Modèles de génération de trajectoires pour l'animation de visages parlants

Oxana Govokhina

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INP Grenoble

Spécialité : Signal, Image, Parole, Télécoms

préparée au : Département de Parole et Cognition du laboratoire GIPSA-Lab

dans le cadre de l'Ecole Doctorale : Electronique, Electrotechnique, Automatique, Traitement du Signal

présentée et soutenue publiquement le 24 octobre 2008
par Oxana Govokhina

TITRE

Modèles de génération de trajectoires pour l'animation de visages parlants

DIRECTEUR DE THESE
Gérard Bailly
CO-DIRECTEUR DE THESE
Gaspard Breton

JURY

Président
Bernard Péroche

Rapporteurs
Sylvie Gibet
Christophe d'Alessandro

Examinateurs
Gérard Bailly
Gaspard Breton

Remerciements

Je remercie tous mes collègues du département Parole et Cognition de l'accueil et plus particulièrement Frédéric Elisei, Antoine Bégault et Christophe Savariaux pour leur aide.

Ce travail a été effectué par le cadre d'une bourse de doctorat financée France Télécom R&D et je remercie les membres de l'équipe IAM pour les conditions de travail dont j'ai bénéficié.

Je remercie Agnès Afriat - pour sa voix d'or et son professionnalisme - et l'ensemble des sujets des tests perceptifs que j'ai effectués pour leur patience et la qualité de leur performance.

Je remercie aussi ma famille pour le soutien de tous les jours.

Résumé

Le travail réalisé durant cette thèse concerne la synthèse visuelle de la parole pour l'animation d'un humanoïde de synthèse. L'objectif principal de notre étude est de proposer et d'implémenter des modèles de contrôle pour l'animation faciale qui puissent générer des trajectoires articulatoires à partir du texte. Pour ce faire nous avons travaillé sur 2 corpus audiovisuels. Tout d'abord, nous avons comparé objectivement et subjectivement les principaux modèles existants de l'état de l'art. Ensuite, nous avons étudié l'aspect spatial des réalisations des cibles articulatoires, pour les synthèses par HMM (*Hidden Markov Model*) et par concaténation simple. Nous avons combiné les avantages des deux méthodes en proposant un nouveau modèle de synthèse nommé TDA (*Task Dynamics for Animation*). Ce modèle planifie les cibles géométriques grâce à la synthèse par HMM et exécute les cibles articulatoires ainsi générées grâce à la synthèse par concaténation. Par la suite, nous avons étudié l'aspect temporel de la synthèse de la parole et proposé un second modèle de synthèse intitulé PHMM (*Phased Hidden Markov Model*) permettant de gérer les différentes modalités liées à la parole. Le modèle PHMM permet de calculer les décalages des frontières des gestes articulatoires par rapport aux frontières acoustiques des allophones. Ce modèle a été également appliqué à la synthèse automatique du LPC (Langage Parlé Complété). Enfin, nous avons réalisé une évaluation subjective des différentes méthodes de synthèse visuelle étudiées (concaténation, HMM, PHMM et TDA).
Mots-clés : synthèse audiovisuelle, coarticulation, animation faciale, HMM, concaténation, évaluation.

Abstract

The work performed during this thesis concerns visual speech synthesis in the context of humanoid animation. Our study proposes and implements control models for facial animation that generate articulatory trajectories from text. We have used 2 audiovisual corpuses in our work. First of all, we compared objectively and subjectively the main state-of-the-art models. Then, we studied the spatial aspect of the articulatory targets generated by HMM-based synthesis and concatenation-based synthesis that combines the advantages of these methods. We have proposed a new synthesis model named TDA (Task Dynamics for Animation). The TDA system plans the geometric targets by HMM synthesis and executes the computed targets by concatenation of articulatory segments. Then, we have studied the temporal aspect of the speech synthesis and we have proposed a model named PHMM (Phased Hidden Markov Model). The PHMM manages the temporal relations between different modalities related to speech. This model calculates articulatory gestures boundaries as a function of the corresponding acoustic boundaries between allophons. It has been also applied to the automatic synthesis of Cued speech in French. Finally, a subjective evaluation of the different proposed systems (concatenation, HMM, PHMM and TDA) is presented.
Key-words : audiovisual synthesis, coarticulation, facial animation, HMM, concatenation, evaluation.

Table des matières

Table des figures

Liste des tableaux

Glossaire

ACP : L'analyse en Composantes Principales est une méthode mathématique d'analyse des données qui consiste à rechercher les directions de l'espace qui représentent au mieux les corrélations entre n variables aléatoires. L'ACP est aussi connue sous le nom de transformée de Karhunen-Loève ou de transformée de Hotelling (en l'honneur d'Harold Hotelling). Lorsqu'on veut compresser un ensemble de N variables aléatoires, la projection des points sur les n premiers axes de l'ACP est optimale, du point de vue de l'inertie expliquée.

ADL : L'analyse discriminante linéaire est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire, ...) d'un ensemble d'observations (individus, exemples, ...) à partir d'une série de variables prédictives (descripteurs, variables exogènes, ...).

Allophone : Un allophone est l'une des réalisations sonores possibles d'un phonème. Au sein d'une même langue, les allophones ne constituent pas des unités distinguées,opposants au sein de paires minimales d'un même phonème. Par exemple, si un locuteur du français roule les /r/, son interlocuteur interprétera ses énoncés de la même façon que s'il ne les roule pas car le /r/ roulé (noté [r] en phonétique) et le /r/ non roulé (le plus souvent [ʁ]) constituent des allophones d'un phonème unique. Les allophones sont désignés par un symbole entre crochets [ ].

Coarticulation : Selon les phonèmes qui l'entourent dans une phrase, un phonème n'est pas articulé de la même manière. La coarticulation est largement planifié : elle résultait d'un compromis entre la production de contrastes acoustiques - et visuels - suffisants pour un effort articulatoire minimal.

Evaluation : Les tests d'évaluation sont effectués pour évaluer la qualité de modèles de synthèse. Dans ce travail, des évaluations objectives (quantitatives) et subjectives (qualitatives) sont réalisées. Les tests objectifs calculent la distorsion entre les trajectoires de synthèse et celles d'origine. Généralement, la mesure de distorsion correspond à l'erreur moyenne ou à la corrélation linéaire. L'évaluation subjective peut être de trois types différents, Theobald *et al.* (2003) :

- Réalisme : le test de réalisme essaie de mesurer la distance subjective entre la parole d'origine (réelle) et celle synthétique.
- Intelligibilité : le test d'intelligibilité mesure la capacité de la parole à être comprise.
- Acceptabilité : le test d'acceptabilité nous montre si un modèle de synthèse est acceptable ou non pour une application donnée.

HMM : Un modèle de Markov caché (MMC) - en anglais *Hidden Markov Models* (HMM) (ou plus correctement, mais moins employé automate de Markov à états cachés) est un modèle statistique dans lequel les signaux observables d'un système sont supposés être émis par une suite d'états « cachés » de ce dernier. Un tel modèle est caractérisé par un triplet de paramètres

spécifiant les probabilités d'émission conditionnelles des observations en fonction de chaque état, les transitions entre états et l'état initial.

Morphage : Transformation progressive d'une image en une autre par un traitement informatique. S'effectue généralement par mise en correspondance des images par un unique maillage déformable puis transformation multilinéaire des pixels de chaque maille. (traduction de l'anglais morphing)

Paramètres visuels ou articulatoires : Les paramètres visuels (ou articulatoires) sont les trajectoires des caractéristiques visuelles (ou articulatoires) en fonction du temps pendant la production de la parole (coordonnées des points de visage, paramètres statistiques généralement obtenus par réduction dimensionnelle des coordonnées des certains points de visage, distances entre différents articulateurs, etc.). Dans notre travail deux types de paramètres sont utilisés :

- Paramètres géométriques (notés ABC) : paramètres correspondants aux distances entre des points caractéristiques des lèvres. Les paramètres géométriques sont : ouverture/fermeture des lèvres, étirement des lèvres et protrusion des lèvres.
- Paramètres articulatoires (notés ART) : paramètres statistiques issus de l'analyse en composantes principales (ACP) appliquée aux différents groupes de coordonnées des points du visage. Les paramètres articulatoires sont : ouverture/fermeture de la mâchoire, arrondissement des lèvres, abaissement/relèvement de la lèvre supérieure, abaissement/relèvement de la lèvre inférieure, avancement/rétraction de la mâchoire, mouvements de la gorge, etc.). Ces paramètres permettent de mettre en forme une géométrie des lèvres. Ils fournissent des degrés de libertés en excès (par exemple : une certaine ouverture des lèvres peut être obtenue par hauteur de la mâchoire et des deux lèvres).

Phone : un phone est un synonyme technique de son, vu sous l'angle de ses propriétés linguistiques. Il peut désigner spécifiquement :

- un son d'une langue (ou un geste dans le cas d'une langue des signes) considéré du point de vue de la physique sans considération de ses propriétés phonologiques
- un segment parlé doté de propriétés physiques ou perceptuelles distinctives
- une occurrence donnée d'un tel segment.

Phonème : Un phonème est la plus petite unité discrète ou distinctive (c'est-à-dire permettant de distinguer des mots les uns des autres) que l'on puisse identifier perceptivement dans la chaîne parlée. Un phonème est en réalité une entité abstraite, qui peut correspondre à plusieurs sons. Il est en effet susceptible d'être prononcé de façon différente selon les locuteurs ou selon sa position et son environnement au sein du mot (voir allophone). On transcrit traditionnellement les phonèmes par des lettres placées entre des barres obliques : /a/, /t/, /r/, etc., selon la règle un phonème = un symbole.

PHMM : *Phased Hidden Markov Model*. Le terme est proposé dans la thèse et correspond à un modèle HMM avec un modèle de déphasage. Plus précisément, un HMM et un modèle de déphasage entre les frontières des gestes articulatoires et des frontières acoustiques sont associés à chaque phone en contexte.

TDA : *Tasks Dynamics for Animation.* Le terme est proposé dans la thèse et correspond à un modèle de synthèse des mouvements faciaux à partir du texte. Le modèle est basé sur la théorie de coarticulation issue de la phonologie articulatoire et notamment sur le modèle proposé par Saltzman et Munhall.

Visème : Un visème est une unité de base de la parole dans le domaine visuel. Dans ce travail un visème est un groupe de phonèmes qui sont proches visuellement (par rapport à une distance statistique définie entre les représentations visuelles des différents phonèmes). Par exemple, les consonnes /p/, /b/ et /m/ forment un visème bilabial. Dans ce travail, nous avons défini 7 visèmes : 4 visèmes pour les consonnes et 3 pour les voyelles.

Chapitre 1

Introduction

Cette thèse a été effectuée en collaboration entre l'équipe Machines Parlantes, Agents Conversationnels et Interaction Face-à-face (MPACIF) du département Parole & Cognition du laboratoire Gipsa-Lab UMR CNRS/Universités de Grenoble et le laboratoire IRIS/IAM de France Télécom Recherche et Développement de Rennes. L'un des objectifs de l'équipe MPACIF est de développer des têtes parlantes virtuelles capables d'engager une conversation face-à-face située avec des partenaires humains. Les modèles de contrôle des gestes développés s'inspire largement de l'observation d'interactions humaines en situation. Quant à elle, l'équipe IAM de France Télécom travaille dans les domaines de création et d'animation automatique des villes en 3D et des personnages virtuels. Les personnages virtuels sont de plus en plus utilisés dans de nombreux domaines : télécommunications, jeux vidéo, divers services interactifs, etc. Le niveau de rendu et d'intelligibilité de ces personnages dépend de l'application proposée. Rares sont les systèmes qui combinent les deux. Le but du travail chez France Télécom R&D est d'avoir des personnages virtuels qui ont l'apparence et le comportement les plus proches possibles de ceux des humains.

L'objectif principal de ce travail de thèse est de proposer des modèles d'animation faciale liée à la parole. Ces modèles doivent produire automatiquement les mouvements faciaux à partir du texte. Pour obtenir un résultat aussi proche que possible du naturel, nous avons utilisé des données audiovisuelles issues de systèmes de capture des mouvements faciaux humains. Les données capturées, et plus exactement les paramètres visuels et acoustiques, sont traitées et analysées en fonction de l'information phonétique. La principale problématique de la synthèse de la parole est liée à la grande variabilité intra- et inter- locuteurs des articulations observées. Non seulement un son n'est pas articulé de la même manière par différents locuteurs, mais il n'est jamais prononcé de la même façon par un même locuteur. De nombreux facteurs influencent en effet l'articulation d'un son : son entourage phonétique, le contenu et la structure de l'énoncé ainsi que d'autres variables physiologiques et paralinguistiques liées au locuteur (sexe, âge, origine géolinguistique, activité socioprofessionnelle, etc.) ou à son état émotionnel, physiologique ou psychologique.

Ainsi, le travail de la thèse consiste à étudier et à modéliser les mouvements faciaux. Plus précisément à générer automatiquement les trajectoires des paramètres articulatoires afin de modéliser au mieux le phénomène de coarticulation spécifique à un locuteur.

Le plan de la thèse est le suivant. Dans le chapitre 2, les principaux systèmes de l'état de l'art s'intéressant à la synthèse visuelle à partir du texte et de l'audio sont décrits. Ensuite, la problématique d'évaluation de ces systèmes est abordée et les modèles les plus représentatifs sont évalués objectivement et subjectivement. Dans le chapitre 3, les données audiovisuelles

qui ont été utilisées dans notre étude sont présentées. Dans le chapitre 4, l'aspect spatial de la génération des trajectoires articulatoires dans le cadre des synthèses par concaténation et par HMM est mis en évidence. Suite à cette étude, notre première contribution consistant en un nouveau modèle de synthèse nommé TDA (*Task Dynamics for Animation*) est détaillée. Dans le chapitre 5, en intégrant l'aspect temporel de la synthèse de la parole, un nouveau modèle de synthèse PHMM (*Phased Hidden Markov Model*) est proposé permettant de gérer les relations temporelles des différentes modalités liées à la parole. Ce modèle est également appliqué à la synthèse automatique du Langage Parlé Complété (LPC) en français. Dans le chapitre 6, les résultats d'une évaluation subjective menée entre les principaux modèles utilisés au cours de ce travail (HMM, concaténation, PHMM et TDA) sont analysés. Enfin, les conclusions et les perspectives du travail sont présentées.

Chapitre 2

Etat de l'art

2.1 Synthèse de la parole

La synthèse de parole à partir du texte a pour but de donner la possibilité aux (micro-) ordinateurs d'émettre des sons de parole à partir de n'importe quel texte tapé au clavier (ou, par exemple, reconnu par un système de reconnaissance de caractères ou de son). L'entrée d'un tel système est une suite de symboles appartenant à un alphabet fini (une chaîne de caractères alphanumériques) et la sortie des signaux (audio et/ou visuel) continus et hautement variables. Le défi majeur de la synthèse est d'identifier les facteurs contextuels de cette variabilité et de paramétrer au mieux des modèles prédictifs de cette variabilité. Ces systèmes peuvent se décomposer en plusieurs modules. De manière générale, les systèmes actuels exploitent les modules suivants (voir Figure 2.1) :

- un module de traitements linguistiques qui fournit des connaissances linguistiques (chaîne phonétique, structure grammaticale, phonologique) sur l'énoncé à prononcer.
- un dictionnaire de segments multi-représentés. Généralement les segments correspondent à des diphones, demi-syllabes, etc. Ils permettent d'accéder à une représentation paramétrique de portions de signaux correspondants. Ces segments sont indexés par des étiquettes de même nature que celles délivrées par le module précédent afin de pouvoir sélectionner grâce à ces étiquettes les segments les plus appropriés à « rendre » ces informations linguistiques. Ces clés peuvent être enrichies par des informations paralinguistiques (émotions, attitudes, etc.) augmentant le texte d'entrée.
- un module de sélection/concaténation des segments. Le coût de sélection se calcule grâce aux distances entre les structures phonologiques du texte à prononcer et des segments du dictionnaire. Le coût de concaténation correspond aux distances entre les segments successifs à concaténer aux points de concaténation.
- un module de traitement prosodique qui calcule une partie des paramètres caractéristiques des segments (durée des sons, variations de fréquence fondamentale ou de spectre, etc.). Ces valeurs peuvent alors être utilisées pour sélectionner de manière plus fine les segments appropriés en intervenant dans le coût de sélection utilisé dans le module précédent. Il faut noter que l'on peut s'affranchir du modèle prosodique si la sélection est effectuée grâce à une structure phonologique riche et détaillée (Taylor & Black, 1999).
- un module de traitement du signal qui récupère les représentations paramétriques des portions de signaux sélectionnés et concaténés et se charge de calculer un signal de synthèse. Ce module peut aussi exploiter la sortie du module prosodique précédent pour mettre en accord les représentations paramétriques avec les paramètres prosodiques calculés plus

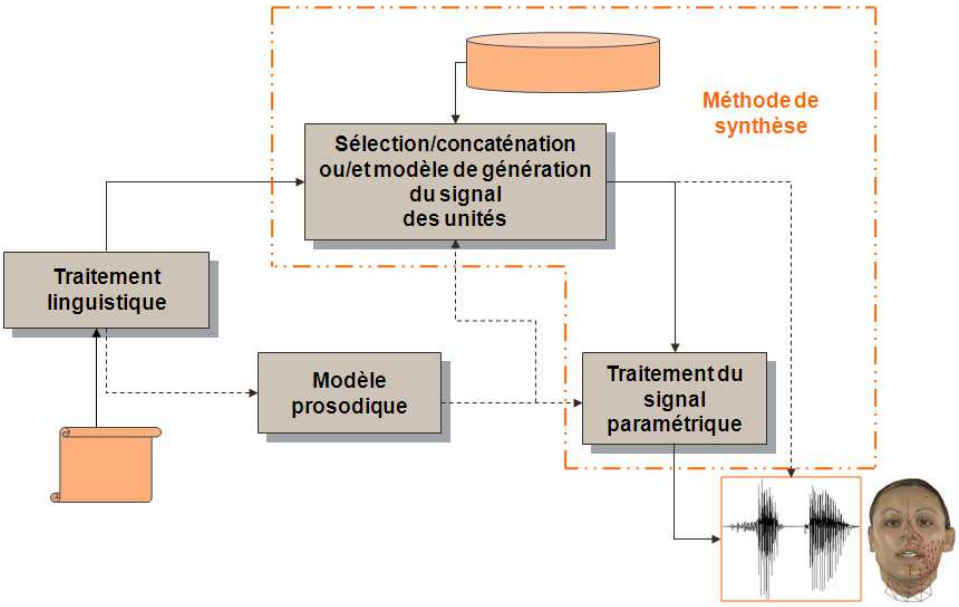


FIG. 2.1 – Schéma général de la synthèse de la parole. Le système illustré ici génère une animation faciale synchrone avec le son étant donné une chaîne de phonèmes marquée en durées.

haut. Ce module peut être au contraire omis dans le cas où le dictionnaire stocke directement des portions de signaux non paramétrés. Une telle option a été d'ailleurs proposée par (Campbell, 1995) pour l'acoustique et par (Weiss, 2005) et (Fagel, 2006) pour l'animation faciale.

Pour une revue complète de la synthèse de parole audio à partir du texte le lecteur pourra se référer à (Calliope, 1989), (d'Alessandro & Tzoukermann, 2001), (van Santen, 1997), (Boite *et al.* , 2000).

Dans le travail de la thèse, l'objectif principal est de faire un modèle de synthèse visuelle de la parole, c'est-à-dire modéliser les mouvements faciaux liés à la parole en fonction d'une chaîne de phonèmes marquée en durées. La plupart des systèmes de synthèse visuelle étudiés dans ce travail acceptent une entrée plus riche qu'une chaîne linéaire de phonèmes, notamment en ajoutant des informations phonotypiques (position dans la syllabe, le mot, etc.) caractérisant chaque allophone. Nous avons choisi de comparer les systèmes dans leur capacité à explorer des entrées minimales.

2.2 Motivations de la synthèse visuelle de la parole

La production et la perception de la parole sont intrinsèquement bimodales. Les humains combinent les informations audio et visuelle pour comprendre ce qui est dit, notamment dans les environnements bruités. La modalité visuelle améliore l'intelligibilité de la parole dans environnements bruités, et cela a été déjà quantifié par Sumby et Pollack (Sumby & Pollack, 1954). L'importance de la fusion correcte audiovisuelle est démontrée dans l'effet de McGurk[1], (Mc-

[1]L'effet McGurk est un phénomène perceptif qui montre une interférence entre l'audition et la vision lors de la perception de la parole. Il suggère que la perception de la parole est intrinsèquement multimodale. Pour montrer l'effet, (McGurk & MacDonald, 1976) ont présenté une vidéo montrant une personne prononçant un phonème (p.ex. /g/) alors que la bande sonore diffuse en synchronie l'enregistrement d'un autre phonème (p.ex.

Gurk & MacDonald, 1976). De plus, la parole visuelle est particulièrement importante pour les sourds et malentendants : les mouvements labiaux jouent un rôle essentiel dans le langage des signes et dans la communication entre les malentendants (Marschark *et al.* , 1998), (Caplier *et al.* , 2007). Il y a trois raisons principales qui expliquent pourquoi la modalité visuelle améliore la perception de la parole (Summerfield, 1987) :

1. l'information visuelle permet la localisation de la source audio
2. elle contient de l'information phonétique complémentaire à la parole acoustique
3. cela permet d'avoir de l'information robuste sur certains lieux d'articulation, notamment labiales

Cette visibilité des organes peut être totale (lèvres, joues) ou partielle (langue, dents). L'information sur le lieu de l'articulation permettrait de lever des ambiguïtés, par exemple, entre les consonnes non-voisées /p/ (bilabiale) et /k/ (vélaire), entre les consonnes voisées /b/ (bilabiale) et /d/ (alvéolaire) et entre la nasale /m/ (bilabiale) et la nasale /n/ (alvéolaire), (Massaro & Stork, 1998). Notons en outre que les personnes sont très sensibles aux incohérences audiovisuelles spatiales (McGurk & MacDonald, 1976) mais aussi temporelles (Dixon & Spitz, 1980). Ces derniers ont montré que leurs sujets détectent l'asynchronie audiovisuelle à partir d'une avance de 200 ms du son sur l'image.

2.3 Animation des visages parlants

Un système d'animation d'un visage parlant comprend généralement trois modèles (Bailly *et al.* , 2003) :

- le modèle de contrôle des paramètres articulatoires qui calcule les trajectoires articulatoires à partie de la chaîne phonétique
- le modèle de forme qui décrit comment change la géométrie faciale en fonction de l'articulation
- le modèle d'apparence qui se charge de calculer le rendu final du visage

Cette chaîne de traitement est illustrée dans la Figure 2.2. Soulignons que les paramètres articulatoires calculés par le modèle de contrôle sont souvent utilisés par le modèle d'apparence (voir notamment les modèles AAM plus bas).

Dans ce qui suit, ces différents aspects de la synthèse des mouvements faciaux liés à la parole sont présentés.

2.3.1 Modèles d'apparence et de forme

Deux principales catégories de modèles de visages parlants sont distingués : systèmes basés image ou vidéo (2D) et systèmes basés modèle (3D), Figure 2.3.

/b/). Lorsque les signaux sont bien synchronisés, le système perceptif est piégé et ne perçoit qu'un unique percept produit de la fusion multimodale (/d/ ou /v/).

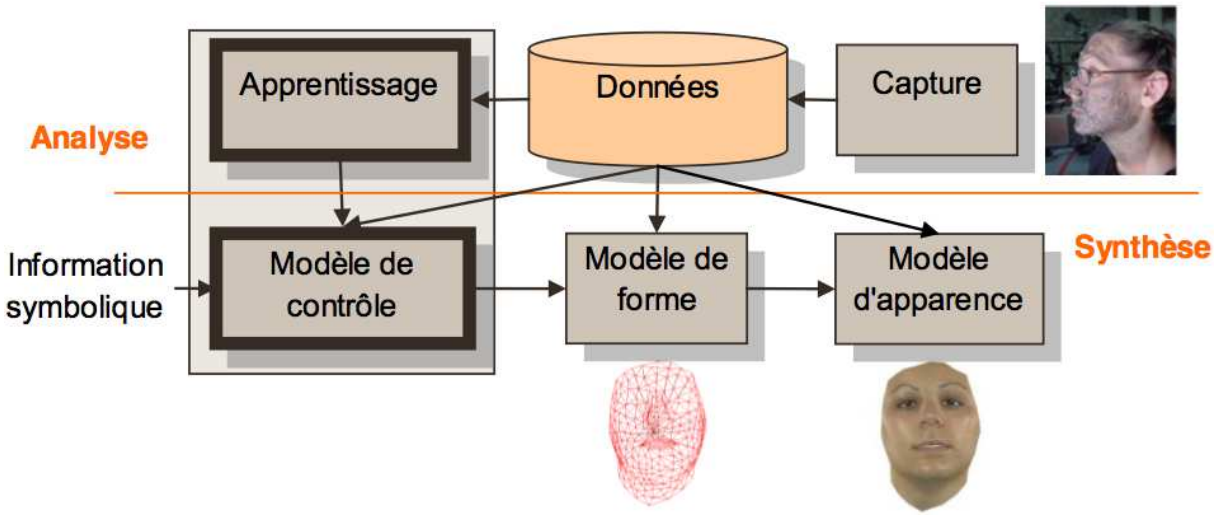


FIG. 2.2 – Modèle général d'une tête parlante. On distingue ici les données et traitements nécessaires à l'analyse hors-ligne et les modules sollicités pour la synthèse en ligne.

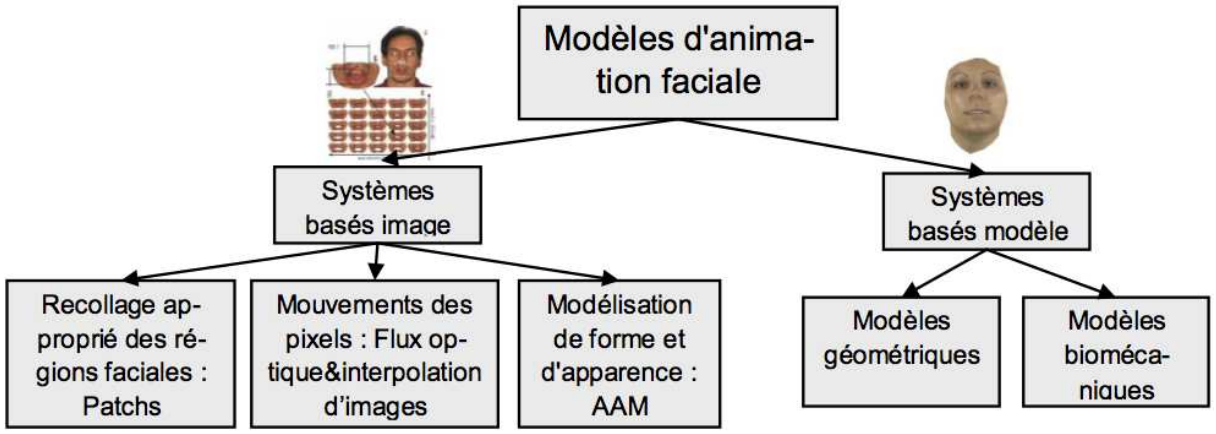


FIG. 2.3 – Systèmes d'animation faciale : systèmes basés image ou vidéo et systèmes basés modèle.

Animation vidéo

Les systèmes basés image estiment la variation de la couleur des pixels en fonction de la parole. Ces systèmes peuvent être regroupés en trois familles (Bailly *et al.* , 2003) : les systèmes qui choisissent les segments appropriés dans une grande base de données et qui superposent les parties choisies sur une image de fond (Bregler, 1997), (Cosatto & Graf, 2000) ; les systèmes qui considèrent les mouvements comme des déplacements de pixels (Ezzat *et al.* , 2002) ; et les systèmes qui calculent l'apparence de chaque pixel en fonction des mouvements faciaux (Brooke & Scott, 1998), (Theobald *et al.* , 2003).

Patchs Le système caractéristique du premier groupe de modèles est celui proposé par Bregler : "*Video Rewrite*" (Bregler, 1997). Ce système utilise une vidéo de fond qui sert de scène aux mouvements des lèvres synthétisés, (Figure 2.4a). Sur la vidéo de fond, la plus longue séquence vidéo liée à la région des lèvres de la base d'apprentissage qui correspond au bon visème, au bon phonème et à la bonne position de tête sont superposés. Le modèle de contrôle est basé sur la concaténation de triphones. Des ajustements sont calculés et appliqués pour pallier les mouvements de tête qui imposent des modifications de l'image de la région des lèvres.

Un autre système de synthèse audiovisuelle, développé dans les laboratoires de ATT par Cosatto et Graf (Cosatto & Graf, 2000), utilise le même principe que Video Rewrite mais avec une décomposition plus complète du visage. Celui-ci est décomposé en six régions : les yeux, la bouche, les dents (supérieures et inférieures), le menton et le front, (Figure 2.4b). Il faut ensuite agencer les mouvements de toutes ces régions de manière cohérente. Le patch de régions a l'avantage de maintenir automatiquement la cohérence des éléments les composant. Par contre, l'éclatement permet de contrôler les éléments séparément et donc de mémoriser les configurations de la base d'apprentissage.

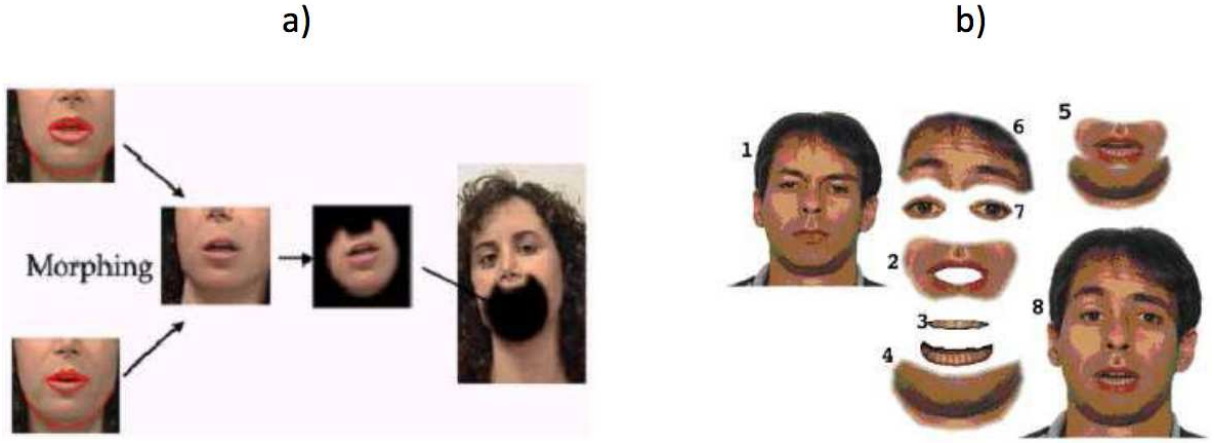


FIG. 2.4 – Exemples de systèmes utilisant la superposition de segments vidéo. a) Système de synthèse Video Rewrite (Bregler, 1997), b) Système de synthèse proposé par (Cosatto & Graf, 2000).

Flux optique et interpolation d'images Dans le deuxième groupe, où la synthèse 2D modélise les mouvements des pixels en fonction du son prononcé, les systèmes suivants peuvent être cités : Actors (Scott *et al.* , 1994), MikeTalk (Ezzat & Poggio, 1998) et Mary101 (Ezzat *et al.* , 2002).

Dans le système MikeTalk (Figure 2.5a), une image prototypique (le visème) représente la cible à atteindre pour le visage pour chaque phonème en contexte. Le modèle d'apparence consiste ensuite en une interpolation entre images-clés : le flux optique est calculé pour chaque passage d'un visème à un autre dans les deux sens, ce qui permet de reconstruire les images intermédiaires par mélange progressif des images le long des flux optiques. Dans le cas de Mary (Figure 2.5b), un phonème en contexte est associé à une distribution sur la base des visèmes par un modèle statistique qui préfigure la synthèse par HMM : le MMM (*Multidimensional Morphable Model*) utilise pour ceci un modèle de forme élaboré par Analyse en Composantes Principales (ACP) des flux optiques reliant une forme neutre et les visèmes sélectionnés. Une méthode de pilotage d'une personne à partir du modèle d'une autre personne est implémentée dans (Chang & Ezzat, 2005).

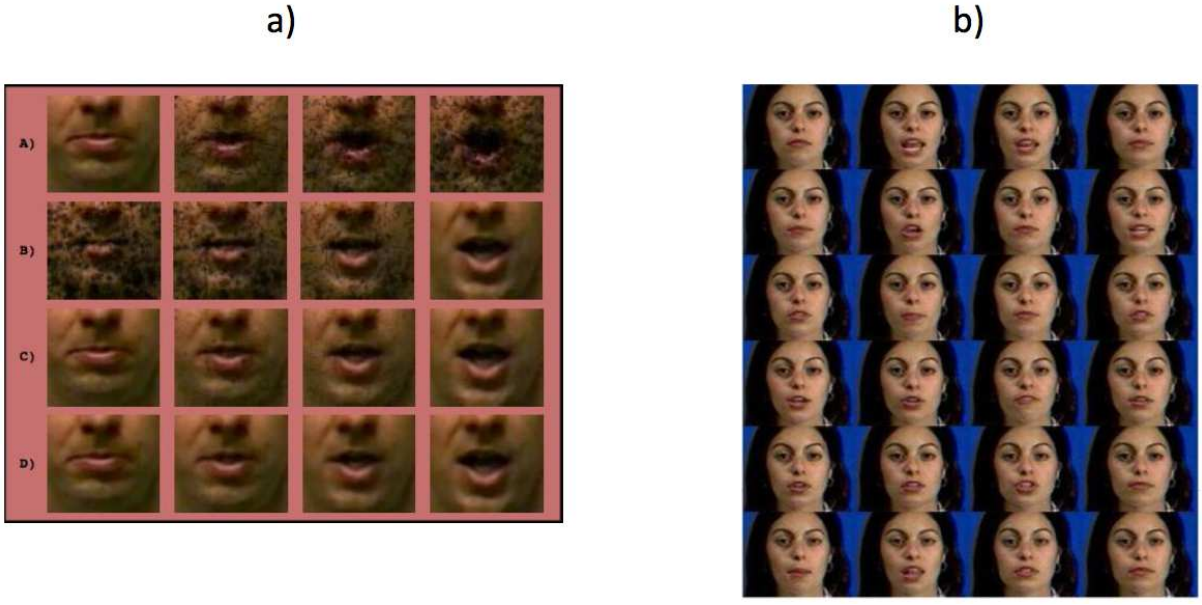


FIG. 2.5 – a) « MikeTalk » : de haut en bas : transformation d'une image I0 (visème0) vers une image1 (visème1), transformation d'une image I1 à une image I0, morphage des images I0 et I1, morphage des images I0 et I1 après un filtrage (Ezzat & Poggio, 1998). b) « Mary101 » : 24 des 46 images prototypiques constituant le MMM (Ezzat *et al.* , 2002).

Active Appearance Models Dans le troisième groupe, les méthodes consistent à créer un modèle statistique de forme et d'apparence du visage (Cootes *et al.* , 2001), (Theobald *et al.* , 2003), (Cosker *et al.* , 2003), voir un exemple dans la Figure 2.6. Dans un premier temps, le modèle statistique de forme est déterminé : on positionne à la main sur un ensemble d'images un maillage déformable et on applique une ACP (Analyse en Composantes Principales) sur ces points de contrôle. Ensuite, un modèle d'apparence est calculé en morphant (voir Glossaire) toutes les images sur la forme moyenne : on obtient ce qu'on appelle des images libres de forme (ou *shape-free images*). Chaque image est ainsi caractérisée par un nombre constant de pixels. On effectue alors une ACP sur les composants RGB de chaque pixels de ces images libres de forme. Cette reduction est appliquée sur des tableaux de grande dimension (par exemple : 768 lignes * 575 colonnes * 3 couleurs (RGB)). Enfin, une ACP de ces deux modèles est effectué. Cependant, construire un modèle statistique d'apparence d'un visage en utilisant une ACP donne lieu à des artefacts : les variations de texture sont notamment non linéaires (apparition/disparition de rides/plis, des dents, de la langue, etc.). Une solution est proposée consistant à modéliser séparément ces différentes régions et créer des MAM (*Multi-segment Appearance Models*) (Theobald

et al. , 2003). Pour construire les MAM, les images sont segmentées en sous-régions perceptivement importantes comme dans dans les systèmes par patches : le visage entier, la bouche et chacun des yeux.

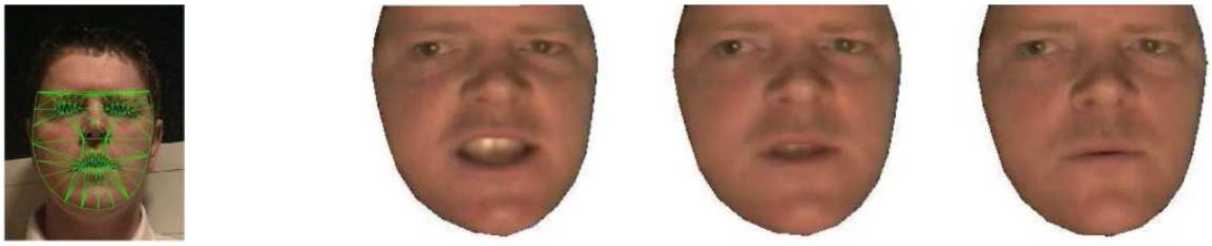


FIG. 2.6 – Modèles de formes et d'apparence : a) le modèle de forme est utilisé pour normaliser les images de la base d'apprentissage ; b) images de synthèse créées à partir du modèle de forme et d'apparence (Theobald *et al.* , 2003).

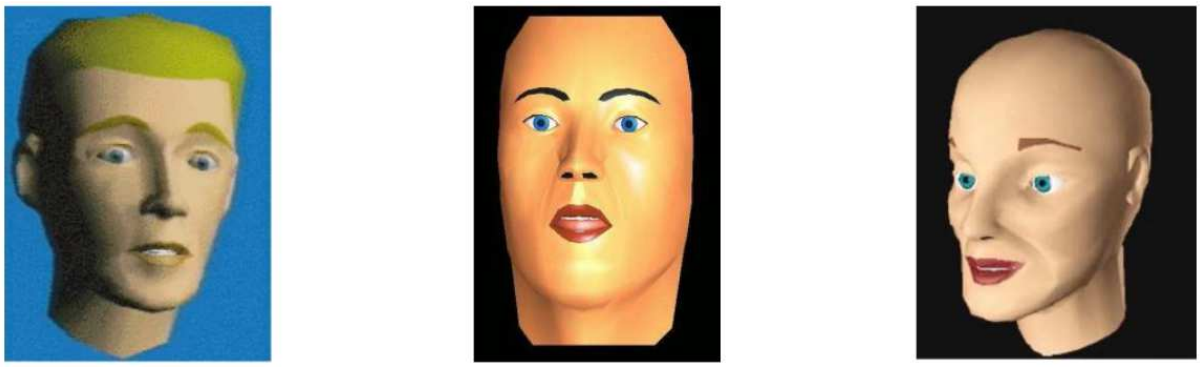


FIG. 2.7 – Exemples de descendants du modèle de Parke (dans l'ordre : Sven, Baldi, LCE).

Animation 3D

Dans les approches basées modèle, la surface faciale est décrite sous la forme d'un maillage polygonal, généralement en 3D. Pendant l'animation, la surface est déformée en déplaçant les sommets du maillage en gardant sa topologie constante. Les mouvements des sommets sont gouvernés par un ensemble de paramètres. Les techniques de l'association des paramètres aux mouvements des sommets peuvent être classées en deux catégories : les approches géométriques et les approches biomécaniques.

Le pionnier de l'approche géométrique est le modèle de Parke (Parke, 1974). De nombreux chercheurs (Beskow, 1995), (Olives *et al.* , 1999), (Massaro, 1998) se basent sur ce modèle pour créer leurs têtes parlantes, (Figure 2.7).

Une norme standardisée MPEG4 (Pandzic & Frorchheimer, 2002) est, également, proposée où les coordonnées 3D des 84 FP (*Feature Points*) sont contrôlées par 68 paramètres (FAP : *Facial Action Parameters*). Ces paramètres sont responsables de la description des mouvements faciaux à deux niveaux : au niveau bas (déplacements de points 3D du visage) et au niveau haut (reproduction des expressions).

L'équipe de l'ICP (Badin *et al.* , 2002), (Elisei *et al.* , 2001), (Revéret *et al.* , 2000) propose

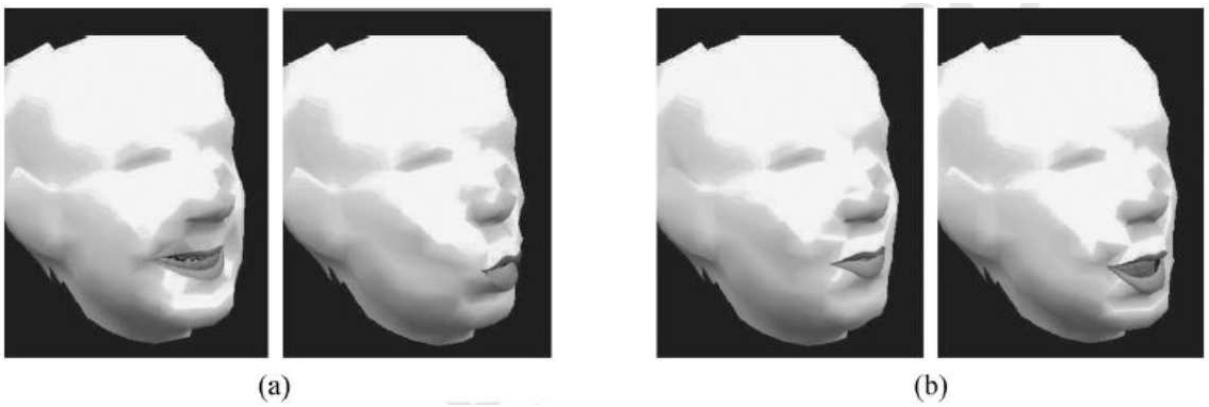


FIG. 2.8 – a) Le premier geste articulatoire statistiquement significatif issu de l'ACP correspondant à l'étirement vs. l'arrondissement des lèvres. b) Le second geste articulatoire correspondant aux mouvements intrinsèques de la lèvre inférieure.

de définir des paramètres articulatoires issus d'une ACP (Analyse en Composantes Principales) guidée (voir la Figure 2.8). La méthodologie consiste en une série de régressions linéaires de l'ensemble des points par des composantes linéaires estimées par des ACP appliquées aux mouvements de différents sous-ensembles de points de peau, supposés déformés par un unique degré de liberté sous-jacent (par exemple, arc mandibulaire pour rotation de la mâchoire). Ces points de peau sont de l'ordre de 200 et rendent compte des variations fines du visage. Pour tous les visages étudiés, 7 paramètres (mouvements de la mâchoire, des lèvres, et du larynx) ainsi obtenus couvrent plus de 95% (Elisei *et al.* , 2001) de la variance des mouvements faciaux liés à la production de parole.

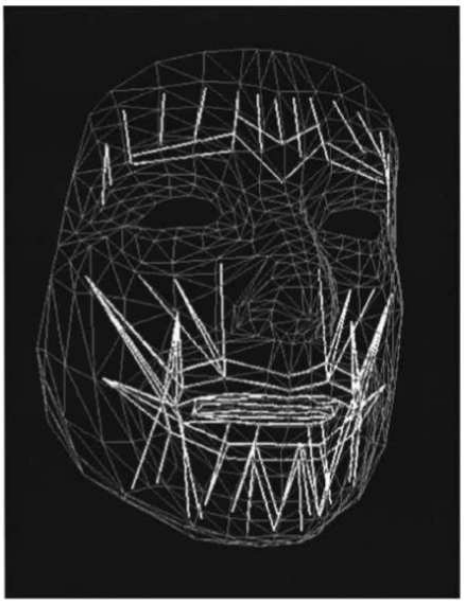


FIG. 2.9 – Lignes d'action des muscles du visage du modèle de Lucero et al. (Lucero *et al.* , 1997).

Dans les approches biomécaniques, le but est de simuler les propriétés biomécaniques des tissus et du système musculo-squelettique (Waters, 1987). Ces modèles sont contrôlés par de nombreux paramètres et sont souvent d'une très grande complexité (Bailly *et al.* , 2003). Les sommets des maillages 3D sont considérés dans ce type de méthodes comme des points de chair.

L'avantage de cette méthode est que les mouvements du visage sont contrôlés par des activations musculaires qui sont supposées être directement connectées à des intentions de communication. On peut citer les travaux de Ekman et Friesen (Ekman & Friesen, 1978) qui ont établi un système, appelé FACS (*Facial Action Coding System*) décrivant les expressions faciales par 66 actions musculaires. Les muscles appliquent des forces à des ensembles de structures géométriques représentant des objets tels que les tissus de peau, (Figure 2.9). En ce qui concerne la modélisation des tissus de peau, l'approche la plus simple consiste à créer une collection de ressorts connectés entre eux en réseaux (Platt & Badler, 1981),(Breton *et al.* , 2001) et organisés en couches (Waters, 1987), (Terzopoulos & Waters, 1990), (Lee *et al.* , 1995). Les inconvénients majeurs de tels systèmes sont (a) la complexité du contrôle (redondance musculaire), (b) l'instabilité des simulations dynamiques (Pitermann, 2004) et (c) la modélisation passive des tissus (protrusion des lèvres souvent simulée par des ressorts externes au maillage du visage.

L'évolution de ce type de méthodes tend vers la modélisation par éléments finis des couches de tissus de peau (Basu *et al.* , 1998), (Couteau *et al.* , 2000), (Groleau *et al.* , 2007), (Nazari *et al.* , 2008).

2.3.2 Modèles de contrôle

Le modèle de contrôle transforme une information phonétique ou acoustique en mouvements articulatoires. Les systèmes utilisant une information phonétique marquée en durée et les systèmes utilisant une information acoustique sont différenciés dans la synthèse audiovisuelle.

Problématique de la coarticulation

La parole continue est caractérisée par une grande variabilité dans les domaines articulatoire et acoustique (J.Hardcastle & Hewlett, 1999). Les effets de la dépendance contextuelle sont un résultat des articulations superposées ou de la coarticulation. Par définition la coarticulation phonétique est un phénomène de la variation de la prononciation (propriétés articulatoires ou acoustiques) d'un segment phonique en fonction des segments voisins dans la chaîne parlée. Par exemple : [t] dans les segments [ti] vs [tu] où il est influencé par les voyelles qui suivent : dans [ti], le [t] est non-arrondi alors que dans [tu], il est arrondi. On distingue deux types de coarticulation : anticipatoire et progressive. Coarticulation anticipatoire (régressive) est un mouvement articulatoire nécessaire à la production d'une unité phonique et qui est déjà amorcé lors de la réalisation phonique précédente dans la chaîne parlée. Par exemple, "Je n'ai pas de sac". - La consonne sonore [d] est généralement assourdie par anticipation de la sourde [s]. Coarticulation par persistance (progressive) est un mouvement articulatoire caractéristique d'une réalisation phonique qui persiste pendant la production de l'unité phonique suivante. Par exemple, dans la phrase "Il est craintif", - la consonne sonore [r] est partiellement assourdie par l'occlusive sourde [k]. Différentes représentations de réalisations consonantiques dans des contextes vocaliques obtenues Rayons X sont présentées dans la Figure 2.10. Sur cette illustration, on voit que les constrictions des occlusives sont très influencées par le contexte vocalique : dans le cas de la consonne /b/ (bilabiale) c'est surtout la position et la forme de la langue qui dépendent du contexte vocalique, pour la consonne /d/ (apico-dentale) c'est la position et la forme de la langue mais aussi du pharynx qui varient, alors que pour la consonne /g/ (dorso-vélaire) c'est surtout les lèvres qui anticipent la voyelle porteuse mais également la racine de la langue et le pharynx.

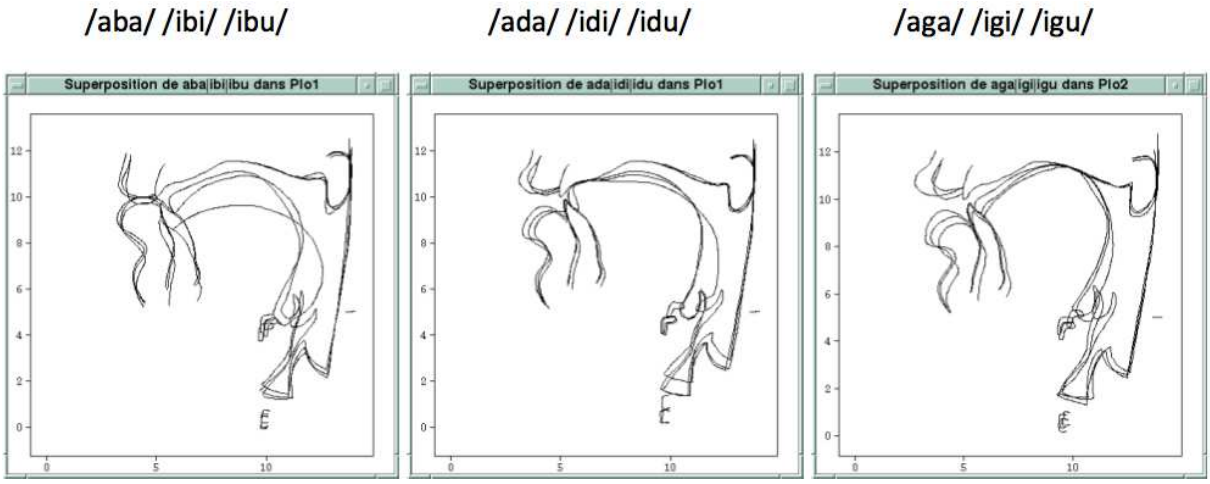


FIG. 2.10 – Réalisations des constrictions consonnantiques dans les différents contextes vocaliques (images par les Rayons X). Dans l'ordre : superposition des constrictions des bilabiales des /aba/, /ibi/, /ibu/ ; apico-dentales /ada/, /idi/, /idu/ ; dorso-vélaires /aga/, /igi/, /igu/.

Parmi les théories classiques de la coarticulation, les modèles se distinguent par le niveau de la variabilité contextuelle et de l'invariance articulatoire et/ou acoustique prises en compte lors de la planification du mouvement. De manière générale, ces modèles considèrent une invariance absolue ou contextuelle d'un jeu de paramètres articulatoires, géométriques ou acoustiques cruciaux pour chaque phonème considéré, la variabilité de surface étant expliquée par une optimisation sous contrainte des paramètres non cruciaux.

Il existe deux principaux modèles numériques de coarticulation, d'une part les modèles basés sur les équations de Lofqvist (Lofqvist, 1990) qui sont développées pour quantifier la coarticulation dans les syllabes du type CV et, d'autre part le modèle d'Öhman (Öhman, 1967) qui est développé pour les séquences du type VCV. Ce modèle suppose l'existence d'un geste vocalique porteur, lisse et lent, sur lequel viennent se greffer des gestes consonantiques rapides. L'équation du mouvement (2.1) est alors assez simple :

$$p\left(x,t\right) = v\left(x,t\right) + k_c\left(t\right) * w_c\left(x\right) * \left[c\left(x\right) - v\left(x,t\right)\right] \tag{2.1}$$

– où x correspond à un paramètre, t est le temps, $p(x,t)$ est la valeur d'un paramètre, $v(x,t)$ est la valeur d'un paramètre du geste vocalique sous-jacent, $c(x)$ est la cible consonantique, $k_c(t)$ la valeur de l'émergence d'une consonne (=1 à la cible consonantique) et $w_c(x)$ est le facteur de coarticulation (=1 quand la fermeture ne dépend pas du contexte vocalique).

Trois modèles ont été proposés pour rendre compte de l'empan temporel de l'anticipation articulatoire : le *look_ahead model* (Kozhevnikov & Chistovich, 1965), le modèle de coproduction (Perkell & Chiang, 1986) et le modèle hybride (Browman & Goldstein, 1990a), voir la Section 4.3.1.

Pour avoir plus de détails sur le phénomène de la coarticulation le lecteur peut se référer à (J.Hardcastle & Hewlett, 1999).

L'objectif principal du travail de la thèse est d'apprendre un modèle de contrôle ou un modèle de génération de trajectoires articulatoires qui puisse modéliser l'effet de la coarticulation spécifique à un locuteur. Dans la Figure 2.11, les réalisations de toutes les trajectoires de la consonne /g/ sont présentées. Ce simple exemple montre l'importante variabilité des trajectoires articulatoires que l'on cherche à expliquer et reproduire en synthèse visuelle automatique de la

parole.

