables q_1 et q_2 comme suit :

$$\begin{gathered} \alpha = q_1 \\ \beta = (1 - q_1)q2 \\ \gamma = (1 - q_1)(1 - q_2) \\ 0 \leq q1, q_2 \leq 1 \end{gathered} \tag{3.11}$$

Il est aussi possible de modéliser une séquence V_1CV_2 où $V_1 \neq V_2$. L'équation 3.9 devient :

$$s(x,t) = v(x,t) + k(t)[c(x) - v(x,t)]w_c(x) \tag{3.12}$$

Dans cette équation, l'absence de symétrie fait apparaître le facteur temps dans l'expression $v(x,t)$

3.3.3.2 Les essais réalisés

Une analyse aux rayons X de différentes séquences comportant les voyelles /y/, /a/ et /u/ avec la consonne /d/ a été réalisée et les différences entre la forme du conduit vocal théorique et réel ont été calculées.

D'une façon générale, la pratique vérifie que le geste de la voyelle a lieu en même temps que celui de la consonne. Dans plusieurs cas de séquences VCV asymétriques, la voyelle n'est pas terminée au relâchement de la constriction, c'est-à-dire quand k commence à décroître. En outre, le paramètre de la voyelle a commencé à prendre des valeurs non nulles avant que la condition de fermeture ($k = 1$) soit atteinte.

3.3.3.3 Cas particulier de la consonne /g/

Une analyse aux rayons X de la séquence /ugu/ et /ygy/ montre que les régions de contact entre la langue et le palais sont différentes et ne se chevauchent pas. Par conséquent, il n'est pas possible de construire une cible indépendante des voyelles $g(x)$ pour cette consonne. Une solution consiste à construire une famille de formes $g(x;\theta)$ et de fonctions de coarticulation $w_g(x;\theta)$ qui dépendent du paramètre voile/palais θ.

En rendant θ égal à q_2, l'équation (3.9) devient

$$s(x,t) = v(x;q_1,q_2) + k(t).[g(x;q_2) - v(x;q_1,q_2)]w_g(x;q_2) \tag{3.13}$$

Cette équation peut faire penser que la notion d'indépendance entre la cible consonantique et la voyelle n'est plus validée, mais Öhman suppose que le point de fermeture de la dorsale doit pouvoir être contrôlé par le locuteur indépendamment de la voyelle.

Pour conclure, le modèle d'Öhman a été défini pour modéliser la coarticulation des mouvements de la langue et ne semble pas parfaitement adapté à la modélisation de la coarticulation labiale. Par exemple, l'échantillonnage du conduit vocal a été réalisé sans tenir compte des lèvres. Par conséquent, le modèle d'Öhman considère de la même façon les voyelles /y/ et /i/ qui ne diffèrent que par la protrusion des lèvres.

3.3.4 Synthèse à base de HMM (Hidden Markov Model)

Les méthodes VQ (Vector Quantification) ou ANN (Artificial Neural Networks) sont basées sur une correspondance échantillon par échantillon entre les paramètres de parole et les paramètres visuels. Ces techniques ne permettent pas de prendre en compte le contexte et peuvent produire des discontinuités dans la synthèse visuelle. En revanche, les HMM constituent une approche statistique très intéressante pour la gestion de l'aspect temporel qui permet sous certaines conditions de prendre en compte l'aspect dynamique, donc la coarticulation.

Un système basé sur les HMM permettant d'animer les mouvements faciaux à partir de données acoustiques a été proposé par Simons et Cox [79]. Ce modèle utilisait un HMM à 16 états, chacun correspondant à une forme labiale prédéfinie. Le nombre de formes de lèvres était donc très limité et aucune information n'était fournie en ce qui concerne les positions des dents ou de la langue. Ce modèle ne prenait pas non plus en compte l'aspect dynamique et ne gérait donc pas de façon convenable le phénomène de coarticulation.

Plutôt que de représenter les états par une forme labiale prédéfinie, Yamamoto [88] a choisi d'utiliser un HMM sur les données acoustiques et d'en déduire les paramètres labiaux en utilisant une synchronisation entre les paramètres labiaux et les paramètres visuels puis en moyennant

les paramètres des images associées avec l'état HMM correspondant. Les figures 3.15 et 3.16 montrent ce processus. L'apprentissage est réalisé en trois étapes :

- Apprentissage des HMM en utilisant la base de données de parole
- Alignement de la parole d'entrée en une séquence d'états HMM en utilisant le décodage de Viterbi
- Moyenne des paramètres visuels des trames associées à chaque état

Le synthèse nécessite également trois étapes :
- Alignement de la parole entrée en une séquence d'états HMM en utilisant le décodage de Viterbi
- Récupération du paramètre image associé
- Concaténation avec les paramètres récupérés

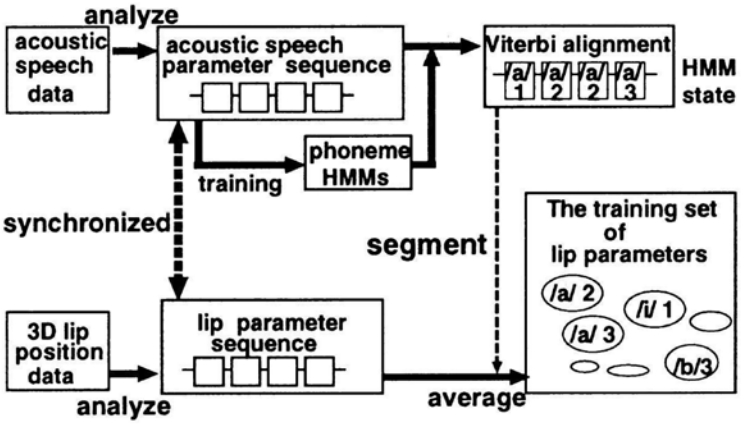

FIG. 3.15 – Schéma de l'apprentissage des paramètres labiaux à partir d'une méthode basée sur les HMM. D'après E. Yamamoto [88].

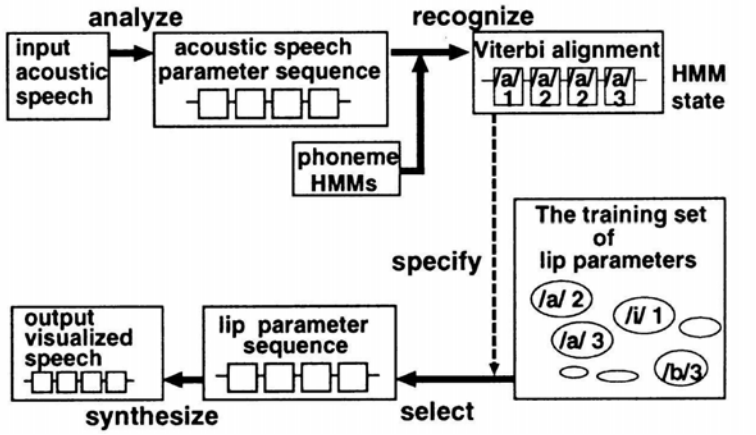

FIG. 3.16 – Schéma de la synthèse des paramètres labiaux à partir d'une méthode basée sur les HMM. D'après E. Yamamoto [88].

Bien que cette technique soit très intéressante, elle ne gère pas non plus l'aspect dynamique. Une première solution consiste à utiliser des biphones ou des triphones comme base d'apprentissage, mais néanmoins, l'effet global de la coarticulation qui peut influencer plusieurs phonèmes adjacents ne peut pas être pris en compte.De plus, l'utilisation de biphones ou triphones augmente considérablement la base d'apprentissage. Yamamoto a proposé une méthode nommée SV_HMM pour prendre en considération le visème suivant tout en conservant une base d'apprentissage basée sur des monophones [89]. Cependant, nous pouvons constater que la seule prise en compte du phonème suivant est insuffisante pour modéliser l'ampleur de la coarticulation. Benguerel et Cowan [7] ont notamment montré dans une étude sur l'anticipation de l'arrondissement du /u/ en français que l'anticipation pouvait commencer 6 segments avant la voyelle cible.

Une solution consiste à intégrer les paramètres dynamiques dans l'équation de résolution du HMM. L'un des premiers à l'avoir réalisé dans le domaine de la parole est Keiichi Tokuda [84]. Soit $O = o_1, o_2, ..., o_T$ le vecteur d'observation et $q = q_1, q_2, ..., q_t$ le vecteur d'état, Tokuda considère le paramètre o_t comme la combinaison des caractéristiques du vecteur statique c_t, c'est-à-dire des coefficients cepstraux et des propriétés dynamiques Δc_t, donc $o_t = \{c_t, \Delta c_t\}$. Tamura [83] a adapté ce modèle dynamique à la synthèse visuelle de la parole. Les paramètres dynamiques Δc_t sont calculés en effectuant une somme pondérée des paramètres adjacents ($\Delta c_t = \sum_{\tau=-L^-}^{L^+} w(\tau) c_{t+\tau}$. Dans la formule, $w(\tau)$ correspond au coefficient de pondération. Le choix de l'intervalle L est délicat car une valeur trop faible risque de ne pas considérer toute l'étendue de la coarticulation alors qu'une trop grande valeur risque de provoquer un lissage exagéré du résultat.

Plutôt que de prendre en compte des triphones, Zhou [90] a choisi de rester indépendant du contexte, mais d'inclure une fonction de trajectoire. Cette technique permet sur le corpus choisi par l'auteur de diminuer le taux d'erreurs de 0.03 à 0.06%. L'avantage de cette technique est de pouvoir tenir compte des effets de la coarticulation s'étendant sur davantage de phonèmes. Néanmoins, les résultats obtenus ne sont pas très significatifs.

Govokhina [36] a proposé une méthode basée à la fois sur les HMM et sur une technique de concaténation. La synthèse de la parole par concaténation consiste en la sélection d'unités pré enregistrées dans un dictionnaire. Tout d'abord, des caractéristiques phonologiques sont utilisées pour sélectionner les unités candidates. Ensuite, un algorithme de programmation dynamique trouve un chemin optimal à travers le treillis de candidats qui minimise un coût cumulé de sélection et de concaténation. L'originalité de la technique utilisée par Govokhina est l'utilisation des HMMs pour déterminer les meilleurs candidats dans le treillis pour la sélection finale du modèle de concaténation. Malheureusement, cette solution semble peu efficace car les résultats obtenus sont moins bons que l'application de la méthode HMM seule. L'auteur remarque cependant que les résultats sont fortement dépendants du contenu phonétique des phrases et que la solution par concaténation génère des trajectoires mieux "coarticulées" que celles produites par HMM.

3.4 Conclusion

Ce chapitre nous a permis de constater la vaste étendue des techniques de modélisation des visages parlants et des méthodes de prédiction de la coarticulation. Les méthodes basées sur les HMM ou les réseaux de neurones ont largement fait leurs preuves dans le domaine de la parole, mais ne semblent pas parfaitement adaptées à la modélisation de la coarticulation. Les modèles d'Öhman ou ceux basés sur les fonctions de dominances sont intéressants mais rendent difficiles le contrôle local des mouvements, comme par exemple le forçage de la fermeture des lèvres pour

les bilabiales. En conséquence, nous avons choisi de développer un modèle de prédiction basé sur des règles. Le chapitre suivant va nous permettre d'identifier nos paramètres de contrôles et de déterminer leurs caractéristiques intrinsèques en étudiant leur variabilité intra et inter locuteurs.

Chapitre 4

Variabilité intra et inter locuteurs du phénomène de coarticulation

Les différentes théories présentées précédemment montrent qu'il n'existe pas de consensus au sujet de la caractérisation du phénomène de coarticulation. Le degré de variabilité de ce phénomène explique sans doute en partie les différences obtenues entre les expériences réalisées. Ne disposant pas de corpus audiovisuel permettant de comparer les mouvements labiaux de plusieurs locuteurs et d'estimer les variations pour un locuteur donné, nous avons choisi d'enregistrer nos propres corpus.

4.1 Choix des corpus

En vue d'étudier la variabilité inter locuteurs, nous avons d'abord enregistré un "mini corpus" avec 10 locuteurs francophones (5 femmes et 5 hommes). Celui-ci est composé de 4 voyelles /i, y, a, o/, de 6 consonnes /p, t, d, s, ʃ, f/ enregistrées isolément, de 8 CV, 20 VCV, 18 VCCV et 2 phrases phonétiquement équilibrées. Chaque séquence prononcée une seule fois est encadrée par des "mots porteurs" destinés à forcer la bouche à prendre une position neutre en début et en fin de séquence. Dans cette étude, les locuteurs ont un style d'articulation relativement neutre et il ne devrait pas y avoir d'hyper ou à d'hypo articulation [54]. La limitation du nombre de combinaisons est due au temps d'enregistrement qui ne devait pas excéder quelques minutes par locuteur et à l'espace de stockage qui pour ce "petit" corpus représente déjà presque 50 Go de données !

Si le "mini corpus" est satisfaisant pour étudier les variations inter locuteurs, nous avons été conduit à enregistrer un "grand corpus" afin d'estimer les variations intra-locuteur. Celui-ci est constitué de toutes les voyelles, semi-voyelles, consonnes et CV, toutes les VCV ayant /i, y , a , u/ comme voyelles possibles, les 70 VCCV les plus courantes en français (d'après le vocabulaire de BREF [49]), 22 VCCV de notre choix afin de pouvoir tester des combinaisons spécifiques de phonèmes et 69 phrases phonétiquement équilibrées extraites de la publication de P. Combescure [20]. L'enregistrement a été fait avec une locutrice spécialiste de la lecture labiale ; de ce fait, l'ensemble des séquences sont légèrement hyper articulées afin de favoriser l'intelligibilité visuelle. La locutrice n'a prononcé qu'une fois chaque séquence ce qui nous a permis d'enregistrer un corpus très varié.

4.2 Acquisitions effectuées

Pour chaque corpus, l'acquisition consiste à l'enregistrement audio et vidéo de la voix et des mouvements faciaux. Deux caméras synchronisées fonctionnant à 120 Hz filmant 200 marqueurs peints sur le visage (46 sur les lèvres) ont permis d'obtenir un maillage tridimensionnel du visage [87]. La figure 4.1 montre une paire d'images stéréo et la maillage associé. Les marqueurs sont visibles en blanc. Les points sur le haut du visage sont utilisés pour compenser le mouvement global de la tête.

FIG. 4.1 – Construction d'un maillage 3D à partir de deux images stéréo.

Une analyse en composantes principales (ACP) réalisée sur le grand corpus a permis d'identifier les trois principaux modes de déformation du visage. Le premier représente la protrusion qui représente à lui seul 66.6% de la variance (cf tableau 4.2) . A ce sujet, plusieurs études divergent au sujet de la mesure de la protrusion : avancée de la lèvre supérieure seule, avancée des deux lèvres ou avancée des deux lèvres et des commissures. L'ACP montre que l'avancée des deux lèvres et des commissures interviennent dans le mouvement (Fig. 4.2). Les deux autres principaux modes sont le mouvement de la mâchoire et l'ouverture des lèvres. Outre ces trois paramètres, nous avons aussi choisi de mesurer la distance (en mm) entre les commissures afin d'avoir une bonne estimation de l'étirement des lèvres.

Les mesures des 4 paramètres articulatoires retenus sont ensuite extraites de l'enregistrement. Compte tenu du résultat de l'ACP qui montre que la protrusion est liée à l'avancée des deux lèvres et des commissures, nous avons estimé ce paramètre en calculant la distance entre le centre de gravité des quatre points A,B, C et D (Fig. 4.3) et un point fixe de référence F. Ce point F est la projection d'un point fixe de la tête sur le vecteur normal $\vec{V_n} = \vec{AB} \wedge \vec{CD}$. Il est important de remarquer que les mesures obtenues correspondent à une distance en millimètres par rapport à ce point de référence, valeurs peu significatives si elles sont prises isolément mais représentatives de l'avancée ou de la rétractation des lèvres si elles sont comparées les unes aux autres.

Le mouvement de la mâchoire est estimé en calculant la distance (en mm) entre un point fixe du visage et une moyenne de 4 points situés sur le menton. Il ne s'agit bien sûr que d'une approximation mais la mesure précise du mouvement de la mâchoire nous aurait forcé à équiper le locuteur de capteurs qui auraient pu perturber sa prononciation naturelle.

En ce qui concerne l'étirement, nous calculons la distance (en mm) entre les deux commissures (Points A et B de la figure 4.3). La mesure de l'ouverture labiale (c'est l'ouverture verticale qui

Mode	Pourcentage de la variance expliquée	Articulateur associé
1	66.7	Protrusion
2	19.9	Mouvement de la mâchoire
3	3.1	Ouverture des lèvres

TAB. 4.1 – Analyse en composantes principales effectuée sur un grand corpus.

nous intéresse ici) a été réalisée en mesurant la distance entre un point de la lèvre supérieure et un point de la lèvre inférieure (Points C et D de la figure 4.3). Cette distance ne caractérise pas complètement l'ouverture labiale mais en donne une bonne estimation. Une étude récente dont les conclusions seront publiées prochainement nous a permis d'extraire avec davantage de précision les paramètres d'ouverture et d'étirement. Les premiers résultats montrent que le degré d'ouverture reste assez voisin de celui estimé avec notre étude.

Les résultats mesurés en millimètres sont ensuite centrés et normalisés afin d'être indépendant des caractéristiques anatomiques de la personne. Chaque paramètre X est converti en $\frac{X-\mu}{\sigma}$. μ représente la valeur moyenne de ce paramètre sur tous les segments de parole du sujet considéré et σ représente l'écart type.

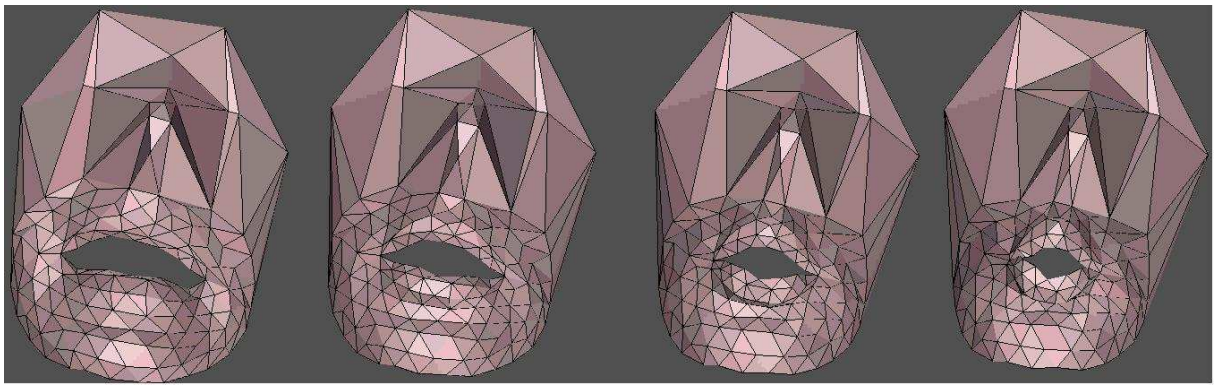

FIG. 4.2 – 1er mode de l'analyse en composantes principales des déformations du visage sur l'ensemble du grand corpus.

4.3 Etude des variations inter-locuteurs

La diversité des théories au sujet de la caractérisation du phénomène d'anticipation nous a poussé à réaliser une étude comparative de ce phénomène avec le mini corpus de 10 locuteurs. Nous avons choisi de caractériser l'anticipation de la protrusion sur des séquences de type VCV et VCCV (prononcées isolément une seule fois par chaque locuteur). Conformément à l'étude de Benguerel et Cowan ([7]), nous avons considéré le début du phénomène de protrusion comme étant l'instant associé au maximum d'accélération. Les figures 4.4 et 4.5 montrent les instants de début et de maximum de protrusion pour les 10 locuteurs. Nous pouvons constater qu'aucune stratégie d'anticipation ne se dégage réellement. Pour les séquences /iCy/ où C correspond à une consonne ayant peu d'influence labiale (par exemple /t/, /s/, /d/), on constate que le mouvement de protrusion commence pour la majorité des sujets au début de la consonne pour se terminer

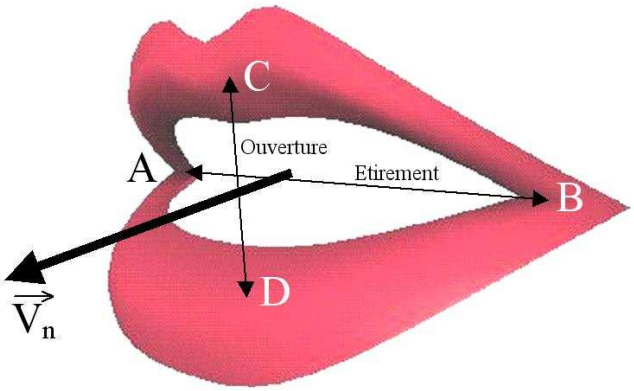

FIG. 4.3 – Mesures de l'ouverture et de l'étirement.

pour 8 locuteurs sur 10 dans la première moitié du /y/. La théorie de Abry et Lallouache ([2]) selon laquelle le mouvement de protrusion est expansible et relativement peu compressible semble vérifiée car les sujets qui ont le débit le plus rapide terminent le mouvement de protrusion à la fin de la voyelle /y/. L'instant de début du mouvement de protrusion semble être davantage lié à la stratégie du locuteur. Par exemple, les locuteurs 5 et 6 prononcent la séquence /isy/ dans le même laps de temps. Alors que le premier commence son mouvement de protrusion au milieu de la consonne, le second l'anticipe beaucoup plus. Ce phénomène n'est pas ponctuel puisque cette différence de stratégie se vérifie sur la grande majorité des séquences étudiées.

Cette analyse multilocuteur a également permis de montrer que l'influence de certains phonèmes varie fortement selon les locuteurs. Par exemple, dans une séquence /ify/, les phonèmes /i/ et /y/ ont toujours une influence sur la protrusion alors que /f/ l'influence seulement pour certains locuteurs. La figure 4.6 montre la séquence /ify/ prononcée par 3 locuteurs ayant des stratégies très différentes. Si le phonème /f/ ne semble pas influencer la protrusion du locuteur 4, ce n'est pas le cas du locuteur 1 pour qui l'augmentation de la protrusion est freinée au niveau de /f/. En ce qui concerne le locuteur 5, le sens de variation de l'arrondissement s'inverse même au niveau de /f/.

Afin de cerner davantage cette variation inter-locuteurs, nous avons réalisé une étude comparative de la corrélation entre les données centrées normalisées. Nous avons choisi de comparer l'évolution de la protrusion pour la séquence /yk∫i/. Cette séquence difficile à prononcer a été choisie volontairement car elle occasionne de fortes transitions du mouvement de protrusion. La mesure de la corrélation entre locuteurs (Fig. 4.7) donne des résultats variant de 0,60 à 0,97. Trois principaux groupes se dégagent. Les locuteurs 1 à 7 ont une forte dépendance entre eux. Les locuteurs 9 et 10 qui ont tendance à davantage marquer la protrusion du phonème /∫/ se détachent des autres locuteurs. Le locuteur 8 représente une situation intermédiaire entre les deux présentées précédemment.

Concernant la coarticulation, cette étude a permis de montrer que les locuteurs ont des stratégies propres ; certains anticipent aussi tôt que possible et peuvent donc être associés au modèle Look Ahead alors que d'autres commencent le mouvement d'anticipation beaucoup plus tard. Aucun des trois principaux modèles théoriques (Look-Ahead, Time-Locked ou Hybrid) n'a

réellement pu être validé par notre étude mais la théorie du mouvement expansionniste ([2]) n'est pas démentie.

4.4 Etudes des variations intra-locuteurs

Afin d'étudier la variabilité intra-locuteur, nous avons examiné une série de séquences appariées de type VCV ou VCCV. Seule une consonne diffère entre chaque paire de logatomes. Par exemple, /ity/ est comparée à /idy/, /ikʃy/ est comparée à /itʃy/, /ykʃi/ est comparée à /ytʃi/. Pour étudier la variabilité intra-locuteur, il aurait été certes plus judicieux de comparer les mêmes séquences répétées par le locuteur, mais la limitation du temps d'enregistrement nous a conduit a faire le choix d'effectuer un seul enregistrement par séquence afin de pouvoir diversifier au maximum le corpus. Pour chaque locuteur et chaque paire de logatomes, nous avons ainsi calculé la corrélation entre l'évolution des paramètres de protrusion, d'ouverture et d'étirement (Fig. 4.8). On peut vérifier que le taux de corrélation intra-locuteur est très important. Ceci est dû d'une part au fait que les consonnes échangées entre les deux logatomes n'ont pas une forte influence labiale et ceci prouve également que la stratégie coarticulatoire d'un locuteur reste sensiblement constante. La figure 4.9 montre l'évolution du mouvement de protrusion de séquences VCV voisines pour deux locuteurs ayant des stratégies propres. Si Odile commence le mouvement de protrusion dès la première voyelle, ce n'est pas le cas de Blaise qui le débute au niveau de la consonne.

Afin d'évaluer la différence entre les variations intra et inter-locuteurs, nous avons calculé la corrélation moyenne entre le profil d'un locuteur et ceux des autres (Fig. 4.8). La corrélation intra-locuteur est très forte (92% en moyenne) et est presque toujours supérieure à la corrélation inter-locuteurs (85 % en moyenne). Seul le locuteur 5 a une stratégie peu stable puisqu'en ce qui le concerne, la corrélation inter-locuteurs est plus forte que la corrélation intra-locuteur.

En conclusion, pour une séquence donnée et comme on pouvait s'y attendre, la variation des paramètres labiaux est prévisible avec une incertitude moins grande si l'on modélise un locuteur précis car dans ce cas, il sera possible de reproduire avec davantage de précision sa stratégie coarticulatoire.

4.5 Variabilité intra et inter-locuteurs des phonèmes

La variation et l'évolution des paramètres labiaux vus précédemment sont fortement dépendantes du degré de liberté de chaque phonème. Certains ont une forte influence sur les paramètres labiaux (par exemple, le phonème /p/ contraint fortement l'ouverture labiale) alors que d'autres semblent beaucoup plus neutres (par exemple /t/). Afin de vérifier cette hypothèse, nous avons réalisé une étude statistique sur le grand corpus enregistré par une seule locutrice afin de dégager les principales contraintes imposées par chacun des phonèmes.

4.5.1 Caractéristiques intrinsèques des phonèmes

Les figures 4.10-4.11-4.12-4.13 montrent les valeurs moyennes centrées réduites et les écarts type des paramètres de protrusion, d'ouverture de mâchoire et d'étirement en fonction du phonème. L'ensemble de la plage de variation temporelle d'un phonème donné est prise en compte lors du calcul de la moyenne. Nous avons fait le choix de ne pas prendre en compte uniquement la

valeur centrale de chaque phonème, car compte-tenu de la coarticulation, celle-ci ne correspond pas forcément à la valeur cible.

En ce qui concerne la protrusion, une nette discrimination est visible entre les voyelles non arrrondies /i, a, e, ɛ / qui sont clairement séparées des autres voyelles. Une protrusion importante est également vérifiée pour les semi voyelles /ɥ/ et /w/. En ce qui concerne les consonnes, /ʃ/ et /ʒ/ ont des valeurs moyennes de protrusion assez élevées alors que les autres consonnes sont relativement neutres vis-à-vis de ce paramètre. Globalement, la variabilité des consonnes pour ce paramètre est plus importante que pour les voyelles ce qui prouve que leur degré de résistance à la coarticulation est moindre.

L'ouverture labiale discrimine fortement les bilabiales (/p , b , m/ et les dentales /f, v/) qui ont une forte résistance à la coarticulation. La forte ouverture de /ʃ/ et /ʒ/ est biaisée car notre mesure de l'ouverture est légèrement couplée à la protrusion. En ce qui concerne les voyelles, on peut vérifier sur les figures une forte ouverture pour le /a/ et /ɛ̃/, mais ces derniers ont une large plage de variabilité , contrairement au phonème /e / qui est beaucoup plus contraint.

La mâchoire n'apporte pas beaucoup plus d'informations que l'ouverture sauf pour le phonème /a/ qui impose une forte utilisation de cet articulateur. La prédiction de l'ouverture des lèvres donnera donc de bonnes indications sur les mouvements de la mâchoire.

L'étirement, quant à lui, permet de bien discriminer les voyelles /i, a, e, ɛ/ et la semi voyelle /w/. Les consonnes, mis à part les labiales et dentales, ont une large variabilité pour ce paramètre.

A la vue de ces résultats, les trois paramètres labiaux : Protrusion, Ouverture et Etirement sont fortement liés à l'identification des phonèmes ayant des contraintes labiales. L'étude multi-locuteur va nous permettre de valider ces conclusions sur les principaux phonèmes.

