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

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 Télécoms 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.

Abstract

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 implémenter des modèles de contrôle pour l'animation faciale qui puissent générer les 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 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 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 (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 aims to propose and implement control models for facial animation that generate articulatory trajectories from text. We have used 2 audiovisual corpuses in our work. First of all, we have compared objectively and subjectively the principal state-of-the-art models. Then, we have studied the spatial aspect of the articulatory targets generated by HMM based synthesis and concatenation based synthesis. We have decided to combine the advantages of these methods and 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 articulatory targets by concatenation. 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 the articulatory gestures boundaries in function of the

corresponding phoneme boundaries. It has been also applied to automatic synthesis of the 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

LISTES DES FIGURES

LISTE DES TABLES

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 le 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, les n premiers axes de l'ACP sont un meilleur choix, 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 pertinentes que le système de la paire minimale permettrait d'opposer. 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 [□]) 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 résulte d'un compromis entre la production de contrastes acoustiques – et visuels - suffisants pour un effort articulatoire minimal.

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.

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 phonèmes sont associés à chaque segment phonétique.

TDA : *Tasks Dynamics for Animation.* Le terme est propose 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 (Saltzman and Munhall 1989).

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)

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 traités et analysés en fonction de l'information phonétique. La principale problématique de la synthèse de la parole est lié à la grande variabilité intra- et inter- locateurs 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 1, 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 2, les données audiovisuelles qui ont été utilisées dans notre étude sont présentées. Dans le chapitre 3, 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 4, 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 5, 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.

1. ETAT DE L'ART

1.1. SYNTHESE 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). L'entrée d'un tel système est une suite de symboles appartenant à un alphabet fini (une chaine 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 1) :

- un module de traitements linguistiques qui fournit des connaissances linguistiques (chaîne phonétique, structure grammaticale, phonologique) sur l'énoncé à prononcer
- un dictionnaire des 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.
- 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 constituant 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 (Taylor and 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 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 and Tzoukermann 2001), (van Santen 1997), (Boite, Boulard et al. 2000).

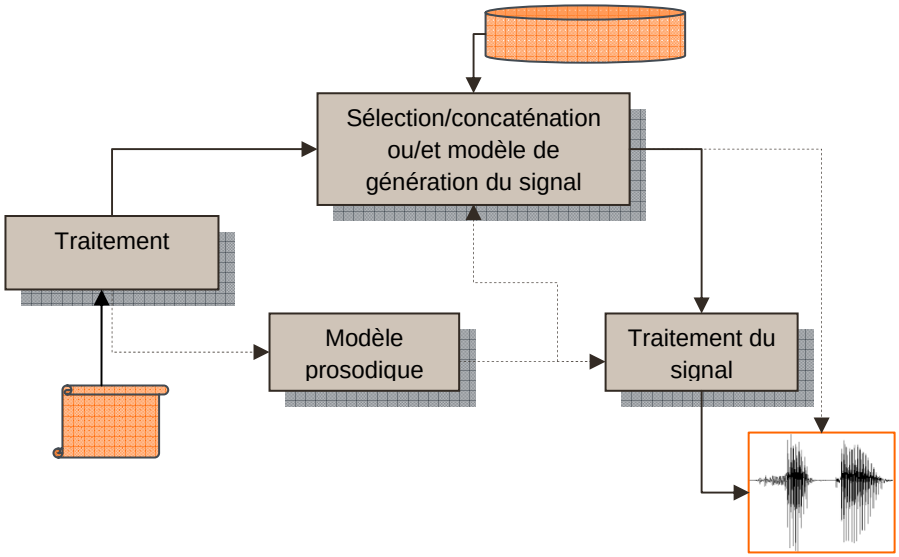


FIGURE 1: SCHEMA GENERAL DE LA SYNTHESE VOCALE

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 chaine de phonèmes marqués en durées.

1.2. MOTIVATIONS DE LA SYNTHESE VISUELLE DA 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 and Pollack

1954). L'importance de la fusion correcte audiovisuelle est démontrée dans l'effet de McGurk[1], (McGurk and 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, LePoutre et al. 1998), (Caplier A., Stillittano et al. 2007). Il y a trois raisons principales pourquoi la modalité visuelle améliore la perception de la parole (Summerfield 1987) : l'information visuelle permet la localisation de la source audio, elle contient de l'information phonétique complémentaire à la parole acoustique et cela permet d'avoir de l'information robuste sur certaines places d'articulation, notamment labiales. Cette visibilité des organes peut être totale (lèvres, joues) ou partielle (langue, dents). L'information sur la place de l'articulation permettrait de lever des ambigüités, 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 and Stork 1998). Les personnes sont très sensibles aux incohérences audiovisuelles spatiales (McGurk and MacDonald 1976) et temporelles (Dixon and Spitz 1980).

1.3. ANIMATION DES VISAGES PARLANTS

Le modèle d'un visage parlant comprend généralement trois modèles (Bailly, Bérar et al. 2003) :

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

---

[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, H. and J. MacDonald (1976). "Hearing lips and seeing voices." Nature **264**: 746-748. ont présenté une vidéo montrant une personne prononçant un phonème (p.ex. /ga/) alors que la bande sonore diffuse l'enregistrement d'un autre phonème (p.ex. /ba/). Lorsque les signaux sont bien synchronisés, le système perceptif est piégé et ne perçoit qu'un unique percept produit de la fusion multimodale (/da/ ou /va/).

Cette chaîne de traitement est illustrée dans la Figure 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.

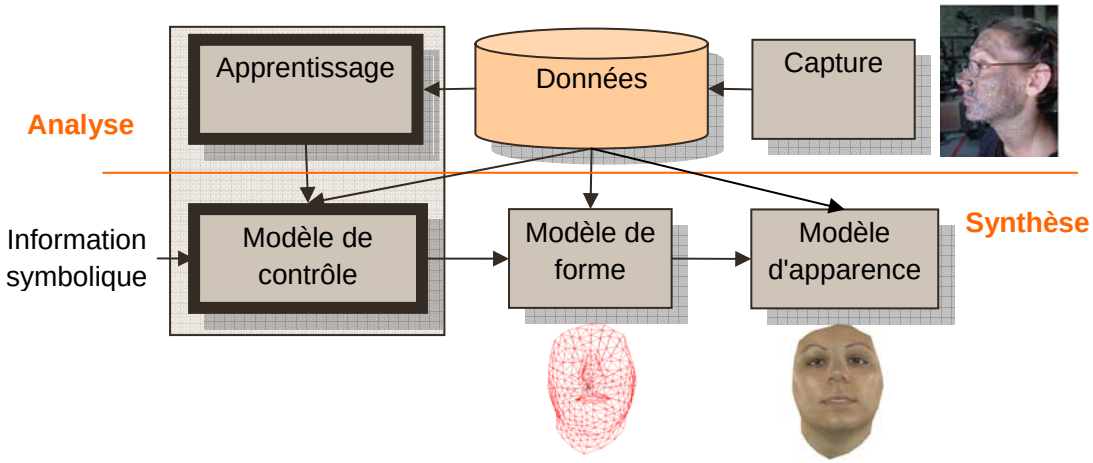


FIGURE 2: MODELE GENERAL D'UNE TETE PARLANTE. ON DISTINGUE ICI LES DONNEES ET TRAITEMENTS NECESSAIRES A L'ANALYSE HORS-LIGNE ET LES MODULES SOLLICITES POUR LA SYNTHESE EN LIGNE.

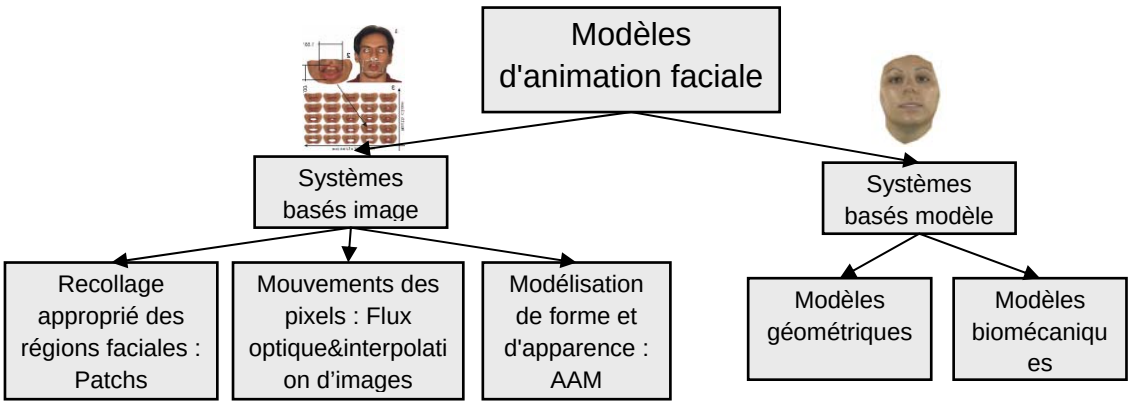


FIGURE 3: SYSTEMES D'ANIMATION FACIALE: SYSTEMES BASES IMAGE OU VIDEO ET SYSTEMES BASES MODELE.

1.3.1. MODELES 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 3.

ANIMATION VIDEO

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, Bérar 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 and Graf 2000); les systèmes qui considèrent les mouvements comme des déplacements de pixels (Ezzat, Geiger et al. 2002); et les systèmes qui calculent l'apparence de chaque pixel en fonction des mouvements faciaux (Brooke and Scott 1998), (Theobald, Bangham et al. 2001).

- Patchs

Dans le premier groupe de modèles un système de synthèse caractéristique de l'approche c'est Video Rewrite proposé par Bregler (Bregler 1997). Ce système utilise une vidéo de fond qui sert de scène aux mouvements des lèvres synthétisés, Figure 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 viseme, 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 and 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 4b. Il faut ensuite agencer les mouvements de toutes ces régions de manière cohérente. Plus une région sera grande, plus la cohérence au niveau de cette région sera maintenue automatiquement.

a) b)

FIGURE 4: EXEMPLES DE SYSTEMES UTILISANT LA SUPERPOSITION DE SEGMENTS VIDEO. A) SYSTÈME DE SYNTHÈSE VIDEO REWRITE (BREGLER 1997), B) SYSTÈME DE SYNTHÈSE PROPOSÉ PAR (COSATTO AND GRAF 2000).

- Flux optique et interpolation d'images

Dans le deuxième groupe de la synthèse 2D modélisant les mouvements des pixels en fonction du son prononcé, les systèmes suivants peuvent être

cités : Actors (Scott, Kagels et al. 1994), MikeTalk (Ezzat and Poggio 1998) et Mary101 (Ezzat, Geiger et al. 2002).

Dans le système MikeTalk (cf. Figure 5.a), une image prototypique (le viseme) 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 viseme 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 (cf. Figure 5.b), 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 and Ezzat 2005).

a) b)

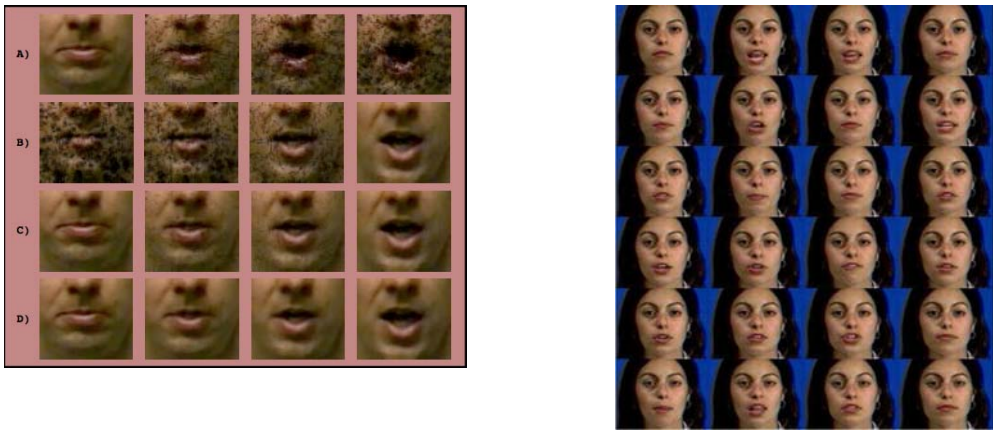


FIGURE 5: A) « MIKETALK » : DE HAUT EN BAS : TRANSFORMATION D'UNE IMAGE I0 (VISEME0) VERS UNE IMAGE1 (VISEME1), TRANSFORMATION D'UNE IMAGE I1 A UNE IMAGE I0, MORPHING DES IMAGES I0 ET I1, MORPHING DES IMAGES I0 ET I1 APRES UN FILTRAGE (EZZAT AND POGGIO 1998). B) « MARY101 » : 24 DES 46 IMAGES PROTOTYPIQUES CONSTITUANT LE MMM (EZZAT, GEIGER ET AL. 2002).

- AAM (*Active Appearance Models*)

Dans le troisième groupe les méthodes consistent à créer un modèle statistique de forme et d'apparence du visage (Cootes, Edwards et al. 2001), (Theobald, Bangham et al. 2003), (Cosker, Marshall et al. 2003). 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 ses 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», voir la Figure 6). Chaque

image est ainsi caractérisée par un nombre constant de pixels. Cependant, construire un modèle statistique de forme et d'apparence d'un visage en utilisant une ACP donne lieu à des artefacts : les variations de texture sont non linéaires (apparition/disparition de rides/plis, des dents, de la langue, etc.). Une solution est proposée à modéliser séparément ces différentes régions et créer des MAM (*Multi-segment Appearance Models*) (Theobald, Bangham et al. 2001). 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.

FIGURE 6: MODELES DE FORMES ET D'APPARENCE : A) LE MODELE DE FORME EST UTILISE POUR NORMALISER LES IMAGES DE LA BASE D'APPRENTISSAGE ; B) IMAGES DE SYNTHESE CREEES A PARTIR DU MODELE DE FORME ET D'APPARENCE (THEOBALD, BANGHAM ET AL. 2001).

FIGURE 7: EXEMPLES DE DESCENDANTS DU MODELE 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, Möttönen et al. 1999), (Massaro 1998) se basent sur ce modèle pour créer leurs têtes parlantes, Figure 7.

Une norme standardisée MPEG4 (Pandzic and R. 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).

FIGURE 8: A) LE PREMIER GESTE ARTICULATOIRE STATISTIQUEMENT SIGNIFICATIF ISSU DE L'ACP CORRESPONDANT A L'ARRONDISSEMENT DES LEVRES. B) LE SECOND GESTE ARTICULATOIRE CORRESPONDANT AUX MOUVEMENTS INTRINSEQUES DE LA LEVRE INFERIEURE

L'équipe de l'ICP (Badin, Bailly et al. 2002), (Elisei, Odisio et al. 2001), (Revéret, Bailly et al. 2000) propose de définir des paramètres articulatoires issus d'une ACP (Analyse en Composantes Principales) guidée (voir Figure 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 ex. arc mandibulaire pour rotation de la mâchoire). Ces points de peau sont de l'ordre de 200 et tiennent 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, Odisio et al. 2001) de la variance des mouvements faciaux liés à la production de parole.