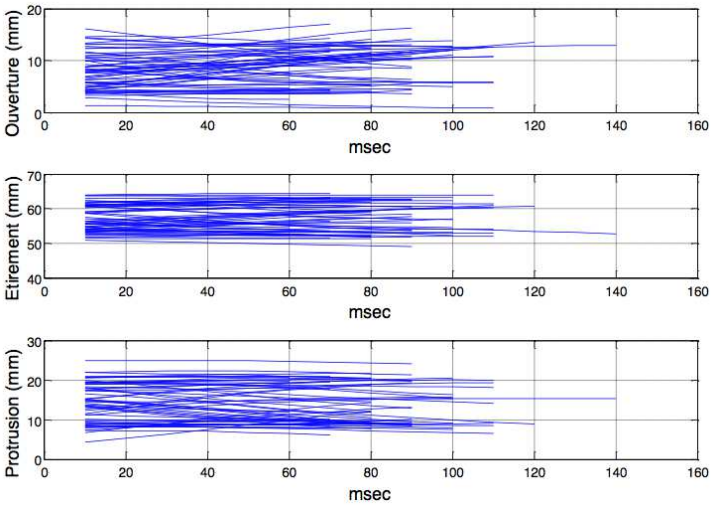


FIG. 2.11 – Toutes les réalisations de la consonne [g] du corpus II selon les trois paramètres géométriques : ouverture, étirement et protrusion des lèvres. On note que les trajectoires de [g] sont influencées par le contexte.

Différents systèmes de formation de trajectoires articulatoires ont été proposés : nous commencerons par détailler ceux qui exploitent directement des théories de coarticulation et, notamment, les modèles numériques de coarticulation et finirons par les méthodes issues des apprentissages automatiques dont le but est de reproduire au mieux au sens des moindres carrés les trajectoires articulatoires originales observées dans un corpus d'apprentissage.

Les différents systèmes de l'état de l'art sont présentés dans la section suivante, Figure 2.12.

Synthèse à partir d'une chaîne phonétique

Modèles basés visèmes Dans les modèles basés visèmes, les trajectoires articulatoires sont obtenues en suivant certaines règles. Les règles sont souvent définies empiriquement ou extraites d'observations ou d'expériences. Le modèle le plus répandu dans l'animation des visages parlants a été proposé par Cohen et Massaro (Cohen & Massaro, 1993). Il est basé sur le modèle gestuel de la production de la parole de Lofqvist (Lofqvist, 1990). Dans ce modèle, les trajectoires articulatoires sont obtenues en superposant des gestes articulatoires élémentaires (voir la Figure 2.13). Chaque segment est associé à une valeur cible et à une fonction dite de dominance, caractérisée par une décroissance exponentielle de part et d'autre de la cible. A chaque instant, la valeur d'un paramètre est calculée comme une somme pondérée de toutes les valeurs cibles pondérées par leurs dominances à cet instant. Chaque fonction de dominance a trois paramètres : sa hauteur à la valeur du pic, ses taux d'accroissement et de décroissance. Ces trois valeurs sont ajustées en fonction de chaque phone et de chaque paramètre articulatoire (voir la Figure 2.13). Notons que si on veut qu'un paramètre atteigne une certaine cible quelque soit le contexte, il faut que les fonctions de dominance adjacentes soient nulles à la cible. Ceci pose notamment des problèmes pour les occlusifs et surtout pour les bilabials.

Un autre modèle basé règles a été proposé par Beskow (Beskow, 1995). Dans ce modèle, la

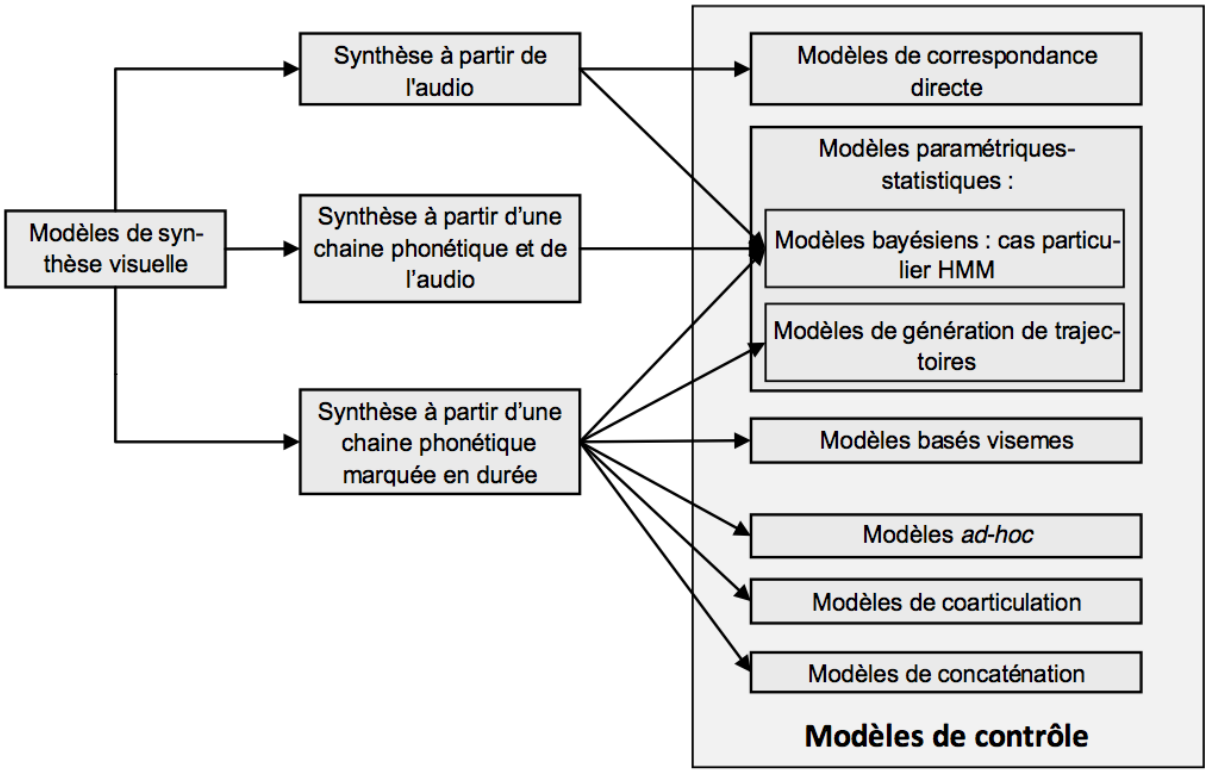


FIG. 2.12 – Modélisation de la parole visuelle : modèles basés sur l'information phonétique et modèles basés sur l'information acoustique.

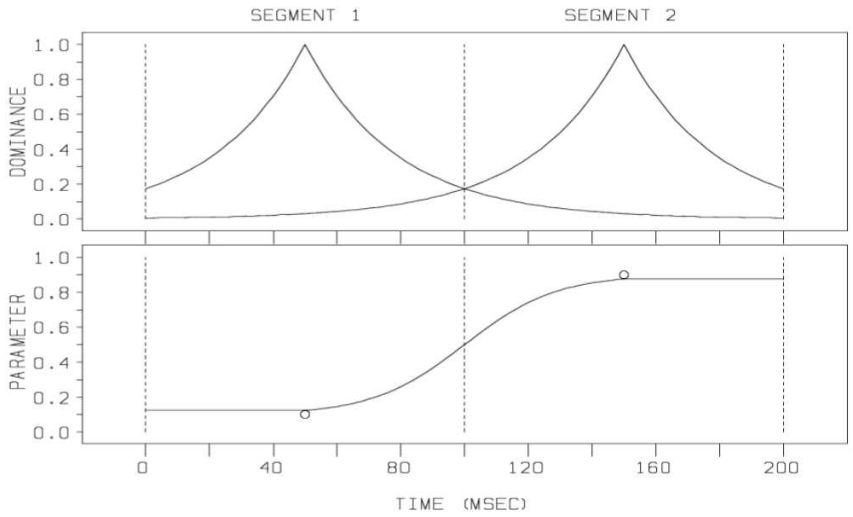


FIG. 2.13 – Modèle de dominance de Cohen&Massaro : la partie du haut correspond aux fonctions de dominance de 2 segments phonétiques ; la partie du bas correspond à la trajectoire d'un paramètre articulatoire obtenue comme une superposition des gestes articulatoires de 2 segments.

caractérisation des cibles visuelles allophoniques des 45 phonèmes du suédois en 10 paramètres articulatoires (mouvements de la mâchoire, des lèvres ...) est factorisée sur 21 visèmes. Une valeur définie ou non définie de chaque paramètre est associée à chaque visème. Si la valeur du paramètre est non définie, cela signifie que le visème est indépendant de ce paramètre. Par exemple, /r/ peut être soit arrondie, soit non arrondie en fonction du contexte, ainsi le paramètre d'arrondissement de lèvres pour ce visème n'est pas défini. Pendant la synthèse les valeurs des paramètres non définis sont calculées grâce à l'interpolation des paramètres voisins. Les règles de ce modèle sont définies empiriquement.

Pelachaud et al. (Pelachaud *et al.* , 1996) proposent aussi un modèle basé règles où les phones sont classées en visèmes. Les visèmes sont divisés en deux groupes : les visèmes qui ne dépendent pas du contexte - les visèmes clés, et les visèmes qui dépendent du contexte - les visèmes de transition. Le but est de calculer les paramètres articulatoires correspondants aux visèmes de transition. Pour le modèle *look-ahead* (Kozhevnikov & Chistovich, 1965), (Cohen & Massaro, 1993), l'expansion d'un mouvement articulatoire est proportionnelle à l'intervalle entre deux segments clés qui commence au début du premier segment clé et se termine au début du deuxième segment clé. Notons l'ancrage sur le début du dernier segment : que ce soit pour la langue ou les lèvres, la configuration cible centrale est atteinte juste à temps (Abry & Boe, 1986).

« MASSY », le modèle de visage proposé par Fagel et al. (Fagel & Clemens, 2004) utilise le modèle de Cohen & Massaro pour son animation. Le modèle est animé par 6 paramètres articulatoires (Figure 2.14). Le modèle d'animation est construit et ajusté à partir de données articulatoires capturées par un système vidéo couplé à un articulagraphe 2D. Les phonèmes de l'allemand sont regroupés en 15 groupes de visèmes (15 voyelles et 9 groupes de consonnes). Pour chaque groupe de visèmes et chaque paramètre, un modèle de dominance est défini.

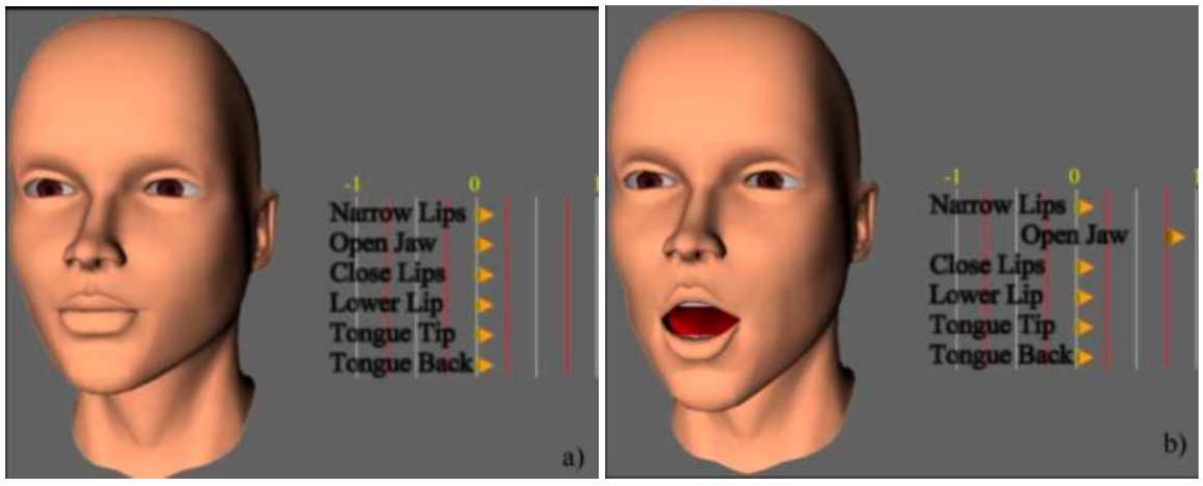


FIG. 2.14 – Un exemple du visage « MASSY » animé par six paramètres avec les articulateurs (a) en position neutre et (b) les articulateurs avec la descente maximale de la mâchoire.

Les avantages des modèles basés règles sont les suivants : d'une part ils sont basés sur des règles mathématiquement simples et donc peuvent être facilement appliqués pour animer une tête parlante et d'autre part ces modèles ne demandent que peu ou pas de données audiovisuelles issues d'une capture de gestes sur un locuteur humain. L'inconvénient principal de ces modèles est le fait qu'ils ne produisent pas les mouvements proches des mouvements naturels, car nous pouvons supposer qu'il soit impossible de reproduire la complexité de l'articulation humaine en utilisant une simple coproduction de gestes et une simple superposition de fonctions élémentaires.

A noter que les modèles présentés ci-dessus peuvent être également paramétrés à partir des données audiovisuelles. (voir notamment le travail de Le Goff et Benoit plus bas).

Modèles basés données Dans les modèles basés données, j'ai classé tous les modèles qui sont produits par apprentissage ou qui utilisent des données pour produire les trajectoires articulatoires. Les modèles basés données peuvent être divisés en deux catégories : ceux qui ne fournissent que de la synthèse visuelle et ceux qui sont capables de générer de la synthèse multimodale. Les systèmes qui ne génèrent que de la synthèse visuelle sont les suivants : les modèles basés visèmes décrits précédemment, les modèles basés sur des théories de coarticulation, modèles de génération de trajectoires et modèles ad-hoc. Les systèmes qui peuvent générer de la synthèse multimodale sont : les modèles basés sur le principe de concaténation et les modèles basés HMM.

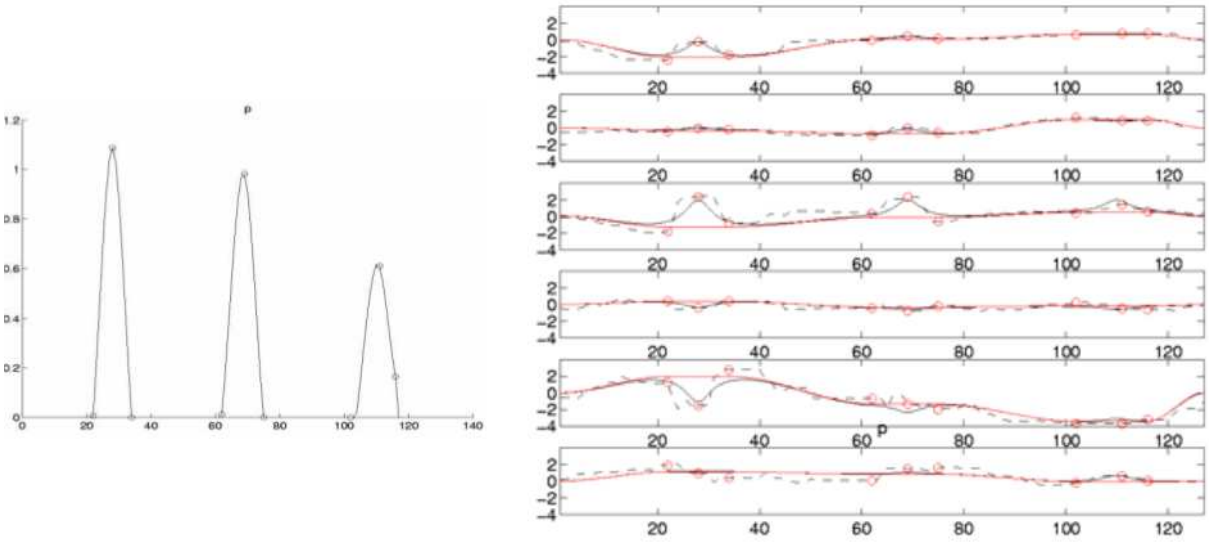


FIG. 2.15 – Synthèse basée sur le modèle d'Öhman (Öhman, 1967). A gauche : Fonctions d'émergence *kc(t)* pour [p] en [apa] [ipi] [upu], à droite : Synthèse des 6 paramètres articulatoires à partir du texte : [apa] [ipi] [upu]. Du haut en bas : ouverture de la mâchoire, protrusion des lèvres, fermeture des lèvres, montée des lèvres, avancement de la mâchoire, mouvements du larynx. Les pointillés correspondent aux mouvements captures, les lignes en rouge - gestes vocaliques et les lignes en noir aux gestes finaux.

Modèles de coarticulation Une approche pour modéliser la parole visuelle à partir des données est de paramétrer les modèles de coarticulation (Cohen & Massaro, 1993), (Öhman, 1967), (Cosi *et al.* , 2002). Le Goff et Benoit (LeGoff & Benoit, 1996) ont ainsi effectué l'apprentissage du modèle de dominance de Cohen et Massaro (Cohen & Massaro, 1993) appliqué aux données d'un locuteur de langue française. Les phones sont classées en 19 visèmes et les mouvements faciaux sont contrôlés par 8 paramètres (Le Goff *et al.* , 1994). Chaque fonction de dominance a trois coefficients. Des problèmes de modélisation sont notamment rencontrés dans le cas des consonnes bilabiales et labiodentales : le modèle de coproduction ne garantit pas la fermeture complète des contacts (entre lèvres ou entre dents et lèvre supérieure).

Le modèle de coarticulation d'Öhman (Öhman, 1967) a été aussi utilisé par l'équipe de l'ICP (Revéret *et al.* , 2000) pour faire de la synthèse de la parole (voir la Figure 2.15 et l'équation 2.1). Les valeurs de gestes vocaliques et de cibles de consonnes ainsi que les fonctions d'émergence et

de coarticulation sont estimées à partir d'un corpus de 24 séquences VCV (8 consonnes dans 3 contextes vocaliques symétriques).

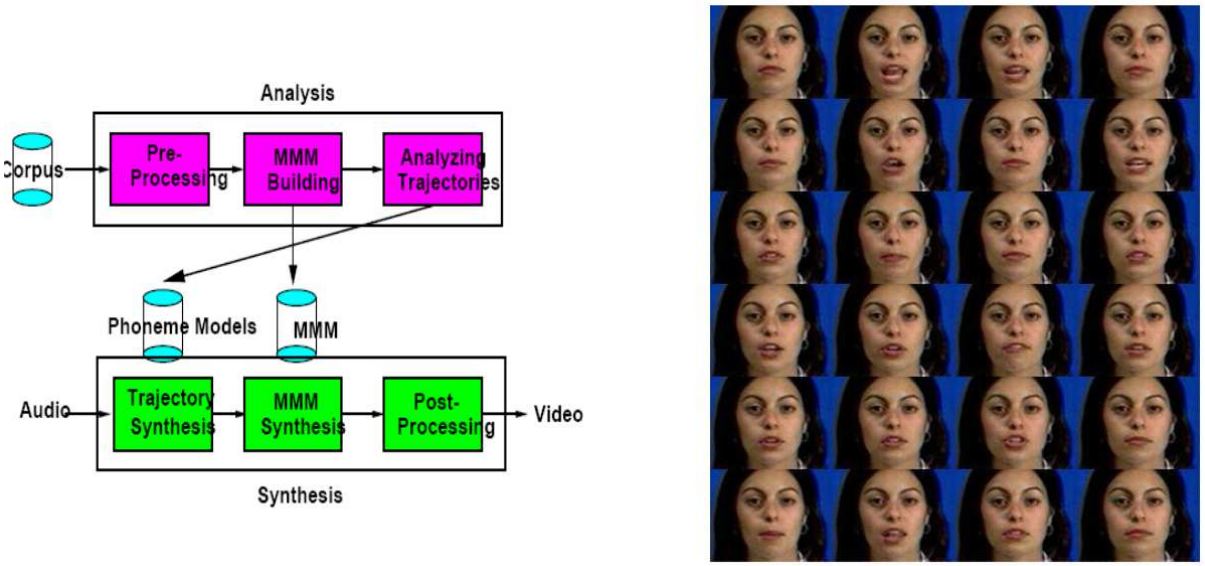


FIG. 2.16 – Modèle de synthèse proposé par T. Ezzat (Ezzat *et al.* , 2002). A gauche : schéma d'analyse et de synthèse de "Mary101", à droite : 24 des 46 images prototypiques constituant le MMM.

Modèles de génération de trajectoires Dans le groupe de modèles de génération de trajectoires, j'ai classé les modèles qui considèrent la production des trajectoires articulatoires comme un problème général de génération de trajectoires. L'exemple typique d'un tel modèle est le modèle *"Mary101"* proposé par Ezzat et al. (Ezzat *et al.* , 2002) (Figure 2.16) qui produit de la synthèse basée-image réaliste et proche du naturel (Geiger *et al.* , 2003).

Le modèle proposé par Ezzat et al. est le MMM (*Morphable Multidimensional Model*) qui est capable de synthétiser une nouvelle image des lèvres à partir d'un ensemble d'images prototypes (46 images). Ces images prototypes sont choisies grâce à un algorithme de *k-moyennes* à partir de l'ensemble des images du corpus. Chaque image est présentée par 2*46 paramètres (46 paramètres pour la forme des lèvres et 46 paramètres pour sa texture). Ensuite, le problème de synthèse des trajectoires des paramètres est vu comme un problème de régularisation (Girosi *et al.* , 1995). Dans le système, chaque phone est représenté par une distribution gaussienne des paramètres articulatoires avec une moyenne et une matrice de covariance qui sont calculées directement à partir de données. L'effet de coarticulation est implicitement lié aux valeurs de la covariance et aux durées de chaque phone. Pour déterminer la trajectoire de synthèse finale, Ezzat et al. minimisent une somme d'un terme cible et d'un terme lissant. Une fois les trajectoires obtenues, les valeurs de la covariance et de la moyenne des gaussiennes sont ajustées. Pour cela, l'erreur euclidienne entre les trajectoires synthétiques et originales est minimisée.

L'avantage de ce type de modèles est le fait qu'ils produisent de la synthèse visuelle très proche du réel car ils sont basés sur des techniques de minimisation qui garantissent une prédiction optimale. De plus, cette optimisation est faite sur des énoncés complets et ne limite donc pas les effets coarticulatoires à un contexte local.

Modèles de concaténation Par analogie avec le domaine de synthèse vocale où les approches par concaténation sont prédominantes, les mêmes méthodes sont proposées pour la synthèse visuelle de la parole. La technique de synthèse par concaténation est basée sur les principes suivants, (voir Figure 2.1) : tout d'abord, on dispose d'un dictionnaire de segments multi-représentés, ensuite les coûts de sélection et de concaténation sont calculés entre les segments candidats, enfin les segments finaux sont choisis grâce au calcul d'un chemin de coût minimal. Les coûts de sélection sont, généralement, calculés en fonction des distances phonologiques et grâce à des cibles paramétriques calculées par un modèle externe, - le modèle prosodique pour la synthèse vocale. Un post-traitement est aussi souvent appliqué sur les trajectoires obtenues, notamment pour assurer la continuité des trajectoires aux points de concaténation.

A l'image de ce qui a été proposé en synthèse audio, la synthèse de concaténation la plus simple consiste à concaténer des segments de vidéos. Les systèmes proposés par (Weiss, 2005) et (Fagel, 2006) proposent aussi de concaténer des segments correspondant à des diphones. Le problème majeur de cette approche réside dans la difficulté de piloter séparément les mouvements de la tête et du visage. Une solution consiste à enregistrer des locuteurs dont la tête est immobile et le fond constant.

Bregler et al. (Bregler, 1997) proposent un système de synthèse *" Video Rewrite"* qui fait la distinction visage/fonds et où les unités de concaténation sont des triphones. Le visage est extrait par une technique de suivi d'un masque (excluant les parties mobiles, yeux et bouche et limitant les mouvements à des translations et rotations planes) initialisé sur une image. Pendant la phase d'analyse, le système utilise la partie audio pour segmenter la vidéo en triphones. Les techniques de vision (Lanitis *et al.* , 1995) permettent de trouver l'orientation de la tête, les formes et les positions de la bouche et de la mâchoire de chaque image. Ensuite pendant la phase de synthèse le système segmente l'audio et l'utilise pour sélectionner les triphones précédemment extraits de la vidéo. Les nouvelles images de la bouche sont ajustées aux images de fond par déformation élastique (Beier & Neely, 1992). *Video Rewrite* étiquette automatiquement les phones pendant la phase d'analyse et pendant la phase de synthèse. L'étiquetage s'effectue par alignement forcé des phones modélisés hors-contexte (Rabiner, 1989). Chaque phone est modélisé avec une chaîne de Markov à trois états. Les auteurs remarquent cependant que avec ce modèle, il y a un problème de synchronisation entre la vidéo et l'audio, notamment, pendant les plosives dont les mouvements ne sont pas synchronisés avec l'audio. Nous reviendrons sur ce problème de synchronisation audiovisuelle dans le chapitre 5. .

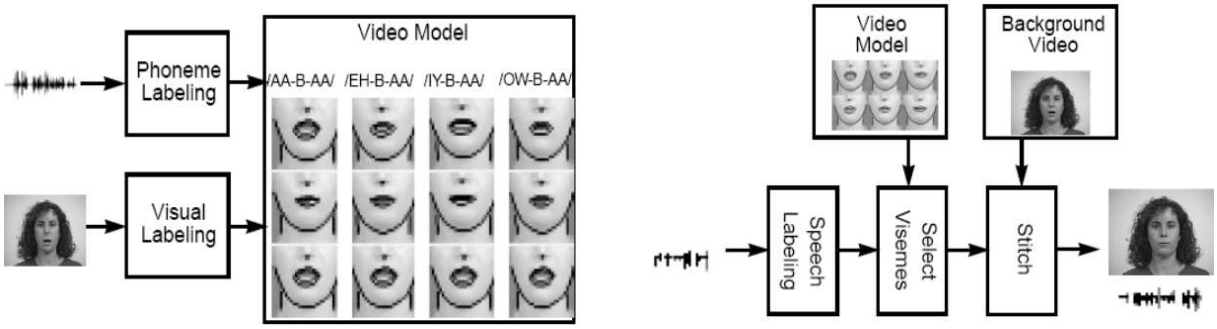


FIG. 2.17 – Schéma d'analyse et de la synthèse de *VideoRewrite* (Bregler, 1997). A gauche : principe de construction du modèle vidéo, à droite : principe de synthèse à partir de l'audio.

Hallgren et Lyberg (Hallgren & Lyberg, 1998) proposent aussi d'utiliser le principe de concaténation pour les systèmes basés modèle. Le système utilise des demi-syllabes comme unités principales pour mieux représenter le processus de coarticulation de la langue suédoise. Les pa-

ramètres visuels sont les trajectoires des marqueurs placés sur un visage et captées par un système optique. Les mouvements enregistrés des marqueurs sont concaténés, et ensuite interpolés.

Minnis et Breen (Minnis & Breen, 1998) considèrent les systèmes de concaténation des mouvements visuels comme une extension de la synthèse vocale par concaténation., c'est-à-dire que la synthèse visuelle est considérée comme un processus de sélection des unités. Généralement, dans la synthèse vocale par sélection des unités, les N-phones sont choisis et déformés en se basant sur des critères linguistiques. Les auteurs proposent d'utiliser les mêmes critères de sélection pour la synthèse visuelle.

Un autre modèle qui modélise les trajectoires des diphones et des triphones est proposé par Deng (Deng *et al.* , 2005). Le principe est le suivant : les modèles des trajectoires (des splines) des diphones et des triphones (appelés les modèles de coarticulation) sont appris à partir des données capturées, ensuite, lors de la synthèse ces trajectoires (des splines) sont concaténées en fonction de la suite phonétique en entrée. Les trajectoires obtenues représentent les mouvements faciaux liés à la parole neutre. De plus, les expressions émotionnelles peuvent être ajoutées aux trajectoires de coarticulations neutres.

FIG. 2.18 – Quelques trames de la parole synthétique (Deng *et al.* , 2005).

Les systèmes de synthèse par concaténation ont plusieurs avantages : tout d'abord ce type de systèmes peut être utilisé dans la synthèse multimodale, ensuite le système concatène des segments qui sont déjà synchrones, donc, la synchronie entre les paramètres est respectée et, enfin, le résultat est très proche du réel car les segments viennent d'une base de données et donc gardent les détails d'articulation. Par contre, cette méthode a aussi plusieurs inconvénients : d'une part, la qualité de la synthèse est proportionnelle à la quantité de données, donc, cette technique demande des grandes bases de données, et, d'autre part, la concaténation des segments finaux n'est pas triviale, ce qui peut donner des résultats non continus et trop saccadés.

Modèles basés HMM La synthèse par HMM a été proposée pour la première fois par Donovan pour la synthèse acoustique (Donovan, 1996), ensuite ce principe est appliqué à la synthèse audiovisuelle par le groupe de travail HTS (*HMM-based Speech Synthesis System*) (Tamura *et al.* , 1999).

Le système de synthèse par HMM comprend deux étapes principales : l'étape d'apprentissage des paramètres des modèles HMM d'un segment de la parole et des modèles des durées d'états ; et l'étape de synthèse des paramètres d'observation (acoustiques, visuels ou audiovisuels) à partir d'une séquence des HMMs concaténés (Figure 2.19). Pendant l'étape d'apprentissage, un HMM (grâce à une estimation basée ML : *Maximum-Likelihood*) est appris pour chaque segment phonétique sans ou avec contexte (Figure 2.20). Les vecteurs d'apprentissage comprennent des paramètres visuels et leurs dérivés, ce que l'on appelle les paramètres statiques et les paramètres dynamiques. Un modèle des durées d'états est également associé à chaque segment phonétique sans ou avec contexte (Yoshimura *et al.* , 1998). Le plus souvent, les modèles des

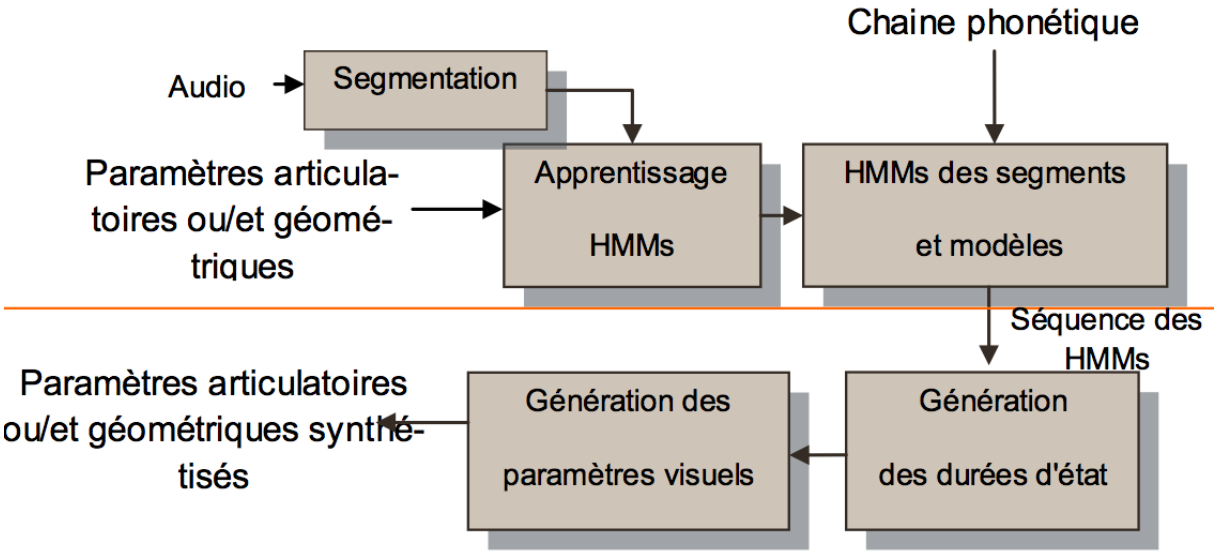


FIG. 2.19 – Principe de synthèse par HMM.

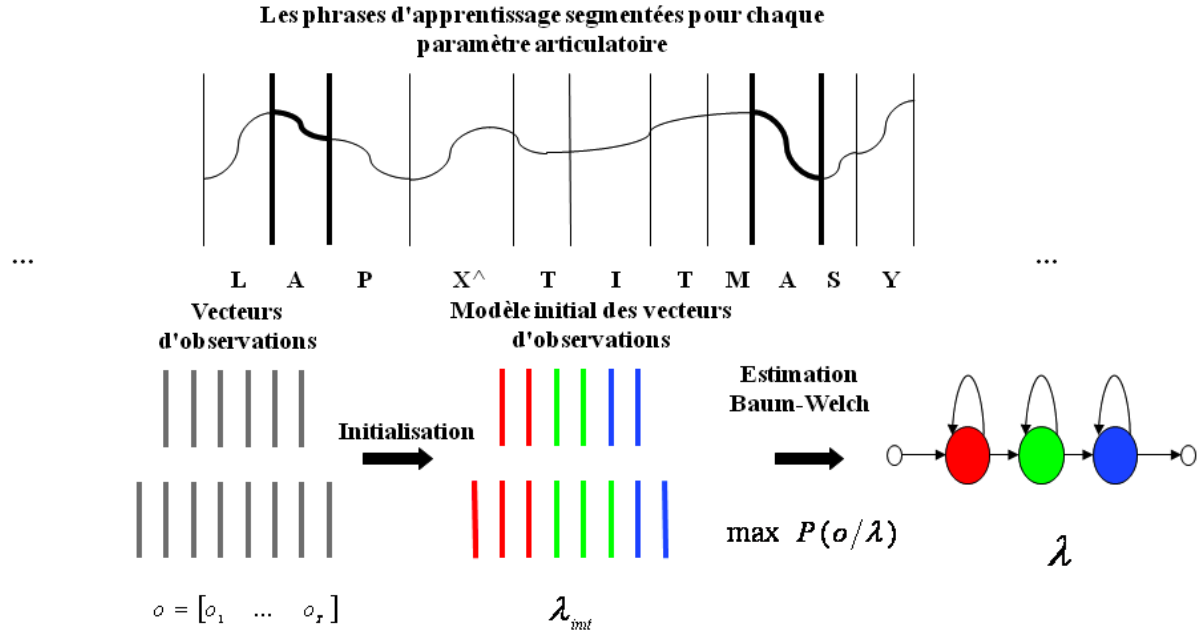


FIG. 2.20 – Principe d'apprentissage des HMM par segment. Dans cet exemple, Collecte des données puis apprentissage de mode hors contexte.

durées d'états correspondent à des modèles gaussiens où la dimension du vecteur correspond au nombre d'états d'un HMM correspondant. C'est analogue au modèle d'élasticité (basé z-score[2]) proposé par Campbell et Isard (Campbell & Isard, 1991). Pendant la phase de synthèse, les HMM correspondant à la séquence des segments sont concaténés, ensuite la trajectoire de synthèse est obtenue grâce aux durées des segments de synthèse et à un algorithme de génération (basé ML) qui est basé sur la connaissance de la relation entre les paramètres statiques et les paramètres dynamiques (Figure 2.21).

Il existe des nombreux travaux sur la synthèse vocale ou audiovisuelle de la parole par HMM dans les laboratoires japonais : (Tokuda *et al.* , 2000), (Zen *et al.* , 2004), (Tamura *et al.* , 1998), (Tamura *et al.* , 1999).

La synthèse par HMM a plusieurs avantages : tout d'abord, cette technique peut être aussi utilisée dans la synthèse multimodale, ensuite cette méthode est une technique permettant de paramétriser complètement et ceci avec un nombre constant de paramètres la chaîne de synthèse. On peut donc aisément faire des opérations sur les paramètres de ces modèles, y compris effectuer des analyses statistiques. Par exemple, dans la synthèse vocale par HMM, des méthodes sont proposées pour changer automatiquement la vitesse de locution, le locuteur, les émotions, la langue (Tachibana *et al.* , 2005), (Nose *et al.* , 2007). L'inconvénient de la synthèse par HMM est le fait que le résultat est moyenné, donc, au final, les trajectoires articulatoires sont correctes à long terme mais sont lissées, ce qui est le contraire de la synthèse par concaténation qui garde les détails de l'articulation. Toda et Tokuda (Toda & Tokuda, 2007) ont récemment proposé une solution possible à ce problème avec une méthode considérant la variance globale des trajectoires obtenues.

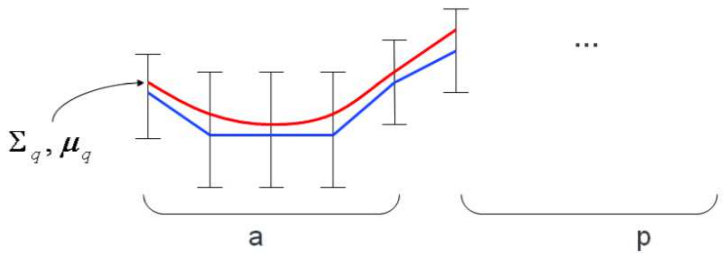


FIG. 2.21 – Principe de génération des trajectoires finales à partir des HMM.

Synthèse à partir de l'audio

L'avantage des systèmes de synthèse à partir de l'audio est le fait qu'ils génèrent directement les paramètres visuels à partir des paramètres acoustiques sans passer par la phase de la segmentation phonétique, l'inconvénient est le fait que la relation acoustique-visuelle est une relation du type *many-to-many*.

Modèles de correspondance directe Les méthodes d'apprentissage qui sont à la base des approches par régression sont fondées sur des algorithmes continus dont le but est d'étudier les relations complexes entre les paramètres acoustiques et visuels. Dans un premier temps on extrait

[2]Le z-score de la durée d'un état (ij, où i :segment, j :état) est un facteur à appliquer à la déviation standard σ_{ij} et sommer avec la moyenne μ_{ij} de la durée de cet état pour obtenir la somme totale des durées d'états donnée d'un segment dur_i. $dur_{ij} = \mu_{ij} + z - score_{ij} * \sigma_{ij}$, $dur_i = \Sigma dur_{ij}$

des paramètres acoustiques (Kakumanu *et al.* , 2002) et articulatoires associés à chaque trame des séquences étudiées. Ensuite, des méthodes d'apprentissage permettent de construire un modèle faisant correspondre les deux types de paramètres. Pour extraire les caractéristiques acoustiques, un prétraitement est d'abord effectué (débruitage,...) sur les séquences audio (Kakumanu *et al.* , 2001). Le signal acoustique est ensuite divisé en trames de longueur assez courte (10-20 msec) pour pouvoir être considérées comme quasi-stationnaires. Chaque trame est fenêtrée pour éviter les distorsions spectrales. Ensuite, une série de paramètres acoustiques est extraite. Généralement trois types de paramètres sont utilisés :

- les caractéristiques du système de production de la parole (*Linear Predictive Coding :* LPC*, Line Spectral Frequencies :* LSF ou *Line Spectral Pairs :* LSP),
- les caractéristiques prosodiques de la parole (énergie, fréquence fondamentale, ...),
- coefficients cepstraux, *Mel-Frequencies Cepstral Coefficients :* MFCC*, ...*.

Yehia et al. (Yehia *et al.* , 1998) étudient un modèle de régression linéaire pour décrire les associations entre 11 paramètres acoustiques (10 LSP et amplitude) et 12 (ou 18) paramètres faciaux. Si le travail est effectué sur un nombre limité de phrases et avec le même sujet, on obtient un coefficient moyen de corrélation de 0.7. Cela montre qu'une redondance d'information existe. Avec un modèle non linéaire et des informations sur le contexte, les résultats peuvent encore être améliorés.

Okadome et al. (Okadome *et al.* , 2000) utilisent 30 coefficients LPC pour coder chaque fenêtre des trames acoustiques. Une autre façon de présenter ces coefficients LPC est de les transformer en coefficients LSF ou LSP. Yehia et al. (Yehia *et al.* , 1998) travaillent avec 12 coefficients LSP qui sont associés aux trames acoustiques fenêtrées. Dans leur étude ces derniers sont préférés aux LPC en raison de leur meilleure interpolation temporelle et à cause du fait que ces coefficients sont mieux corrélés aux fréquences résonantes (Schroeder, 1967). Dans de nombreux travaux, les coefficients cepstraux sont utilisés (Massaro *et al.* , 1999), (Curinga *et al.* , 1996), (Hong *et al.* , 2002).

En ce qui concerne la modélisation, des modèles linéaires (Yehia *et al.* , 1998) ou non-linéaires sont utilisés (Massaro *et al.* , 1999), (Curinga *et al.* , 1996), (Hong *et al.* , 2002), (Kakumanu, 2003). Les réseaux connexionnistes sont aussi utilisés car ils sont bien adaptés à ce type d'apprentissage : ils peuvent modéliser les associations non linéaires et leurs couches cachées sont capables de modéliser les relations complexes entres les entrées et les sorties (Beskow, 2003).

Berthommier (Berthommier, 2003) propose un modèle linéaire de transformation des paramètres acoustiques en paramètres visuels. Les paramètres acoustiques sont 16 paramètres MFCC extraits avec la fréquence de 50 Hz dans une fenêtre de 40 ms. Les paramètres visuels sont des paramètres DCT (*Discrete Cosinus Transformation*) extraits à partir des images RGB sur la région de la bouche avec une fréquence de 50 Hz. Le modèle linéaire proposé est une matrice de transformation linéaire. Des exemples des images générées sont représentés dans la Figure 2.22.

Lavagetto (Curinga *et al.* , 1996) utilise des réseaux connexionnistes à quatre couches avec des retards (TDNN *: Time Delay Neural Network*s) pour modéliser 4 paramètres labiaux à partir de 12 coefficients cepstraux. Pour chaque paramètre facial, un TDNN séparé est construit. Cette étude prend en compte non seulement la trame courante (20 ms) en entrée mais aussi les 5 trames précédentes et les 5 trames suivantes pour prendre en compte le phénomène de coarticulation.

Massaro et al. (Massaro *et al.* , 1999) utilisent aussi des TDNN à trois couches pour modéliser 39 paramètres de contrôle à partir de 13 coefficients cepstraux en prenant en compte le contexte

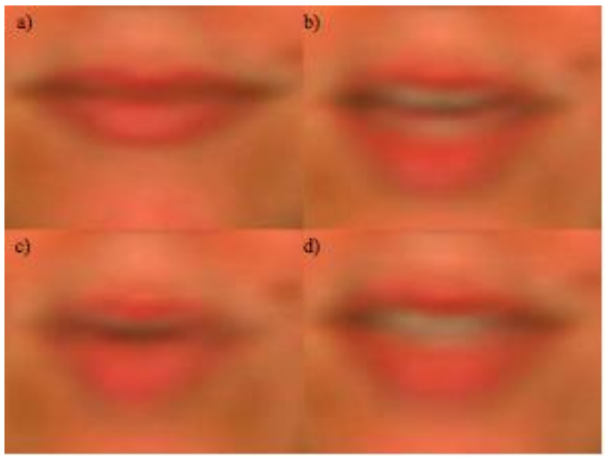


FIG. 2.22 – Exemples des trames générées à partir de l'audio grâce au modèle de transformation linéaire proposé par (Berthommier, 2003).

(les 5 trames précédentes et les 5 trames suivantes de la trame courante). Le premier apprentissage est fait sur 400 mots et avec 600 neurones cachés. Dans ce cas, la corrélation moyenne obtenue est de 0.77 pour l'ensemble d'apprentissage et de 0.64 pour l'ensemble de test.

Hong et al. (Hong *et al.* , 2002) utilisent aussi des réseaux connexionnistes pour modéliser les paramètres faciaux à partir de 10 paramètres cepstraux. Dans un premier temps les trames audio sont classées en 44 sous-ensembles. Chaque sous-ensemble représente un modèle gaussien. Ensuite, un apprentissage par un TDNN à trois couches avec le contexte (7 trames) est effectué pour chaque sous-ensemble. Le corpus enregistré est de 100 phrases. Le coefficient de corrélation sur les données de test est de 0.974. Le test montre que les résultats de synthèse sont dépendants du corpus et de la langue.

Arslan and Talkin (Arslan & Talkin, 1998) construisent une table audiovisuelle pour estimer les trajectoires des points faciaux en fonction de la parole. Cette table est constituée des paires "paramètres acoustiques - paramètres faciaux". Un nouveau vecteur acoustique est comparé aux paramètres acoustiques de la table, et la partie visuelle correspondante est calculée comme une somme pondérée des paramètres faciaux où les poids sont calculés en fonction de la similarité acoustique. L'apprentissage est effectué sur un corpus de dix minutes et le coefficient de corrélation est estimé pour évaluer cette méthode. La corrélation sur les données d'apprentissage est de 0.92 et sur les données de test de 0.73.

Kakumanu (Kakumanu, 2003) dans son travail de thèse compare les différentes méthodes d'apprentissage : réseaux RBF (*Radial Basis Functions*), réseaux I-RBF (*Incremental Radial Basis Functions*), SVM (*Support Vector Machines*), les chaînes de Markov et KNN (*K-Nearest Neighbor*). Trois paramètres faciaux obtenus grâce à PCA sont modélisés en fonction de 10 paramètres acoustiques obtenus grâce à l'analyse linéaire discriminant de la combinaison des coefficients LPC, LSF, MFCC, PCBF, de l'énergie et de la fréquence fondamentale. La comparaison est faite par rapport aux coefficients moyens de corrélation et à l'erreur moyenne entre les trajectoires originales et celles de synthèse. En conclusion, les réseaux RBF, SVM et KNN donnent les meilleurs résultats, suivis par les réseaux I-RBF et enfin, suivent les chaînes de Markov.

De nouvelles techniques d'estimation non-linéaires basées sur les systèmes de conversion de voix par modélisation statistique semblent maintenant prometteuses. Toda et al. (Toda *et al.* , 2008) ont notamment proposé d'utiliser des modèles de mixtures de Gaussiennes (GMM)

pour apprendre la correspondance entre signaux et articulation. Le signal est caractérisé par la projection sur les premiers plans factoriels d'une ACP de paramètres acoustiques collectés sur une large fenêtre (>100ms) centrée sur la trame articulatoire courante. Les expériences sont menées sur la base MOCHA d'articulations linguales.

Synthèse à partir de l'audio avec de l'information phonétique Dans le cas de l'apprentissage à partir de l'audio augmenté de l'information sur la chaîne phonétique, le signal audio est d'abord représenté sous une forme discrète intermédiaire, laquelle est ensuite transformée en paramètres visuels.

Okadome et al. (Okadome *et al.* , 2000) utilisent le principe de table de correspondances et améliorent la synthèse en ajoutant de l'information phonétique. Dans en premier temps, une table de paires "12 paramètres LPC - 9 paramètres articulatoires" est construite. Ensuite, pour un nouveau vecteur acoustique une zone de recherche des paramètres acoustiques qui lui sont les plus similaires est calculée grâce aux calculs de la distance spectrale. Dans un deuxième temps l'information phonétique est ajoutée : une trajectoire de synthèse d'accélération minimale est calculée grâce au modèle cinématique de triphones (Okadome *et al.* , 1999). Dans la zone de recherche une trajectoire de synthèse finale qui minimise la distance entre elle et la trajectoire d'accélération minimale est choisie. Cette méthode est évaluée en comparant les erreurs de synthèse avec ou sans l'information phonétique et avec les différentes méthodes de calcul des temps d'articulation des phones. Le meilleur résultat (l'erreur de 1.6 mm) est atteint dans le cas de la synthèse à partir de l'audio avec de l'information phonétique et avec les temps d'articulation observés. Des erreurs de 1.8 mm sont obtenues dans le cas de la synthèse sans l'information phonétique et avec de l'information phonétique mais avec les temps d'articulation estimés.

Hiroya et Honda (Hiroya & Honda, 2004) proposent aussi un système de synthèse des trajectoires articulatoires en se basant sur les données acoustiques et phonétiques. Leur modèle de production de la parole est basé sur les HMM. Une chaîne de Markov à trois états est associée à chaque phone. L'état représente un paramètre articulatoire auquel un paramètre acoustique est associé grâce à une fonction linéaire. Pendant la phase de synthèse, dans un premier temps les paramètres acoustiques de chaque trame sont calculés. Ensuite, une séquence d'états de HMM optimale pour une séquence acoustique donnée est estimée en utilisant l'algorithme de Viterbi (Cornuéjols & Miclet, 2002). Enfin, les transitions entre les moyennes des états sont lissées en utilisant l'approche de Tokuda et al. (Tokuda *et al.* , 2000).

Tamura et al. (Tamura *et al.* , 1998) modélisent les syllabes comme des séquences des états HMM. Pendant la phase d'apprentissage un vecteur des paramètres audio-visuels, qui est constitué des paramètres cepstraux et leurs dérivées, ainsi que des paramètres faciaux et leurs dérivées, est défini. Ensuite à partir des ces vecteurs audio-visuels les séquences HMM associées à chaque syllabe sont construites. La phase de synthèse comprend deux étapes : l'étape d'identification pendant laquelle les séquences des syllabes sont obtenues grâce aux calculs des HMMs les plus similaires des données acoustiques, et l'étape de synthèse des paramètres visuels à partir des séquences HMM générées. Le corpus est constitué de 216 mots. Les données de validation comprennent 5 mots et une phrase. Les tests DMOS sont utilisés *(Degradation Mean Opinion Score)* pour évaluer la qualité (Klaus *et al.* , 1993) de synthèse à partir d'information acoustique et à partir de l'information phonétique avec ou sans les dérivées des paramètres audio et visuels. Quasiment les mêmes résultats sont obtenus dans les cas de synthèse à partir de l'audio ou à partir du texte avec l'utilisation des dérivées (3.62 sur 5), par contre, sans les dérivées le coefficient DMOS est de 2.1 sur 5.

2.4 Problématique de l'évaluation

La question de l'évaluation des résultats de la synthèse audiovisuelle est très importante. Ce sont les tests d'évaluation objective et subjective qui permettent de dire comment les systèmes de synthèse répondent au cahier de charges prévu. Avant de commencer la construction d'un système de synthèse de la parole, il serait idéal de comparer les systèmes existants de l'état de l'art. La comparaison de ces systèmes est problématique (Beskow, 2003) car les modèles sont construits à partir de différents corpus (locuteur, langue, conditions d'enregistrement, corpus, etc.) et leurs méthodologies d'évaluation sont différentes. De plus, l'approche modulaire est rarement possible car les modèles de contrôle, de forme et d'apparence sont souvent très liés. La qualité du rendu influence aussi beaucoup la qualité des modèles de contrôle (Pandzic *et al.* , 1999). Actuellement, deux types d'évaluation sont utilisés : l'évaluation dite objective et l'évaluation perceptive ou subjective. L'évaluation objective comprend généralement l'erreur RMS *(Root Mean Square)* et/ou la corrélation entre les paramètres de synthèse et les originaux. Ceci peut être effectué au niveau des paramètres articulatoires, des coordonnées des points 3D décrivant la géométrie faciale ou même des pixels de l'image finale produite comme cela est possible dans les AAM (on parle alors de PSNR). L'évaluation subjective comprend les tests sur l'intelligibilité, sur le réalisme et sur la reproductibilité de l'effet McGurk (McGurk & MacDonald, 1976). Dans ce qui suit les différents travaux sur l'évaluation des visages parlants sont présentés.

Pandzic et al. (Pandzic *et al.* , 1999) font une série d'expérimentations pour évaluer l'utilité potentielle de l'animation faciale dans les services interactifs combinée avec un TTS. Les études objectives et subjectives sont faites sur 190 personnes. Les visages parlants utilisés sont de trois types : un visage d'un maillage polygonal 3D, un visage avec une texture collée et un visage basé image (Cosatto & Graf, 2000). Le modèle de contrôle est le modèle de dominance de Cohen&Massaro. La conclusion générale est que l'animation faciale n'améliore pas (en moyenne) la compréhension de l'audio d'où la nécessité de progresser dans le domaine de la modélisation visuelle de la parole (l'année 1999).

Engwall (Engwall, 2002) évalue objectivement un système de synthèse par concaténation des paramètres linguaux 3D. Les segments de concaténation sont des diphones. Le corpus utilisé est la base de données MOCHA-TIMIT[3] qui est faite sur 40 sujets prononçant 460 phrases phonétiquement équilibrées. Les données de synthèse sont comparées avec des données réelles obtenues avec un EMA (*Electromagnetic articulographe*) et à partir des images radiographiques. Le modèle reproduit globalement les mouvements naturels et restreint la synthèse des mouvements locaux comme, par exemple, ceux du bout de langue.

Yamamoto et al. (Yamamoto *et al.* , 1998) effectuent des évaluations objectives et subjectives sur des modèles de contrôle basés sur l'information acoustique (modèle de régression et HMM). D'après les évaluations objectives les HMM donnent de meilleurs résultats que les modèles de régression. Les évaluations subjectives sont : le test d'intelligibilité et le test d'acceptabilité. Les tests sont effectués sur 10 personnes. Les évaluations subjectives ne montrent pas de grandes différences entre les différents modèles. Cela provient peut-être des conditions expérimentales des essais (le nombre limité de sujets, le nombre limité de phrases de test, ...).

Geiger et al. (Geiger *et al.* , 2003) effectuent une série très complète d'évaluations subjectives sur le réalisme et l'intelligibilité du système *"Mary"* de T. Ezzat (Ezzat *et al.* , 2002) avec la participation de 24 sujets. Le test de Turing est effectué pour distinguer les images réelles et celles animées avec ou sans l'audio. En moyenne, les sujets ne font pas de différence entre les images de synthèse et celles réelles. Pour évaluer l'intelligibilité du système, les sujets doivent lire

[3]http ://www.cstr.ed.ac.uk/artic/mocha.html

sur les lèvres de l'image de synthèse. Contrairement au modèle naturel, le taux de reconnaissance des mots et des phones est beaucoup plus faible dans le modèle de synthèse.