4.5.2 Robustesse de la discrimination

Pour estimer la robustesse des conclusions précédentes réalisées sur l'enregistrement du grand corpus avec une seule locutrice, nous avons extrait les valeurs moyennes de la protrusion, de l'ouverture des lèvres et de leur étirement du corpus multilocuteur. Nous pouvons remarquer sur la figure 4.14 que les conclusions du paragraphe précédent restent valables. Par exemple, la voyelle /i/ est caractérisée en moyenne par un fort étirement, une faible protrusion et une ouverture assez moyenne. La comparaison entre locuteurs des figures 4.15 et 4.16 vient confirmer que même si les stratégies des locuteurs sont différentes, les principes généraux qui caractérisent les phonèmes ayant une influence labiale sont respectés. On constate par exemple que chez tous les locuteurs, le phonème /y/ se caractérise par une forte protrusion, une faible ouverture et un faible étirement. En ce qui concerne les bilabiales, par exemple le phonème /p/, on peut vérifier que le niveau d'ouverture est très faible, l'étirement est faible et la protrusion moyenne. Notons que dans ce cas, la protrusion est liée à la fermeture des lèvres dont le phénomène de "pincement" pour les bilabiales génère indirectement une avancée des lèvres. Ce n'est pas le cas des labio-dentales dont la fermeture des lèvres s'accompagne d'un très faible niveau de protrusion. Dans le cas des phonèmes qui n'ont pas une forte influence labiale, par exemple /t/ ou /d/, on constate que les valeurs moyennes des paramètres choisis sont proches des valeurs neutres sans véritablement d'influence d'un paramètre particulier.

4.5.3 Classification des phonèmes

L'ensemble de ces résultats ainsi que les positions articulatoires des différents phonèmes du français (Fig. 4.17) extraite de l'ouvrage [57]) nous a conduit à établir une classification des niveaux de protrusion, d'ouverture et d'étirement pour les phonèmes ayant des contraintes labiales. Le tableau 4.18 présente ce classement sur une échelle de 0 à 4 choisie arbitrairement. Nous avons volontairement réservé le niveau 0 aux phonèmes ayant une contrainte forte empêchant l'augmentation du paramètre considéré. Nous pouvons constater qu'en ce qui concerne le paramètre d'ouverture, seules les consonnes bilabiales imposent la fermeture des lèvres. Il est clair que ce tableau ne donne qu'un ordre de grandeur des différents paramètres articulatoires ; une adaptation en fonction de la stratégie du locuteur devra être réalisée.

Nos données acquises à la fois sur le mini-corpus de 10 locuteurs et le grand corpus d'une locutrice montrent en outre une forte dépendance entre les paramètres articulatoires. En particulier, le phénomène de protrusion est anticorrélé avec l'étirement des lèvres (coefficient de corrélation de -0.85 sur l'ensemble des données du mini corpus). De plus, comme le montre la figure 4.17, la variation de la protrusion est anticorrélée avec celle de l'ouverture pour les voyelles postérieures et les bilabiales. Dans une moindre mesure, on peut également constater que l'étirement et l'ouverture varient en sens opposé pour les voyelles non arrondies.

4.6 Conclusion

Ce chapitre a permis de constater que même si chaque locuteur a une stratégie propre, des grands principes restent toujours vérifiés en ce qui concerne les mouvements des articulateurs. Ces conclusions vont nous permettre dans le prochain chapitre d'établir un algorithme de prédiction des mouvements coarticulatoires en fonction d'une suite quelconque de phonèmes. Cet algorithme purement formel sera ensuite adapté pour modéliser un locuteur précis ; ce qui permettra ainsi de ne pas construire une tête parlante générique, mais un modèle beaucoup plus complexe. L'intérêt sera par exemple de pouvoir choisir un locuteur ayant une stratégie hyper articulatoire dans le cas où l'on souhaite faciliter la lecture labiale ou de choisir un locuteur ayant une stratégie plus neutre si le visage parlant est destiné à une application grand public dont le seul but est de rendre le message plus attractif.

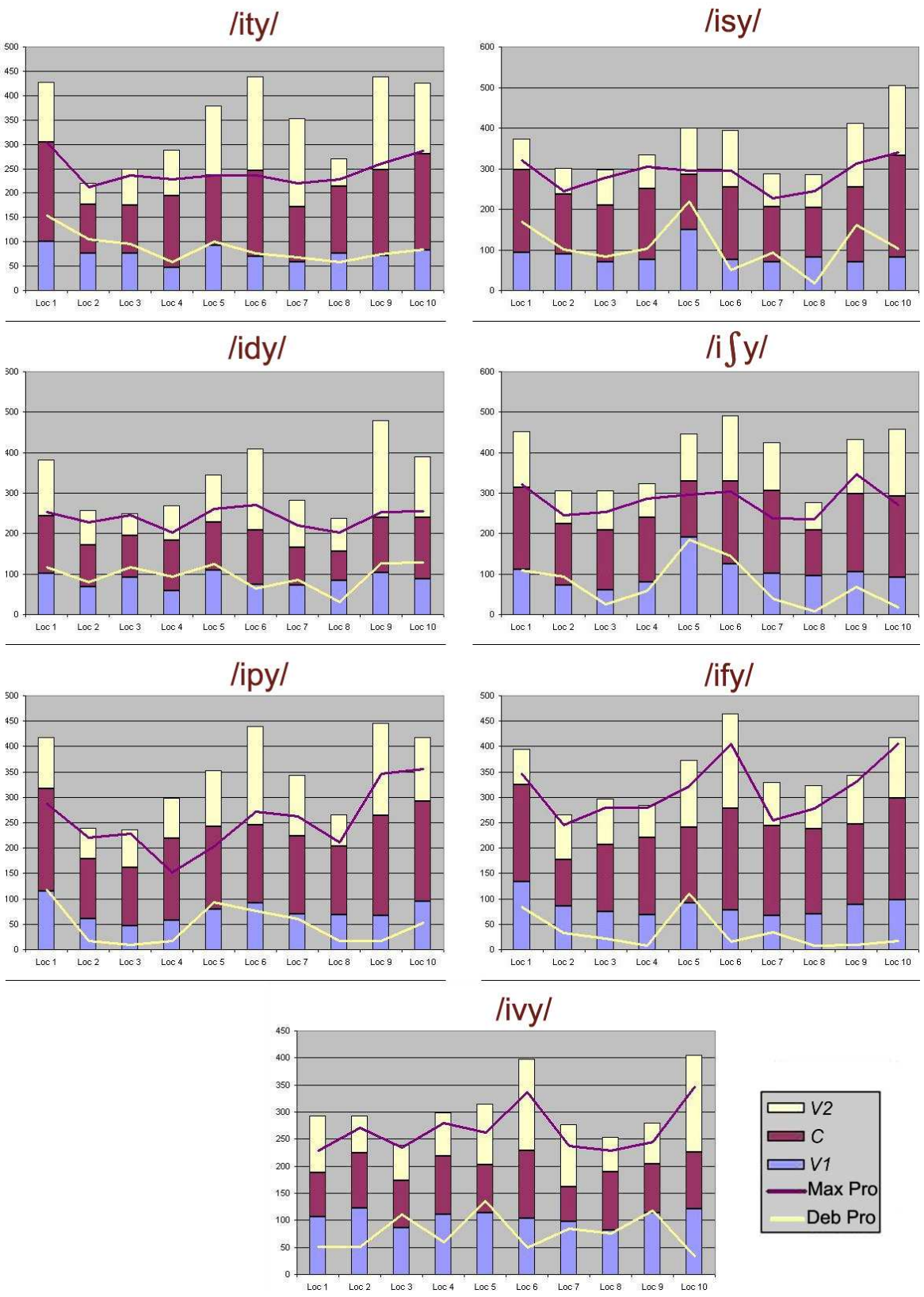

FIG. 4.4 – Instants de début et de maximum de protrusion des VCV pour les 10 locuteurs (ordonnée en ms).

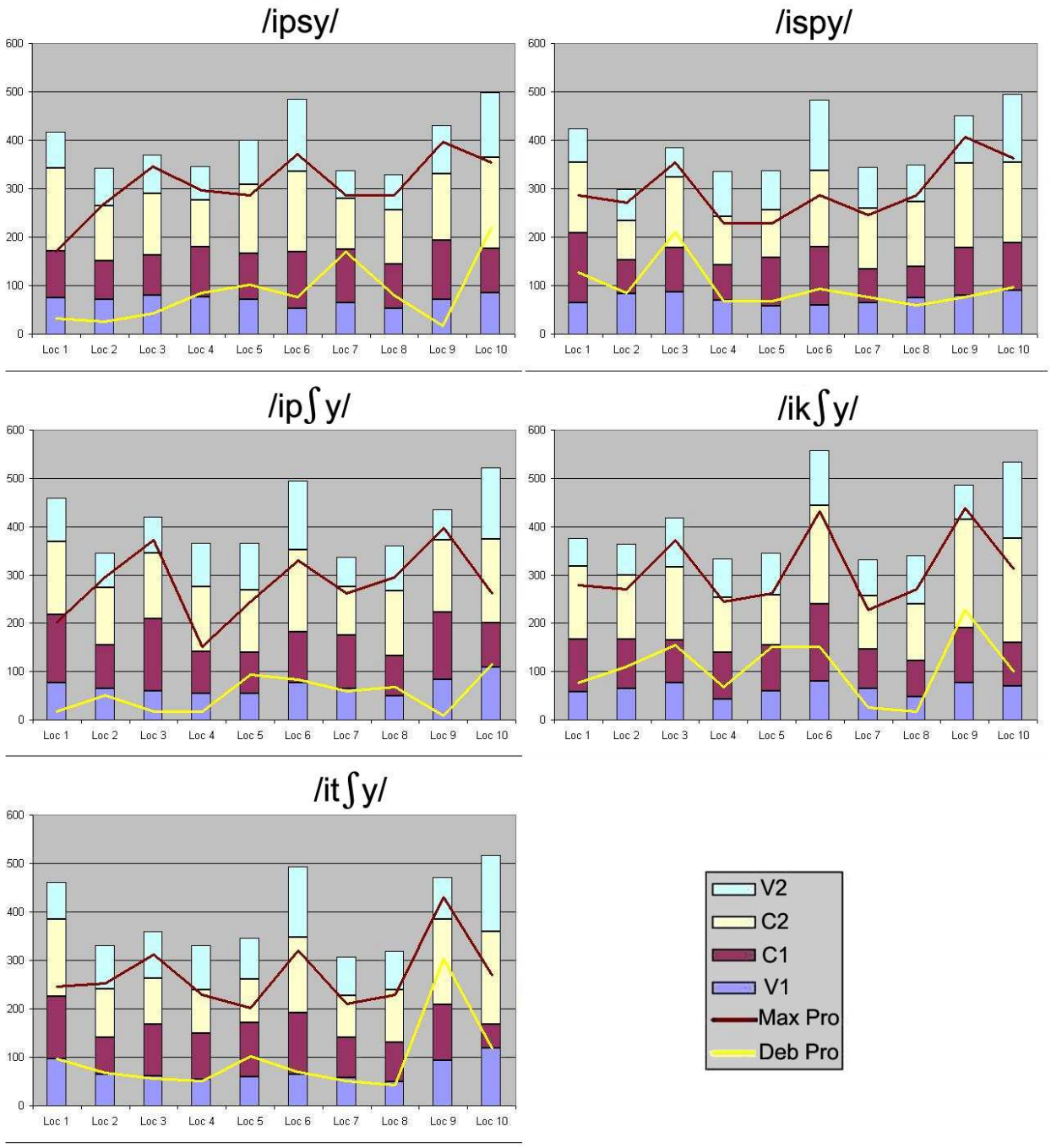

FIG. 4.5 – Instants de début et de maximum de protrusion des VCCV pour les 10 locuteurs (ordonnée en ms).

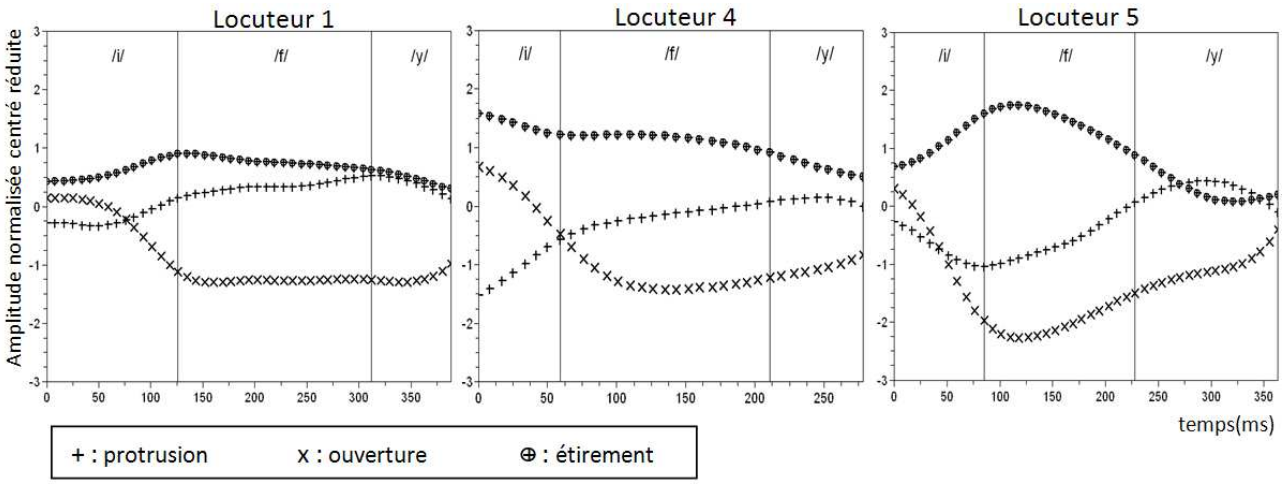

FIG. 4.6 – Evolution des paramètres labiaux pour la séquence /ify/ et pour 3 locuteurs.

	Loc1	Loc2	Loc3	Loc4	Loc5	Loc6	Loc7	Loc8	Loc9	Loc10
Loc1	1	0.97	0.88	0.9	0.97	0.96	0.86	0.85	0.73	0.73
Loc2	0.97	1	0.9	0.95	0.96	0.95	0.89	0.79	0.75	0.81
Loc3	0.88	0.9	1	0.94	0.88	0.91	0.92	0.68	0.78	0.79
Loc4	0.9	0.95	0.94	1	0.89	0.89	0.89	0.64	0.69	0.79
Loc5	0.97	0.96	0.88	0.89	1	0.98	0.85	0.88	0.74	0.75
Loc6	0.96	0.95	0.91	0.89	0.98	1	0.88	0.85	0.78	0.77
Loc7	0.86	0.89	0.92	0.89	0.85	0.88	1	0.73	0.86	0.8
Loc8	0.85	0.79	0.68	0.64	0.88	0.85	0.73	1	0.71	0.6
Loc9	0.73	0.75	0.78	0.69	0.74	0.78	0.86	0.71	1	0.9
Loc10	0.73	0.81	0.79	0.79	0.75	0.77	0.8	0.6	0.9	1

FIG. 4.7 – Corrélations du mouvement de protrusion, entre les 10 locuteurs du corpus pour la séquence /ykʃi/.

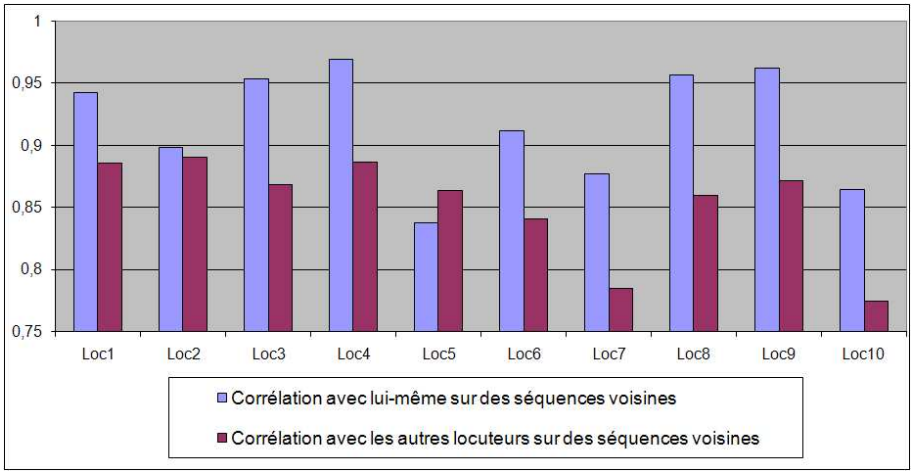

FIG. 4.8 – Corrélations du mouvement de protrusion, d'ouverture et d'étirement (moyenne) entre les 10 locuteurs du corpus pour des séquences voisines.

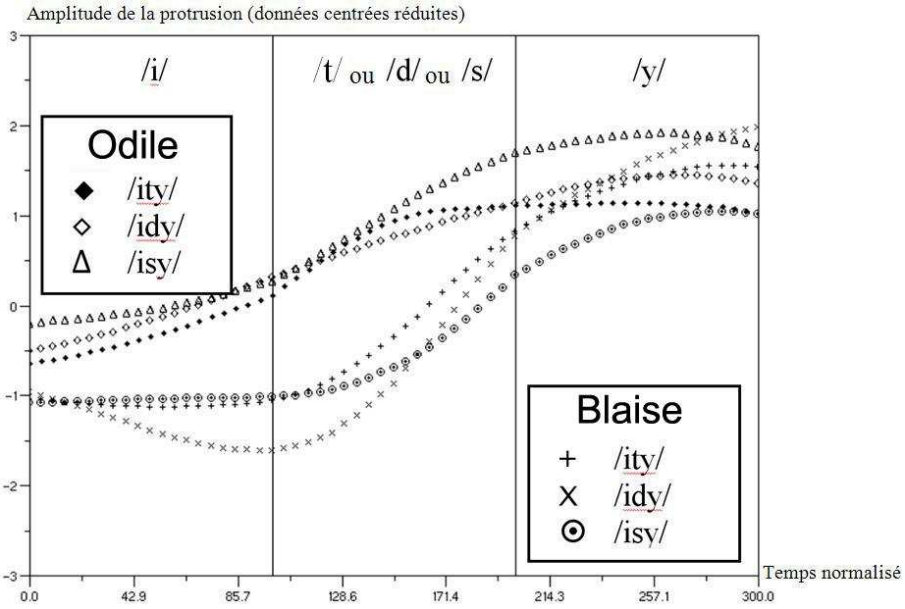

FIG. 4.9 – Evolution du mouvement de protrusion de séquences VCV voisines pour deux locuteurs ayant des stratégies différentes.

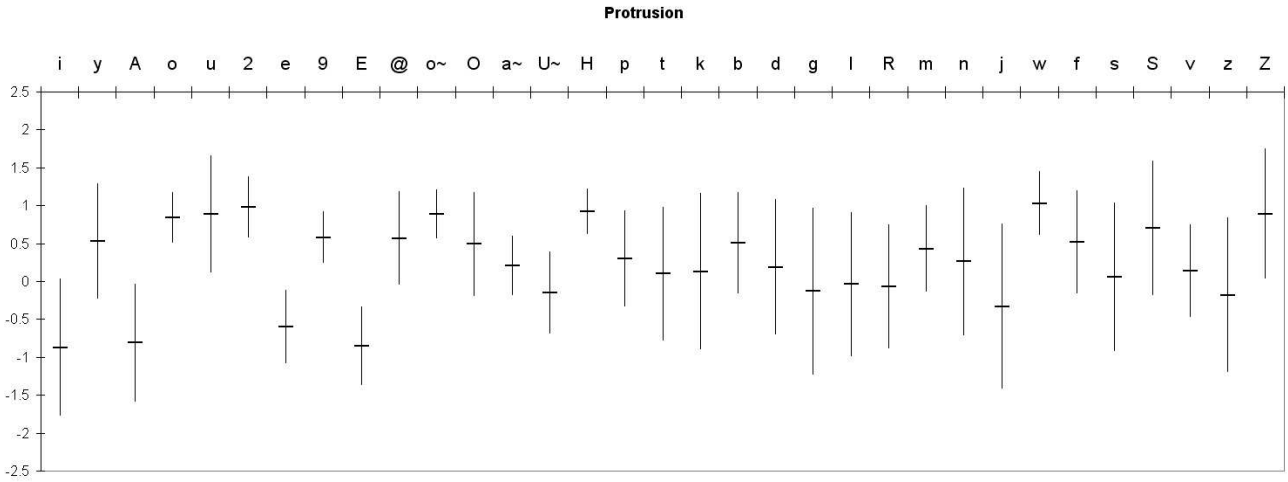

FIG. 4.10 – Mesures de la protrusion (valeurs centrées réduites de la moyenne et de l'écart type) pour chaque phonème d'un grand corpus (Alphabet phonétique SAMPA - cf. Annexe A).

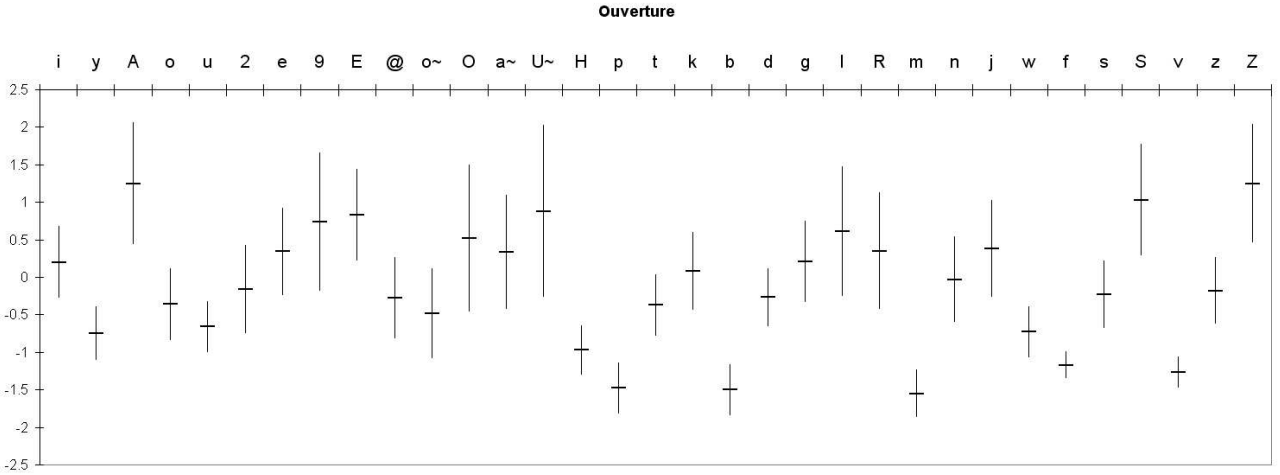

FIG. 4.11 – Mesures de l'ouverture labiale (valeurs centrées réduites de la moyenne et de l'écart type) pour chaque phonème d'un grand corpus (Alphabet phonétique SAMPA - cf. Annexe A).

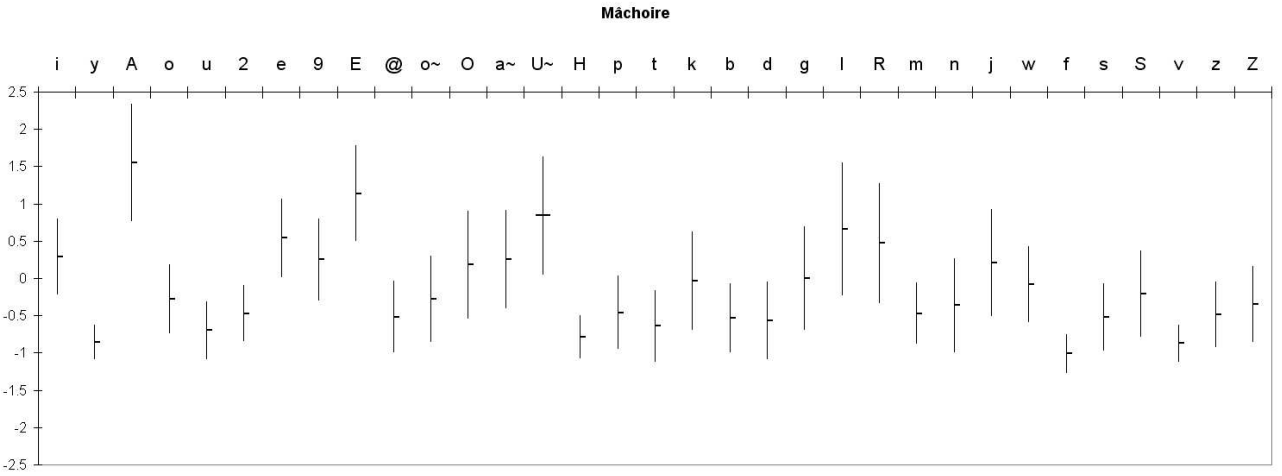

FIG. 4.12 – Mesures de l'ouverture de la mâchoire (valeurs centrées réduites de la moyenne et de l'écart type) pour chaque phonème d'un grand corpus (Alphabet phonétique SAMPA - cf. Annexe A).

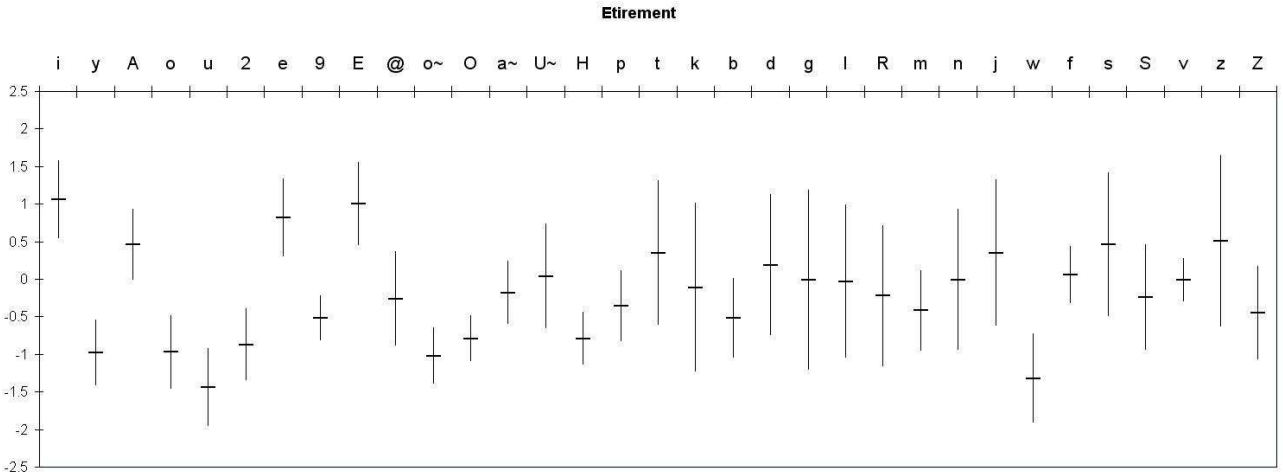

FIG. 4.13 – Mesures de l'étirement des lèvres (valeurs centrées réduites de la moyenne et de l'écart type) pour chaque phonème d'un grand corpus (Alphabet phonétique SAMPA - cf. Annexe A).

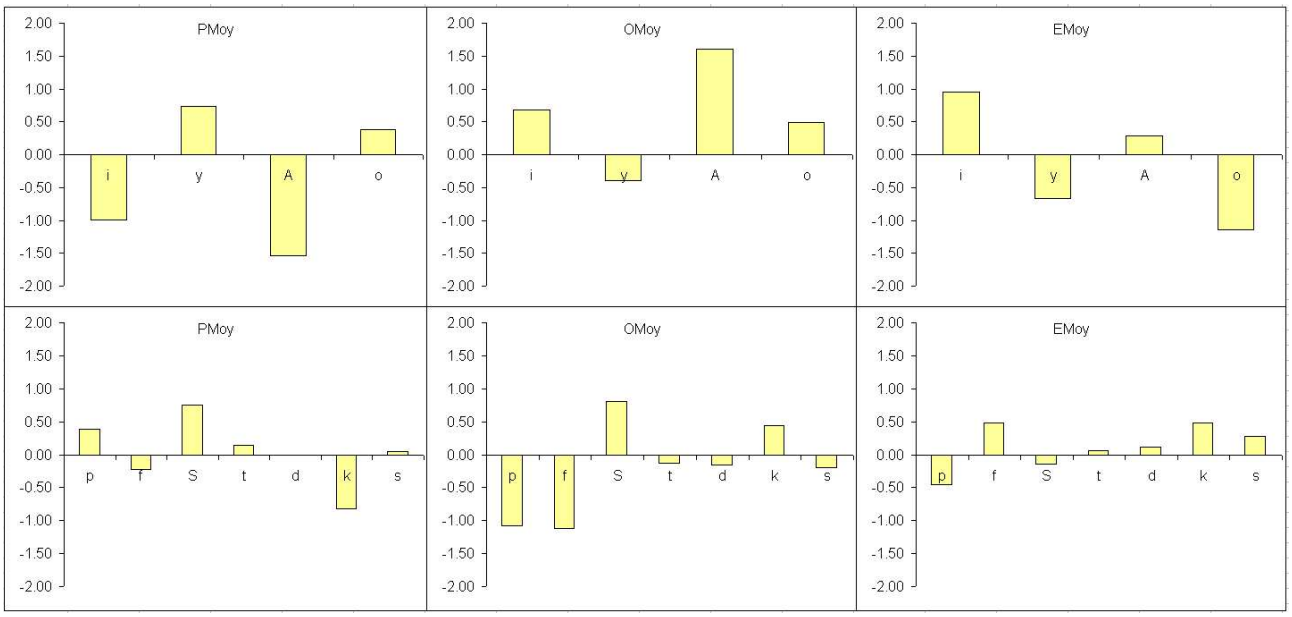

FIG. 4.14 – Moyennes centrées réduites des valeurs d'ouverture, de protrusion et d'étirement d'un mini corpus de 10 locuteurs.