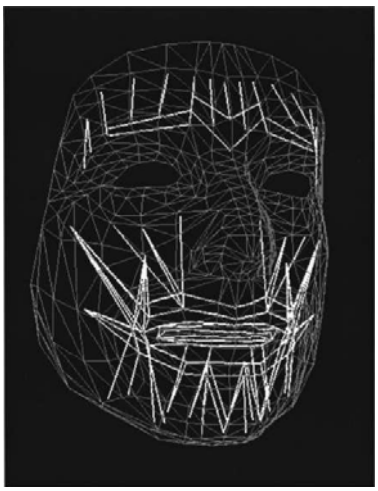


FIGURE 9: LIGNES D'ACTION DES MUSCLES DU VISAGE DU MODELE DE LUCERO ET AL. (LUCERO, MUNHALL ET AL. 1997).

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 nombreux paramètres et sont souvent de très grande complexité (Bailly, Bérar 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 and 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 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 and Badler 1981) (Breton, Bouville et al. 2001) et organisés en couches (Waters 1987), (Terzopoulos and Waters 1990), (Lee, Terzopoulos 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, Oliver et al. 1998), (Couteau, Payan et al. 2000), (Groleau J., Chabanas M. et al. 2007).

1.3.2. MODELES DE CONTROLE

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

PROBLEMATIQUE DE LA COARTICULATION

La parole continue est caractérisée par une grande variabilité dans les domaines articulatoire et acoustique (J.Hardcastle and 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 chaine parlée. Par exemple: [t] dans les segments [ti] vs [tu] où il est influencé par les voyelles qui suivent : dans le [ti] le [t] est non-arrondi et dans le [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 chaine 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, "Il est craintif". - La consonne sonore [r] est partiellement assourdie par l'occlusive sourde [k]. Un exemple des différentes représentations des réalisations consonantiques dans des contextes vocaliques obtenues grâce aux Rayons X sont dans la Figure 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) - la position et la forme de la langue mais aussi du pharynx qui varient, 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.

/aba/ /ibi/ /ibu/ /ada/ /idi/ /idu/ /aga/ /igi/ /igu/

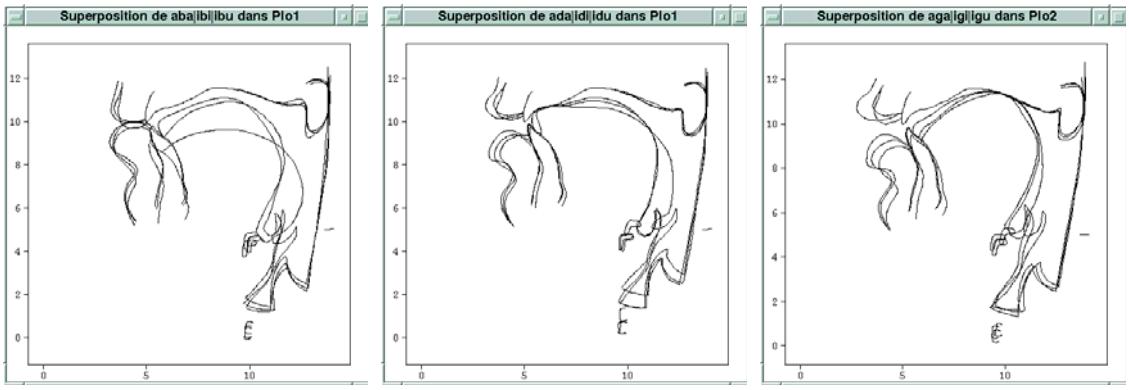


FIGURE 10 : REALISATIONS DES CONSTRICTIONS CONSONNANTIQUES DANS LES DIFFERENTS CONTEXTES VOCALIQUES (IMAGES PAR LES RAYONS X). DANS L'ORDRE : SUPERPOSITION DES CONSTRICTIONS DES BILABIALES DES /ABA/, /IBI/, /IBU/ ; APICO-DENTALES /ADA/, /IDI/, /IDU/ ; DORSO-VELAIRES /AGA/, /IGI/, /IGU/.

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 : modèles basés sur les équations de Lofqvist (Lofqvist 1990) qui sont

développées pour quantifier la coarticulation dans les syllabes du type CV et le modèle d'Öhman (Öhman 1967) qui est développé pour les séquences du type VCV. Ce dernier 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 est alors assez simple

$$p(x,t) = v(x,t) + kc(t) * wc(x) * [c(x) - v(x,t)]$$

EQUATION 1 : EQUATION D'ÖHMAN

- 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 d'un pure geste vocalique, c(x) est la cible consonantique, kc(t) la valeur de l'émergence d'une consonne (=1 à la fermeture des lèvres) et wc(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 and Chistovich 1965), le modèle de coproduction (Perkell and Chiang 1986) et le modèle hybride(Browman and Goldstein 1990a), voir la section 3.3.1.

Pour avoir plus de détails sur le phénomène de la coarticulation le lecteur peut se référer à (J.Hardcastle and 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 cet effet de la coarticulation spécifique à un locuteur. Dans la Figure 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.

Différentes systèmes de formation de trajectoires articulatoires ont été proposées: nous commencerons par détailler celles 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.

Dans ce qui suit, les différents systèmes de synthèse de la parole visuelle de l'état de l'art existant sont présentés, Figure 12.

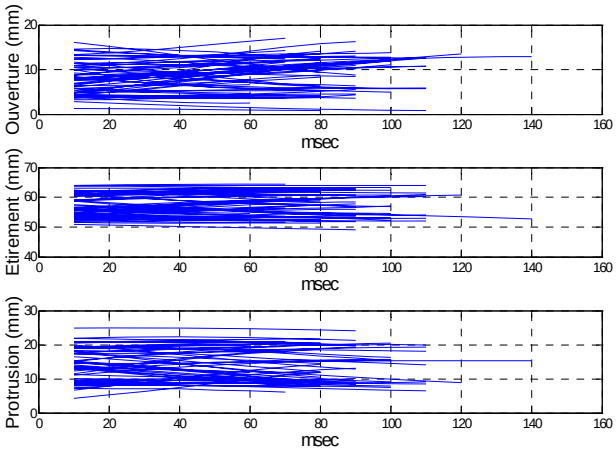


FIGURE 11 : TOUTES LES REALISATIONS DE LA CONSONNEES /G/ DU CORPUS II SELON LES TROIS PARAMETRES GEOMETRIQUES : OUVERTURE, ETIREMENT ET PROTRUSION DES LEVRES.

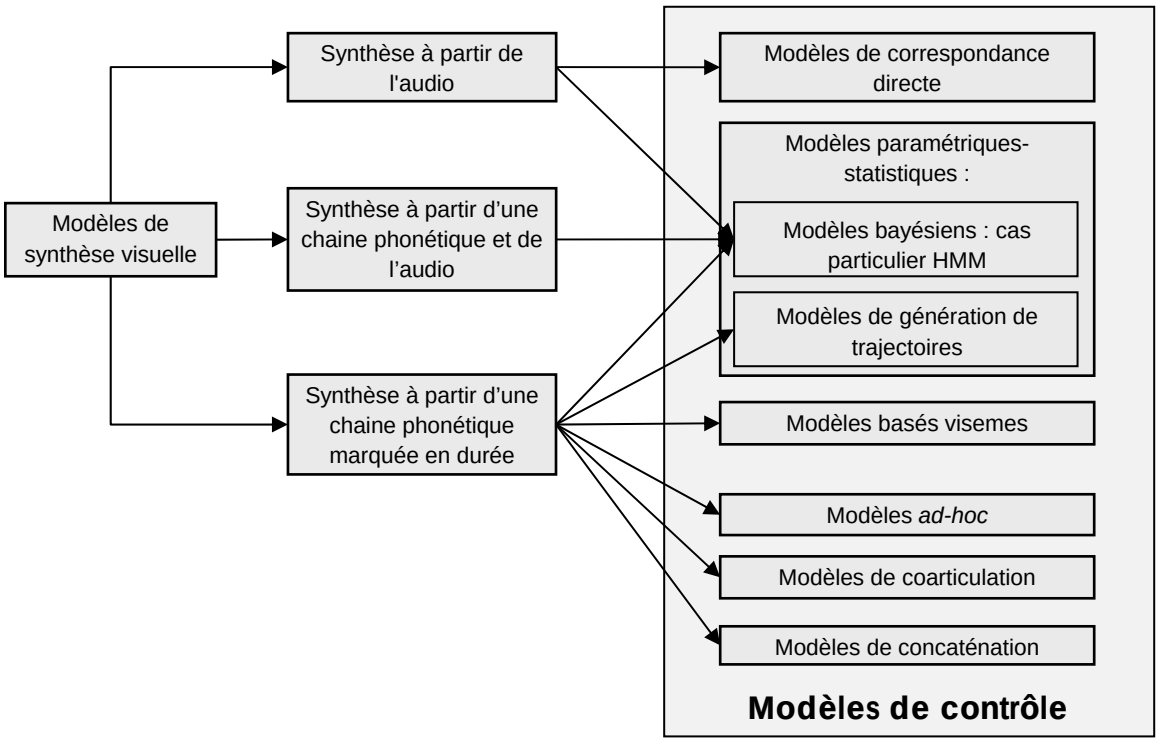


FIGURE 12: MODELISATION DE LA PAROLE VISUELLE : MODELES BASES SUR L'INFORMATION PHONETIQUE ET MODELES BASES SUR L'INFORMATION ACOUSTIQUE.

SYNTHESE A PARTIR D'UNE CHAINE PHONETIQUE

MODELES BASES VISEMES

Dans les modèles basés visèmes, les trajectoires articulatoires sont obtenues en suivant certaines règles. Les règles sont souvent définies empiriquement ou extraites à partir des observations ou expériences. Le modèle le plus répandu dans l'animation des visages parlants a été proposé par Cohen et Massaro (Cohen and 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 Figure 11). 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 décroissement. Ces trois valeurs sont ajustées en fonction de chaque phone et de chaque paramètre articulatoire (voir

Figure 13).

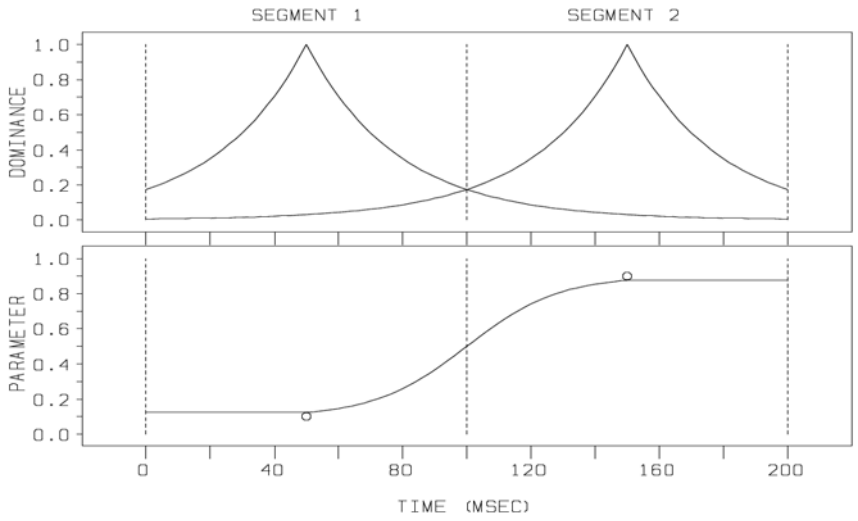


FIGURE 13: MODELE DE DOMINANCE DE COHEN&MASSARO : LA PARTIE DU HAUT CORRESPOND AUX FONCTIONS DE DOMINANCE DE 2 SEGMENTS PHONETIQUES; LA PARTIE DU BAS CORRESPOND A LA TRAJECTOIRE D'UN PARAMETRE ARTICULATOIRE OBTENUE COMME UNE SUPERPOSITION DES GESTES ARTICULATOIRES DE 2 SEGMENTS.

Un autre modèle basé règles a été proposé par Beskow (Beskow 1995). Dans ce modèle, la 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, Badler 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. Le modèle *look-ahead* est utilisé (Kozhevnikov and Chistovich 1965; Cohen and Massaro 1993). Selon ce modèle l'expansion d'un mouvement articulatoire est proportionnelle à l'intervalle entre deux segments clés qui commence à *l'offset* du premier segment clé et se termine au début du deuxième segment clé.

« MASSY », le modèle de visage proposé par Fagel et al (Fagel and Clemens 2004) utilise le modèle de Cohen & Massaro pour son animation. Le modèle est animé par 6 paramètres articulatoires (cf. Figure 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 d'allemand sont regroupés en 15 groupes de visemes (15 voyelles et 9 groupes de consonnes). Pour chaque groupe de visèmes et chaque paramètre, un modèle de dominance est défini.

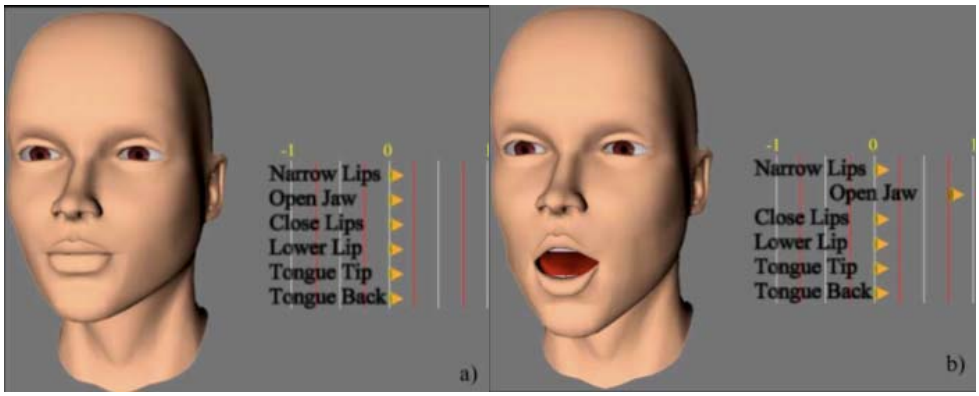


FIGURE 14 : UN EXEMPLE DU VISAGE « MASSY » ANIME PAR SIX PARAMETRES AVEC LES ARTICULATEURS (A) EN POSITION NEUTRE ET (B) LES ARTICULATEURS AVEC LA DESCENTE MAXIMALE DE LA MACHOIRE.

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 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 assez naturels et proches des mouvements réels, car il est 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).

MODELES BASES DONNEES

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 visemes 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.