Beskow (Beskow, 2004) compare quatre différents modèles de contrôle, qui sont modélisés avec le même corpus, en les évaluant objectivement et subjectivement. Les quatre modèles sont : le modèle de dominance de Cohen&Massaro, le modèle d'Öhman, un modèle ANN avec un contexte symétrique de 15 trames et un modèle ANN avec un contexte de 2 trames précédentes et des 28 suivantes de la trame courante. D'après les évaluations objectives (l'erreur RMS et la corrélation) les différences entre les modèles sont non significatives (l'ensemble de test est de 89 phrases). Le test d'intelligibilité donne le même résultat, qui est la quasi égalité de tous les modèles, et une intelligibilité faible (l'identification correcte est de 74%). Ce test est aussi effectué sur un modèle basé règles (Beskow, 1995) et l'identification correcte obtenue est alors de 81%. Les modèles basés règles sont développés dans le but d'avoir une articulation claire et une intelligibilité forte, alors que ceux basés données sont construits dans le but de reproduire un style de production de la parole d'un sujet.

Bailly et al. (Bailly *et al.* , 2002) proposent d'utiliser un modèle d'apparence par des points lumineux (Figure 2.23a) pour s'affranchir de l'influence de la qualité du modèle d'apparence sur l'appréciation du modèle de contrôle. Ils évaluent objectivement (corrélation) et subjectivement (MOS test : *Mean Opinion Score*) quatre différents modèles de contrôle (Figure 2.23) : deux modèles de concaténation de diphones (*Syn et Sinl*), un modèle par règles utilisant le modèle d'Öhman (*Reg*) et un modèle de régression linéaire (avec des données de test comprises dans les données d'apprentissage *Mltst* et avec des données de test les mêmes que les données d'apprentissage *Mlapp*). D'une part, le test d'acceptabilité montre que le modèle *Synl* donne le meilleur résultat suivi par les *Reg* et *Syn*. Les modèles linéaires sont jugés comme non acceptables. D'autre part, le modèle *Mlapp* a la meilleure corrélation suivi par les *Syn* et *Synl*. Le modèle *Reg* a un coefficient de corrélation très faible. Les évaluations objectives et subjectives ne donnent donc pas les mêmes résultats. Cela montre que la perception audiovisuelle est très sensible aux passages sur les valeurs cibles des paramètres articulatoires. Ces phases de passage sont préservées dans les modèles de concaténation, simplifiées dans les modèles de coarticulation, et les modèles de régression linéaire ne permettent pas de les atteindre.

Dans beaucoup de cas, l'évaluation subjective des systèmes donne des résultats très pauvres par rapport au bon réalisme obtenu. Cette contradiction peut être expliquée (Odisio & Bailly, 2004) par la complexité accumulée à par de chaque module de la chaîne de synthèse audiovisuelle. Les mouvements corrects peuvent être jugés inadéquats si on a un mauvais modèle d'apparence. De même, un mauvais modèle de contrôle engendre des mouvements qui peuvent être jugés comme non acceptables.

En conclusion, il est difficile de déterminer le meilleur modèle de contrôle d'animation lié à la parole à partir des tests d'évaluations disponibles dans la littérature. Il faut tout de même souligner l'importance des tests subjectifs, ce sont eux qui permettent d'évaluer le résultat final d'un système de synthèse pour une application donnée. Par la suite, nous proposons d'évaluer les principaux modèles de contrôle d'état de l'art avec le corpus I. Cette étude diffère de celle d'Odisio et al. (Odisio & Bailly, 2004) par le fait qu'un modèle de technique vidéoréaliste est ici utilisé et que nous avons réimplementé les systèmes testés - notamment HMM.

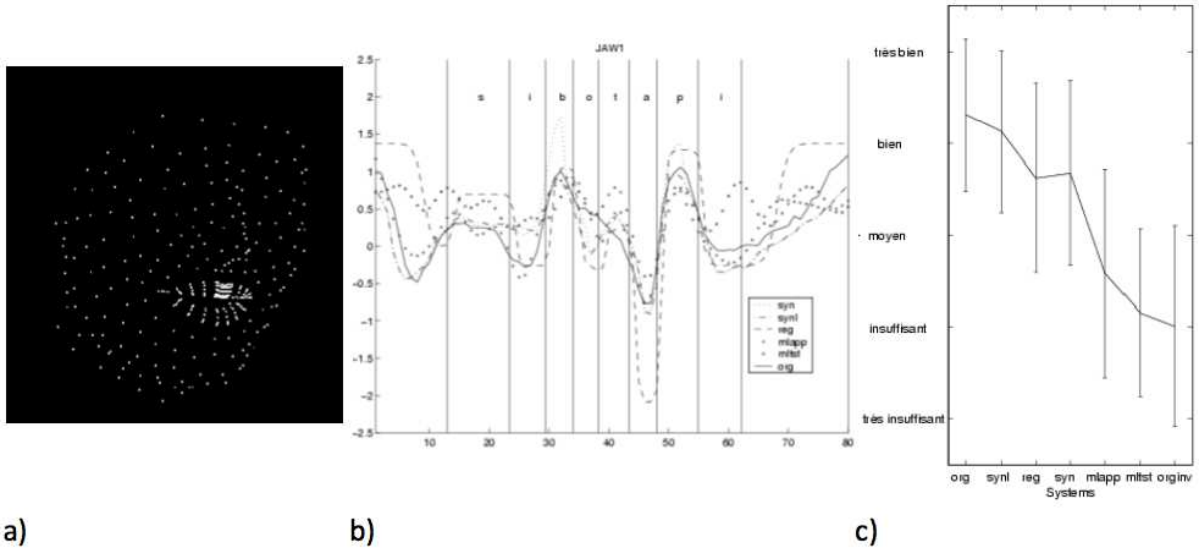


FIG. 2.23 – Evaluations objectives et subjectives par Bailly et al. (Bailly *et al.* , 2002) des systèmes de synthèse visuelle (concaténation sans et avec lissage *Syn* et *Synl*, régression linéaire pour les données d'apprentissage et de test *Mlapp* et *Mltst*, modèle d'Ohman *Reg*). a) modèle faciale en *Point Lights* ; b) exemple de la synthèse du paramètre articulatoire Jaw1 mouvements de la mâchoire pour la phrase "Six beaux tapis" ; c) résultats du test MOS pour les différents modèles.

2.5 Evaluation des modèles de l'état de l'art

Le corpus I de 238 phrases est utilisé dans cette étude. 228 phrases sont utilisées dans l'apprentissage des modèles et 10 phrases sont utilisées dans les tests. Les 10 phrases de test sont choisies pour que tous leurs diphones soient présents au moins une fois dans le corpus d'apprentissage. Les paramètres visuels utilisés sont les sept paramètres articulatoires. Le travail réalisé a été présenté et publié aux Journées d'Etudes sur la Parole (JEP2006) (Govokhina *et al.* , 2006a).

2.5.1 Modèles de contrôle utilisés

Nous avons implémenté les modèles de contrôle suivants : modèle de régression linéaire entre les paramètres acoustiques et paramètres articulatoires, modèle basé HMM, concaténation simple et concaténation guidée par les HMM.

Le modèle linéaire guidé par l'acoustique

Les paramètres articulatoires (120 Hz) sont calculés à partir des paramètres acoustiques (120 Hz) grâce au modèle linéaire de correspondance directe. Les signaux de 16.7 ms (fenêtre) sont extraits à la fréquence de 120 Hz à partir du signal d'origine en synchronie avec les données visuelles. Douze paramètres LSP (*Line Spectrum Pairs*) et l'énergie sont calculés et lissés (Yehia *et al.* , 1998). Chaque trame acoustique est représentée par 13 paramètres acoustiques et chaque trame visuelle est représentée par 7 paramètres articulatoires Enfin, un modèle de régression linéaire qui relie les paramètres acoustiques aux paramètres articulatoires trame par trame est

estimé. Lors de la synthèse, les paramètres articulatoires sont générés à partir des paramètres acoustiques grâce au modèle obtenu.

Le modèle statistique-paramétrique. Le modèle HMM

Apprentissage. Un HMM et un modèle des durées d'états sont appris pour les paramètres articulatoires de chaque phone en contexte de la base d'apprentissage (pour avoir plus de détails sur l'apprentissage et la synthèse par HMM voir Annexe 8.2). Les vecteurs d'observation sont constitués des paramètres visuels statiques et dynamiques, c'est-à-dire, des valeurs des paramètres articulatoires et de leurs dérivées. L'estimation des paramètres des HMMs est basée sur le calcul de maximum de vraisemblance (*Maximum-Likelihood criterion*) (Donovan, 1996). Cette estimation est effectuée par un algorithme spécifique de EM (*Expectation/Maximisation*) connu comme algorithme récursif de Baum-Welch. Ainsi, un modèle gauche-droit à 3 états avec les distributions gaussiennes simples est appris pour chaque diphone.

Synthèse. La synthèse est effectuée comme suit. D'abord, la chaîne phonétique à synthétiser est découpée en diphones (avec leurs durées respectives). Ensuite, une séquence des HMMs correspondant est construite. Les durées des états sont déterminées (Yoshimura *et al.* , 1998). Une fois la séquence d'états spécifiée, la trajectoire des paramètres articulatoires est estimée grâce à un algorithme spécifique de génération des paramètres (Zen *et al.* , 2004). Cet algorithme exploite la dépendance entre paramètres statiques et dynamiques. Ainsi, ce système est théoriquement adéquat pour prendre en compte l'effet de coarticulation.

Concaténation

Ici, les candidats sont des diphones multi-représentés. La sélection est effectuée en considérant les contextes gauche et droit. Si aucun diphone en contexte n'est représenté, les diphones hors-contexte sont alors considérés par l'algorithme de programmation dynamique. Aucun coût de sélection n'est considéré. Les coûts de concaténation sont égaux aux distances euclidiennes entre les paramètres articulatoires aux frontières des unités pondérées par la variance globale expliquée (voir la Table 3.5). Enfin, les trajectoires des unités sélectionnées sont élargies/compressées non linéairement pour correspondre aux durées des diphones puis un algorithme spécifique de lissage anticipatoire est appliqué (Bailly *et al.* , 2002).

Concaténation basée HMM

Un nouveau modèle qui utilise la prédiction par HMMs pour présélectionner les candidats a été implémenté. Les diphones contextuels (tri-diphones) présélectionnés à la première étape du système de concaténation sont ensuite classés dans l'ordre décroissant du coefficient de corrélation entre les trajectoires des diphones de la base des données et celles prédites par HMMs. Les N meilleurs candidats sont retenus dans le treillis pour la sélection finale du modèle de concaténation. Notons que $N = \infty$ correspond au modèle de concaténation initial et qu'une méthode de sélection moins brutale aurait consisté à utiliser le coût de sélection pour pénaliser les segments les moins corrélés.

Numéro phrase/modèle(Corrélation)	Nat	Inv	Lin	HMM	Conc
1	1,00	-1,00	0,17	**0,55**	0,50
2	1,00	-1,00	0,26	**0,63**	0,47
3	1,00	-1,00	0,26	**0,58**	0,30
4	1,00	-1,00	0,18	**0,70**	0,66
5	1,00	-1,00	0,41	0,56	**0,64**
6	1,00	-1,00	**0,62**	0,54	0,56
7	1,00	-1,00	0,12	**0,60**	0,41
8	1,00	-1,00	0,39	**0,55**	0,20
9	1,00	-1,00	0,33	0,49	**0,56**
10	1,00	-1,00	0,40	0,59	**0,67**
Global	1,00	-1,00	0,31	**0,58**	0,50

TAB. 2.1 – Corrélation moyenne entre les trajectoires de synthèse et celle d'origine pour les différents modèles et phrases.

2.5.2 Evaluation objective

Les modèles de synthèse visuelle proposés sont paramétrés à partir de la base d'apprentissage. Les dix phrases de test sont synthétisées. Le coefficient de corrélation linéaire (coefficient de Pearson) entre les trajectoires synthétiques et celles d'origine est utilisé pour l'évaluation objective, (Table 2.5.2). Cette première évaluation est mise à profit pour commencer à paramétrer de manière optimale les systèmes. Les corrélations moyennes dans le cas de la synthèse par HMMs augmentent si les paramètres dynamiques sont pris en compte pendant les phases d'apprentissage et de synthèse. La corrélation est significativement plus importante quand la dérivée première est utilisée. L'utilisation de la dérivée seconde n'augmente cette corrélation que de manière marginale. La corrélation moyenne dans le cas de la synthèse par concaténation en fonction des différentes valeurs du nombre N de Gaussiennes atteint une valeur optimale pour N=3 pour ce corpus.

Les trajectoires articulatoires des dix phrases sont générées par trois modèles : (a) le système de synthèse basé HMM avec les vecteurs articulatoires comprenant la dérivée première (HMM) ; (b) le système de synthèse par concaténation avec la méthode de présélection proposée et N=3 (Conc) ; (c) le système de synthèse par modèle de régression linéaire (Lin). Cet ensemble est complété par les trajectoires originales (Nat) et leurs inverses (Inv) où les paramètres originaux sont multipliés par -1 de manière à fournir aux sujets une gamme assez large de qualité. Les résultats de l'évaluation objective correspondant aux modèles retenus sont dans la Table 2.5.2. La corrélation moyenne est maximale pour la synthèse par HMMs. Dans le cas d'une seule phrase, la corrélation est plus importante pour le modèle linéaire que pour les autres modèles.

2.5.3 Evaluation subjective

Le but du test subjectif utilisé est d'évaluer la préférence globale des modèles proposés par rapport aux mouvements faciaux d'origine. Il faut noter que cette référence - souvent absente dans l'ensemble des stimuli utilisés dans les tests publiés - est très importante (Bailly *et al.* , 2002), (Geiger *et al.* , 2003).

Le signal acoustique original est joué en synchronie avec les mouvements faciaux. Ici, le test

Vote (%, Nb)	Nat	Inv	Lin	HMM	Conc (N=3)
1	14,30 (3)	4,80 (1)	0	14,30 (3)	**66,70 (14)**
2	33,30 (7)	0	0	28,60 (6)	**38,10 (8)**
3	19,00 (4)	0	0	**57,10 (12)**	23,80 (5)
4	28,60 (6)	0	0	**52,40 (11)**	19,00 (4)
5	**85,70 (18)**	0	0	9,50 (2)	4,80 (1)
6	0	0	0	38,10 (8)	**61,90 (13)**
7	**71,40 (15)**	0	0	28,60 (6)	0
8	**57,10 (12)**	0	0	38,10 (8)	4,80 (1)
9	**81,00 (17)**	0	0	14,30 (3)	4,80 (1)
10	38,10 (8)	0	0	**57,10 (12)**	4,80 (1)
Global	**42,9**	0,5	0	33,8	22,8

TAB. 2.2 – Résultats des évaluations subjectives.

de préférence moyenne (*Mean Preference Score* : MPS) est utilisé. Chaque participant doit alors choisir la séquence qu'il préfère parmi cinq pour chaque phrase. Les 21 sujets qui ont participé à l'expérience n'ont aucune pathologie audiovisuelle. Les sujets peuvent jouer les stimuli autant de fois qu'ils le désirent et peuvent changer leurs choix. L'ordre initial des séquences pour chaque phrase est aléatoire. Le test est effectué dans un environnement de luminance contrôlé. Les conditions de la luminance de fond sont basées sur la ITU-R BT.500-9 (ITU-R, 1998).

Les résultats du test subjectif sont dans la Table 2.5.3. Le modèle préféré est l'original (42.9%) suivi par le modèle HMM (33.8%) et le modèle de concaténation (22.9%). Les scores de préférence pour les modèles linéaire et inverse sont très bas, 0% et 0.5% respectivement. La méthode de synthèse par HMM est jugée comparable aux mouvements originaux ; les phrases générées par HMM étant de plus toujours préférées par au moins deux personnes. La synthèse par concaténation guidée HMMs est moins performante mais les résultats dépendent des phrases. Il est intéressant de constater que les mouvements de synthèse (HMM ou concaténation) sont préférés aux originaux pour six des dix phrases. Cela peut provenir des imperfections des modèles de forme et d'apparence mais les mouvements générés par ces deux modèles de prédiction sont jugés globalement comme équivalents aux mouvements originaux. Le modèle linéaire a le score le plus bas (voir aussi les résultats précédents obtenus par Gibert et al. (Bailly *et al.* , 2002)) même si sa corrélation objective est parfois importante et même proche de celle obtenue par le modèle de concaténation pour certaines phrases. Ce résultat confirme l'importance des tests subjectifs et que les résultats des tests objectifs ne sont pas toujours confirmés par les tests subjectifs. Il faut donc être prudent avec les résultats des tests objectifs, par exemple, dans l'état de l'art de la synthèse à partir de l'audio où beaucoup de systèmes ont les coefficients de corrélation de 0.7 et plus, mais ces résultats ne sont pas validés subjectivement.

2.5.4 Discussion

Dans ce chapitre, différentes méthodes de synthèse visuelle ont été évaluées objectivement et subjectivement. Une nouvelle méthode proposée concatène les segments articulatoires présélectionnés grâce à une méthode basée HMM. L'utilisation de cette méthode augmente considérablement la corrélation entre les trajectoires synthétiques et originales. Ce gain ne permet pas cependant d'atteindre ceux de la synthèse purement HMM. Dans l'ensemble, les résultats de l'évaluation objective sont confirmés par l'évaluation subjective. Le système HMM semble être le plus efficace et le mieux accepté. L'étude des résultats montre cependant que les résultats des

évaluations dépendent du contenu phonétique des phrases. Le modèle HMM, s'il est meilleur en moyenne partout, génère des trajectoires moins articulées que celles produites par le système par concaténation. C'est dans cet esprit que nous avons décidé de coupler la solide charpente construite par HMM avec la richesse des détails phonétiques capturés par la synthèse par concaténation. Nous allons continuer à suivre cette idée qui devrait à terme produire un système à la fois robuste et fin.

Chapitre 3

Données audiovisuelles

3.1 De la capture des mouvements au clonage d'une tête parlante

Dans cette partie, les principes globaux de la construction d'une tête parlante sont décrits, (voir Figure 3.1).

3.1.1 Construction du corpus

La base de tout système de synthèse de la parole est le corpus. Dans un premier temps, l'objectif est de choisir les phrases, les mots ou les type de sons à enregistrer. Le corpus enregistré doit représenter au maximum tous les mouvements faciaux et les types de sons correspondants que l'on trouve dans la parole naturelle. Pour ce faire les phrases du corpus doivent être au moins phonétiquement équilibrées[1] ou s'en rapprocher. De plus, la taille du corpus est limitée, surtout dans le cas d'un corpus audiovisuel car l'enregistrement est limité à une journée - comme nous en avons fait l'expérience, le replacement précis des marqueurs d'un enregistrement à l'autre aux mêmes endroits du visage est très problématique. Dans ce travail, deux corpus de deux différents locuteurs ont été utilisés. La construction et l'analyse de ces corpus seront présentées dans les sections 3.2 et 3.3.

3.1.2 Choix du matériel et enregistrement

Une fois le corpus théorique à enregistrer préparé, il faut choisir le matériel pour l'enregistrer et pour obtenir des autres données nécessaires à la construction d'une tête parlante. Il faut dire, qu'il n'existe pas une méthode ou un système unique qui peut fournir toutes les informations pour la construction et la modélisation d'une tête parlante. Souvent différentes méthodes d'enregistrement sont combinées. Parmi les méthodes d'enregistrement des données visuelles, on distingue (Beskow, 2003) les méthodes statiques vs. dynamiques avec diverses résolutions temporelles, les méthodes basées vidéo (2D) vs. basées marqueurs (3D) avec diverses résolutions

[1] Une liste des phrases est dite phonétiquement équilibrée lorsque la distance du χ^2 entre la distribution de ses phonèmes approche celle observée sur de grandes banques de données phonétiques de la langue française (Combescure, 1981).

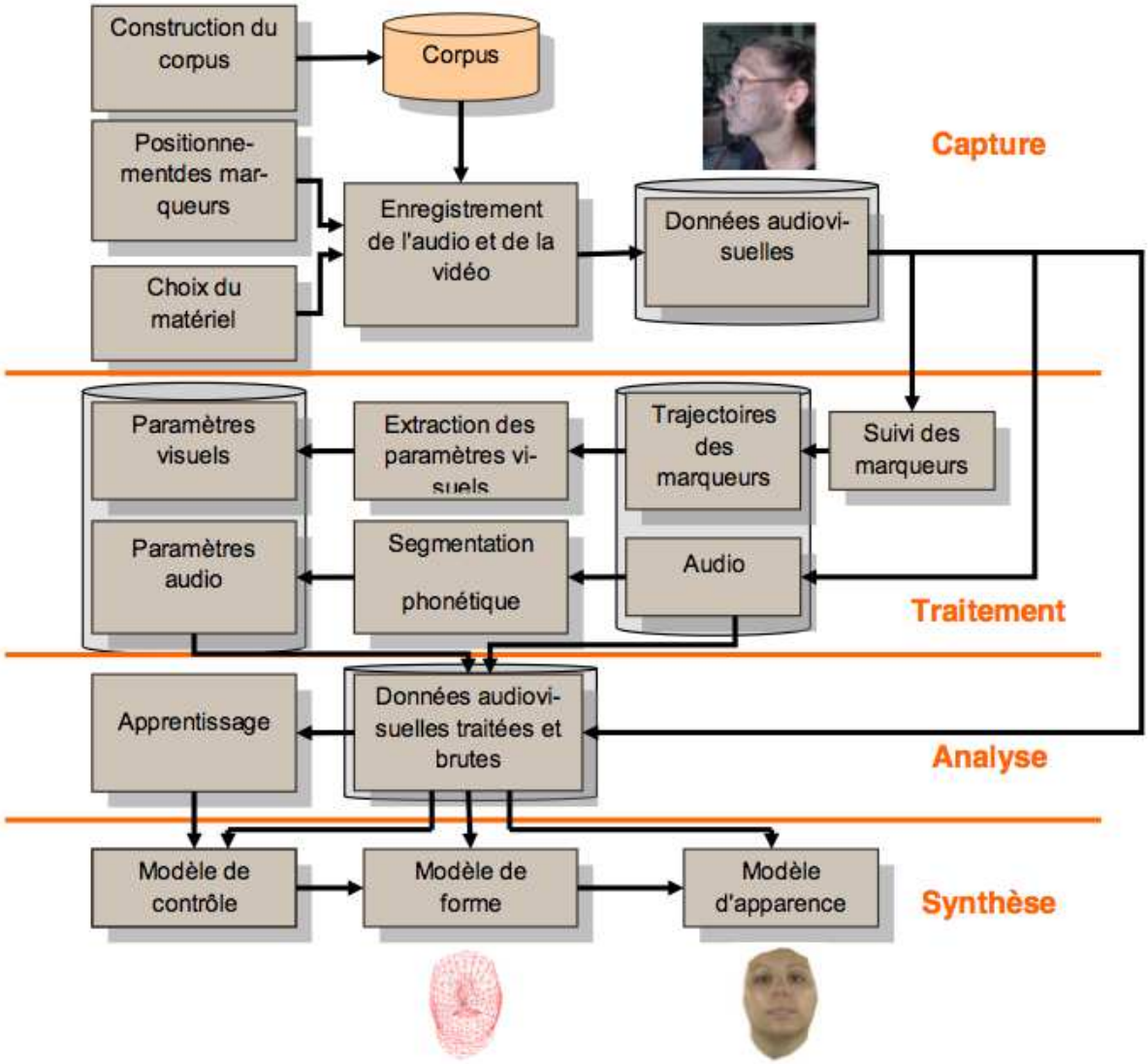


FIG. 3.1 – Schéma global du système de synthèse audiovisuelle : de l'acquisition des données audiovisuelles à la synthèse des mouvements liés à la parole.

Méthode	**Statique/ Dynamique**	**Information fournie**	**2D/ 3D**	**Interne/ Externe**	**Publications**
Photogrammétrie 3D	Statique ou dynamique	Marqueurs+texture	3D	externe	(Parke, 1982), (Elisei *et al.* , 2001), (Pighin *et al.* , 1998)
Scan Laser 3D	statique	Géométrie+texture	3D	externe	(Cohen *et al.* , 2002)
Ultrasons	statique	Géométrie	3D	interne	(Stone, 1990)
IRM	statique	Géométrie	3D	interne	(Engwall, 2000)
Basées vidéo	dynamique	Texture	2D ou 3D	externe	(Öhman, 1998), (Basu *et al.* , 1998)
Photogrammétrie infrarouge	dynamique	Marqueurs	3D	externe	ELITE, QUALISYS, VICON, ...
Electromyographie, Electropalatographie, Articulographie électromagnétique	dynamique	Marqueurs	2D ou 3D	interne	(Lucero *et al.* , 1997), (Engwall, 2000)

TAB. 3.1 – Synthèse des méthodes d'enregistrement des données visuelles pour la construction d'une tête parlante.

spatiales et les méthodes internes (par exemple, l'enregistrement des données des paramètres linguaux) vs. externes. Une synthèse des méthodes d'enregistrement est présentée dans la 3.1.2.

Dans ce travail, deux méthodes sont utilisées : des dispositifs de photogrammétrie vidéo (le système de l'ICP basé sur trois stations DPS et le système FacePox inspiré du précédent et développé par France Télécoms R&D) et un système optique (Vicon). L'audio est enregistré en synchronie avec la partie visuelle. L'enregistrement des corpus est détaillé dans les sections 3.2 et 3.3.

3.1.3 Suivi des marqueurs et extraction des paramètres visuels

Une fois les données audiovisuelles brutes enregistrées, il faut, dans un premier temps, effectuer le suivi des marqueurs, c'est-à -dire extraire les coordonnées 2D ou 3D des marqueurs en fonction du temps. Généralement, les coordonnées 2D des marqueurs sont extraites grâce aux méthodes de segmentation et de traitements des images en 2D, ensuite ces coordonnées 2D sont traduites en coordonnées 3D si nécessaire.

Dans un deuxième temps, en général, une réduction dimensionnelle est appliquée aux trajectoires des marqueurs car l'information contenue dans toutes les trajectoires est redondante (Potamianos *et al.* , 2004). Ainsi les trajectoires des paramètres visuels sont obtenues. Dans ce travail, le suivi et les paramètres visuels sont obtenus grâce à une méthode de suivi utilisant des modèles de forme et d'apparence (Odisio *et al.* , 2004), construits à partir de données étiquetées

semi-automatiquement par un suivi automatique de marqueurs (Bailly *et al.* , 2006) puis vérifiés à la main.

L'audio est enregistré en synchronie avec la modalité visuelle. Cette synchronisation est garantie par l'utilisation d'une procédure de synchronisation (génération électronique d'un bip audio et d'un flash visuel).

3.1.4 Segmentation phonétique

La segmentation en phonèmes est nécessaire pour pouvoir construire les modèles de synthèse audiovisuelle à partir du texte. En général, la segmentation phonétique est faite grâce à un alignement automatique de modèles acoustiques HMM appris grâce à l'application HTK[2] (*Hidden Markov Model Toolkit*) et vérifiée à la main.

3.1.5 Analyse et synthèse

Pour construire le modèle de contrôle, l'analyse des données audiovisuelles en fonction de la suite phonétique est effectuée. L'objectif de la construction du modèle de synthèse visuelle est de trouver les relations entre la chaîne phonétique et les paramètres visuels correspondants. Pendant la phase de synthèse les paramètres visuels sont générés en fonction de la suite phonétique grâce au modèle de contrôle obtenu pendant la phase d'analyse, (Figure 3.2).

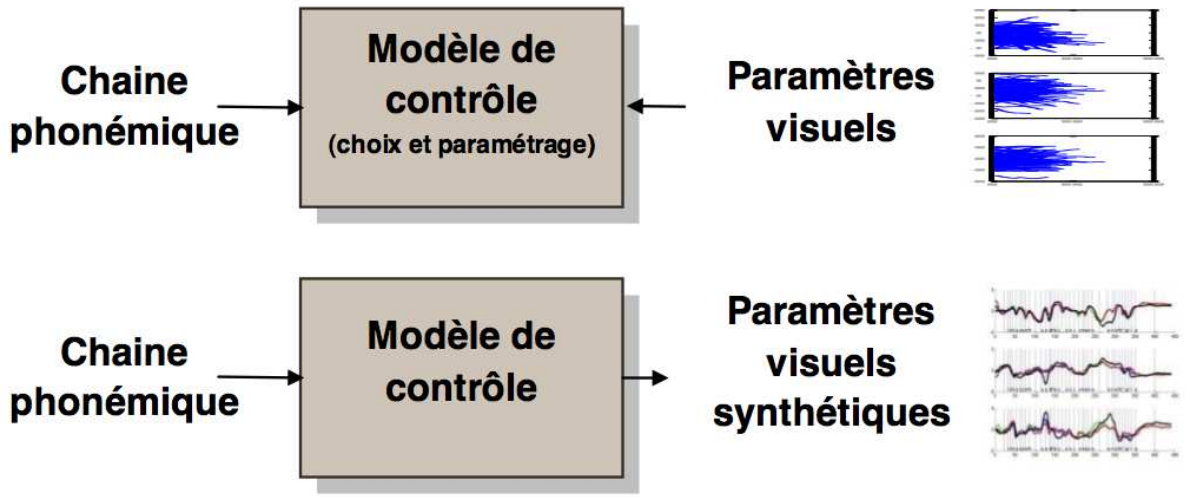
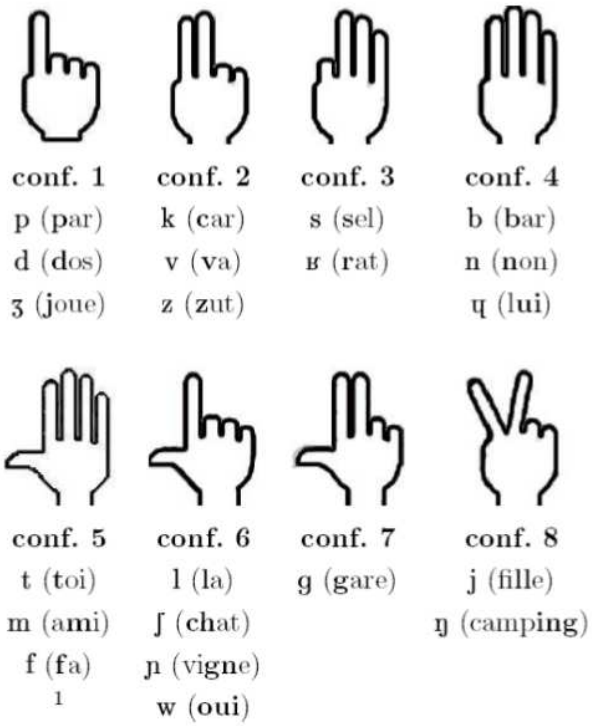


FIG. 3.2 – Les objectifs de l'analyse et de le synthèse de la parole visuelle.

3.2 Corpus I

Le corpus I est, à la base, construit pour la modélisation audiovisuelle du Langage Parlé Complété (LPC). Il est enregistré dans le cadre du projet ARTUS et a servi de base à la thèse de Gibert (Gibert, 2006). Dans ce qui suit, l'acquisition, le traitement et la modélisation des paramètres visuels du corpus I sont décrits.

[2]http ://htk.eng.cam.ac.uk/



Tab. 3.2 – Formes de la main du code LPC pour le français.

3.2.1 LPC : Langage Parlé Complété

Le Langage Parlé Complété, renommé récemment Langue française Parlée Complétée (ou code LPC), a été créée par le Dr Orin R. Cornett en 1967 sous le nom de *Cued Speech* pour l'anglais américain. Ce système manuel complétant la lecture labiale a été adapté depuis à plus de 50 langues (Cornett, 1988). Ce système est basé sur l'association articulation faciale/clés (formées par la main). Le découpage temporel est basé sur la série CV (Consonne-Voyelle). Lorsque le locuteur parle, il utilise une forme de main (déterminant un sous-ensemble de consonnes cf. Table 3.2) pour indiquer une position sur le visage (déterminant un sous-ensemble de voyelles cf. Table 3.3) pour chaque unité CV qu'il prononce (si le locuteur se retrouve à prononcer une consonne non suivie immédiatement par une voyelle, il existe une position neutre, la position "côté", de même lorsqu'il s'agit de prononcer des voyelles isolées, il existe une forme de main neutre, la configuration 5). Les clés sont définies de telle sorte que les phonèmes ayant des représentations visuelles semblables (sosies labiaux) soient associés à des clés différentes. Ainsi, les deux informations, celle délivrée par les lèvres et celle délivrée par la main, sont complémentaires et nécessaires. Elles fournissent un matriçage quasi-optimal des indices phonétiques et, par conséquent, un codage robuste des éléments phonétiques distinctifs du discours.

3.2.2 Couverture phonétique

Le corpus I se compose de 238 phrases phonétiquement équilibrées (référencées en 8.3). Il est construit pour mettre en oeuvre des systèmes de synthèse de parole par concaténation de polysons (multimodaux). Généralement, la répartition des diphones d'un corpus est comparée avec la répartition des diphones dans la langue à synthétiser pour évaluer la couverture pho-

Côté	Bouche	Menton	Pommette	Gorge
a (ma)	i (mi)	ɛ (mais)	ɛ̃ (main)	õe (un)
ʊ (maux)	ɔ (on)	u (mou)	ø (feu)	y (tu)
œ (teuf)	ã (temps)	ɔ (fort)		e (fée)

2

TAB. 3.3 – Positions de la main par rapport au visage du code LPC pour le français.

nétique de ce corpus pour la synthèse vocale. Dans le cas de la synthèse visuelle, il serait donc également intéressant d'évaluer la couverture des visèmes (voir le glossaire). Les tableaux de la couverture phonétique (répartition des phonèmes et des visèmes en contextes) du corpus I sont présentés dans l'annexe 8.3. Les histogrammes des nombres de représentants des diphones et des visèmes en contexte sont présentés dans les Figure 3.3 et Figure 3.4 et les fréquences d'apparition des visèmes en contexte pour les deux corpus et le dictionnaire sont présentées dans la Figure 3.5. Nous avons comparé la fréquence de l'apparition des diphones dans le corpus I à celle des diphones dans le dictionnaire composé de 500000 mots, en supposant que ce dernier reflète la fréquence d'apparition des diphones dans la langue française et donc phonétiquement équilibré. L'histogramme montre que le nombre (307) des diphones non représentés dans le corpus I est comparable à celui (164) du dictionnaire et que le nombre (31) des visèmes en contexte non représentés dans le corpus I est comparable à celui (21) du dictionnaire. Avec ces données, il est possible de confirmer que le corpus I est phonétiquement équilibré.

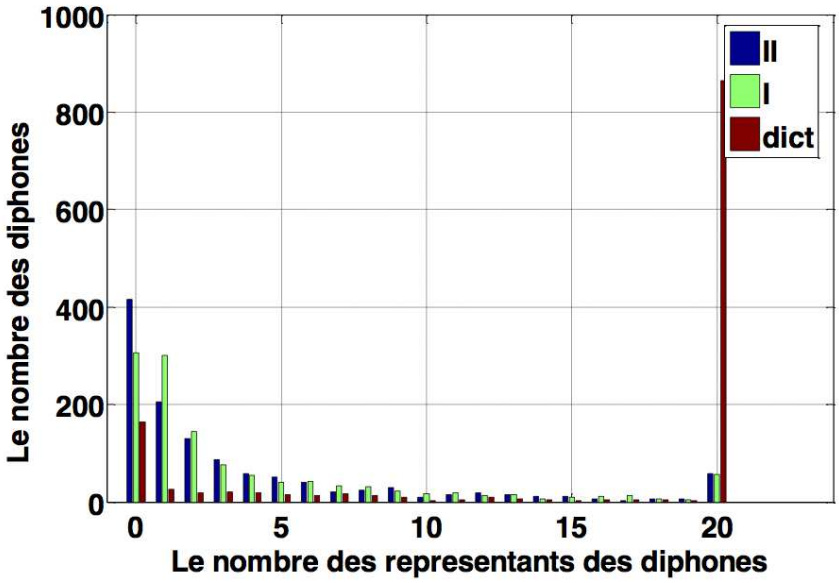


FIG. 3.3 – Nombre des diphones en fonction du nombre des représentants de ces diphones.

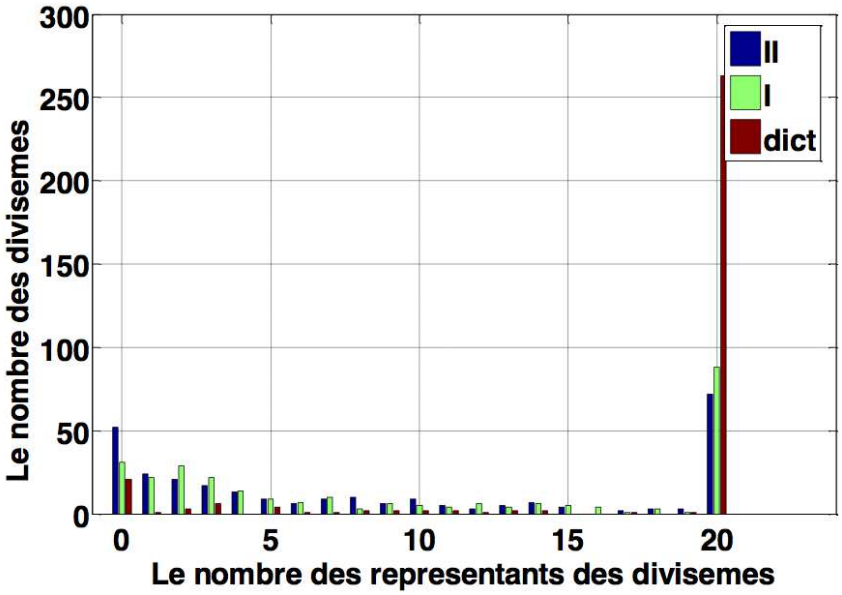


FIG. 3.4 – Nombre des divisèmes en fonction du nombre des représentants de ces divisèmes. Divisème : équivalent de diphone pour les visèmes.

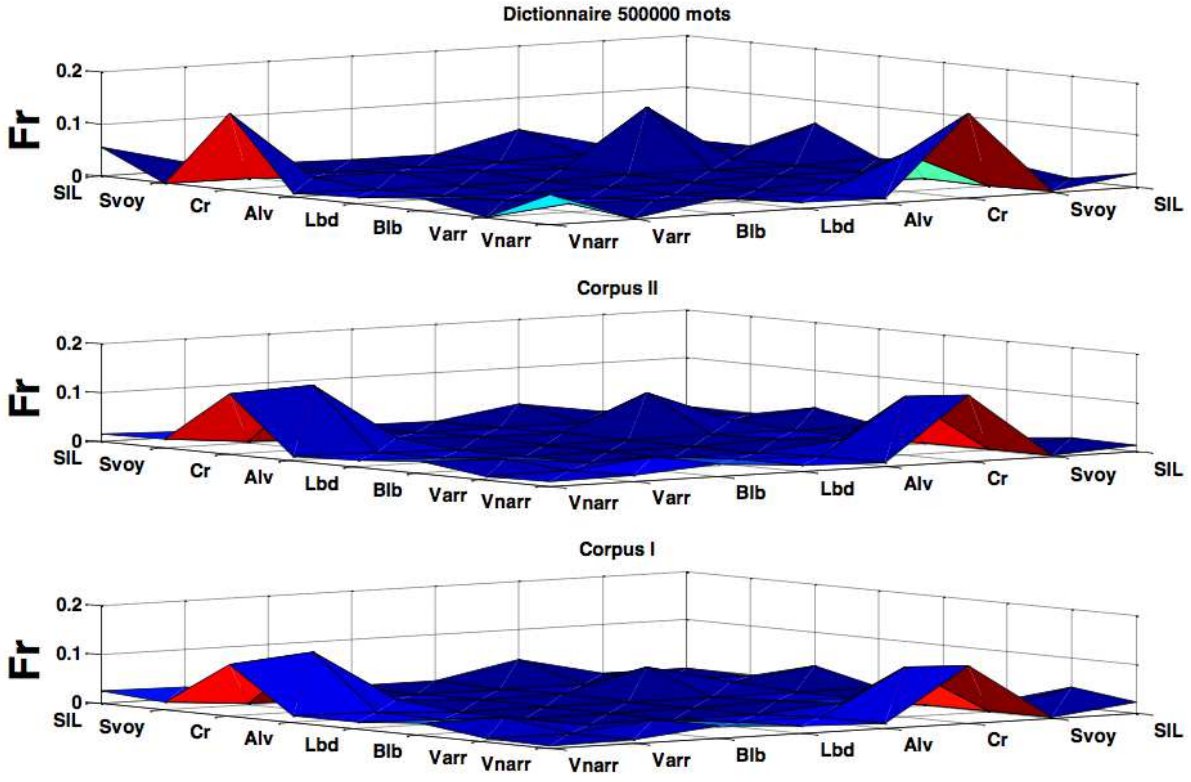


FIG. 3.5 – Fréquence d'apparition des visèmes en contexte des différents corpus. Liste des visèmes : Vnarr : voyelles non arrondies, Varr : voyelles arrondies, Blb : bilabiales, Lbd : labiodentales, Alv : post-alvéolaires, Cr : le reste des consonnes, Svoy : semi-voyelles, SIL : silences.

	0	**1**	**2**	**3**	**3**	**5**
0	-	71	33	27	47	56
1	126	720	305	174	167	243
2	32	306	76	46	36	91
3	15	285	20	21	10	27
4	19	142	75	46	42	47
5	46	211	77	64	66	124

TAB. 3.4 – Nombre de représentants lors des transitions de position à position. La position 0 correspond à la position de la main en début et fin de phrase (position "repos").

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	-	34	12	27	12	75	61	1	3
1	27	46	55	98	47	116	61	14	22
2	26	44	47	65	40	95	68	7	29
3	47	90	68	79	61	157	65	11	30
4	28	43	38	46	27	74	74	13	20
5	47	120	93	183	105	250	128	15	31
6	35	78	85	72	46	135	47	17	23
7	6	5	9	20	10	13	16	4	1
8	19	22	15	18	13	52	17	3	6

TAB. 3.5 – Nombre de représentants lors des transitions de forme à forme. La forme 0 correspond à la forme de la main en début et fin de phrase (position "repos").

3.2.3 Répartition des diclés

En ce qui concerne la main, il y a au moins une fois toutes les transitions de main, tant au niveau de la forme que de la position (cf. Table 3.4 et Table 3.5, les formes de main sont au nombre de 8 plus une forme "repos" - codée « 0 » - cf. Table 3.2 ; les positions de la main par rapport au visage sont au nombre de 5 plus une position "repos" - codée « 0 » - cf. Table 3.3). En revanche, toutes les transitions (forme + position) 1 vers (forme + position) 2 ne sont pas présentées (cf. Annexe 8.3). Elles sont en effet en très grand nombre (1680 transitions) et il est impossible d'obtenir toutes ces transitions dans un corpus de taille raisonnable. Ces transitions, dès à présent, sont nommées diclés par analogie avec les diphones, afin de pouvoir générer toutes les transitions possibles du code LPC.

3.2.4 Acquisition des données

Ce corpus a été enregistré, traité et exploité à des fins d'animation utilisant une technique de concaténation lors de la thèse de G. Gibert (Gibert, 2006). On rappelle ici les principales caractéristiques de ce corpus.

La locutrice - codeuse

La codeuse (20 ans au moment de l'enregistrement) pratique quotidiennement la Langue Française Parlée Complétée depuis 7 ans avec sa jeune soeur sourde. Elle effectue également du

codage en lycée pour d'autres sourds. Il s'agit d'une personne entendante et oralisante. Elle n'a pas encore le diplôme de codeuse professionnelle pour des raisons de disponibilité mais nous a été recommandée par le service d'orthophonie du service ORL du CHU de Grenoble. Elle suit une formation de linguistique à Grenoble, a de bonnes connaissances en phonétique et souhaite avoir le diplôme de codeuse très prochainement.

Le matériel

La phase d'enregistrement s'est déroulée dans les locaux d'Attitude Studio[3]. Une première phase a consisté à valider et calibrer le principe d'enregistrement par capture du mouvement optique des gestes de la Langue française Parlée Complétée. Pour effectuer la capture de mouvement optique, des capteurs retro-réfléchissants sont utilisés (ils sont hémisphériques de diamètre 2.5 mm) d'un système Vicon (Oxford Metrics) (composé de 12 caméras MCAM capables d'enregistrer à 120 images/s et d'une résolution d'un million de pixels). Les capteurs sont placés sur le visage et la main de la codeuse comme représenté sur la Figure 3.6. Le nombre de marqueurs est de 50 sur la main (extérieur des doigts et dos de la main) et 63 sur le visage (uniquement sur la moitié gauche (partie haute) du visage et principalement sur le bas du visage). On peut remarquer que le pouce est pourvu de plus de capteurs que le reste des doigts de la main car il est plus mobile (il possède plus de degrés de liberté). Quant au visage, on ne place pas de capteurs sur le coté droit (à l'exception du cou) afin d'éviter toute interférence avec les capteurs placés sur la main de la codeuse.

FIG. 3.6 – Position des marqueurs sur la codeuse lors de l'enregistrement.

Outre les marqueurs, le système de caméras est disposé selon deux configurations différentes en fonction des corpus à enregistrer afin d'éviter les occlusions. Ainsi, une première disposition des caméras est mise en place pour l'enregistrement du corpus main seule et une deuxième pour l'enregistrement des corpus visage seul et main + visage comme représenté sur la Figure 3.7. Dans le cas du corpus main seule, un axe principal est imposé à la main : elle est positionnée de telle sorte qu'en position poing fermé pouce ouvert, celui-ci se trouve à la verticale. Ainsi, les mouvements de rotation des doigts se trouvent alors dans un plan horizontal. Les caméras sont disposées suivant deux arcs de cercles horizontaux et des caméras supplémentaires sont ajoutées pour pouvoir suivre le pouce. Dans le cas des corpus visage seul et main + visage, une configuration dissymétrique est utilisée pour tenir compte du mouvement de la main droite lors du codage.

[3]Attitude Studio (http ://www.attitude-studio.com) est une entreprise leader dans le domaine des agents virtuels et de l'animation par capture de mouvements.

(a) configuration main seule (b) configuration main + visage

FIG. 3.7 – Configurations des caméras pour les enregistrements.

Notons que dans le même temps, le son est enregistré de façon synchrone ainsi que la vidéo de face de la locutrice.

Le protocole d'enregistrement

Les phrases du corpus sont d'abord présentées sur un écran placé en face de la codeuse. Puis une personne énonce la phrase à haute voix à un rythme normal d'élocution. La locutrice-codeuse prononce et code cette phrase. Après chaque phrase, on passe immédiatement à la phrase suivante. En cas d'erreur (évaluée par la codeuse uniquement) la phrase est mise de coté et représentée en fin de session. L'ensemble des 238 phrases et des éléments complémentaires du corpus ont été enregistrés en un après-midi à l'exception du corpus main seule qui fut enregistré la veille. La codeuse conservait les billes collées sur sa main pendant toute la nuit - la main étant protégée par un gant plastifié. Ainsi, une seule configuration de marqueurs sur le visage est utilisée alors que pour la main, la codeuse a conservé les marqueurs sur la main en les protégeant pendant la nuit par un gant entre l'enregistrement du corpus main seule et du corpus main + visage.

3.2.5 Extraction des paramètres visuels et de la main

Prétraitements

La phase de prétraitement débute par la segmentation du signal audio. Il s'agit d'une segmentation semi-automatique. Pour extraire la suite de phonèmes contenue dans la phrase et le signal audio, tout d'abord un système de reconnaissance forcée est appliqué (basé sur HTK (Woodland *et al.* , 1994)). Ainsi, en sortie, une première segmentation grossière est obtenue, qu'il s'agit dans un deuxième temps d'affiner à la main (en ajustant les frontières des consonnes et des voyelles).

Une fois cette phase de segmentation accomplie, les données sont nettoyées afin de connaître précisément les trajectoires des marqueurs de la main et du visage. Les données délivrées par les systèmes de capture de mouvements ne sont pas sans erreurs, il y a des occlusions, des

fausses détections de marqueurs, des confusions entre marqueurs... La solution envisagée est de construire des modèles statistiques des objets visage et main.

Modélisation statistique

Le visage et la main sont traités de manière séparée. Ils sont articulés par rapport à la tête et à l'avant bras, eux-mêmes considérés comme des objets rigides à 6 degrés de liberté dans l'espace.

Le visage

Paramètres articulatoires La méthodologie utilisée à l'ICP pour construire des têtes parlantes animées par des paramètres articulatoires consiste en une série d'analyses en composantes principales guidées appliquée aux mouvements de différents sous-ensembles de points de peau (Revéret *et al.* , 2000), (Elisei *et al.* , 2001), (Badin *et al.* , 2002), (Bailly *et al.* , 2003). Pour la parole, on s'intéresse plus particulièrement à la contribution de la rotation de la mâchoire, du geste d'arrondissement des lèvres, du mouvement vertical propre de la lèvre supérieure et inférieure, de celui des coins des lèvres et au mouvement de la gorge.

Toutes les opérations nécessaires au calcul du modèle sont réalisées sur les mouvements du visage où tous les marqueurs sont visibles. Une quantification vectorielle assurant un minimum de distance 3D entre les trames sélectionnées (égal ici à 2 mm), est mis en oeuvre avant la modélisation. Au final 4938 trames sont retenues comme base d'apprentissage du modèle.

A partir des mouvements de ces 63 points et, plus particulièrement, de ceux des lèvres et de la mâchoire (leurs mouvements étant supposés prépondérants), un modèle linéaire composé de 7 degrés de liberté de la parole visuelle est calculé :

1. montée/descente de la mâchoire (paramètre Jaw1). L'ACP est appliquée aux coordonnées y des points de la mâchoire et des dents du bas ;
2. étirement/protrusion des lèvres (paramètre Lips1). L'ACP est appliquée aux coordonnées (résidu) xyz des points lèvres ;
3. montée/descente de la lèvre inférieure (paramètre Lips2). L'ACP est appliquée aux coordonnées (résidu) y des points de la lèvre inférieure ;
4. montée/descente de la lèvre supérieure (paramètre Lips3). L'ACP est appliquée aux coordonnées (résidu) y des points de la lèvre supérieure ;
5. montée/descente des commissures (paramètre Lips4). L'ACP est appliquée aux coordonnées (résidu) y des points des lèvres ;
6. avancée/rétraction de la mâchoire (paramètre Jaw2). L'ACP est appliquée aux coordonnées (résidu) z des points de la mâchoire et des dents du bas ;
7. montée/descente du larynx (paramètre Lar1). L'ACP est appliquée aux coordonnées (résidu) y de tous les points sauf des lèvres et les dents du bas.

Pour chaque paramètre la variance du mouvement total expliquée par celui-ci est obtenue, comme référencée dans la Table 3.6. Ainsi les mouvements faciaux peuvent être contrôlés par un ensemble des 7 paramètres. $X = [x_1 y_1 z_1 ... x_n y_n z_n]^T = X_0 + M_x * \alpha$ où α : 7 paramètres articulatoires, M_x : modèle statistique, X : coordonnées des points faciaux (marqueurs).

Nom du paramètre	Variance expliquée	Variance cumulée
jaw1	0,462	0,462
lips1	0,187	0,649
lips2	0,038	0,687
lips3	0,032	0,719
lips4	0,016	0,735
jaw2	0,046	0,781
lar1	0,013	0,794
mvtV1	0,480	0,480
mvtV2	0,340	0,820
mvtV3	0,079	0,899
mvtV4	0,064	0,963
mvtV5	0,029	0,992
mvtV6	0,008	1

TAB. 3.6 – Variance expliquée et cumulée des paramètres articulatoires et de roto-translation pilotant le modèle de visage.

Paramètres géométriques Les paramètres géométriques sont aussi extraits. Les trois paramètres géométriques utilisés dans ce travail sont :

1. Ouverture/fermeture des lèvres - A (la distance (mm) entre les points centraux de la lèvre supérieure et de la lèvre inferieure)
2. Etirement des lèvres - B (la distance (mm) entre les coins des lèvres)
3. Protrusion des lèvres - C (la distance (mm) moyenne selon l'axe cote des points centraux de la lèvre supérieure et de la lèvre inferieure)

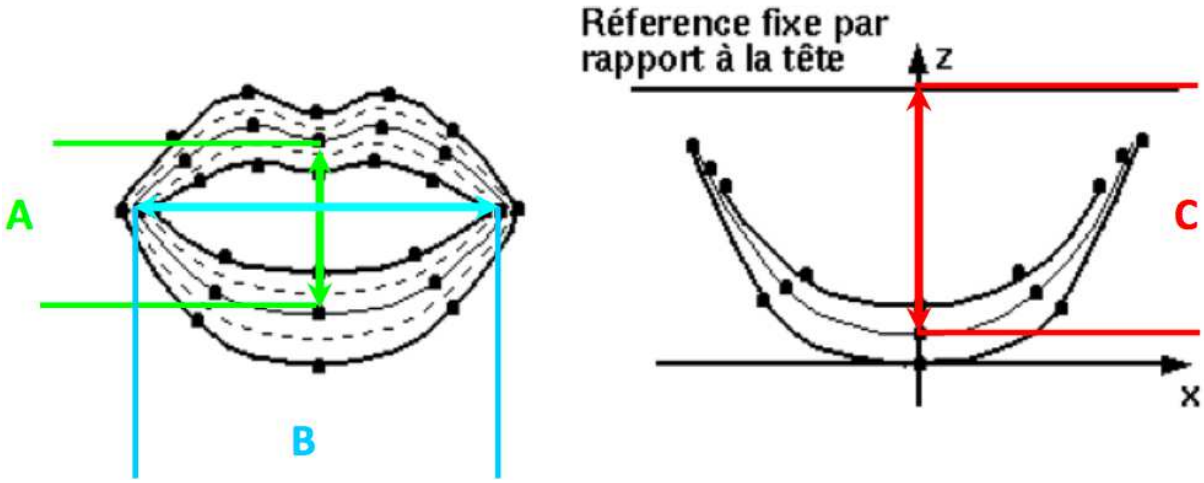


FIG. 3.8 – Paramètres géométriques utilisés. Contour des lèvres.