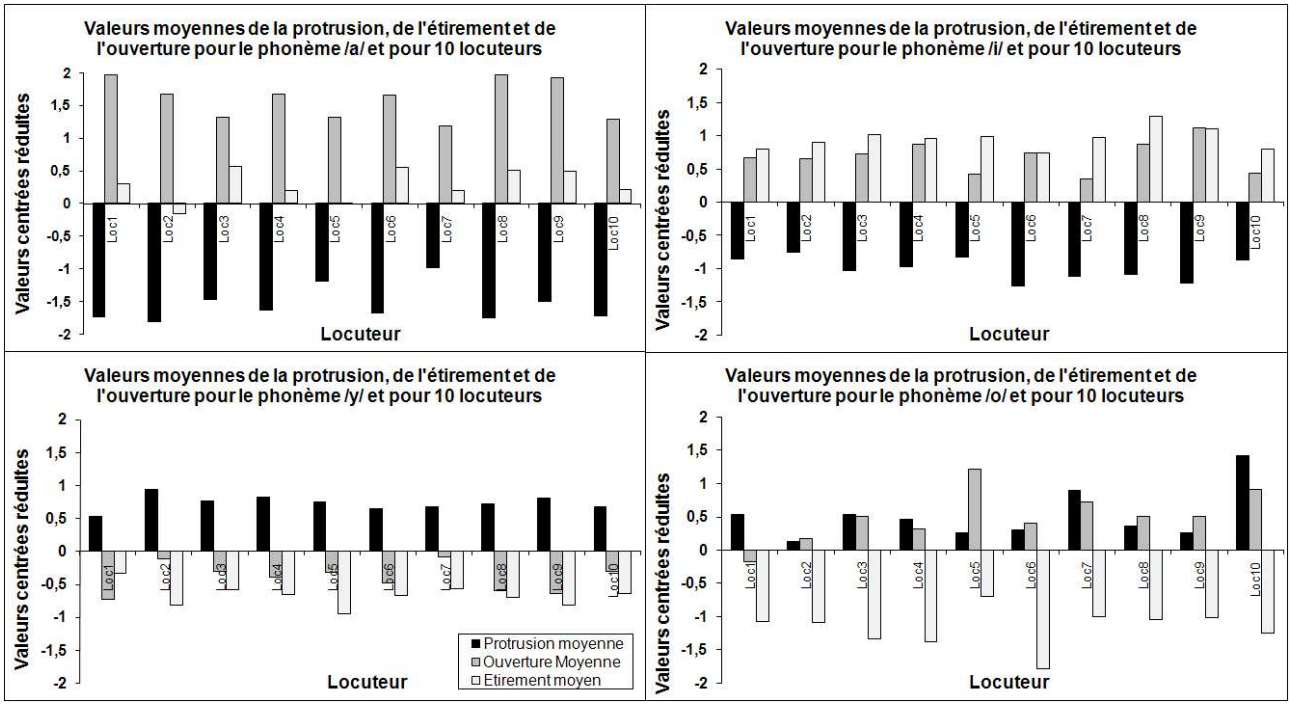

FIG. 4.15 – Moyennes centrées réduites des valeurs d'ouverture, de protrusion et d'étirement des voyelles pour chaque locuteur.

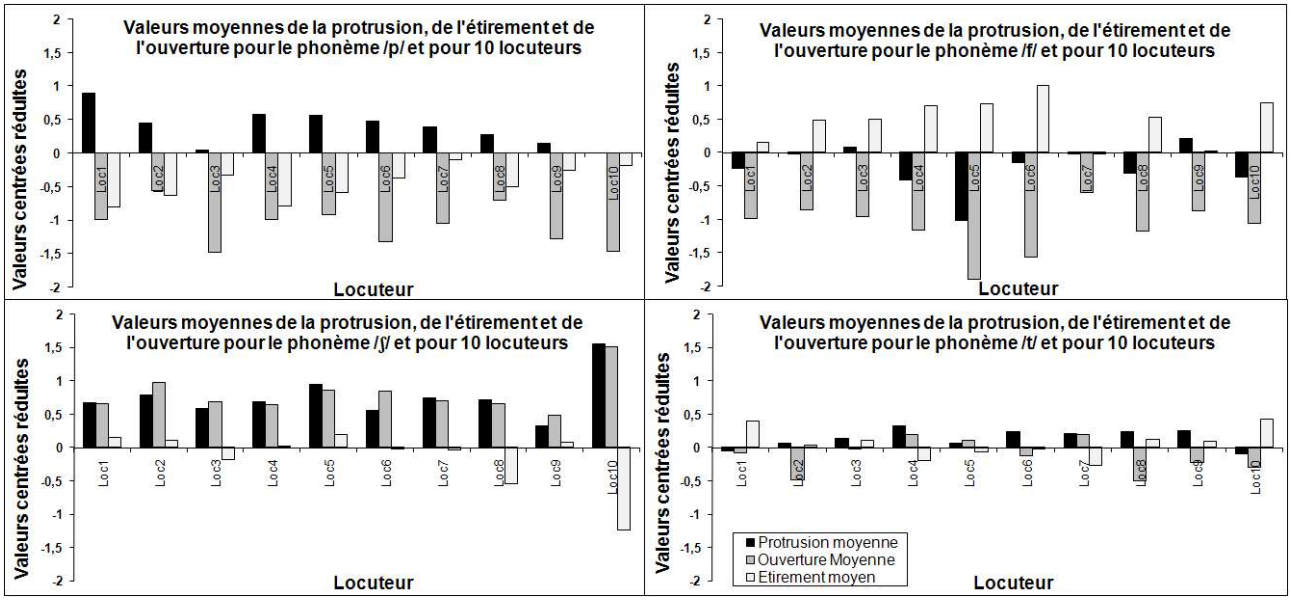

FIG. 4.16 – Moyennes centrées réduites des valeurs d'ouverture, de protrusion et d'étirement des consonnes pour chaque locuteur.

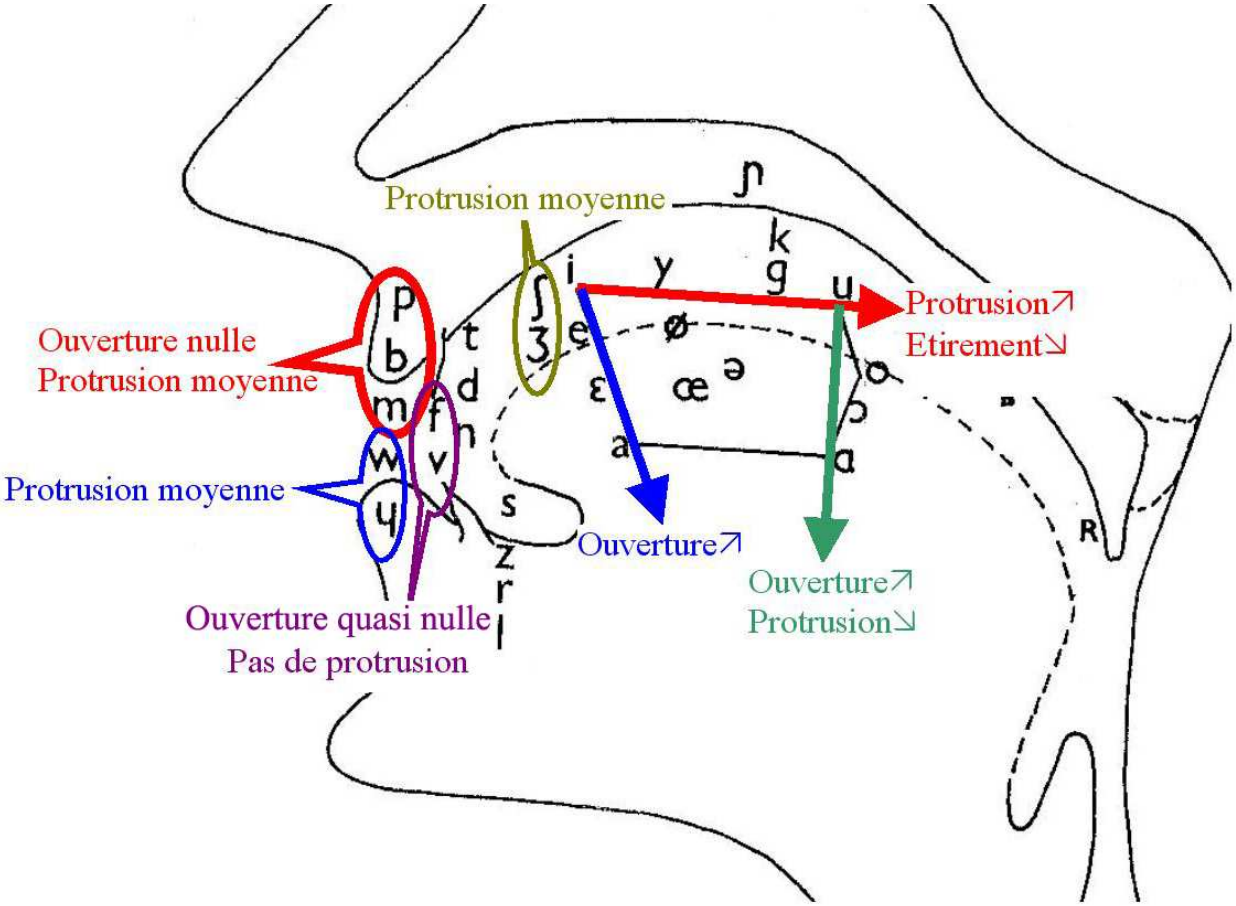

FIG. 4.17 – Quantification de l'ouverture, de la protrusion et de l'étirement en fonction des phonèmes.

Phonème	Protrusion	Ouverture	Etirement
[i]	1	2	4
[e]	1	2	4
[ɛ]	1	3	4
[a]	1	4	3
[y]	3	1	1
[ø]	3	2	1
[œ]	3	3	1
[u]	4	1	1
[o]	4	2	1
[ɔ]	2	3	2
[ɛ̃]	2	3	2
[ã]	2	3	2
[õ]	3	2	1
[w]	3	1	1
[ɥ]	3	1	1
[j]	1	3	3
[p]	2	0	
[b]	2	0	
[m]	2	0	
[t]			
[d]			
[n]			
[k]			
[g]			
[ŋ]			
[f]		0.5	
[v]		0.5	
[s]			
[z]			
[l]			
[ʃ]	3		
[ʒ]	3		
[r]			
[R]			
[ɹ]			
[ʁ]			

FIG. 4.18 – Tableau de classification des paramètres de protrusion, d'ouverture et d'étirement des phonèmes du français.

Chapitre 5

Etude d'une méthode de prédiction de la coarticulation

Le chapitre précédent nous a permis d'effectuer une classification des paramètres d'ouverture, d'étirement et de protrusion pour les différents phonèmes du français. Certains phonèmes qui n'ont qu'une faible influence labiale, comme par exemple [t] n'ont donc pas été quantifiés. Nous avons également extrait les règles d'interaction suivantes entre les paramètres :

- La protrusion et l'étirement varient en sens opposé (vrai pour tous les phonèmes).
- Dans le cas des voyelles postérieures et les bilabiales, la protrusion varie en sens opposé de l'ouverture.
- L'étirement et l'ouverture varient en sens opposé pour les voyelles non arrondies.

Dans ce chapitre, nous allons présenter une méthode de prédiction de la coarticulation basée sur les règles phonétiques énoncées précédemment.

5.1 Algorithme de prédiction de la coarticulation

A partir de notre classification phonétique et en tenant compte de l'influence mutuelle des paramètres labiaux, nous avons défini un algorithme de prédiction de la coarticulation d'une séquence quelconque. Cet algorithme prend en compte le fait que la coarticulation anticipatrice est prédominante en français. Il s'agit d'un algorithme formel permettant de connaître la nature de l'évolution des paramètres labiaux (augmentation, stagnation ou diminution). Un corpus d'apprentissage sera nécessaire par la suite pour quantifier précisément les variations des différentes grandeurs.

L'idée de base de cet algorithme consiste à vérifier la concordance des mouvements de protrusion, d'étirement et d'ouverture en empêchant la génération de mouvements incohérents : par exemple, l'augmentation de la protrusion ne peut pas aller de pair avec l'augmentation de l'étirement. Pour privilégier la coarticulation anticipatrice, l'étude de la séquence à modéliser se fait en commençant par la fin de cette dernière. Avant de détailler précisément l'algorithme, nous allons étudier un exemple :

5.1.1 Exemple d'application de l'algorithme

Prenons l'exemple de la séquence /apRi/. La phase d'initialisation consiste tout d'abord à reporter sur la séquence considérée la classification phonétique du tableau 4.18. Ensuite, une

détermination du sens de variation des paramètres labiaux est réalisée.

Phase d'initialisation

Phonème	/a/	/p/	/ʀ/	/i/
P	1	2		1
O	4	0		2
E	3			4

⇒

Phonème	/a/	/p/	/ʀ/	/i/
P	1	2	↘	1
O	4	0	↗	2
E	3	↗	↗	4

Corps de l'algorithme. L'algorithme se déroule ensuite de la façon suivante : en partant de la fin de la séquence, une vérification des règles de dépendance entre étirement, protrusion et ouverture est effectuée. Dans le cas où une incohérence est détectée, un blocage de l'évolution du paramètre concerné a lieu. Compte tenu que le mouvement de la mâchoire est fortement lié à celui de l'ouverture, sa variation sera calquée sur ce dernier.

- De /R/ à /i/ : pas d'incohérence

Phonème	/a/	/p/	/ʀ/	/i/
P	1	2	↘	1
O	4	0	↗	2
E	3	↗	↗	4

⇒

Phonème	/a/	/p/	/ʀ/	/i/
P	1	2	↘	1
O	4	0	↗	2
E	3	↗	↗	4

- De /p/ à /R/ : pas d'incohérence

Phonème	/a/	/p/	/ʀ/	/i/
P	1	2	↘	1
O	4	0	↗	2
E	3	↗	↗	4

⇒

Phonème	/a/	/p/	/ʀ/	/i/
P	1	2	↘	1
O	4	0	↗	2
E	3	↗	↗	4

- De /a/ à /p/ : la protrusion augmente, donc l'étirement ne doit pas augmenter. On force donc une stabilisation de l'étirement au niveau du phonème /p/.

Phonème	/a/	/p/	/ʁ/	/i/
P	1	2	↘	1
O	4	0	↗	2
E	3	↗	↗	4

⇒

Phonème	/a/	/p/	/ʁ/	/i/
P	1	2	↘	1
O	4	0	↗	2
E	3	3	↗	4

L'augmentation de la protrusion de /a/ à /p/ est incompatible avec celle de l'étirement
⇒ on bloque la valeur de l'étirement de /a/ à /p/

5.1.2 Algorithme

L'exemple précédent peut être généralisé par l'algorithme suivant.

5.1.2.1 Notation

- Ph : Phonème
- P(Ph) : Protrusion du phonème Ph
- O(Ph) : Ouverture du phonème Ph
- E(Ph) : Etirement du phonème Ph

5.1.2.2 Initialisation

pour $Ph_i = Ph_1$ à Ph_n **faire**
  Recopier les paramètres du tableau 4.18 pour le phonème considéré
  si un des paramètres de Ph_i est non spécifié **alors**
    Déterminer l'évolution de ce paramètre (croissance, décroissance ou stagnation)
  finsi
fin pour

5.1.2.3 Corps de l'algorithme

pour $Ph_i = Ph_{n-1}$ à Ph_1 **faire**
  si P(Ph_i) quantifié **alors**
    si l'étirement varie dans le même sens que la protrusion **alors**
      E(Ph_i)=E(Ph_{i+1}) {On force la stagnation de l'étirement}
    finsi
    si Ph_i est une voyelle postérieure ou une bilabiale et O(Ph_i) varie dans le même sens que la protrusion **alors**
      O(Pho_i) = O(Pho_{i+1}) {On force la stagnation de l'ouverture}
    finsi
  sinon si E(Ph_i) quantifié **alors**
    si la protrusion varie dans le même sens que l'étirement **alors**
      P(Ph_i)=P(Ph_{i+1}) {On force la stagnation de la protrusion}
    finsi

si Ph_i est une voyelle non arrondie et O(Ph_i) varie dans le même sens que l'étirement **alors**
O(Pho_i) = O(Pho_{i+1}) {On force la stagnation de l'ouverture}
finsi
sinon si O(Ph_i) quantifié **alors**
si Ph_i est une voyelle non arrondie et E(Ph_i) varie dans le même sens que l'ouverture **alors**
E(Pho_i) = E(Pho_{i+1}) {On force la stagnation de l'étirement}
finsi
si Ph_i est une voyelle postérieure ou une bilabiale et P(Ph_i) varie dans le même sens que l'ouverture **alors**
P(Pho_i) = P(Pho_{i+1}) {On force la stagnation de la protrusion}
finsi
finsi
fin pour

Cet algorithme bloque l'anticipation s'il existe des antagonismes au niveau des mouvements articulatoires. Dans des séquences VC...CV où les consonnes n'ont pas d'influence labiale, ce modèle est de de type Look-Ahead et compatible avec le modèle proposé par Henke [40]. En revanche la présence de bilabiales ou labio-dentales par exemple va "bloquer" le phénomène d'anticipation de l'ouverture des lèvres. L'originalité de notre démarche est la prise en compte de l'interdépendance des articulateurs.

5.2 Stratégie de prédiction de la coarticulation labiale

L'algorithme présenté au paragraphe précédent donne une indication sur la variation des paramètres articulatoires sans en préciser les grandeurs et variations possibles. D'un point de vue phonologique, les variations potentielles d'un paramètre s'étendent tant qu'aucun son contradictoire n'est produit. La quantification des grandeurs quant à elle est fortement dépendante du locuteur. La distinction entre le caractère phonétique et phonologique est mis en évidence par Keating [44]. Comme nous l'avons évoqué au paragraphe 2.5.2, P. Keating établit une fenêtre de variation possible pour les phonèmes au niveau d'un articulateur donné. Ainsi, une séquence donnée est associée à une suite de fenêtres et le mouvement de l'articulateur est défini par un "chemin" entre fenêtres qui en minimise l'effort articulatoire.

Le modèle de Keating nous semble intéressant mais insuffisamment précis pour quantifier la dynamique du mouvement. De plus, les différents modèles théoriques et expérimentaux existants montrent que l'évolution d'un paramètre donné a souvent lieu par palier. Une forme de sigmoïde (Equation 5.1) semble bien adaptée pour représenter cette évolution et pour pouvoir contrôler parfaitement la dynamique. C Pelachaud et al [70] qui ont construit un modèle facial à partir d'une approche pseudo-musculaire ont d'ailleurs imaginé une fonction de déformation (fonction à base radiale) ayant une forme voisine d'une sigmoïde.

V_i et V_f représentent les valeurs initiale et finale de la sigmoïde. c pilote la vitesse de transition et t_0 correspond à la position du centre de la sigmoïde. Ainsi, comme le montre la figure 5.1 les courbes obtenues peuvent facilement être adaptées à différents profils d'évolutions dynamiques. Dans cet exemple, une valeur de 0.05 pour c correspond au modèle "Look-Ahead" alors qu'une valeur de c plus grande que 0.10 avec un choix approprié de t_0 ressemble davantage au modèle

"Time-Locked". Comme on le devine, les sigmoïdes peuvent aussi permettre d'approximer le modèle expansionniste. De plus, ce type de courbe permet facilement de s'adapter à des vitesses d'élocution différentes et l'hyper-articulation peut-être simulée en modifiant V_i et V_f. Le choix des sigmoïdes va donc permettre de s'adapter le mieux possible au locuteur tout en respectant les contraintes phonologiques et phonétiques.

$$Sig(t) = V_i + \frac{(V_f - V_i)}{1 + e^{-c.(t-t_0)}} \tag{5.1}$$

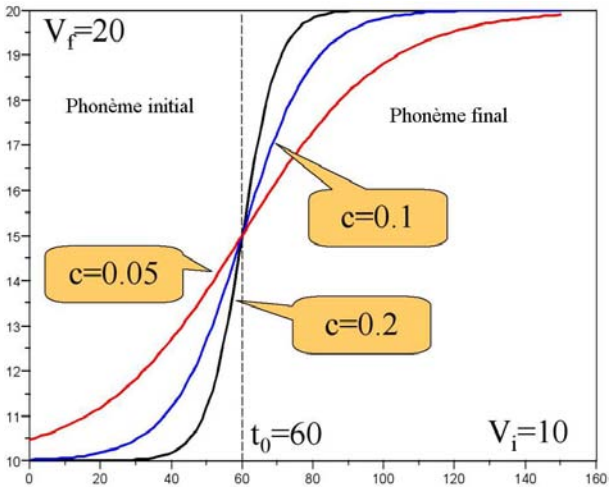

FIG. 5.1 – Influence du paramètre c sur la forme des sigmoïdes.

La figure 5.2 représente l'allure des mouvements de protrusion, d'ouverture et d'étirement de l'exemple du paragraphe précédent (/apRi/). La courbe du bas représente les mouvements réels de ces articulateurs (extraits du grand corpus). On constate que les mouvements réels correspondent aux estimations sauf pour le phonème /R/ qui semble influencer plus largement l'ouverture que prévu.

Pour une séquence donnée, l'algorithme de prédiction de la coarticulation développé au paragraphe précédent a permis de déterminer l'ordre de grandeur ou l'évolution des paramètres articulatoires. Pour chaque paramètre, entre deux phonèmes quantifiés (pour lesquels un numéro entre 0 et 4 a été attribué), nous allons rechercher dans la phase d'apprentissage préliminaire à la synthèse la sigmoïde "collant" le mieux au mouvement réel. Une méthode de minimisation sera appliquée à cette fin. Compte-tenu des hypothèses simplificatrices que nous avons faites, il est possible que la recherche échoue : par exemple, en ce qui concerne la séquence /apRi/ que nous avons étudiée précédemment, l'évolution réelle des paramètres 5.2 montre que l'évolution du mouvement d'ouverture entre /p/ et /i/ peut difficilement être approximé par une seule sigmoïde car le phonème /R/ a une influence non négligeable qui n'était pas prévue. Si les seuils que nous allons appliquer à la minimisation sont dépassés, chaque phonème sera alors quantifié par sa valeur moyenne (moyenne de l'amplitude pour l'articulateur concerné sur l'ensemble des intervalles de temps associés à ce phonème). Bien sûr, dans ce cas, la modélisation sera donc très grossière, mais néanmoins, si le nombre de phonèmes quantifiés isolément est faible, nous verrons dans le chapitre suivant que la phase finale de lissage chargée d'approximer l'ensemble du mouvement limite les erreurs.

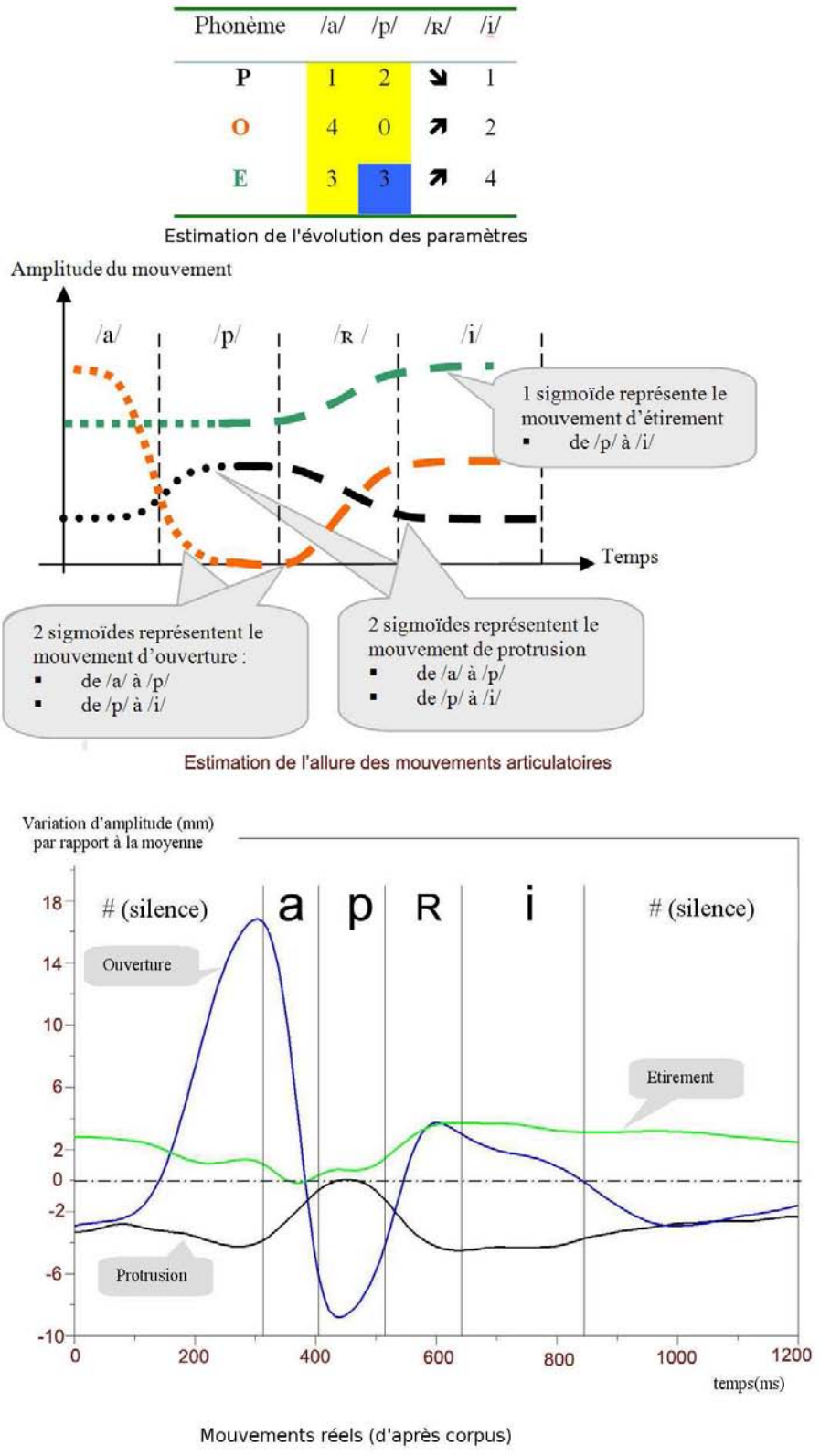

FIG. 5.2 – Identification des sigmoïdes de la séquence apRi.

Chapitre 6

Mise en œuvre de la synthèse de la coarticulation labiale

6.1 Introduction

Le but essentiel de ces travaux de recherche est le développement d'une tête parlante pouvant par exemple améliorer la compréhension des personnes sourdes ou malentendantes. En partant d'un signal audio, notre système doit être capable de générer un visage parlant dont les mouvements des lèvres et de la mâchoire augmentent l'intelligibilité du signal d'origine. Il est bien sûr indispensable que le signal sonore et le signal vidéo soient parfaitement synchronisés ; le pire serait que la vidéo dégrade l'intelligibilité du signal sonore seul. De plus, dans l'optique future d'une application temps réel, il est nécessaire que notre système de synthèse dispose d'une infrastructure légère et rapide.

La figure 6.1 présente l'ensemble de la chaîne de production, depuis le signal audio d'origine jusqu'à la tête parlante. Les principales étapes qui seront détaillées dans la suite sont les suivantes :

- Le décodage acoustico-phonétique qui permet d'obtenir une ensemble de phonèmes alignés sur le signal acoustique.
- La décomposition des séquences obtenues en VC...CV.
- L'extraction de la base de données d'apprentissage des VC...CV cherchées si elles s'y trouvent.
- La complétion dans le cas où les VC...CV cherchées ne font pas partie de la base d'apprentissage.
- L'adaptation temporelle des suites à concaténer afin de les synchroniser avec le signal acoustique.
- L'adaptation d'amplitude basée sur des contraintes phonétiques et phonologiques.
- La concaténation des différents éléments obtenus qui permet après lissage d'obtenir la synthèse des mouvements des articulateurs retenus.
- La génération d'une tête parlante pilotée par les paramètres articulatoires synthétisés et synchronisée avec le message audio d'origine.

Pour réaliser ces travaux, notre corpus monolocuteur nous a servi de base d'apprentissage et de test pour mettre en œuvre la synthèse de la coarticulation basée sur l'algorithme présenté au chapitre précédent. Parmi les 69 phrases du corpus, chaque phrase a été extraite successivement du corpus d'apprentissage pour servir de séquence de test, le reste des phrases servant à

apprendre les paramètres nécessaires. Même si la taille de la base de données d'acquisition est très importante (plusieurs dizaines de Giga octets sont nécessaires pour stocker l'ensemble des informations visuelles et sonores), le nombre total de séquences et de phrases reste faible par rapport à toute l'étendue des combinaisons qu'offre la langue française. Il est donc évident que notre synthèse ne pourra pas prendre en compte toutes les subtilités articulatoires. La limitation de l'espace de stockage et du temps d'enregistrement ne nous ont pas non plus permis d'enregistrer les séquences avec plusieurs contextes prosodiques. Même s'il est évident que la prosodie influence fortement les mouvements des articulateurs, nous avons délibérément choisi de découpler notre travail. Cette thèse présente la première phase de la synthèse par concaténation et des travaux futurs devraient nous permettre de l'affiner en prenant notamment en compte les effets prosodiques.

Malgré toutes ces limitations, nous allons montrer que notre technique de synthèse par concaténation permet de prédire avec une assez bonne précision les mouvements des articulateurs choisis.