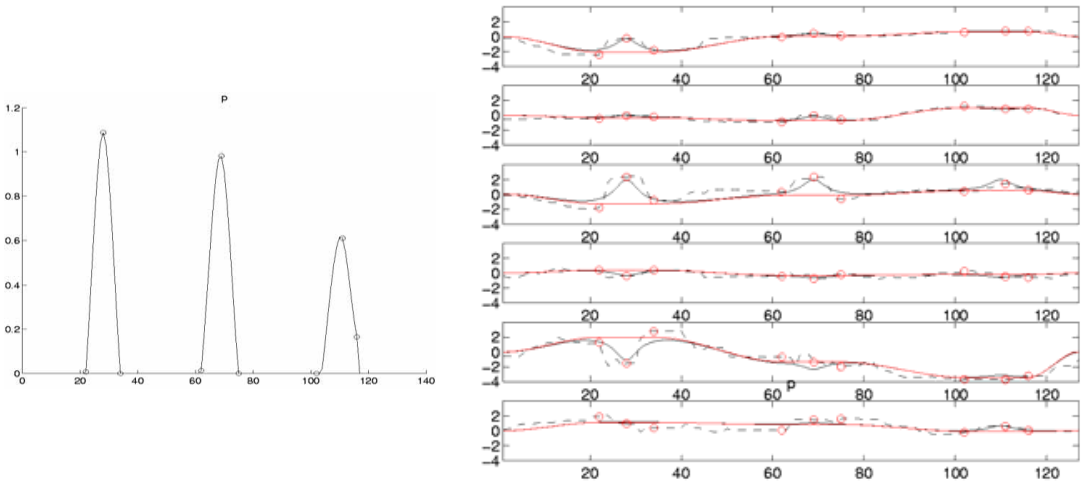


FIGURE 15: SYNTHESE BASEE SUR LE MODELE D'ÖHMAN (ÖHMAN 1967). A GAUCHE: FONCTIONS D'EMERGENCE *KC(T)* POUR [P] EN [APA] [IPI] [UPU], A DROITE: SYNTHESE DES 6 PARAMETRES ARTICULATOIRES A PARTIR DU TEXTE: [APA] [IPI] [UPU]. DU HAUT EN BAS : OUVERTURE DE LA MACHOIRE, PROTRUSION DES LEVRES, FERMETURE DES LEVRES, MONTEE DES LEVRES, AVANCEMENT DE LA MACHOIRE, MOUVEMENTS DU LARYNX. LES POINTILLES CORRESPONDENT AUX MOUVEMENTS CAPTURES, LES LIGNES EN ROUGE – GESTES VOCALIQUES ET LES LIGNES EN NOIR AUX GESTES FINAUX.

MODELES DE COARTICULATION

Une approche pour modéliser la parole visuelle à partir des données est de paramétrer les modèles de coarticulation (Cohen and Massaro 1993), (Öhman 1967), (Cosi, Caldognetto et al. 2002). Le Goff et Benoit (LeGoff and Benoit 1996) ont ainsi effectué l'apprentissage du modèle de dominance de Cohen et Massaro (Cohen and Massaro 1993) appliqué à aux données d'un locuteur de langue française. Les phones sont classées en 19 visemes et les mouvements faciaux sont contrôlés par 8 paramètres (Le Goff, Guiard-Marigny 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 : lle 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) est aussi utilisé par l'équipe de l'ICP (Revéret, Bailly et al. 2000) pour faire de la synthèse de la parole (cf. Figure 15 et équation 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).

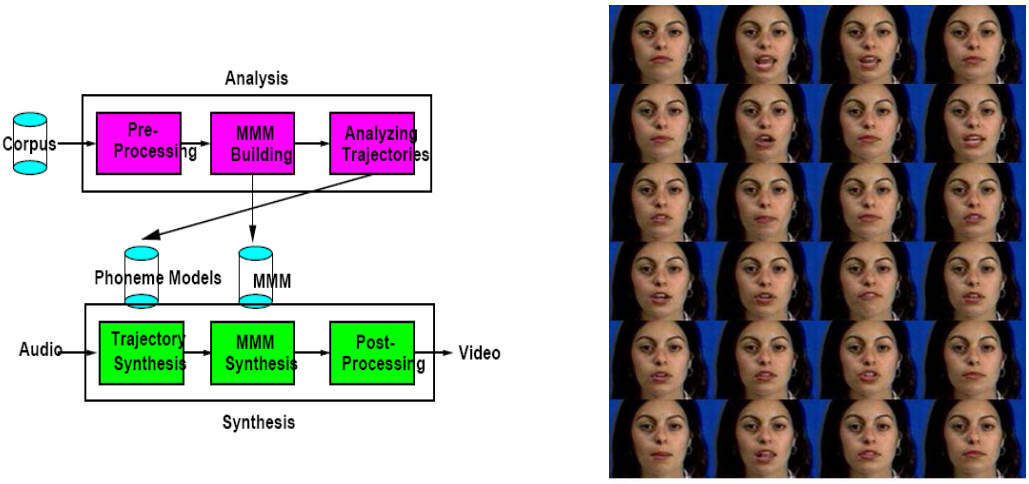


FIGURE 16: MODELE DE SYNTHESE PROPOSE PAR T. EZZAT (EZZAT, GEIGER ET AL. 2002). A GAUCHE: SCHEMA D'ANALYSE ET DE SYNTHESE DE "MARY101", A DROITE: 24 DES 46 IMAGES PROTOTYPIQUES CONSTITUANT LE MMM

MODELES DE GENERATION DE TRAJECTOIRES

Dans le groupe de modèles génériques, 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, Geiger et al. 2002) qui produit de la synthèse basée-image réaliste et proche du naturel (Geiger, T. Ezzat 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 (46 images) d'images prototypes. 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, Jones 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 propres 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 trouver la trajectoire de synthèse, 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 des 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.

MODELES DE CONCATENATION

En 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, Figure 1: tout d'abord, on dispose d'un dictionnaire de segments multi-représentés, ensuite les couts de sélection et de concaténation sont calculés entre les segments candidats, enfin les segments finaux sont choisis grâce au calcul d'un chemin de couts minimaux. Les couts de sélection sont, généralement, calculés en fonction des distances phonologiques et grâce à un modèle externe, le plus souvent, c'est 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 le plus simple consiste à concaténer des segments de vidéos. Les systèmes proposés par (Weiss 2005) et (Fagel 2006) proposent de concaténer des segments correspondants 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 fonds 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, Taylor 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 par déformation élastique (Beier and Neely 1992) aux images de fond. "*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 que avec ce modèle il y a un problème de synchronisation entre la vidéo et l'audio, notamment, pendant les plosives, les mouvements ne sont pas synchronisés avec l'audio.

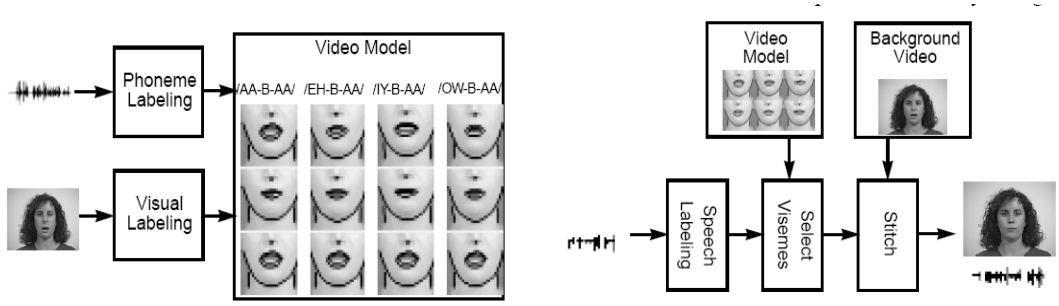


FIGURE 17: SCHEMA D'ANALYSE ET DE LA SYNTHESE DE "VIDEOREWRITE" (BREGLER 1997). A GAUCHE: PRINCIPE DE CONSTRUCTION DU MODELE VIDEO, A DROITE: PRINCIPE DE SYNTHESE A PARTIR DE L'AUDIO.

Hallgren et Lyberg (Hallgren and 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 paramètres 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 and 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, Lewis 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 des mouvements faciaux liés à la parole neutres. De plus, les expressions émotionnelles sont rajoutées aux trajectoires de coarticulations neutres.

FIGURE 18: QUELQUES TRAMES DE LA PAROLE SYNTHETIQUE (DENG, LEWIS 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, et 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 des 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.

MODELES BASES 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, Kondo et al. 1999).

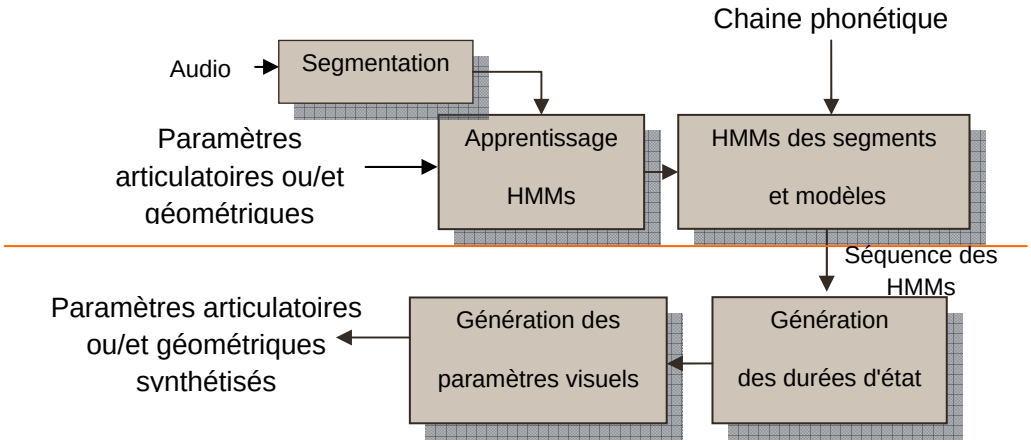


FIGURE 19: PRINCIPE DE SYNTHESE PAR HMM

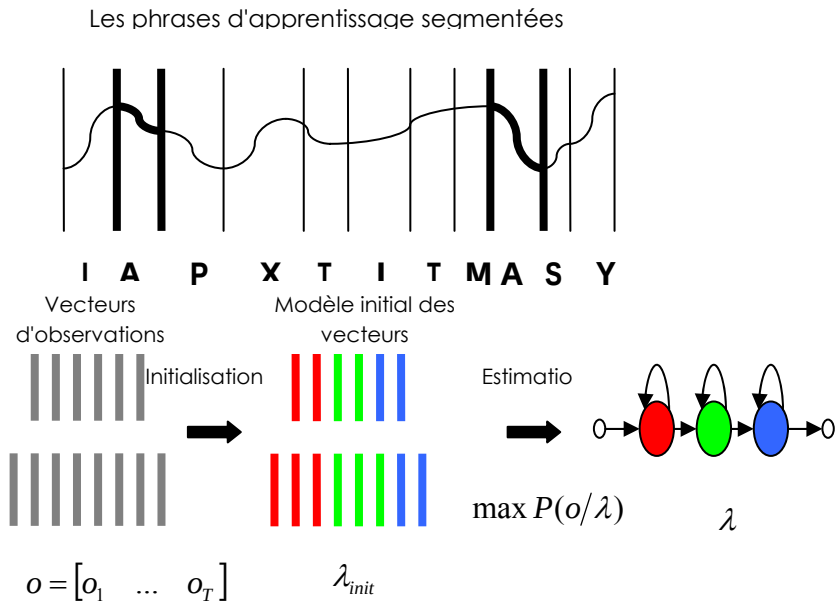


FIGURE 20: PRINCIPE D'APPRENTISSAGE DES HMM PAR SEGMENT. DANS CET EXEMPLE, COLLECTE DES DONNEES PUIS APPRENTISSAGE DE MODEHORS CONTEXTE.

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 (cf. Figure 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 (cf. Figure 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, Tokuda et al. 1998). Le plus souvent, les modèles des 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 à un modèle d'élasticité proposé par Campbell et Isard (Campbell and Isard 1991).Pendant la phase de synthèse, les HMM corresponds 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 (basée ML) qui est basé sur la connaissance de la relation entre les paramètres statiques et les paramètres dynamiques (cf. Figure 21).

Il existe des nombreux travaux sur la synthèse vocale ou audiovisuelle de la parole par HMM dans les laboratoires japonais : (Tokuda, Yoshimura et al. 2000), (Zen, Tokuda et al. 2004), (Tamura, Masuko et al. 1998), (Tamura, Kondo 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 d'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, Yamagishi et al. 2005), (T. Nose 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 and Tokuda 2007) ont récemment proposé une solution possible à ce problème avec une méthode considérant la variance globale des trajectoires obtenues.

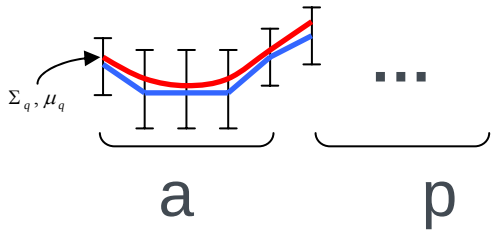


FIGURE 21: PRINCIPE DE GENERATION DES TRAJECTOIRES FINALES A PARTIR DES HMM.

SYNTHESE A 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*.

MODELES 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 d'étudier les relations complexes entre les paramètres acoustiques et visuels. Dans en premier temps on extrait des paramètres acoustiques (Kakumanu, Gutierrez-Osuna 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, Gutierrez-Osuna et al. 2001). Le signal acoustique est ensuite divisé en trames de longueur assez courte (10-20 msec) pour pouvoir être considérée comme quasi-stationnaire.

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 Parameters :* LSP),
- les caractéristiques prosodiques de la parole (énergie, fréquence fondamentale, …),
- celles de perception du système d'audition (coefficients cepstraux, *Mel-Frequencies Coefficients :* MFCC, …).

Yehia et al. (Yehia, Rubin et al. 1998) étudient un modèle de régression linéaire pour décrire les associations entre 11 paramètres acoustiques (10 LSP et 1 rms amplitude) et 12 (ou 18) paramètres faciaux. En effet, 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 grandement améliorés.

Okadome et al.(Okadome, Suzuki et al. 2000) utilisent 30 coefficients LPC pour coder chaque fenêtre des trames acoustiques. L'idée d'utilisation des coefficients LPC est fondée sur le fait que la glotte agit comme un filtre sur la source de parole qui peut être caractérisé justement par ces coefficients. Une autre façon de présenter ces coefficients LPC est de les transformer en coefficients LSF ou LSP. Yehia et al. (Yehia, Rubin 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 nombreux travaux, les coefficients cepstraux sont utilisés (Massaro 1999) (Curinga, Lavagetto et al. 1996), (Hong, Wen et al. 2002).

En ce qui concerne la modélisation, des modèles linéaires (Yehia, Rubin et al. 1998) ou non-linéaires sont utilisés (Massaro 1999), (Curinga, Lavagetto et al. 1996), (Hong, Wen 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 cachés 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 avec 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 22.

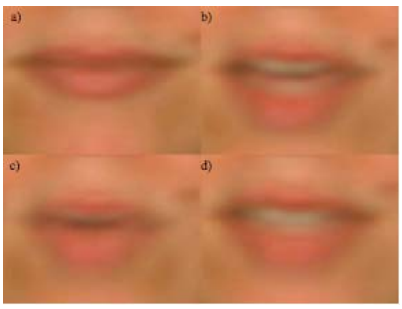


FIGURE 22 : EXEMPLES DES TRAMES GENEREES A PARTIR DE L'AUDIO GRACE AU MODELE DE TRANSFORMATION LINEAIRE PROPOSE PAR (BERTHOMMIER 2003).

Lavagetto (Curinga, Lavagetto 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 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 (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 de test et de 0.64 pour l'ensemble de validation.