Les trois paramètres géométriques sont normalisés (z-score) et tout le travail est ensuite effectué sur les paramètres géométriques normalisés.

La main Des modèles statistiques non-linéaires sont utilisés dans la modélisation de la main (Bowden, 2000) car ils permettent de modéliser au mieux ces mouvements. Pour plus de détails

Nom du paramètre	Variance expliquée	Variance cumulée
ang01	0,648	0,648
ang02	0,172	0,820
ang03	0,093	0,913
ang04	0,032	0,945
ang05	0,018	0,963
ang06	0,013	0,976
ang07	0,007	0,983
ang08	0,005	0,988
ang09	0,003	0,991
mvtM1	0,464	0,464
mvtM2	0,333	0,797
mvtM3	0,143	0,940
mvtM4	0,052	0,992
mvtM5	0,007	0,999
mvtM6	0,001	1

TAB. 3.7 – Variance expliquée et cumulée des paramètres articulatoires et de roto-translation pilotant le modèle de main.

sur la modélisation de la main se référer à la thèse de G. Gibert (Gibert, 2006).

Les opérations de modélisation sont faites sur les mouvements de la main où tous les marqueurs sont visibles. Comme précédemment, une quantification vectorielle est calculée afin d'assurer un minimum de distance 3D entre les trames sélectionnées (égal ici à 2 mm). 8446 trames sont conservées comme base d'apprentissage.

A partir de ces points, 24 angles sont calculés : deux angles pour le poignet (un dans le plan abscisse-ordonnée et un dans le plan abscisse-cote), un angle dans le plan abscisse-cote pour chaque phalange de l'index, du majeur, de l'annulaire et de l'auriculaire (soit 12 angles), un angle pour ces mêmes doigts dans le plan abscisse-ordonnée pour l'écartement et enfin deux angles par phalange pour le pouce dans les plans abscisse-ordonnée et abscisse-cote. Une ACP sur les angles permet de ne conserver que 9 paramètres expliquant 99% (seuil qui est fixé) du mouvement articulatoire de la main. Les variances du mouvement expliqué par chacun de ces paramètres sont référencées dans la Table 3.7.

Le modèle de forme final est alors obtenu en effectuant une régression linéaire des coordonnées 3D des 50 marqueurs de main dans le référentiel avant-bras avec les cosinus et sinus des angles ainsi prédits. L'erreur finale de prédiction est de l'ordre du millimètre.

Résumé

La modélisation statistique des objets 3D a deux objectifs : nettoyer les données de capture de mouvement dans lesquelles apparaissent des trous, des inversions de points, etc., et réduire l'information à transmettre. Deux modèles statistiques, que l'on pilote à l'aide de 7 paramètres articulatoires (et 3 paramètres géométriques) en ce qui concerne le modèle du visage, et à l'aide de 9 paramètres articulatoires pour celui de la main, sont créés. A ces paramètres, il faut rajouter les 6 paramètres de roto-translation pour chacun des objets. Il y a donc un ensemble de paramètres

articulatoires et de roto-translations pour chaque trame de chaque phrase qui permet via les deux modèles de reconstruire les coordonnées 3D des points de ces deux objets.

3.3 Corpus II

3.3.1 Couverture phonétique

Le corpus II est constitué de 301 phrases. Les phrases utilisées sont dans l'annexe 8.3. 201 phrases choisies aléatoirement sont utilisées dans le corpus d'apprentissage et 100 phrases sont utilisées dans le corpus de test. Les tableaux de couverture phonétique (répartition des phonèmes et des visèmes en contextes) du corpus II sont présentés dans l'annexe 8.3. Les histogrammes des nombres de représentants des diphones et des visèmes en contexte sont présentés dans les Figure 3.3 et Figure 3.4 et les fréquences d'apparition des visèmes en contexte pour les deux corpus et le dictionnaire sont présentées dans la Figure 3.5. Nous avons comparé la fréquence d'apparition des diphones dans le corpus II à celle des diphones dans le dictionnaire composé de 500000 mots en supposant que ce dernier reflète la fréquence d'apparition des diphones dans la langue française et donc phonétiquement équilibré. L'histogramme montre qu'il y a 416 diphones et 52 visèmes en contexte non représentés dans le corpus II. Ainsi, il est possible de conclure que le corpus II est phonétiquement équlibré.

3.3.2 Acquisition des données

Le corpus II est enregistré grâce au système de capture de mouvements FacePox. Le système FacePox est développé au sein du laboratoire IRIS/IAM de France Télécom R&D. Le système est composé de trois caméras analogiques (une caméra de face, une caméra de face contre-plongeante et une camera de profil), (Figure 3.9), d'un microphone, d'un système de synchronisation des données audio et vidéo, d'un écran affichant les phrases à prononcer et d'un ordinateur.

La locutrice est une actrice de voix professionnelle âgée de 35 ans. 155 marqueurs sont disposés sur le visage et le cou de la locutrice.

FIG. 3.9 – Exemples des captures des trois caméras utilisés lors de l'acquisition du corpus II.

Pendant l'acquisition des données, les phrases à prononcer sont affichées sur un écran devant la locutrice. Si une phrase n'est pas prononcée correctement, la locutrice la refait car il y a une possibilité de revenir d'une phrase à l'autre.

Pour que les données audio et visuelles soient synchrones, nous avons mis en place un dispositif de synchronisation. Ce dispositif est composé d'un LED et d'un « bip ». Avant le début de chaque phrase la locutrice appuyait sur un bouton sur le dispositif et ainsi un signal audio de durée très courte (un bip) et un allumage du LED ont été enregistrés simultanément. Ainsi, nous avons pu synchroniser l'audio et le vidéo lors du traitement des séquences du corpus.

Les vidéos sont enregistrées à la fréquence 50 Hz sans entrelacement. L'audio est enregistré à la fréquence 32kHz avec le format PCM.

Ainsi, à l'issue de la phase d'acquisition du corpus II, on dispose des trois séquences vidéo et d'une séquence audio pour chaque phrase enregistrée en synchronie.

3.3.3 Extraction des paramètres visuels

Le même principe de modélisation statistique et d'extraction des paramètres visuels est utilisé que pour le corpus I, voir la section 3.2.5. Ainsi six paramètres articulatoires et trois paramètres géométriques sont extraits pour chaque phrase. Les six paramètres articulatoires sont : Jaw1 et Jaw2 (mouvements de la mâchoire), Lips1-Lips3 (mouvements des lèvres), sourc1 (mouvements des sourcils). Les trois paramètres géométriques sont : A (ouverture/fermeture des lèvres), B (étirement des lèvres) et C (protrusion des lèvres), voir la section 3.2.5. Les séquences des paramètres visuels sont sur-échantillonnées pour obtenir des séquences à 100 Hz.

3.4 Résumé

Dans ce chapitre l'acquisition et la modélisation des deux corpus sont présentées. Les deux corpus sont multimodaux : le corpus I contient l'audio, la vidéo des mouvements du visage et de la main de 238 phrases ; le corpus II contient l'audio et la vidéo du visage de 301 phrases. Les données audiovisuelles brutes issues de l'acquisition audiovisuelle sont traitées et les paramètres visuels sont extraits. Ces paramètres visuels sont : 7 paramètres articulatoires (les degrés de liberté issus de l'analyse ACP guidée) et 3 paramètres géométriques (distances en mm représentants ouverture, étirement et protrusion des lèvres) pour le corpus I, 6 paramètres articulatoires et 3 paramètres géométriques pour le corpus II. L'audio est enregistré en synchronie avec la modalité visuelle. La segmentation phonétique semi-automatique est appliqué à l'audio des deux corpus. Ainsi pour la modélisation du modèle de contrôle on dispose des séquences des paramètres visuels et l'audio correspondant segmenté phonétiquement.

Chapitre 4

Synthèse par TDA (*Task Dynamics for Animation*). Aspect configurationnel

Dans ce chapitre, la synthèse par HMM et par concaténation sont étudiées plus en détails. En particulier, les réalisations des cibles des phonèmes pour les différents paramètres des deux modèles sont analysées. Suite à cette analyse et aux résultats d'évaluation obtenus, nous allons proposer un nouveau modèle de synthèse nommé TDA (*Task Dynamics for Animation*).

4.1 Groupement des phonèmes en classes des visèmes

Pour analyser la réalisation spatiale des allophones lors de la synthèse, la réalisation des cibles articulatoires correspondantes est étudiée. Dans ce travail, une cible articulatoire est la valeur des paramètres articulatoires prise à la moitié de la durée du phone correspondant. Dans un premier temps, l'objectif est d'étudier la ressemblance articulatoire entre les réalisations des différents allophones. La distance de Bhattacharyya (4.1) est souvent utilisée pour calculer la distance entre deux distributions gaussiennes (Mak & Banard, 1996), (Hazen, 2006). Ici, la distance de Bhattacharyya est utilisée pour calculer les distances entre les distributions gaussiennes des cibles articulatoires des allophones correspondants. Une fois ces distances calculées, les dendrogrammes[1] pour les consonnes et les voyelles sont construits. Ici, les phonèmes proches visuellement sont appelés visèmes. Sur les Figure 4.1 et Figure 8.1 (en Annexe 8.1) les dendrogrammes correspondants aux consonnes et aux voyelles sont représentés pour les paramètres articulatoires et géométriques des deux corpus.

$$d_{ij} = \frac{1}{8}(\mu_i - \mu_j)^t \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2}\ln \frac{\left|\left(\Sigma_i + \Sigma_j\right)/2\right|}{\left(|\Sigma_i|^{\frac{1}{2}}|\Sigma_j|^{\frac{1}{2}}\right)} \tag{4.1}$$

où d_{ij} est la distance de Bhattacharyya entre deux phonèmes i et j, μ_i, μ_j sont les moyennes des cibles articulatoires des phonèmes i et j respectivement, Σ_i, Σ_i sont les variances des cibles articulatoires des phonèmes i et j respectivement.

[1] La représentation graphique d'une classification hiérarchique. Dans notre cas le dendrogramme correspond à un graphe de proximité (par rapport à une distance minimale) par pairs.

Ainsi, les consonnes peuvent être groupées en quatre groupes de visèmes (entre les parenthèses sont indiqués les symboles équivalents utilisés par Gipsa-lab) :

- Les bilabiales : p, b, m, (p, b, m) : Blb
- Les labiodentales : f, v, (f, v) : Lbd
- Les post-alvéolaires : ʃ, ʒ, (s^, z^) : Alv
- Les consonnes restantes : t, d, n, s, z, k, g, l, ʁ, (t, d, n, s, z, k, g, l, r) : Cr

Les voyelles peuvent être groupées en trois groupes de visèmes :

- Les voyelles ouvertes : a, e, i, ɛ, j, œ̃, ɛ̃, (a, e, i, e^, j, x⌒, e⌒) : Vnarr
- Les voyelles mi-ouvertes : ɔ, ɑ̃, (o^, a⌒) : Vnarr
- Les voyelles fermées et arrondies : o, œ, ø, u, y, ɔ̃, et ɥ, w , (o, x^, x, u, y, o⌒ et h, w) : Varr

Il intéressant de remarquer que les mêmes résultats sont obtenus pour les deux corpus et pour différents ensembles de paramètres visuels.

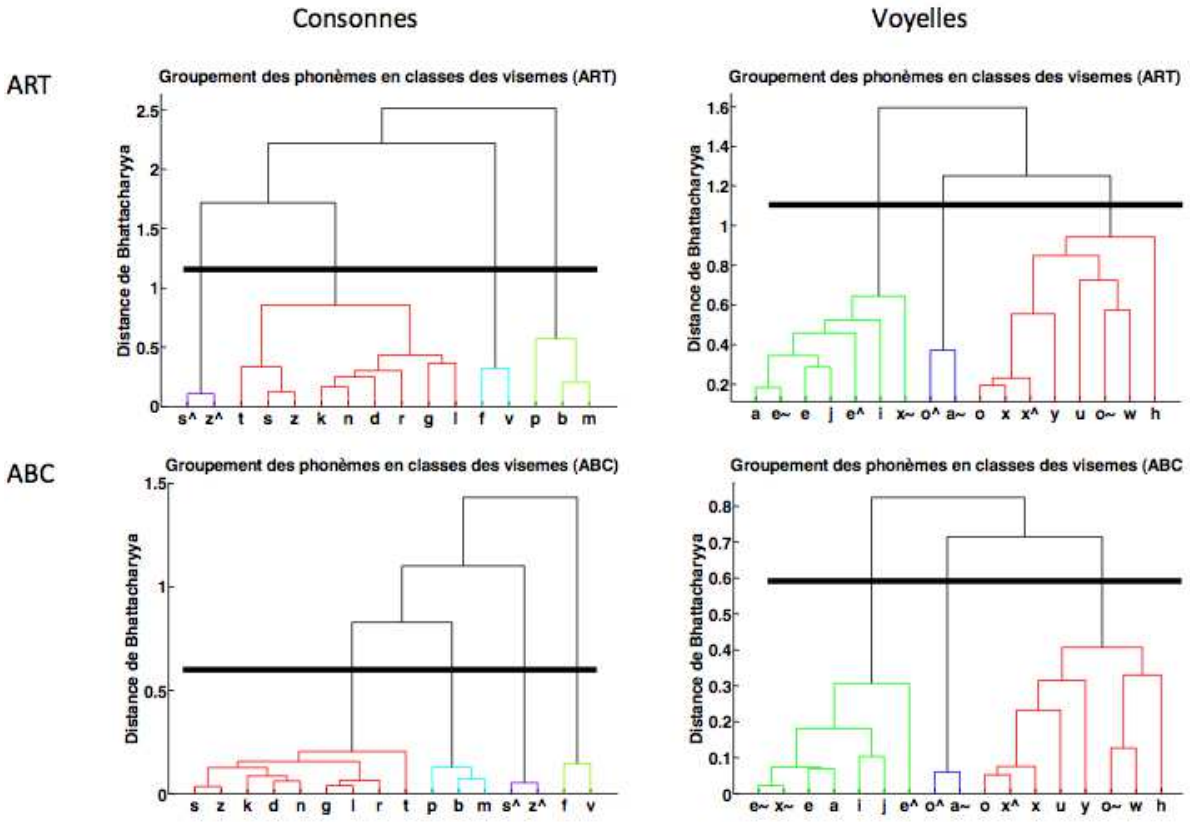


FIG. 4.1 – Corpus I. Groupement des consonnes et voyelles en classes des visèmes grâce à la distance de Bhattacharyya pour les paramètres articulatoires et géométriques. Le trait horizontal gras figure le seuil choisi pour déterminer les classes de visèmes.

4.2 Réalisation des cibles par la synthèse par HMM et par concaténation

Une fois que les réalisations visuelles des phonèmes sont classées en visèmes, l'objectif est d'étudier les réalisations des visèmes pour les différentes méthodes de synthèse. La corrélation linéaire et l'analyse discriminante ADL (Analyse Discriminante Linéaire)[2] sont utilisées pour évaluer les réalisations des visèmes. L'ADL permet de réduire l'espace de représentation et de proposer une représentation graphique qui permet de visualiser les proximités entre les observations. Pour cette raison, l'ADL est utilisée pour analyser la réalisation des cibles. D'une part, cela permet d'étudier la séparation des visèmes grâce au coefficient ADL (rapport entre inter- et intra-distances). D'autre part, cela permet de visualiser la réalisation spatiale des visèmes grâce à la projection des données sur le premier plan discriminant (deux premiers paramètres de l'ADL). Le modèle ADL est tout d'abord calculé pour les données d'origine (Nat). Ensuite, les données d'origine ainsi que les données de synthèse obtenus par HMM et par concaténation (Conc) sont projetées sur le premier plan discriminant. Les ellipses de dispersion des cibles des visèmes selon ces deux premiers paramètres ADL pour les différentes méthodes de synthèse sont représentées dans les Figure 4.2 et Figure 8.2 (en Annexe 8.1) pour le corpus I et dans les Figure 8.3 (en Annexe 8.1) et Figure 8.4 (en Annexe 8.1) pour le corpus II. Sur ces figures, l'intra-distance (taille des ellipses) de la synthèse par concaténation est proche des données d'origine. Par contre, l'intra-distance de la synthèse par HMM est très réduite par rapport aux donnes naturelles. Selon les résultats des Table 4.1 et Table 8.1 (en Annexe 8.1), le coefficient ADL est plus grand dans le cas de synthèse par HMM que dans celui de la synthèse par concaténation : la première offrant une meilleure séparation des cibles des visèmes. Ce résultat confirme les tests objectifs et subjectifs réalisés auparavant. En moyenne, les HMM réalisent mieux les cibles articulatoires (coefficient ADL : meilleure séparation des cibles que dans le cas de synthèse par concaténation). Cependant, ces cibles sont prototypiques, pas assez coarticulées. La concaténation, quant à elle, autorise de conserver la variabilité des cibles en contexte (l'intra-distance : taille des ellipses de dispersion est plus importante que dans le cas de synthèse par HMM).

Les coefficients de corrélation linéaire sont calculés pour les synthèses par concaténation et par HMM dans les différents espaces des paramètres visuels, Figure 4.3. L'espace ART_abc correspond aux paramètres géométriques calculés à partir des paramètres articulatoires déjà synthétisés. La synthèse HMM est mieux réalisée dans l'espace des paramètres géométriques que dans l'espace des paramètres articulatoires. Elle a également la corrélation plus grande que la synthèse par concaténation. Dans la Figure 4.4, un exemple de la synthèse de la phrase "Celui qui joue" est représenté. Dans cet exemple (selwikiʒu), l'amplitude de la synthèse par concaténation est très proche du naturel. Malheureusement, son timing est plus décalé de la cible que celui de la synthèse par HMM. Par contre, le timing de HMM est bon même si son amplitude est lissée.

[2]L'objectif de l'analyse discriminante linéaire est de maximiser la rapport entre la matrice de l'inter-distance notée S_b et la matrice de l'intra-distance notée S_w : $max(det(S_b)/det(S_w)) = max(d_{intra}/d_{inter})$. La transformation linéaire est donnée par une matrice U dont les colonnes sont les vecteurs propres du ${S_w}^{-1}/S_b$. Les vecteurs propres sont la solution du : $S_b u_k = \lambda_k S_w u_k$. Notons :

- C est le nombre des classes (visèmes)
- μ_i est la moyenne d'une classe i, $i = 1, 2, ..., C$
- M_i est le nombre d'échantillons dans une classe i, $i = 1, 2, ..., C$
- $M = \sum_{i=0}^{C} M_i$ est le nombre total d'échantillons
- $S_w = \sum_{i=1}^{C} \sum_{j=1}^{M_i} (x_j - \mu_i)(x_j - \mu_i)^T$
- $S_b = \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T$ où $\mu = 1/C \sum_{i=1}^{C} \mu_i$

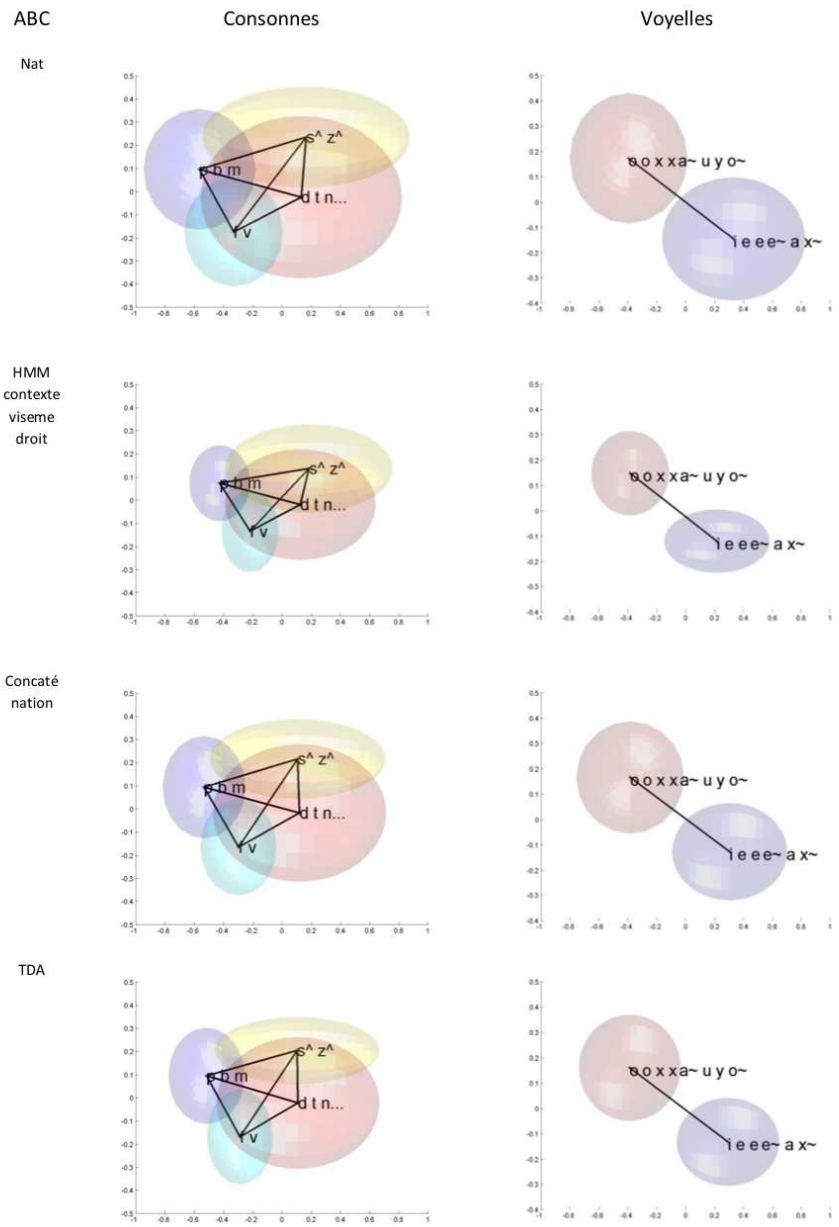


FIG. 4.2 – Ellipses de dispersion des cibles géométriques pour les principales classes des consonnes et des voyelles selon ADL pour les données naturelles, la synthèse par HMM, la synthèse par la concaténation et la synthèse par TDA. Corpus I

Cibles	**Modèle**	**d_inter**	**d_intra**	**d_inter/ d_intra**	**Taux de reconnaissance %**
ABC voyelles	NAT	1,25	1	1,25	94
	HMM	0,85	0,34	2,53	94
	Conc	1,01	0,64	1,58	94
	TDA	0,99	0,52	1,89	94
ABC consonnes	NAT	0,29	1	0,29	67
	HMM	0,17	0,43	0,39	64
	Conc	0,25	0,67	0,37	68
	TDA	0,24	0,59	0,41	69
ART voyelles	NAT	0,45	1	0,45	94
	HMM	0,30	0,32	0,96	92
	Conc	0,40	0,67	0,60	94
	TDA	0,38	0,55	0,70	94
ART consonnes	NAT	0,20	1	0,20	70
	HMM	0,10	0,39	0,25	65
	Conc	0,17	0,71	0,24	70
	TDA	0,16	0,62	0,26	71

TAB. 4.1 – Les caractéristiques de la ADL (inter-distance, intra-distance et leur rapport) des consonnes et voyelles dans les espaces géométrique et articulatoire pour les données naturelles, la synthèse par HMM, la synthèse par la concaténation et la synthèse par TDA. Corpus I. Données d'apprentissage et de test. Le taux de reconnaissance est obtenu par calcul de la distance de Mahalanobis des cibles aux centres des ellipses de dispersion des visèmes.

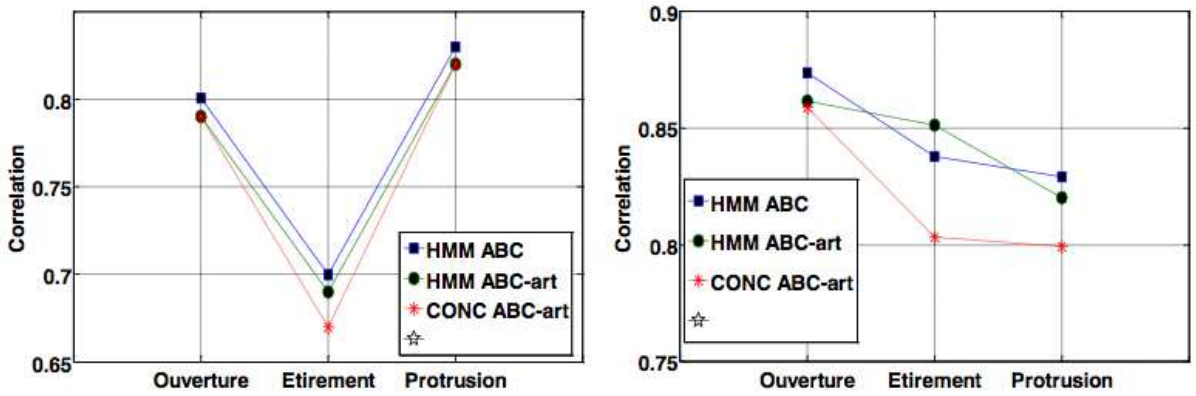


FIG. 4.3 – Les coefficients de corrélation des synthèses par HMM et par concaténation dans l'espace géométrique. Corpus I (gauche) et II (droit). Données d'apprentissage et de test. HMM ABC : paramètres géométriques de synthèse par HMM, HMM ABC-art : paramètres géométriques obtenus par la synthèse par HMM sur les paramètres articulatoires, CONC ABC-art : paramètres géométriques obtenus par la synthèse par concaténation sur les paramètres articulatoires.

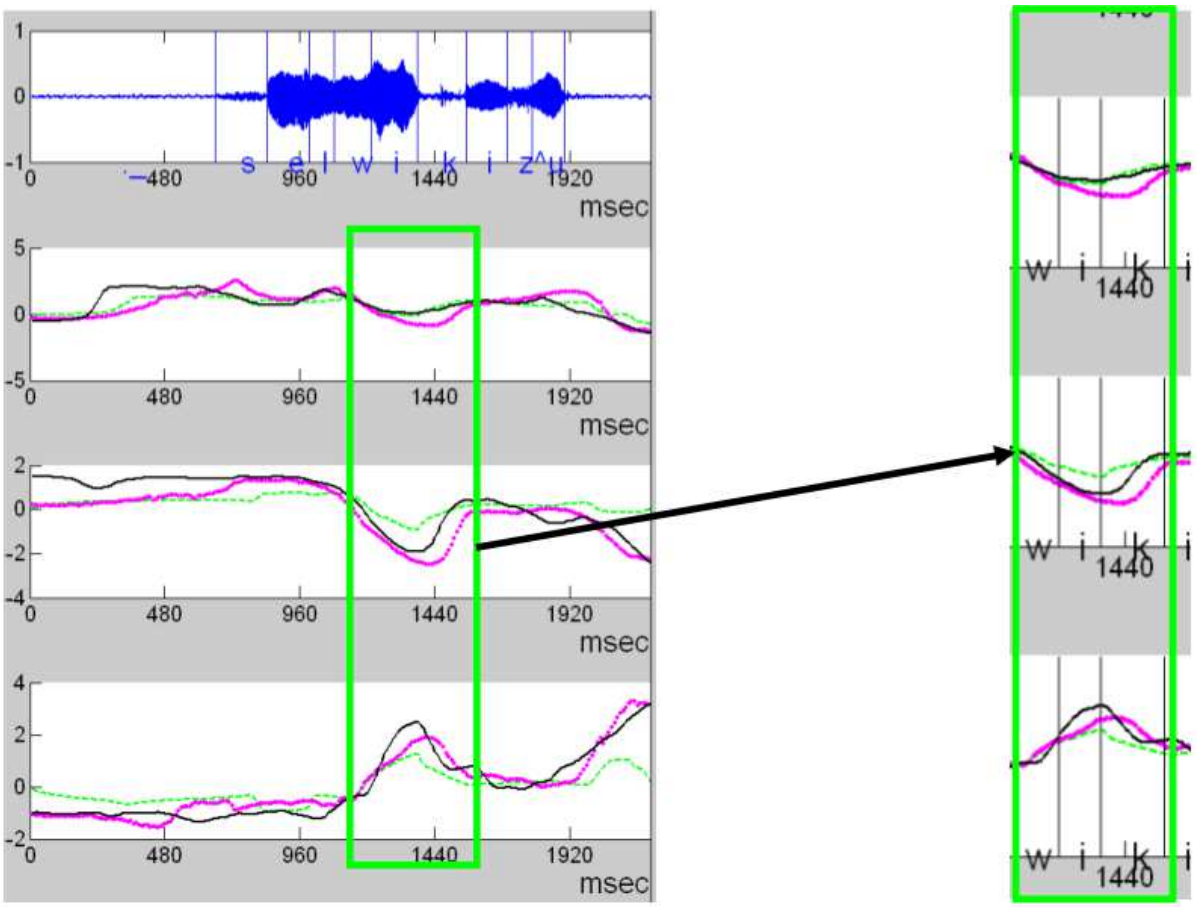


FIG. 4.4 – Les trajectoires des paramètres géométriques pour les synthèses par HMM en vert et par concaténation en rose, naturel est en noir. Phrase "Celui qui joue". A souligner, les trajectoires moins articulées pour la synthèse par HMM et les trajectoires plus articulées mais le timing décalé pour la synthèse par concaténation.

4.3 Synthèse par TDA

4.3.1 Planification de la coarticulation

La principale problématique de la modélisation des mouvements faciaux liés à la parole est sa grande variabilité. L'objectif de la synthèse de la parole est de modéliser la relation entre la chaîne de symboles en entrée et le signal audiovisuel en sortie. Cette relation est du type *one-to-many*, donc l'objectif de la synthèse de la parole est de trouver les lois, les contraintes qui gouvernent la sélection des trajectoires observées dans l'espace des possibilités. Quelques théories ont été proposées pour réconcilier cette apparente variabilité avec l'existence d'invariabilités acoustiques, géométriques ou articulatoires.

Il existe ainsi une théorie de la coarticulation qui postule que la parole peut être représentée comme une séquence des gestes articulatoires présentant un certain nombre de facettes invariantes (Browman & Goldstein, 1990a), (Browman & Goldstein, 1990), les constellations articulatoires étant donc fortement planifiées (Whalen, 1990). La théorie de la phonologie articulatoire (Browman & Goldstein, 1990a), (Browman & Goldstein, 1990) essaie de trouver des invariabilités qui existent dans la production et la perception de la parole. Cette théorie s'inscrit dans une réflexion sur la nature des primitives phonologiques et sur les relations entre phonologie et phonétique. La théorie de phonologie articulatoire originale est basée sur une seule unité, le geste articulatoire, servant à la fois de primitive dans les représentations phonologiques et d'unité d'action dans la production de la parole. Un geste articulatoire est aussi : l'action de formation et de relâchement d'une constriction à un endroit spécifique dans le conduit vocal. Le mot "pas", par exemple, commence par un geste d'occlusion labiale, alors que le mot "cas" commence par un geste d'occlusion du corps de la langue. Un geste va alors se caractériser par :

- des informations sur les articulateurs qui le forment,
- des informations sur la constriction qui est son but (sa tâche),
- des paramètres dynamiques spécifiant comment cette constriction est faite : quels sont les articulateurs utilisés pour l'accomplir.

Ainsi, les gestes possibles dans le conduit vocal et potentiellement distinctifs vont être définis en types catégoriquement distincts. Pour cela les gestes sont premièrement catégorisés en fonction des "structures coordinatives" qui forment la constriction dans le conduit. Un geste est toujours spécifique à une structure coordinative, et cette structure est composée d'une série d'articulateurs indépendants travaillant en synergie. Par exemple, le geste d'occlusion labiale emploie une structure coordinative composée de trois articulateurs dont l'action est coordonnée : la lèvre inférieure, la lèvre supérieure et la mâchoire. Deuxièmement, les gestes sont caractérisés par un certain nombre de "variables du conduit vocal". Ces variables spécifient à la fois l'objectif fonctionnel (la tâche) de chaque geste (les caractéristiques de la constriction, où et comment elle est formée) et définissent les paramètres dynamiques qui lui sont associés. Dans l'exemple du mot "pas", le premier geste se caractérisera par une double variable du conduit vocal qui spécifie que la constriction doit être totale et au niveau labial et qui y associe des caractéristiques dynamiques spatio-temporelles déterminant les trajectoires articulatoires de l'ensemble des articulateurs compris dans la structure coordinative concernée (ici, lèvres inférieure et supérieure, mâchoire).

En résumé, cette théorie postule qu'il existe des invariants gestuels dans l'espace des primi-

Dans notre cas, nous appliquons d'abord LDA sur les données naturelles (S_w, S_b, λ_k, u_k), ensuite nous projetons les données de synthèse dans l'espace LDA obtenu précédemment : λ_k, u_k restent les mêmes, c'est S_w, S_b qui changent.

tives phonologiques malgré la variabilité surfacique des gestes articulatoires (voir la Figure 4.5a). La réalisation du phonème /b/ est donnée comme exemple de l'application de la théorie (voir Figure 4.5b) : si la fermeture labiale (variable géométrique) est invariante, ses réalisations articulatoires spatio-temporelles (par les articulateurs recrutés : mâchoire, lèvre inférieure et supérieure) varient en fonction du contexte, du locuteur et etc.

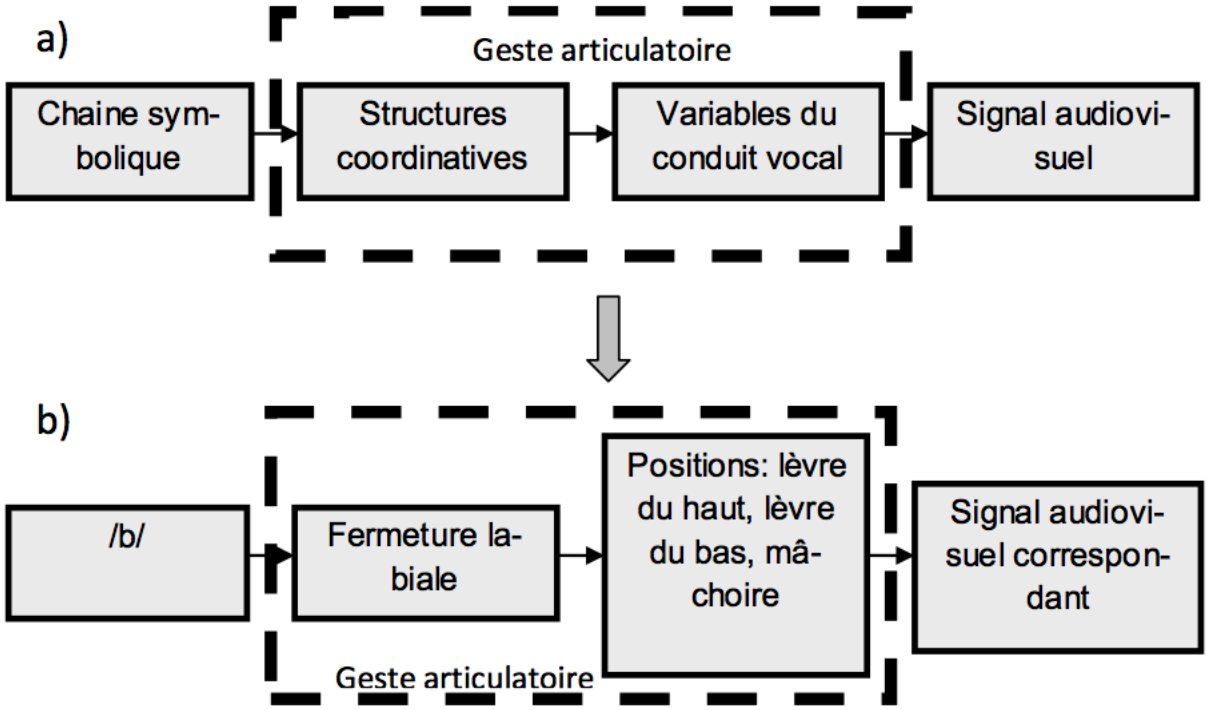


FIG. 4.5 – a) Production de la parole selon la théorie de la phonologie articulatoire ; b) Exemple de la production de la parole selon la théorie de la phonologie articulatoire pour le phonème /b/.

Une simulation numérique *Task Dynamics* de la théorie de la phonologie articulatoire a été proposée par des chercheurs de *Haskins Laboratories* (Saltzman & Munhall, 1989). Ce système calcule les trajectoires évoluant dans le temps des différents articulateurs des structures coordinatives. Les variables du conduit sont spécifiques aux structures coordinatives qui forment la constriction. Pour les structures coordinatives "labiale", "pointe de la langue", "corps de la langue", les gestes se caractérisent par deux variables du conduit vocal couvrant les deux dimensions de degré et de lieu de constriction le long du conduit vocal. Dans le modèle proposé, les modes et lieux d'articulation sont implémentés sous forme d'étiquettes (*descriptors*) correspondant aux plages de valeurs associées aux paramètres contrôlant les degrés et les lieux de constriction. Ces étiquettes sont :

- degré de constriction : occlusion, critique, étroite, moyenne, large ;
- lieu de constriction : labiale, dentale, alvéolaire, palatale, vélaire, uvulaire, pharyngale.

Les trajectoires sont calculées à partir d'un paramètre spécifiant la cible (le point d'équilibre), la rapidité des mouvements (raideur) et le degré de rigidité de la constriction (amortissement). Il est à noter que ces valeurs n'ont pas de statut théorique particulier (contrairement à la catégorisation des gestes en structures coordinatives). Les valeurs numériques correspondant à ces étiquettes sont obtenues empiriquement. Les variables du conduit sont spécifiques aux structures coordinatives qui forment la constriction.

4.3.2 TDA : Concaténation guidée HMM

Nous proposons d'utiliser la théorie de la planification des gestes articulatoires proposée par Browman et Goldstein dans le système de synthèse de la parole visuelle étudié dans cette thèse. Dans le système proposé, les gestes articulatoires sont planifiés grâce à la génération par HMM dans l'espace géométrique (paramètres de fermeture labiale, étirement et protrusion). Ensuite, les gestes articulatoires sont exécutés par concaténation dans l'espace articulatoire (paramètres articulatoires statistiques issus de l'ACP guidée) grâce aux partitions géométriques et articulatoires (voir la Figure 4.6). L'étape de planification est par nature adaptée à l'espace géométrique. En effet, les trois paramètres géométriques utilisés correspondent aux principaux mouvements visibles des lèvres. De plus, il est démontré que les cibles articulatoires sont mieux discriminées par HMM dans l'espace géométrique, voir les résultats dans les table 4.1 et table 8.1. Les paramètres articulatoires correspondent aux variables visibles d'un visage parlant 3D. En effet, c'est l'ensemble des paramètres articulatoires (variables articulatoires) qui forme une constellation articulatoire détaillée. La synthèse par HMM est utilisée pour planifier les mouvements articulatoires, elle fournit des gabarits spatio-temporels. Il est démontré par des tests objectifs et subjectifs que la synthèse par HMM en moyenne donne les meilleurs résultats que la synthèse par Concaténation (voir la Section 2.5) et, en plus, la synthèse par HMM a la capacité de planifier la coarticulation à long terme (voir la Section 4.3.3). La synthèse par concaténation joue le rôle d'exécution des mouvements articulatoires. La méthode de concaténation fournit de l'articulation détaillée, voir la Section 4.2. La synthèse par concaténation guidée HMM devrait produire des mouvements plus proches des trajectoires naturelles et mieux agencés que la concaténation simple car les coûts de sélection (partitions géométriques) sont utilisés en plus des coûts de concaténation.

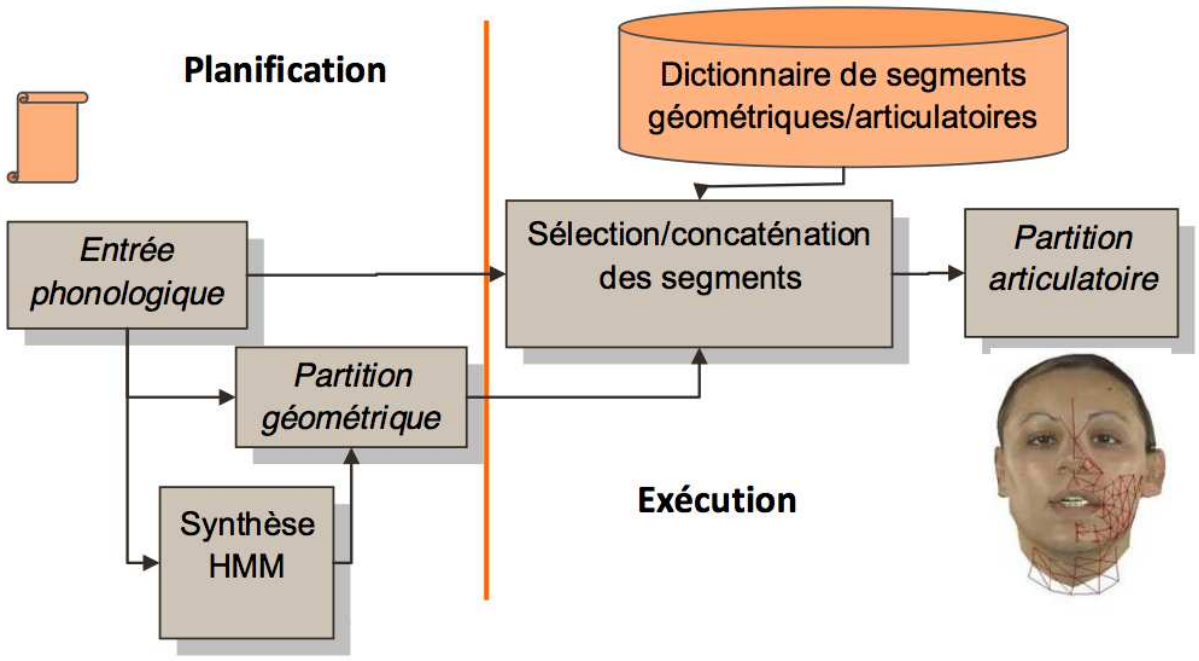


FIG. 4.6 – Schéma du système de la synthèse par TDA : Planification par HMM dans l'espace géométrique et exécution par concaténation dans l'espace articulatoire.

4.3.3 Planification par HMM

Dans cette partie, l'analyse des divers paramètres de la synthèse par HMM est présentée.

Analyse par rapport à la structure optimale des HMM

Dans un premier temps, un modèle HMM est appris pour chaque monophone (phonème sans contexte). Notons que le graphe de transition est imposé : nous utilisons des chaînes d'ordre 1. L'erreur moyenne et la corrélation entre les trajectoires de synthèse et celles d'origine sont calculées en fonction du nombre d'états des HMM. L'erreur moyenne (et son écart-type) est calculée sur les valeurs des paramètres géométriques sur toutes les phrases du corpus (test ou/et apprentissage) et exprimée en mm. Le test Anova[3] est appliqué pour comparer les moyennes des différents modèles. Pour le corpus I, la distorsion minimale est obtenue avec quatre états (pas de différence significative à partir de quatre états, p=0.05[4], Figure 4.7). Pour le corpus II, la distorsion minimale est obtenue avec cinq états (p=0.05). Nous avons donc choisi de travailler avec les HMM à cinq états pour les corpus I et II.

Dans un deuxième temps, un modèle HMM est appris pour chaque monophone en contexte (contexte visème droit, détails du test contexte gauche ou droit sont dans la section suivante). Enfin, l'analyse par rapport aux paramètres dynamiques, l'utilisation des paramètres acoustiques dans le vecteur d'apprentissage et l'utilisation des mélanges des gaussiennes est effectuée, (Figure 4.8). Il n'y a pas de différence significative (p>0.05) entre les résultats de synthèse avec le paramètre dynamique du 1er ordre et avec les paramètres dynamiques du 1er et du 2ème ordre. De plus, si les paramètres acoustiques (3 paramètres acoustiques calculés par ACP de 20 paramètres MFFC) sont ajoutés dans le vecteur d'apprentissage, la distorsion n'est pas diminuée. Lorsque les probabilités d'émission des modèles HMM sont représentées comme des mélanges des gaussiennes, la distorsion n'est pas diminuée. Désormais, un modèle HMM est appris par phonème avec le contexte droit visème, il est constitué de cinq états et les paramètres dynamiques du 1er ordre.

Analyse par rapport à l'information contextuelle

L'erreur moyenne est calculée pour les différents modèles HMM en fonction du contexte utilisé, Figure 4.9 et Figure 4.10. Plusieurs types de contexte sont étudiés : contexte phonème gauche ou droit, contexte visème gauche ou droit, contexte gauche et droit et contexte visème droit avec l'information syllabique. La distorsion est moins importante (p=0.05) dans le cas de synthèse avec le contexte droit qu'avec la synthèse avec le contexte gauche pour les deux corpus. Ce résultat confirme la théorie de coarticulation sur le fait que la coarticulation anticipatoire est prédominante sur la coarticulation progressive. Il n'y a pas de différence significative (p>0.05)

[3]L'analyse de la variance (terme souvent abrégé par le terme anglais ANOVA : *ANalysis Of VAriance*) est une technique statistique permettant de comparer les moyennes de deux populations ou plus. Cette méthode utilise des mesures de variance afin de déterminer le caractère significatif, ou non, des différences de moyenne mesurées sur les populations.

[4]« La différence est significative » veut tout simplement dire qu'il y a une évidence statistique qu'il existe une différence. Dans les cas simples, un test statistique des hypothèses est défini comme probabilité de faire une décision pour rejeter l'hypothèse nulle quand celle-ci est vraie. La décision est souvent prise grâce à une valeur dite de p-value (noté aussi comme α). Si la valeur de p est plus petite d'un seuil significatif alors, l'hypothèse nulle est rejetée. Les valeurs traditionnelles de p sont 0.05, 0.01 et 0.001.

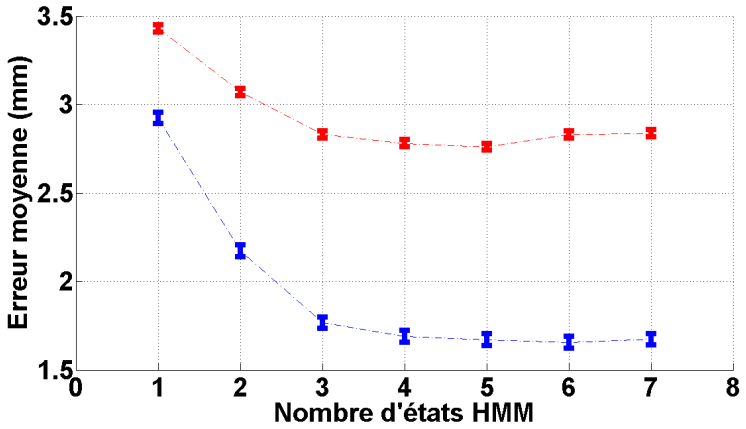


FIG. 4.7 – L'erreur moyenne (mm) entre les trajectoires de synthèse et originales pour les différents nombres d'états HMM. HMM monophone (hors contexte). Corpus I (bleu) et II (rouge). Données d'apprentissage et de test.

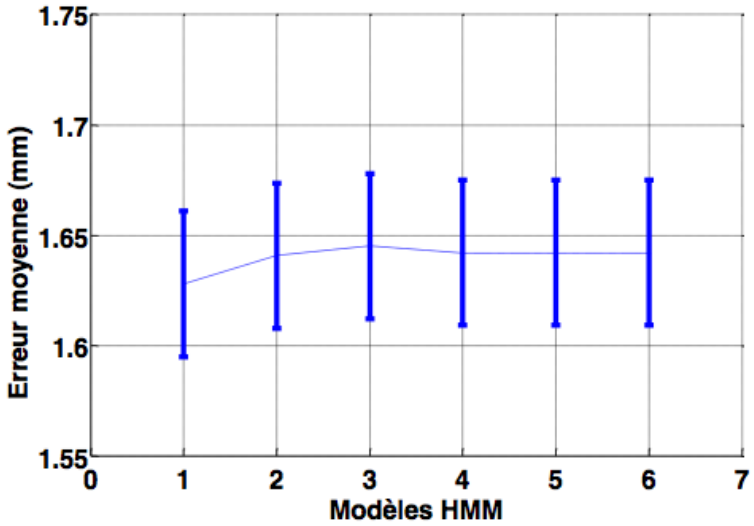


FIG. 4.8 – L'erreur moyenne (mm) entre les trajectoires de synthèse et originales pour les différents modèles HMM, dans l'ordre : 1) HMM phonème en contexte visème droit (avec les paramètres dynamiques du 1er ordre), 2) HMM phonème en contexte visème droit avec mélange de gaussiennes d'ordre 2, 3) HMM phonème en contexte visème droit avec mélange de gaussiennes d'ordre 4, 4) HMM phonème en contexte visème droit avec mélange de gaussiennes d'ordre 6, 5) HMM phonème en contexte visème droit (avec les paramètres dynamiques du 1er ordre et 2ème ordre), 6) HMM phonème en contexte visème droit avec les paramètres visuels et acoustiques. Données d'apprentissage et de test. Corpus I.

si l'on rajoute de l'information syllabique ou de l'information contextuelle triphone (gauche et droit) dans les modèles HMM. Il n'y a pas de différence significative ($p>0.05$) entre l'utilisation de l'information contextuelle visémique ou phonémique. De plus, l'utilisation de l'information visémique permet d'avoir plus de représentants pour apprendre les modèles HMM que l'utilisation de l'information phonémique. Suite à ces résultats, nous avons choisi de travailler avec les modèles HMM par phonème en contexte droit visème.

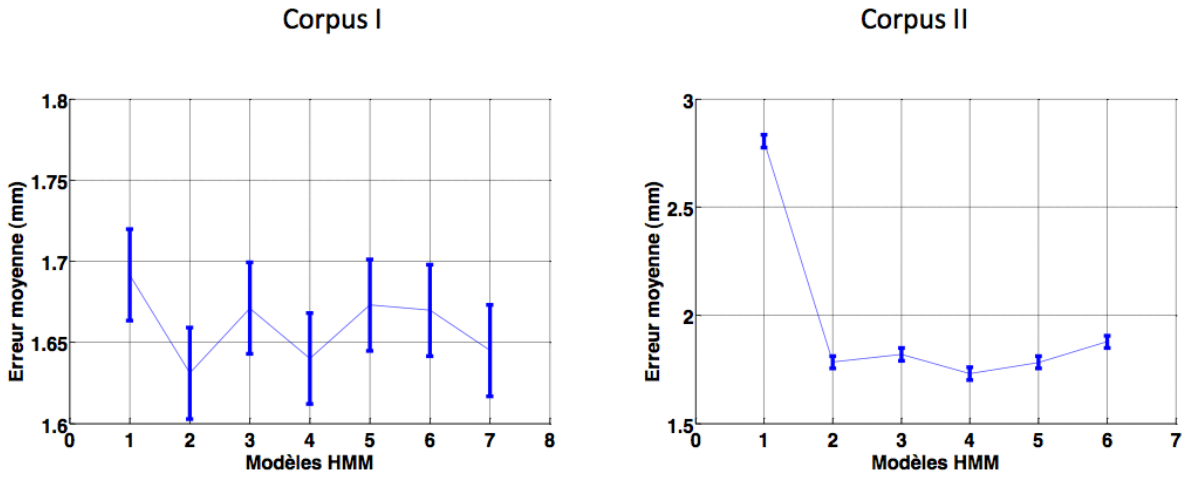


FIG. 4.9 – L'erreur moyenne (mm) de la synthèse par HMM pour les différents modèles : dans l'ordre : 1) phonème sans contexte, 2) phonème contexte droit phonème, 3) phonème contexte gauche phonème, 4) phonème contexte droit visème, 5) phonème contexte gauche visème, 6) phonème contexte gauche et droit phonème et 7) information syllabique pour le corpus I seulement. Corpus I et II. Données d'apprentissage.