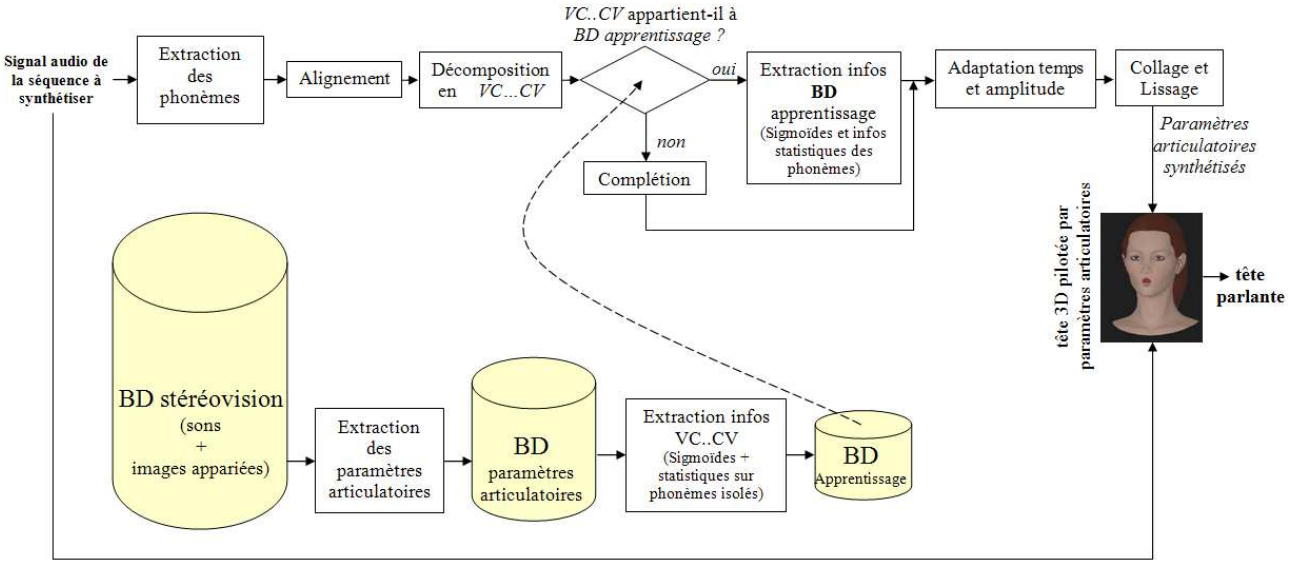

FIG. 6.1 – Chaîne de production de la vidéo de synthèse.

6.2 Phase d'apprentissage

Cette phase a pour but d'extraire le maximum d'informations pertinentes sur les mouvements des articulateurs en accord avec la méthode de prédiction de la coarticulation du chapitre précédent. Les différentes étapes sont les suivantes :

- L'acquisition en stéréovision permet d'obtenir les séquences sonores et les déformations des points du maillage constituant le visage. Ensuite, nous effectuons une mesure des paramètres articulatoires choisis (Protrusion, Ouverture des lèvres, Etirement des lèvres et mouvement de la mâchoire comme indiqué au paragraphe 4.2). Un alignement phonétique semi-automatique nous permet de mettre en correspondance les phonèmes avec les données 3D.

- Nous effectuons ensuite une extraction des données statistiques du corpus : minimum, moyenne et maximum des paramètres de protrusion, d'ouverture, d'étirement et du mouvement de la mâchoire à la fois pour l'ensemble du corpus et pour chaque phonème (sur toute l'étendue de ce dernier). Comme nous le verrons ultérieurement, ces informations vont être utiles pour déterminer les bornes des sigmoïdes, pour caractériser les phonèmes dans le cas où la recherche de sigmoïde va échouer et pour la phase de recalage qui sera effectuée lors de la synthèse.

- Pour chaque séquence du corpus d'apprentissage (suite de phonèmes sans silence), nous appliquons l'algorithme de prédiction de la coarticulation présenté au chapitre précédent. Ceci nous permet d'obtenir pour chaque phonème et chaque paramètre articulatoire soit une quantification sur l'échelle de 0 à 4 de l'ampleur du mouvement, soit une indication sur l'évolution du mouvement (croissance, décroissance ou stagnation).

- Pour chaque paramètre et chaque suite de phonèmes $Pho_1...Pho_2$ où Pho_1 et Pho_2 sont quantifiés (une valeur de 0 à 4 leur a été attribuée par l'algorithme de prédiction), une sigmoïde est recherchée. Afin que les coefficients des sigmoïdes "collent" le plus au corpus d'apprentissage, nous appliquons une minimisation de type Powell. Le but est d'estimer les valeurs des 4 paramètres caractéristiques des sigmoïdes V_i, V_f, c et t_0 (Eq. 5.1) qui minimisent l'erreur quadratique avec les données réelles sur l'ensemble des frames comprises entre Pho_1 et Pho_2. Afin d'éviter de trouver des minima locaux, une bonne estimation des valeurs initiales de ces paramètres est nécessaire. L'algorithme de prédiction nous permet de savoir si nous cherchons une sigmoïde croissante ou décroissante. Si celle-ci est croissante, V_i est initialisée avec l'amplitude minimale au niveau de Pho_1 et Vi_f est initialisée avec l'amplitude maximale au niveau de Pho_2. Si elle est décroissante, c'est l'inverse (Fig. 6.2). La moyenne des instants associés à V_i et V_f nous sert d'estimation pour t_0. Après avoir vérifié que les valeurs trouvées pour V_i et V_f sont cohérentes avec le sens supposé du mouvement, la minimisation est lancée. La figure 6.3 montre un exemple des sigmoides obtenues pour le paramètre ouverture des lèvres d'une phrase du corpus d'apprentissage. Si la recherche échoue, c'est-à-dire si l'erreur quadratique de la sigmoïde la mieux approximée dépasse un seuil fixé ou si le sens supposé n'est pas correct, nous mémorisons les informations statistiques élémentaires des phonèmes concernés (minimum, moyenne, maximum) qui serviront pour le recalage d'amplitude et le lissage lors de la synthèse.

- Enfin, nous procédons à un archivage des coefficients des sigmoïdes. Dans le cas, où aucune sigmoïde n'approchant convenablement les données réelles n'a été trouvée, nous mémorisons les informations liées aux phonèmes isolés.

6.3 La synthèse

Le but général de notre étude étant la construction d'une tête parlante, notre travail doit permettre une bonne animation des mouvements faciaux de façon à augmenter l'intelligibilité de la séquence prononcée par un locuteur. Le travail d'apprentissage réalisé précédemment a permis d'extraire des informations pertinentes sur les mouvements des lèvres et de la mâchoire en fonction d'un ensemble de séquences prononcées. Nous devons maintenant utiliser ces résultats pour modéliser les mouvements faciaux d'une séquence quelconque. Nous avons choisi une approche par concaténation, adaptable à la stratégie du locuteur. Ce type d'approche a déjà largement été utilisé dans le domaine de la parole. En ce qui concerne la génération de têtes

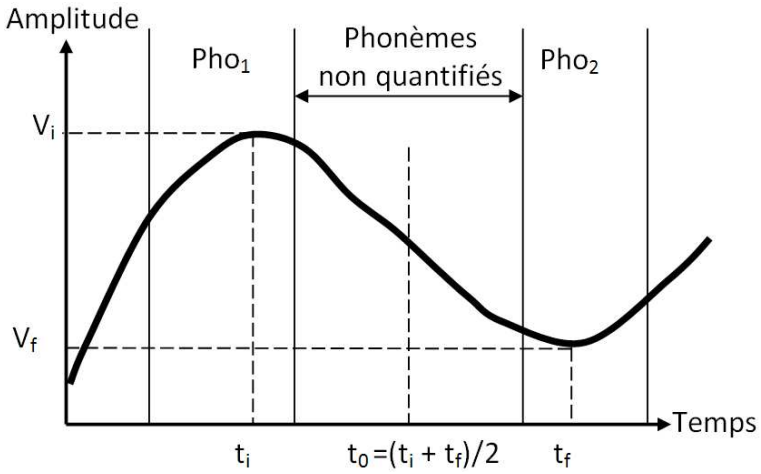

FIG. 6.2 – Estimation avant minimisation des paramètres des sigmoïdes.

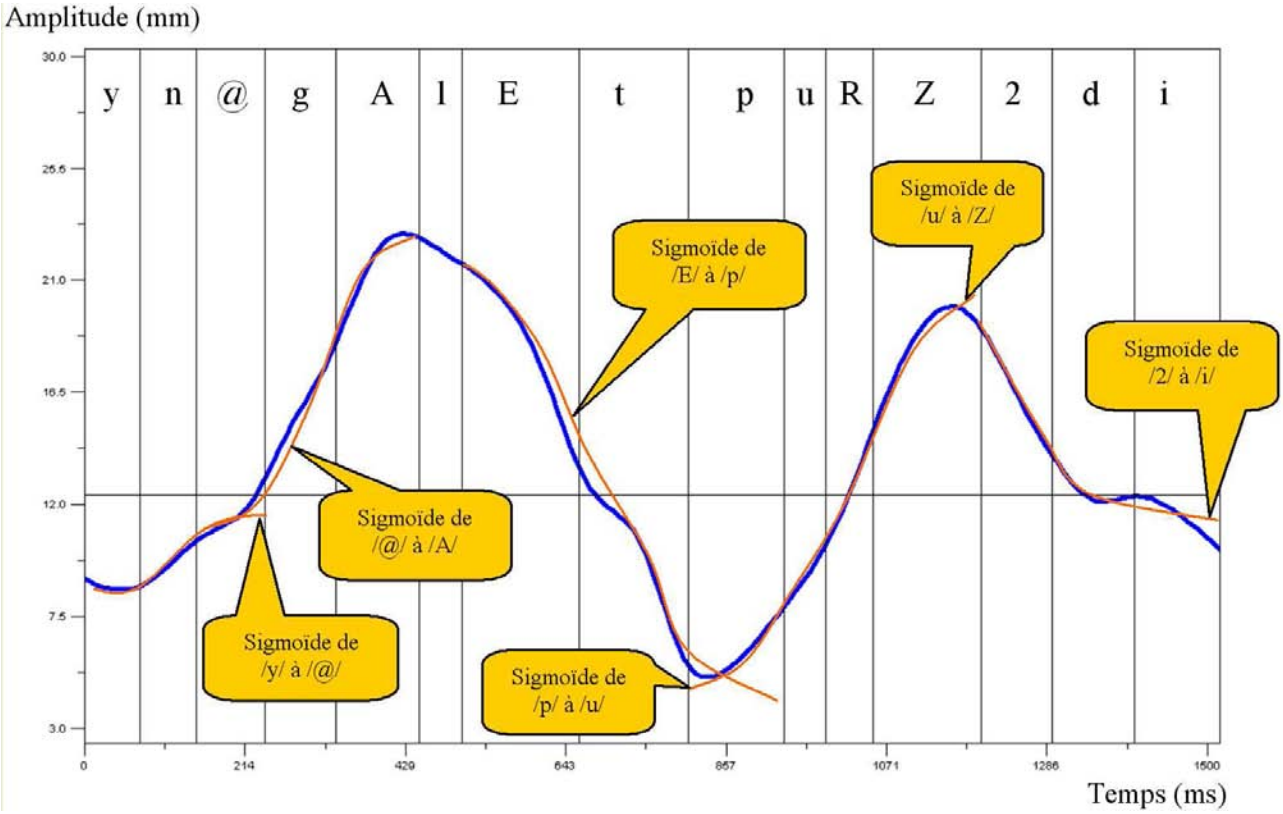

FIG. 6.3 – Extraction des sigmoïdes de la phrase "Une galette pour jeudi" pour l'ouverture labiale. Le mouvement réel est en trait gras et les sigmoïdes en traits fins. L'alphabet SAMPA est utilisé - cf. Annexe A.

parlantes, nous avons déjà parlé du système Video Rewrite de Bregler et al [16] qui concatène une suite de formes de bouches puis applique une superposition sur l'image de fond (le reste du visage) afin de créer une vidéo réaliste. En ce qui concerne les modèles en 3D, Breen at al [15] ont mis au point une tête parlante construite à partir d'une base de di-Visèmes (suite de deux

visèmes). Le process de génération est basé sur l'identification des voyelles de la séquence qui sont considérées comme points d'articulation cibles. Une fois les voyelles trouvées, le di-visème correspondant est recherché et une concaténation des visèmes est générée. Si l'utilisation des di-visèmes réduit énormément la base de données à utiliser, ce choix ne peut pas permettre de prendre en compte toute l'étendue de l'anticipation qui peut s'étendre jusqu'à 6 visèmes en deçà du phonème courant comme l'affirment Benguerel et Cowan [7]. Notre choix a été de conserver un corpus raisonnable tout en prenant en compte toute l'étendue potentielle de la coarticulation. Le modèle proposé par S. Minnis et A. Breen [62] propose une synthèse par concaténation basée sur des N-Phones et tenant compte de critères linguistiques : par exemple, si la fréquence fondamentale d'un phonème ou sa durée change, Minnis et al proposent que les mouvements labiaux en tiennent compte. Pour cela, ils enregistrent les mouvements d'un ensemble de points caractéristiques du visage. Au niveau acoustique, les consonnes sont classés en trois catégories : les phonèmes invisibles /t/, /d/, /g/, /k/ qui ne sont pas associés à une forme précise de bouche, les phonèmes protégés /f/, /v/, /m/, /b/, /p/ qui sont associés à une forme de bouche très précise et les phonèmes normaux qui sont associés à une forme de bouche relativement flexible. Le principe de cette décomposition, cohérent avec notre tableau de classification des phonèmes (Fig. 4.18) fait référence au concept de résistance à la coarticulation [12]. La concaténation proposée par Minnis et Breen [62] tient compte de la classe des phonèmes lors de la concaténation des différents segments par attribution d'un poids spécifique et selon le contexte anticipatoire ou rétentif. Néanmoins, l'identification des mouvements d'un ensemble de points caractéristiques du visage occasionne l'enregistrement d'un très grand nombre d'informations et reste associée à un seul locuteur.

La technique de synthèse que nous mettons en œuvre n'a pas le défaut évoqué précédemment. Le fait d'avoir choisi la sigmoïde comme forme élémentaire permet une adaptation temporelle et d'amplitude aisée, donc adaptable à un locuteur précis, tout en restant très légère en terme de quantité d'informations à stocker. Voici les principales phases de cette synthèse.

6.3.1 Découpage de la suite de phonèmes à synthétiser

La suite de phonèmes à synthétiser est tout d'abord découpée en séquences où chacune d'elle représente une suite de phonèmes sans silence.

Notre classification des phonèmes (fig 4.18) montre que les paramètres de protrusion, d'ouverture des lèvres et d'étirement sont définis pour toutes les voyelles alors qu'ils ne le sont pas ou seulement partiellement pour les consonnes. Cette constatation nous a amené à décomposer chaque séquence à synthétiser en VC...CV afin que les phonèmes extrêmes des éléments à concaténer soient le plus stables possible. Dans le cas où la séquence commence ou finit par une consonne, une voyelle construite artificiellement (dont les coefficients des paramètres articulatoires modélisent au mieux une forme neutre de bouche) est ajoutée en début ou en fin de séquence.

Exemple : Soit la phrase à synthétiser : "*La vaisselle propre est mise sur l'évier.*"

La suite de phonèmes associés à cette phrase est : l a v e s ɛ l p R ɔ p R ə # e m i z ə s y R l e v j e"

Le symbole "#" représentant un silence, cette phrase est découpée en deux séquences :
– l a v e s ɛ l p R ɔ p R ə

– e m i z ə s y R l e v j e

Ensuite, chaque séquence est décomposée en VC...CV et un phonème artificiel que nous noterons "ϕ" est ajouté en début ou fin des séquences commençant ou finissant par une consonne. Nous obtenons donc la décomposition suivante :

l a v e s ɛ l p R ɔ p R ə est décomposée sous la forme :
– ϕ l a
– a v e
– e s ɛ
– ɛ l p R ɔ
– ɔ p R ə

e m i z ə s y R l e v j e est décomposée sous la forme :
– e m i
– i z ə
– ə s y
– y R l e
– e v j e

6.3.2 Extraction des éléments trouvés dans le corpus d'apprentissage

Après décomposition, une recherche des VC...CV trouvées est effectuée dans la base d'apprentissage qui contient les caractéristiques des sigmoïdes et les données statistiques des phonèmes. Plusieurs cas peuvent se présenter :
– la suite VC...CV cherchée n'appartient pas au corpus d'apprentissage
– la suite VC...CV cherchée apparait une ou plusieurs fois dans le corpus d'apprentissage.

Dans ce paragraphe, nous allons nous intéresser au cas où la suite VC...CV cherchée se trouve dans le corpus d'apprentissage. Dans l'ensemble des séquences à synthétiser, notons A une séquence donnée et B celle qui lui succède.

Si A se trouve plusieurs fois dans le corpus d'apprentissage, nous allons privilégier par ordre décroissant :
– la VC...CV trouvée à la position i du corpus d'apprentissage dans le cas où B se trouve à la position $i+1$ dans la même phrase du corpus.
– la VC...CV du corpus d'apprentissage dans le cas où B appartient à la même phrase du corpus, mais n'est pas à la position suivante.
– la VC...CV du corpus d'apprentissage dans le cas où B se trouve dans le corpus, mais pas dans la même phrase.

Si plusieurs suites VC...CV ont la même priorité, nous conservons celle dont les durées des phonèmes sont les plus proches de la séquence à synthétiser.

6.3.3 Complétion des éléments manquants

Compte tenu de la grande quantité de stockage et du temps nécessaire pour extraire les informations des enregistrements, même notre grand corpus ne contient pas toutes les séquences VC...CV possibles de la langue française qui comporte plus de 5000 syllabes différentes. Même en limitant à 4 le nombre de voyelles possibles, nous avons recensé dans le grand corpus BREF [49] 291 VCCV différentes commençant et finissant par les phonèmes /a, i, y, u/. Il est donc fort

probable qu'une ou plusieurs suites VC...CV d'une phrase à reconstruire ne puissent pas être trouvées dans le corpus d'apprentissage. Dans ce cas, nous réalisons une complétion.

6.3.3.1 Complétion des voyelles manquantes

Comme il n'est pas possible d'enregistrer toutes les VCV, VCCV, VCCCV du français, nous avons seulement enregistré les voyelles /a, i, u, y/ dans ce type de contexte. Les autres séquences sont ainsi interpolées linéairement à partir des séquences comportant /a, i, u, y/ . Les poids sont les coordonnées barycentriques de la voyelle inconnue par rapport à /a/ , /i/, /y/, /u/. Ceci nécessite de quantifier les différentes voyelles. Nous avons choisi d'utiliser comme mesure l'amplitude moyenne du paramètre considéré sur toute l'étendue de la voyelle enregistrée isolément. Nous avons fait le choix de prendre en compte toute l'étendue de la voyelle et non sa valeur centrale car nous n'avons aucune certitude sur la position réelle de la valeur cible.

Prenons l'exemple de la reconstruction de la séquence $/\varepsilon tu/$. Ceci nécessite le calcul des coefficients barycentriques α_i de $/\varepsilon/$ par rapport à /a, i, u, y/ en utilisant les données obtenues lors de l'enregistrement des voyelles isolées. Ainsi, pour chacun des paramètres de protrusion, d'ouverture des lèvres, d'étirement et d'ouverture de la mâchoire, les sigmoïdes sont estimées par $\alpha_1.f([atu]) + \alpha_2.f([itu]) + \alpha_3.f([utu]) + \alpha_4.f([ytu])$. Les coefficients α_i (i de 1 à 4 correspondant aux phonèmes /a, i, u, y/) doivent satisfaire : $\sum \alpha_i = 1$. Le coeficient α_i est d'autant plus proche de 1 que la "distance" (différence d'amplitude moyenne) du phonème à estimer avec le phonème i est faible. L'équation 6.1 présente le calcul de ce coefficient.

$$\alpha_{i=1}^{4} = \frac{\frac{1}{|param(PhonemeCherche) - param(Phoneme_i)|}}{\sum_{j=1}^{4} \frac{1}{|param(PhonemeCherche) - param(Phoneme_j)|}} \tag{6.1}$$

Dans le cas d'une séquence $V_1C...CV_2$ où aucune des deux voyelles n'appartient à /a, i , u , y/, /aC..CV_2/, $/iC..CV_2/$, $/uC..CV_2/$, $/yC..CV_2/$ sont d'abord reconstruits en utilisant la technique exposée précédemment, puis $V_1C...V_2$ est linéairement interpolé à partir de $/aC..CV_2/$, $/iC..CV_2/$, $/uC..CV_2/$, $/yC..CV_2/$.

6.3.3.2 Complétion des consonnes manquantes

Toutes les combinaisons CV et VCV avec $V \in /a, i, u, y/$ ont été enregistrées, mais seulement 92 $VCCV$ ont été retenues afin de limiter la taille du corpus. Les $VCCV$ choisies ont été sélectionnées à partir d'une analyse statistique sur un grand corpus construit afin de maximiser la couverture phonétique de français [49]. Les VC_1C_2V qui ne sont pas dans le corpus d'apprentissage sont reconstruites par superposition des séquences VC_1V and VC_2V. Chaque séquence reconstruite est ensuite intégrée au corpus d'apprentissage ; ainsi, le corpus s'étoffe au fur et à mesure réduisant le temps nécessaire à la synthèse. Cette technique qui sépare clairement le rôle des voyelles et des consonnes est en accord avec la théorie d'Öhman[42] qui dans son étude de la coarticulation des mouvements de la langue a déduit que les mouvements articulatoires sont pilotés par les voyelles auxquels se superpose le geste consonantique.

La séquence /ikty/, par exemple n'appartient pas au corpus d'apprentissage et elle est reconstruite à partir de /iky/ et /ity/ qui appartiennent au corpus. Même si la protrusion augmente de /i/ à /y/ dans les deux séquences, les mouvements sont différents et notre complétion permet d'obtenir une bonne estimation du mouvement de /ikty/. La figure 6.4 montre ce processus.

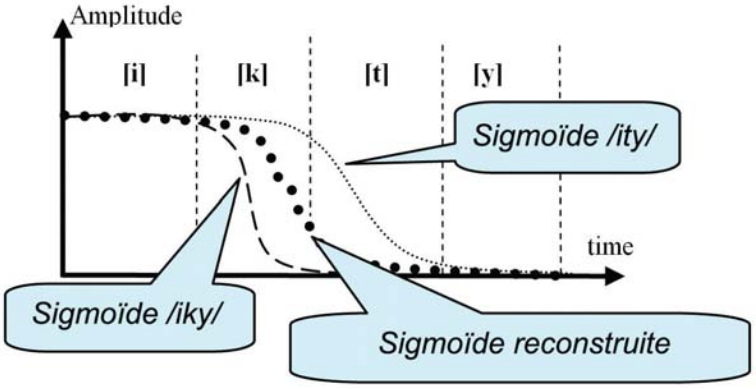

FIG. 6.4 – Reconstruction d'une séquence VCCV à partir de deux séquences VCV.

6.3.4 Adaptation temporelle

Les durées des phonèmes enregistrés dans le corpus n'étant pas nécessairement identiques à celles des séquences à synthétiser, il est nécessaire d'adapter les durées dans la phase de reconstruction. Notre choix des sigmoïdes comme formes élémentaires du mouvement nous offre une grande souplesse au niveau des déformations temporelles. Pour chaque sigmoïde, la position relative du centre de la sigmoïde (t_0) est conservée invariante proportionnellement à la fin du phonème de départ de la sigmoïde et le début du phonème d'arrivée. La figure 6.5 en montre le principe sur une séquence VCV. En ce qui concerne la dynamique de la sigmoïde, contrôlée par le paramètre c de l'équation 5.1, nous avons choisi de conserver celle de l'apprentissage afin de rester le plus cohérent possible avec la stratégie du locuteur dans un contexte prosodique neutre.

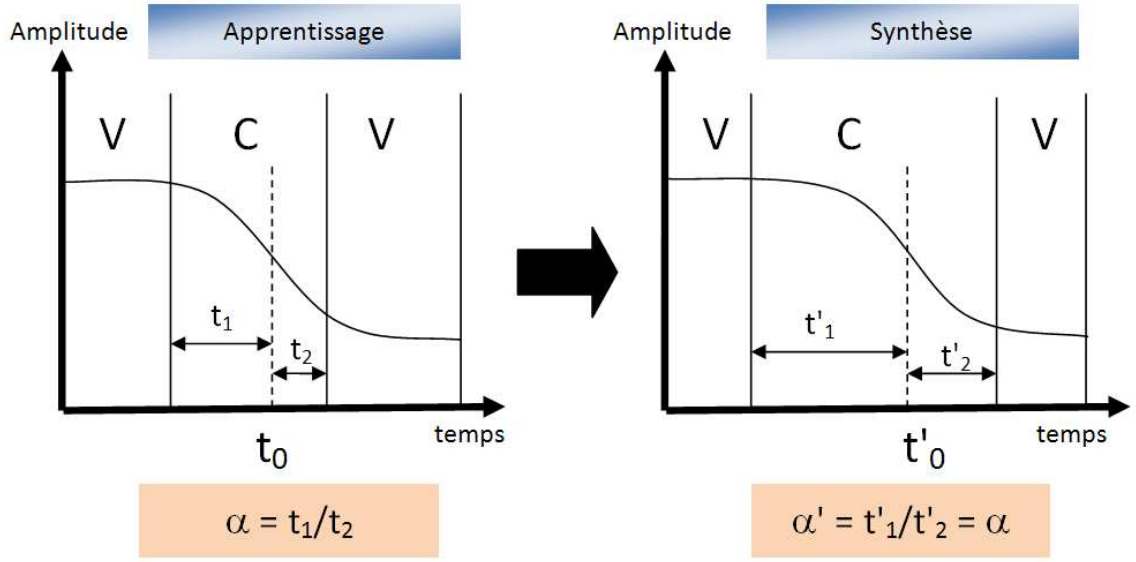

FIG. 6.5 – Détermination du centre de la sigmoïde lors de la synthèse.

6.3.5 Adaptation de l'amplitude

La reconstruction des paramètres labiaux repose sur la concaténation de sigmoïdes. Il faut garantir la cohérence paradigmatique, c'est-à-dire les caractéristiques intrinsèques des sons, mais aussi la cohérence syntagmatique, c'est-à-dire respecter les contrastes distinctifs entre les sons à l'intérieur de la phrase à synthétiser. Les axes syntagmatiques et paradigmatiques s'adaptent en fonction des éléments en présence. Selon la phrase à synthétiser, il est donc nécessaire d'ajuster la position des paramètres articulatoires en tenant compte de ces deux axes. L'approche de Keating [44] repose sur le même principe car l'évolution du paramètre articulatoire consiste à trouver un chemin continu entre un ensemble de fenêtres constituant la séquence. La largeur de la fenêtre peut être associée à l'axe paradigmatique et la juxtaposition des fenêtres à l'axe syntagmatique.

Par conséquent, les données extraites du corpus d'apprentissage (sigmoïdes ou valeurs moyennes de phonèmes) doivent être corrigées en amplitude lors de la phase de synthèse. Pour respecter la cohérence syntagmatique, il faut recaler les sigmoïdes consécutives afin que les valeurs finales des premières correspondent aux valeurs initiales des secondes. Néanmoins, ce recalage doit être modéré afin de conserver les caractéristiques du son (axe paradigmatique). Pour cela, une méthode de minimisation multidimensionnelle de type Powell est appliquée sur l'ensemble des sigmoïdes d'une séquence donnée afin de déterminer le décalage optimal à appliquer sur chaque sigmoïde pour que l'écart avec les sigmoïdes voisines soit minimal (effet syntagmatique). Parallèlement, afin de tenir compte du critère paradigmatique, la minimisation tient compte pour chaque phonème caractéristique (ayant un certain degré de résistance à la coarticulation) de sa variation par rapport à ce même phonème enregistré isolément.

L'équation 6.2 détaille l'expression à minimiser. $Sig_i(t_{min})$ et $Sig_i(t_{max})$ correspond à la valeur initiale et finale de la sigmoïde numéro i. La première partie de l'équation représente la différence entre la valeur finale d'une sigmoïde et la moyenne des points de départ des sigmoïdes suivantes (quand le phonème initial de ces sigmoïdes correspond au phonème final de la sigmoïde courante). Il est important de rappeler que dans le cas des complétions de consonnes manquantes dans une $VCCV$, plusieurs sigmoïdes estiment la séquence manquante, donc il peut y avoir plusieurs sigmoïdes dans les mêmes zones temporelles ce qui explique la sommation $\sum_{j=1/Sig_j \in Next(i)}^{k} Sig_j(t_{min})$ dans l'équation 6.2. Cette première partie de l'expression qui minimise l'écart entre les sigmoïdes voisines correspond à l'axe syntagmatique.

Le second et le troisième terme de l'expression correspondent aux différences entre les extrémités des sigmoïdes (correspondant aux voyelles) et la valeur moyenne des voyelles isolées ; il correspond donc à l'axe paradigmatique. Nous appliquons cette minimisation uniquement aux phonèmes ayant un certain degré de résistance à la coarticulation, c'est-à-dire ceux présentant des valeurs caractéristiques pour les paramètres labiaux (ceux qui sont quantifiés dans le tableau 4.18).