Hong et al. (Hong, Wen et al. 2002) utilise 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 and 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 validation est 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, Black 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 menées sur al base MOCHA d'articulations linguales sont impressionnantes.

SYNTHESE A PARTIR DE L'AUDIO AVEC DE L'INFORMATION PHONETIQUE

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, Suzuki 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, Kaburagi 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 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 and Miclet 2002). Enfin, les transitions entre les moyennes des états sont lissées en utilisant l'approche de Tokuda et al. (Tokuda, Masuko et al. 1995).

Tamura et al. (Tamura, Masuko 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, Klix 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.

1.4. PROBLEMATIQUE DE L'EVALUATION

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 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, Ostermann 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 and MacDonald 1976). Dans ce qui suit les différents travaux sur l'évaluation des visages parlants sont présentés.

Pandzic et al. (Pandzic, Ostermann 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 polygonale 3D, un visage avec une texture collée et un visage basé image (Cosatto and 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 progrèsser dans le domaine de 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[2] qui est faite sur 40 sujets prononçant 460 phrases phonétiquement équilibrées. Les données de synthèse sont comparées avec

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

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.

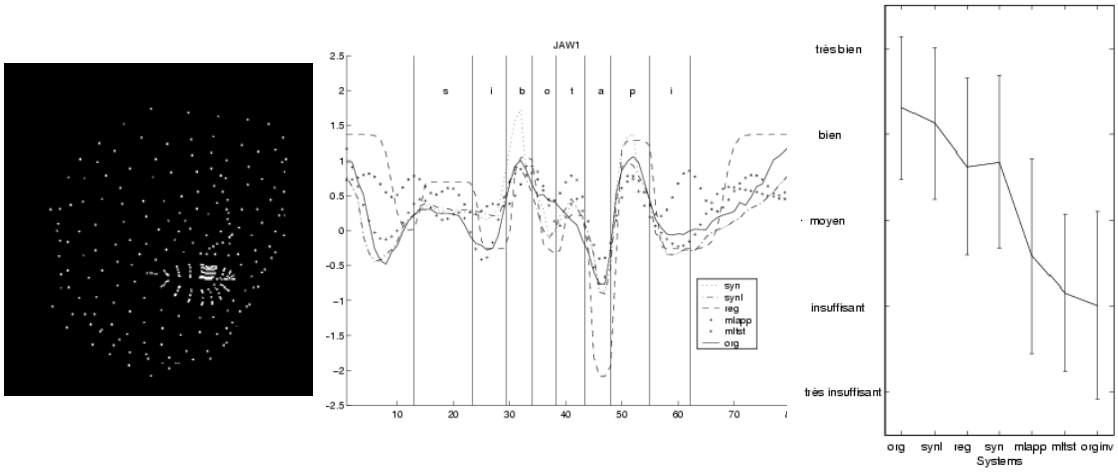
Yamamoto et al. (Yamamoto, Nakamura et al. 1998) effectuent des évaluations objectives et subjectives sur des modèles de contrôle basés sur l'information acoustique (modèle de régression et HMM). D'après les évaluations objectives les HMM donnent de meilleurs résultats que les modèles de régression. 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, T. Ezzat et al. 2003) effectuent une série d'évaluations subjectives sur le réalisme et l'intelligibilité du système *"Mary"* de T. Ezzat (Ezzat, Geiger et al. 2002) avec la participation de 24 sujets. Trois tests de Turing sont effectués pour distinguer les images réelles et celles animées avec ou sans l'audio. En moyenne, les sujets ne font de différence entre les images de synthèse et celles réelles. Pour évaluer l'intelligibilité du système, les sujets doivent lire 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 signifiantes (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 (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 sur les données sont construits dans le but de reproduire un style de production de la parole d'un sujet.

Bailly et al. (Bailly, Gibert et al. 2002) proposent d'utiliser un modèle d'apparence par des points lumineux (Figure 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 23) : deux modèles de concaténation de diphones (*Syn et Sinl*), un modèle d'Öhman (*Reg*) et un modèle de régression linéaire (avec des données de test différentes des 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 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.



a) b) c)

FIGURE 23: EVALUATIONS OBJECTIVES ET SUBJECTIVES PAR BAILLY ET AL. (Bailly, Gibert et al. 2002) DES SYSTEMES DE SYNTHESE VISUELLE (CONCATENATION SANS ET AVEC LISSAGE *SYN* ET *SYNL*, REGRESSION LINEAIRE POUR LES DONNEES D'APPRENTISSAGE ET DE TEST *MLAPP* ET *MLTST*, MODELE D'OHMAN *REG*). A) MODELE FACIALE EN *POINT LIGHTS*; B) EXEMPLE DE LA SYNTHESE DU PARAMETRE ARTICULATOIRE JAW1 MOUVEMENTS DE LA MACHOIRE POUR LA PHRASE "SIX BEAUX TAPIS"; C) RESULTATS DU TEST MOS POUR LES DIFFERENTS MODELES.

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 and Bailly 2004) par la complexité accumulée à partir de chaque modèle de la chaîne de synthèse audiovisuelle. Les mouvements corrects peuvent être jugés inadéquats si on a un mauvais modèle d'apparence. De plus, 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 existantes. 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.

1.5. EVALUATION DES MODELES DE L'ETAT 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 l'apprentissage. Les paramètres visuels utilisés sont les sept paramètres articulatoires.

1.5.1. MODELES DE CONTROLE UTILISES

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 MODELE LINEAIRE GUIDE PAR L'ACOUSTIQUE

Les paramètres articulatoires (120 Hz) sont calculés directement à partir des paramètres acoustiques (120 Hz). 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 Pair*) et l'énergie sont calculés et lissés (Yehia, Rubin 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 MODELE STATISTIQUE-PARAMETRIQUE. LE MODELE 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 B). Les vecteurs d'observation sont constitués des paramètres visuels statiques et dynamiques, c'est-à-dire, des valeurs des paramètres articulatoires et des 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.

SYNTHESE. La synthèse est effectuée comme suit. D'abord, la chaine phonétique à synthétiser est découpée en diphones (avec leurs durées respectives). Ensuite, une séquence des HMMs correspondants est construite. Les durées des états sont déterminées (Yoshimura, Tokuda 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, Tokuda 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.

CONCATENATION

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 7). 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, Gibert et al. 2002).

CONCATENATION BASEE 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.

1.5.2. 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 1. 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.

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

TABLE 1 : CORRELATION MOYENNE ENTRE LES TRAJECTOIRES DE SYNTHESE ET CELLE D'ORIGINE POUR LES DIFFERENTS MODELES ET PHRASES.

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 1. La corrélation moyenne est maximale pour la synthèse par HMMs. Dans le

cas d'une phrase, la corrélation est plus importante pour le modèle linéaire que pour le modèle de concaténation.

1.5.3. 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, Gibert et al. 2002), (Geiger, T. Ezzat et al. 2003).

Le signal acoustique original est joué en synchronie avec les mouvements faciaux. Ici, le test 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 tant de fois qu'ils 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. Le modèle le plus 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, Gibert 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 évalués subjectivement.

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

TABLE 2: RESULTATS DES EVALUATIONS SUBJECTIVES.

1.5.4. DISCUSSION

Dans ce chapitre, des différentes méthodes de synthèse visuelle sont 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.

2. DONNEES AUDIOVISUELLES

2.1. DE LA CAPTURE DES MOUVEMENTS AU CLONAGE D'UNE TETE PARLANTE

Dans cette partie, les principes globaux de la construction d'une tête parlante sont décrits, (voir Figure 24).

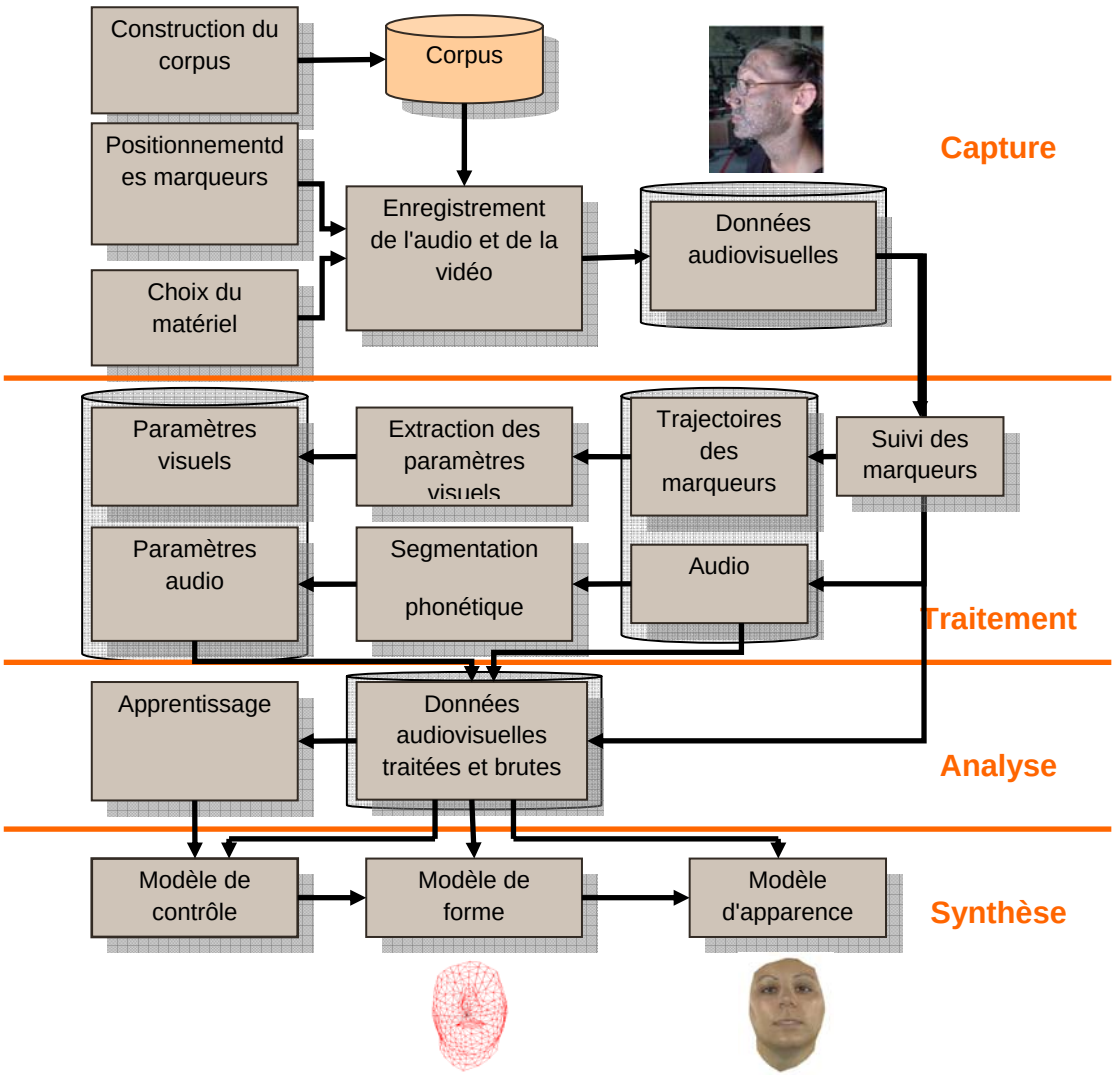


FIGURE 24: SCHEMA GLOBAL DU SYSTEME DE SYNTHESE AUDIOVISUELLE: DE L'ACQUISITION DES DONNEES AUDIOVISUELLES A LA SYNTHESE DES MOUVEMENTS LIES A LA PAROLE.

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[3] 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 2.3.1 et 2.2.2.

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 image (2D) vs. basées vidéo (3D) avec diverses résolutions spatiales et les méthodes internes (par exemple, l'enregistrement des données du visage comme la langue) vs. externes. Une synthèse des méthodes d'enregistrement est présentée dans la Figure 25.

Méthode	Statique/Dynamique	Information fournie	2D/3D	Interne/externe	Publications
Photogrammétrie 3D	Statique ou dynamique	Marqueurs +texture	3D	externe	(Parke 1982), (Elisei, Odisio et al. 2001), (Pighin, Hecker et al. 1998)
Scan Laser 3D	statique	Géométrie +texture	3D	externe	(Cohen, Massaro et al. 2002)
Ultrasons	statique	Géométrie	3D	interne	(Stone 1990)
IRM	statique	Géométrie	3D	interne	(Engwall 2000)

[3] 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).

Basées vidéo	dynamique	Texture	2D ou 3D	externe	(Öhman 1998), (Basu, Oliver et al. 1998)
Photogrammétrie infrarouge	dynamique	Marqueurs	3D	externe	ELITE, QUALISYS, VICON, …
Electromyographie, Electropalatographie, Articulographie électromagnétique	dynamique	Marqueurs	2D	externe	(Lucero and Munhall 1999), (Engwall 2000)

FIGURE 25 : SYNTHESE DES METHODES D'ENREGISTREMENT DES DONNEES VISUELLES POUR LA CONSTRUCTION D'UNE TETE PARLANTE.

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 2.3.2 et2.2.4.

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, Neti et al.). 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, Bailly et al. 2004), construits à partir de données étiquetées semi-automatiquement par un suivi automatique de marqueurs (Bailly, Elisei et al. 2006) puis vérifiés à la main (voir la section 2.2.5).

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).

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[4] et vérifiée à la main.

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 pouvoir de trouver les relations entre la chaine phonétique et les paramètres visuels correspondants.

Pendant la phase de synthèse les paramètres visuels sont générés en fonction de la suite phonétique en entrée grâce au modèle de contrôle obtenu.

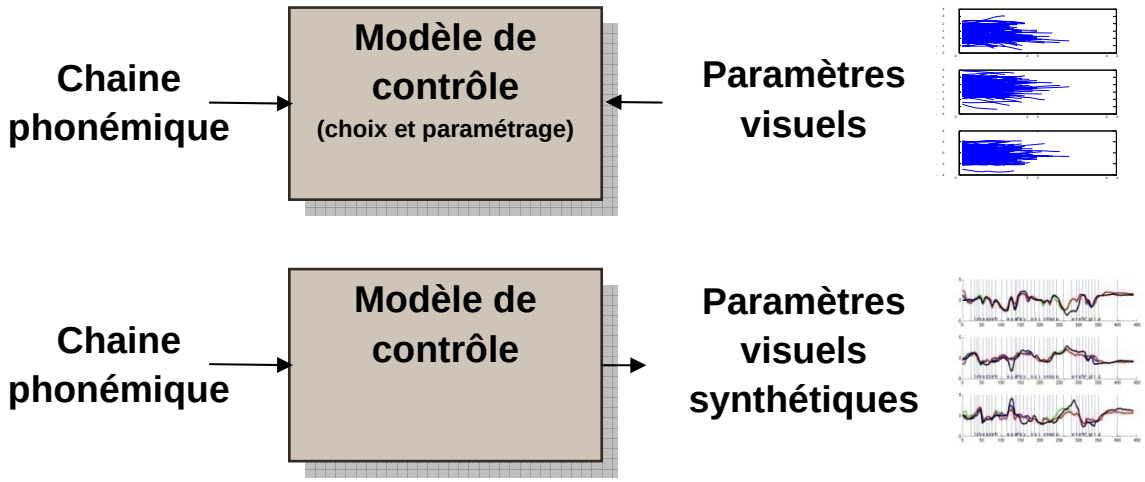


FIGURE 26 : LES OBJECTIFS DE L'ANALYSE ET DE LE SYNTHESE DE LA PAROLE VISUELLE.