4.3.4 Exécution par concaténation

L'objectif, ici, est de sélectionner dans le dictionnaire les meilleurs (au sens d'un coût à définir) diphones à concaténer. Les diphones candidats sont multi-représentés et sont représentés dans deux espaces, géométrique et articulatoire. Les candidats finaux sont choisis grâce aux coûts de sélection et de concaténation. La première étape de la synthèse par HMM fournit une préestimation des trajectoires des paramètres visuels (partition géométrique). Le coût de sélection choisi correspond à la distance entre les segments candidats et les trajectoires estimées par HMM. Cette distance correspond à la distance moyenne quadratique entre les paramètres géométriques du segment préestimé par HMM et les segments candidats. Le coût de concaténation quantifie la gêne perspective engendrée par la juxtaposition du segment avec le segment précédent. Dans notre cas, le coût correspond à la distance moyenne quadratique entre les paramètres articulatoires (pondérée par la variance du mouvement expliquée par chaque paramètre) aux points de concaténation. A noter que la concaténation originale utilise seulement le coût de concaténation pour calculer la distance entre les candidats. Ensuite, la somme des deux coûts est calculée et considérée comme le coût élémentaire du choix d'un diphone, et un algorithme de programmation dynamique recherche dans le treillis les candidats finaux, ceux réalisant la distance cumulée minimale. Bien que les segments soient sélectionnés pour correspondre au mieux avec leurs voisins, il reste encore des artefacts liés à la concaténation. Pour éviter les sauts liés à la méthodologie de concaténation, une procédure de lissage anticipatoire sur les paramètres articulatoires est appliquée, (Figure 4.11). Cette procédure compense les sauts aux frontières

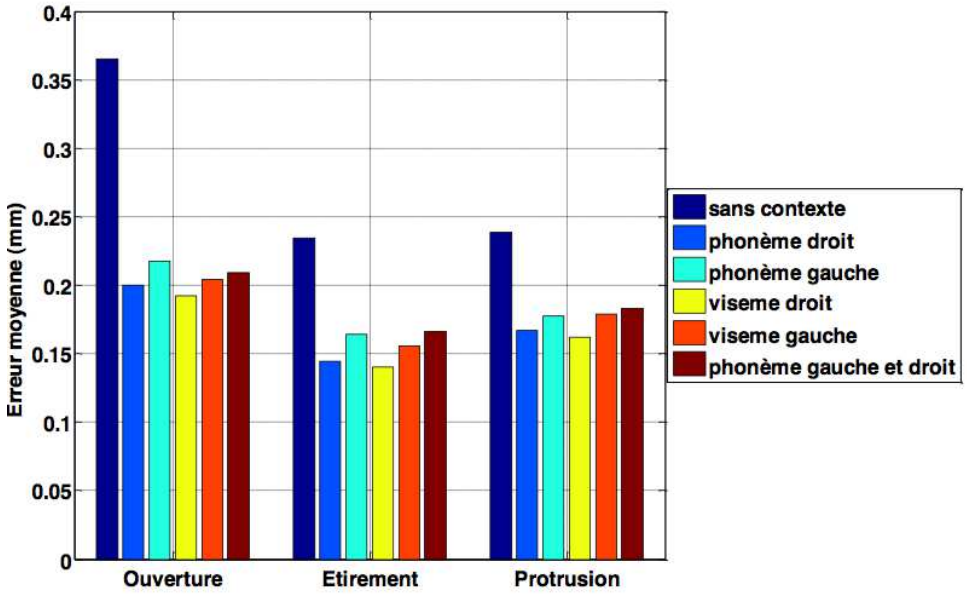


FIG. 4.10 – L'erreur moyenne (mm) de la synthèse par HMM pour les différents modèles et pour les différents paramètres : dans l'ordre : 1) phonème sans contexte, 2) phonème contexte droit phonème, 3) phonème contexte gauche phonème, 4) phonème contexte droit visème, 5) phonème contexte gauche visème, 6) phonème contexte gauche et droit. Corpus II. Données d'apprentissage.

inter-diphones : une interpolation linéaire est calculée sur le saut observé durant le diphone précédent.

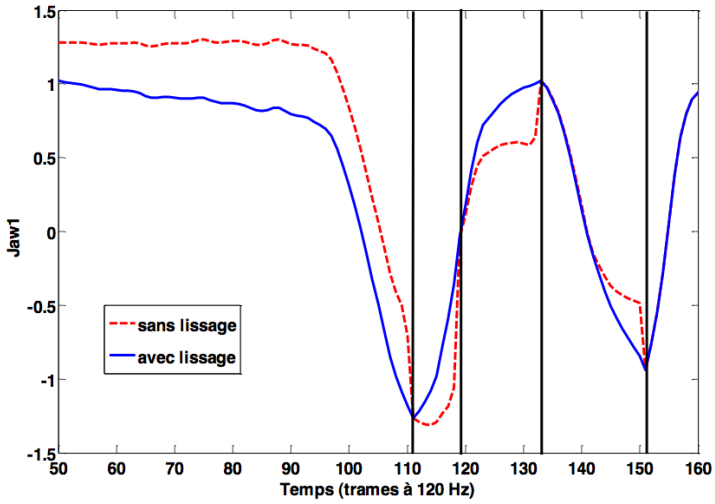


FIG. 4.11 – Un extrait de phrase. L'exemple du lissage anticipatoire pour le paramètre Jaw1.

4.4 Résultats

Les résultats de synthèse par TDA sont présentés dans les Figure 4.12, Figure 4.13. La corrélation moyenne pour les différents modèles de synthèse appliqués aux deux corpus est présentée dans les figures 4.12 et 4.13. La distorsion est moins importante dans le cas de la synthèse par

TDA que par concaténation simple indépendamment du corpus et des paramètres (p=0.05). Des exemples de génération des trajectoires géométriques sont présentés dans la Figure 4.14 pour les phrases "Du thon huileux" et "Il se garantira". Dans ces figures, la correction des trajectoires de synthèse par concaténation est très visible : grâce au coût de sélection basé HMM, les trajectoires TDA se rapprochent le plus des trajectoires d'origine. L'analyse ADL est appliquée aux trajectoires de synthèse par TDA et représentée dans les Figure 4.2, Figure 8.2, Figure 8.3 et Figure 8.4 du paragraphe 4.2 et dans l'Annexe 8.1. La synthèse par TDA fournit de l'articulation détaillée (l'intra-distance ADL : taille des ellipses de dispersion des cibles) grâce à l'exécution par concaténation. De plus, la synthèse TDA s'approche plus des données d'origine que la concaténation grâce au coût de sélection proposé basé HMM. Les résultats de synthèse par TDA ainsi que l'étude sur la synthèse par HMM et concaténation ont été publés à Interspeech Govokhina *et al.* (2006c), Govokhina *et al.* (2006b).

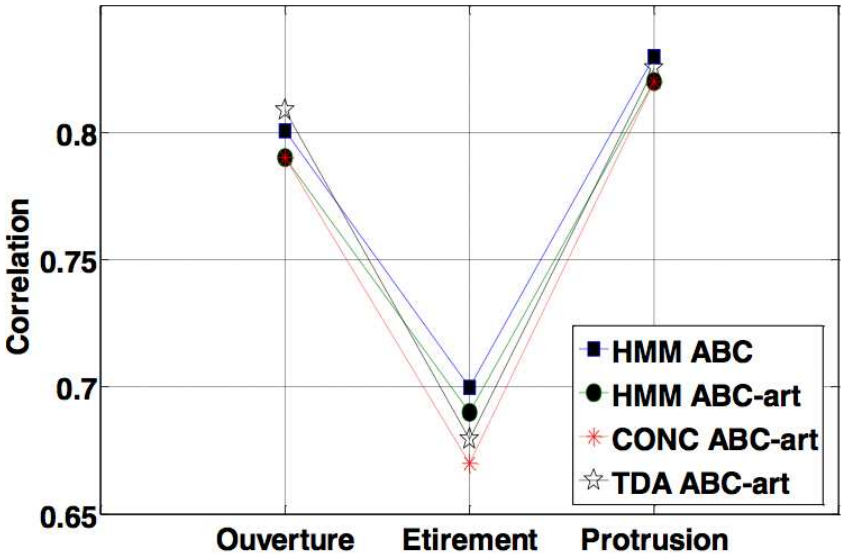


FIG. 4.12 – Les coefficients de corrélation pour les synthèses par HMM et par concaténation dans l'espace géométrique. Corpus I. Données d'apprentissage et de test. HMM ABC : paramètres géométriques de synthèse par HMM, HMM ABC-art : paramètres géométriques obtenus par la synthèse par HMM sur les paramètres articulatoires, CONC ABC-art : paramètres géométriques obtenus par la synthèse par concaténation sur les paramètres articulatoires, TDA ABC-art : paramètres géométriques obtenus par la synthèse par TDA sur les paramètres articulatoires.

4.5 Résumé

Dans ce chapitre, l'analyse détaillée de la synthèse par HMM et par concaténation ainsi que la nouvelle méthode de synthèse TDA a été présentée. Les techniques de synthèse par HMM et par concaténation sont deux méthodes très différentes par leurs principes de base de synthèse. La synthèse par concaténation décrit l'ensemble des réalisations en extension. La synthèse par HMM décrit l'ensemble en compréhension. Elle a donc intrinsèquement des capacités de génération et de surgénération. Ces deux techniques ont leurs avantages et inconvénients. La synthèse par HMM a la capacité de modéliser la coarticulation à long terme et de planifier la coarticulation, par contre, elle génère une articulation moyennée ou lissée (voir la solution proposée par Toda (Toda & Tokuda, 2007) consistant à compenser ceci par une amplification à posteriori de la

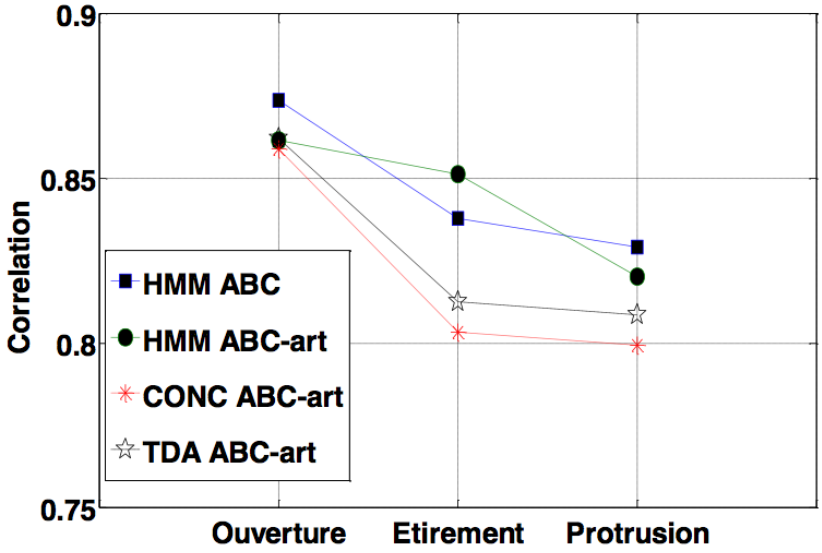


FIG. 4.13 – Les coefficients de corrélation pour les synthèses par HMM et par concaténation dans l'espace géométrique. Corpus II. Données d'apprentissage et de test. HMM ABC : paramètres géométriques de synthèse par HMM, HMM ABC-art : paramètres géométriques obtenus par la synthèse par HMM sur les paramètres articulatoires, CONC ABC-art : paramètres géométriques obtenus par la synthèse par concaténation sur les paramètres articulatoires, TDA ABC-art : paramètres géométriques obtenus par la synthèse par TDA sur les paramètres articulatoires.

dynamique des trajectoires). La synthèse par concaténation fournit une articulation détaillée car elle concatène des segments existants mais cette méthode produit des artefacts liés à la concaténation qui peuvent être très gênants visuellement. La nouvelle méthode de synthèse TDA combine les deux techniques de synthèse et en tire les avantages. Au final, le système TDA fournit une articulation détaillée grâce à la synthèse par concaténation et planifie mieux la coarticulation grâce à la préestimation par HMM.

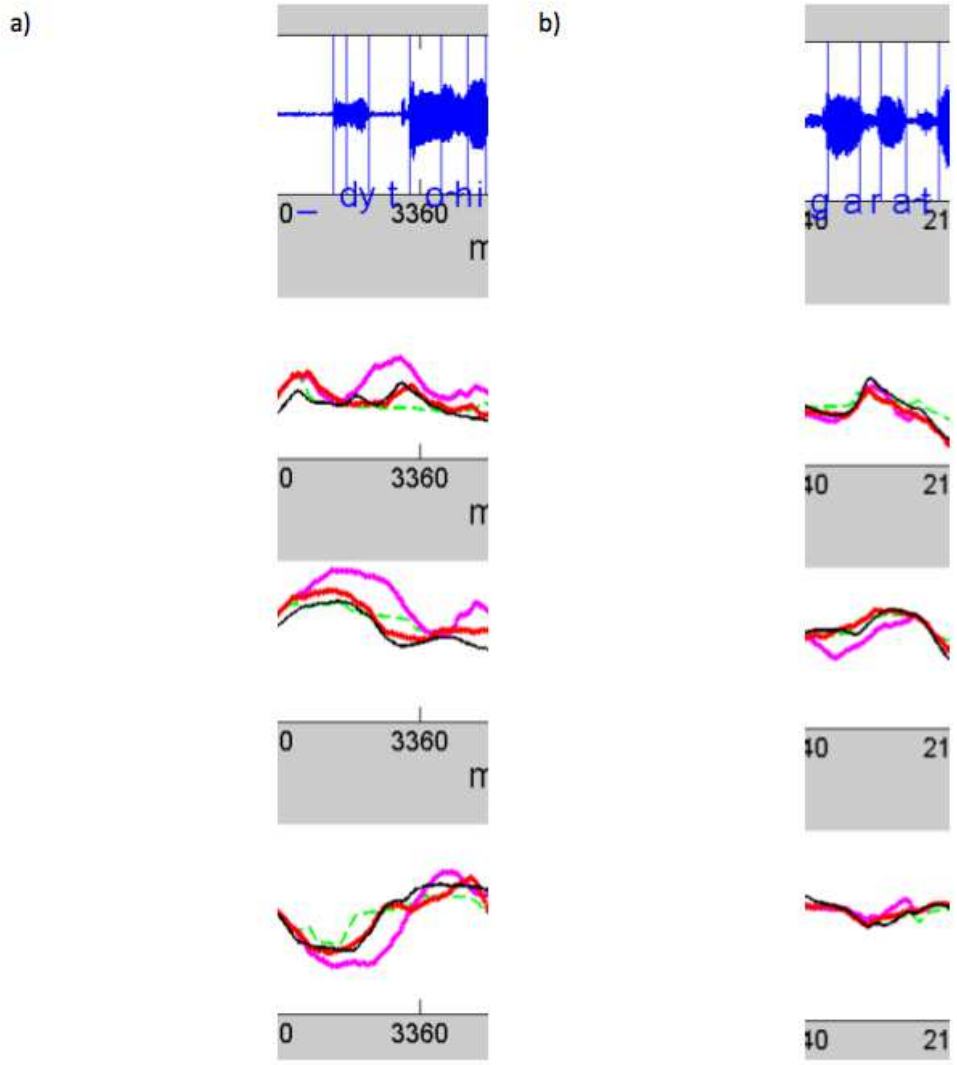


FIG. 4.14 – De haut en bas : signal audio, ouverture (mm) des lèvres, étirement (mm) des lèvres et protrusion (mm) des lèvres. Les segments des phrases a) "Du thon huileux" et b) "Il se garantira". En noire : données d'origine, en rouge : TDA, en vert : HMM et en mauve : concaténation.

Chapitre 5

Synthèse par PHMM (*Phased Hidden Markov Model*). Aspect Temporel

Jusqu'à présent nous avons travaillé sur l'aspect configurationnel de la synthèse visuelle de la parole, dans l'objectif de mieux reconstruire les trajectoires articulatoires par rapport aux trajectoires capturées. Dans ce chapitre nous proposons d'étudier l'aspect temporel de la synthèse de la parole et notamment l'asynchronie audiovisuelle.

5.1 Asynchronie audiovisuelle

Actuellement, dans la synthèse audiovisuelle de la parole les frontières entre allophones générées par la synthèse audio (Bailly, 2001) sont utilisées telles quelles comme repères de transition entre les mouvements faciaux associés à l'articulation des phonèmes, (Figure 5.1a). Les repères acoustiques ne sont pas forcément optimaux pour la synthèse des mouvements sous-jacents car :

- Certains mouvements ne laissent peu ou pas de trace dans le son (ex : mouvements préphonatoires, anticipation des mouvements labiaux dans les occlusives, etc.) ;
- Les gestes précèdent leur conséquence acoustique (Eriksson *et al.* , 2002).

Les problèmes posés par la synchronisation absolue des segments acoustiques et gestuels sont les suivants : d'une part, théoriquement les marques des frontières audio et visuelles ne doivent pas être les mêmes car il y a notamment un problème d'anticipation des caractéristiques phonétiques (Abry *et al.* , 1990), (Perkell & Chiang, 1986). Nous sommes particulièrement sensibles aux mouvements labiaux et la non prise en compte des phénomènes anticipatoires introduit souvent un décalage gênant. De ce fait, la qualité de la synthèse audiovisuelle est détériorée. D'autre part, la détermination des marques des frontières des phonèmes sur les paramètres visuels sans la connaissance de segmentation audio est difficile car ces frontières visuelles sont souvent peu marquées.

Nous proposons un algorithme qui modélise le décalage des frontières visuelles à partir des frontières audio issues d'un système de synthèse vocale classique, (Figure 5.1b).

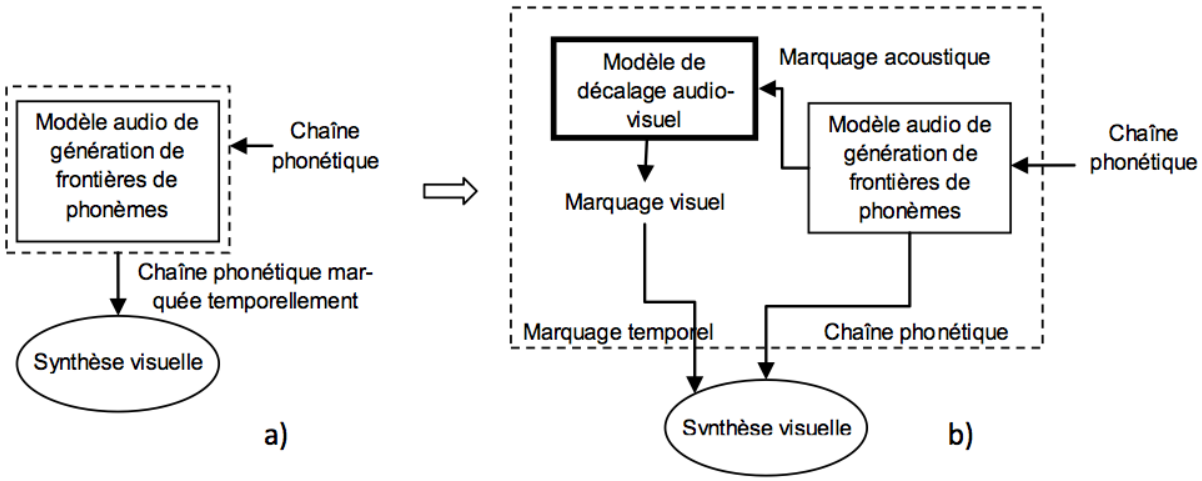


FIG. 5.1 – Principe de génération des frontières temporelles des phonèmes à partir d'une chaîne phonétique pour la synthèse audiovisuelle : a) Modèle de marquage de phonèmes basé audio (état de l'art existant) ; b) Modèle de marquage de phonème basé audio et visuel (algorithme proposé).

5.2 Segmentation temporelle en gestes visuels

5.2.1 Algorithme de repositionnement des frontières des phonèmes par l'analyse par la synthèse

L'algorithme de repositionnement des frontières de phonèmes pour la synthèse audiovisuelle proposé consiste à apprendre un modèle de décalage entre les frontières acoustiques et gestuelles de manière à ce que les modèles HMM gestuels de visèmes en contexte génèrent au mieux les trajectoires articulatoires observées. Il est composé de deux phases, Figure 5.2 :

1. La phase d'apprentissage du modèle de décalage des frontières (hors-ligne) par boucle d'analyse-synthèse de modèles HMM gestuels exploitant deux procédures principales :
   (a) Apprentissage et alignement forcé de modèles HMM gestuels
   (b) Paramétrage d'un modèle de décalage audiovisuel
2. La phase d'utilisation du modèle de décalage obtenu dans la synthèse audiovisuelle (en-ligne)

Par la suite nous détaillerons l'algorithme de repositionnement de frontières de phonèmes. Nous introduisons les notations suivantes :
- i est le numéro de l'itération
- SA est la segmentation acoustique en phonèmes
- SV_i est la segmentation gestuelle (visuelle) en phonèmes à l'itération i
- SV_{vi} est la segmentation gestuelle (visuelle) en phonèmes intermédiaire après l'alignement non forcé par Viterbi à l'itération i
- λ_i sont les modèles HMMs à l'itération i, λ_{iseg} est un modèle HMM par segment à l'itération i

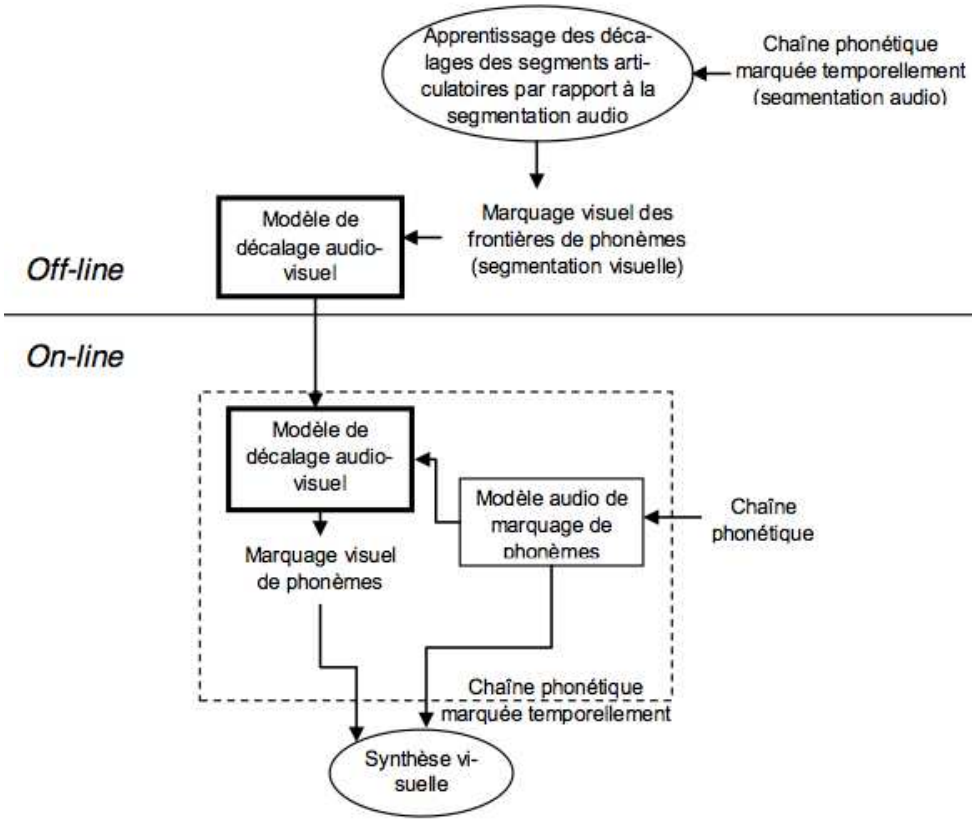


FIG. 5.2 – Schéma global de l'algorithme de repositionnement des frontières de phonèmes pour la synthèse audiovisuelle. Hors-ligne : apprentissage du modèle de décalage audiovisuel à partir de la segmentation audio et des paramètres visuels. En-ligne : utilisation du modèle de décalage audiovisuel dans la synthèse audiovisuelle.

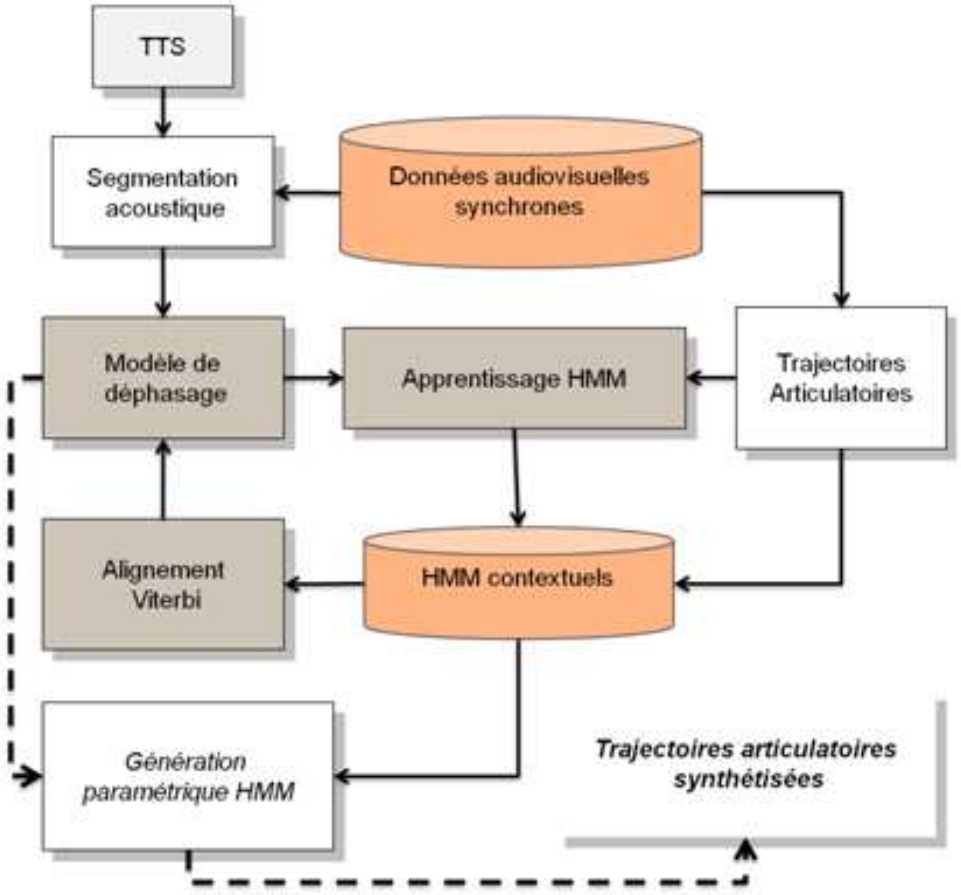


FIG. 5.3 – Schéma global de la synthèse par PHMM.

– $decm_{seg}$ est une moyenne de décalages correspondant à un segment (ici un phonème en contexte visème suivant)

L'algorithme de repositionnement des frontières des phonèmes pour la synthèse visuelle par HMM (voir la Figure 5.3) comprend les étapes suivantes, Figure 5.4 :

1. Apprentissage des modèles HMM des phonèmes en contexte visème suivant (λ_i) sur les paramètres articulatoires à partir de la segmentation audio (SA, itération 0) ou visuelle (SV_i, itération >0)
2. Réalignement non forcé par Viterbi de ces modèles HMM (λ_i) sur les trajectoires articulatoires observées. Nous obtenons à cet étape une nouvelle segmentation SV_{vi}
3. Calcul des décalages audiovisuels entre la segmentation obtenue intermédiaire SV_{vi} et la segmentation acoustique SA. Ces décalages se calculent exactement sur le début de chaque segment (phonème en contexte visème suivant). Ainsi à chaque segment correspond un ensemble de décalages.
4. Calcul d'un modèle moyen de décalage par segment. Nous obtenons, alors, un décalage moyen $decm_{seg}$ par segment.
5. Segmentation visuelle (SV_i) effectuée grâce aux modèles de décalage obtenus précédemment $decm_{seg}$ (4), c'est-à-dire que nous recalculons la segmentation visuelle (SV_i) à partir de la segmentation acoustique SA et les modèles moyens de décalage $decm_{seg}$. Contrainte : durée minimale d'un phonème doit être au moins de $D_{min} = 41$ ms pour le corpus II ou de $D_{min} = 50$ ms pour le corpus I (Nombre d'états dans un HMM (ici $N_{st} = 5$) multiplié par la durée d'une trame visuelle (ici $D_{frame} = 8,33$ ms ou $D_{frame} = 10$ ms) : $D_{min} = N_{st} * D_{frame}$).
   (a) S'il y a une stabilisation de la segmentation visuelle entre celle obtenue pendant la boucle courante (SV_i) et celle de la boucle précédente (SV_{vi}) aller à l'étape (6). Critère de stabilisation : coefficient de corrélation entre les frontières de segments entre deux itérations successives doit être > 0.99.
   (b) S'il n'y pas de stabilisation aller à l'étape (1)
6. Le modèle moyen de décalage par segment (intermédiaire) obtenu à l'étape (4) devient le modèle moyen de décalage final $decm_{seg}$.

Ainsi nous obtenons un modèle HMM λ_{seg} et un modèle moyen de décalage $decm_{seg}$ par segment, ce que l'on appelle un PHMM (*Phased Hidden Markov Model*).

Lors de la synthèse, les frontières visuelles sont générées à partir des frontières audio, de l'information phonémique et contextuelle et des modèles de décalages audiovisuels obtenus lors de l'analyse. Plus précisément, on déplace la fin d'un phonème pour obtenir la fin d'un visème, ainsi les frontières des segments se repositionnent.

5.2.2 Etude de la segmentation visuelle temporelle

L'algorithme de repositionnement de frontières de phonèmes proposé permet de segmenter les mouvements articulatoires en "allophones visuels" et de calculer les modèles HMM en contexte avec cette segmentation visuelle. Cette approche devrait améliorer les résultats de la synthèse visuelle avec segmentation acoustique, et notamment faire une synthèse par HMM plus dynamique. L'analyse de la synthèse par PHMM est présentée dans ce qui suit. La stabilisation de l'algorithme de repositionnement est atteinte à partir de la 2ème itération. Sur la Figure 5.5

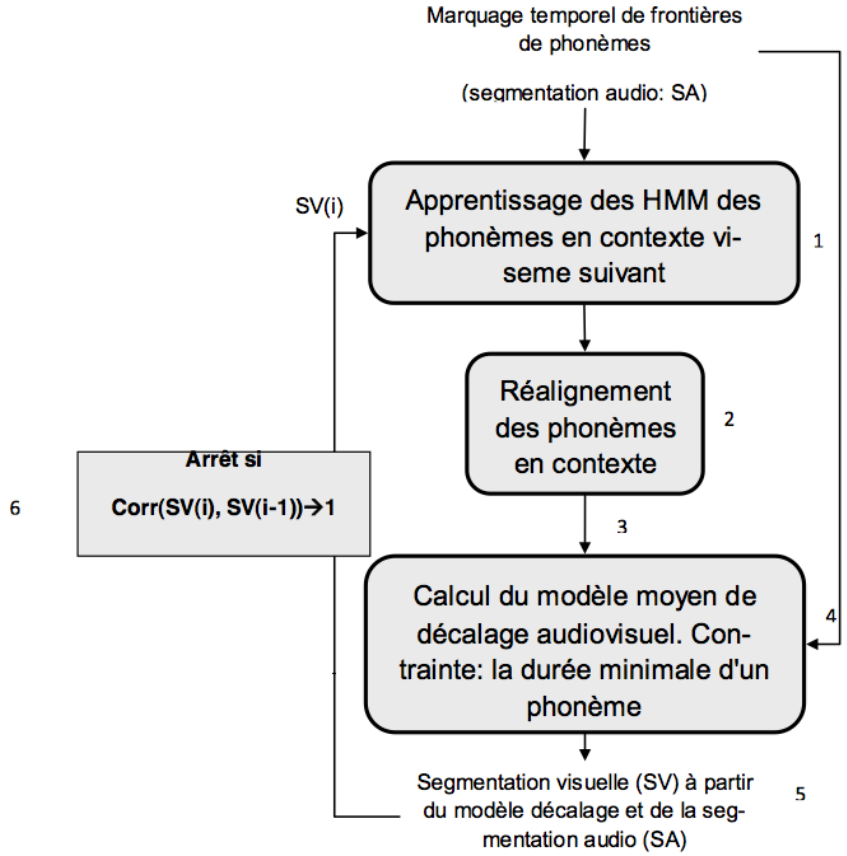


FIG. 5.4 – Exemple du procédé d'apprentissage du modèle de décalage audiovisuel basé HMM.

l'erreur moyenne est représentée en fonction du nombre d'itérations pour la synthèse par PHMM par monophone et par phonème contexte droit visème. L'erreur moyenne diminue considérablement à partir de la 2ème itération pour la synthèse sans et avec contexte pour les deux corpus, et cette différence est significative (p=0.05).

De plus, grâce à l'algorithme proposé, la segmentation automatique en "allophones visuels" est effectuée. Sur la Figure 5.6, l'augmentation et la diminution des durées de ces "allophones visuels" par rapport aux durées des allophones acoustiques corresspondants sont présentées. L'articulation du premier phone des phrases dure en moyenne 100 - 150 ms de plus par rapport au son correspondant. L'articulation du dernier phone des phrases dure aussi en moyenne 100 ms - 150 ms de plus que le son correspondant. Ces résultats montrent que les PHMM captent bien les mouvements pré-phonatoires et post-phonatoires des phrases. L'articulation des voyelles (surtout des voyelles arrondies) est plus longue que leurs traces acoustiques en moyenne de 30 à 60 ms. L'articulation des consonnes bilabiales, des post-alvéolaires et des labiodentales est aussi plus longue que leurs traces acoustiques d'environ de 10 à 40 ms. L'articulation du reste des consonnes (labiodentales, alvéolaires, vélaires, uvulaire) est plus rapide que les sons correspondants. Ces résultats confirment la théorie numérique de coarticulation de Öhman (Öhman, 1967) qui dit que les gestes vocaliques sont des gestes lents et les consonnes représentent des constrictions rapides superposées sur les gestes vocaliques.

L'exemple de génération de la phrase "Un huis-clos" est présenté dans la Figure 5.7. Ici, le geste préphonatoire du [œ] et le geste de la fin [o] sont bien présents. Le geste d'arrondissement pour le [ɥ] est mieux prédit par les PHMM que par les HMM classiques. Notez que les durées

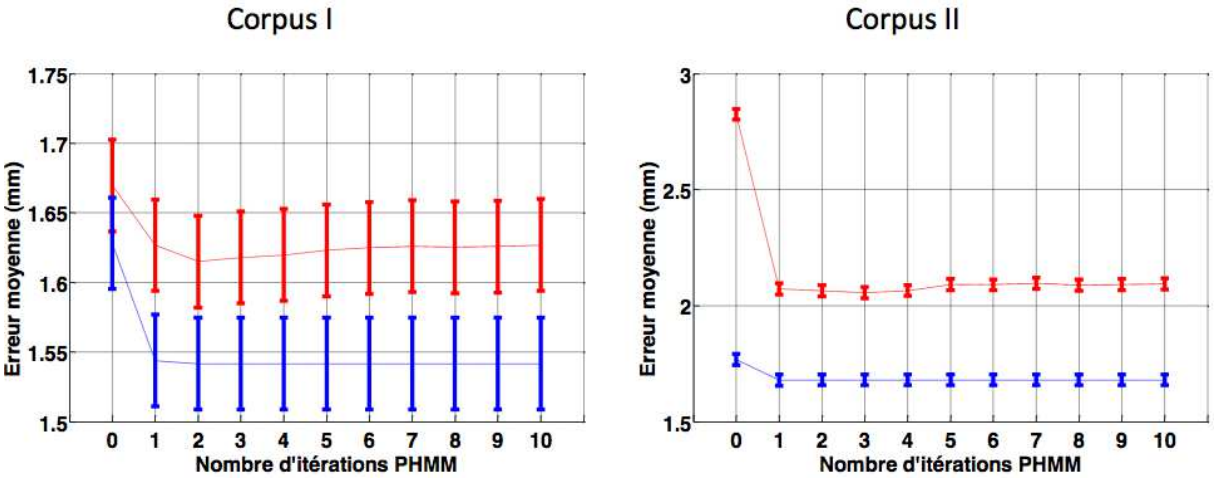


FIG. 5.5 – Erreur moyenne (mm) ($p<0.05$) pour la synthèse par HMM sans contexte et avec le contexte droit visème en fonction du nombre d'itérations de l'algorithme de décalage. Corpus I (gauche) et II (droit). Données d'apprentissage et de test.

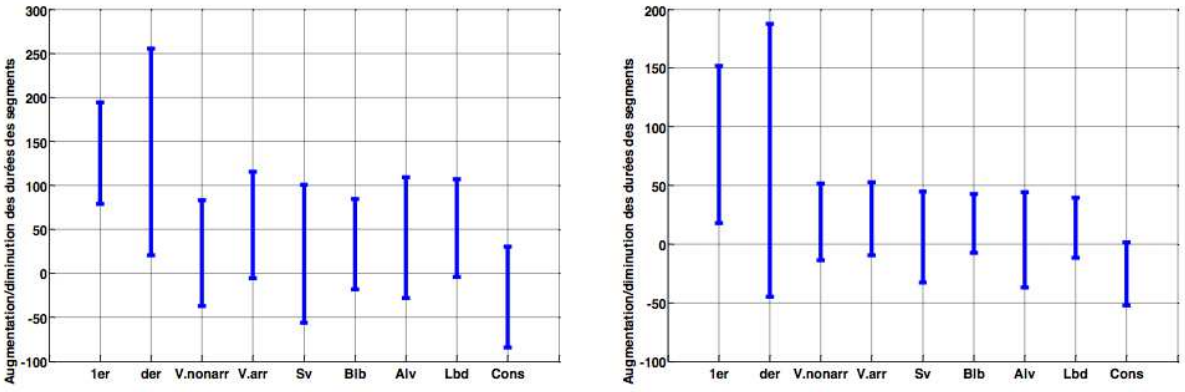


FIG. 5.6 – L'augmentation/diminution des durées des gestes articulatoires (ms) par rapport à leurs durées acoustiques. Corpus I (gauche) et II (droit). 1er : premier phonème, der : dernier phonème, V.nonarr : voyelles non arrondies, V.arr : voyelles arrondies, Blb : bilabiales, Lbd : labiodentales, Alv : post-alvéolaires, Cr : le reste des consonnes, Sv : semi-voyelles.

de l'articulation des consonnes [k] et [l] sont réduites.

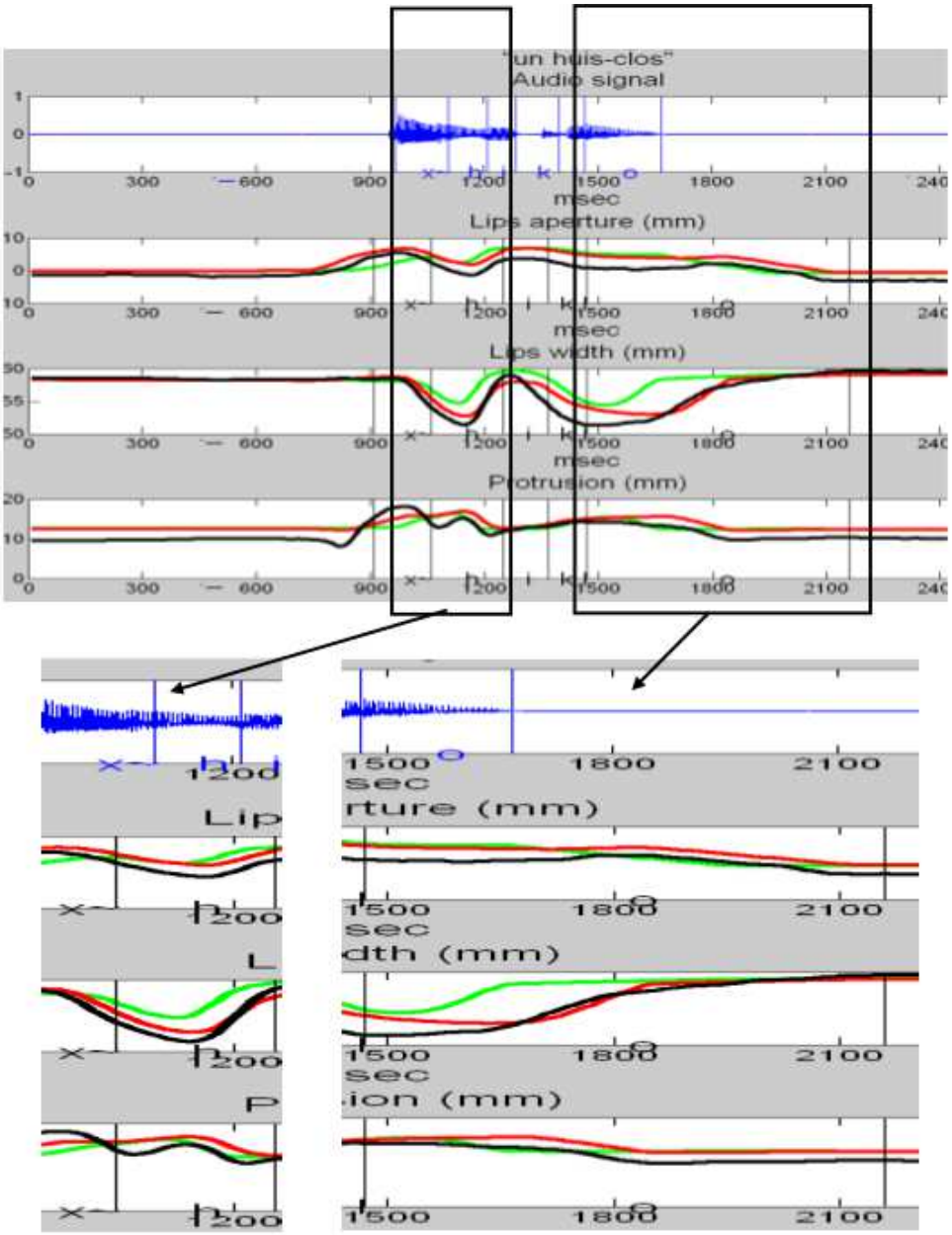


FIG. 5.7 – L'exemple de génération de la phrase "Un huis-clos". Notons l'amélioration importante de la synthèse par PHMM notamment pour le paramètre d'étirement des lèvres. En noir : trajectoires d'origine, en vert : HMM et en rouge : PHMM.

5.3 Synthèse par TDA avec la planification par PHMM

Nous obtenons trois systèmes différents basés concaténation :

- Concaténation (Conc)
- TDA avec la planification par HMM (TDA)
- TDA avec la planification par PHMM (TDA-phmm)

Les phrases des deux corpus sont synthétisées pour tous ces types de synthèse. L'erreur moyenne est calculée entre les trajectoires de synthèse et celles d'origine et présentée dans la Figure 5.8. La synthèse par TDA-phmm donne de meilleurs résultats que la synthèse par concaténation et la synthèse par TDA pour les deux corpus et pour tous les paramètres, (p=0.05). Ce résultat confirme que le système TDA profite de l'étape de planification par HMM ou PHMM et sa distorsion est moins importante que dans le cas de la synthèse par concaténation simple.

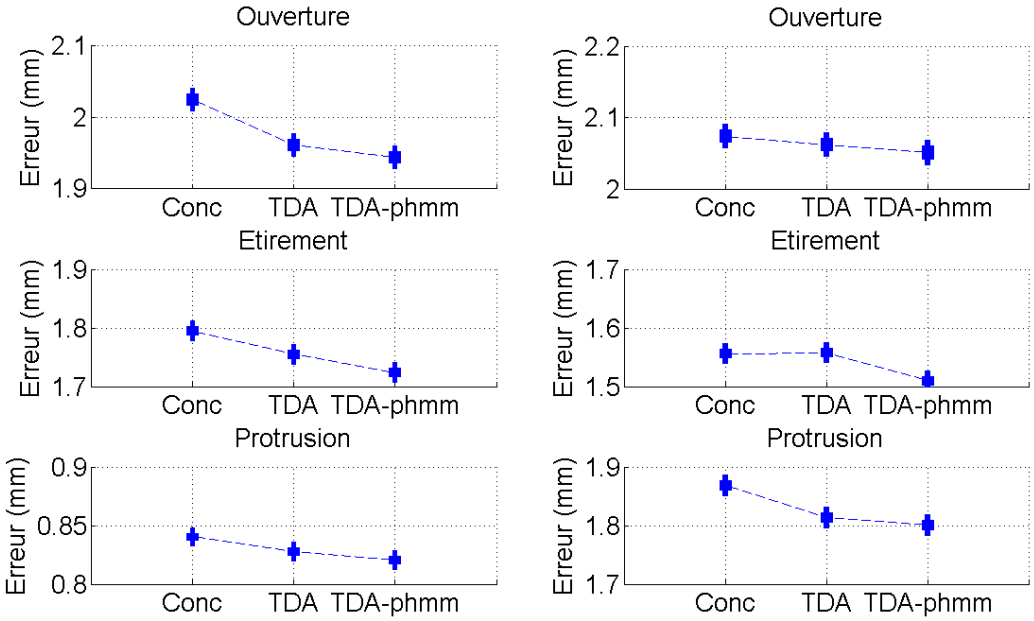


FIG. 5.8 – L'Erreur moyenne (mm) pour les systèmes de synthèse de gauche à droite : Concaténation, TDA, TDA avec la planification PHMM. Corpus I (gauche) et II (droit).

5.4 Application au Langage Parlé Complété

La synthèse et la segmentation en gestes par PHMM sont appliquées au Langage Parlé Complété en français (LPC). L'étude sur la segmentation en gestes LPC en fonction des durées acoustiques des phonèmes correspondants s'avère être très importante pour la synthèse du LPC. Dans les études précédentes, les relations s'organisent autour d'un cadre général moyen où l'attention de la cible manuelle se synchronise avec la cible acoustique de la consonne (Gibert, 2006), (Attina, 2005). Cependant, une grande variabilité accompagne ce phasage moyen et aucun modèle quantitatif de dépendance du phasage entre le geste de main et le signal acoustique en fonction du contenu phonétique n'a été proposé. On se propose ici d'appliquer le PHMM à la synchronisation entre visage et main en partant du patron de phasage moyen proposé par Gibert et Attina (Gibert, 2006), (Attina, 2005). Nous introduisons les notations suivantes dans le cadre du LPC :

- SM_{man} est la segmentation en gestes de main effectuée à la main
- SM_{auto} est la segmentation en gestes de main obtenue automatiquement à partir de la segmentation acoustique en phonèmes (SA). Le début du geste de main correspond au début de la consonne correspondant (car un geste LPC correspond à une séquence CV (consonne-voyelle))
- SM_{mod} est la segmentation en gestes de main obtenue grâce au repositionnement basé PHMM des gestes de main à partir de la segmentation automatique SM_{auto}

– $\lambda_{SM_{man}}$, $\lambda_{SM_{auto}}$ ou $\lambda_{SM_{mod}}$ sont les modèles HMM calculés avec les segmentations manuelles, automatiques et basé modèle (HMM ou PHMM) respectivement

Dans notre étude il y a plusieurs types de données sur les gestes en entrée :

– Segmentation en gestes de main LPC
  – Etiquetage manuel des cibles gestuelles (SM_{man})
  – Prédiction automatique du positionnement des cibles à partir de la segmentation acoustique suivant le phasage moyen donné par (Gibert, 2006) et (Attina, 2005) (SM_{auto})
– Les paramètres de LPC (voir la section 3.2.5)
  – Les paramètres des mouvements de la tête (6 paramètres)
  – Les paramètres des mouvements de la main (9 paramètres)
  – Les paramètres de la position de la main (6 paramètres)

En sortie, les types de données sont :

– La segmentation en gestes LPC obtenue grâce à l'algorithme de repositionnement par PHMM (SM_{mod})
– Les modèles PHMM appris par diclé sur les paramètres LPC ($\lambda_{SM_{mod}}$)

5.4.1 Reconnaissance des cibles des gestes LPC comme moyen d'évaluation de la synthèse LPC

Le code LPC est un geste de désignation : la main désigne un lieu dans l'espace egocentré avec une certaine clé de doigts. A part la position côté, le geste consiste en une constriction main-visage : la locutrice étudiée effectue effectivement un mouvement de tête anticipant les lieux sur le visage visés par la main. Le lieu dépend de la voyelle V de la série CV et la forme de la main utilisée pour effectuer ce placement de la consonne C. Les 238 phrases du corpus sont segmentées manuellement (SM_{man}) aux instants de constriction maximale en utilisant le système d'animation MOTHER de l'ICP (Revéret *et al.* , 2000) et étiquetées avec les valeurs des clés appropriées, c'est-à-dire un chiffre entre 0 et 8 pour les formes de la main. L'étiquetage en positions de la main est ajouté : un chiffre entre 0 et 5. La position de la main pour chaque cible est caractérisée comme la position 3D du doigt le plus long (référentiel 3D rattaché à la tête). Ainsi, les 3831 gestes élémentaires LPC sont obtenus et des modèles gaussiens sont calculés sur les cibles des positions et des formes de la main. Ensuite, les valeurs des trajectoires des paramètres de main sont projetées sur ces modèles gaussiens obtenus. Les taux de reconnaissance obtenus sont respectivement de 98,36% et 95,89% pour les positions et les formes de la main, voir les Table 5.1 et Table 5.2. La reconnaissance des cibles avec les modèles gaussiens calculés à partir de la segmentation manuelle est utilisée comme moyen d'évaluation de la synthèse par PHMM.

5.4.2 Résultats de la synthèse LPC par PHMM

Dans un premier temps, les modèles HMM sont appris pour les segments LPC avec la segmentation manuelle SM_{man}. La cible est supposée atteint au 3ème état de chaque HMM. Les taux de reconnaissance obtenus avec $\lambda_{SM_{man}}$ sont 95,08% et 98,27% pour les positions et les formes de la main respectivement, voir les Table 8.2 et Table 8.3, Section 8.1. Ce résultat est très intéressant car les taux de reconnaissance sont plus grands que pour les données d'origine (sauf pour les positions de la main 2,3 et 5), cela veut dire que les HMM peuvent éventuellement

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	99,35	0,00	0,00	0,00	0,00	0,00	0,00	0,65	0,00
1	1,24	98,14	0,00	0,21	0,00	0,00	0,41	0,00	0,00
2	0,47	0,00	91,94	0,00	0,71	0,00	0,24	0,00	6,64
3	0,00	0,17	0,00	99,67	0,00	0,17	0,00	0,00	0,00
4	0,00	0,28	0,28	0,00	98,89	0,55	0,00	0,00	0,00
5	0,10	0,00	0,00	0,10	0,00	99,69	0,00	0,10	0,00
6	2,42	8,38	0,00	0,00	0,00	0,00	89,20	0,00	0,00
7	0,00	0,00	0,00	0,00	0,00	3,57	1,19	95,24	0,00
8	1,22	7,93	0,00	0,00	0,00	0,00	0,00	0,00	90,85

TAB. 5.1 – Les taux de reconnaissance des formes de main. Pour une configuration segmentée des données d'origine SM_{man} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**
0	97,26	2,74	0,00	0,00	0,00	0,00
1	0,64	98,25	0,23	0,47	0,35	0,06
2	0,17	0,51	98,98	0,00	0,17	0,17
3	0,26	0,53	0,00	99,21	0,00	0,00
4	0,27	1,35	0,27	0,00	97,84	0,27
5	1,02	0,00	0,00	0,17	0,17	98,63

TAB. 5.2 – Les taux de reconnaissance des positions de main. Pour une configuration segmentée des données d'origine SM_{man} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

« corriger » les erreurs de codage commises par la codeuse : on voit que les HMM effectuent une modélisation des données qui les nettoie en cohérence avec la partition des données.

Dans un deuxième temps, les modèles HMM sont appris pour les segments LPC avec la segmentation automatique SM_{auto}. Les taux de reconnaissance obtenus avec $\lambda_{SM_{auto}}$ sont 69,58% et 78,04% pour les positions et les formes de la main respectivement, voir les Table 8.4 et Table 8.5, Section 8.1.

Enfin, l'algorithme PHMM est appliqué aux $\lambda_{SM_{auto}}$ avec la segmentation automatique. L'algorithme de repositionnement PHMM donne une nouvelle segmentation en segments LPC SM_{mod} et les modèles HMM correspondants $\lambda_{SM_{mod}}$. Les taux de reconnaissance obtenus avec $\lambda_{SM_{mod}}$ sont 75,81% et 85,91% pour les positions et les formes de la main respectivement, voir les Table 8.6 et Table 8.7, Section 8.1. Ces taux sont plus grands que dans le cas des $\lambda_{SM_{auto}}$. De plus, les frontières des segments LPC obtenues par PHMM se rapprochent de la segmentation manuelle, voir les histogrammes des décalages par rapport à la segmentation manuelle représentés dans la Figure 5.9. La configuration $\lambda_{SM_{mod-mix}}$ correspond aux modèles PHMM calculés avec l'initiation avec les SM_{man} mais toujours à partir de la segmentation automatique SM_{auto}, voir les Table 8.8 et Table 8.9, Section 8.1. Les résultats de synthèse en LPC par PHMM sont résumés dans les Figure 5.10 et Figure 5.11. La synthèse par PHMM permet de segmenter automatiquement les trajectoires articulatoires en gestes LPC et fournit de la synthèse automatique à partir de la segmentation acoustique. La segmentation automatique par PHMM fournit des meilleurs résultats que la segmentation basée seulement sur l'acoustique et fournit des durées des gestes de la main en fonction des durées des allophones acoustiques.