α_1 correspond au poids accordé à l'axe syntagmatique et α_2 au poids de l'axe paradigmatique.

$$\alpha_1 \sum_{i=1}^{n-1} \left| Sig_i(t_{max}) - \frac{\sum_{j=1/Sig_j \in Next(i)}^{k} Sig_j(t_{min})}{k} \right| + \alpha_2 \sum_{i=1}^{n} \left| Sig_i(t_{min}) - \overline{IsolatedVowel(t_{min})} \right| + \alpha_2 \sum_{i=1}^{n} \left| Sig_i(t_{max}) - \overline{IsolatedVowel(t_{max})} \right| \quad (6.2)$$

Les figures 6.6 et 6.7 montrent l'exemple de la synthèse de l'extrait "Au muret de ce pont". Dans la figure 6.7, une correction d'amplitude est effectuée ; on constate alors que l'allure des données estimées s'approche davantage des données réelles. L'apport de la correction est encore plus frappant dans les figures 6.8 et 6.9. Il s'agit de la synthèse du mouvement de protrusion de la fin de la phrase "La voiture s'est arrêtée". En l'absence de recalage, des mouvements de protrusion injustifiés ont lieu dans la partie "est arrêtée". En revanche, le recalage permet de corriger en grande partie ce problème (Fig. 6.9).

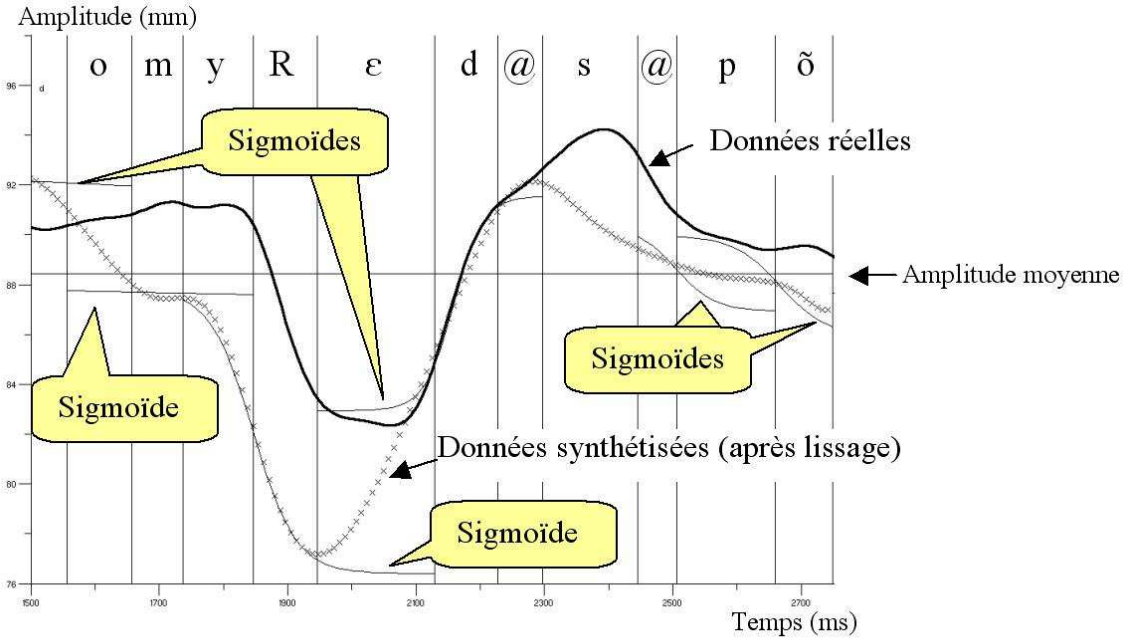

FIG. 6.6 – Synthèse du mouvement de protrusion de la séquence "Au muret de ce pont" sans recalage d'amplitude (Alphabet phonétique SAMPA - cf. Annexe A).

6.3.6 Lissage final

A la fin de cette étape de notre travail, la séquence à synthétiser est modélisée au mieux par une suite de sigmoïdes et d'informations sur des phonèmes isolés quand aucune sigmoïde n'a pu être trouvée. A partir de ces informations, nous construisons une spline cubique d'approximation

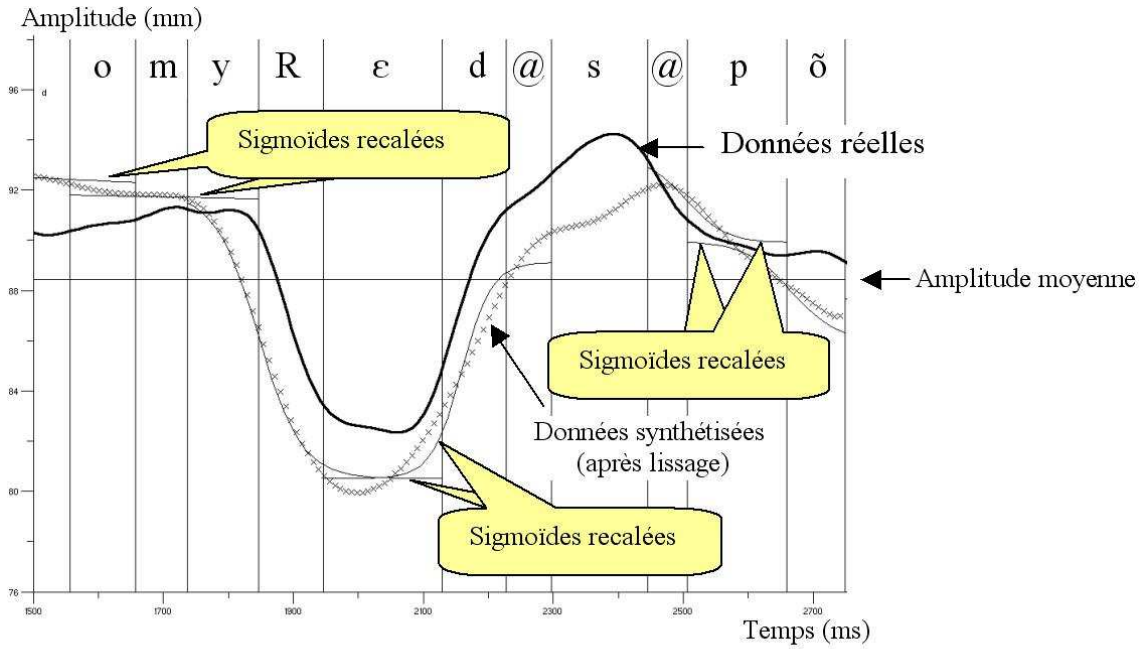

FIG. 6.7 – Synthèse du mouvement de protrusion de la séquence "Au muret de ce pont" avec recalage d'amplitude (Alphabet phonétique SAMPA - cf. Annexe A).

afin de modéliser l'évolution des mouvements articulatoires. Les points de contrôle de cette spline sont les valeurs centrales des phonèmes. Des poids sont attribués pour privilégier l'anticipation et les phonèmes résistants à la coarticulation, notamment les bilabiales en ce qui concerne l'ouverture (par exemple, un poids plus important est attribué pour l'ouverture des lèvres au niveau du phonème /p/ pour forcer la spline à bien s'approcher de la valeur cible). En ce qui concerne les phonèmes qui n'ont pas pu être approchés par une sigmoïde et qui ont été approximés par leur valeur moyenne, nous attribuons un poids plus important au début du phonème pour privilégier le phénomène d'anticipation.

Si S est l'espace des fonctions splines cubiques avec des points de contrôles $x1 < x2 < ... < xn$ alors la spline d'approximation s sur m points d'abscisse xd et d'ordonnée yd vérifie l'équation 6.3 pour toutes les fonctions $f \in S$, ce qui signifie que s minimise la somme des erreurs quadratiques sur toutes les fonctions de S. Dans cette équation, $wd(k)$ représente le poids accordé au point $xd(k), yd(k)$.

$$\sum_{k=1}^{m} wd(k)(s(xd(k)) - yd(k))^2 \leq \sum_{k=1}^{m} wd(k)(f(xd(k)) - yd(k))^2 \tag{6.3}$$

6.4 Conclusion

La synthèse que nous avons développée repose sur le paradigme de concaténation de suites $VC...VC$ stockées sous formes de sigmoïdes. Les données d'apprentissage stockées sont ainsi très

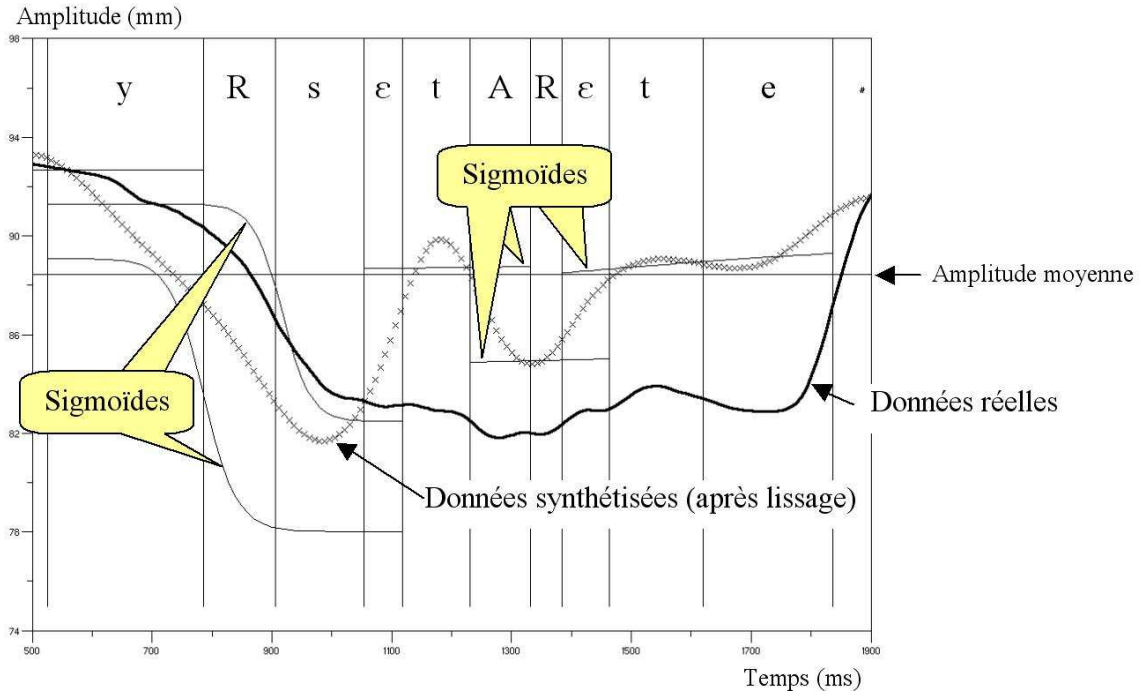

FIG. 6.8 – Synthèse du mouvement de protrusion de la fin de la phrase "La voiture s'est arrêtée" sans recalage d'amplitude (Alphabet phonétique SAMPA - cf. Annexe A).

"lègères" tout en conservant les informations utiles permettant de s'adapter à une vitesse d'élocution particulière. De même, un réglage de l'amplitude peut permettre de modéliser une hyper ou une hypo-articulation [54] tout en conservant une cohérence syntagmatique de l'ensemble. Enfin, le lissage final permet d'affiner encore les résultats obtenus par l'attribution de poids aux points à prendre en compte. La figure 6.10 montre la synthèse d'une phrase complète. On constate que l'allure globale des mouvements est respectée ; nous allons maintenant analyser plus précisément l'ensemble des résultats pour évaluer au mieux notre méthode.

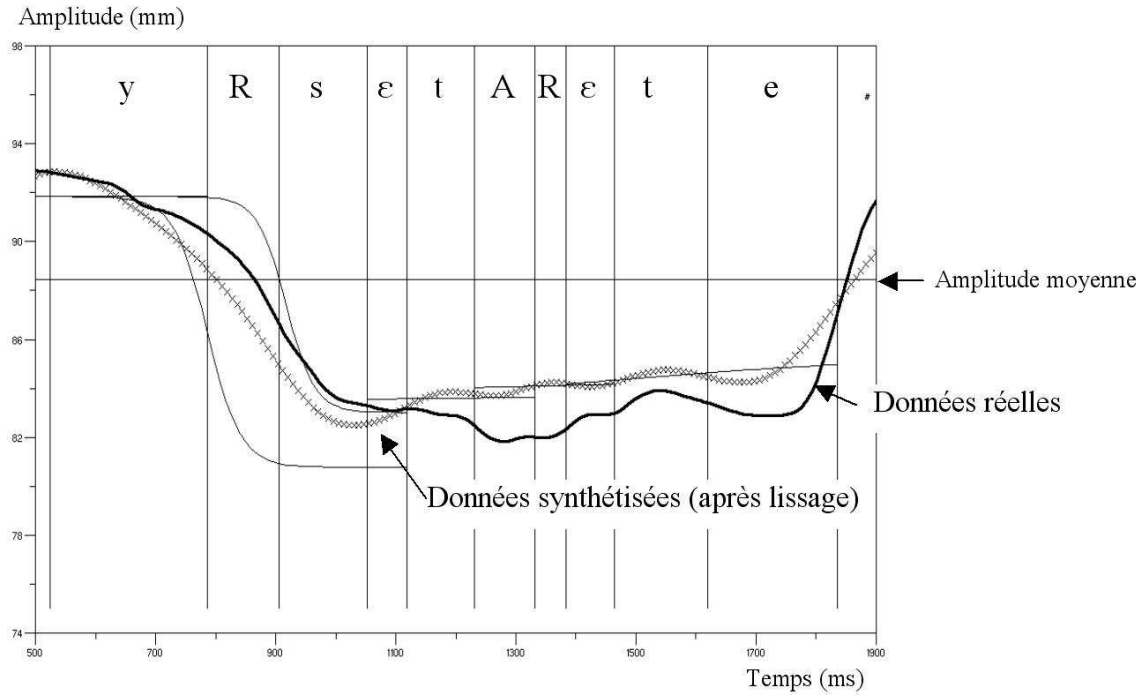

FIG. 6.9 – Synthèse du mouvement de protrusion de la fin de la phrase "La voiture s'est arrêtée" avec recalage d'amplitude (Alphabet phonétique SAMPA - cf. Annexe A).

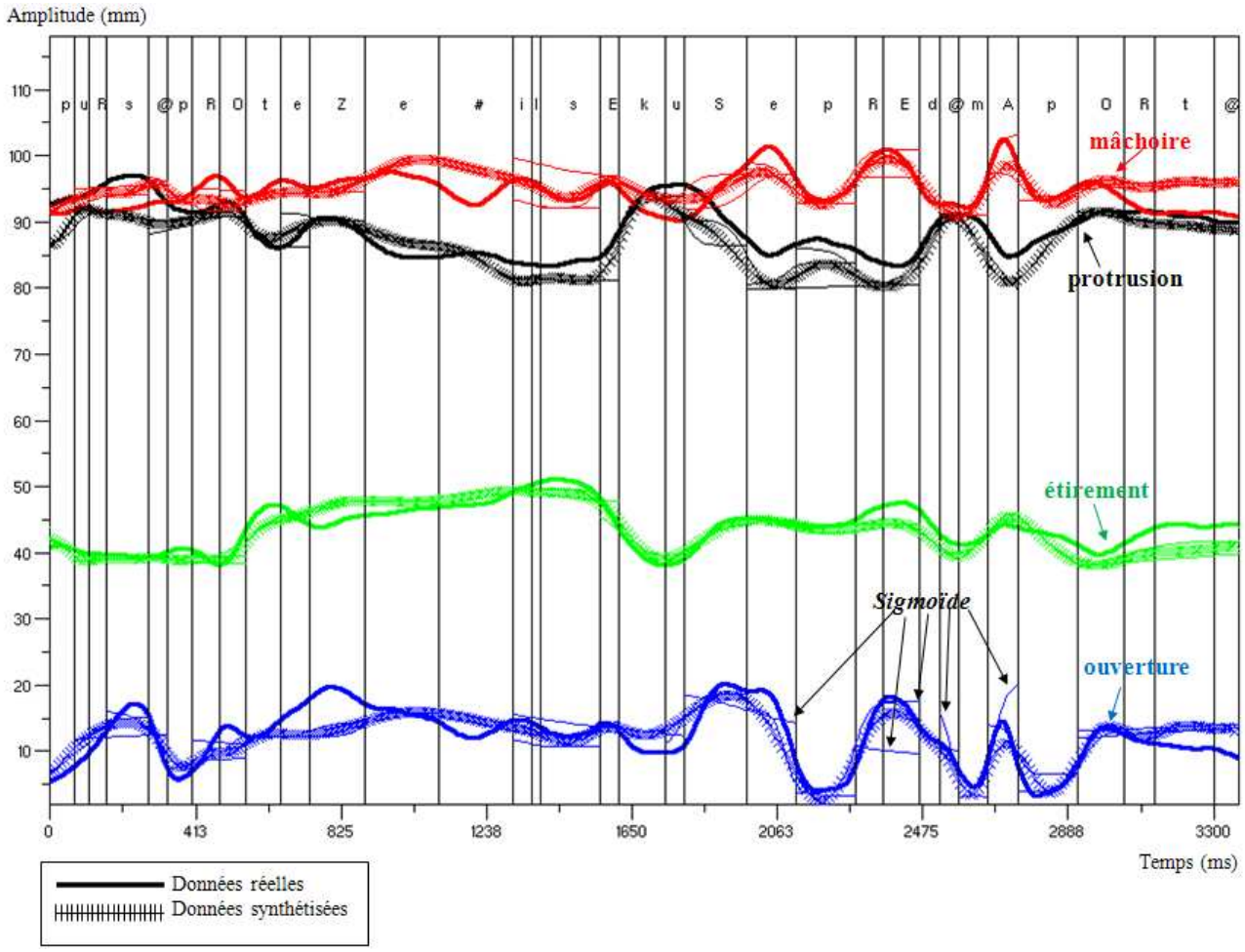

FIG. 6.10 – Synthèse de la phrase "Pour se protéger, il s'est couché près de ma porte"(Alphabet phonétique SAMPA - cf. Annexe A).

Chapitre 7

Analyse statistique de la qualité de la synthèse

Le chapitre précédent nous a permis de constater que l'allure des mouvements articulatoires générés par synthèse approchait correctement les mouvements articulatoires réels. Nous allons quantifier dans ce chapitre de façon objective la qualité de cette synthèse, puis, dans le chapitre suivant, nous réaliserons des tests de perception qui vont permettre de quantifier de façon subjective la qualité des messages fournis par une tête parlante pilotée par nos paramètres articulatoires.

7.1 Evaluation de la qualité de la phase d'apprentissage

Afin d'évaluer notre système, nous allons tout d'abord mesurer la dégradation générée par la phase d'apprentissage. Suite à l'application de notre algorithme de prédiction de la coarticulation, cette phase analyse les données brutes et extrait soit un ensemble de sigmoïdes soit des informations statistiques sur les phonèmes (minimum, moyenne et maximum) dans le cas où la recherche de sigmoïde a échoué. La transformation des données initiales du corpus en sigmoïdes offre l'énorme avantage de faciliter l'adaptation lors de la synthèse à une vitesse d'élocution particulière ou à un mouvement articulatoire plus ou moins ample. En outre, la quantité d'informations stockées est considérablement réduite. Néanmoins, les mouvements articulatoires ne correspondent pas tous exactement à des sigmoïdes et par conséquent, les données obtenues dégradent légèrement les données brutes. De plus, notre algorithme de prédiction ne génère pas la recherche de sigmoïdes entre tous les phonèmes consécutifs et dans certains cas, il est possible qu'un phonème considéré comme "neutre" ne le soit pas réellement. Pour évaluer la dégradation provoquée par l'apprentissage, nous avons calculé l'erreur RMSE (Root Mean Square Error) et la corrélation entre les données réelles et les données estimées. L'erreur RMSE est calculée sous forme d'un pourcentage par rapport à l'étendue complète du paramètre concerné sur les zones de parole. L'erreur RMSE étant fortement influencée par les zones de large amplitude où des erreurs importantes risquent davantage d'apparaître, alors que de faibles déviations risquent d'être cruciales pour la perception (par exemple la fermeture des lèvres pour les bilabiales), nous avons choisi de compléter cette mesure par celle de la corrélation entre les deux signaux. Le tableau 7.1 montre les résultats obtenus si l'on considère l'ensemble de notre grand corpus monolocuteur. On constate que l'erreur moyenne est de 5.95% et que les 4 paramètres choisis subissent approximativement la même dégradation. Cette perte d'informations est due à plusieurs causes et la figure 7.1 en donne des exemples.

	Protrusion	Ouverture	Mâchoire	Etirement
Corrélation	90.69	90.20	88.58	91.27
RMSE	5.72	5.84	6.67	5.57

TAB. 7.1 – Evaluation de la dégradation générée par la modélisation des mouvements articulatoires sous forme de juxtaposition de sigmoïdes.

- Une sigmoïde ne peut être qu'une approximation du mouvement articulatoire qui par son essence est dû à une multitude de facteurs. Quand aucune sigmoïde n'a été trouvée, l'apprentissage mémorise la valeur moyenne du(des) phonème(s) concerné(s) et dans ce cas, l'approximation est donc plus mauvaise, puisqu'aucune coarticulation n'est prise en compte.
- Certains phonèmes ont été considérés comme neutres dans notre étude théorique alors qu'ils influencent légèrement le mouvement. En fait, leur influence est d'autant plus grande quand le mouvement articulatoire s'écarte de la position moyenne.
- Les séquences sont très influencées par les silences et les souffles au début, en cours ou en fin de phrase et compte-tenu de leur large variabilité, il nous est difficile de modéliser convenablement les mouvements articulatoires durant ces périodes.

7.2 Evaluation statistique de la qualité de la synthèse

7.2.1 Evaluation globale

Pour évaluer la qualité de notre synthèse coarticulatoire, nous avons synthétisé l'ensemble des phrases de notre grand corpus (après les avoir successivement enlevées de la base de données d'apprentissage). En comparant les données réelles et les données synthétisées sur l'ensemble de ces phrases, nous avons une bonne estimation de l'erreur générée par la synthèse. La mesure de l'erreur RMSE et de la corrélation constituent nos critères de contrôle afin de pouvoir comparer avec la dégradation due à l'apprentissage et aussi pour être cohérent avec l'étude comparative de Beskow [10] qui a comparé plusieurs modèles coarticulatoires (Modèles de Cohen et Massaro [19] , modèle d'Öhman [42], deux modèles basés sur des réseaux de neurones et une méthode à base de règles développée par Beskow lui même [9]). Cette étude a montré que la méthode de Cohen et Massaro obtenait le meilleur score statistique (Taux d'erreur RMSE de 8.63% et taux de corrélation de 0.689). C'est donc tout naturellement avec cette dernière méthode que nous avons choisi de comparer nos résultats.

Dans son corpus, Beskow a choisi 200 phrases. En ce qui nous concerne, nous en avons choisi 100, mais elles sont plus longues. Par conséquent, il nous a semblé possible en première approche de comparer notre méthode testée sur notre corpus avec celle testée par Beskow sur un corpus différent. Pour cela, nous avons calculé l'erreur RMSE et la corrélation entre les données brutes et estimées sur notre corpus. Nous obtenons un taux d'erreur de 9.3% et un taux de corrélation de 74.44%. En comparant avec les résultats de Beskow, on peut en déduire que notre solution a un taux d'erreur proche de celui obtenu avec la synthèse utilisant la méthode de Cohen et Massaro et un taux de corrélation supérieur. Cette première constation montre que notre technique de synthèse par concaténation est efficace.

Néanmoins, la comparaison entre une méthode de prédiction basée sur un corpus suédois et une autre basée sur un corpus français est forcément faussée et ne peut donc fournir qu'une

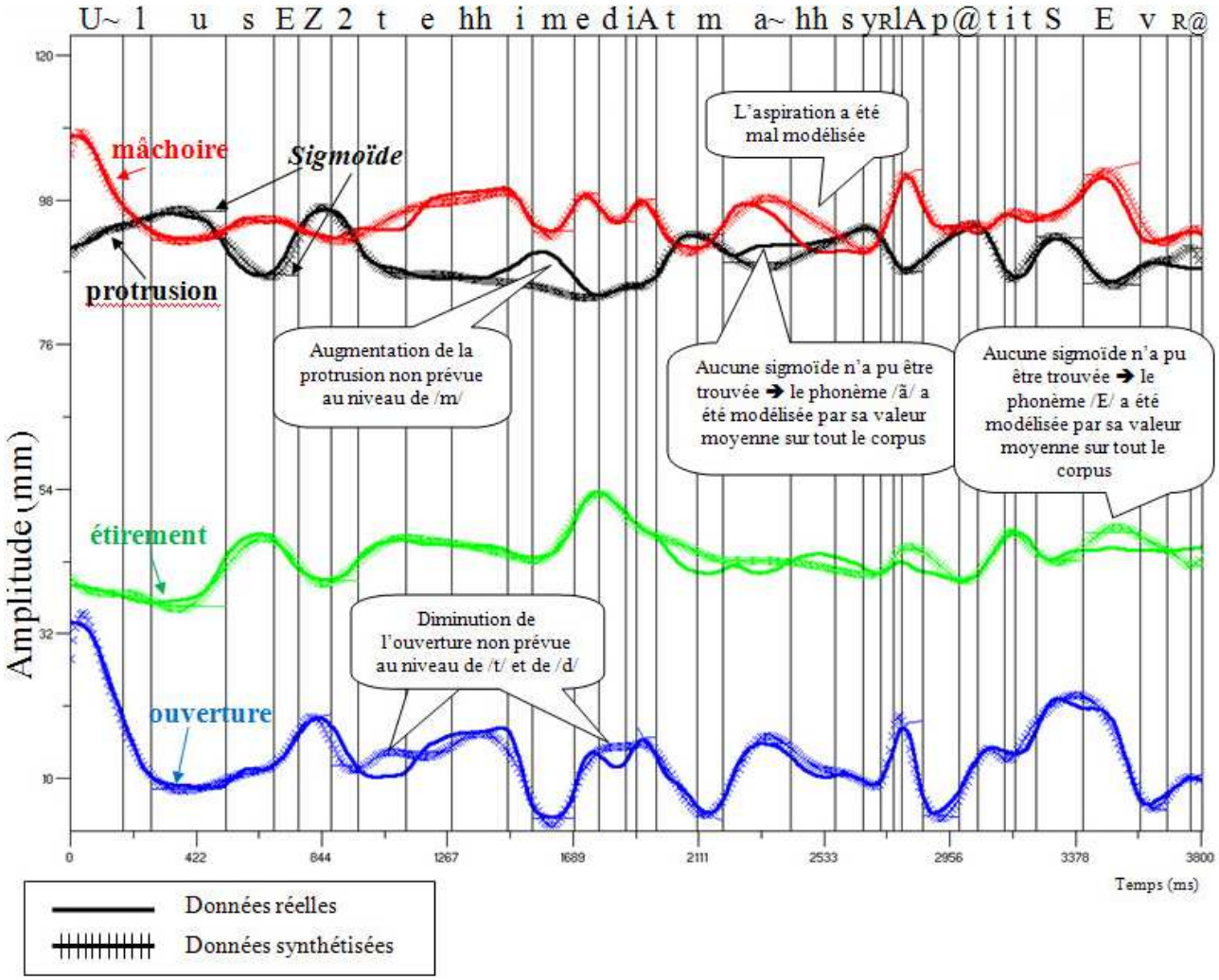

FIG. 7.1 – Modélisation de la phrase "Un loup s'est jeté immédiatement sur la petite chèvre" par un ensemble de sigmoïdes (Alphabet phonétique SAMPA - cf. Annexe A).

première approximation de la qualité de la synthèse. Afin d'évaluer les performances de notre synthèse avec davantage de précision, nous avons choisi de comparer nos résultats avec ceux obtenus par J. Feldmar, un membre de notre équipe qui a implanté une version améliorée de l'algorithme de Cohen et Massaro. Le grand corpus monolocuteur a servi de base à cette modélisation ; 29 phrases de ce corpus ayant été retirées pour réaliser la synthèse avec la méthode de Cohen et Massaro. Le tableau 7.2 montre que la méthode de Cohen et Massaro présente globalement de meilleurs taux de corrélation et une erreur moindre que notre technique basée sur la concaténation. La protrusion est le paramètre dont le taux de corrélation et l'erreur RMSE se rapprochent le plus de ceux calculés en appliquant la méthode de Cohen et Massaro, mais c'est aussi le paramètre qui est globalement le moins bien modélisé. La figure 7.2 montre le détail des valeurs de corrélation et d'erreur sur les 29 phrases qui ont servi de test pour le paramètre "Protrusion". On constate que le meilleur score n'est pas systématiquement obtenu avec la méthode de Cohen et Massaro. En revanche, celle-ci semble plus stable alors que notre technique présente pour la phrase n° 13 un taux de corrélation catastrophique (36,4 %) et pour la phrase

		Protrusion	Ouverture	Mâchoire	Etirement
Corrélation	Cohen et Massaro	77.61	85.86	85.33	85.58
	Concaténation	73.62	74.44	73.99	75.57
RMSE	Cohen et Massaro	8.37	6.48	7.62	6.79
	Concaténation	10.32	9.3	9.91	9.28

TAB. 7.2 – Comparaison entre la méthode de prédiction de la coarticulation de Cohen et Massaro et notre technique basée sur la concaténation.

n°29 une erreur de 21%. Examinons plus en détail ces deux phrases afin de cerner la cause de ces erreurs importantes. La figure 7.3 montre l'évolution de la protrusion pour la phrase n°13 "La poire est un fruit à pépins". On constate que l'erreur la plus importante apparaît au niveau du phonème $\tilde{\varepsilon}$ dont la protrusion estimée est beaucoup plus faible que la protrusion réelle. Cette erreur est due au fait que la suite /e_t_$\tilde{\varepsilon}$/ n'apparaît pas dans le corpus d'apprentissage, donc une complétion a été réalisée en utilisant l'enregistrement isolé du phonème /$\tilde{\varepsilon}$/. Or, en mode isolé, la protrusion enregistrée est de $81.37mm$ (Rappelons qu'il s'agit d'une distance par rapport à un point de référence arbitraire, donc seuls les écarts sont significatifs), c'est-à-dire très faible, donc l'estimation qui en découle se rapproche de cette valeur. Une telle erreur semble donc assez facile à corriger si l'on prend soin d'enregistrer plusieurs fois les voyelles isolées qui ne font pas partie des VCV ou VCCV retenues. Etudions maintenant la phrase n°29, "Ce soir, nous nous coucherons plus tard", qui présente le taux d'erreur RMSE le plus important pour la protrusion. Notons que dans le même temps, cette phrase a un taux de corrélation de 92.3% qui prouve que l'allure estimée "colle" presque parfaitement à l'allure réelle comme le montre la figure 7.4. L'algorithme de correction d'amplitude que nous avons appliqué a bien joué son rôle en ce qui concerne la conservation de l'effet syntagmatique. L'effet paradigmatique semble avoir souffert de ce recalage, mais seuls des tests de perception pourront nous apprendre si cela nuit réellement à la compréhension car la protrusion dispose d'un degré de liberté assez grand.