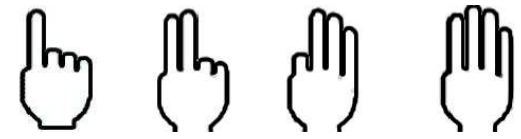
2.2. CORPUS I

Le corpus I est, à la base, construit pour la modélisation audiovisuelle du Langage Parlé Compété (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.

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

2.2.1.LPC: LANGAGE PARLE COMPLETE

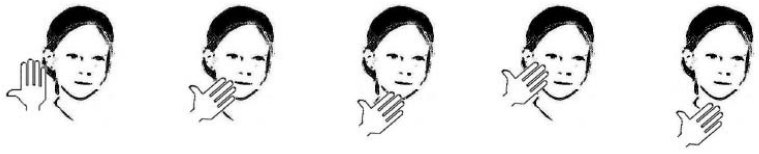
Le Langage Parlé Complété, renommé récemment Langue française Parlée Complétée (ou code LPC), a été créée par le Dr Orin R. Cornett en 1967 sous le nom de ***Cued Speech*** pour l'anglais américain. Ce système manuel complétant la lecture labiale a été adapté depuis à plus de 50 langues (Cornett 1988). 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) pour indiquer une position sur le visage (déterminant un sous-ensemble de voyelles cf. Table 4) pour chaque unité CV qu'il prononce (si le locuteur se retrouve à prononcer une consonne non liée à 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 de l'indice phonétique et par conséquent la détermination de façon univoque du discours.



conf. 1	conf. 2	conf. 3	conf. 4
p (par)	k (car)	s (sel)	b (bar)
d (dos)	v (va)	ʁ (rat)	n (non)
ʒ (joue)	z (zut)		ɥ (lui)

conf. 5	conf. 6	conf. 7	conf. 8
t (toi)	l (la)	g (gare)	j (fille)
m (ami)	ʃ (chat)		ŋ (camping)
f (fa)	ɲ (vigne)		
[1]	w (oui)		

TABLE 3 FORMES DE LA MAIN DU CODE LPC POUR LE FRANÇAIS.



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

TABLE 4: POSITIONS DE LA MAIN PAR RAPPORT AU VISAGE DU CODE LPC POUR LE FRANÇAIS.

2.2.2. COUVERTURE PHONETIQUE

Le corpus I se compose de 238 phrases phonétiquement équilibrées (référencées en 6.3.1). Il est construit pour mettre en œuvre des systèmes de synthèse de parole par concaténation de polysons (multimodaux). Les tableaux de la couverture phonétique (répartition des phonèmes et des phonèmes en contextes) du corpus I sont présentés dans l'annexe 6.3.1. L'histogramme des nombres de représentants des diphones est présenté dans les Figure 30 et Figure 31 et les fréquences d'apparition des phonèmes en contexte pour les deux corpus et le dictionnaire sont présentées dans la Figure 32. Comme pour le corpus II, nous avons comparé la fréquence de l'apparition des diphones dans le corpus I à celle des diphones dans le dictionnaire composé de 500.000 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 phonèmes en contexte non représentés dans le corpus I est comparable à celui (21) du dictionnaire. Les fréquences d'apparition des phonèmes en contexte du corpus I sont proches de celles du dictionnaire (Figure 32). Avec ces données, il est possible de confirmer que le corpus I est phonétiquement équlibré.

2.2.3. REPARTITION DES DICLES

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 5 et Table 6, les formes de main sont au nombre de 8 plus une forme "repos" - codée « 0 » - cf. Table 3 ; les positions de la main par rapport au visage sont au nombre de 5 plus une position "repos" - codée « 0 » - cf. Table 4). En revanche, toutes les transitions (forme + position) 1 vers (forme + position) 2 ne sont pas présentées (cf. Corpus I. Main. en Annexe C). 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.

	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

TABLE 5 : NOMBRE DE REPRESENTANTS LORS DES TRANSITIONS DE POSITION A POSITION. LA POSITION 0 CORRESPOND A LA POSITION DE LA MAIN EN DEBUT 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

TABLE 6: NOMBRE DE REPRESENTANTS LORS DES TRANSITIONS DE FORME A FORME. LA FORME 0 CORRESPOND A LA FORME DE LA MAIN EN DEBUT ET FIN DE PHRASE (POSITION "REPOS").

2.2.4. ACQUISITION DES DONNEES

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.

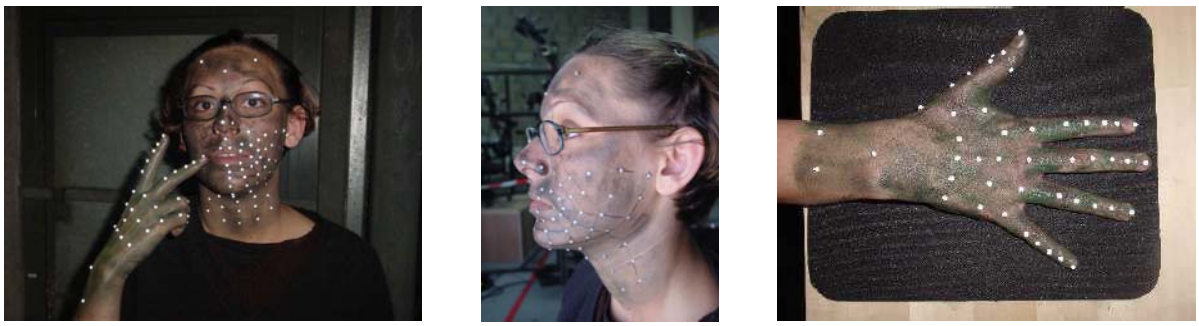
LA LOCUTRICE - CODEUSE

La codeuse (20 ans au moment de l'enregistrement) pratique quotidiennement la Langue Française Parlée Complétée depuis 7 ans avec sa jeune sœur 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 codeur très prochainement.

LE MATERIEL

La phase d'enregistrement s'est déroulée dans les locaux d'Attitude Studio[5]. Une première phase a consistée 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'1 million de pixels). Les capteurs sont placés sur le visage et la main de la codeuse comme représenté sur la Figure 27. 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.



(a) visage : vue de face (b) visage : vue de profil (c) main : vue du dessus

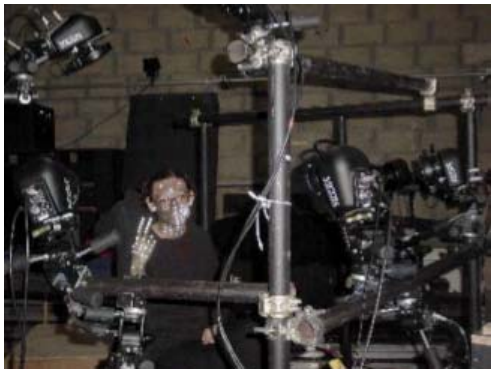
FIGURE 27: POSITION DES MARQUEURS SUR LA CODEUSE LORS DE L'ENREGISTREMENT.

Outre les marqueurs, le système de caméras est disposé selon deux configurations différentes en fonction des corpus à enregistrer afin d'éviter les

[5] 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.

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 28. 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.

(a) configuration main seule



(b) configuration main + visage

FIGURE 28: CONFIGURATIONS DES CAMERAS 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. 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.

2.2.5. EXTRACTION DES PARAMETRES VISUELS ET DE LA MAIN

PRETRAITEMENTS

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, Odell 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 connaitre 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.

MODELISATION 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

PARAMETRES 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, Bailly et al. 2000), (Elisei, Odisio et al. 2001), (Badin, Bailly et al. 2002), (Bailly, Bérar 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 œuvre 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) ;
2. étirement/protrusion des lèvres (paramètre Lips1) ;
3. montée/descente de la lèvre inferieure (paramètre Lips2) ;
4. montée/descente de la lèvre supérieure (paramètre Lips3) ;
5. montée/descente des commissures (paramètre Lips4) ;
6. avancée/rétraction de la mâchoire (paramètre Jaw2) ;
7. montée/descente du larynx (paramètre Lar1).

Pour chaque paramètre la variance du mouvement total expliquée par celui-ci est obtenue, comme référencée dans la Table 7.

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

TABLE 7: VARIANCE EXPLIQUEE ET CUMULEE DES PARAMETRES ARTICULATOIRES ET DE ROTO-TRANSLATION PILOTANT LE MODELE DE VISAGE.

PARAMETRES GEOMETRIQUES

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)

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

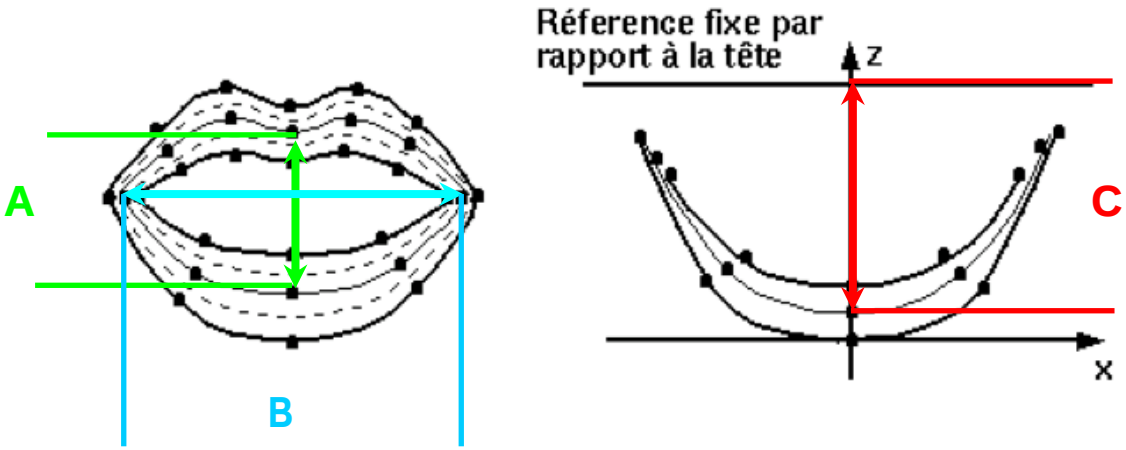


FIGURE 29: PARAMETRES GEOMETRIQUES UTILISES.

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 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 8.

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.

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

TABLE 8 : VARIANCE EXPLIQUEE ET CUMULEE DES PARAMETRES ARTICULATOIRES ET DE ROTO-TRANSLATION PILOTANT LE MODELE DE MAIN.

RESUME

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.

2.3. CORPUS II

2.3.1. COUVERTURE PHONETIQUE

Le corpus II est constitué de 301 phrases. Les phrases utilisées sont dans l'annexe 6.3.1. 201 phrases choisies aléatoirement sont utilisées dans le corpus de test et 100 phrases sont utilisées dans le corpus d'apprentissage. Les tableaux de couverture phonétique (répartition des phonèmes et des phonèmes en contextes) du corpus II sont présentés dans l'annexe 6.3.1. L'histogramme des nombres de représentants des diphones est présenté dans les Figure 30 et Figure 31 et les fréquences d'apparition des phonèmes en contexte pour les deux corpus et le dictionnaire sont présentées dans la Figure 32. Nous avons comparé la fréquence d'apparition des diphones dans le corpus II à celle des diphones dans le dictionnaire composé de 500.000 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 divisèmes non représentés dans le corpus. Les fréquences d'apparition des phonèmes en contexte du corpus II sont proches de celles du dictionnaire (Figure 3). Ainsi, il est possible de conclure que le corpus II est phonétiquement équlibré.

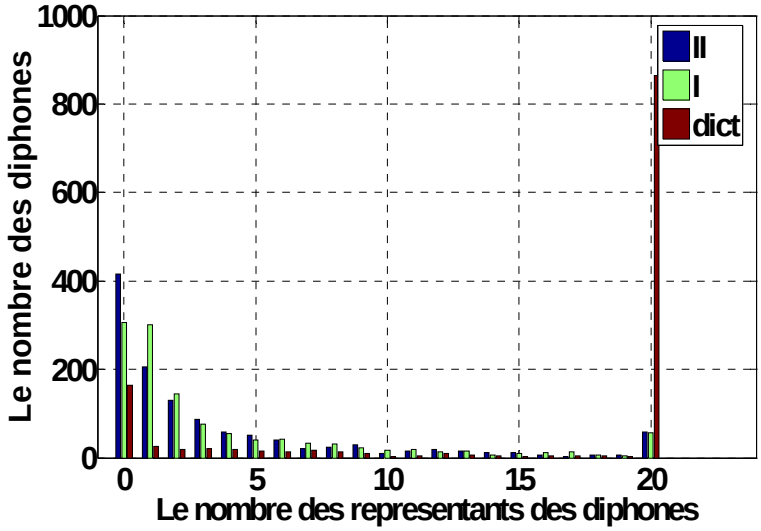


FIGURE 30: NOMBRE DES DIPHONES EN FONCTION DU NOMBRE DES REPRESENTANTS DE CES DIPHONES.

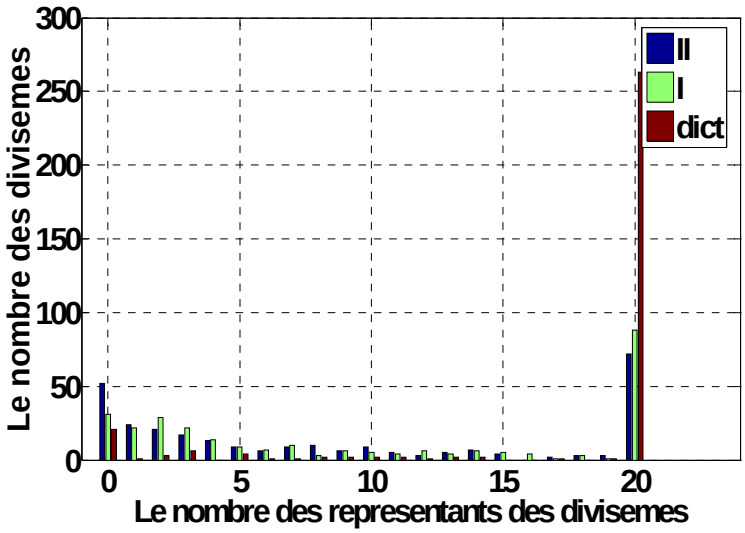


FIGURE 31: NOMBRE DES DIVISEMES EN FONCTION DU NOMBRE DES REPRESENTANTS DE CES DIVISEMES.