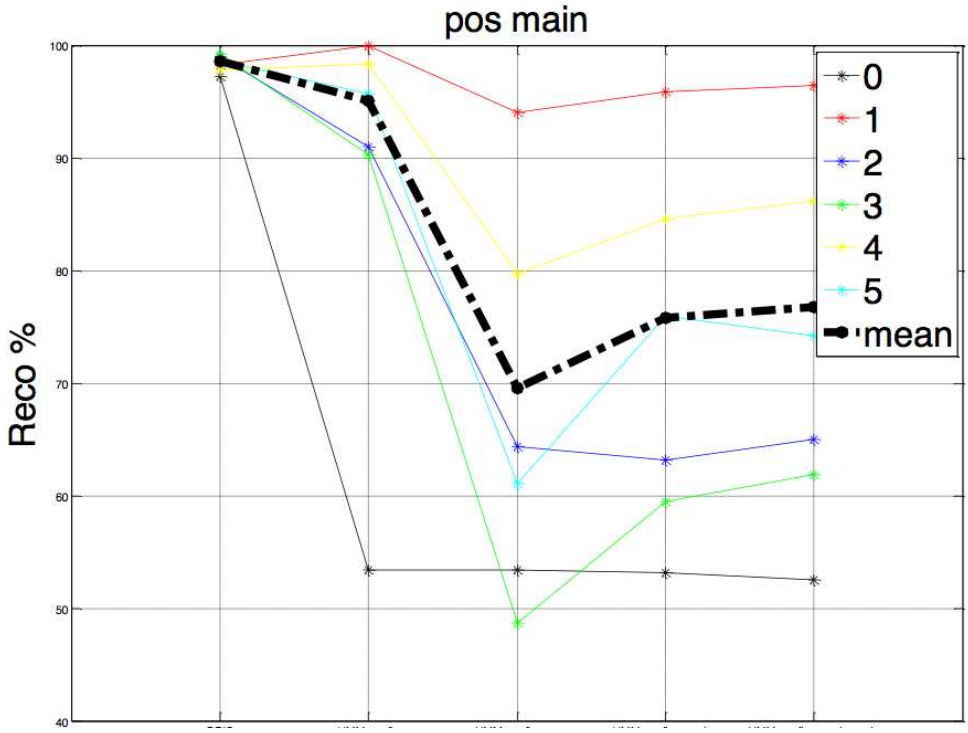


FIG. 5.10 – Les taux de reconnaissance des positions de la main pour les différents modèles et les différentes segmentations. De gauche à droite : Données d'origine, $\lambda_{SM_{man}}$, $\lambda_{SM_{auto}}$, $\lambda_{SM_{mod}}$, $\lambda_{SM_{mod-mix}}$. De bas en haut : taux de reconnaissance de 40% à 100%.

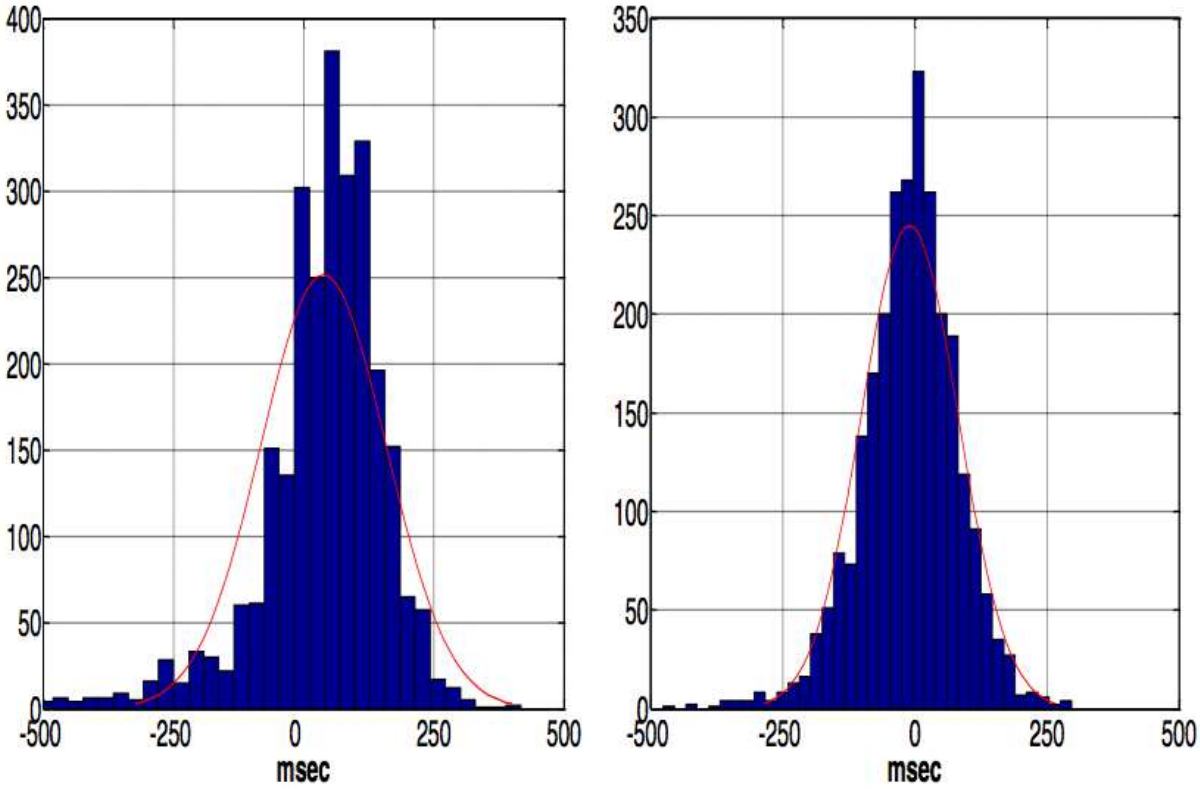


FIG. 5.9 – Histogrammes de décalage des frontières des gestes LPC SM_{auto} (à gauche) et SM_{mod} (à droite) par rapport à la segmentation manuelle SM_{man}. A gauche, les cibles gestuelles sont en retard par rapport aux cibles réalisées. A droite, le recalage est effectué. La moyenne des cibles gestuelles est en synchronie par rapport aux cibles réalisées.

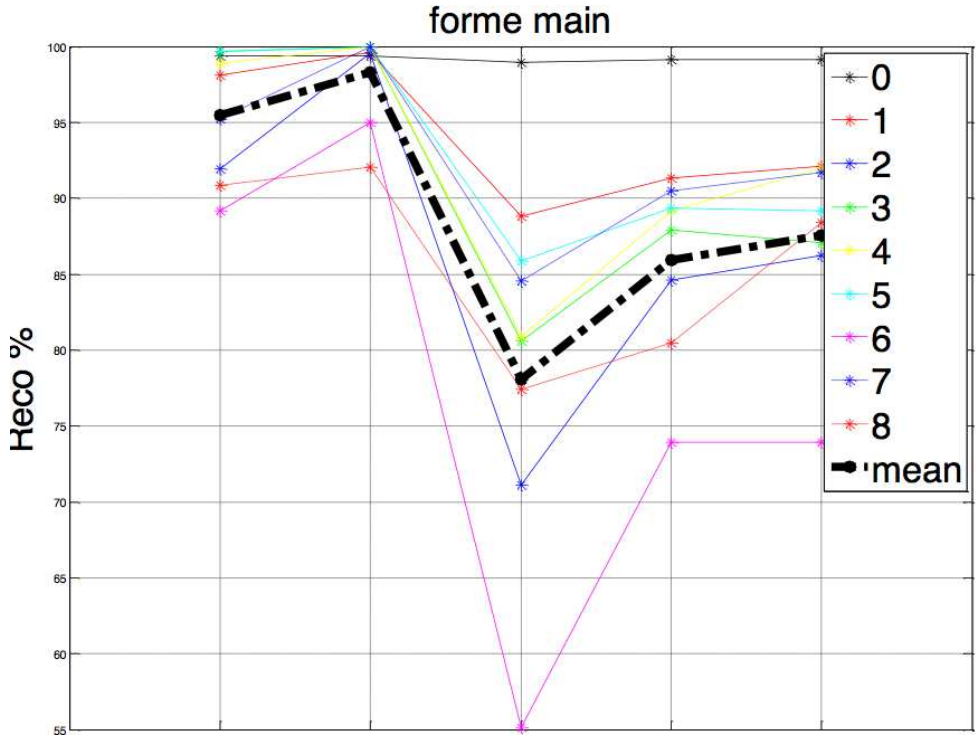


FIG. 5.11 – Les taux de reconnaissance des formes de la main pour les différents modèles et les différentes segmentations. De gauche à droite : Données d'origine, $\lambda_{SM_{man}}$, $\lambda_{SM_{auto}}$, $\lambda_{SM_{mod}}$, $\lambda_{SM_{mod-mix}}$. De bas en haut : taux de reconnaissance de 55% à 100%.

5.5 Résumé

La synthèse classique par HMM fournit une articulation correcte en moyenne mais trop lissée. Cela peut être dû au fait que les frontières entre allophones générées par la synthèse audio sont utilisées telles quelles comme repères de transition entre les mouvements faciaux associés à l'articulation des phonèmes. Or, ces repères acoustiques ne sont pas forcement optimaux pour la synthèse des mouvements articulatoires. Un algorithme de repositionnement des frontières de phonèmes pour la synthèse visuelle est proposé. Cet algorithme PHMM est un algorithme d'analyse par la synthèse qui fournit conjointement les décalages et les modèles HMM par segment phonétique. Les résultats montrent que la synthèse par PHMM améliore considérablement la synthèse par HMM et confirme la théorie numérique de coarticulation proposé par Öhman (Öhman, 1967). Le principe de synthèse par PHMM peut être appliqué aux autres modalités liées à la parole comme, par exemple, la synthèse du LPC. L'application de l'algorithme basé PHMM au LPC permet de segmenter automatiquement en gestes LPC et fournir de la synthèse automatique en LPC à partir des phonèmes marqués en durées avec une meilleure prise en compte des contextes.

Le travail sur la synthèse par PHMM a été présenté et publié dans le Workshop Interspeech 2007 (Govokhina *et al.* , 2007), dans Interspeech 2008 (Bailly *et al.* , 2008b), dans Acoustics 2008 (Bailly *et al.* , 2008a) et dans ASSISTH 2007 Beautemps *et al.* (2007). Un brevet sur le procédé de décalage par PHMM a été également déposé (FR0757063).

Chapitre 6

Evaluation

Nous avons évalué subjectivement les différents modèles de contrôle étudiés dans la thèse grâce au test MOS (*Mean Opinion Score*). Nous avons utilisé les résultats de modélisation du corpus II de 301 phrases. 10 phrases sont choisies pour le test subjectif dans le corpus de test de 100 phrases. Le déroulement et les résultats du test sont présentés dans ce chapitre.

6.1 Modèle de forme et d'apparence utilise

Le modèle de forme est celui du corpus II. Le modèle d'apparence est un modèle d'apparence actif développé par Antoine Bégault lors de son master recherche : un modèle linéaire est appris par régression des paramètres articulatoires avec des images de face libres de forme (voir AAM dans la section 2.3.1) des cibles gestuelles des allophones de toutes les phrases originales.

6.2 Modèles de contrôle utilisés

Cinq modèles de génération de trajectoires articulatoires ont été évalués subjectivement. Les modèles sont :

- Modèle Naturel
- Modèle HMM (HMM par phonème en contexte visème droit)
- Modèle PHMM (HMM et modèle moyen de decalage par phonème en contexte visème droit)
- Modèle Concaténation
- Modèle TDA (et plus exactement, TDA-phmm avec la planification PHMM)

Le modèle naturel correspond aux trajectoires d'origine capturées et animées avec le modèle d'apparence utilisé. Les modèles utilisés sont ceux présentés auparavant dans les chapitres 4 et 5.

6.3 Déroulement du test

20 sujets ont participé au test. Les sujets sont des adultes, répartition homme-femme : 60/40%, âge 33±10 ans, de professions diverses et naïfs (n'exerçant aucune activité liée à l'animation graphique).

Nous avons choisi d'effectuer un test MOS à 5 valeurs (Très insuffisant, Insuffisant, Moyen, Bon, Très bon) où les sujets répondent à la question : « Les mouvements du visage sont calculés par ordinateur à partir du son et utilisés pour animer un visage de synthèse. Sont-ils bien en cohérence avec la phrase prononcée ?»

L'interface du test a été développée en Matlab Guide, un exemple de capture d'écran est présenté dans la Figure 6.1. 60 séquences sont présentées successivement à un sujet. 10 phrases de début servent à habituer les sujets au test et ne sont pas considérées dans les résultats. Ainsi les sujets n'évaluent que 50 séquences qui correspondent aux 10 phrases avec 5 modèles. Les sujets ont la possibilité de passer à la phrase suivante (avec le bouton Suivant), de jouer la phrase courante (bouton Jouer), de valider la réponse choisie (bouton Valider) parmi les 5 réponses possibles. Les sujets ne peuvent pas rejouer les phrases et ni revenir aux phrases une deuxième fois. Les séquences sont des vidéos du type avi, avec les images de taille 480x640 pixels (réellement la taille présentée est de 410x290 pixels) et la fréquence de 50 trames par seconde. Nous avons en effet choisi de jouer dans les tests une partie de la tête parlante sans les yeux. Ce choix est fait pour que les sujets soient plus concentrés sur les mouvements labiaux et ne soient pas gênés par l'absence de mouvement de la partie haute du visage.

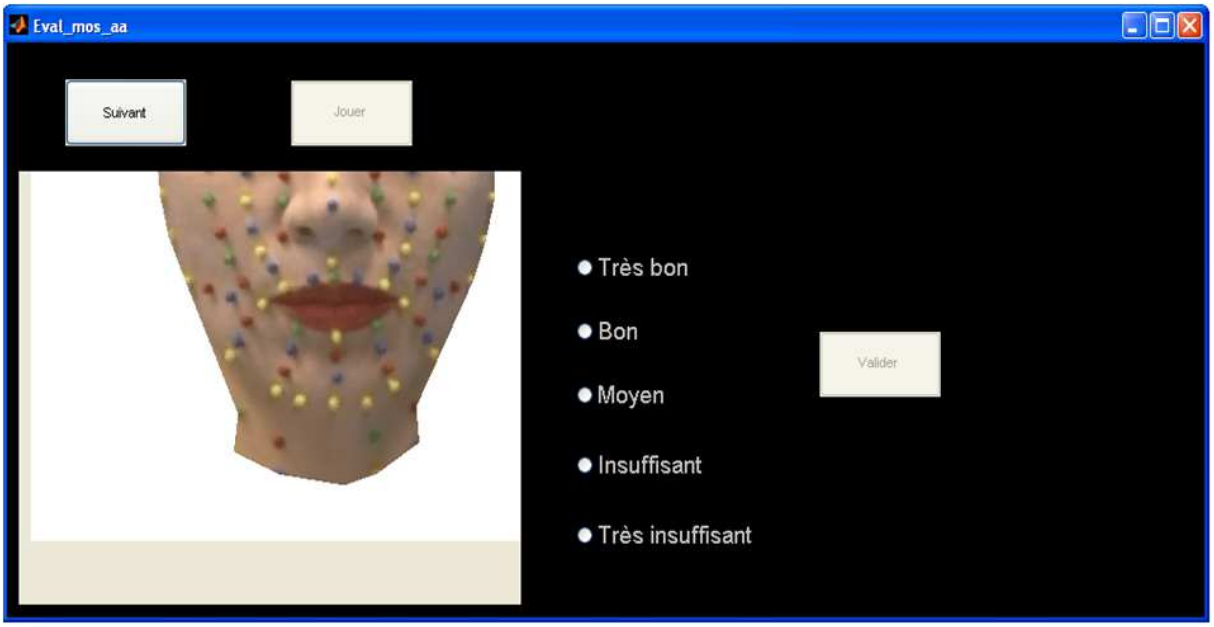


FIG. 6.1 – Une capture d'écran de l'interface du test MOS de l'évaluation subjective des différents modèles de contrôle : Nat, HMM, PHMM, concaténation et TDA.

6.4 Résultats

Les résultats d'évaluation MOS sont les notes d'évaluation (par séquence et par sujet) et le temps de réflexion des sujets (par séquence et par sujet) qui correspond au temps entre la fin d'une séquence et le moment de validation de la note. Les moyennes et les écarts-types des résultats sont présentés dans la Figure 6.2. Les résultats sont les suivants : le modèle naturel donne les meilleurs résultats, suivi par les modèles PHMM, TDA, HMM et concaténation. Nous pouvons distinguer trois groupes de modèles pour lesquels la différence est non significative : tout d'abord ce sont le modèle naturel, PHMM et TDA, ensuite ce sont PHMM, TDA et HMM et enfin le plus mauvais modèle est celui par concaténation. Les résultats obtenus du test MOS globalement confirment les résultats des tests objectifs.

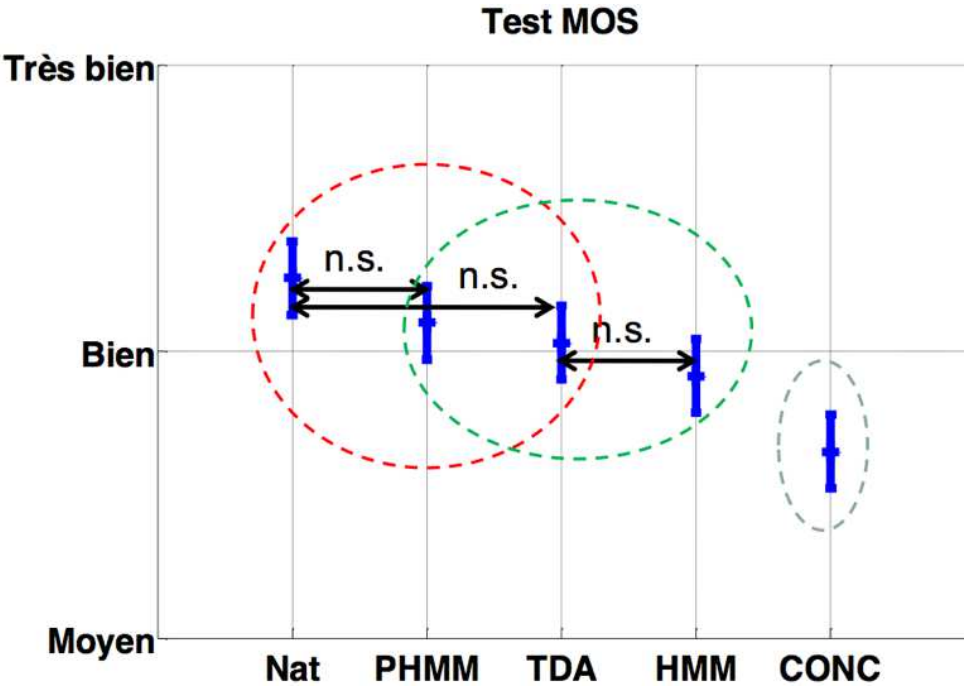


FIG. 6.2 – Résultats du test MOS du corpus II. Moyennes et écarts-types des notes des sujets pour les différents modèles de génération.

Chapitre 7

Conclusions et perspectives

Dans le travail effectué, différents modèles de synthèse visuelle de la parole ont été implémentés et comparés objectivement et subjectivement. Tout au long de la thèse, les synthèses par concaténation et par HMM sont analysées et confrontées. Le choix des ces deux modèles se fonde sur le fait qu'ils sont actuellement les plus utilisés dans la synthèse de la parole à partir du texte et qu'ils permettent d'obtenir une synthèse multimodale. La synthèse par concaténation garde la richesse des détails articulatoires lors de la synthèse car les segments concaténés viennent d'une base de données audiovisuelles capturées. Par contre, la qualité de la synthèse par concaténation est proportionnelle à la taille du corpus utilisé et le manque d'un segment approprié ou la mauvaise mise en commun de ces segments peuvent être gênants visuellement. La synthèse par HMM est une synthèse statistique-paramétrique qui peut nous donner un modèle de contrôle gérer par un ensemble de paramètres : toute la synthèse visuelle, du texte au rendu final du visage parlant, devient complètement paramétrable. D'après les tests objectifs et subjectifs, la synthèse par HMM donne, en moyenne, de meilleurs résultats que la synthèse par concaténation. Néanmoins, la synthèse par HMM a tendance à moyenner, à lisser les trajectoires articulatoires ce qui donne de l'animation moins articulée que les mouvements d'origine.

Suite à cette comparaison entre modèles, nous avons introduit un nouveau modèle de synthèse que nous avons nommé TDA qui combine les deux approches. La synthèse par TDA est fondée sur deux étapes de planification et d'exécution issues de la théorie de la phonologie articulatoire. Ainsi, pendant la phase de préestimation, les trajectoires géométriques sont planifiées grâce à la synthèse par HMM. Ensuite, les trajectoires articulatoires sont exécutées grâce à la synthèse par concaténation. La synthèse par TDA donne de meilleurs résultats que la synthèse par concaténation. Cela démontre que le TDA profite de la phase de planification par HMM. La variance de dispersion des cibles articulatoires est meilleure pour la synthèse par TDA que pour la synthèse par HMM. Cela veut dire que TDA profite aussi de la phase d'exécution par concaténation. La synthèse par TDA combine donc les avantages des deux méthodes. Certes, il y a des exceptions, surtout quand l'étape de planification par HMM est moins bonne que la concaténation simple mais en moyenne TDA est supérieur aux deux synthèses pour les deux corpus.

Nous avons également étudié l'aspect temporel dans la synthèse visuelle de la parole et nous avons proposé un nouveau modèle de synthèse PHMM. Ce modèle permet d'estimer des décalages entre les gestes articulatoires et les frontières des phonèmes. Ainsi les modèles HMM classiques sont réestimés avec les nouvelles frontières des segments phonétiques et la distorsion globale avec PHMM est significativement diminuée par rapport aux HMM classiques. La nouvelle segmentation obtenue permet de modéliser les effets pré-phonatoires et post-phonatoires par

PHMM. Nous observons aussi l'augmentation des durées des voyelles et la diminution des durées de la plupart des consonnes ce qui en accord avec la théorie numérique d'Öhman. Le modèle PHMM permet de gérer automatiquement différentes modalités liées à la parole. Nous avons donc appliqué avec succès la synthèse par PHMM à la génération automatique du LPC en français et notamment pour la gestion des relations temporelles entre les mouvements de main, de lèvres et les mouvements globaux de la tête.

Nous avons effectué le test d'évaluation subjective MOS pour comparer les modèles proposés : HMM, PHMM, concaténation et TDA. Cette évaluation montre que la synthèse par PHMM donne les meilleurs résultats (et cette différence est significative) que la synthèse par concaténation simple. La synthèse par PHMM donne aussi les meilleurs résultats que la synthèse par HMM et la synthèse par TDA donne les meilleurs résultats que la synthèse par concaténation simple. Les résultats du test subjectif sont confirmés par les résultats des tests objectifs.

En perspective, nous pouvons encore améliorer la synthèse par HMM et notamment essayer de résoudre ce problème de trajectoires moyennées grâce à la solution récemment proposée par Toda (Toda & Tokuda, 2007). Dans cette solution, une réestimation de la variance des modèles HMM est proposée. En ce qui concerne la synthèse par PHMM, ici les modèles de déphasage des frontières peuvent être plus élaborés. Pour le moment, ces modèles de déphasage correspondent à des modèles moyens par segment en contexte. L'idée serait de faire l'estimation de modèles de déphasage prenant en compte un contexte plus large comme dans le cas des calculs des modèles des durées des phonèmes pour la synthèse vocale à partir du texte. Cela suppose d'avoir accès à des bases de données de gestes plus importantes.

Une autre perspective serait d'étudier les réalisations des différents phonèmes en fonction des paramètres visuels. Par exemple, on peut supposer que le paramètre d'ouverture des lèvres sera plus important pour la synthèse des bilabials que les paramètres d'étirement ou de protrusion des lèvres, que le paramètre articulatoire des mouvements de la mâchoire sera important pour les labiodentals, etc. Ainsi, on pourrait affiner les modèles de synthèse des différents paramètres en fonction du phonème à prononcer (en introduisant, par exemple, des poids de pondération ou autres).

Dans notre travail nous avons utilisé les méthodes d'évaluation subjective, pouvant être qualifiées de basiques, comme les tests MOS (*Mean Opinion Score*) ou MPOS (*Mean Preference Opinion Score*). Une nouvelle perspective serait d'envisager d'autres tests d'évaluation subjective (réalisme, intelligibilité, acceptabilité) plus adaptés aux applications envisagées (jeux vidéo, sourds et malentendants, ...).

A long terme, le principal axe serait très certainement d'adapter les modèles de synthèse (notamment HMM) aux différents locuteurs, afin de s'approcher de plus en plus d'un modèle multi-locuteur paramétrable suivant des paramètres anatomiques et idiosyncratiques.

Chapitre 8

Annexes

8.1 Annexe A - Résultats

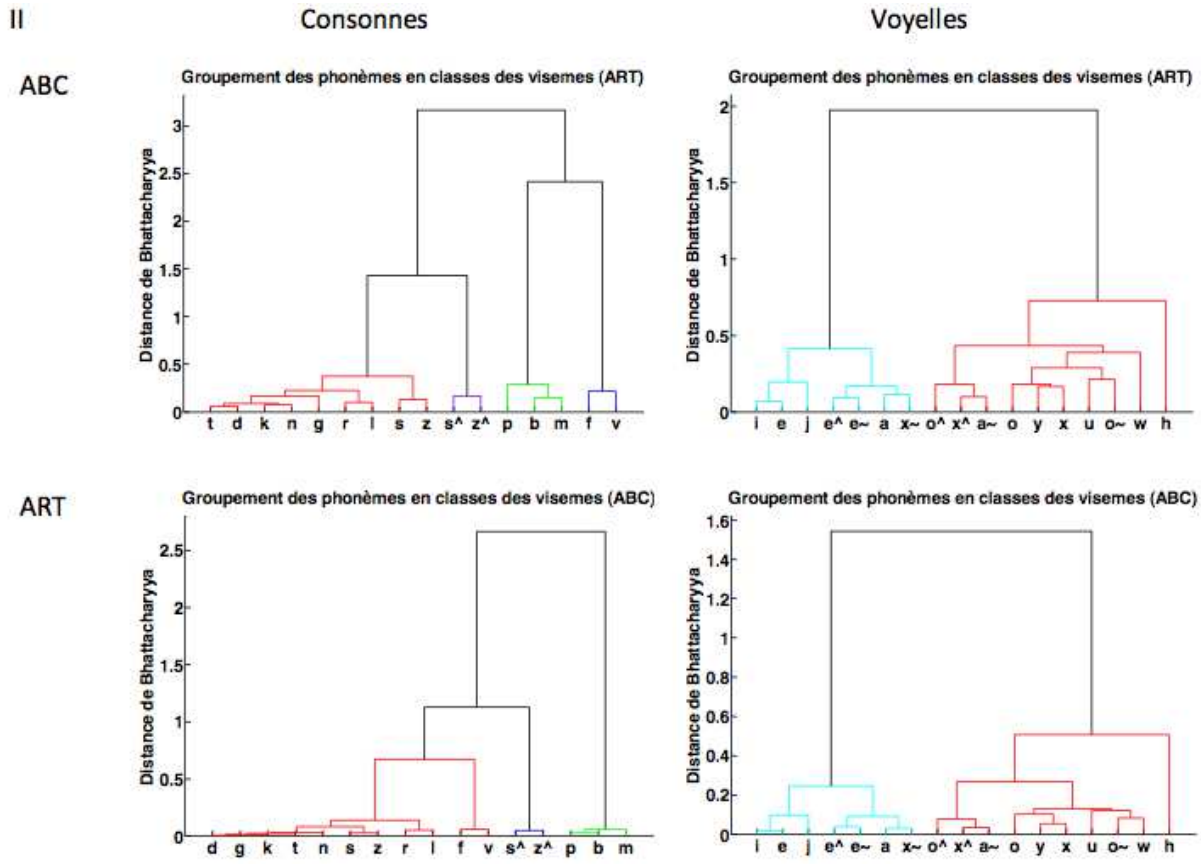


FIG. 8.1 – Groupement des consonnes et voyelles en classes des visèmes grâce à la distance de Bhattacharyya pour les paramètres articulatoires et géométriques. Corpus II.

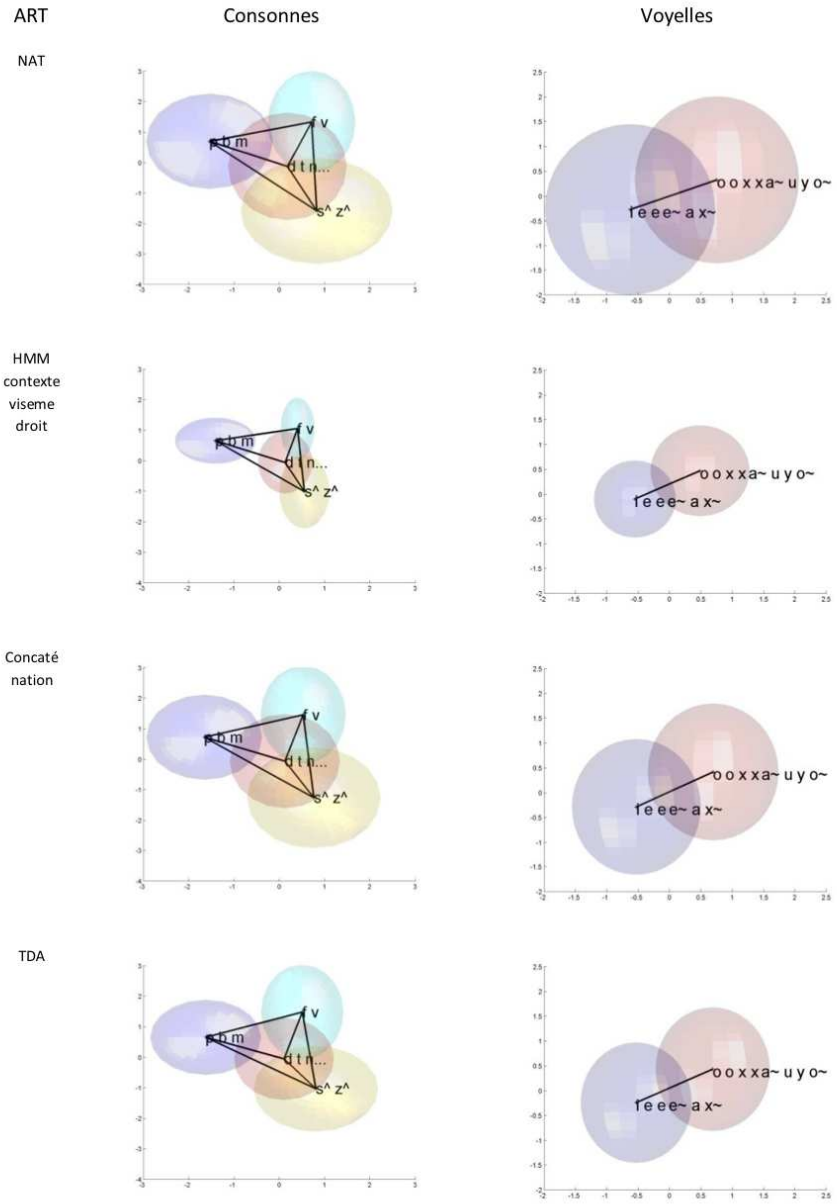


FIG. 8.2 – Ellipses de dispersion des cibles articulatoires pour les principales classes des consonnes et des voyelles avec la ADL pour les données naturelles, la synthèse par HMM, la synthèse par la concaténation et la synthèse par TDA. Corpus I.

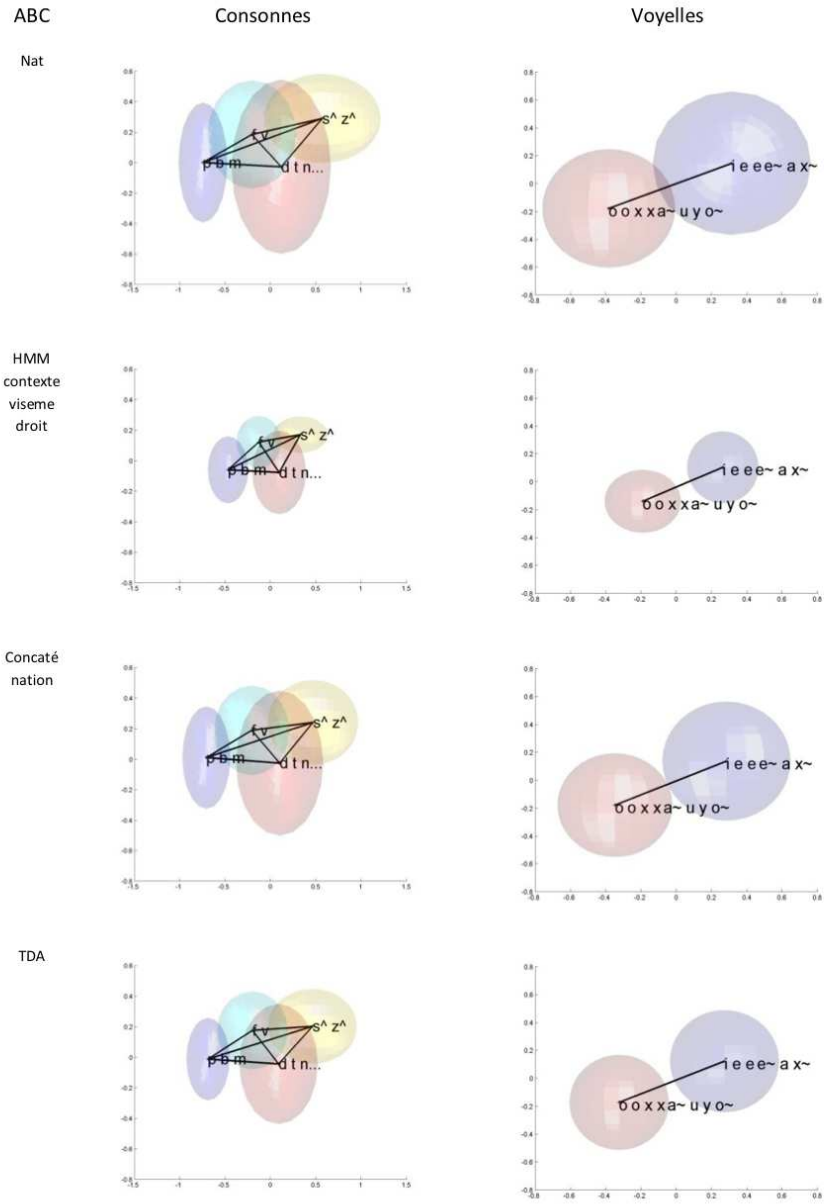


FIG. 8.3 – Ellipses de dispersion des cibles géométriques pour les principales classes des consonnes et des voyelles avec la ADL pour les données naturelles, la synthèse par HMM, la synthèse par la concaténation et la synthèse par TDA. Corpus II.

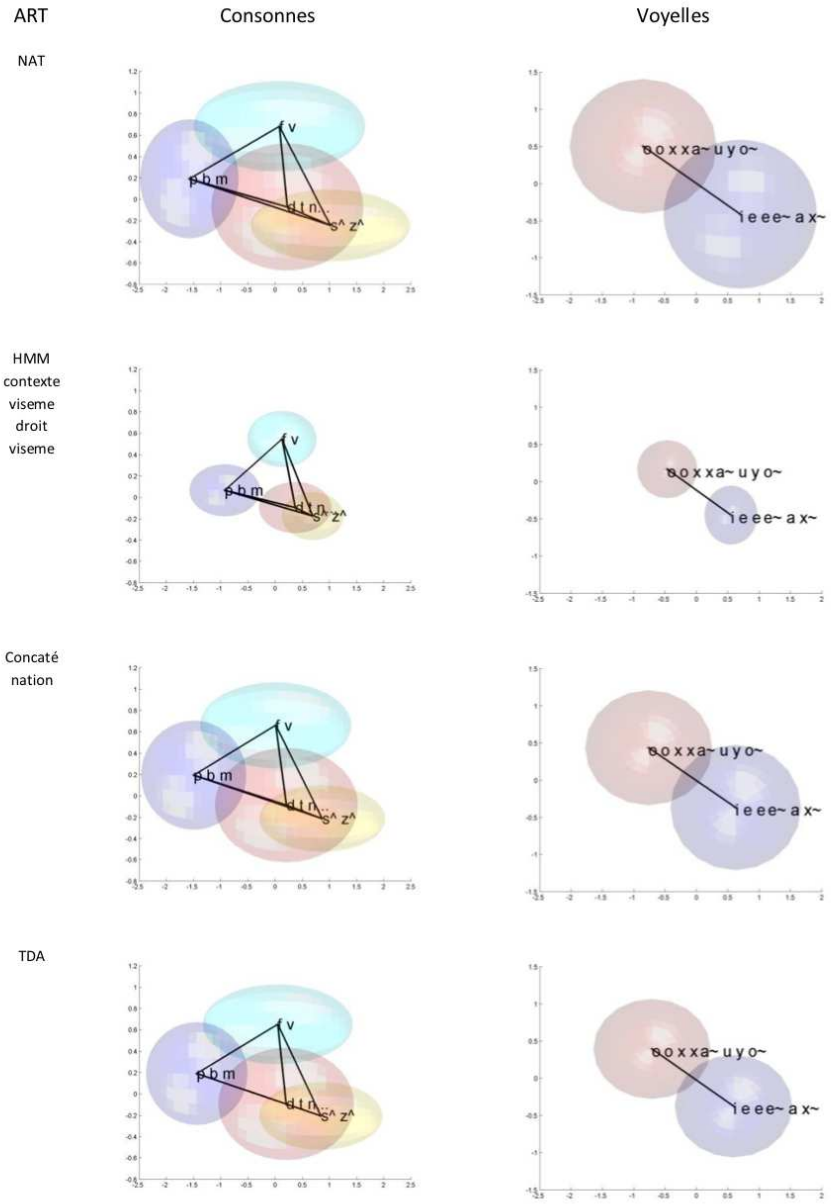


FIG. 8.4 – Ellipses de dispersion des cibles articulatoires pour les principales classes des consonnes et des voyelles avec la ADL pour les données naturelles, la synthèse par HMM, la synthèse par la concaténation et la synthèse par TDA. Corpus II.

Cibles	**Modèle**	**D_inter**	**D_intra**	**D_inter/ D_intra**	**Taux de reconnaissance %**
ABC voyelles	NAT	1,81	1	1,81	94
	HMM	1,32	0,26	5,12	95
	Conc	1,59	0,74	2,16	94
	TDA	1,59	0,56	2,86	94
ABC consonnes	NAT	0,20	1	0,20	61
	HMM	0,07	0,41	0,17	59
	Conc	0,19	0,77	0,22	60
	TDA	0,14	0,65	0,22	61
ART voyelles	NAT	0,13	1	0,13	96
	HMM	0,09	0,23	0,39	95
	Conc	0,22	0,78	0,29	96
	TDA	0,27	0,66	0,40	96
ART consonnes	NAT	0,11	1	0,11	80
	HMM	0,04	0,32	0,13	69
	Conc	0,08	0,82	0,10	78
	TDA	0,08	0,73	0,11	79

TAB. 8.1 – Les caractéristiques de la ADL (inter-distance, intra-distance et leur rapport) des consonnes et voyelles dans les espaces géométrique et articulatoire pour les données naturelles, la synthèse par HMM, la synthèse par la concaténation et la synthèse par TDA. Corpus II. Données d'apprentissage et de test. Le taux de reconnaissance est obtenu par calcul de la distance de Mahalanobis des cibles aux centres des ellipses de dispersion des visèmes.

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	99,35	0,00	0,00	0,00	0,00	0,22	0,00	0,43	0,00
1	0,41	99,59	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0,00	0,00	99,53	0,00	0,00	0,00	0,00	0,00	0,47
3	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00	0,00
4	0,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00
5	0,00	0,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00
6	1,68	3,35	0,00	0,00	0,00	0,00	94,97	0,00	0,00
7	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00	0,00
8	0,00	7,93	0,00	0,00	0,00	0,00	0,00	0,00	92,07

TAB. 8.2 – Les taux de reconnaissance des formes de main. Pour une configuration segmentée SM_{man} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**
0	53,38	46,62	0,00	0,00	0,00	0,00
1	0,00	100,00	0,00	0,00	0,00	0,00
2	0,17	8,83	91,00	0,00	0,00	0,00
3	0,00	9,47	0,26	90,26	0,00	0,00
4	0,00	1,62	0,00	0,00	98,38	0,00
5	1,71	0,00	0,00	2,56	0,00	95,73

TAB. 8.3 – Les taux de reconnaissance des positions de main. Pour une configuration segmentée SM_{man} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	98,92	0,00	0,00	0,00	0,00	0,22	0,43	0,43	0,00
1	1,45	88,82	0,00	5,80	0,21	1,45	1,24	0,21	0,83
2	1,18	1,66	71,09	2,84	2,84	2,37	1,18	0,47	16,35
3	1,00	1,49	1,82	80,60	4,81	9,12	0,00	0,33	0,83
4	1,11	2,49	0,55	9,97	80,89	3,32	0,55	0,00	1,11
5	5,68	0,93	0,21	5,16	0,83	85,86	0,72	0,31	0,31
6	6,70	26,26	0,37	3,91	0,00	3,91	55,12	2,61	1,12
7	1,19	0,00	0,00	1,19	1,19	7,14	3,57	84,52	1,19
8	1,83	7,32	1,83	6,10	1,22	3,05	0,61	0,61	77,44

TAB. 8.4 – Les taux de reconnaissance des formes de main. Pour une configuration segmentée SM_{auto} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**
0	53,38	46,62	0,00	0,00	0,00	0,00
1	1,46	94,04	2,46	1,64	0,23	0,18
2	0,68	34,63	64,35	0,34	0,00	0,00
3	0,53	46,84	3,68	48,68	0,00	0,26
4	0,54	15,95	3,51	0,27	79,73	0,00
5	0,85	3,92	6,48	27,47	0,17	61,09

TAB. 8.5 – Les taux de reconnaissance des positions de main. Pour une configuration segmentée SM_{auto} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	99,14	0,00	0,00	0,00	0,00	0,22	0,22	0,43	0,00
1	0,83	91,30	0,00	5,38	0,00	0,41	1,24	0,41	0,41
2	1,18	0,95	84,60	2,13	2,37	0,95	0,95	0,24	6,64
3	0,17	1,99	1,16	87,89	1,49	6,47	0,33	0,33	0,17
4	0,83	1,11	0,83	5,26	89,20	2,49	0,00	0,00	0,28
5	1,65	0,83	0,31	4,44	1,24	89,37	1,03	0,83	0,31
6	5,59	13,22	0,19	2,23	0,19	1,86	73,93	2,05	0,74
7	1,19	0,00	0,00	1,19	0,00	3,57	2,38	90,48	1,19
8	1,22	7,32	1,83	6,71	0,00	0,61	1,83	0,00	80,49

TAB. 8.6 – Les taux de reconnaissance des formes de main. Pour une configuration segmentée SM_{auto} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne du haut).

Seg/reco (%)	**0**	**1**	**2**	**3**	**4**	**5**
0	53,16	46,84	0,00	0,00	0,00	0,00
1	0,88	95,91	1,46	1,05	0,29	0,41
2	0,17	35,99	63,16	0,17	0,51	0,00
3	0,26	39,47	0,79	59,47	0,00	0,00
4	0,27	14,05	1,08	0,00	84,59	0,00
5	0,51	3,07	2,22	18,26	0,00	75,94

TAB. 8.7 – Les taux de reconnaissance des positions de main. Pour une configuration segmentée SM_{auto} (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

Seg/reco (%)	0	1	2	3	4	5	6	7	8
0	99,14	0,00	0,00	0,00	0,00	0,22	0,22	0,43	0,00
1	0,62	92,13	0,00	4,97	0,00	0,41	0,83	0,41	0,62
2	0,95	1,18	86,26	1,90	1,42	0,95	0,95	0,24	6,16
3	0,17	1,82	1,00	87,06	1,99	6,80	0,33	0,50	0,33
4	0,55	1,11	0,83	2,77	91,97	2,77	0,00	0,00	0,00
5	1,65	1,24	0,52	3,92	1,65	89,16	1,03	0,62	0,21
6	4,66	14,34	0,56	2,23	0,56	1,49	73,93	2,05	0,19
7	1,19	0,00	0,00	0,00	0,00	4,76	1,19	91,67	1,19
8	0,00	3,66	1,83	3,66	0,61	0,61	1,22	0,00	88,41

TAB. 8.8 – Les taux de reconnaissance des formes de main. Pour une configuration segmentée $SM_{auto-mix}$ (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

Seg/reco (%)	0	1	2	3	4	5
0	52,53	47,47	0,00	0,00	0,00	0,00
1	0,64	96,43	1,35	0,76	0,41	0,41
2	0,17	34,30	65,03	0,17	0,34	0,00
3	0,00	37,89	0,26	61,84	0,00	0,00
4	0,81	12,16	0,81	0,00	86,22	0,00
5	1,71	2,56	2,22	19,28	0,00	74,23

TAB. 8.9 – Les taux de reconnaissance des positions de main. Pour une configuration segmentée $SM_{auto-mix}$ (colonne de gauche), le nombre de représentants reconnus par configuration est représenté (ligne de haut).

8.2 Annexe B - Les algorithmes d'apprentissage et de synthèse par HMM

8.2.1 Les Modèles de Markov

Une introduction théorique des HMMs est présentée dans cette partie, (Rabiner, 1989).

Notations

n	Le nombre d'états du modèle de Markov caché
$S = \{s_1, s_2, ..., s_n\}$	Les états du HMM
A	La matrice des probabilités de transitions entre les états
$a_{ij}, i, j \in [1, n]$	Un élément de A
B	La matrice des probabilités d'observation
$b_j, j \in [1, n]$	Un élément de B
π	Le vecteur des probabilités initiales du HMM
$\lambda = (A, B, \pi)$	Un HMM
T	La longueur d'une séquence observée
$O = O_1...O_t...O_T$	Une séquence observée
$q_1...q_t...q_T$ avec $q_t \in S$	Une suite des états qui a émis une séquence

Modèles de Markov observables

Un ***modèle stochastique observable*** est un processus aléatoire qui peut changer d'état s_i, $i = 1, ..., n$ au hasard, aux instants $t = 1, 2, ..., T$. Le résultat observé est la suite des états dans lesquels il est passé. Chaque séquence est émise avec une probabilité $P(S) = P(s_1, s_2, ..., s_T)$. Pour calculer $P(S)$, il faut se donner la probabilité initiale $P(s_1)$ et les probabilités d'être dans état s_t, connaissant l'évolution antérieure.

Un processus stochastique est ***markovien*** (ou de ***Markov***) si son évolution est entièrement déterminée par une probabilité initiale et des probabilités de transitions entre états. Autrement dit, en notant $(q_t = s_i)$ le fait que l'état observé à l'instant t est s_i

$$\forall t, P(q_t = s_i \mid q_{t-1} = s_j, q_{t-2} = s_k ...) = P(q_t = s_i \mid q_{t-1} = s_j) \tag{8.1}$$

d'où :

$$P(q_1 ... q_T) = P(q_1) \times P(q_2 \mid q_1) \times ... \times P(q_T \mid q_{T-1}) \tag{8.2}$$

Pour simplifier les processus de Markov auxquels nous avons affaire sont généralement ***stationnaires*** c'est-à-dire que leurs probabilités de transition ne varient pas dans le temps. Ainsi une matrice de probabilité de transitions $A = [a_{ij}]$ est définie telle que :

$$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i), 1 \leq i \leq n, 1 \leq j \leq n \tag{8.3}$$

avec :

$$\forall i, j, a_{ij} \geq 0, \forall i, \sum_{j=1}^{j=n} a_{ij} = 1 \tag{8.4}$$

Un modèle de Markov observable λ est un processus stochastique observable, markovien et stationnaire. Un tel modèle est décrit par :

- son nombre d'états n
- sa matrice de transitions A
- son vecteur des probabilités initiales π

$$\lambda = (A, \pi) \tag{8.5}$$

HMMs

Le modèle de Markov caché généralise le modèle de Markov observable car il produit une séquence en utilisant deux suites de variables aléatoires ; l'une cachée et l'autre observable.

- La suite cachée correspond à la suite des états $q_1, q_2, ..., q_T$, notée $Q(1 : T)$, où les q_i prennent leur valeur parmi l'ensemble des n états du modèle $s_1, s_2, ..., s_n$.
- la suite observable correspond à la ***séquence des observations*** $O_1, O_2, ..., O_T$, notée $O(1 : T)$.

Un HMM est donc notée $\lambda = (A, B, \pi)$ et se défini par :

- Ses états, en nombre n, qui composent l'ensemble $S = \{s_1, s_2, ..., s_n\}$. L'état où se trouve le HMM à l'instant t est noté $q_t (q_t \in S)$.
- Une matrice A de probabilités de transition entre les états : a_{ij} représente la probabilité que le modèle évolue de l'état i vers l'état j

– Une matrice B de probabilités d'observation des symboles dans chacun des états du modèle. C'est-à-dire qu'à chaque instant donné on observe une réalisation d'une variable aléatoire suivant la loi de probabilité associée à l'état visité à cet instant. Ces lois donc appelées les ***lois d'émission***.
– Un vecteur π de probabilités initiales.

Suivant la typologie des lois d'émissions les HMMs discrets et les HMMs continus sont distingués.

HMMs discrets Un HMM est discret si les lois d'émission sont discrètes et les variables aléatoires correspondantes ont des valeurs dans le même ensemble fini d'observations possibles. Cet ensemble est souvent appelé "alphabet". Si l'alphabet est noté comme $V = V_1, V_2, ..., V_M$, ces lois sont décrites par une matrice B de taille $(n \times M)$:

$$B = b_i(k) \tag{8.6}$$

avec $b_j(k)$ représentant la probabilité que l'on observe le symbole v_k alors que le modèle se trouve dans l'état j, soit :

$$b_j(k) = P(O_t = v_k \mid q_t = s_j), 1 \leq j \leq n, 1 \leq k \leq M \tag{8.7}$$

HMMs continus Un HMM est continu si les lois d'émission sont absolument continues sur $\mathbb{R}^N$. Une densité continue sur $\mathbb{R}^N$ est associée à chaque état q_i de λ notée $f_i(\cdot)$. On calcule donc la vraisemblance $p(O_t \mid q_t = s_i) \triangleq f_i(O_t)$, qui est appelée par analogie ***vraisemblance d'émission***.

L'hypothèse d'indépendance s'écrit maintenant :

$$p\left(O_1, O_2, \ldots, O_T \mid q_1 = s_i, q_2 = s_j, \ldots q_T = s_n\right) = \\ p\left(O_1 \mid q_1 = s_i\right) p\left(O_2 \mid q_2 = s_j\right) \ldots p\left(O_T \mid q_T = s_n\right) \tag{8.8}$$

Ensuite on se pose la question du choix des densités continues pour la modélisation. Souvent, en première approximation, on choisit les ***densités monogaussiennes multidimensionnelles*** :

$$f_i(O_t) = \frac{1}{(2\pi)^{N/2} \left|\Sigma_i\right|^{1/2}} \exp\left(-\frac{1}{2}\left(O_t - \mu_i\right)^\top \Sigma_i^{-1}\left(O_t - \mu_i\right)\right) \tag{8.9}$$

où μ_i et Σ_i représentent respectivement le vecteur moyen et la matrice de covariance associés à l'état q_i. Le modèle dans ce cas est paramètré ainsi :

$$\lambda_c = \{\pi, A, \mu_i, \Sigma_i \mid i = 1, \ldots, n\} \tag{8.10}$$

Le choix d'une telle distribution est justifié par le ***théorème limite central*** et par le fait, que l'estimation des paramètres de cette distribution est beaucoup plus simple que pour les autres distributions. D'un autre côté ce choix est, quand même, assez restrictif : on ne peut pas approcher n'importe quelle distribution par une gaussienne. C'est pour cela que les ***mélanges de gaussiennes*** sont utilisés préférentiellement.

Modèle de mélanges de gaussiennes GMM Un modèle de mélanges de gaussiennes (GMM *Gaussian Mixture Model*) peut être construit comme suit :

1. On tire une variable aléatoire discrète η à valeurs dans $\{1, 2, \ldots, K\}$ où K désigne le nombre de ***composantes du mélange***, on note $c_k = P\{\eta = k\}$ pour $k = 1, 2, \ldots, K$ les probabilités respectives de tirer chacune des composantes.

2. Conditionnellement à l'événement $\{\eta = k\}$, O est une réalisation de variable aléatoire, qui est distribuée selon la loi gaussienne multidimensionnelle $\mathcal{N}_N(\mu_k, \Sigma_k)$ dont la densité $g_k(O)$ est définie par (8.9).