7.2.2 Influence du paramètre articulatoire

Notre étude montre que la correction d'amplitude réalisée afin de satisfaire la cohérence syntagmatique est efficace seulement pour le paramètre "Protrusion" ; les autres paramètres obtenant des valeurs de corrélation et d'erreur RMSE très légèrement meilleures si aucune correction n'est effectuée. En revanche, en ce qui concerne la protrusion, la correction d'amplitude fait passer la corrélation de 54.69% à 73.62% et l'erreur RMSE de 18.79% à 10.32%. L'effet paradigmatique semble donc prédominant pour les paramètres d'étirement ou d'ouverture des lèvres et de la mâchoire alors que l'effet syntagmatique est prédominant pour la protrusion qui semble beaucoup moins soumise que les autres paramètres à des contraintes articulatoires. Or, la méthode de Cohen et Massaro, basée sur les fonctions de dominance prend peu en considération l'effet syntagmatique et c'est notamment pour cela que son score est le moins bon pour la protrusion.

Les figures 7.5, 7.6 et 7.7 montrent la comparaison phrase par phrase pour les autres paramètres. L'erreur RMSE reste relativement stable de phrase en phrase alors que le taux de corrélation est plus variable. Les taux de corrélation de la mâchoire et de l'ouverture labiale sont très voisins ce qui montre que nous avons sans doute eu raison d'appliquer le même algorithme de gestion de la coarticulation pour ces deux paramètres.

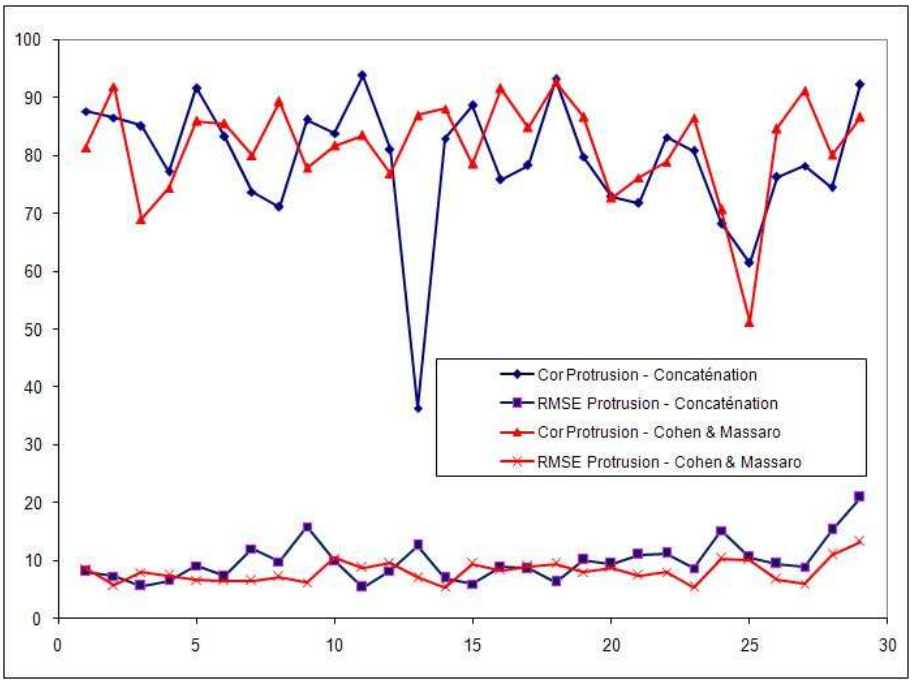

FIG. 7.2 – Corrélations et erreurs RMSE obtenues pour les 29 phrases de test pour le paramètre "Protrusion" avec la technique de modélisation par concaténation et celle de Cohen et Massaro.

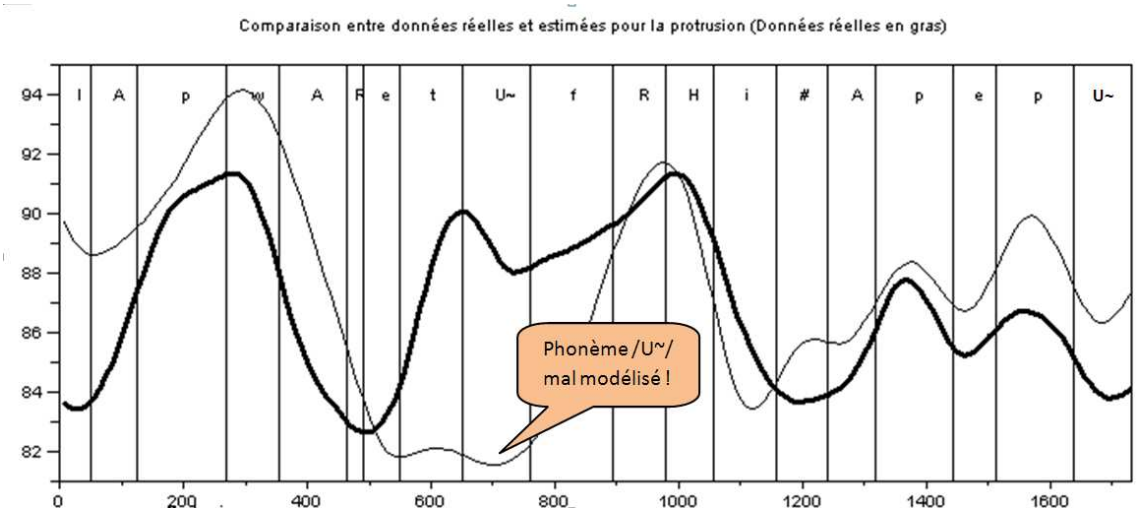

FIG. 7.3 – Evolution de la protrusion pour la phrase "La poire est un fruit à pépins" (Alphabet phonétique SAMPA - cf. Annexe A).

7.2.3 Influence des phonèmes

Si nous comparons les deux méthodes de prédiction de la coarticulation phonème par phonème, on constate que notre technique par concaténation peut rivaliser avec la méthode de Cohen et Massaro au niveau de la protrusion. Pour ce paramètre, nos résultats sont aussi bons voire

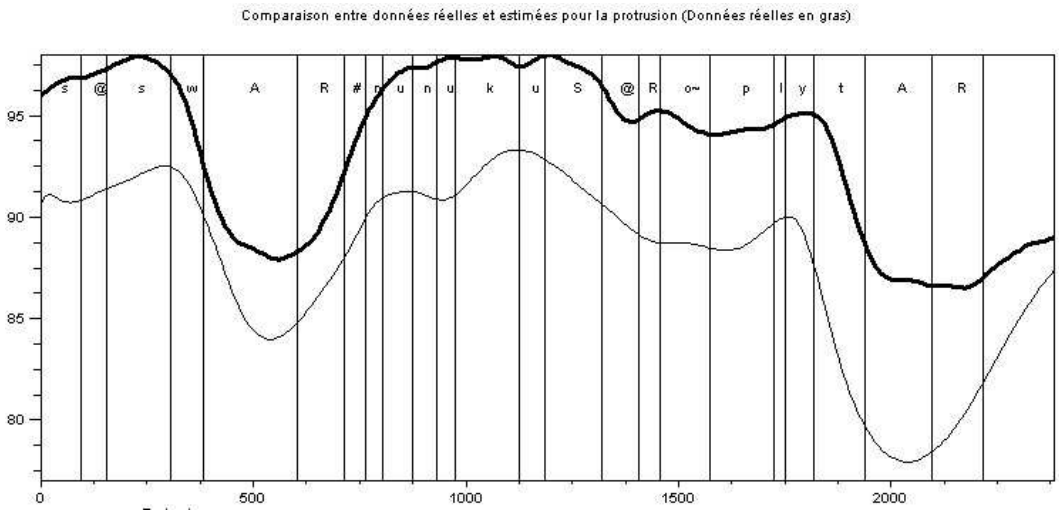

FIG. 7.4 – Evolution de la protrusion pour la phrase "Ce soir, nous nous coucherons plus tard" (Alphabet phonétique SAMPA - cf. Annexe A).

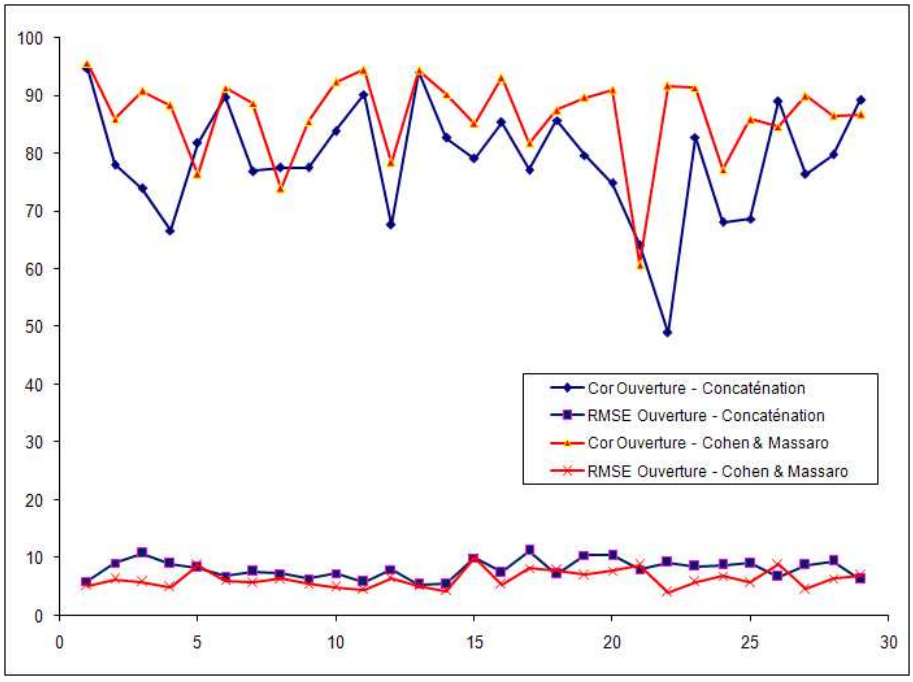

FIG. 7.5 – Corrélations et erreurs RMSE obtenues pour les 29 phrases de test pour le paramètre "Ouverture" avec la technique de modélisation par concaténation et celle de Cohen et Massaro.

meilleurs si l'on considère les phonèmes liés fortement à la protrusion (notamment /i/, /y/ , /u/ , /ʃ/)comme le montre la figure 7.8. En revanche, la méthode de Cohen et Massaro présente de meilleurs résultats pour les autres paramètres. En ce qui concerne l'ouverture, et contrairement aux affirmations de Cosi [22] qui a corrigé le modèle de Cohen et Massaro, les bilabiales obtiennent de bons scores avec la méthode basée sur les fonctions de dominance (Fig. 7.9)

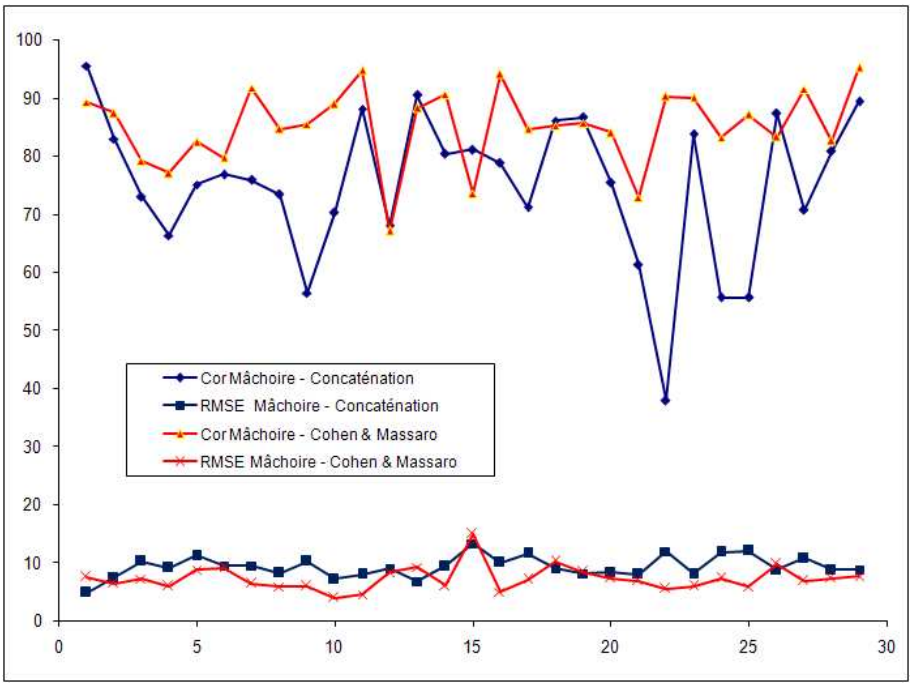

FIG. 7.6 – Corrélations et erreurs RMSE obtenues pour les 29 phrases de test pour le paramètre "Mâchoire" avec la technique de modélisation par concaténation et celle de Cohen et Massaro.

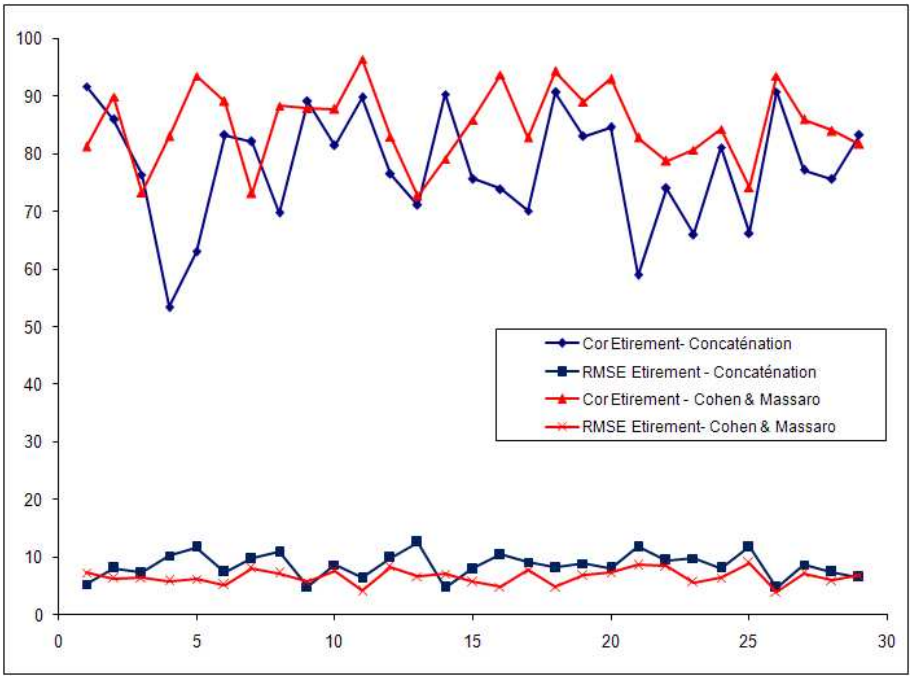

FIG. 7.7 – Corrélations et erreurs RMSE obtenues pour les 29 phrases de test pour le paramètre "Etirement" avec la technique de modélisation par concaténation et celle de Cohen et Massaro.

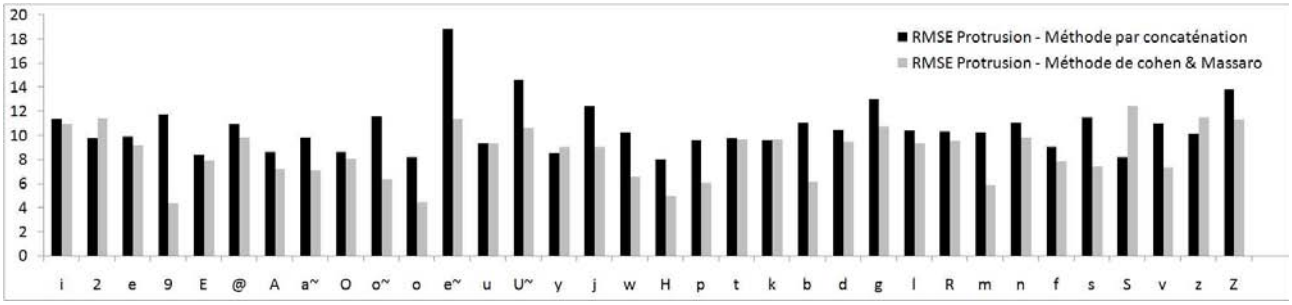

FIG. 7.8 – Erreurs RMSE obtenues en fonction de chaque phonème pour le paramètre "Protrusion" pour les deux méthodes de prédiction de la coarticulation (Alphabet phonétique SAMPA - cf. Annexe A).

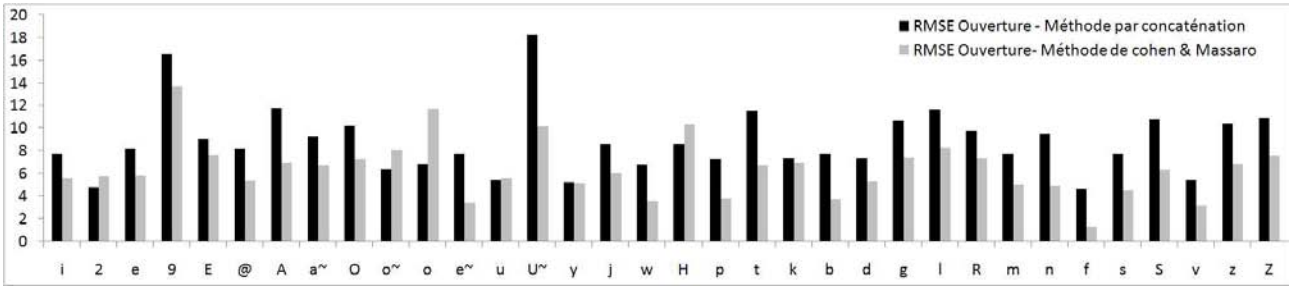

FIG. 7.9 – Erreurs RMSE obtenues en fonction de chaque phonème pour le paramètre "Ouverture" pour les deux méthodes de prédiction de la coarticulation (Alphabet phonétique SAMPA - cf. Annexe A).

7.2.4 Influence des séquences

La méthode de Cohen et Massaro obtient globalement des meilleurs scores que notre technique basée sur la concaténation. Dans ce paragraphe, nous allons essayer de vérifier s'il en est de même pour des séquences complexes de type $VCCV$ où l'effet de la coarticulation est davantage marqué que dans le cas de VCV. De nombreuses études ont été menées sur des $VCCV$, notamment en ce qui concerne l'arrondissement des lèvres pour déterminer l'ampleur de la coarticulation anticipatrice; nous pouvons citer par exemple l'étude de Abry et Lallouache qui a débouché sur le modèle expansionniste [2] . Il est intéressant sur de telles séquences de comparer notre méthode basée sur la concaténation, proche du modèle théorique Look-Ahead avec le modèle de Cohen et Massaro, basé sur les fonctions de dominance définies par Löfqvist [50] et se rapprochant davantage du modèle théorique de type Time-Locked en ce qui concerne la prédiction de l'anticipation. Les mesures de la corrélation et de l'erreur RMSE sur ces séquences sont reportées dans le tableau 7.3. Le corpus utilisé contient 92 $VCCV$, les plus courantes de la langue française. Les résultats montrent pour les deux méthodes des résultats très proches en ce qui concerne les paramètres d'étirement et d'ouverture des lèvres et de la mâchoire. En revanche, notre technique par concaténation obtient un score nettement meilleur pour la protrusion (20% de plus pour la corrélation et un taux d'erreur diminué de 4%). A partir du corpus existant, notre méthode a donc permis d'estimer avec une bonne précision les séquences complexes de type $VCCV$ préalablement retirées du corpus pour réaliser la synthèse. Ceci prouve que notre

		Protrusion	Ouverture	Mâchoire	Etirement
Corrélation	Cohen et Massaro	57.17	87.19	88.68	89.56
	Concaténation	78.40	88.63	88.74	91.04
RMSE	Cohen et Massaro	14.81	6.70	7.50	7.51
	Concaténation	10.59	6.79	8.06	6.60

TAB. 7.3 – Comparaison entre la méthode de prédiction de la coarticulation de Cohen et Massaro et notre technique basée sur la concaténation pour les VCCV de notre grand corpus.

technique de complétion et d'adaptation (du temps et de l'amplitude) est efficace. En ce qui concerne la locutrice choisie pour l'enregistrement de notre grand corpus, ces résultats montrent également que le phénomène d'arrondissement des lèvres se rapproche davantage de la stratégie Look-Ahead que du modèle Time-Locked.

7.2.5 Premières conclusions

Même si les résultats avec notre méthode restent globalement moins bons que ceux obtenus après application de la méthode de Cohen et Massaro, les résultats sont encourageants et perfectibles. Tout d'abord, notre technique est fortement liée à la taille et à la qualité du corpus. Plus le nombre de $VCCV$ enregistrées est important, plus le nombre de complétions sera réduit, diminuant ainsi le risque d'erreur. Afin d'améliorer les résultats obtenus, il semble aussi très important d'enregistrer plusieurs fois les voyelles isolées qui ne font pas partie des $VC...CV$ retenues dans le corpus. En effet, si V_1 est une voyelle enregistrée seulement de façon isolée, $V_1C..CV$ ou $VC...CV_1$ est estimé en tenant compte de la position de V_1 par rapport aux autres voyelles fréquemment utilisées dans le corpus. Une autre erreur semble due à notre tableau de classification des phonèmes (Fig. 4.18) pour lequel nous avons considéré comme neutres un certains nombre de phonèmes qui en fait ne le sont pas réellement. Par exemple, le phonème /n/ dans une séquence $/V_1nV_2/$ reste neutre au niveau de l'ouverture des lèvres uniquement si celle-ci n'est pas trop forte. Si /n/ semble avoir peu d'influence dans la suite /ani/, il en a une grande dans /ana/ car il force les lèvres à se refermer entre les deux voyelles comme le montre l'extrait de la phrase "Ce petit canard apprend à nager" (Fig. 7.10). En ce qui concerne la protrusion, un phénomène similaire est relaté dans les études sur le Suédois de Mc Allister [59] et Engstrand [28]. Il ont montré un relâchement de la protrusion dans une séquence $/uCu/$ où C n'a pas de contrainte labiale. Selon eux, la présence ou l'absence de cible pour les consonnes non spécifiées dépend du locuteur et de la langue utilisée - d'après leurs conclusions, les fricatives alvéolaires sont produites en français et en Allemand avec les lèvres rétractées alors qu'elles n'ont pas d'influence dans plusieurs autres langues étudiées -

Notre tableau de classification (tab. 4.18) et notre algorithme ne tiennent pas compte de ce phénomène et notre estimation est donc erronée. Il est possible de corriger ceci si l'on considère qu'un phonème perd sa neutralité dans le cas où les valeurs de la sigmoïde trouvée au niveau de ce dernier sont en dehors de sa plage de variation (déterminée en utilisant le corpus d'apprentissage). Dans le cas d'une séquence /ana/, si l'on constate que l'ouverture des lèvres au niveau du /a/ n'est pas dans la plage de variation possible pour /n/, deux sigmoïdes vont être recherchées, l'une entre /a/ et /n/ et l'autre entre /n/ et /a/. Ceci est tout à fait cohérent avec la théorie de Keating [44] qui utilise la notion de fenêtres pour quantifier le degré de variation possible d'un paramètre articulatoire. La figure 7.11 montre sur un exemple que le raccordement entre une séquence VCV, modélisée par trois fenêtres (au sens de Keating) peut être réalisé en utilisant une ou deux sigmoïdes.

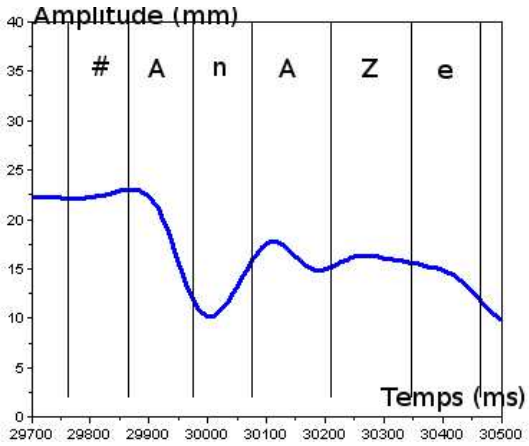

FIG. 7.10 – Evolution de l'ouverture des lèvres pour la séquence /ana/ (extrait de la phrase "Ce petit canard apprend à nager") (Alphabet phonétique SAMPA - cf. Annexe A).

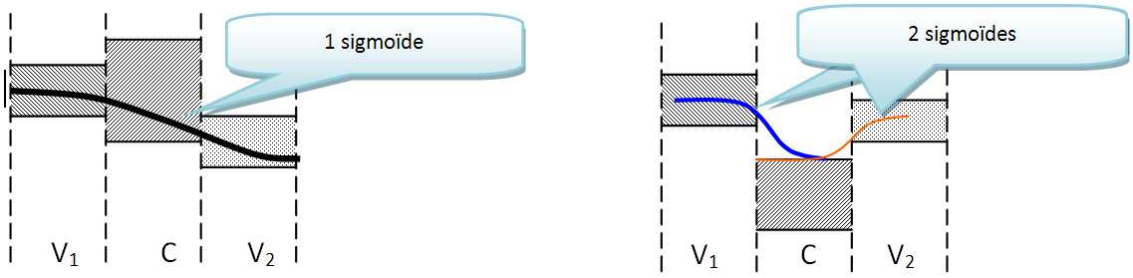

FIG. 7.11 – Raccordement entre fenêtres (au sens de Keating) par une ou deux sigmoïdes.

7.2.6 Raffinement de la méthode par concaténation

Afin de tenir compte du défaut constaté dans le paragraphe précédent, nous avons modifié notre algorithme de prédiction. Soit C, une consonne considérée neutre dans une séquence $Pho_1...C...Pho_2$ où Pho_1 et Pho_2 sont non neutres pour le paramètre considéré. Si la valeur de la sigmoïde estimée au niveau du phonème C n'est pas dans la fenêtre de variation de C, alors, nous supprimons la neutralité de ce phonème. En conséquence, deux sigmoïdes vont être recherchées, l'une de Pho_1 à C et l'autre de C à $Pho2$ (Fig. 7.11).

Pour estimer la fenêtre de variation de C, nous nous sommes basé sur la position centrale du phonème. En explorant l'ensemble du corpus, nous avons recherché la valeur moyenne et l'écart type σ du centre du phonème. Afin de ne pas prendre en considération les valeurs extrêmes qui peuvent être causées par une hyper ou une hypo articulation temporaire du locuteur, nous avons défini une fenêtre de largeur $2.\sigma$ autour de la moyenne. Si la répartition était gaussienne, cela signifierait que nous prenons en compte 68% de la population ; cette proportion nous a semblé acceptable.

En ce qui concerne l'apprentissage, les sigmoïdes trouvées approchent mieux les données réelles. Le gain au niveau de la corrélation est de 0.75% pour la protrusion et de 1.5% pour les

autres paramètres. L'erreur RMSE baisse quant à elle de 0.5%. Le taux de corrélation global est maintenant proche de 92% et l'erreur moyenne est de 5.45% ce qui prouve que le choix des sigmoïdes comme modèle principal est justifié.