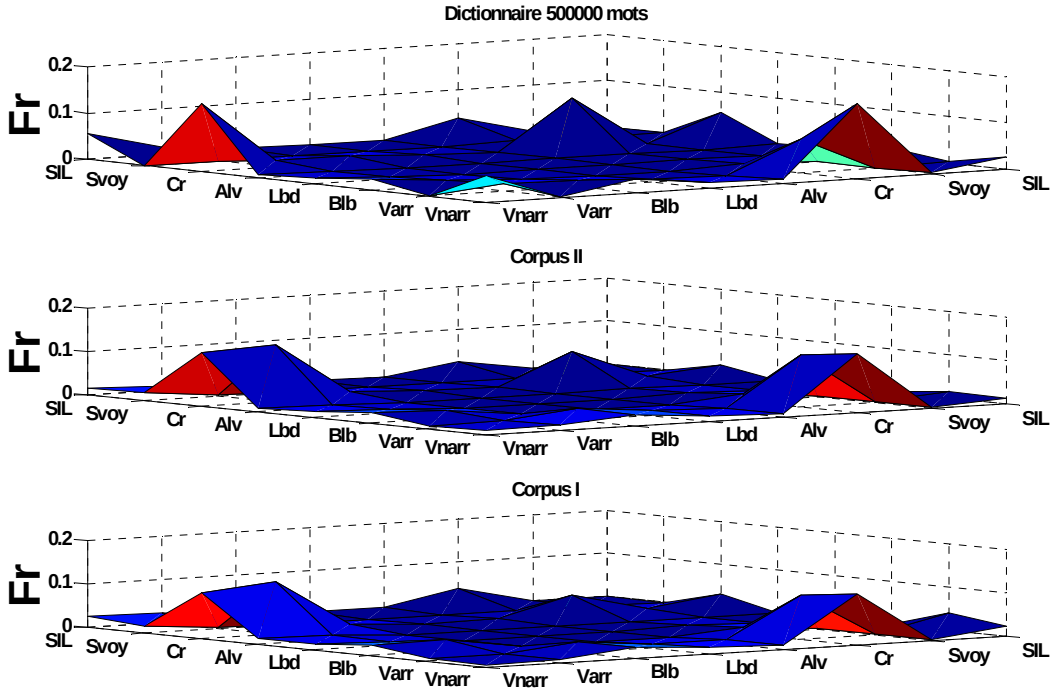


FIGURE 32: FREQUENCE D'APPARITION DES DIVISEMES DES DIFFERENTS CORPUS.

2.3.2. ACQUISITION DES DONNEES

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 plongeante et une camera de profile)

Figure 33, 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, du sexe féminin et agée de 35 ans. 155 marqueurs sont disposés sur le visage et le cou de la locutrice.

FIGURE 33 : EXEMPLES DES CAPTURES DES TROIS CAMERAS UTILISES 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. L'audio est enregistré à la fréquence 32kHz et 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 enregistrées en synchronie.

2.3.3. EXTRACTION DES PARAMETRES VISUELS

Le même principe de modélisation statistique et d'extraction des paramètres visuels est utilisé que pour le corpus I, voir la section 2.2.5. 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 2.2.5. Les séquences des paramètres visuels sont sur-échantillonnées pour obtenir des séquences à 100 Hz.

2.4. RESUME

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 dégrées de liberté issues 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.

3. SYNTHÈSE PAR TDA (*TASK DYNAMICS FOR ANIMATION*). ASPECT CONFIGURATIONNEL

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

3.1. GROUPEMENT DES PHONEMES EN CLASSES DES VISEMES

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 (Équation 1) est souvent utilisée pour calculer la distance entre deux distributions gaussiennes (Mak and 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 sont calculées, les dendrogrammes pour les consonnes et les voyelles sont construits. Ici, les phonèmes qui sont proches visuellement sont appelés visèmes. Sur les Figure 34 et Figure 65 (en Annexe A) les dendrogrammes correspondants aux consonnes et aux voyelles sont représentés pour les paramètres articulatoires et les paramètres géométriques pour les deux corpus.

$$d_{Bhattacharyya} = \frac{1}{8}(\mu_1 - \mu_2)^t \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_1 - \mu_2) + \frac{1}{2}\ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{\left(|\Sigma_1|^{\frac{1}{2}} |\Sigma_2|^{\frac{1}{2}}\right)}$$

ÉQUATION 1 : DISTANCE DE BHATTACHARYYA.

Ainsi, les consonnes peuvent être groupées en quatre groupes de visèmes:

- Les bilabiales: p, b, m,
- Les labiodentales: f, v
- Les post-alvéolaires: □, □
- Les consonnes restantes: t, d, n, s, z, k, g, l,

Les voyelles peuvent être groupées en trois groupes de visèmes:

- Les voyelles ouvertes: a, e, i, □, j, œ̃, □̃

- Les voyelles mi-ouvertes : □, ã
- Les voyelles fermées et arrondies: o, œ, ø, u, y, □̃, et □, w

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

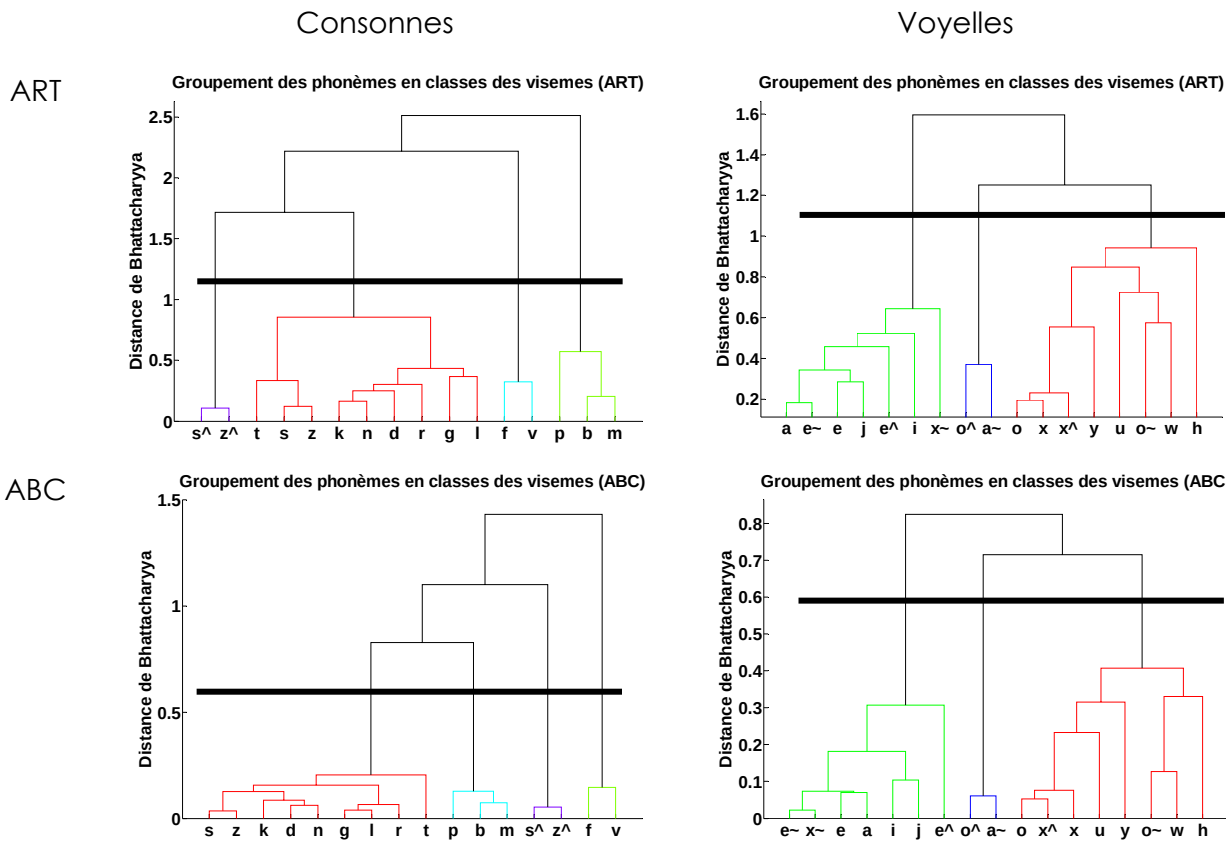


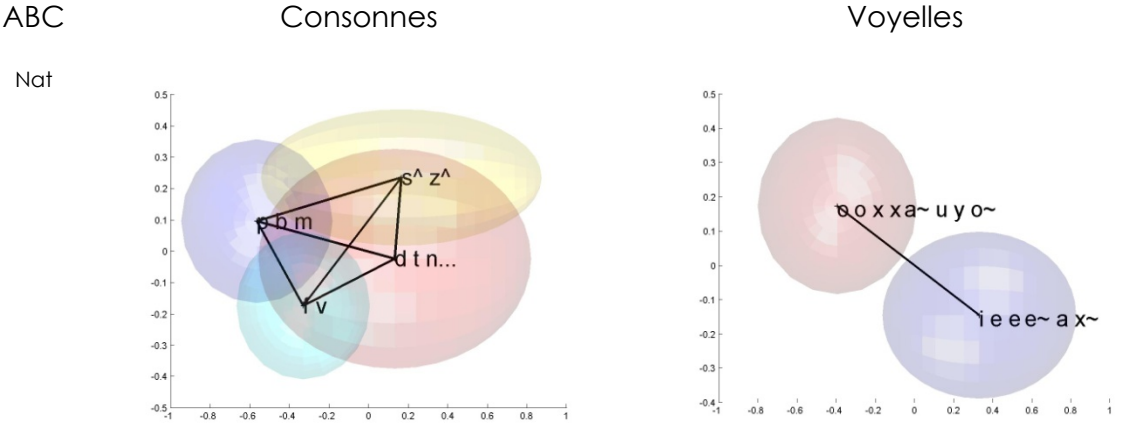
FIGURE 34: CORPUS I. GROUPEMENT DES CONSONNES ET VOYELLES EN CLASSES DES VISEMES GRACE A LA DISTANCE DE BHATTACHARYYA POUR LES PARAMETRES ARTICULATOIRES ET GEOMETRIQUES. LE TRAIT HORIZONTAL GRAS FIGURE LE SEUIL CHOISI POUR DETERMINER LES CLASSES DE VISEMES.

3.2. REALISATION DES CIBLES PAR LA SYNTHESE PAR HMM ET PAR CONCATENATION

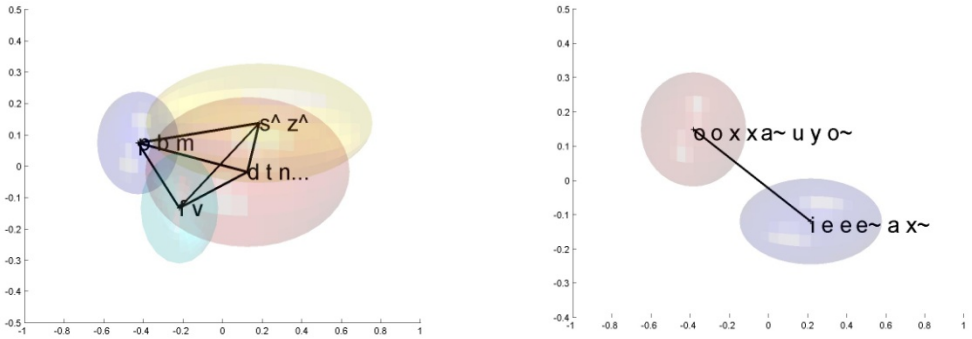
Une fois les réalisations visuelles des phonèmes sont classées en visèmes, l'objectif est d'étudier les réalisations des visèmes pour les différentes méthodes de synthèse. Pour ce faire, la corrélation linéaire et l'analyse discriminante ADL (Analyse Discriminante Linéaire) sont utilisées. L'ADL permet de réduire l'espace de représentation et de proposer une représentation graphique qui permet de visualiser les proximités entre les observations. C'est pour cette raison que l'ADL est utilisée pour analyser la réalisation des cibles car d'une part cela permettrait d'étudier la séparation des visèmes grâce au coefficient ADL (rapport entre inter et intra distances) et d'autre part cela permettrait de visualiser la réalisation spatiale des visèmes grâce à la projection des données sur le premier plan discriminant. Tout d'abord le modèle ADL est calculé pour les données d'origine (Nat), ensuite,

les données d'origine et les données de synthèse par HMM (HMM) et par concaténation (Conc) sont projetées sur le premier plan discriminant. Les ellipses de dispersion des cibles des visèmes selon deux premiers paramètres ADL pour les différentes méthodes de synthèse sont représentées dans les Figure 35 et Figure 66 (en Annexe A) pour le corpus I et dans les Figure 67 (en Annexe A) et Figure 68 (en Annexe A) 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 Figure 36 et Figure 69 (en Annexe A) le coefficient ADL est plus grand dans le cas de synthèse par HMM que par la synthèse par concaténation, c'est-à-dire qu'il y a une meilleure séparation des cibles des visèmes dans la synthèse par HMM que par concaténation. 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) mais ses cibles sont prototypiques, pas assez coarticulées, alors que, la concaténation 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).

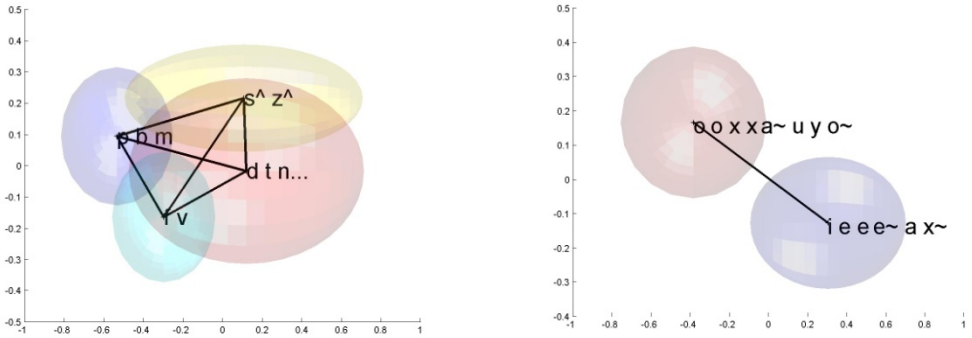
Les coefficients de corrélation linéaire sont calculés pour les synthèses par concaténation et par HMM dans les différents espaces des paramètres visuels, Figure 37. L'espace ART_abc correspond aux paramètres géométriques calculés à partir des paramètres articulatoires déjà synthétisés. La synthèse HMM est mieux réalisée dans l'espace des paramètres géométriques que dans l'espace des paramètres articulatoires. La synthèse par HMM a la corrélation plus grande que la synthèse par concaténation. Dans la Figure 38 un exemple de la synthèse de la phrase "Celui qui joue" est représenté. Dans cet exemple (voir la partie soulignée / sel<u>wiki</u>□u /), l'amplitude de la synthèse par concaténation est très proche du naturel mais son timing est plus décalé de la cible que celui de la synthèse par HMM par contre, le timing de HMM est bon mais son amplitude est lissée.