On peut montrer, que O est une réalisation de variable aléatoire de densité :

$$g(O) = \sum_{k=1}^{K} P(\eta = k) p(O \mid \eta = k) = \sum_{k=1}^{K} c_k g_k(O) \tag{8.11}$$

Maintenant, on associe à chaque état i d'un HMM une densité de mélange de gaussiennes, qui s'écrit comme :

$$f_i(O_t) = \sum_{k=1}^{K} \frac{c_{ik}}{(2\pi)^{N/2} |\Sigma_{ik}|^{1/2}} \exp\left(-\frac{1}{2}(O_t - \mu_{ik})^\top \Sigma_{ik}^{-1}(O_t - \mu_{ik})\right) \tag{8.12}$$

avec les contraintes :

$$\begin{cases} c_{ik} \geq 0, & \forall i, k \\ \sum_{k=1}^{K} c_{ik} = 1, & \forall i \end{cases} \tag{8.13}$$

Un HMM à densités continues est alors paramétrisé comme :

$$\Lambda_G = \{\pi, A, \mu_{ik}, \Sigma_{ik}, c_{ik} \mid i = 1, \ldots, n, k = 1, \ldots, K\} \tag{8.14}$$

Bien qu'un modèle de mélanges de gaussiennes soit décrit par un mécanisme très simple, sa distribution approche bien n'importe quelle autre distribution, si le nombre de composantes du mélange est suffisamment grand. Ces distributions sont alors largement utilisées en reconnaissance automatique de la parole.

8.2.2 La théorie de la synthèse visuelle de la parole par HMMs

Dans cette partie la théorie de la synthèse visuelle de la parole par HMM est présentée. Dans en premier temps les problématiques de la synthèse visuelle et le principe de la méthode de synthèse sont donnés. Dans en deuxième temps les principaux algorithmes d'apprentissage et de synthèse de paramètres visuels sont donnés.

Problématiques de la synthèse visuelle par HMM

Dans cette partie la théorie de la synthèse de la parole par HMM est présentée. Dans en premier temps les problématiques de la synthèse et le principe de la méthode de synthèse sont donnés. Dans en deuxième temps les principaux algorithmes d'apprentissage et de synthèse de paramètres visuels sont donnés. L'apprentissage et la synthèse est appliquée aux paramètres visuels.

1. **Choix des unités de modélisation.** Dans la plupart des cas un HMM est construit pour une unité phonétique. Les unités phonétiques sont choisies selon deux critères : l'un est basé sur des études phonétiques et l'autre dépend de la quantité de données disponibles.
2. **Trois problèmes pour HMMs.** Les tâches associées aux HMMs sont généralement formulées sous la forme de trois problèmes (Rabiner, 1989).

   Problème 1 (Calcul de la vraisemblance d'une séquence observée). *Étant donné la modèle λ, défini par (8.10), comment calcule-t-on $p(O|\lambda)$, la vraisemblance de la séquence d'observations $O = O_1, O_2, ..., O_T$?*

   Puisque l'ensemble de tous les événements $q = q_1, q_2, ..., q_T$ possibles est une partition de l'espace probabiliste, la vraisemblance peut être réécrite comme :

$$p(O, \lambda) = \sum_q p(O|q, \lambda) P(q, \lambda) \tag{8.15}$$

où la somme est faite sur toutes les séquences d'états q possibles. En utilisant l'hypothèse d'indépendance (8.8) :

$$p(O|q, \lambda) = \prod_{t=1}^{T} p(O_t|s_t = q_t, \lambda) = \prod_{t=1}^{T} f_{q_t}(O_t) \tag{8.16}$$

Ensuite, en utilisant la définition de la probabilité conditionnelle et l'hypothèse (8.1), le deuxième terme de (8.15) peut être réécrit comme :

$$\begin{aligned} P(q \mid \lambda) &= P(s_1 = q_1 \mid \lambda) \prod_{t=2}^{T} P(s_t = q_t \mid s_{t-1} = q_{t-1}, \ldots, \lambda) \\ &= P(s_1 = q_1 \mid \lambda) \prod_{t=2}^{T} P(s_t = q_t \mid s_{t-1} = q_{t-1}, \lambda) \\ &= \pi_{q_1} \prod_{t=2}^{T} a_{q_t q_{t-1}} \end{aligned} \tag{8.17}$$

en déduisant enfin l'expression de la vraisemblance $p(O \mid \lambda)$

$$p(O \mid \lambda) = \sum_q \left[\pi_{q_1} f_{q_1}(O_1) \prod_{t=2}^{t} a_{q_t q_{t-1}} f_{q_t}(O_t) \right] \tag{8.18}$$

Pour calculer cette vraisemblance en utilisant directement l'équation (8.18), il faut effectuer $(2T-1)n^T$ multiplications (chaque terme de la somme demande $2T-1$ multiplications et il existe n^T séquences différentes d'états, c'est-à-dire n^T termes). En pratique c'est bien sûr impossible, même pour des valeurs de T assez petites, ce qui pose un réel problème.

Problème 2 (Recherche de la séquence d'états optimale). *Étant donné le modèle λ, comment choisir la séquence d'états $q = q_1, q_2, \ldots, q_T$ maximisant $p(O, q \mid \lambda)$, la vraisemblance conjointe de la séquence d'observations $O = O_1, O_2, \ldots, O_T$ et de la séquence d'états ? Donc on cherche q^*, tel que*

$$q^* = \underset{q}{\operatorname{argmax}}\, p(O, q \mid \lambda) \tag{8.19}$$

Selon (8.16) et (8.17) $p(O, q \mid \lambda)$ s'écrit comme :

$$\begin{aligned} p(O, q \mid \lambda) &= p(O \mid q, \lambda) P(q \mid \lambda) \\ &= \pi_{q_1} f_{q_1}(O_1) \prod_{t=2}^{T} a_{q_t q_{t-1}} f_{q_t}(O_t) \end{aligned} \tag{8.20}$$

Si on fait la recherche de q^* en utilisant directement l'expression (8.20), on voit bien qu'il faudra calculer cette expression pour chaque séquence d'états possible, c'est-à-dire n^T fois. On se retrouve alors avec la même complexité de calcul que pour le problème 1.

Problème 3 (Estimation des paramètres du modèle). *Comment ajuster les paramètres λ du modèle HMM d'une façon telle que la vraisemblance $p(O \mid \lambda)$ soit maximale ? On cherche alors λ^*, satisfaisant*

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, p(O \mid \lambda) \tag{8.21}$$

3. **Génération des paramètres visuels avec la synthèse par HMM.** L'objectif de la phase de synthèse est de générer les paramètres visuels à partir de la suite phonétique étiquittée temporellement et à partir des HMMs appris pour chaque segment phonétique. La première étape consite en génération des durées des états de chaque segment phonétique à partir de la durée totale du segment phonétique et à partir des modèles de durées d'états appris pendant la phase d'analyse. La deuxième étape consiste en génération des séquences des paramètres visuels à partir de la durée de chaque état et à partir des HMMs.

Principe de la synthèse visuelle par HMM

Le système de synthèse visuelle par HMM comprend deux étapes principales : l'étape d'apprentissage de paramètres des modèles HMM et l'étape de synthèse de paramètres visuels à partir d'une séquence de HMMs. Le principe d'un système de synthèse visuelle est schématisé sur la figure 8.5.

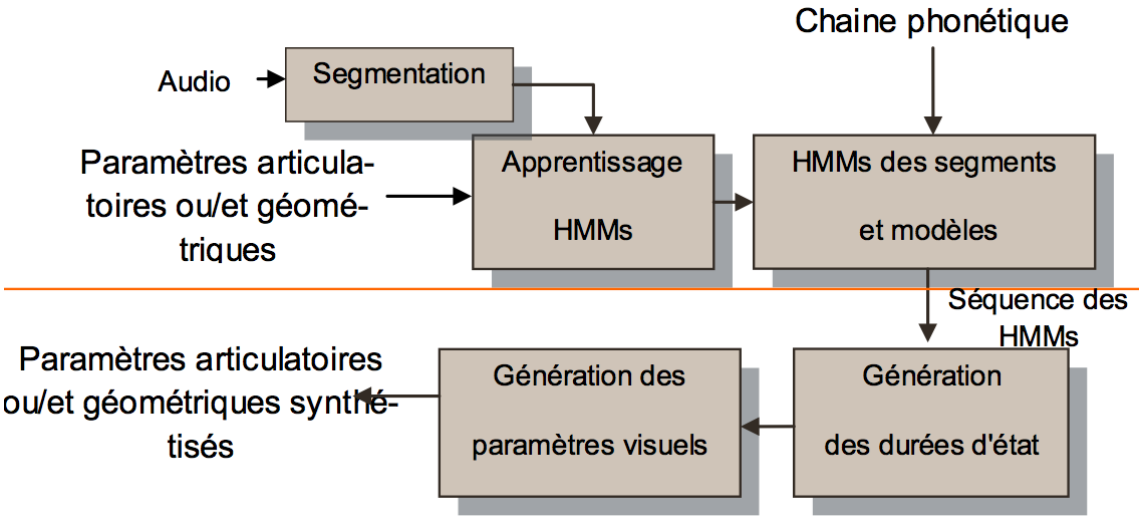


FIG. 8.5 – Le principe de la synthèse par HMMs de la parole visuelle.

Apprentissage des HMM En entrée de la phase d'apprentissage il y a une suite de vecteurs acoustiques et visuels, chaque paire de vecteurs correspond à une trame. Les vecteurs sont classés en groupes. Chaque groupe de vecteurs correspond à une unité phonétique, figure 8.6. Ensuite, un HMM est construit pour chaque unité. Un ensemble de séquences de vecteurs O^k(acoustiques ou visuels) $O = \{O^1, O^2, ..., O^m\}$ est utilisé dans l'apprentissage d'un HMM λ, figure 8.7. Le but de l'apprentissage est de déterminer les paramètres d'un HMM d'architecture fixée : $\lambda = (A, B, \pi)$, qui maximisent la probabilité $P(O|\lambda)$ (**problème 3**) (Rabiner, 1989). *L'algorithme EM* (Expectation - Maximisation) est une solution très générale d'un tel problème. Cet algorithme est itératif. Chaque itération se décompose en deux étapes : expectation et maximisation, respectivement.

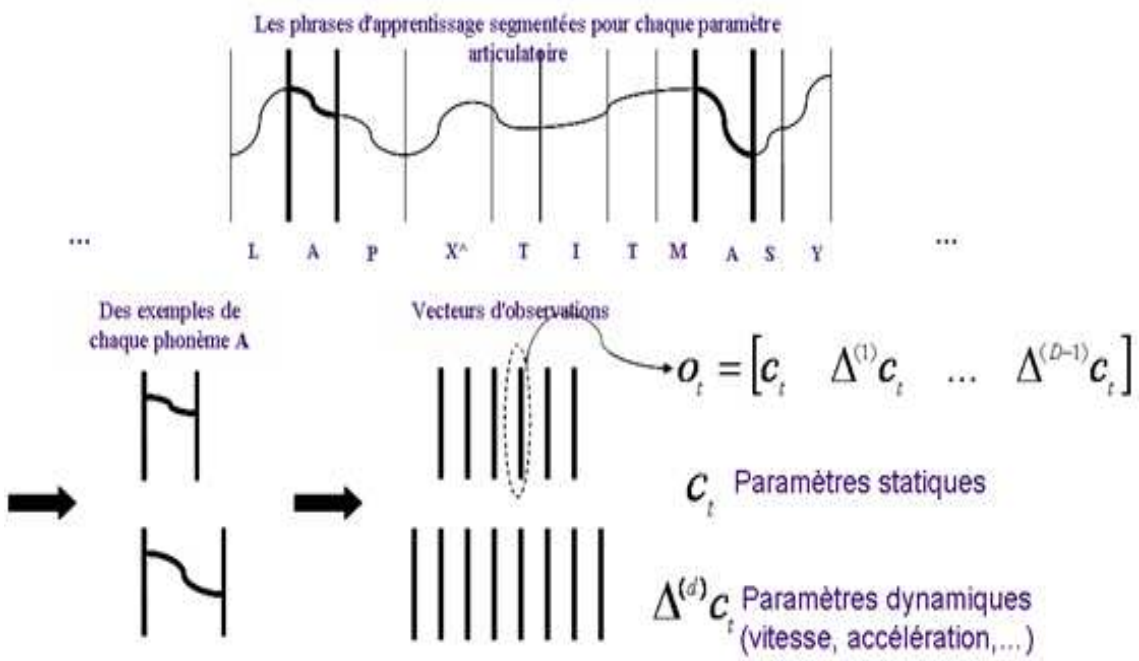


FIG. 8.6 – Exemple de construction d'un vecteur d'observation.

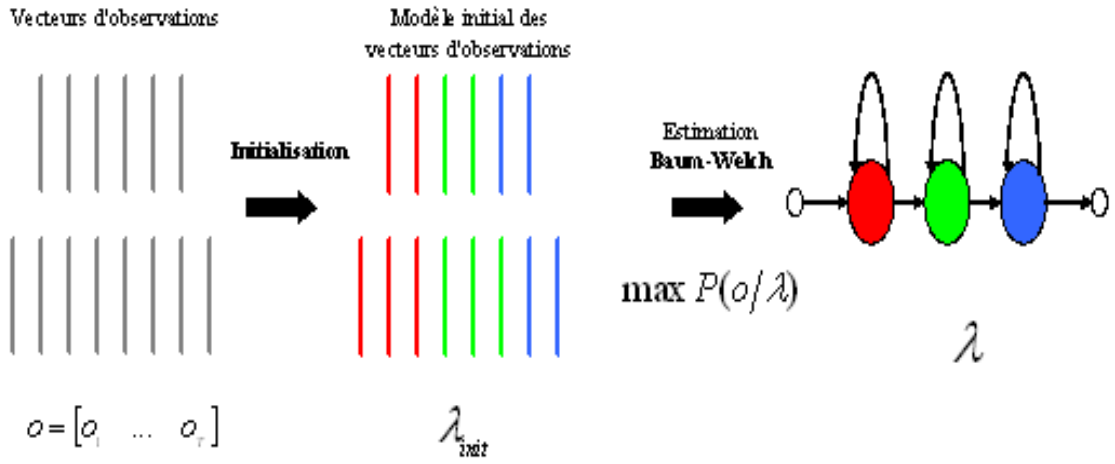


FIG. 8.7 – Principe de l'apprentissage d'un HMM.

Introduisons les notations suivantes :

- $\lambda = \{\pi, A, \mu_i, \Sigma_i \mid i = 1, \ldots, n\}$ les paramètres du modèle estimés à l'itération précédente,
- $\hat{\lambda} = \left\{\hat{\pi}, \hat{A}, \hat{\mu}_i, \hat{\Sigma}_i \mid i = 1, \ldots, n\right\}$ les paramètres du modèle estimés à l'itération courante,
- $\gamma_t(i) = P(s_t = q_i \mid O, \lambda)$ la probabilité d'être dans l'état q_i à l'instant t, étant donnés la séquence d'observations O et le modèle λ,
- $\xi_t(i, j) = P(s_t = q_i, s_{t+1} = q_j \mid O, \lambda)$ la probabilité de passer de l'état q_i à l'instant t à l'état q_j à l'instant $t + 1$ sachant O et λ.

Solution du problème 1 : Procedure "avant-arrière"
Pour résoudre le **problème 1**, on introduit d'abord deux quantités :

$$\alpha_t(i) = p\left(O_1, O_2, \ldots, O_t, s_t = q_i \mid \lambda\right) \tag{8.22}$$

la vraisemblance de la séquence d'observations partielle jusqu'à l'instant t et de l'état q_i à l'instant t, et

$$\beta_t(i) = p\left(O_{t+1}, O_{t+2}, \ldots, O_T \mid s_t = q_i, \lambda\right) \tag{8.23}$$

la vraisemblance de la séquence d'observations partielle allant de $t + 1$ jusqu'à T, sachant, que l'on était à l'état q_i à l'instant t.

La vraisemblance $p\left(O \mid \lambda\right)$ peut se calculer à partir de ces deux quantités, donc on a deux façons de calculer cette vraisemblance :

Procédure "avant"

1. Initialisation, pour $1 \leq i \leq n$:

$$\alpha_1(i) = \pi_i f_i(O_1) \tag{8.24}$$

2. Recurrence "avant", pour $t = 1, 2, \ldots, T - 1, 1 \leq j \leq n$:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{n} \alpha_t(i) a_{ij}\right] f_j(O_{t+1}) \tag{8.25}$$

3. Calcul de vraisemblance :

$$p\left(O \mid \lambda\right) = \sum_{i=1}^{n} \alpha_T(i) \tag{8.26}$$

Procédure "arrière"

1. Initialisation, pour $1 \leq i \leq n$:

$$\beta_T(i) = 1 \tag{8.27}$$

2. Recurrence "arrière", pour $t = T - 1, T - 2, \ldots, 1, 1 \leq i \leq n$:

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij} f_j(O_{t+1}) \beta_{t+1}(j) \tag{8.28}$$

3. Calcul de vraisemblance :

$$p\left(O \mid \lambda\right) = \sum_{i=1}^{n} \pi_i f_i(O_1) \beta_1(i) \tag{8.29}$$

Dans l'équation (8.25) de la procédure "avant" on calcule tout d'abord $p(O_1, \ldots, O_T, s_{t+1} = q_j \mid \lambda)$ la vraisemblance d'émettre la séquence d'observations partielle $O_1, O_2, \ldots, O_t$ et de passer à l'état q_j à l'instant $t+1$, indépendamment de l'état précédent à l'instant t. On somme donc sur tous les états précédents possibles i. Ensuite, on ajoute la vraisemblance d'émission d'observation O_{t+1}, qui ne dépend pas de l'état précédent. Donc, on a $f_j(O_{t+1})$ dehors des parenthèses. L'équation (8.26) est juste la sommation sur tous les états finaux possibles, pour obtenir la vraisemblance désirée. La procédure "arrière" s'explique de la même façon.

On voit bien, que chacun de ces deux algorithmes demande $O(n^2T)$ multiplications au lieu de $O(2Tn^T)$ multiplications pour le calcul direct.

En utilisant la définition de la probabilité conditionnelle et les expressions (8.30) et (8.31) on a :

$$\alpha_t(i) = p\left(O_1, O_2, \ldots, O_t, s_t = q_i \mid \lambda\right) \tag{8.30}$$

$$\beta_t(i) = p\left(O_{t+1}, O_{t+2}, \ldots, O_T \mid s_t = q_i, \lambda\right) \tag{8.31}$$

$$\gamma_t(i) = \frac{p(s_t = q_i, O \mid \lambda)}{p(O \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{p(O \mid \lambda)} \tag{8.32}$$

De la même façon on peut montrer, que $\xi_t(i,j)$ s'exprime comme :

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}f_j(O_{t+1})\beta_{t+1}(j)}{p(O \mid \lambda)} \tag{8.33}$$

Ensuite, si on somme $\gamma_t(i)$ et $\xi_t(i,j)$ de $t = 1$ jusqu'à $T-1$, les quantités obtenues peuvent être considérées comme :

$$\begin{aligned} \sum_{t=1}^{T-1} \gamma_t(i) &= \text{Estimation du nombre de transitions effectuées à partir de } i \\ \sum_{t=1}^{T-1} \xi_t(i,j) &= \text{Estimation du nombre de transitions de } i \text{ vers } j \end{aligned}$$

Maintenant il est assez naturel de calculer les probabilités de transition $\hat{a}_{ij}$ du nouveau modèle $\hat{\lambda}$ comme le rapport entre le nombre de transitions de i vers j et le nombre de transitions effectuées à partir de i. Les vecteurs moyens $\hat{\mu}_i$ et les matrices de covariance $\hat{\Sigma}_i$ sont calculées de la manière habituelle, mais en pondérant selon les probabilités $\gamma_t(i)$. On obtient alors l'algorithme suivant :

Algorithme de Baum-Welch.

1. Initialisation : choisir une approximation initiale $\lambda = \lambda^0$,
2. Estimation des probabilités (l'étape d'expectation de l'algorithme EM) : calculer $\gamma_t(i)$ et $\xi_t(i,j)$ en utilisant les expressions (8.32) et (8.33) avec la procédure "avant-arrière" (en annexe).
3. Réestimation des paramètres (l'étape de maximisation de l'algorithme EM) :

$$\hat{\pi}_i = \gamma_1(i) \tag{8.34}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{8.35}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} \gamma_t(i) O_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{8.36}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(O_t - \hat{\mu}_i)(O_t - \hat{\mu}_i)^T}{\sum_{t=1}^{T} \gamma_t(i)} \tag{8.37}$$

4. Poser $\lambda = \hat{\lambda}$ et passer à l'étape 2, ou bien arrêter selon un critère d'arrêt (par exemple un nombre d'itérations fixé).

L'algorithme EM, qui est au fond de cet algorithme, assure la convergence vers un minimum local selon l'approximation initiale λ^0. En plus, la vraisemblance maximisée ne peut qu'augmenter à chaque itération, c'est-à-dire

$$p(O \mid \hat{\lambda}) \geq p(O \mid \lambda) \tag{8.38}$$

Synthèse de séquences d'observation

Construction de la séquence d'états La séquence d'états peut être obtenue soit à partir des durées des unités (Yamamoto *et al.* , 1998), (Tamura *et al.* , 1999) (les durées des états sont parfois aussi modélisées par des distributions monogaussiennes correspondantes à chaque état d'un HMM), soit à partir des paramètres acoustiques (Hiroya & Honda, 2004). Dans tous les cas il se pose le problème d'estimation de la séquence optimale d'états q pour une observation O et un modèle λ données. La séquence est obtenue en maximisant la probabilité de sortie **(Problème 2)** :

$$q^* = \underset{q}{\operatorname{argmax}}\, p(O, q \mid \lambda) \tag{8.39}$$

Tout d'abord, en prenant le logarithme de (8.20), on définit

$$U(O, q \mid \lambda) \triangleq \log p(O, q \mid \lambda) = \log\left(\pi_{q_1} f_{q_1}(O_1)\right) + \sum_{t=2}^{T} \log\left(a_{q_t q_{t-1}} f_{q_t}(O_t)\right) \tag{8.40}$$

Puisque le logarithme est une fonction croissante, le problème (8.39) est équivalent au problème suivant :

$$q^* = \underset{q}{\operatorname{argmax}}\, U(O, q \mid \lambda) \tag{8.41}$$

Ce passage au logarithme sert seulement à simplifier les calculs. Effectivement, dans l'expression (8.20) on a un produit d'un grand nombre de vraisemblances et de probabilités. Pour une valeur de T assez importante ce produit devient trop petit ou bien trop grand, provoquant des problèmes de dépassement des possibilités de représentation numérique (*underflow* ou *overflow*). Dans le domaine logarithmique ce n'est plus le cas.

Imaginons maintenant que l'on construit un graphe orienté à nT noeuds. Chaque noeud $(q_i t)$ représente le fait d'être dans l'état q_i à l'instant t en émettant l'observation O_t, et on peut aller du noeud $(q_i[t-1])$ au noeud $(q_j t)$ avec un coût $\log(a_{ij}) + \log(f_{q_j}(O_t))$. Le coût d'un chemin dans ce graphe est la somme des coûts de tous les déplacements successifs. L'exemple d'un tel graphe pour un HMM à 3 états et $T = 4$ est représenté dans la figure 8.8. On voit bien que la solution du problème (8.41) consiste à trouver dans le graphe le chemin avec le coût maximal. Un tel problème se résout à l'aide de la méthode de *Programmation Dynamique*. Dans le cadre des HMMs cette méthode s'appelle *l'algorithme de Viterbi*.

Notons par $\delta_t(i)$ le coût maximal accumulé à l'état i à l'instant t, c'est-à-dire le coût du meilleur chemin qui s'arrête au noeud $(q_i t)$, et par $\psi_t(i)$ l'état à l'instant $t-1$ qui donne le coût maximal pour la transition à l'état i à l'instant t.

Algorithme de Viterbi

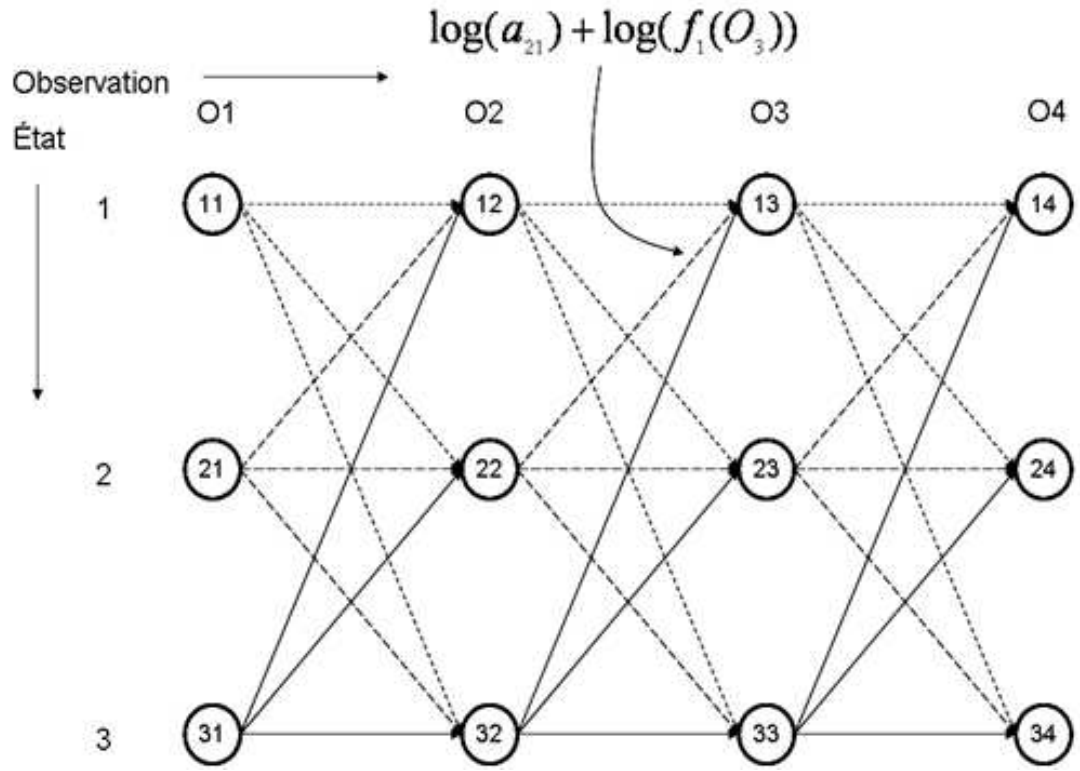


FIG. 8.8 – Graphe pour la recherche de Viterbi (n = 3, T = 4).

1. Initialisation, pour $1 \leq i \leq n$:

$$\delta_1(i) = \log(\pi_i) + \log(f_i(O_1)) \tag{8.42}$$
$$\psi_1(i) = 0 \tag{8.43}$$

2. Calcul récursif, pour $t = 2, 3, \ldots, T$, pour $1 \leq j \leq n$:

$$\delta_t(j) = \max_{1 \leq i \leq n} [\delta_{t-1}(i) + \log(a_{ij})] + \log(f_j(O_t)) \tag{8.44}$$
$$\psi_t(j) = \operatorname*{argmax}_{1 \leq i \leq n} [\delta_{t-1}(i) + \log(a_{ij})] \tag{8.45}$$

3. Terminaison :

$$U^* = \max_{1 \leq i \leq n} [\delta_T(i)] \tag{8.46}$$
$${q_{i_T}}^* = \operatorname*{argmax}_{1 \leq i \leq n} [\delta_T(i)] \tag{8.47}$$

4. Tracement en arrière de la séquence d'états optimale, pour $t = T-1, T-2, \ldots, 1$:

$${q_{i_t}}^* = \psi_{t+1}({q_{i_{t+1}}}^*) \tag{8.48}$$

Donc l'algorithme de Viterbi donne la séquence d'états optimale $q^* = q_1^*, q_2^*, \ldots, q_T^*$. La vraisemblance maximisant (8.20) peut être aussi calculée comme $\exp(U^*)$.

On peut facilement estimer que, comme pour la procédure "avant - arrière" (en annexe), la complexité de calcul est d'ordre $O(L^2T)$ au lieu de $O(2TL^T)$ pour le calcul direct.

Génération de paramètres Cette phase de synthèse a pour but de générer les paramètres visuels en ayant la suite des HMMs et les séquences d'états optimales pour chaque HMM. La séquence des paramètres visuels $O = \left[O_1^\top, ..., O_T^\top\right]$ est obtenue en maximisant la vraisemblance $P(O|\lambda)$ par rapport à O.

$$P(O|\lambda) = \sum_q P(O|q,\lambda)P(q,\lambda) \tag{8.49}$$

Où le vecteur O_t est constitué d'un vecteur statique et des vecteurs dynamiques :

$$O_t = \left[c_t^\top, \Delta^{(1)}c_t^\top, ..., \Delta^{(D-1)}c_t^\top\right]^\top \tag{8.50}$$

$$c_t = \left[c_t(1), c_t(2), ..., c_t(M)\right]^\top \tag{8.51}$$

$$\Delta^{(d)}c_t = \sum_{\tau=-L_-^{(d)}}^{L_+^{(d)}} w^{(d)}(\tau)c_{t+\tau} \tag{8.52}$$

Les $w^{(d)}(\tau)$ représentent les coefficients de la fenêtre de calcul d'un paramètre dynamique de l'ordre d.

A chaque état d'un HMM une distribution Gaussienne $P(O|q,\lambda)$ est associée :

$$P(O|q,\lambda) = \prod_{t=1}^{T} N(O_t|\mu_{q_t}, \Sigma_{q_t}) = N(O|\mu_q, \Sigma_q) \tag{8.53}$$

où μ_{q_t} et Σ_{q_t} sont des vecteurs de la dimension $DM \times 1$ et $DM \times DM$ respectivement associés à l'état q_t :

$$\mu_q = \left[\mu_{q_1}^\top, \mu_{q_2}^\top, ..., \mu_{q_T}^\top\right]^\top$$

$$\mu_{q_t} = \left[\Delta^{(0)}\mu_{q_t}^\top, \Delta^{(1)}\mu_{q_t}^\top, ..., \Delta^{(D-1)}\mu_{q_t}^\top\right]^\top$$

$$\Delta^{(d)}\mu_{q_t} = \left[\Delta^{(d)}\mu_{q_t}^\top(1), \Delta^{(d)}\mu_{q_t}^\top(2), ..., \Delta^{(d)}\mu_{q_t}^\top(M)\right]^\top$$

$$\Sigma_q = diag\left[\Sigma_{q_1}^\top, \Sigma_{q_2}^\top, ..., \Sigma_{q_T}^\top\right]$$

$$\Sigma_{q_t} = diag\left[\Delta^{(0)}\Sigma_{q_t}^\top, \Delta^{(1)}\Sigma_{q_t}^\top, ..., \Delta^{(D-1)}\Sigma_{q_t}^\top\right]$$

$$\Delta^{(d)}\Sigma_{q_t} = diag\left[\Delta^{(d)}\Sigma_{q_t}^\top(1), \Delta^{(d)}\Sigma_{q_t}^\top(2), ..., \Delta^{(d)}\Sigma_{q_t}^\top(M)\right]$$

La condition (8.52) peut être exprimée comme suit :

$$O = Wc \tag{8.54}$$

où

$$c = \left[c_1^\top, c_2^\top, ..., c_T^\top\right]^\top \tag{8.55}$$

$$W = \left[W_1, W_2, ..., W_T\right]^\top \otimes I_{M\times M} \tag{8.56}$$

$$W_t = \left[w_t^{(0)}, w_t^{(1)}, ..., w_t^{(D-1)}\right] \tag{8.57}$$

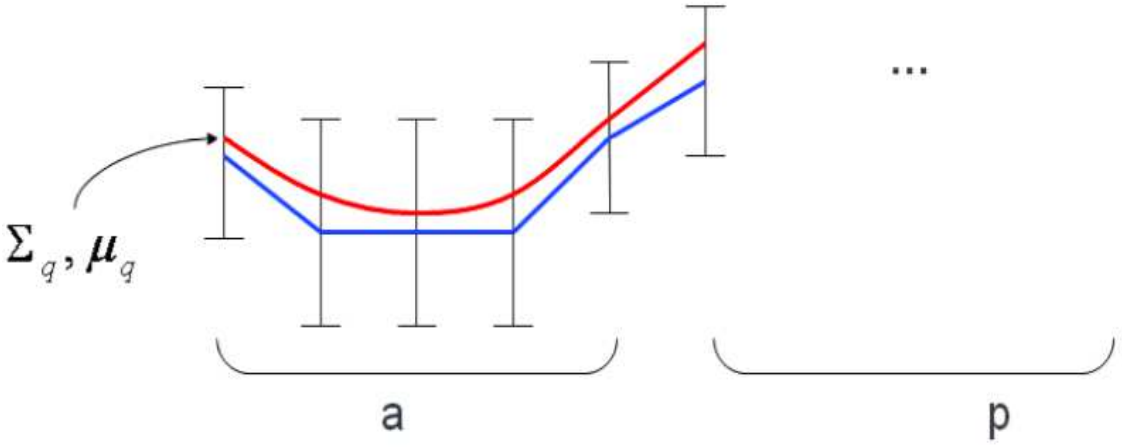


FIG. 8.9 – Principe de l'algorithme de "lissage".

$$w_t^{(d)} = \left[\underbrace{0, ..., 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), w^{(d)}(L_+^{(d)}), \underbrace{0, ..., 0}_{T-(t+L_+^{(d)})} \right]^\top, \tag{8.58}$$

$$L_-^{(0)} = L_+^{(0)} = 0$$

$$w^{(0)}(0) = 1$$

Il est evident que la vraisemblance (8.53) est maximisée si $O = \mu_q$. Ainsi la séquence des paramètres visuels devient une séquence des moyennes. Cela est du au fait que les paramètres statiques et les paramètres dynamiques sont considérés comme indépendants. Pour éviter ce problème la contrainte (8.54) qui existe entre ces paramètres est prise en compte, figure 8.9. C'est pour cette raison que la maximisation de la (8.53) par rapport à O revient à la même chose si l'on fait par rapport à c :

$$\frac{\delta \log P(Wc|q, \lambda)}{\delta c} = 0 \tag{8.59}$$

$$\log P(O|Q, \lambda) = -\frac{1}{2} O^\top \Sigma^{-1} O + O^\top \Sigma^{-1} M + K \tag{8.60}$$

En calculant la dérivée du $logP(O|Q, \lambda)$ on obtient :

$$W^\top \Sigma^{-1} W c = W^\top \Sigma^{-1} M \tag{8.61}$$

d'où

$$c = (W^\top \Sigma^{-1} W)^{-1} W^\top \Sigma^{-1} M \tag{8.62}$$

Ainsi on obtient les séquences de paramètres visuels $c = \left[c_1^\top, c_2^\top, ..., c_T^\top\right]^\top$

8.3 Annexe C

8.3.1 Corpus I. Visage

Liste des phrases

1. Ma chemise est roussie.
2. Voilà des bougies !
3. Donne un petit coup !
4. Il a du goût.
5. Elle m'étripa.
6. Une réponse ambiguë.
7. Louis pense à ça.
8. Un four touffu.
9. Un tour de magie.
10. Voilà du filet cru.
11. La force du coup.
12. Prête-lui seize écus.
13. Il fait des achats.
14. Chevalier du gué.
15. Le jeune hibou.
16. Il fume son tabac.
17. Un piège à poux.
18. L'examen du cas.
19. Je suis à bout.
20. Elle a chu.
21. Je vais chez l'abbé.
22. Deux jolis boubous.
23. Une belle rascasse.
24. Il part pour Vichy.
25. Faire la nouba.
26. C'est Louis qui joue.
27. C'est ma tribu.
28. Gilles m'attaqua.
29. Une rocaille moussue.
30. Un pied fourchu.
31. La chaise du bout.
32. Trop d'abus.
33. J'en ai assez.
34. Jean est fâché.
35. Le pied du gars.
36. Vous avez réussi.
37. Ils n'ont pas pu.
38. Le vent mugit.
39. Une autre roupie.
40. Deux beaux bijoux.
41. Tu ris beaucoup.
42. Dés que le tambour bat les gens accourent.
43. Annie s'ennuie loin de mes parents.
44. Vous poussez des cris de colère.
45. Mon père m'a donné l'autorisation.
46. Un loup s'est jeté immédiatement sur la petite chèvre.
47. J'ai un scorpion sec dans mon talon aiguille.

48. Nos dalmatiens campaient au camping à la montagne.
49. Vend-on un cake intact à Hong-Kong.
50. Noam Chomsky balaie encore le club ce soir.
51. L'avoué a besoin d'un joint sous huitaine.
52. La sueur suinte du thon huileux.
53. Le beau ouistiti suit le riche huissier à Waterloo.
54. Tout Winipeg attend Wendy sur le parking ouest.
55. Bud et Buck font un bon whist à Maubeuge.
56. Youri fouette l'ail ionique de Kohoutek.
57. Beung j'ai heurté le puits dans la lueur.
58. Vuitton fait cuire dix wapitis goûteux.
59. David Bowie s'est rué sur le quai où j'ai organisé ce must.
60. Young fait un petit huit avec un joueur noueux.
61. Jean Nohain a chargé Watson de louer le huitième buisson.
62. Li-peng met du nuoc-mam dans son amuse-gueule.
63. J'ai Eugène au téléphone qui cueille joliment du gui.
64. Les keums du wharf rament évidemment dans le paysage.
65. Ivanhoé a fait un bug au huitième essai.
66. Tu huiles l'étui du buzzer de deux watts.
67. J'ai étudié le parking huit à Plancoët.
68. Eh oui les forums de l'accueil sont chouettes.
69. Des Ewoks habitent la maison en paille du centre spatial.
70. La famille ouistiti a éternué sous les dolmens.
71. J'ai identifié un mohican dans un western pyrénéen.
72. Le balai a fait un looping sur la toundra.
73. Ce tuyau a voyagé très haut chez les martiens.
74. Les caïds jouent au ping-pong avec l'équipe de Bosnie.
75. Je souhaite que sa peau usée ne reçoive jamais cette greffe ridicule.
76. Ce fou ordinaire cache le turban indien dans le bain optionnel.
77. Une agrafe géante a pu heurter son beau hors-bord.
78. De mauvaises gens privent Victor de sa coiffe bretonne.
79. La grive perchée sur l'if noir couve toujours ce canif chinois.
80. Le vase zen a perdu aussi un anneau en roche grise.
81. La houle lave les hublots d'une case déserte.
82. Il abrase chaque jour un pneu ancien avec ses griffes pointues.
83. Le photographe garantit un gag tordu au goût incertain.
84. Le bateau heurta les housses du hublot un peu humides.
85. La feuille fut sertie avec une dent usée de la biche docile.
86. Le géologue trouve finalement la houille en vrac dans le gave de Pau.
87. Le loup oublie son plan astucieux dans une poche chinoise.
88. Le prof mielleux triche souvent à ce jeu idiot.
89. Ce jazz rythmé est un cadeau inespéré.
90. Le veau heureux attend Eudes dans le hameau indien.
91. L'âne bègue voit que la vache de Joseph se vexe.
92. Tu houspilles ton amant onctueux qui louche réellement.
93. La gaffe fabrique une ruche carrée si tu y coopères.
94. Cette phrase particulière étouffée toute une strophe vertueuse.
95. Ce chant hideux rase son héros venu en hâte.
96. Le camp hostile coordonne le putsch dans la cohue.
97. Cette pêche fameuse a vu onduler l'endive blanche.
98. Il se lève chaque jour et attend Hercule qui oublie.

99. Il n'arrive nullement qu'une vague surgisse du hors-d'oeuvre.
100. Un zébu heureux ne touche jamais au houblon.
101. Son gant entoure la valise trouvée sur la digue droite.
102. Jean heurta une cuve large pleine de gouache verte.
103. Le vent établi sèche bien le houx où crèche mon hibou.
104. Tchang ôte sa toge cintrée d'une main innocente.
105. Dom Juan drague finalement une jeune fille mal faite.
106. à eux la soif zoologique du bourgeon ouvert.
107. Au yen la tâche pénible de ce prêt embarrassant.
108. En haut la guêpe pense aux heures.
109. L'anglaise lui offre ce qu'elle a au doigt ou à l'oreille.
110. Elle joue uniquement avec la neige chantante.
111. On tua onze ou douze torchons archaïques.
112. Oudini ignore le train où doit se produire le spectacle.
113. Il est parti illico en avion ou en gondole.
114. Il gobe douze fèves et blèche tout mon jardin.
115. La caisse seule a en sur le ring en bois.
116. Votre crêpe chaude vise bien le haut du feu.
117. Tailles-en un bien haut et travaille chaque nom.
118. Fernand oublie de moudre son café.
119. L'abeille n'engrange pas de miel sur un chemin.
120. Eole aide sa robe fendue à se soulever.
121. Bashung oublie aussi qu'il lègue quelque chose.
122. Je passe chercher ce que j'ai lu avec vous.
123. Un zoom ferait ce que neuf demis pensent faire.
124. Le fou immerge son aiguille et brode finement.
125. Chaque bout du rail carré est une tige ténue.
126. Un argument élogieux échappe bien au rosbif.
127. Le malade guéri attrape mon solide microbe.
128. Zola demande notamment du bon lait à un mage zurichois.
129. Cette dame veut galber un tube vertical.
130. Nous traquions bien Euler pendant son footing urbain.
131. J'ai vu un holding important sur un terre plein escarpé.
132. Pain et pudding gallois aident le petit hussard oublieux.
133. Une bouteille de Riesling heurta le balcon humide.
134. Ce jeu invite un type joueur et une dame riche.
135. Miss Zazie effectue un travelling heureux sur un machin imposant.
136. Une vache normande dirige rarement un jumping zélé.
137. Le viking honteux a mal chuté sur cette petite nappe.
138. Le pape vient en Yamaha dans une bourgade curieuse.
139. Le lapin utilise son yoyo et a besoin d'aide.
140. Le dumping l'incite à jeter les prunes tombées.
141. Les yétis mal rasés ont la bouille pâteuse.
142. Ils oublièrent Chuck dans un tube carré.
143. Le King charmeur porte une chemise rouge foncé.
144. Yasmine aime ton standing japonais.
145. Gaspard blague mollement sur le leasing omniprésent.
146. Eux aussi aiment la tripe glorieuse un peu euphorique.
147. Oeuvrez pour l'ove du globe bleu des yeux.
148. Mes juges vont manger ce cache yaourt à la truelle.
149. Cet oeil globuleux porte une lentille luisante.

150. La sage baleine zoophile n'a aucune patte valide.
151. Un pâle zébu agnostique mange normalement une solide pizza.
152. Le prieur brade tout centime gagné.
153. La caille revient sans eux dans l'herbage gourmand.
154. Une guenon heureuse a vu un balcon ombragé.
155. Chaque garçon aime que le soleil brille.
156. Il y a un truc qui ondule dans la cage murale.
157. Tapes-en au noir sur une petite zone.
158. La fausse reine en tailleur agace Guy.
159. Nous tuons chaque chiot qui a été heureux.
160. Flambes-y une crêpe bretonne de gamme moyenne.
161. Chaque zéro est un looping tordu.
162. La meilleure omelette du Larzac peut rivaliser avec le yachting normand.
163. Un nain heurta une bogue charnue un onze janvier.
164. Une tombe Ming ne passe jamais pour un karting belge.
165. Un homme jeune ne tombe pas pendant cette java.
166. Des rides charmantes aèrent cette robe choisie dans les pages jaunes.
167. La foule a afflué quand mon neveu heurta lever.
168. Le thon heurta un bleuet.
169. Ceux des gueux bigleux veulent libérer Bob Taylor.
170. Où était Oxymel.
171. Le jeu ôtait illico au parfum oublié un fin bouquet d'embrun.
172. Le cousin chinois du tribun évalue au jugé autrement le tissu invendu.
173. Moreau étale immanquablement un déchet commun à la queue de l'UE.
174. Aladin élève chacun en symbiose avec le vieil ouzbek.
175. Chacun ignore son c.e. un peu un moment.
176. Avec un aplomb imparable nous avons chacun un c.e. énergique.
177. Cette énergie insensée grève un quinzième de Ugines.
178. Sa tape un peu impolie heurta Bernache un peu trop violemment.
179. Sylvain ne suit pas le parfum imprévu.
180. Ce cabot ombrageux fête son accession au pouvoir.
181. Un noir de jais évoque le front eurasien.
182. Ce suspect heurta le bibelot ancien un peu lourdement.
183. Le bedeau euphorique secoue l'anneau un jour par an.
184. Aux lilas violet européens Corot Eugène préfère vingt-et-un Oeillets.
185. Jojo heurta le défunt et le tua.
186. Le l.p.e. insiste et les p.m.e. ont signé.
187. Regardes il zigzague un peu vite.
188. Un huit dans l'eau a huilé l'un des tiroirs.
189. Railles un bourrin oisif.
190. Prends- le Euclide.
191. Tailles huit brins ouatés.
192. Je m'huile le corps dans ce lieu iodé.
193. Jourdain rajoute un pneu huileux.
194. Il se ouate le tein rebelle.
195. J'ai reçu ton dessin hier.
196. Quantum suédois ou rituel wolof.
197. La secoueuse fait des percings linguaux.
198. J'ai oublié ton message.
199. Tiens-toi assis !
200. Vous êtes exclue.

201. Pas plus de quatre rubis.
202. C'est lui qui me poussa.
203. Il se garantira du froid avec ce bon capuchon.
204. Les deux camions se sont heurtés de face.
205. La vaisselle propre est mise sur l'évier.
206. Les gangs indigents des bings et des bangs périlleux sur une aile.
207. Huit jésuites très huileux se font un brushing yougoslave.
208. J'avais honte car la fille huait les Who.
209. L'africa song s'emballe en juillet sur un walkman muet.
210. C'est Hervé qui fuit dans un yacht en leasing.
211. Walid a hué les Pink Floyd à Rouen.
212. Nous jouons aux billes dans les ruines muettes.
213. J'ai huilé un rayon du train huit à l'équinoxe.
214. J'ai eu les symptômes de la presbytie en huit jours.
215. Pose calmement ta dague pointue sur cette étoffe carrée.
216. Va dans une cave quelconque et caches-y ce drapeau honteux.
217. Rêves-y car l'extase vient de cette bague gracieuse.
218. Il élague curieusement la houppe qui est récalcitrante.
219. Quand je soulève ma hache le banc ondule.
220. La horde de hors-la-loi alpague bientôt l'épave galloise.
221. Un très bon vin en bouteille exige un planning idoine.
222. Objectez à Neuilly contre le gaz nocif des hommes.
223. La pin-up feind de tomber chez toi mais ne blague jamais.
224. Cherche où est le thon obtus que je trouve sot.
225. Ce buveur balte augmente sa masse veineuse à heure régulière.
226. Le moteur du Boeing ronronne dans la brouette.
227. Le rotring exige une page carrée dans une feuille verte.
228. Léon range le parking vendéen où on aime zoner.
229. Nous draguions le torrent pour trouver des crabes noirs.
230. Ce soldat un peu honteux fait un job glorieux.
231. Il a été heurté par un prêcheur.
232. Intonnes un u ou un euh à intervalles réguliers.
233. Le c.e. isole les engins communs aux deux charlots.
234. Une québécoise pleurnicheuse brandit Euclide lors des réunions.
235. Un coup heureux et impétueux amodié un vulgaire pain onctueux en gnome.
236. Sur le zing chacun interprète l'atlas humblement posé sur l'ancien jabot.
237. à jeun Antoine le heurte et cet accident le hantera.
238. Antoine avait ouint son numéro huit.