En ce qui concerne la synthèse, la modification effectuée apporte une amélioration notable pour les paramètres d'ouverture des lèvres et de la mâchoire ainsi que pour l'étirement. Le taux moyen de corrélation pour ces trois paramètres passe de 74.66% à 76.9% et l'erreur RMSE moyenne de 9.5% à 9.1%. En ce qui concerne ces trois paramètres, les meilleurs résultats sont obtenus sans effectuer la correction d'amplitude, c'est-à-dire sans tenir compte de l'effet syntagmatique car nous l'avons évoqué précédemment, l'effet paradigmatique est fortement prédominant pour ces paramètres. En revanche, comme la protrusion dépend beaucoup plus de l'effet syntagmatique, la modification que nous avons apportée dégrade les performances pour ce paramètre ; la corrélation passe de 73.62% à 69.39% et l'erreur RMSE de 10.32% à 10.55%. Les fenêtres que nous avons utilisées sont en fait beaucoup trop liées au paradigme et ne sont pas adaptées pour la protrusion qui dispose d'un grand degré de liberté. La solution consiste à traiter différemment les paramètres en accordant davantage d'importance à l'effet syntagmatique pour la protrusion et davantage d'importance à l'effet paradigmatique pour le mouvement de la mâchoire, l'ouverture et l'étirement des lèvres.

7.3 Conclusion

Les mesures objectives réalisées montrent que les mouvements articulatoires peuvent de façon très satisfaisante être modélisés par une concaténation de sigmoïdes ce qui permet de réduire considérablement la base de données à manipuler sans perte importante d'information. La modification des paramètres de contrôle des sigmoïdes permet d'adapter la synthèse à des vitesses d'élocution spécifiques et de pouvoir contrôler l'amplitude des mouvements afin de tenir compte de la prosodie de la phrase ou pour gérer l'hyper ou l'hypo articulation.

Cependant, notre méthode est fortement tributaire d'un bon enregistrement du corpus. En particulier, une bonne acquisition des voyelles isolées est primordial car la position de ces voyelles détermine les coefficients de la combinaison linéaire qui permet de reconstruire par complétion les VC...CV dont l'une au moins des voyelles ne fait pas partie des VC...CV enregistrées lors de l'apprentissage. De plus, le maintien d'un style articulatoire neutre de la part du locuteur qui a enregistré le corpus est indispensable pour éviter que les sigmoïdes trouvées lors de l'apprentissage soient trop tributaires d'un effet d'hyper ou d'hypo articulation. Un bon enregistrement et un corpus plus dense devrait aussi limiter le nombre de sigmoïdes qui n'ont pas été trouvées lors de l'apprentissage. Comme nous l'avons dit précédemment, si une séquence VC...CV n'a pas pu être approximée par une sigmoïde, les phonèmes caractéristiques sont quantifiés par leur valeur moyenne ce qui constitue une approximation beaucoup moins fidèle. La méthode développée par Cohen et Massaro est globale et subit moins les conséquences des erreurs ponctuelles.

L'erreur résiduelle est également en partie due au choix de $VC...CV$ comme séquence de décomposition. Ce choix nous a conduit à considérer que les voyelles extrêmes n'étaient pas influencées par les phonèmes voisins car pour les voyelles, tous les paramètres sont quantifiés. Ceci n'est qu'en partie vrai. Abry et al [1] ont montré que dans une syllabe CV, l'arrondissement des consonnes /ʃ/ et /ʒ/ se propageait au niveau de la voyelle quand celle-ci n'était pas protruse. Par exemple, lors de la prononciation du mot "Chat", la voyelle /a/ a une protrusion beaucoup

plus marquée que la normale. Lors de la concaténation des $VC...CV$, notre algorithme gère en partie ce phénomène grâce à la correction d'amplitude (recalage syntagmatique).

Il est également important de noter que la technique par concaténation atteint des performances voisines de celle de Cohen et Massaro en ce qui concerne le mouvement de protrusion et même nettement meilleures si l'on considère uniquement les $VCCV$. Nous avons montré que la protrusion est beaucoup plus sensible que les autres paramètres à la cohérence syntagmatique. La technique des fonctions de dominance ne permet de prendre en compte que l'effet paradigmatique car, avant mélange, des fonctions de dominance sont attribuées à chaque phonème sans tenir compte a priori de la séquence qui devra être modélisée. L'importance de l'axe syntagmatique est d'autant plus forte en français, car une analyse en composantes principales sur notre grand corpus a montré que la protrusion explique à elle seule 66% de la variance (Tab. 4.2). En Anglais, en revanche, le mouvement de la mâchoire est largement dominant.

D'une façon générale, même si les résultats statistiques sont bons, les mouvements générés par la méthode de Cohen et Masaro présentent cependant des défauts. D'une part, les mouvements sont plus saccadés que ceux obtenus avec la synthèse par concaténation. D'autre part, certains changements de direction ou certaines transitions très brutales ne pourraient pas être réalisés par les articulateurs réels. La figure 7.12 montre la synthèse avec les deux méthodes du mouvement de protrusion pour la phrase "Maman a préparé une galette pour jeudi". Les résultats statistiques sont très voisins pour cette phrase, mais on distingue à plusieurs reprises les défauts constatés ci-dessus au niveau du mouvement synthétisé par la technique basée sur les fonctions de dominance.

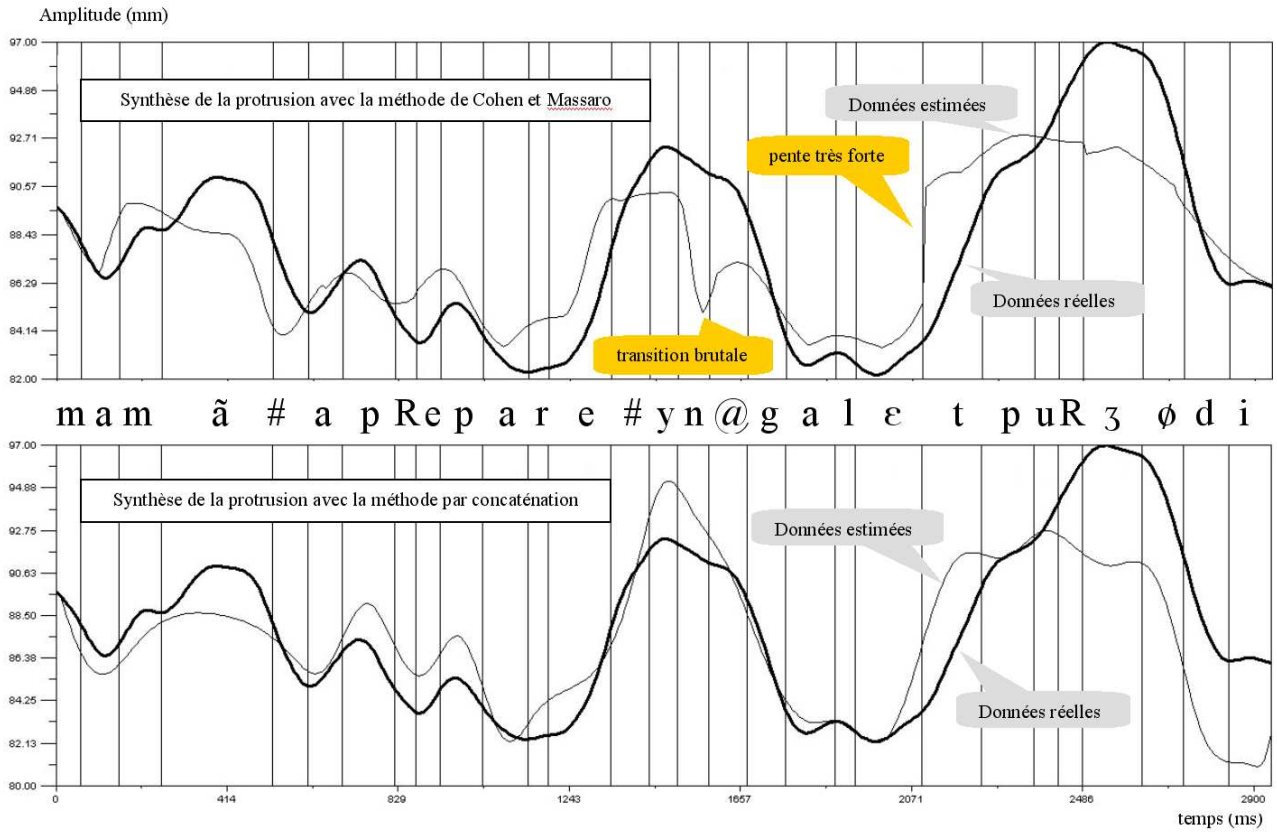

FIG. 7.12 – Synthèse avec les deux méthodes de la phrase "Maman a préparé une galette pour jeudi" (Alphabet phonétique SAMPA - cf. Annexe A).

Chapitre 8

Tests de perception

Cette thèse s'inscrit dans un projet dont l'un des buts est de faciliter la communication avec les sourds et les malentendants. L'évaluation objective du chapitre précédent a permis d'obtenir des indications statistiques sur notre système de synthèse des mouvements articulatoires. Mais la perception de la qualité de l'animation du visage parlant n'est pas forcément proportionnelle à ces informations. Beskow [10] a montré dans son étude comparative de plusieurs modèles coarticulatoires que le modèle de Cohen et Massaro qui avait obtenu le meilleur score statistique était devancé par celui d'Öhman lors des tests de perception, ce dernier étant lui même devancé par le modèle à base de règles de Beskow [9]. Le but de ce chapitre est de vérifier si une tête parlante pilotée par nos paramètres articulatoires améliore effectivement la compréhension de l'auditeur et de comparer, en terme de perception, l'apport de notre modèle par rapport à celui de Cohen et Massaro.

8.1 Construction d'une tête parlante pilotée par nos paramètres articulatoires

Il serait utopique de penser que nos 4 paramètres articulatoires permettent de reconstruire une tête parlante sophistiquée présentant toutes les subtilités des mouvements faciaux. Néanmoins, la modélisation des mouvements labiaux et de la mâchoire devraient améliorer de façon significative l'intelligibilité du message oral. Ouni et al [66] ont montré dans leur étude comparative sur l'intelligibilité des têtes parlantes que la bimodalité signal acoustique avec les lèvres seules obtenait quasiment les mêmes scores de perception que la bimodalité avec le visage complet prouvant ainsi que les mouvements labiaux renferment la majeure partie de l'information visuelle utile à l'intelligibilité. Nous avons néanmoins choisi de modéliser une tête complète même si dans notre étude, l'animation conserve seulement les lèvres. La tête générée offrira ainsi ultérieurement la possibilité de rajouter des informations utiles au visage afin de pouvoir simuler les émotions à travers les mouvements des yeux, des sourcils, etc...

C'est un membre de notre équipe, J. Feldmar qui a réalisé la tête parlante pilotée par nos paramètres articulatoires. Il a réalisé un morphing 3D linéaire en utilisant comme forme de base une tête artificielle issue du logiciel commercial POSER. Une opération manuelle a été nécessaire pour construire par déformation du maillage les formes cibles associées à nos paramètres articulatoires (Ouverture des lèvres et de la mâchoire, étirement des lèvres et protrusion). A l'issue de la phase de calibration qui permet de faire correspondre les valeurs extrêmes de ces formes avec les données mesurées du corpus, nos 4 paramètres permettent donc de contrôler les déformations

du visage parlant comme le montre la figure 8.1.

FIG. 8.1 – Visage 3D contrôlé par nos paramètres articulatoires.

8.2 Procédure de test

Notre but est double. D'une part, nous devons vérifier si le visage parlant piloté par nos paramètres articulatoires augmente réellement la compréhension du message audio d'origine et quantifier l'apport d'intelligibilité par rapport au visage réel. D'autre part, nous souhaitons comparer les performances de notre méthode de prédiction de la coarticulation avec celle de Cohen et Massaro. A cette fin, la tête parlante que nous générons sera pilotée par les 4 paramètres articulatoires déduits soit de l'algorithme de prédiction par concaténation, soit de l'algorithme basé sur les fonctions de dominance. Nous aurions pu utiliser Baldi, la tête parlante développée par Cohen et Massaro à l'université de Santa Cruz en Californie pour tester la deuxième méthode de prédiction de la coarticulation, mais la comparaison de deux visages parlants très différents aurait été peu significative.

D'une façon générale, deux objectifs différents ont conduit à la création de visages parlants : l'élaboration d'un modèle le plus réaliste possible - notamment pour le monde du cinéma ou des films d'animation - ou l'augmentation de l'intelligibilité du message transmis, notamment pour les lecteurs labiaux.

Comme nous l'avons vu au paragraphe 3.2.1, les premiers modèles [16][21][30], s'attachant davantage au rendu, négligent souvent l'effet de la coarticulation labiale et sont lourds à mettre en œuvre du fait de la taille conséquente des données audiovisuelles à manipuler. Les concepteurs de ces modèles réalisent des tests de perception avec des questions "ouvertes" ou "fermées". Les questions "ouvertes" [16][21] font l'objet d'évaluations subjectives du modèle. Des vidéos sont présentées à des participants et ils doivent donner leurs impressions sur le réalisme, la synchronisation, ... De tels tests, longs et compliqués à dépouiller rendent difficiles la comparaison entre modèles. Dans le cas des questions "fermées" ou "à choix forcé" [29][38], les tests les plus couramment utilisés consistent à diffuser à la fois des vidéos réelles et synthétiques. Les sujets doivent identifier la nature de chaque vidéo en indiquant si elle est naturelle ou de synthèse. Ces tests à "choix forcé" ont deux inconvénients majeurs. D'une part, les participants peuvent

se forger une opinion a priori des animations pendant le test ce qui peut ainsi influencer leur décision. D'autre part, les bonnes réponses peuvent résulter d'une sélection au hasard.

En ce qui concerne les modèles dont le but principal est d'augmenter l'intelligibilité du message à transmettre, plusieurs types de test existent. Massaro [58] utilise un ensemble de 264 mots monosyllabiques pour tester l'efficacité de la tête parlante *Baldi*. Les vidéos réelles et synthétisées de ces mots sont présentées à un ensemble de sujets qui sont chargés de les identifier. En comparant les proportions de mots correctement identifiés et en analysant les confusions, Massaro en tire des conclusions sur la qualité de la tête parlante. Plutôt que de diffuser seulement la vidéo, d'autres tests [66][21] y superposent un signal audio bruité. Cette solution a pour avantage de pouvoir quantifier l'apport de la vidéo en fonction du rapport signal/bruit.

Cosker [23] a quant à lui l'idée originale d'utiliser l'effet McGurk [60] pour réaliser des tests de perception. McGurk a constaté que si l'on diffuse le message audio /ba/ pendant qu'un locuteur prononce /ga/, alors on perçoit /ga/. Au niveau du cerveau, ce test montre que les canaux audio et vidéo se combinent pour produire la perception qui les associe le mieux. Cosker propose un test où il présente aux participants un ensemble de couples audio/vidéo comportant soit la vidéo réelle, soit la vidéo synthétique. Dans chaque couple, l'audio et la vidéo ne correspondent pas au même mot. Lors de la diffusion, il est simplement demandé aux participants d'indiquer le mot qu'ils ont entendu. L'auteur suppose que si l'algorithme de synchronisation des lèvres est bon alors une combinaison des mots issus de l'audio et de la vidéo devrait être perçue par le sujet alors que dans le cas contraire le signal audio devrait l'emporter sur le signal vidéo. Cette méthode est intéressante, mais elle nécessite un choix judicieux des mots. Par exemple, en diffusant l'audio "Bat" et la vidéo "Vet", le sujet devrait percevoir "Vat".

En ce qui nous concerne, nous souhaitons valider notre méthode de prédiction de coarticulation sur des mots de taille variable comportant des séquences VCV ou $VCCV$. La diffusion d'un signal vidéo avec un signal audio plus ou moins bruité nous a semblé la meilleure solution. En effet, notre tête parlante n'est pas suffisamment réaliste ; donc les tests consistant à demander aux participants si la tête présentée est réelle ou non ne sont pas adaptés. Le test avec l'effet McGurk ne convient pas non plus car celui-ci impose un choix très précis des mots ; des études anciennes sur l'effet McGurk [25] ont permis de définir un certain nombre de combinaisons "audio/vidéo/perception" valides, mais il s'agit essentiellement de monosyllabes associées à la langue anglaise. Or, nous nous intéressons au français et nous souhaitons valider certaines séquences $VCCV$ très précises pour lesquels l'effet de la coarticulation est davantage marqué.

Nous avons donc créé un corpus composé de 51 mots monosyllabiques issus de la liste de Lafon [48] et de 29 mots de notre choix, destinés à évaluer des combinaisons VCCV spécifiques. Le test des listes «cochléaires» du Professeur Lafon est très utilisé en audiophonologie. Elles sont censées évaluer les distorsions effectuées par les cochlées pathologiques sur le message phonétique. Ces listes phonétiquement équilibrées ont été étalonnées sur une population qu'on peut considérer représentative de la population française. Le test cochléaire comporte 20 listes de 17 mots de 3 phonèmes, correspondant à des mots monosyllabiques. En théorie, la fréquence d'occurrence des phonèmes dans le français parlé est la même à l'intérieur de chaque liste. Chaque liste peut ainsi être considérée comme un modèle réduit de la distribution des phonèmes dans la langue. Nous avons seulement retenu 3 listes de 20 mots pour notre test de perception afin de limiter la durée du test. Une locutrice dont la langue maternelle est le français a prononcé ces mots et nous avons enregistré le signal audio et vidéo. Notons que nous avons délibérément choisi une locutrice

différente de celle qui a enregistré notre corpus d'apprentissage afin de vérifier si notre méthode de prédiction de la coarticulation est robuste au changement de locuteur. Chaque mot enregistré est brouillé avec un bruit de type «coktail party» d'un niveau de 1 à 3. Au niveau 1, le mot est encore audible alors qu'au niveau 3, la visualisation labiale est indispensable pour discerner le mot prononcé. Nous avons fixé à l'aide de tests préliminaires les trois niveaux de bruits dans la région où la pente de la courbe qui représente le taux de reconnaissance en fonction du rapport Signal/Bruit est maximale (Fig. 8.2).

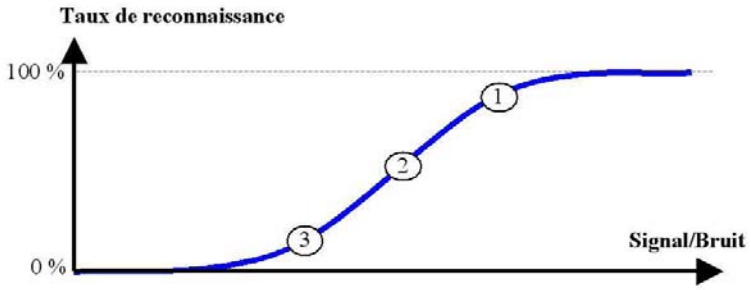

FIG. 8.2 – Choix des niveaux de bruit.

Lors du test, les mots sont mélangés entre eux et au final, chaque mot bruité est diffusé aléatoirement au moins 4 fois au sujet :

- une fois avec le son uniquement.
- une fois avec la vidéo réelle de la locutrice.
- une fois avec le visage parlant reconstruit en utilisant la méthode de prédiction de la coarticulation de Cohen et Massaro.
- une fois avec la tête parlante générée après application de la méthode de prédiction de la coarticulation basée sur la concaténation.

Afin d'éviter que le sujet s'attende à avoir exactement quatre diffusions de chaque mot, certaines diffusions sont répétées pour obtenir au total 350 diffusions par sujet. Le test est réalisé avec 20 personnes dont la langue maternelle est le français. Les mots bruités sont diffusés les uns après les autres et de façon aléatoire aux sujets, tous munis d'un casque (dont le volume est étalonné au début du test). Après chaque mot, les personnes doivent saisir sur le clavier d'un micro-ordinateur ce qu'elles pensent avoir compris. Un test d'entraînement composé d'une dizaine de mots leur est proposé avant le test final pour qu'elles se familiarisent avec le logiciel que nous avons développé.

8.3 Conclusion

A l'heure actuelle, les tests de perception n'ont pas encore été réalisés complètement. Les phases réalisées sont les suivantes :

- L'enregistrement du corpus
- Le décodage acoustico-phonétique qui permet d'obtenir une ensemble de phonèmes alignés sur le signal acoustique.
- Le bruitage des mots.

- La synthèse des mouvements de la mâchoire, de protrusion, d'ouverture et d'étirement des lèvres avec notre technique de prédiction de la coarticulation par concaténation.
- La génération des vidéos pilotées par nos paramètres articulatoires.

La synthèse des mouvements articulatoires pilotés par l'algorithme de Cohen et Massaro n'est pas terminée. En effet, pour que les deux synthèses puissent être comparées, il est indispensable que les formes de bouche initiale et finale soient identiques avec les deux méthodes. Pour cela, nous avons mis au point une voyelle artificielle (dont les coefficients des paramètres articulatoires modélisent au mieux une forme neutre de bouche) et nous ajoutons cette voyelle en début et fin de séquence avant de réaliser la synthèse. Ainsi, la tête parlante synthétisée commence et finit chaque mot avec les lèvres et la mâchoire en position neutre. Comme le montre la figure 8.3, nous avons essayé de définir une position neutre qui ressemble le plus à celle adoptée par notre locutrice et nous avons orienté la tête parlante comme le visage réel afin que les deux visages soient perçus de la même manière par les sujets.

Malheureusement, la création du phonème neutre est plus complexe à mettre en œuvre avec la technique de synthèse de Cohen et Massaro car cela exige un réglage manuel des fonctions de dominance qui n'a pas encore été réalisé à l'heure actuelle. Une fois cette adaptation faite, les tests pourront être réalisés.

FIG. 8.3 – visage réel et virtuel en position initiale ou finale (en début ou fin de mot).

Conclusion et perspectives

Cette thèse, qui s'inscrit dans une étude plus vaste sur la modélisation d'une tête parlante destinée à améliorer l'intelligibilité du message transmis, porte essentiellement sur la caractérisation et la modélisation du phénomène de coarticulation labiale. L'état de l'art a permis de constater que de nombreuses théories rentrent en concurrence au sujet de l'explication de l'origine de l'anticipation alors qu'un certain consensus semble régner au sujet de l'effet rétentif considéré comme essentiellement dû à des effets inertiels. C'est pour cela que nous nous sommes concentré principalement sur la caractérisation de l'anticipation. Entre des théories basées essentiellement sur des critères phonétiques et phonologiques et des théories fondées uniquement sur des critères articulatoires, le modèle expansionniste de Abry et Lallouache [2] qui considère que la coarticulation est un mouvement fortement expansible et relativement peu compressible nous a semblé un bon compromis et n'a pas été démenti par nos mesures expérimentales sur un corpus multilocuteur. De même, la solution proposée par Keating [44], qui consiste à quantifier par une fenêtre le degré de variation d'un paramètre articulatoire donné, nous a semblé intéressante pour prendre en compte le coefficient de résistance à la coarticulation, mais ce modèle est imprécis car il ne définit pas la manière dont le mouvement articulatoire doit évoluer d'une fenêtre à l'autre.

Parmi toutes les solutions expérimentales modélisant la coarticulation, les méthodes à base de règles ont retenu notre attention. En effet, contrairement aux solutions statistiques comme les ANN(Artificial Neural Networks) ou les HMM(Hidden Markov Model), elles sont capables de modéliser toute l'étendue du phénomène coarticulatoire et offrent la possibilité d'un contrôle local et précis des mouvements. Le modèle de Löfqvist [50] ou sa version adaptée par Cohen et Massaro [19] ne nous ont pas non plus complètement convaincu. En effet, la quantification par des fonctions de dominance de l'influence des différents phonèmes nous a semblé intéressante, mais aussi trop globale pour permettre le contrôle précis du mouvement imposé ponctuellement par certains phonèmes (par exemple la fermeture complète des lèvres pour les bilabiales). En outre, ces modèles de type time-locked ne peuvent pas complètement prendre en compte l'effet syntagmatique, c'est à dire le respect de contrastes distinctifs entre les sons.

Par rapport aux méthodes à base de règles existantes, notre approche permet de réaliser une synthèse audiovisuelle avec un corpus audiovisuel vaste (pour l'audiovisuel, mais petit si on le rapporte à la taille des corpus acoustiques pour la synthèse à partir du texte) et de modélisation de la coarticulation très compacte. En effet, l'algorithme de prédiction a permis de sélectionner les phonèmes pertinents de chaque geste. De plus, le fait d'avoir modélisé le mouvement par des sigmoïdes diminue énormément le nombre de données à stocker tout en favorisant l'adaptation à des vitesses d'élocution différentes et des mouvements articulatoires plus ou moins amples. L'acquisition d'un nouveau corpus encore plus grand, constitué avec une attention toute particulière sur l'enregistrement des voyelles isolées, permettra d'étoffer la base d'apprentissage et de mieux réaliser les complétions. Il nous semble aussi important d'estimer avec davantage de précision

les adaptions temporelles à réaliser sur les sigmoïdes. Actuellement, seule une modification de la position de la sigmoïde est réalisée pour s'adapter à la durée des phonèmes. Nous ne modifions pas le paramètre qui contrôle la vitesse de transition car aucune information du corpus actuel ne permet de le quantifier. Si, dans le cadre d'un nouveau corpus, le locuteur prononce plusieurs fois et plus ou moins rapidement certaines séquences, il deviendra possible d'estimer l'influence de la vitesse d'élocution sur ce paramètre.

Une critique qui pourrait être formulée à l'encontre de notre modèle est l'absence de prise en compte de la coarticulation rétentive. S'il est vrai que notre algorithme de prédiction de la coarticulation ne tient compte que de l'anticipation, la phase de lissage que nous effectuons à la fin de la synthèse permet de simuler partiellement l'effet inertiel. Des études ultérieures nous semblent néanmoins utiles pour quantifier avec davantage de précision la coarticulation progressive. Par exemple, les organes plus massifs ont une inertie plus grande que les organes plus légers. En conséquence, il pourrait être intéressant, lors de la synthèse, d'affiner l'attribution des poids de la spline de lissage afin d'accorder davantage d'importance aux effets inertiels de la mâchoire (organe plus lourd que les lèvres).

Avoir choisi seulement quatre paramètres articulatoires pour contrôler la tête parlante peut aussi constituer une critique de notre modèle, mais plusieurs éléments plaident en notre faveur. Tout d'abord, le choix des paramètres de contrôle provient d'une analyse en composantes principales qui montre que la protrusion par exemple explique 66.6% de la variance (sur le grand corpus). D'autre part, certaines études sur la perception [66] ont montré que l'essentiel de l'intelligibilité du visage se concentre au niveau des lèvres. Néanmoins, nos mesures ne prennent pas en compte toute la complexité du mouvement articulatoire. Par exemple, la quantification de la protrusion est en fait beaucoup plus complexe que la simple mesure de l'avancée des lèvres. L'arrondissement est associé à un mouvement de "dépliement" qui augmente la surface visible des lèvres et contribue à des déformations complexes au voisinage des lèvres. Seuls quelques marqueurs peints sur le visage nous ayant servi de référence pour calculer la protrusion, ce phénomène n'est pas pris en compte. En ce qui concerne la mesure de l'ouverture verticale des lèvres, nous l'estimons par la distance entre un point sur la lèvre supérieure et un point sur la lèvre inférieure. Ceci n'est donc aussi qu'une approximation et des travaux récents de notre équipe qui seront publiés prochainement ont corrigé ce défaut. Une solution pour corriger les problèmes évoqués (notamment le phénomène de dépliement des lèvres) serait d'utiliser le maillage complet avec toute la difficulté que cela comporte.

Une autre piste d'amélioration concerne la prise en compte de la prosodie. Pour limiter la taille des données et privilégier la variété des séquences, nous avons délibérément choisi de ne pas considérer ce phénomène lors de l'enregistrement du corpus. Désormais, il serait intéressant d'étudier comment réaliser l'adaptation de l'algorithme de prédiction et de la synthèse pour en tenir compte. Le fait que nous contrôlions complètement l'axe syntagmatique devrait permettre de gérer en partie les variations d'amplitudes générées par les effets prosodiques, mais nous devrons enregistrer un nouveau corpus pour valider cette théorie.

Enfin, même si la visualisation des mouvements labiaux constitue l'essentiel du gain d'intelligibilité, il serait intéressant de modéliser d'autres mouvements ; ceux de la langue par exemple pourraient être utiles pour l'apprentissage des langues étrangères ; les mouvements des yeux, des sourcils pourraient être couplés aux effets prosodiques etc. De la même manière que l'équipe du "Perceptual Science Laboratory (Université de Californie - Santa Cruz)" qui a développé la tête

parlante "Baldi", nous envisageons de progresser par étapes successives afin de gérer séparément les problèmes. L'étape ultime serait de développer un visage parlant parfaitement réaliste et adapté au locuteur.

Annexe A

Correspondances entre API (Alphabet Phonétique International) et SAMPA(Speech Assessment Methods Phonetic Alphabet)

API	SAMPA	Exemple
i	i	l**i**t
e	e	pr**é**
ɛ	E	s**ei**ze
a	a	p**a**tte
ɑ	A	p**â**te
y	y	d**u**
ø	2	d**eu**x
œ	9	n**eu**f
ə	@	just**e**ment
u	u	f**ou**
o	o	p**o**t
ɔ	O	p**o**rt
ɛ̃	U~	pép**in**
ɑ̃	a~	v**en**t
õ	o~	b**on**
w	w	q**u**oi [kwa]
ɥ	H	n**u**it [nHi]
j	j	**y**o**y**o
p	p	**p**ont
b	b	**b**on
m	m	**m**on
t	t	**t**emps
d	d	**d**ent
n	n	**n**om
k	k	**c**ar
g	g	**g**are
ŋ	N	a**gn**eau
f	f	**f**eu
v	v	**v**ie
s	s	**s**oir
z	z	ro**s**e
l	l	**l**ong
ʃ	S	mou**ch**e
ʒ	Z	**j**oie
ʁ ou R	R	**r**ond
hh		Aspiration
#		Silence

Annexe B

Constitution du mini corpus

Ce mini corpus a été enregistré par 10 locuteurs francophones. Mis à part les 4 voyelles /i/, /y/, /a/ et /o/ et les consonnes /p/, /t/, /d/, /S/, /s/ ,/f/, les CV, VCV, VCCV ont été enregistrées avec des mots "porteurs" en début et fin pour forcer le locuteur à se rapprocher d'une forme neutre de bouche. Les mots choisis étaient "trois" [tRwa] en début de phrase et "lave" [lav] en fin de phrase. Le principe d'utiliser des "mots porteurs" n'a pas donné entière satisfaction en ce qui concerne les formes de bouches initiales et finales et n'a pas été reconduit pour l'enregistrement du grand corpus.