HMM context e viseme droit



Concat énation



TDA

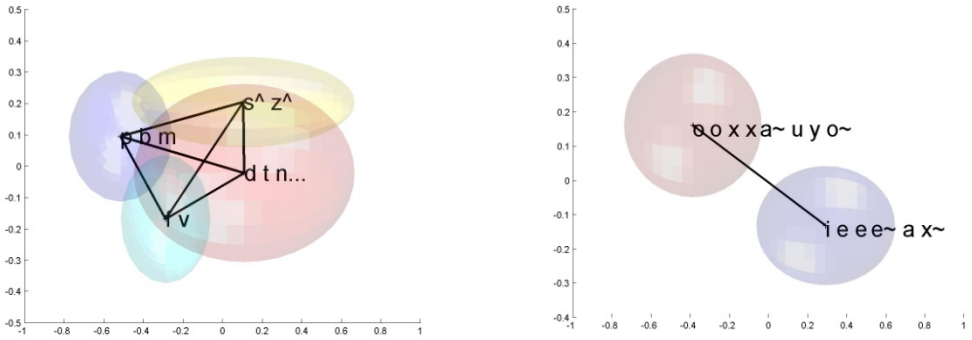


FIGURE 35: ELLIPSES DE DISPERSION DES CIBLES GEOMETRIQUES POUR LES PRINCIPALES CLASSES DES CONSONNES ET DES VOYELLES ASELON ADL POUR LES DONNEES NATURELLES, LA SYNTHESE PAR HMM, LA SYNTHESE PAR LA CONCATENATION ET LA SYNTHESE PAR TDA. CORPUS I

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

FIGURE 36: LES CARACTERISTIQUES DE LA ADL (INTER-DISTANCE, INTRA-DISTANCE ET LEUR RAPPORT) DES CONSONNES ET VOYELLES DANS LES ESPACES GEOMETRIQUE ET ARTICULATOIRE POUR LES DONNEES NATURELLES, LA SYNTHESE PAR HMM, LA SYNTHESE PAR LA CONCATENATION ET LA SYNTHESE PAR TDA. CORPUS I. DONNEES 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 VISEMES.

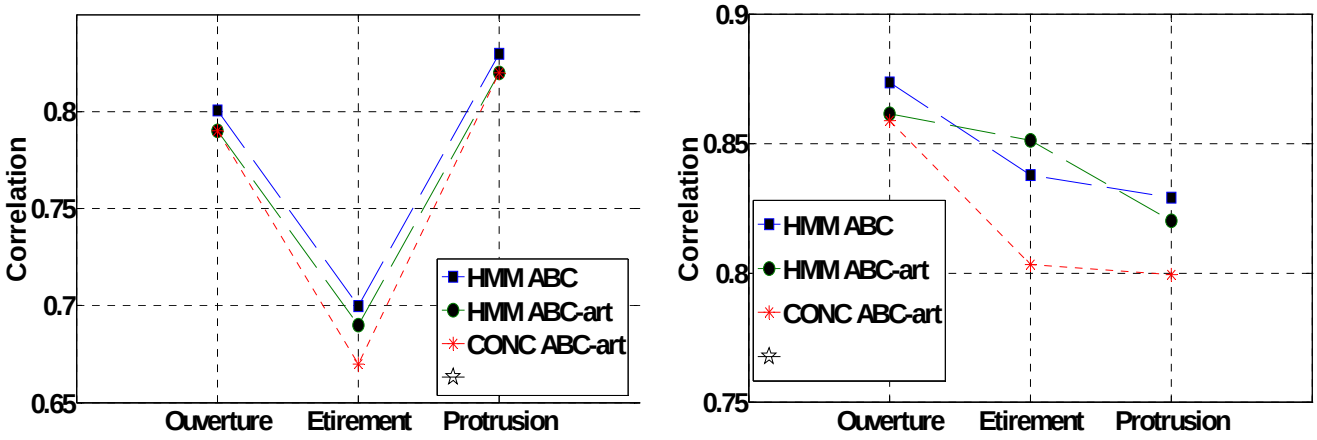


FIGURE 37: LES COEFFICIENTS DE CORRELATION DES SYNTHESES PAR HMM ET PAR CONCATENATION DANS L'ESPACE GEOMETRIQUE. CORPUS I (GAUCHE) ET II (DROIT). DONNEES D'APPRENTISSAGE ET DE TEST.

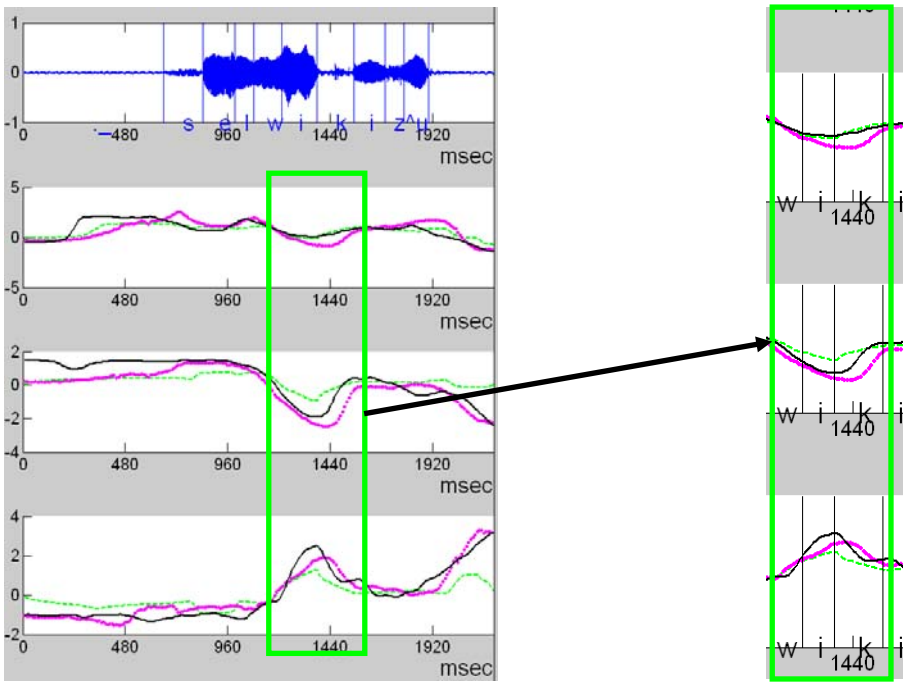


FIGURE 38: LES TRAJECTOIRES DES PARAMETRES GEOMETRIQUES POUR LES SYNTHESES PAR HMM EN VERT ET PAR CONCATENATION EN ROSE, NATUREL EST EN NOIR. PHRASE "CELUI QUI JOUE". A SOULIGNER, LES TRAJECTOIRES MOINS ARTICULEES POUR LA SYNTHESE PAR HMM ET LES TRAJECTOIRES PLUS ARTICULEES MAIS LE TIMING DECALE POUR LA SYNTHESE PAR CONCATENATION.

3.3. SYNTHESE PAR TDA

3.3.1. 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 chaine 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 and Goldstein 1990), (Browman and Goldstein 1989), les constellations articulatoires étant donc fortement planifiées (Whalen 1990). La théorie de la phonologie articulatoire (Browman and Goldstein 1990), (Browman and Goldstein 1989) 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 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.

Cette théorie postule qu'il existe des invariants gestuels dans l'espace des primitives phonologiques malgré la variabilité surfacique des gestes articulatoires (voir la Figure 39a). La réalisation du phonème /b/ est donnée comme exemple de l'application de la théorie (cf. Figure 39b) : 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.

Une simulation numérique *Task Dynamics* de la théorie de la phonologie articulatoire est proposée par des chercheurs de *Haskins Laboratories* (Saltzman and 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.

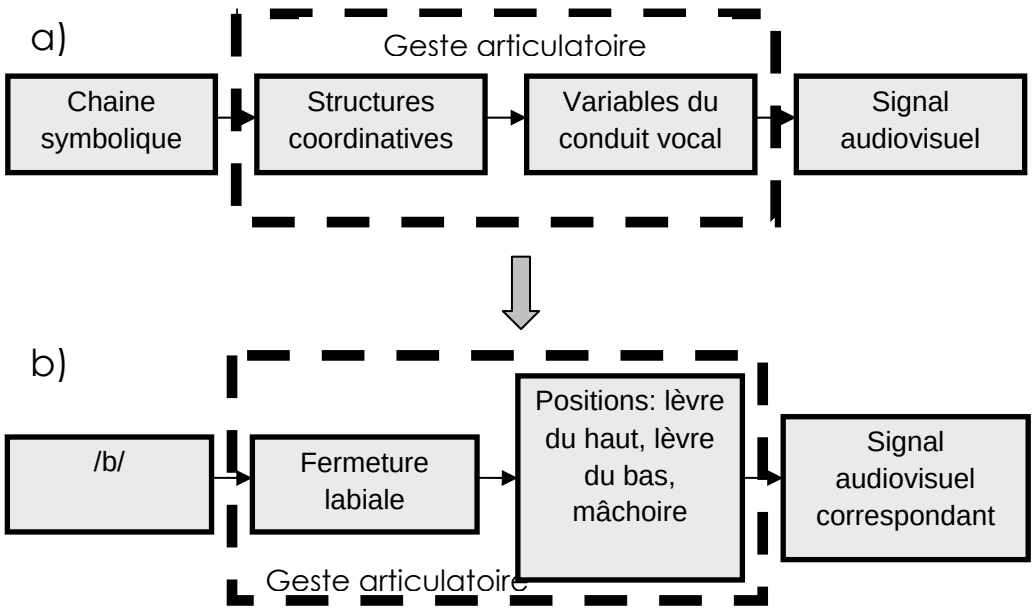


FIGURE 39: A) PRODUCTION DE LA PAROLE SELON LA THEORIE DE LA PHONOLOGIE ARTICULATOIRE; B) EXEMPLE DE LA PRODUCTION DE LA PAROLE SELON LA THEORIE DE LA PHONOLOGIE ARTICULATOIRE POUR LE PHONEME /B/.

3.3.2.TDA: CONCATENATION GUIDEE 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 étudiée dans la 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 la concaténation dans l'espace articulatoire (paramètres articulatoires statistiques issus de l'ACP guidée) grâce aux partitions géométriques et articulatoires, (voir la Figure 40). Par sa nature, l'espace géométrique correspond à la partie de la planification car les trois paramètres géométriques utilisés correspondent aux principaux mouvements visibles des lèvres. En plus, il est démontré que les cibles articulatoires sont mieux discriminées par HMM dans l'espace géométrique, voir les résultats dans les Figure 36 et Figure 69. 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 3.2) et, en plus, la synthèse par HMM a la capacité de planifier la coarticulation à long terme (voir 3.3.3). La synthèse par concaténation joue le rôle d'exécution des mouvements articulatoires. La méthode de concaténation fournit de l'articulation détaillée, voir 3.2. La synthèse par concaténation guidée HMM devrait produire des mouvements plus proches des trajectoires naturelles et mieux

agencés que la concaténation simple car les couts de sélection (partitions géométriques) sont utilisés en plus des couts de concaténation.

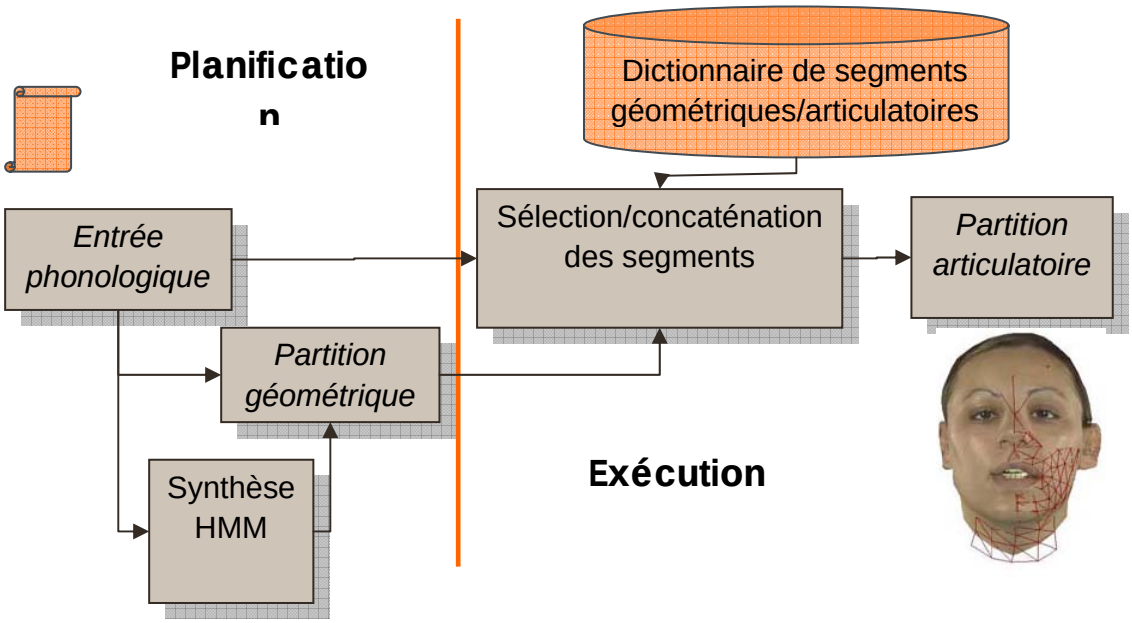


FIGURE 40: SCHEMA DU SYSTEME DE LA SYNTHESE PAR TDA: PLANIFICATION PAR HMM DANS L'ESPACE GEOMETRIQUE ET EXECUTION PAR CONCATENATION DANS L'ESPACE ARTICULATOIRE.

3.3.3. PLANIFICATION PAR HMM

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

ANALYSE PAR RAPPORT A LA STRUCTURE MATHEMATIQUE DES HMM

Dans un premier temps, un modèle HMM est appris pour chaque monophone (phonème sans contexte). L'erreur moyenne et la corrélation entre les trajectoires de synthèse et celles d'origine sont calculées en fonction du nombre d'états dans un HMM. Notons que le graphe de transition est imposé : nous utilisons des chaines d'ordre 1. Pour le corpus I la distorsion minimale est obtenue avec quatre états (pas de différence significative à partir de quatre états, $p \leq 0.05$[6], cf.

[6] « 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 a). 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.

Figure 41). Pour le corpus II la distorsion minimale est obtenue avec cinq états ($p \leq 0.05$). 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 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 42. 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 grâce l'analyse ACP (Analyse en Composantes Principales) à partir de 20 paramètres MFFC (*Mel Frequency Cepstral Coefficients*) sont ajoutés dans le vecteur d'apprentissage la distorsion n'est pas diminuée. Lorsque les gaussiennes des modèles HMM sont représentées comme 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.

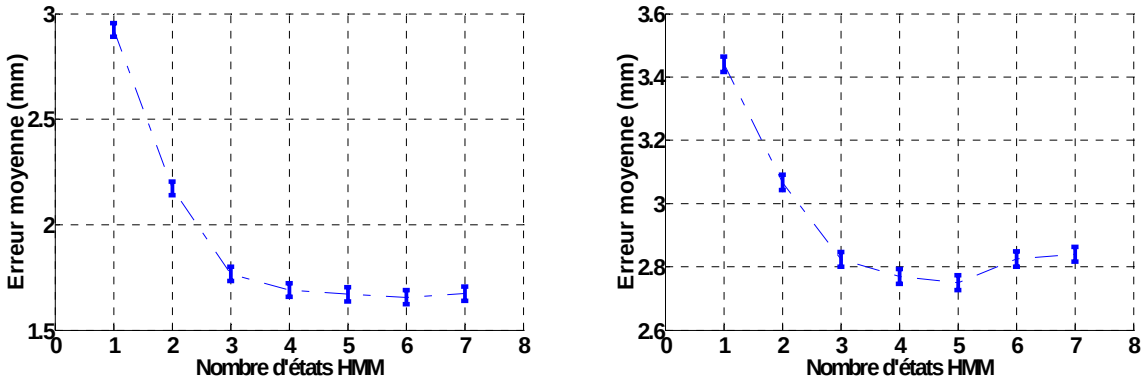


FIGURE 41: L'ERREUR MOYENNE (MM) ENTRE LES TRAJECTOIRES DE SYNTHESE ET ORIGINALES POUR LES DIFFERENTS NOMBRES D'ETATS HMM. HMM MONOPHONE (HORS CONTEXTE). CORPUS I (GAUCHE) ET II (DROIT). DONNEES D'APPRENTISSAGE ET DE TEST.