Corpus I. Visage

Tableau des diphones

	i	e	eˆ	a	oˆ	o	u	y	x	xˆ	a˜	e˜	o˜	x˜	p	t	k	b	d	g	f	v	s	z	sˆ	zˆ	r	l	m	n	h	j	w	sild	silf	q	sil	nˆ	ntˆ	#	Tot.
i	0	1	0	1	0	1	1	1	0	0	0	0	1	2	10	25	26	10	22	4	9	6	16	15	8	9	6	39	7	17	1	19	1	0	13	0	26	30	3	0	330
e	1	0	2	3	2	2	1	4	2	3	1	3	2	7	12	21	17	10	16	11	8	6	17	9	4	3	14	21	3	7	2	3	5	0	28	1	47	0	0	0	298
eˆ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	23	17	0	3	3	2	7	6	4	4	2	43	17	8	12	0	5	0	0	0	0	1	0	0	0	160
a	2	3	1	4	1	2	3	1	1	1	1	1	1	4	22	25	29	20	16	17	16	32	26	16	15	16	51	49	29	15	1	12	1	0	17	0	12	0	2	0	465
oˆ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	3	8	1	1	5	1	2	1	2	2	29	9	6	8	0	1	0	0	0	0	0	0	0	0	82
o	2	1	1	1	1	4	2	1	0	0	1	2	1	1	5	12	7	5	9	4	2	3	10	5	2	3	7	14	6	8	2	2	2	0	7	0	21	0	0	0	154
u	0	3	1	1	0	1	0	0	2	1	2	0	2	1	5	9	2	11	7	1	2	7	8	7	2	5	24	6	2	1	1	2	1	0	14	0	7	2	0	0	140
y	1	6	2	3	0	1	0	1	5	2	1	0	2	2	1	6	4	10	2	5	3	2	6	4	2	4	29	13	5	36	1	1	1	0	17	0	14	1	0	0	193
x	0	2	0	0	0	1	0	1	0	0	0	0	1	1	2	2	3	3	8	2	2	2	4	12	2	4	6	2	4	1	0	2	1	0	11	0	27	0	0	0	106
xˆ	0	0	1	1	0	1	1	1	0	1	1	0	0	0	14	21	18	14	4	6	8	17	20	2	1	11	44	16	20	6	1	7	1	0	0	0	20	1	0	0	259
a˜	1	1	0	2	0	2	0	0	0	0	0	0	1	0	4	21	3	8	13	6	1	2	13	8	1	7	1	17	3	4	1	1	1	0	13	0	27	1	0	0	163
e˜	0	1	0	0	0	1	0	0	0	0	0	0	0	0	6	12	1	2	8	1	1	2	6	1	0	1	2	1	2	1	1	1	1	0	8	0	23	1	0	0	84
o˜	1	1	0	0	0	0	0	1	0	0	0	0	0	1	3	10	6	8	5	1	2	1	7	3	3	2	2	2	3	7	1	1	1	0	6	0	18	2	0	0	98
x˜	0	0	0	0	1	0	0	1	1	0	0	0	1	0	16	8	5	8	4	1	2	1	2	2	1	3	1	3	4	6	1	1	2	0	1	0	13	0	0	0	89
p	13	11	10	31	5	7	11	3	9	7	5	3	2	1	0	0	0	1	1	0	0	1	1	1	1	0	15	8	1	2	1	4	2	0	1	0	6	0	0	0	164
t	31	19	10	27	10	6	10	24	5	2	9	2	13	4	2	0	2	1	2	0	1	0	4	1	4	1	33	2	2	1	3	4	4	0	2	27	13	0	0	0	281
k	19	3	5	33	3	11	9	10	2	12	6	1	5	5	0	7	0	1	0	1	0	1	10	1	0	1	7	9	2	1	1	1	4	0	2	7	12	0	0	1	193
b	11	6	8	15	3	9	16	8	2	7	3	2	4	1	1	2	1	0	1	0	1	1	0	1	0	1	14	22	1	1	1	7	1	0	1	0	2	0	0	0	154
d	12	21	2	9	4	4	2	36	7	34	28	3	1	1	1	0	0	1	0	1	1	0	0	1	1	3	5	0	1	1	1	5	4	0	3	4	8	0	0	0	205
g	2	2	2	18	2	1	3	4	1	6	1	0	1	0	0	0	0	0	0	0	0	0	0	4	0	0	9	6	1	2	2	1	2	0	1	2	12	0	0	1	86
f	11	9	7	5	3	4	6	3	1	3	1	2	3	3	1	1	1	1	2	0	1	0	1	0	1	1	6	6	1	1	1	1	1	0	3	0	4	0	0	0	95
v	8	9	17	13	2	1	4	5	2	4	6	3	2	0	0	1	1	0	1	0	1	0	1	1	2	1	4	3	1	1	1	9	6	0	1	1	5	0	0	0	117
s	19	16	18	12	2	5	5	25	3	30	10	4	18	0	6	13	5	2	2	1	1	0	1	1	1	1	1	2	1	1	5	12	4	0	4	0	4	0	0	0	235
z	11	13	4	11	2	6	3	5	2	2	6	2	1	3	0	0	0	2	2	1	1	0	3	1	0	2	0	1	0	0	1	3	2	0	8	2	13	0	0	0	113
sˆ	5	6	4	18	1	2	2	4	1	6	3	1	2	0	0	1	1	1	2	0	1	1	1	1	1	1	1	1	1	1	1	1	2	0	2	0	5	0	0	0	81
zˆ	9	21	3	11	1	6	13	2	6	11	7	1	1	2	1	1	0	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	3	0	5	2	5	0	0	0	123
r	35	26	11	32	11	13	8	9	8	13	10	4	2	7	7	30	5	8	15	3	3	4	10	1	5	7	0	17	8	6	1	4	2	0	20	1	26	0	0	0	372
l	31	40	13	74	9	11	7	7	15	73	9	3	3	1	4	1	5	1	3	2	3	1	4	3	1	2	1	5	11	3	4	6	7	0	12	2	9	0	0	0	386
m	12	17	3	20	1	8	2	5	2	4	32	3	7	2	2	2	1	1	2	1	1	1	4	1	1	1	1	2	1	1	1	3	1	0	2	0	2	0	0	0	150
n	14	7	6	15	5	12	7	6	3	16	2	1	3	2	4	3	4	5	7	2	2	1	2	0	1	0	2	2	4	0	1	0	8	0	8	1	10	0	0	0	166
h	40	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43
j	1	16	11	8	0	13	3	0	12	3	1	18	8	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	3	1	7	0	0	0	125
w	8	2	5	45	2	2	1	0	0	1	2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75
sild	17	2	3	5	1	2	2	10	1	1	2	1	1	15	5	10	2	3	9	1	3	9	23	1	6	22	3	62	5	6	1	3	1	0	0	0	0	0	0	0	238
silf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	2	5	0	0	0	0	0	0	23	0	18	0	0	1	52
sil	13	29	9	45	8	14	18	15	13	16	13	16	8	22	23	11	23	5	35	8	11	6	25	2	4	4	8	22	5	6	1	2	2	0	0	0	0	0	0	0	442
nˆ	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1	25	0	0	0	38
ntˆ	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	5
#	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Tableau des phonèmes en contexte

	Vnarr	Varr	Blb	Lbd	Alv	Cr	Svoy	SIL
i	4	4	27	15	17	222	2	39
e	13	21	25	14	7	136	7	75
e^	0	0	11	9	6	133	0	1
a	14	12	71	48	31	258	2	29
o^	0	0	16	6	4	56	0	0
o	6	12	16	5	5	78	4	28
u	6	8	18	9	7	69	2	21
y	14	12	16	5	6	107	2	31
x	3	3	9	4	6	42	1	38
x^	2	5	48	25	12	145	2	20
a~	4	3	15	3	8	88	2	40
e~	1	1	10	3	1	35	2	31
o~	3	1	14	3	5	46	2	24
x~	0	4	28	3	4	33	3	14
p	66	52	2	1	1	32	3	7
t	91	108	5	1	5	49	7	15
k	65	66	3	1	1	37	5	15
b	41	54	2	2	1	49	2	3
d	45	123	3	1	4	13	5	11
g	24	21	1	0	0	22	4	14
f	35	26	3	1	2	19	2	7
v	47	30	1	1	3	22	7	6
s	65	102	9	1	2	39	9	8
z	42	31	2	1	2	11	3	21
s^	33	22	2	2	2	10	3	7
z^	46	50	3	1	2	7	4	10
r	111	79	23	7	12	91	3	46
l	159	139	16	4	3	33	11	21
m	54	64	4	2	2	18	2	4
n	48	59	13	3	1	22	9	19
h	42	1	0	0	0	0	0	0
j	37	59	3	2	2	10	2	10
w	60	15	0	0	0	0	0	0
sild	42	21	13	12	28	120	2	0
silf	0	0	0	0	0	0	0	0
q	0	0	1	0	2	7	0	42
sil	119	123	35	17	10	146	4	26

Tableau des visèmes en contexte

	Vnarr	Varr	Blb	Lbd	Alv	Cr	Svoy	SIL
Vnarr	31	41	162	89	65	782	14	158
Varr	39	45	163	63	56	673	17	275
Blb	161	170	8	5	4	99	7	14
Lbd	82	56	4	2	5	41	9	13
Alv	79	72	5	3	4	17	7	17
Cr	686	788	80	21	34	331	59	206
Svoy	102	16	0	0	0	0	0	0
SIL	162	143	46	29	36	262	5	0

8.3.2 Corpus I. Main.

	0	**11**	**12**	**13**	**14**	**15**	**21**	**22**	**23**	**24**	**25**	**31**	**32**	**33**	**34**	**35**	**41**	**42**	**43**	**44**	**45**	**51**	**52**	**53**	**54**	**55**	**61**	**62**	**63**	**64**	**65**	**71**	**72**	**73**	**74**	**75**	**81**	**82**	**83**	**84**	**85**
0	-	7	6	2	7	12	6	2	4	0	0	2	2	6	11	6	6	1	4	1	0	17	19	6	2	31	29	0	1	22	7	1	0	0	0	0	1	0	2	0	0
11	13	15	4	4	3	1	8	6	1	0	4	26	6	4	0	7	8	3	1	2	3	7	8	1	3	10	9	4	2	7	1	3	0	1	0	1	3	1	1	3	3
12	4	1	2	0	0	0	4	0	0	0	8	4	4	1	1	2	5	1	1	0	1	3	4	1	1	2	8	0	1	4	3	0	0	0	0	0	11	0	0	0	0
13	3	1	1	0	0	0	3	0	0	0	0	19	0	0	0	2	5	0	0	0	1	4	1	0	0	2	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0
14	1	3	0	0	2	1	5	1	0	0	1	4	2	2	1	0	1	1	1	0	0	8	14	3	2	9	6	0	2	0	0	2	0	0	0	0	0	0	0	0	0
15	6	2	1	1	1	3	8	0	3	1	2	2	5	0	4	2	4	4	3	0	2	9	7	5	5	7	6	1	0	0	3	2	1	1	1	1	0	0	0	0	0
21	18	5	6	3	3	5	9	5	3	0	2	19	6	5	2	10	8	7	3	0	2	17	6	7	2	15	28	6	1	5	5	5	0	0	0	0	7	2	1	6	2
22	2	5	3	1	2	3	5	1	0	0	0	0	0	0	1	0	2	0	1	0	0	8	6	1	1	8	2	1	1	0	0	1	0	0	0	0	9	0	0	0	0
23	3	0	0	2	0	0	14	0	1	1	1	12	0	0	0	0	2	0	0	0	0	4	0	0	2	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	1	0	1	2	1	0	0	0	0	1	0	0	0	0	1	0	0	1	1	1	4	0	0	1	2	1	0	2	0	1	0	0	0	0	1	0	0	0	0
25	3	0	0	0	1	1	2	1	0	0	1	3	1	2	1	2	7	0	0	2	3	1	4	1	1	4	6	0	1	1	2	0	0	0	0	0	1	0	0	0	0
31	29	18	8	5	9	17	20	5	2	2	5	15	7	6	3	4	10	12	2	2	2	32	17	7	4	18	14	7	3	9	5	3	1	0	0	1	4	1	0	10	2
32	7	7	3	1	1	0	12	1	0	1	0	3	0	1	0	2	6	1	0	0	4	15	4	4	1	3	5	0	0	1	1	0	0	0	0	0	8	0	0	1	0
33	0	5	2	0	0	0	4	1	0	0	1	2	1	1	0	0	7	1	0	0	0	15	1	1	0	2	5	0	1	1	1	0	0	0	0	0	1	0	0	0	0
34	2	4	0	0	2	1	4	0	3	2	1	2	1	2	0	3	1	2	1	1	1	5	3	5	0	1	4	1	1	0	0	2	0	0	0	0	0	0	1	0	0
35	9	0	2	1	1	3	2	1	0	0	1	15	1	0	1	9	2	4	1	0	1	3	2	5	5	4	4	0	0	2	0	2	0	0	0	0	2	0	2	0	0
41	12	10	4	2	4	2	10	2	2	0	3	12	6	1	2	3	5	6	5	1	0	9	11	2	2	7	23	16	3	7	3	5	0	0	1	0	2	0	0	6	0
42	3	3	3	3	0	2	3	3	0	2	1	4	3	0	0	0	3	0	1	0	0	11	3	3	2	0	6	1	0	4	3	2	0	0	0	1	7	0	0	0	2
43	5	2	1	1	0	2	3	0	0	0	0	10	0	0	1	0	3	0	1	0	0	3	0	2	1	1	4	0	0	0	0	3	0	0	0	0	2	0	0	0	0
44	1	3	0	0	0	0	3	0	0	1	1	1	0	0	1	0	1	1	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
45	7	0	0	0	1	0	4	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2	1	0	2	6	2	0	0	1	1	1	0	0	0	0	0	0	0	0	0
51	21	10	8	4	8	8	15	3	13	3	3	47	20	6	7	5	9	5	1	4	0	36	14	3	5	15	32	6	6	7	7	7	0	1	0	0	6	1	4	1	3
52	8	14	4	1	0	8	14	2	4	0	3	8	3	1	1	5	13	3	4	1	3	25	5	3	5	15	21	4	3	1	4	2	1	0	0	0	10	0	0	0	0
53	1	5	1	1	0	1	4	0	1	0	0	34	1	1	1	1	14	0	0	0	1	14	5	3	1	4	13	1	2	0	0	0	0	0	0	0	3	0	0	0	0
54	7	7	0	4	1	1	6	2	0	0	0	4	2	1	7	1	1	0	1	0	0	8	6	4	3	3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
55	8	14	2	4	9	5	7	3	4	2	4	11	4	5	0	7	35	4	3	2	1	21	10	11	9	22	1	1	6	4	6	4	0	0	0	0	1	1	0	0	0
61	22	11	6	1	6	7	27	9	3	5	4	21	1	3	6	3	11	7	7	0	3	25	12	14	2	10	12	4	5	4	5	6	0	3	0	0	4	0	3	1	1
62	3	9	1	0	1	0	4	2	0	0	1	4	1	1	0	1	4	1	1	0	0	5	4	1	1	1	2	0	0	0	1	1	0	0	0	0	4	0	1	1	2
63	1	2	2	0	0	0	8	0	0	0	0	13	0	0	0	1	3	0	0	0	1	6	0	0	0	2	2	0	3	0	0	2	0	0	0	0	3	0	0	0	0
64	2	10	0	3	3	5	4	4	2	1	2	3	5	1	1	2	3	3	0	1	0	10	9	4	4	5	1	1	1	0	0	3	0	0	0	0	1	0	0	0	0
65	7	4	2	1	2	2	5	3	0	0	0	1	0	0	2	2	0	0	0	1	0	9	2	3	2	4	2	1	1	0	2	1	1	0	0	0	1	0	0	0	1
71	2	3	0	0	0	0	5	2	1	0	1	8	6	2	0	1	3	4	0	0	0	4	2	2	1	2	9	0	3	1	1	3	0	0	0	0	0	1	0	0	0
72	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
73	1	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
75	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
81	7	5	2	0	1	5	5	3	1	0	1	3	2	1	1	1	1	4	3	1	0	8	7	4	2	10	6	1	1	2	1	2	0	0	0	0	3	0	1	0	0
82	2	0	0	0	0	1	0	1	0	0	0	0	0	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
83	0	1	0	0	0	0	0	0	0	0	0	4	1	0	0	0	2	0	0	0	0	2	0	0	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0
84	6	1	2	0	1	0	2	1	0	1	0	0	1	0	0	1	0	0	0	0	0	5	1	1	1	2	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0
85	4	0	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	2	1	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0

Corpus I. Position.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	-	34	12	27	12	75	61	1	3
1	27	46	55	98	47	116	61	14	22
2	26	44	47	65	40	95	68	7	29
3	47	90	68	79	61	157	65	11	30
4	28	43	38	46	27	74	74	13	20
5	47	120	93	183	105	250	128	15	31
6	35	78	85	72	46	135	47	17	23
7	6	5	9	20	10	13	16	4	1
8	19	22	15	18	13	52	17	3	6

Corpus I. Forme.

	0	**1**	**2**	**3**	**3**	**5**
0	-	71	33	27	47	56
1	126	720	305	174	167	243
2	32	306	76	46	36	91
3	15	285	20	21	10	27
4	19	142	75	46	42	47
5	46	211	77	64	66	124

8.3.3 Corpus II. Visage.

Liste des phrases

0. De quoi je me mêle.
1. a eu huit enfants.
2. l'univers courbe d'Einstein.
3. du 6 au 8.
4. Un honteux film.
5. former un huit.
6. Et oeil disant.
7. Un hibou grand-duc.
8. Quinze heureux élus.
9. sont les bienvenues.
10. un huis-clos.
11. Les huit hiboux hululent en haut des hêtres.
12. les bombes ont gagné.
13. des sandwich huileux.
14. ces huit hérons usent follement leurs huppes.
15. c'est du snobisme.
16. C'est un fou euphorique.
17. surface sphérique.
18. l'Incroyable Hulk.
19. en taille XXL.
20. devenir hugolien.
21. vous êtes triste hein.
22. ou ouvrir une enquête.
23. donnent froid dans le dos.
24. ou brunch le dimanche.
25. à la Tête de goinfre.
26. de voluptueux aussi.
27. le secteur conjugue croissance.
28. et oeillet à la boutonnière.
29. Le premier est haletant.
30. et les gnomes de l'intégration financière.

31. sa houppe blonde descendant.
32. stagne dans les sondages.
33. ai-je reçu le prix.
34. et sa fameuse Ombre jaune.
35. surnommé le Sphinx.
36. puis singapourienne ou hongkongaise.
37. ou n'importe lequel de ses sbires.
38. un happy end.
39. une tranche de pain beurré.
40. A la différence du western-spaghetti.
41. Au-delà du scoutisme.
42. dans une housse de canapé.
43. Au hit-parade des classiques.
44. Il y a un hic.
45. Il est huit heures.
46. avec une double houlette.
47. A bord de cette voiture.
48. Comme un hamster.
49. pousse un hurlement euphorique.
50. Aie un peu d'imagination.
51. si on analyse froidement.
52. nappée d'un bouillon onctueux et truffé ?
53. galopant comme la phtisie.
54. et l'un des monstrueux gamins.
55. Les uns ont un réseau dédié.
56. était pour eux une première.
57. Une perspective unique.
58. ne lui a-t-il rien laissé à lui ?
59. y eurent la leur.
60. sans la heurter.
61. avec les avantages obtenus.
62. on joue uniquement au whist.
63. garde-robe irréprochable.
64. au plan européen est inutile.
65. où l'on accueille 820 élèves.
66. Sur une vidéo enregistrée.
67. devant le pays un peu surpris.
68. la capitale tchétchène.
69. est morte à l'âge de 55 ans.
70. cette humble maison.
71. un institut de langue yiddish.
72. qui n'ont plus une once d'idéologie.
73. présente trop d'ombres ombres juridiques.
74. Il travaille oeuvre après oeuvre.
75. grand psaume de la République.
76. celle d'un hussard.
77. Inracontable.
78. elle se parfume huit fois par jour.
79. les formats et les prix n'enflent pas.
80. bonne année, Il est 1 heure du matin.
81. mais leur présence n'empêche rien.

82. à jeun et heureux.
83. vidéo ou audio.
84. la formation chiite libanaise.
85. et dérange heureusement ses habitudes.
86. + 8
87. La faim oeuvre sur le coup de midi.
88. en tant que pays hôte.
89. ourdissant des complots mystérieux.
90. un bricolage institutionnel.
91. Il compte comme autres actionnaires.
92. Immanquablement.
93. mais tout cela est assez psychologique.
94. Seule ombre au tableau.
95. vaut mieux qu'une mauvaise hutte.
96. L'Europe chinoise 1.
97. Les premières salves.
98. en tchatche et en costumes.
99. que certains opposants jugent xénophobe.
100. presque heurté parfois.
101. prennent une posture humble et mystérieuse.
102. en même temps que l'industrie de la houille.
103. ou plutôt une vision un peu distante et humble.
104. mais tout de même heureux d'avoir dix ans.
105. La vengeance est un plat qui se mange froid.
106. évoluait entre 16 et 17 euros.
107. s'est renforcé face à l'euro et au yen.
108. guinéen ou français.
109. marié à une humble paysanne aux pieds bandés.
110. Depuis sept ou huit ans.
111. D'une carrure ample.
112. Le coucou heureux se balade dans les bois.
113. ces bigres de bougres.
114. un tantinet boy-scout.
115. et pour l'amitié entre les peuples.
116. l'obervatoire présidentiel.
117. aucune équipe institutionnalisée.
118. l'homme à l'allure ascétique oeuvrait depuis comme.
119. ivre de colère.
120. est le maître d'oeuvre depuis août dernier.
121. chants de gorge et accordéons interviennent.
122. veut acheter un meuble en hêtre avec un heurtoir.
123. où il engloutit une huitre sans coup férir.
124. Dans le film Huit femmes.
125. qui fonctionne avec succès outre-Manche.
126. fabrique ainsi depuis sept ans une version.
127. avoir eu accès aux comptes de l'entreprise.
128. Ses délégués y sont restés une grosse heure.
129. Puis-je avancer un début d'explication ?
130. Deux cents heures de rushes.
131. trier et détecter les produits obtenus.
132. Combien y a-t-il de bougies sur le gâteau ?

133. et se dirige droit vers le composteur.
134. Ils ont l'air si heureux.
135. Elle joue avec art de ce sfumato vocal.
136. L'un d'entre eux invite un couple de retraités.
137. à l'université d'Aix-Marseille III.
138. accrochent un peu.
139. ou a eu envie.
140. les nymphes de l'Ariane.
141. l'inspire et la scande.
142. à 139 yens.
143. laser et micro-ondes.
144. n'appartiennent pas au pays organisateur.
145. qui n'atteignirent jamais leurs destinataires.
146. et le groove organique des vrais instruments.
147. filiale de journaux gratuits du groupe Ouest.
148. Cette huitième édition.
149. une taupe orange.
150. La proche communauté nouvelle musique-jazz.
151. néoélitisme urbain.
152. vont stagner cette année.
153. hormis aux hautes latitudes.
154. avec un haussement d'épaules.
155. échevinage dans les tribunaux correctionnels.
156. dans les méandres.
157. et se paye huit mille francs.
158. néo-impressionnisme.
159. et de cohues au soleil.
160. envenimées jusque-là par le dossier tchétchène.
161. Voilà enfin un chien heureux.
162. s'il est oui ou non candidat.
163. nauséeux ou démagogique.
164. Le 29 août.
165. L'oeil d'un yéti luit huit fois.
166. un réceptif installé.
167. Depuis onze ans.
168. que son réseau haute tension.
169. et a heurté un poteau causant.
170. qui ont hanté son septennat.
171. le 25 août.
172. il était redevenu intenable.
173. aura eu chaud.
174. à oeuvrer dans ce sens.
175. teint blanchâtre et ongles rongés.
176. humble et généreuse.
177. Les néons et les machines à sous.
178. lorsque les géants IBM.
179. quelqu'un qui ne fiche rien.
180. un hollandais.
181. il y eut jusqu'au bout.
182. Et le Val d'Aoste.
183. les pièces du puzzle.

184. Aujourd'hui il part.
185. n'a heureusement rien d'innocent.
186. Huit bienheureux ont mangé heure après heure.
187. Avec son yoyo.
188. son époux regimbe.
189. raille un programme européen indécis.
190. rassemble sur une page Web.
191. Lorsqu'il eut fini.
192. il faut bien autre chose que la.
193. mais la situation outre-Rhin.
194. attirent tellement d'oiseaux.
195. En grimpant en haut d'un troène.
196. Il a sauté une haie.
197. Trop perméable aux rumeurs.
198. où on trouve quelques gorilles yankees.
199. préférait rendre hommage au hip-hop.
200. dernier round de la guerre froide.
201. Qu'est-ce que vous regardez comme oiseau.
202. Le loup oisif visite les bergeries.
203. la mise à l'écart du chef historique.
204. La voix est intacte.
205. on leur prodigue une.
206. agnostique préoccupé de théologie.
207. cherche inspiration.
208. C'est un hold-up.
209. et le fait qu'elle ait ou non des enfants.
210. un commerçant coréen en pleurs.
211. le moteur devenait instable.
212. n'ai-je pas été informé.
213. l'homme exhibe un convertisseur.
214. éclipse du pays.
215. Un jour rive droite.
216. L'un d'entre eux en tout cas.
217. dont il est hanté et qui le hante.
218. un job impossible.
219. Vertueux juges.
220. y imaginait en effet qu'après.
221. Ces Indiens-là vivent entre mer.
222. à de rares exceptions près.
223. vrais clivages gauche-droite.
224. Les chansons hymnes.
225. on mélange joyeusement huit choses !
226. picotements et spasmes.
227. La longue odyssée.
228. assez heureux.
229. une pour les humbles.
230. Le mur est haut et escarpé.
231. juillet et août.
232. déjà très prégnante.
233. un hors-série plomb.
234. la série en huit volets conçue.

235. stagnent ou régressent.
236. seule l'étude de la roche in situ.
237. et enfin inscrit sur ordinateur.
238. caméléone.
239. s'étire sur huit régions.
240. il aura 8 ans.
241. et de la honte.
242. garances ou caille-lait.
243. les militants d'Act Up.
244. et week-ends actuellement.
245. punk primitif.
246. de 7,5
247. A la même heure.
248. et le reste principalement en yens.
249. Il signe là une très étrange histoire d'amour.
250. leur plus haut niveau en huit ans.
251. une culture handball.
252. le galbe des sculptures.
253. qui brigue un nouveau mandat.
254. est éclairant.
255. un pays entre en récession.
256. Il est 8 heures passées.
257. En un peu plus d'un an.
258. habillé en hâte.
259. dont huit de l'opposition.
260. réunissant en un seul lieu recherche.
261. au cours d'une une battue aux sangliers.
262. fait autant d'heureux que d'émules.
263. et le dégoût du grunge.
264. et un rendez-vous entre le Nord et le Sud.
265. ne pas être un intellectuel heureux.
266. parfaitement oiseuse.
267. le pressing ou le coiffeur.
268. ou européo-asiatiques.
269. Dix ans ont passé.
270. En version windows.
271. les orgues gloutons de Gainsbourg.
272. Soudain sa voix enfle.
273. dont un néon avec les lettres Action.
274. le pare-brise haut.
275. en a eu assez.
276. Vue sous cet angle.
277. souffla aussitôt la lampe à huile.
278. erreur au-delà .
279. Un hère est un jeune cerf de plus de six mois.
280. anges cannibales.
281. Nous étions 700 ou 800.
282. Harponnera-t-elle parmi eux un futur mari.
283. qui avoisine 8000 tonnes.
284. par le consortium européen Eurodif.
285. c'est un zingueur qu'il nous faut.

286. ou des razzias en horde.
287. propose huit histoires de paternité.
288. ou par e-mail.
289. une oasis de l'âme.
290. encore des petits trous.
291. qu'elles se terminent lundi.
292. puis gagne un hublot ouvert.
293. dumping fiscal.
294. pas de cloisons étanches entre les bureaux.
295. européen ou mondial.
296. Une hiérarchie.
297. l'appui européen a été important.
298. en djellaba.
299. surfent sur les mêmes thèmes.
300. Avec 11 défaites pour 4 victoires.

Tableau des diphones

	i	**e**	**eˆ**	**a**	**oˆ**	**o**	**u**	**y**	**x**	**xˆ**	**a˜**	**e˜**	**o˜**	**x˜**	**p**	**t**	**k**	**b**	**d**	**g**	**f**	**v**	**s**	**z**	**sˆ**	**zˆ**	**r**	**l**	**m**	**n**	**h**	**j**	**w**	**sild**	**silf**	**q**	**sil**	**Tot.**
i	3	1	1	4	2	2	1	2	3	0	1	0	3	1	6	35	29	10	11	4	9	9	32	13	3	2	13	33	9	27	1	6	0	0	18	0	1	295
e	7	3	4	3	1	9	0	2	4	2	6	9	4	6	13	16	9	8	11	7	5	3	18	15	1	5	13	21	9	7	3	0	1	0	22	0	1	248
eˆ	1	1	1	3	1	1	3	2	0	0	2	2	0	1	5	35	20	2	3	0	1	1	11	8	2	2	42	21	8	13	2	3	1	0	6	0	1	205
a	2	2	2	0	1	1	2	5	1	3	3	0	1	1	14	25	12	14	9	7	5	20	22	15	2	12	37	31	12	11	2	6	1	0	14	0	0	295
oˆ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	5	1	1	3	0	0	0	1	7	0	3	0	26	4	11	8	0	1	0	0	0	0	0	72
o	2	1	1	2	0	2	2	2	0	0	2	1	1	0	14	11	6	3	11	2	1	1	7	8	2	5	8	13	6	7	1	2	2	0	13	0	0	139
u	1	0	0	2	0	1	1	2	3	0	1	0	2	0	5	11	3	2	6	2	4	6	4	2	0	1	19	3	2	3	2	2	1	0	3	0	0	94
y	0	0	2	2	0	3	0	0	3	0	1	1	0	0	5	4	3	2	8	3	2	0	11	3	2	3	23	8	5	36	1	0	1	0	6	0	0	138
x	0	1	0	0	0	1	1	1	0	0	1	1	1	1	8	1	5	1	4	4	6	4	6	8	0	1	21	4	5	2	1	1	0	0	8	0	0	98
xˆ	0	0	1	0	0	0	1	0	0	0	1	0	0	1	17	4	8	4	13	6	2	12	13	1	3	4	42	20	12	5	0	4	1	0	0	0	0	175
a˜	1	1	1	1	1	3	1	1	2	0	1	0	1	1	6	25	10	1	19	5	7	4	14	2	5	9	3	9	1	4	3	1	1	0	23	0	0	167
e˜	1	0	1	1	0	1	2	0	2	1	1	2	0	1	5	7	6	3	4	2	2	3	12	3	0	0	1	3	1	2	0	0	0	0	8	0	0	75
o˜	1	1	0	1	0	0	1	1	0	0	1	1	1	1	5	9	4	6	6	4	1	1	7	3	1	3	2	3	2	3	1	1	1	0	12	0	0	84
x˜	2	1	1	2	4	0	0	3	0	1	0	0	2	0	9	2	4	8	6	0	2	0	1	1	1	2	3	0	1	5	2	1	0	0	1	0	0	65
p	6	16	5	26	8	6	9	1	5	3	2	2	0	0	0	6	0	1	0	0	0	0	4	0	1	0	29	15	0	0	9	2	1	0	4	0	0	161
t	28	22	16	15	4	6	4	18	5	10	18	4	1	9	2	2	1	0	4	0	0	2	1	0	7	0	42	3	3	3	4	4	5	0	20	1	0	264
k	13	0	9	15	10	9	10	6	1	12	4	2	12	3	3	9	0	0	2	0	0	0	14	0	0	0	5	8	1	0	0	0	2	0	14	0	0	164
b	6	1	2	5	3	1	8	3	1	2	1	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	9	22	0	0	0	4	1	0	3	0	0	77
d	22	31	5	9	1	3	1	14	17	37	12	1	4	5	0	0	0	1	5	0	0	0	0	0	0	3	6	0	1	0	2	4	2	0	8	0	0	194
g	1	2	2	12	1	5	2	2	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	2	0	0	13	5	0	6	0	1	1	0	0	0	0	59
f	8	5	5	5	8	3	3	2	0	2	2	5	1	0	0	1	0	0	0	0	0	1	0	0	0	0	7	3	0	0	0	1	3	0	2	0	0	67
v	10	0	19	4	2	6	3	2	1	0	5	3	1	0	0	0	1	0	2	0	0	0	0	0	0	0	10	0	0	5	0	1	7	0	2	0	0	84
s	19	20	29	6	2	3	5	14	3	13	19	5	8	1	6	31	13	0	5	0	3	0	1	0	0	0	0	0	7	2	1	24	0	0	4	0	0	244
z	9	9	4	6	1	7	0	3	2	4	9	3	4	2	0	0	1	2	1	0	2	0	0	0	0	0	0	1	3	0	0	3	0	0	9	1	0	86
sˆ	5	1	7	3	0	3	0	0	0	2	3	1	0	1	0	2	1	0	2	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	5	0	0	40
zˆ	9	6	0	4	2	2	6	6	1	4	1	2	0	1	1	0	1	0	3	2	1	0	0	0	0	1	1	0	0	0	1	1	1	0	4	0	0	61
r	29	25	20	22	7	23	4	11	19	21	18	3	3	2	8	12	4	3	21	4	6	2	16	0	3	5	2	12	9	6	1	9	10	0	37	5	0	382
l	12	39	14	47	5	11	4	15	12	41	10	4	5	4	2	6	4	1	6	0	2	1	4	1	0	1	0	1	7	1	3	3	1	0	27	1	0	295
m	13	11	13	13	1	3	1	5	4	4	21	2	2	1	1	2	0	0	2	0	0	0	1	0	0	0	0	1	0	1	1	4	2	0	11	1	0	121
n	15	12	8	19	3	9	5	6	3	9	6	4	4	2	6	3	3	1	5	2	2	3	5	0	2	1	1	1	1	0	2	7	1	0	15	2	0	168
h	45	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48
j	3	14	16	5	2	8	0	0	4	2	2	11	18	2	0	1	0	0	2	0	0	0	1	0	0	0	0	2	0	0	2	1	0	0	2	0	0	98
w	5	0	3	36	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	47
sild	15	21	11	22	1	7	12	7	1	0	12	2	3	16	15	3	15	1	22	5	4	9	29	0	2	1	3	46	5	10	1	0	0	0	0	0	0	301
silf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
q	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	1	0	0	1	1	0	0	1	0	0	0	0	0	2	11
sil	1	0	0	0	0	0	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5

Tableau des phonèmes en contexte

	Vnarr	Varr	Blb	Lbd	Alv	Cr	Svoy	SIL
i	10	14	25	18	5	203	1	19
e	23	37	30	8	6	117	4	23
eˆ	7	11	15	2	4	156	3	7
a	7	17	40	25	14	175	3	14
oˆ	1	0	19	1	3	48	0	0
o	6	10	23	2	7	75	3	13
u	3	10	9	10	1	55	3	3
y	4	8	12	2	5	99	2	6
x	2	6	14	10	1	56	1	8
xˆ	2	2	33	14	7	116	1	0
a˜	5	10	8	11	14	92	4	23
e˜	4	9	9	5	0	40	0	8
o˜	4	5	13	2	4	42	2	12
x˜	6	10	18	2	3	23	2	1
p	53	36	1	0	1	56	10	4
t	90	71	5	2	7	60	9	20
k	40	66	4	0	0	38	2	14
b	15	23	0	0	0	35	1	3
d	72	90	2	0	3	15	4	8
g	18	13	0	0	0	27	1	0
f	23	26	0	1	0	12	3	2
v	33	23	0	0	0	19	7	2
s	75	72	13	3	0	76	1	4
z	30	34	5	2	0	6	0	9
sˆ	17	9	0	1	0	7	1	5
zˆ	20	24	1	1	1	8	2	4
r	98	114	20	8	8	86	11	37
l	116	108	10	3	1	26	4	27
m	51	44	1	0	0	11	3	11
n	56	51	8	5	3	27	3	15
h	47	1	0	0	0	0	0	0
j	40	47	0	0	0	7	2	2
w	44	2	0	0	0	1	0	0
sild	85	45	21	13	3	133	1	0
silf	0	0	0	0	0	0	0	0
q	0	1	0	0	0	7	1	2
sil	1	4	0	0	0	0	0	0

Tableau des visèmes en contexte

	Vnarr	Varr	Blb	Lbd	Alv	Cr	Svoy	SIL
Vnarr	53	89	128	55	32	674	13	64
Varr	31	61	140	57	42	630	17	75
Blb	119	103	2	0	1	102	14	18
Lbd	56	49	0	1	0	31	10	4
Alv	37	33	1	2	1	15	3	9
Cr	635	666	67	23	22	368	37	136
Svoy	91	3	0	0	0	1	0	0
SIL	86	49	21	13	3	133	1	0

Bibliographie

Abry, C., & Boe, L.-J. 1986. Laws for Lips. *Speech Communication*, **5**, 97–104.

Abry, C., Orliaguet, J.-P., & Sock, R. 1990. Patterns of speech phasing. Their robustness in the production of a timed linguistic task : single versus double (abutted) consonants in French. *European Bulletin of Cognitive Psychology*, **10**, 263–288.

Arslan, L.M., & Talkin, D. 1998. 3-D Face Point Trajectory Synthesis using an Automatically derived Visual Phoneme Similarity matrix. *Pages 175–180 of : AVSP Proceedings.*

Attina, V. 2005. *La Langue française Parlée Complété (LPC) : production et perception.* Ph.D. thesis, INPG.

Badin, P., Bailly, G., Revèret, L., Baciu, M., Segebarth, C., & Savariaux, Christophe. 2002. Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images. *Journal of Phonetics*, **30**(3), 533–553.

Bailly, G. 2001. Audiovisual speech synthesis. *Pages 1–10 of :* Taylor, Paul (ed), *ETRW on Speech Synthesis.*

Bailly, G., Gibert, G., & Odisio, M. 2002. Evaluation of movement generation systems using the point-light technique. *Pages 27–30 of : IEEE Workshop on Speech Synthesis.*

Bailly, G., Bérar, M., Elisei, F., & Odisio, M. 2003. Audiovisual speech synthesis. *International Journal of Speech Technology*, **6**, 331–346.

Bailly, G., Elisei, F., Badin, P., & Savariaux, C. 2006. *Degrees of freedom of facial movements in face-to-face conversational speech.*

Bailly, G., Govokhina, O., & Breton, G. 2008a. Multimodal control of talking heads. *In : Acoustics.*

Bailly, G., Govokhina, O., Breton, G., Elisei, F., & Savariaux, C. 2008b. The trainable trajectory formation model TD-HMM parameterized for the LIPS 2008 challenge. *In : Interspeech.*

Basu, S., Oliver, N., & Pentland, A. 1998. 3D lip shapes from video : a combined physical-statistical model. *Speech Communication*, **26**, 131–148.

Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.-A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.-F., Tribout, M., & Vidal, S. 2007. TELMA : Téléphonie à l'usage des malentendants. Des modèles aux tests d'usage. *In : Conférence Internationale sur l'Accessibilité et les systèmes de suppléance aux personnes en situation de Handicaps (ASSISTH).*

Beier, T., & Neely, S. 1992. Feature-based image metamorphosis. *Computer graphics*, 35–42.

Berthommier, F. 2003. *Direct Synthesis of Video from Speech Sounds for New Telecommunication Applications.*

Beskow, J. 1995. Rule-based Visual Speech Synthesis. *Pages 299–302 of : Proceedings of Eurospeech '95.*

Beskow, J. 2003. *Talking Heads - Models and Applications for Multimodal Speech Synthesis.* Ph.D. thesis.

Beskow, J. 2004. Trainable Articulatory Control Models for Visual Speech Synthesis. *Journal of Speech Technology*, **7(4)**, 335–349.

Boite, R., Boulard, H., Dutoit, T., Hancq, J., & Leich, H. 2000. *Traitement de la Parole.* Lausanne : Presses Polytechniques et Universitaires Romandes.

Bowden, R. 2000. *Learning non-linear Models of Shape and Motion.* Ph.D. thesis, Brunel University.

Bregler, C. 1997. Video rewrite : driving visual speech with audio. *Pages 353–360 of :* SIGGRAPH, ACM (ed), *Proceedings of Computer Graphics.*

Breton, G., Bouville, C., & Pelé, D. 2001. FaceEngine a 3D facial animation engine for real time applications : a 3D facial animation engine for real time applications. *Web3D*, 15–22.

Brooke, N., & Scott, S. D. 1998. Two and Three-Dimensional Audio-Visual Speech Synthesis. *Pages 213–218 of : AVSP'98.*

Browman, C. P., & Goldstein, L. 1990a. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, **18**, 299–320.

Browman, C. P., & Goldstein, L. M. 1990. Gestural specification using dinamically-defined articulatory structures. *Journal of Phonetics*, **18**(3), 299–320.

Calliope. 1989. *La parole et son traitement automatique.* Paris, France : Masson.

Campbell, N. 1995. CHATR : A High-Definition Speech Re-Sequencing System. *Pages 1637–1647 of : Eurospeech'95.*

Campbell, N., & Isard, S. D. 1991. Segment durations in a syllable frame. *Journal of Phonetics*, **19**, 37–47.

Caplier, A., Stillittano, S., Aran, O., Akarun, L., Bailly, G., Beautemps, D., Aboutabit, N., & Burger, T. 2007. Image and video for hearing impaired people. *EURASIP Journal on Image and Video Processing*, 14.

Chang, Y.-J., & Ezzat, T. 2005. Transferable videorealistic speech animation. *Pages 143 – 151 of : ACM Siggraph/Eurographics Symposium on Computer Animation.*

Cohen, M. M., & Massaro, D. W. 1993. Modeling coarticulation in synthetic visual speech. *Pages 139–156 of :* Thalmann, N.M., & Thalmann, D. (eds), *Models and Techniques in Computer Animation.* Tokyo : Springer-Verlag.

Cohen, M. M., Massaro, D. W., & Clark, R. 2002. Training a talking head. *Pages 14–16 of : Fourth International Conference on Multimodal Interfaces.*

Combescure, P. 1981. 20 listes de dix phrases phonétiquement équilibrées. *Pages 34–38 of : Revue d'Acoustique*, vol. 56.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. 2001. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(6), 681–685.

Cornett, R. O. 1988. Cued Speech, manual complement to lipreading, for visual reception of spoken language. *Principles, practice and prospects for automation. Acta Oto-Rhino-Laryngologica Belgica*, **42**(3), 375–384.

Cornuéjols, A., & Miclet, L. 2002. *Apprentissage Artificiel.*

Cosatto, E., & Graf, H. P. 2000. Photo-realistic talking-heads from image samples. *Pages 152–163 of : Trans. on Multimedia*, vol. 2.

Cosi, P., Caldognetto, E. Magno, Perin, G., & Zmarich, C. 2002. Labial Coarticulation Modeling for Realistic Facial Animation. *Pages 505–510 of : ICMI.*

Cosker, D., Marshall, D., Rosin, P., & Hicks, Y. 2003. Video realistic talking heads using hierarchical non-linear speech-appearance models. *Pages 2–7 of : Mirage.*

Couteau, B., Payan, Y., & Lavallée, S. 2000. The Mesh-Matching algorithm : an automatic 3D mesh generator for finite element structures. *Journal of Biomechanics*, **33**(8), 1005–1009.

Curinga, S., Lavagetto, F., & Vignoli, F. 1996. Lips Movements Synthesis using Time-Delay Neural Networks. *Pages 183–186 of : Signal Processing VIII Theory and Applications.*

d'Alessandro, C., & Tzoukermann, E. (eds). 2001. *Synthèse de la parole à partir du texte.* Traitement automatique des langues, vol. 42. Paris : Hermès.

Deng, Z., Lewis, J. P., & Neumann, U. 2005. Synthesizing speech animation by learning compact speech co-articulation models. *CGI '05 : Proceedings of the Computer Graphics International 2005*, 19–25.

Dixon, N. F., & Spitz, L. 1980. The detection of audiovisual desynchrony. *Perception*, **9**, 719–721.

Donovan, R. 1996. *Trainable Speech Synthesis.* Ph.D. thesis, University of Cambridge.

Ekman, P., & Friesen, W. 1978. *Facial Action Coding System (FACS) : A technique for the measurement of facial action.* Palo Alto, California. : Consulting Psychologists Press.

Elisei, F., Odisio, M., Bailly, G., & Badin, P. 2001. Creating and controlling video-realistic talking heads. *Pages 90–97 of : Auditory-Visual Speech Processing Workshop.*

Engwall, O. 2000. Are statistical MRI data representative of dynamic speech ? Results from a comparative study using MRI, EMA and EPG. *Pages 17–20 of : International Conference on Speech and Language Processing*, vol. 1.

Engwall, O. 2002. Evaluation of a system for concatenative articulatory visual speech synthesis. *Pages 665–668 of : Proc of ICSLP'2002.*

Eriksson, E. J., Sullivan, K. P. H., & Czigler, P. E. 2002. The importance of anticipatory coarticulation in the perception of round in Swedish front vowels : an investigation comparing natural speech with diphone synthesis. *Pages 665–668 of : Cognitive Science Conference.*

Ezzat, T., & Poggio, T. 1998. MikeTalk : a talking facial display based on morphing visemes. *Pages 96–102 of : Computer Animation.*

Ezzat, T., Geiger, G., & Poggio, T. 2002. MARY101 : A trainable videorealistic speech animation system. *Pages 57 – 64 of :* Vatikiotis-Bateson, E. (ed), *Audiovisual Speech Processing.* MIT Press.

Fagel, S. 2006. Joint Audio-Visual Unit Selection - The JAVUS Speech Synthesizer. *In : International Conference on Speech and Computer*.

Fagel, S., & Clemens, C. 2004. An Articulation Model for Audiovisual Speech Synthesis - Determination, Adjustment, Evaluation. *Speech Communication*, **44**(Special issue on auditory-visual speech processing), 141–154.

Geiger, G., Ezzat, T., & Poggio, T. 2003. *Perceptual evaluation of videorealistic speech*. Tech. rept. Massachusetts Institute of Technology.

Gibert, G. 2006. *Conception et Evaluation d'un système de synthèse 3D de Langue française Parlée Complétée (LPC) à partir du texte*. Ph.D. thesis, INPG.

Girosi, F., Jones, M., & Poggio, T. 1995. Regularization theory and neural networks architectures. *Neural Computation*, **7**, 219–269.

Govokhina, O., Bailly, G., Breton, G., & Bagshaw, P. 2006a. Evaluation de systèmes de génération de mouvements faciaux. *Pages 305–308 of : Journées d'Etudes sur la Parole*.

Govokhina, O., Bailly, G., Breton, G., & Bagshaw, P. 2006b. A new trainable trajectory formation system for facial animation. *Pages 25–32 of : ISCA Workshop on Experimental Linguistics*.

Govokhina, O., Bailly, G., Breton, G., & Bagshaw, P. 2006c. TDA : A new trainable trajectory formation system for facial animation. *Pages 2474–2477 of : InterSpeech*.

Govokhina, O., Bailly, G., & Breton, G. 2007. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. *Pages 305–308 of : ISCA Speech Synthesis Workshop*.

Groleau, J., Chabanas, M., Marecaux, C., Payrard, N., Segaud, B., Rochette, M., P., Perrier, & Y., Payan. 2007. A biomechanical model of the face including muscles for the prediction of deformations during speech production. *Pages 173–176 of : Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA'2007*.

Hallgren, A., & Lyberg, B. 1998. Visual speech synthesis with concatenative speech. *Pages 181–184 of : AVSP 1998*.

Hazen, T. J. 2006. Visual model structures and synchrony constraints for audio-visual speech recognition. *Pages 1082–1089 of : IEEE Transactions on Audio, Speech and Language Processing*.

Hiroya, S., & Honda, M. 2004. Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model. *Pages 175–185 of : IEEE Transactions on Speech and Audio Processing*.

Hong, P., Wen, Z., & Huang, T. S. 2002. Real-Time Speech-Driven Face Animation. *Pages 115–124 of : MPEG-4 Facial Animation*. John Wiley Sons, Ltd.

J.Hardcastle, W., & Hewlett, N. 1999. *Coarticulation : Theory, Data, and Techniques*. Press Syndicate of the University of Cambridge.

Kakumanu, P. 2003. *Analysis and evaluation of factors affecting speech driven facial animation*. Ph.D. thesis, Computer Science Department, WSU.

Kakumanu, P., Gutierrez-Osuna, R., Esposito, A., Bryll, R., Goshtasby, A., & Garcia, O. N. 2001. Speech-driven Facial Animation. *Pages 1–5 of : Proceedings of the ACM conference on Perceptual User Interfaces (PUI 2001).*

Kakumanu, P., Gutierrez-Osuna, R., Esposito, A., & Garcia, O. N. 2002. Comparing Different Acoustic Data-Encoding for Speech-Driven Facial Animation. *Speech Communication*, 598–615.

Klaus, H., Klix, H., Sotscheck, J., & Fellbaumn, K. 1993. An evaluation system for ascertaining the quality of synthetic speech based on subjective category rating tests. *Pages 1679–1682 of : Proceedings of the Third European Conference on Speech Communication and Technology*, vol. 3.

Kozhevnikov, V., & Chistovich, L. 1965. Speech : Articulation and Perception. *Joint Publications Research Service*, 1779–1793.

Lanitis, A., Taylor, C.J., & Cootes, T.F. 1995. A unified approach for coding and interpreting face images. *Pages 368–373 of : Proc. Int. Conf. Computer Vision.*

Le Goff, B., Guiard-Marigny, T., Cohen, M., & Benoît, C. 1994. Real-Time Analysis-Synthesis and Intelligibility of Talking Faces. *Pages 53–56 of : Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis.*

Lee, Y., Terzopoulos, D., & Waters, K. 1995. Realistic modeling for facial animation. *Pages 55–62 of : SIGGRAPH.*

LeGoff, B., & Benoit, C. 1996. A text-to-audiovisual-speech synthesizer for french. *Pages 2163–2166 of : International Conference on Spoken Language Processing (ICSLP).*

Lofqvist, A. 1990. Speech as audible gestures. *Speech Production and Speech Modeling*, 289–322.

Lucero, J. C., Munhall, K. G., Vatikiotis-Bateson, E., Gracco, V. L., & Terzopoulos, D. 1997. Muscle-based modeling of facial dynamics during speech. *The Journal of the Acoustical Society of America*, **101**(5), 3175–3176.

Mak, B., & Banard, E. 1996. Phone clustering using Bhattacharyya distance. *Proceedings of the International Conference on Spoken Language Processing*, **4**, 2005–2008.

Marschark, M., LePoutre, D., & Bement, L. 1998. *Mouth movement and signed communication.* Hove, United Kingdom : Psychology Press Ltd. Publishers.

Massaro, D.W. 1998. *Perceiving Talking Faces : From Speech Perception to a Behavioral Principle.* Cambridge, MA : MIT Press.

Massaro, D.W., & Stork, D.G. 1998. Speech recognition and sensory integration. *American Scientist*, **86**(3), 236–244.

Massaro, D.W., Beskow, J., Cohen, M.M., C.L., Fry, & Rodriquez, T. 1999. Picture My Voice : Audio to Visual Speech Synthesis using Artificial Neural Networks. *Pages 133–138 of : Proceedings from AVSP'99.*

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

Minnis, S., & Breen, A.P. 1998. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. *Pages 759–762 of : International Conference on Speech and Language Processing*, vol. 2.

Nazari, M.A., Payan, Y., Perrier, P., M., Chabanas, & Lobos, C. 2008. A continuous biomechanical model of the face : a study of muscles coordinations for speech lip gestures. *In : Proc. of ISSP Strasbourg.*

Nose, T., Yamagishi, J., Masuko, T., & Kobayashi, T. 2007. A Style Control Technique for HMM-based Expressive Speech Synthesis. *IEICE Trans. Inf. Syst.*, **E90-D**(9), 1406–1413.

Odisio, M., & Bailly, G. 2004. Audiovisual perceptual evaluation of resynthesised speech movements. *Pages 2029–2032 of : Proceedings of the International Conference on Spoken Language Processing.*

Odisio, M., Bailly, G., & Elesei, F. 2004. Tracking talking faces with shape and appearance models. *Speech Communication*, 63–82.

Öhman, S. E. G. 1967. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 310–320.

Öhman, T. 1998. *An audio-visual speech database and automatic measurements of visual speech.* Tech. rept. 1-2. Quaterly Progress and Status Report, Department of Speech, Music and Hearing - KTH.

Okadome, T., Kaburagi, T., & Honda, M. 1999. Articulatory movement formation by kinematic triphone model. *Pages 469–474 of : SMC '99.*

Okadome, T., Suzuki, S., & Honda, M. 2000. Recovery of articulatory movements from acoustics with phonemic information. *Pages 229–232 of : Proceedings of the 5th Seminar on Speech Production.*

Olives, J-L., Möttönen, R., Kulju, J., & Sams, M. 1999. Audio-Visual Speech Synthesis for Finnish. *Pages 157–162 of : Auditory-visual Speech Processing Workshop.*

Pandzic, I., & Frorchheimer, R. 2002. *MPEG4 - Facial Animation.*

Pandzic, I, Ostermann, J., & Millen, D. 1999. Users evaluation : synthetic talking faces for interactive services. *The Visual Computer*, **15**, 330–340.

Parke, F.I. 1974. A parametric model for human faces.

Parke, F.I. 1982. A parametrized model for facial animation. *IEEE Computer Graphics and Applications*, **2**(9), 61–70.

Pelachaud, C., Badler, N., & Viaud, M.-L. 1996. Generating Facial Expressions for Speech. *Cognitive Science*, **20**(1), 1–46.

Perkell, J.S., & Chiang, C.-M. 1986. Preliminary support for a "hybrid model" of anticipatory coarticulation. *In : XII International Congress of Acoustics.*

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., & Salesin, D. H. 1998. Synthesizing Realistic Facial Expressions from Photographs. *Pages 75–84 of : Proceedings of Siggraph.*

Pitermann, M. 2004. Chaos dans la modélisation des tissus mous. *Pages 401–404 of : Journées d'Etude sur la Parole (JEP).*

Platt, S.M., & Badler, N.I. 1981. Animating facial expressions. *Computer Graphics*, **15**(3), 245–252.

Potamianos, G., Neti, C., Luettin, J., & Matthews, I. 2004. Audiovisual automatic speech recognition : an overview. *In : Audiovisual speech processing.* MIT Press.

Rabiner, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Pages 267–296 of :* Waibel, A., & Lee, K. F. (eds), *Readings in Speech Recognition.* San Mateo : CA : Morgan Kaufmann Publishers.

Revéret, L., Bailly, G., & Badin, P. 2000. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *Pages 755–758 of : International Conference on Speech and Language Processing*, vol. 2.

Saltzman, E. L., & Munhall, K. G. 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**(4), 1615–1623.

Schroeder, M. 1967. Determination of the geometry of the human vocal tract by acoustic measurements. *J.Acoust.Soc.Am.*, **41(4)**, 1002–1010.

Scott, K.C., Kagels, D.S., Watson, S.H., Rom, H., Wright, J.R., Lee, M., & Hussey, K.J. 1994. Synthesis of speaker facial movement to match selected speech sequences. *Pages 620–625 of : Australian Conference on Speech Science and Technology.*

Stone, M. 1990. A three dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *Journal of the Acoustical Society of America*, **87**, 2207–2217.

Sumby, W. H., & Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212–215.

Summerfield, Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. *Pages 3–51 of :* Dodd, B., & Campbell, R. (eds), *Hearing by eye : the pyschology of lipreading.* Hillsdale, NJ - USA : Lawrence Erlbaum Associates.

Tachibana, M., Yamagishi, J., Masuko, T., & Kobayashi, T. 2005. Speech synthesis with various emotional expressions and speaking styles by style Interpolation and morphing. *EICE Trans. Inf. Syst.*, **E88-D**(11), 2484–2491.

Tamura, M., Masuko, T., Kobayashi, T., & Tokuda, K. 1998. Visual speech synthesis based on parameter generation from HMM : speech-driven and text-and-speech-driven approaches. *Pages 3745–3748 of : AVSP 1998.*

Tamura, M., Kondo, S., Masuko, T., & Kobayashi, T. 1999. Text-to-audiovisual speech synthesis based on parameter generation from HMM. *Pages 959–962 of : European Conference on Speech Communication and Technology*, vol. 2.

Taylor, P., & Black, A. W. 1999. Speech synthesis by phonological structure matching. *Pages 1531–1534 of : EuroSpeech*, vol. 4.

Terzopoulos, D., & Waters, K. 1990. Physically-based facial modeling, analysis and animation. *The Journal of Visual and Computer Animation*, **1**, 73–80.

Theobald, B.-J., Bangham, J. A., Matthews, I., & Cawley, G. 2003. Evaluation of a talking head based on appearance models. *Pages 187–192 of : Auditory-visual Speech Processing Workshop.*

Toda, T., & Tokuda, K. 2007. A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. *IEICE Transactions on Information and Systems*, 816–824.

Toda, T., Black, A.W., & Tokuda, K. 2008. Statistical Mapping between Articulatory Movements and Acoustic Spectrum with a Gaussian Mixture Model. *Speech Communication*, **50**(3), 215–227.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. *Pages 1315–1318 of : ICASSP.*

van Santen, J. P. H. 1997. Segmental duration and speech timing. *Pages 225–249 of :* Sagisaka, Yoshinori, Campbell, Nick, & Higuchi, Norio (eds), *Computing prosody : Computational models for processing spontaneous speech.* Springer Verlag.

Waters, K. 1987. A muscle model for animating three-dimensional facial expression. *Computer Graphics*, **21**(4), 17–24.

Weiss, C. 2005. FSM and k-nearest-neighbor for corpus based video-realistic audio-visual synthesis. *Pages 2537–2540 of : INTERSPEECH.*

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

Woodland, P. C., Odell, J. J., Valtchev, V., & Young., S. J. 1994. Large vocabulary continuous speech recognition using HTK. *Pages 125–128 of : ICASSP'94.*

Yamamoto, Eli, Nakamura, Satoshi, & Shikano, Kiyohiro. 1998. Subjective evaluation for HMM-based speech-to-lip movement synthesis. *Pages 227–232 of : AVSP-1998.*

Yehia, H.C., Rubin, P.E., & Vatikiotis-Bateson, E. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**, 23–43.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. 1998. Duration modeling for HMM-based speech synthesis. *Pages 692–693 of : ICSLP.*

Zen, H., Tokuda, K., & Kitamura, T. 2004. An introduction of trajectory model into HMM-based speech synthesis. *Pages 191–196 of : ISCA Speech Synthesis Workshop.*