Voici ci-dessous les séquences constituant ce corpus (Alphabet IPA)

/i/	/y/	/a/	/o/	/pø/	/tø/	/dø/	/sø/	/ʃø/	/fø/
/sy/	/ʃy/	/ty/	/dy/	/si/	/ʃi/	/ti/	/di/	/isy/	/iʃy/
/ity/	/ipy/	/idy/	/ysi/	/yʃi/	/yti/	/ypi/	/ydi/	/ify/	/ifi/
/yfy/	/yfi/	/yvi/	/ivy/	/ivi/	/yvy/	/osy/	/oʃy/	/ipsy/	/ipʃy/
/ispy/	/iʃpy/	/itʃy/	/ikʃy/	/ypʃi/	/ypsi/	/yʃpi/	/yspi/	/ytʃi/	/ykʃi/
/ospy/	/opsy/	/oʃpy/	/opʃy/	/osty/	/otsy/				

Les deux phrases suivantes ont été aussi incluses au corpus :
– /ilsøgarantiRadyfRwaavɛksøbonkapyʃon/ (Il se garantira du froid avec ce bon capuchon).
– /løʒohajeabRwajelekajudølavwajaʒøz/ (Le joaillier a broyé les cailloux de la voyageuse).

Annexe C

Constitution du grand corpus

Le grand corpus a été enregistré par une locutrice spécialiste de la lecture labiale dont la langue maternelle est le français. Chaque séquence n'a été enregistrée qu'une seule fois afin de limiter la durée d'enregistrement et la taille de l'espace de stockage.

Le corpus a été présenté écrit en français car la locutrice n'est pas une spécialiste de la lecture phonétique. Les éléments en italique sont simplement donnés à titre d'information à la locutrice et ne sont pas prononcés.

Les voyelles isolées

- **a** *comme dans patte*
- **i** *comme dans pie*
- **ou** *comme dans pou*
- **u** *comme dans pu*
- **e** *comme dans peur*
- **è** *comme dans père*
- **é** *comme dans pépé*
- **o** *comme dans peau*
- **ain** *comme dans patin*
- **an** *comme dans pan*
- **on** *comme dans pont*
- **e** *comme dans peu*

Les consonnes

pe	te	ke	fe	se	che	le	be	re	me
de	ne	gue	ve	je	ze				

Les VV

i-a	ou-a	ou-i	u-a	u-i	u-ou	e-a	e-i	e-ou	e-u
è-a	è-i	è-ou	è-u	è-e					

Les CV

pa	pi	pou	pu	pe	pè	ta	ti	tou	tu
te	tè	ka	ki	kou	ku	ke	kè	fa	fi
fou	fu	fe	fè	sa	si	sou	su	se	sè
cha	chi	chou	chu	che	chè	la	li	lou	lu
le	lè	ra	ri	rou	ru	re	rè	wa	wi
wè	ya	yi	you	yu	ye	yè	oua	oui	ba
bi	bou	bu	be	bè	ma	mi	mou	mu	me
mè	da	di	dou	du	de	dè	na	ni	nou
nu	ne	nè	ga	gui	gou	gu	gue	guè	va
vi	vou	vu	ve	vè	ja	ji	jou	ju	je
jè	za	zi	zou	zu	ze	zè			

Les VCV

apa	api	apou	apu	ata	ati	atou	atu	aka	aki
akou	aku	afa	afi	afou	afu	asa	asi	asou	asu
acha	achi	achou	achu	ala	ali	alou	alu	ara	ari
arou	aru	ipa	ipi	ipou	ipu	ita	iti	itou	itu
ika	iki	ikou	iku	ifa	ifi	ifou	ifu	isa	isi
isou	isu	icha	ichi	ichou	ichu	ila	ili	ilou	ilu
ira	iri	irou	iru	oupa	oupi	oupou	oupu	outa	outi
outou	outu	ouka	ouki	oukou	ouku	oufa	oufi	oufou	oufu
ousa	ousi	ousou	ousu	oucha	ouchi	ouchou	ouchu	oula	ouli
oulou	oulu	oura	ouri	ourou	ouru	upa	upi	upou	upu
uta	uti	utou	utu	uka	uki	ukou	uku	ufa	ufi
ufou	ufu	usa	usi	usou	usu	ucha	uchi	uchou	uchu
ula	uli	ulou	ulu	ura	uri	urou	uru	aya	ayi
ayou	ayu	iya	iyi	iyou	iyu	ouya	ouyi	ouyou	ouyu
uya	uyi	uyou	uy	aba	abi	abou	abu	ama	ami
amou	amu	ada	adi	adou	adu	ana	ani	anou	anu
aga	agui	agou	agu	ava	avi	avou	avu	aja	aji
ajou	aju	aza	azi	azou	azu	iba	ibi	ibou	ibu
ima	imi	imou	imu	ida	idi	idou	idu	ina	ini
inou	inu	iga	igui	igou	igu	iva	ivi	ivou	ivu
ija	iji	ijou	iju	iza	izi	izou	izu	ouba	oubi
oubou	oubu	ouma	oumi	oumou	oumu	ouda	oudi	oudou	oudu
ouna	ouni	ounou	ounu	ouga	ougui	ougou	ougu	ouva	ouvi
ouvou	ouvu	ouja	ouji	oujou	ouju	ouza	ouzi	ouzou	ouzu
uba	ubi	ubou	ubu	uma	umi	umou	umu	uda	udi
udou	udu	una	uni	unou	unu	uga	ugui	ugou	ugu
uva	uvi	uvou	uvu	uja	uji	ujou	uju	uza	uzi
uzou	uzu								

Les VCCV

aktu	uska	usti	uktu	iskou	ouvri	artou	isku	atrou	arkou
ardu	urzi	urba	oursi	ouvra	asky	ourpa	uski	arpou	arsu
ispu	ourdi	ourtou	usta	argu	azdu	ourki	upri	abru	oursa
urdu	outra	ukti	urvi	usku	azdu	urki	ursa	ipsu	ipchu
ispu	ichpu	itchu	ikchu	upchi	upsi	uchpi	uspi	utchi	ukchi
apsou	aspou	oupsa	ouspa	atrou	artou	outra	ourta	akrou	oukra
arkou	ourka	apli	ipla	alpi	ilpa	apri	ipra	arpi	irpa
ospu	opsu	ochpu	opchu	ostu	otsu	arka	utra	isti	atra
ikti	arta	agra	asta	aspa	iski	afra	aska	atri	itra
arti	irta								

Des mots utiles si on veut tester l'effet McGurk

- baba
- dada
- vava
- gaga

Les phrases

Les phrases enregistrées sont des phrases phonétiquement équilibrées choisies par P Combescure [20].

1. Leur chienne a hurlé toute la nuit.
2. Annie s'ennuie loin de ses parents.
3. Les deux camions se sont heurtés de face.
4. Un loup s'est jeté immédiatement sur la petite chèvre.
5. Dès que le tambour bat, les gens accourent.
6. Vous poussez des cris de colère.
7. Ce petit canard apprend à nager.
8. La voiture s'est arrêtée au feu rouge.
9. Souvent, je m'accoude au muret de ce pont.
10. Mon père m'a donné l'autorisation.
11. Pour se protéger, il s'est couché près de ma porte.
12. La vaisselle propre est mise sur l'évier.
13. Sa voisine est inimitable.
14. Le renard se hâte vers son gite.
15. Le bouillon fume dans les assiettes.
16. Le caractère de cette femme est moins calme.
17. Le camp d'été s'est passé au bord du fleuve.
18. Un train entre déja en gare.

19. A l'ouest, mes pommiers donnent peu.
20. Lentement des canes se dirigent vers la mare.
21. Une goélette déploie ses voiles.
22. Le facteur va porter le courrier.
23. Bien sûr, je connais son nom.
24. Maman prend un verre et une assiette.
25. Désormais, je me tournerai quand il partira.
26. Les avions tournent au dessus de la place.
27. Mettez la faux, ici sous ma tente.
28. Je suis resté sourd à ses cris.
29. Le chameau est loin de son abri.
30. Je pense être de retour ici avant la nuit.
31. Des chiens nous montraient leurs crocs pointus.
32. La jeune fille se peigne devant la glace.
33. Il a été condamné pour un vol de voiture.
34. Je ne veux pas que vous changiez pour le moment.
35. Nous avons pris froid en jouant au tennis
36. Il est désormais accablé par le travail.
37. Ce bonbon contenait trop de sucre.
38. A la hâte, le métayer ansilait ses récoltes avant l'hiver
39. Une brume épaisse s'est formée sur la mer
40. Le menuisier a scié une planche et l'a rabotée.
41. Maman a préparé une galette pour jeudi.
42. Le foot-ball, voila ce qui l'intéresse.
43. C'est un charmant spectacle, je t'assure.
44. Ils m'ont apporté des friandises à mon anniversaire.
45. Ces élèves prendront l'autocar tout à l'heure.
46. Parfois, mon épicière vend à crédit.
47. Personne n'a applaudi ce beau discours.
48. Je me demande pourquoi on court sans cesse.
49. Il se repend de ce qu'il vient de faire.
50. Des gens se sont levés dans les tribunes.
51. Vous éplucherez les légumes du pot au feu.
52. Ce chasseur projette encore de partir d'ici, ce matin.
53. La poire est un fruit à pépin.
54. Plus nous le connaissons, plus nous le respectons.
55. Là-haut, monte la voix du pâtre qui ramène ses moutons.
56. Le courrier arrive en retard en ce moment.

57. Cette cage contient mon oiseau.
58. Des lièvres jouent à l'orée du bois.
59. Je te dis que ma bouteille s'abime à la cave.
60. Il s'est réfugié dans ma chambre.
61. Le troupeau s'abreuvait au ruisseau.
62. Le client s'attend à ce que vous fassiez une réduction.
63. Chaque fois que je me lève, ma plaie me tire.
64. Une rançon est exigée par les ravisseurs.
65. Ainsi cette comédie est en un acte.
66. Papa aime mon vin quand il est bon.
67. Le ciel est tout noir, il va tomber des cordes.
68. On dit que l'essor de ce village est important.
69. Ce soir, nous nous coucherons plus tard.

Annexe D

Constitution du corpus de perception

Ce corpus est constitué d'un ensemble de mots. Nous en avons extrait 51 des listes de Lafon [48] et nous en avons choisi 29 autres afin d'évaluer certaines combinaisons VCCV spécifiques. 34 autres extraits aussi des listes de Lafon ont été enregistrés afin de servir de test d'entraînement.

Le corpus

Mots extraits des listes de Lafon.

bouée	rôde	fente	tige	grain	cave	bulle	somme	maine	preux
bord	souille	site	sauve	oser	chance	gagne	bouche	rôde	feinte
jute	ligne	cure	moule	sème	anis	pour	gris	somme	sente
sève	rase	poche	agneau	buse	code	foule	fange	gaule	cran
allait	monte	nasse	près	brou	sac	tord	rave	bise	sèche
vigne									

Mots de notre choix (pour tester les VCCV)

actuel	brusqua	fustiger	discours	ouvrier	cartouche	disculpe
attroupe	ardu	charpie	casse cou	pourparlers	harpon	dispute
assourdi	supprime	abrupt	outrage	fructifie	survie	sapristi
fluctue	patrouille	parcours	courba	court-circuit	urticant	applique
extirpa						

Le corpus d'entraînement

buée	ride	foc	agis	vague	croc	lobe	mieux	natte	col
fort	soupe	tonte	vèle	nage	souche	rogne	bile	dore	fil
sage	gaine	cru	boule	mule	bonne	cale	rive	sol	tempe
fauve	phase	chatte	règne						

Bibliographie

[1] C. Abry, J.L.Boë, and R. Descout. Voyelles arrondies et voyelles protruses en français. *Labialité et Phonétique. Publications de l'Université des Langues et Lettres de Grenoble*, pages 203–215, 1980.

[2] C. Abry and T. Lallouache. Le MEM : un modèle d'anticipation paramétrable par locuteur : Données sur l'arrondissement en français. *Bulletin de la communication parlée*, 3(4) :85–89, 1995.

[3] A. Bell. Visible speech, universal alphabetics or self-interpreting physiological letters for the writing of all languages in one alphabet. London : Simpkin and Marschall, 1867.

[4] F. Bell-Berti. Velopharyngeal function : A spatio-temporal model. *Speech and Language : Advances in Basic Research and Practice - New York : Academic Press.*, pages 291–316, 1980.

[5] F. Bell-Berti and K. S. Harris. A temporal model of speech production. *Phonetica*, 38 :9–20, 1981.

[6] F. Bell-Berti and R. A. Krakow. Anticipatory velar lowering : a co-production account. *Journal of the Acoustical Society of America*, 90 :112–123, 1991.

[7] A. P. Benguerel and H. A. Cowan. Coarticulation of upper lip protrusion in french. *Phonetica*, 30 :41–55, 1974.

[8] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry. A set of french visemes for visual speech synthesis. *Les Cahiers de l'Institut de la communication parlée*, 1997.

[9] J. Beskow. Rule-based visual speech synthesis. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 299–302, Madrid, Spain, 1995.

[10] J. Beskow. Trainable articulatory control models for visual speech synthesis. *International Journal Of Speech Technology 7*, pages 335–349, 2004.

[11] C.S Blackburn and S Young. A self-learning predictive model of articulator movements during speech production. *Journal of the Acoustical Society of America*, 107 :1659–1670, 2000.

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

[13] S. Boyce. Coarticulatory organization for lip rounding in turkish and english. *JASA*, 88 :2584–2595, 1990.

[14] E. Brücke. Grundzüge der physiologie und systematik der sprachlaute für linguisten und taubstummenlehrer. Vienna : Gerold, 1856.

[15] A. P. Breen, E. Bowers, and W. Welsh. An investigation into the generation of mouth shapes for a talking head. In *Proc. ICSLP '96*, volume 4, pages 2159–2162, Philadelphia, PA, 1996.

[16] C. Bregler, M. Covell, and M. Slaney. Video rewrite : Driving visual speech with audio. In *Proceedings of ACM SIGGRAPH'97*, pages 353–360, 1997.

[17] C. P. BROWMAN and L. GOLDSTEIN. Articulatory gestures as phonological units. *Phonology, 6*, pages 201–251, 1989.

[18] C. P. Browman and L. Goldstein. Dynamics and aticulatory phonology. Haskins Laboratories Status Reports on Speech Research 113 : 51-62, 1993.

[19] M. Cohen and D. Massaro. Modeling coarticulation in synthetic visual speech, 1993.

[20] P. Combescure. 20 listes de dix phrases phonétiquement équilibrées. *Revue d'acoustique 56*, 1981.

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

[22] P. Cosi and al. Labial coarticulation modeling for realistic facial animation. In *Proceedings of ICMI'02, 4th International Conference on Mutlimodal Interfaces*, pages 505–510, Pittsburgh, PA, USA, 2002.

[23] D. Cosker, S. Paddock, D. Marshall, and P. L. Rosin. Towards perceptually realistic talking heads : models, methods and mcgurk. In *In APGV 2004*, pages 151–157, 2004.

[24] R. G. Daniloff and R. E. Hammarberg. On defining coarticulation. *Journal of Phonetics*, 1 :239–48, 1973.

[25] D. Dekle, C. Fowler, and M. Funnel. Audio-visual integration in perception of real words. *Perception and Psychophysics*, 51, 4 :355–362, 1992.

[26] P. Delattre, A.M. Liberman, and F.S. Cooper. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27 :769–774, 1955.

[27] F. Elisei, M. Odisio, G. Bailly, and P. Badin. Creating and controlling video-realistic talking heads. In *in Audio Visual Speech Processin Workshop*, pages 90–97, Aalborg, Danemark, 2001.

[28] O. Engstrand. Acoustic constraints or invariant input representation ? an experimental study of selected articulatory movements and targets. *Reports from Uppsala University Department of Linguistic*, 7 :67–95, 1981.

[29] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of ACM SIGGRAPH 2002*, pages 388–398, Sans Antonio, TX, 2002.

[30] Tony Ezzat and Tomaso Poggio. Miketalk : A talking facial display based on morphing visemes. In *CA*, pages 96–102, 1998.

[31] G. Fant. Stops in cv syllables. In *Speech Sounds and Features*, pages 110–139. MIT Press, Cambridge, MA, 1973.

[32] E. Farnetani and D. Recasens. *Coarticulation. Theory, Data and Techniques*, chapter Coarticulation models in recent speech production theories, pages 31–65. Cambridge University Press, 1999.

[33] C. A. Fowler. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8 :1113–133, 1980.

[34] C. Gelfer, F. Bell-Berti, and K. Harris. Determining the extent of coarticulation : Effects of experimental design. *Journal of the Acoustical Society of America*, 86 :2443–2445, 1989.

[35] B. Le Goff. Automatic modeling of coarticulation in text-to-visual speech synhtesis. In *Eurospeech-1997*, pages 1667–1670, 1997.

[36] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw. Evaluation de systèmes de génération de mouvements faciaux. In *Proceedings of JEP 2006 (Journées d'Etude sur la Parole)*, Dinard, FRANCE, 2006.

[37] F. H. Guenther. Speech sound acquisition, coarticulation and rate effects in a neural network model os speech production. *Psychological Review*, 102, No.3 :594–621, 1995.

[38] C. Hack and C. J. Taylor. Modelling talking head behavior. In *Proceeedings of British Machine Vision Conference*, 2003.

[39] W. J. Hardcastle and N. Hewlett. *Coarticulation. Theory, Data and Techniques.* Cambridge University Press, 1999.

[40] W. L. Henke. *Dynamic articulatory model of speech production using computer simulation , Unpublished doctoral dissertation.* PhD thesis, MIT Cambridge, 1966.

[41] S. Öhman. Coarticulatio in vcv utterances : spectrographic measurements. *Journal of the Acoustical Society of America*, 39 :151–168, 1966.

[42] S. Öhman. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41 :310–320, 1967.

[43] D. Jones. Outline of english phonetics. Cambridge University Press (3rd edn), 1932.

[44] P. Keating. The window model of coarticulation : articulatory evidence. In J. Kingston and M. Beckman, editors, *Papers in Laboratory Phonology I*, pages 451–470. Cambridge University Press, 1990.

[45] D. H. Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, 82 :737–793, 1987.

[46] V. Kozhevnikov and L. Chistovich. Speech; articulation and perception. In *DC : Joint Publications Research Service*, volume Translation 30,543, pages 115–135, Washington, 1965.

[47] D. Krull. Second formant locus patterns as a measure of consonant-vowel coarticulation. In *Phonetic Experimental Research at the Institute of Linguistics*, pages 43–61, University of Stockholm, Suède, 1987.

[48] J. C. Lafon. Le test phonétique. *Bulletin d'Audiophonologie - Faculté de Médecine et de Pharmacie de Besançon - Université de Franche-Comté*, 9, 1958.

[49] L.F. Lamel, J.-L. Gauvain, and M. Eskénazi. BREF, a Large Vocabulary Spoken Corpus for French. In *Proceedings of European Conference on Speech Technology*, pages 505–508, Genova, Italy, September, 1991.

[50] A. Löfqvist. *Speech as audible gestures.* Hardcastle, W.J. and Marchal, A. (eds). Dordrecht : Kluwer Academic Publishers, 1990.

[51] A. Löfqvist and H. Yoshioka. Laryngeal activity in icelandic obstruent production. *Nordic Journal of Linguistics*, 4, 1981.

[52] B. Lindblom. On vowel reduction, report no 29. The Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, 1963.

[53] B. Lindblom. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.*, 35(11) :1773–1781, November 1963.

[54] B. Lindblom. *Speech Production and Speech Modelling*, chapter Explaining Phonetic Variation. A sketch of the H&H Theory, pages 403–439. Kluwer Academic Publishers, w.j. hardcastle and a. marchal(eds) edition, 1990.

[55] J. F. Lubker. Temporal aspects of speech production : anticipatory labial coarticulation. *Phonetica*, 38 :51–55, 1981.

[56] J. C. Lucero and K. G. Munhall. A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America*, 106(5), 1999.

[57] A. Marchal. *Les sons et la parole*. Guérin. Montreal, 1980.

[58] D. Massaro. *Perceiving Talking Faces. From Speech Perception to a Behavioral Principle*. Cognitive Psychology series. A Bradford Book., 1998.

[59] R. McAllister. Temporal asymmetry in labial co-articulation. *Papers from the Institute of Linguistics, University of Stockholm*, 35 :1–29, 1938.

[60] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, pages 746–748, 1976.

[61] P. Menzerath and A. De Lacerda. Koarticulation, steuerung und lautabgrenzung. *Berlin and Bonn : Fred. Dummlers*, 1933.

[62] S. Minnis and A. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *Proc. ICSLP '2000*, pages 759–762, Beijing, China, 2000.

[63] K. Moll and R. Daniloff. Investigation of the timing of velar movements during speech. *Journal of the Acoustical Society of America*, 50 :678–684, 1971.

[64] P. D. Mac Neilage. *Speech and Cortical Functionning*. J. H. Gilbert. New York : Academic, 1972.

[65] J. Ostermann. Animation of synthetic faces in mpeg-4. *Computer Animation*, pages 49–51, 1998.

[66] S. Ouni, M. M. Cohen, H. Ishak, and D. W. Massaro. Visual contribution to speech perception : Measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech and Music Processing*, 2007, 2007.

[67] F. I. Parke. Parametrized models for facial animation. *IEEE Computer Graphics, 2(9)*, pages 61–68, 1982.

[68] H. Paul. Prinzipien der sprachgeschichte (3rd edn). Halle : Niemeyer, 1898.

[69] C. Pelachaud, N.I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive science, 20(1)*, pages 1–46, 1996.

[70] C. Pelachaud, E.Magno-Caldognetto, C. Zmarich, and P. Cosi. An approach to an italian talking head. In *InterSpeech*, pages 1035–1038, Aalborg, Danemark, 2001.

[71] J.S. Perkell and C.M. Chiang. Preminінary support for a 'hybrid model' of anticipatory coarticulation. In *Proceedings of the XIIth International Congress of Acoustics*, Toronto : Canadian Acoustical Association, A3-6, 1986.

[72] J.S. Perkell and M.L. Matthies. Temporal measures of anticipatory labial coarticulation for the vowel /u/ : Within-and cross-subject variability. *Journal of the Acoustical Society of America*, 1992.

[73] S. M. Platt and N. I. Badler. Animating facial expression. *Computer Graphics, 15(3)*, pages 245–252, 1981.

[74] L. Reveret, G. Bailly, and P. Badin. Mother : A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *Proceedings of the 6th International Conference on spoken Language Processing*, pages 755–788, Bejing , China, 2000.

[75] P. J. Rousselot. Principe de phonétique expérimentale, i-ii. Paris : H. Welter, 1897.

[76] K. C. Scott, D. S. Kagels, S. H. Watson, H. Rem, J. R. Wright, M. Lee, and K. J. Hussey. Synthesis of speaker facial movement to match selected speech sequences. *Speech Science and Technology*, 1994.

[77] E. Scripture. The elements of experimental phonetics. New York : Charles Scribner's Sons, 1902.

[78] E. Sievers. Grundzuge der lautphysiologie zur einfuhrung in das studium der lautlehre der indogermanischen sprachen. Leipzig : Breitkopf and Hartel, 1876.

[79] A.D. Simons and S.J. Cox. Generation of mouthshapes for a synthetic talking head. In *Proceedings of the Institute of Acoustics*, pages 475–482, 1990.

[80] R. Stetson. Motor phonetics : a study of speech movements in action (2nd edn). Amsterdam : North Holland, 1951.

[81] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2) :212–215, 1954.

[82] H. M. Sussman and J. Westbury. The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research*, 24 :16–24, 1981.

[83] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda. Visual speech synthesis based on parameter generation from hmm : Speech-driven and text-and-speech-driven approaches, 1998.

[84] K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from hmm using dynamic features. In *ICASSP 95*, pages 660–663, Detroit, MI, USA, 1995.

[85] W. Wickelgren. Context-sensitive coding, associative memory, and serial order in (speech) behaviour. *Physiological Review*, 76 :1–15, 1969.

[86] W. Wickelgren. *Discussion paper on speech perception*, volume Speech and Cortical Functionning, pages 237–262. NY : Academic Press, New York, 1972.

[87] B. Wrobel-Dautcourt, M.O. Berger, B. Potard, Y. Laprie, and S. Ouni. A low-cost stereovision based system for acquisition of visible articulatory data. In *AVSP 2005*, 2005.

[88] E. Yamamoto, S. Nakamura, and K. Shikano. Speech to lip movement synthesis by hmm. In *AVSP 1997*, pages 137–140, 1997.

[89] E. Yamamoto, S. Nakamura, and K. Shikano. Lip movement synthesis from speech based on hidden markov models. *Speech Communication*, 26 :105–115, 1998.

[90] J.L. Zhou, F. Seide, and L. Deng. Coarticulation modeling by embedding a target-directed hidden trajectory model into hmm - model and training. In *ICASSP 2003*, pages 744–747, 2003.

Publications personnelles

[1] V. Robert, A. Bonneau, B. Wrobel-Dautcourt, and Y. Laprie. Prédiction phonétique de la coarticulation labiale. In Béatrice Vaxelaire, Rudolph Sock, Georges Kleiber, and Fabrice Marsac, editors, *Perturbations et réajustements : langue et langage*, pages 155–167. Publications de l'Université Marc Bloch - Strasbourg 2, 2007.

[2] V. Robert, J. Feldmar, and Y. Laprie. Comparaison between two predicting methods of labial coarticulation. In *International Seminar of Speech Production (ISSP08)*, 2008.

[3] V. Robert, Y. Laprie, and A. Bonneau. A phonetic concatenative approach of labial coarticulation. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1402–1405, 2007.

[4] V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau. Inter speaker variability of labial coarticulation with the view of developing a formal coarticulation model for french. In *Proceedings of Auditory-Visual Speech Processing (AVSP)*, pages 65–70, 2005.

[5] V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau. Strategies of labial coarticulation. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1021–1024, 2005.

Résumé

Cette thèse s'inscrit dans une étude sur l'élaboration d'une tête parlante. Nous nous intéressons tout particulièrement à la prédiction du mouvement de coarticulation des lèvres et de la mâchoire. Après avoir analysé les variations intra et interlocuteur des paramètres labiaux de deux corpora audiovisuels, nous avons conçu un algorithme de prédiction de la coarticulation basé sur des règles phonétiques et prenant en considération l'interaction entre les articulateurs. Nous avons ensuite proposé une technique pour synthétiser les mouvements articulatoires des lèvres et de la mâchoire en utilisant un corpus monolocuteur. Le principe de base est la concaténation de séquences élémentaires de type VC...CV qui ont été jugées pertinentes par notre algorithme de prédiction phonétique, et qui sont soit extraites du corpus, soit obtenues par complétion. Nous avons modélisé les mouvements articulatoires par des sigmoïdes qui offrent l'avantage de réduire considérablement la taille du modèle construit et permettent de s'adapter facilement à des vitesses d'élocution ou des stratégies articulatoires particulières tout en conservant les contrastes distinctifs entre les sons successifs et leurs caractéristiques intrinsèques. Afin d'estimer la qualité de notre synthèse, nous avons mesuré les différences entre les signaux réels et synthétisés sur l'ensemble des phrases du corpus et nous avons comparé notre solution avec l'algorithme de Cohen et Massaro. Nous avons montré que notre synthèse est meilleure pour certaines séquences spécifiques de type VCCV où l'anticipation est plus complexe.

Mots-clés: coarticulation, tête parlante, anticipation, synthèse audiovisuelle.

Abstract

This thesis comes within the scope of talking heads. We are particularly interested in the prediction of labial and jaw coarticulation movements. After analyzing intra and inter speaker variability using two corpora, we defined a prediction algorithm for anticipatory coarticulation based on phonetic rules which takes into account interactions between articulators. We then proposed a solution to estimate labial and jaw movements using a one speaker corpus. It consists in concatenating elementary VC...CV sequences selected by our prediction algorithm and either extracted from the corpus or rebuilt by completion. We modeled articulatory movements using sigmoids which offer the advantage of considerably reducing the model size and which are adaptable to speaking rate or articulatory strategies. Additionally, sigmoids are able to keep distinctive contrasts between neighboring segments as well as intrinsic characteristics of the sounds. With the aim of estimating the quality of our synthesis process, we measured differences between real and predicted data for all the sentences of the corpus et we compared our solution with Cohen and Massaro 's algorithm. It turns out that our solution is better for specific VCCV sequences in which anticipation is more complex.

Keywords: coarticulation, talking head, anticipation, audiovisual synthesis.