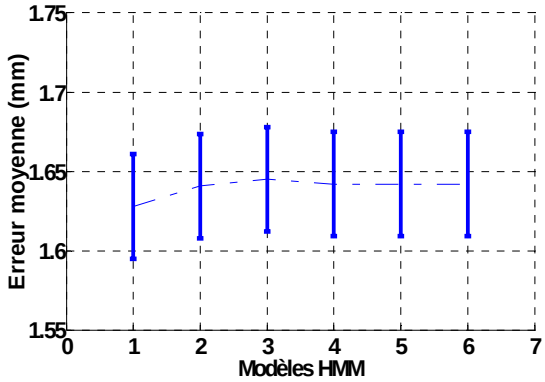


FIGURE 42: L'ERREUR MOYENNE (MM) ENTRE LES TRAJECTOIRES DE SYNTHESE ET ORIGINALES POUR LES DIFFERENTS MODELES HMM, DANS L'ORDRE: 1) HMM PHONEME EN CONTEXTE VISEME DROIT (AVEC LES PARAMETRES DYNAMIQUES DU 1ER ORDRE), 2) HMM PHONEME EN CONTEXTE VISEME DROIT AVEC MELANGE DE GAUSSIENNES D'ORDRE 2, 3) HMM PHONEME EN CONTEXTE VISEME DROIT AVEC MELANGE DE GAUSSIENNES D'ORDRE 4, 4) HMM PHONEME EN CONTEXTE VISEME DROIT AVEC MELANGE DE GAUSSIENNES D'ORDRE 6, 5) HMM PHONEME EN CONTEXTE VISEME DROIT (AVEC LES PARAMETRES DYNAMIQUES DU 1ER ORDRE ET 2EME ORDRE), 6) HMM PHONEME EN CONTEXTE VISEME DROIT AVEC LES PARAMETRES VISUELS ET ACOUSTIQUES. DONNEES D'APPRENTISSAGE ET DE TEST. CORPUS I.

ANALYSE PAR RAPPORT A L'INFORMATION CONTEXTUELLE

L'erreur moyenne est calculée pour les différents modèles HMM en fonction du contexte utilisé, Figure 43 et Figure 44. Plusieurs types de contexte sont étudiés: contexte phonème gauche ou droit, contexte viseme gauche ou droit, contexte gauche et droit et contexte viseme droit avec l'information syllabique. La distorsion est moins importante ($p \leq 0.05$) 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 ($p > 0.05$) 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 visemique ou phonémique. De plus, l'utilisation de l'information visemique permet d'avoir plus de représentants pour apprendre les modèles HMM que l'utilisation de l'information phonémique. Suite à ces résultats on choisit de travailler avec les modèles HMM par phonème en contexte droit viseme.

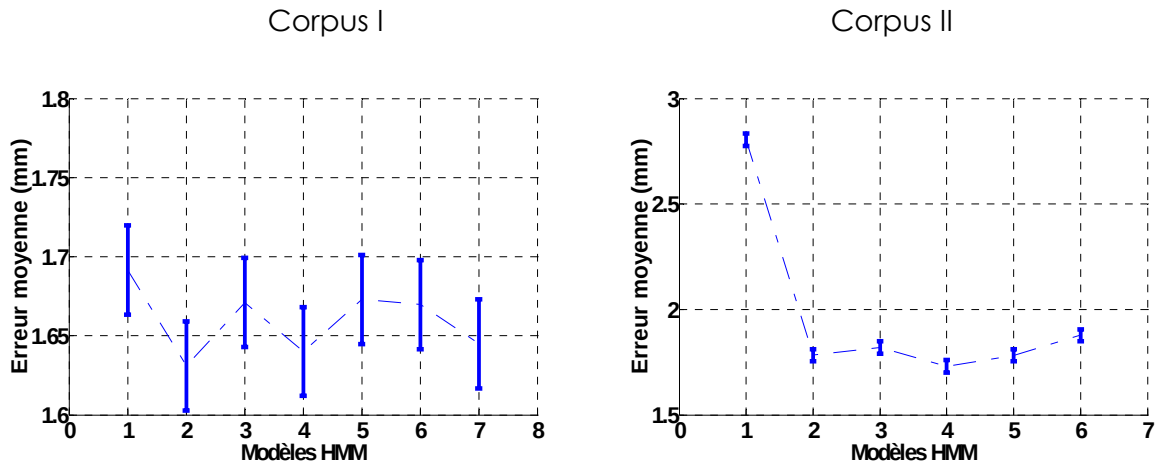


FIGURE 43: L'ERREUR MOYENNE (MM) DE LA SYNTHESE PAR HMM POUR LES DIFFERENTS MODELES: DANS L'ORDRE: 1) PHONEME SANS CONTEXTE, 2) PHONEME CONTEXTE DROIT PHONEME, 3) PHONEME CONTEXTE GAUCHE PHONEME, 4) PHONEME CONTEXTE DROIT VISEME, 5) PHONEME CONTEXTE GAUCHE VISEME, 6) PHONEME CONTEXTE GAUCHE ET DROIT PHONEME ET 7) INFORMATION SYLLABIQUE POUR LE CORPUS I SEULEMENT. CORPUS I ET II. DONNEES D'APPRENTISSAGE.

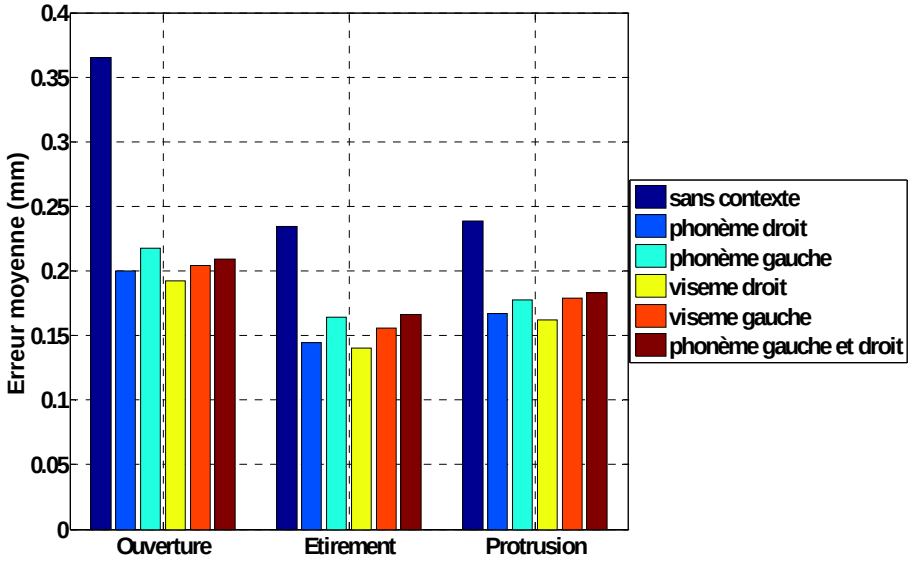


FIGURE 44 : L'ERREUR MOYENNE (MM) DE LA SYNTHESE PAR HMM POUR LES DIFFERENTS MODELES ET POUR LES DIFFERENTS PARAMETRES: DANS L'ORDRE: 1) PHONEME SANS CONTEXTE, 2) PHONEME CONTEXTE DROIT PHONEME, 3) PHONEME CONTEXTE GAUCHE PHONEME, 4) PHONEME CONTEXTE DROIT VISEME, 5) PHONEME CONTEXTE GAUCHE VISEME, 6) PHONEME CONTEXTE GAUCHE ET DROIT. CORPUS II. DONNEES D'APPRENTISSAGE.

3.3.4. EXECUTION PAR CONCATENATION

L'objectif, ici, est de sélectionner dans le dictionnaire les meilleurs (au sens du 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 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 simple 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 45. Cette

procédure compense les sauts aux frontières inter-diphones: une interpolation linéaire est calculée sur le saut observé durant le diphone précédent.

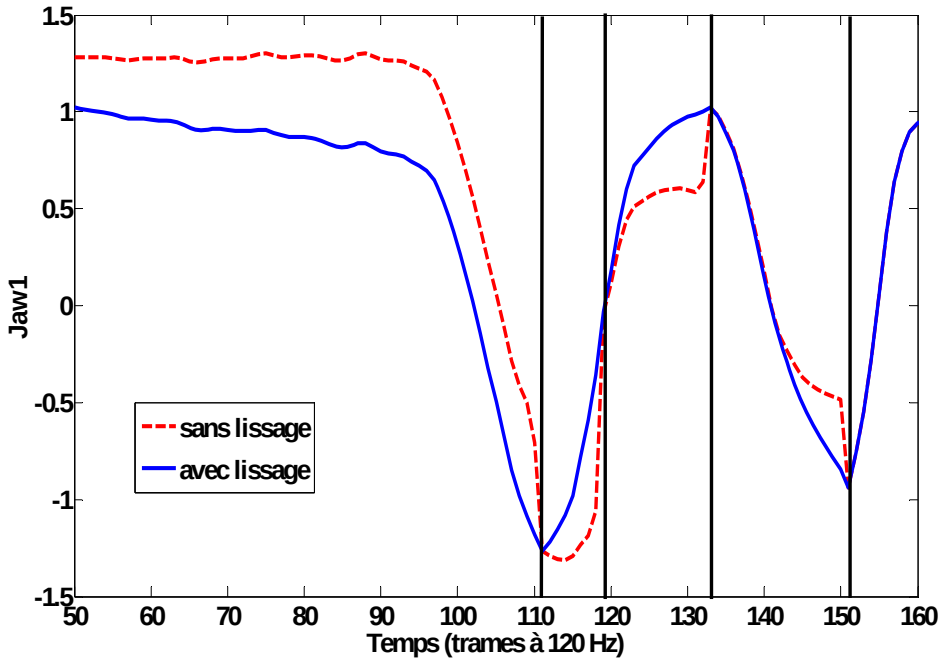


FIGURE 45: L'EXEMPLE DU LISSAGE ANTICIPATOIRE POUR LE PARAMETRE JAW1.

3.4. RESULTATS

Les résultats de synthèse par TDA sont présentés dans les Figure 46, Figure 47 et Figure 48. La distorsion est moins importante dans le cas de la synthèse par TDA que par concaténation simple pour le corpus I pour tous les paramètres ($p \leq 0.05$) et pour le corpus II pour tous les paramètres ($p \leq 0.05$). Les exemples de génération des trajectoires géométriques sont présentés dans la Figure 49 pour les phrases "Du thon huileux" et "Il 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 35, Figure 66, Figure 67 et Figure 68 du paragraphe 3.2. 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.

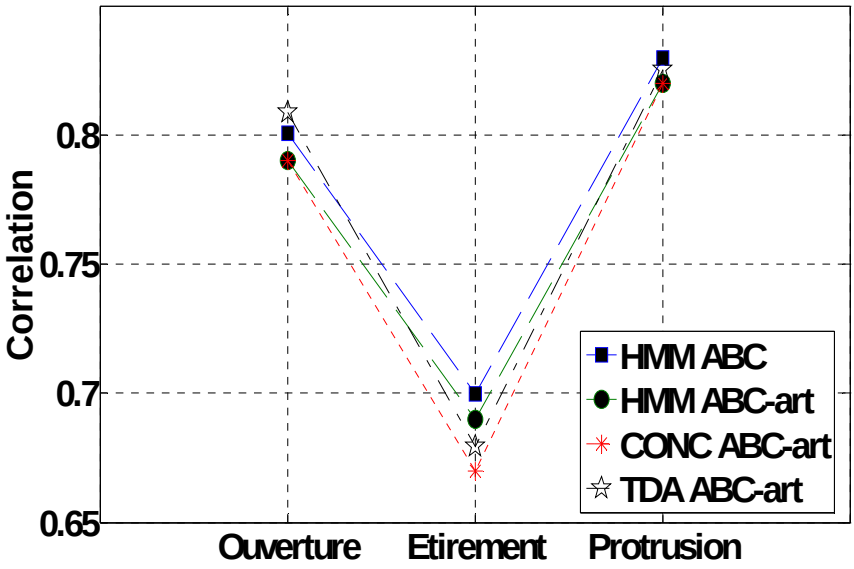


FIGURE 46: LES COEFFICIENTS DE CORRELATION POUR LES SYNTHESES PAR HMM ET PAR CONCATENATION DANS L'ESPACE GEOMETRIQUE. CORPUS I. DONNEES D'APPRENTISSAGE ET DE TEST.

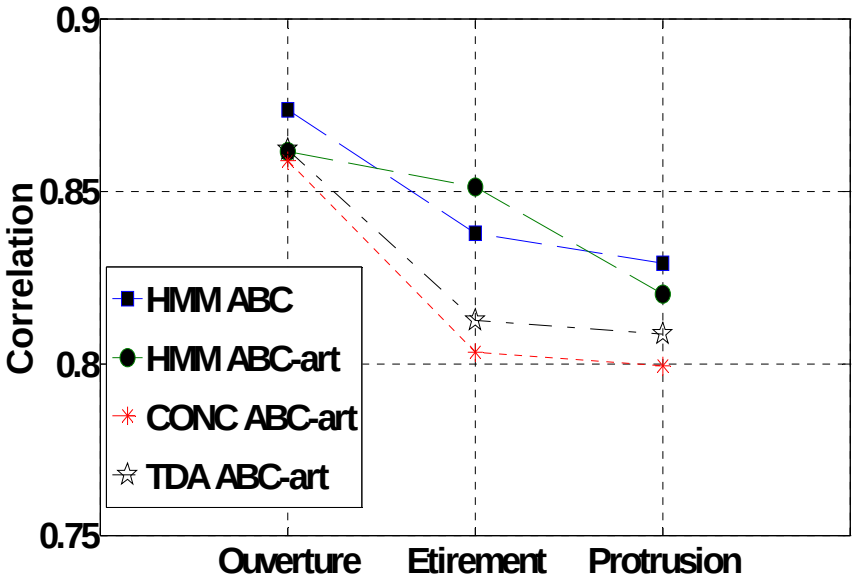


FIGURE 47: LES COEFFICIENTS DE CORRELATION POUR LES SYNTHESES PAR HMM ET PAR CONCATENATION DANS L'ESPACE GEOMETRIQUE. CORPUS II. DONNEES D'APPRENTISSAGE ET DE TEST.

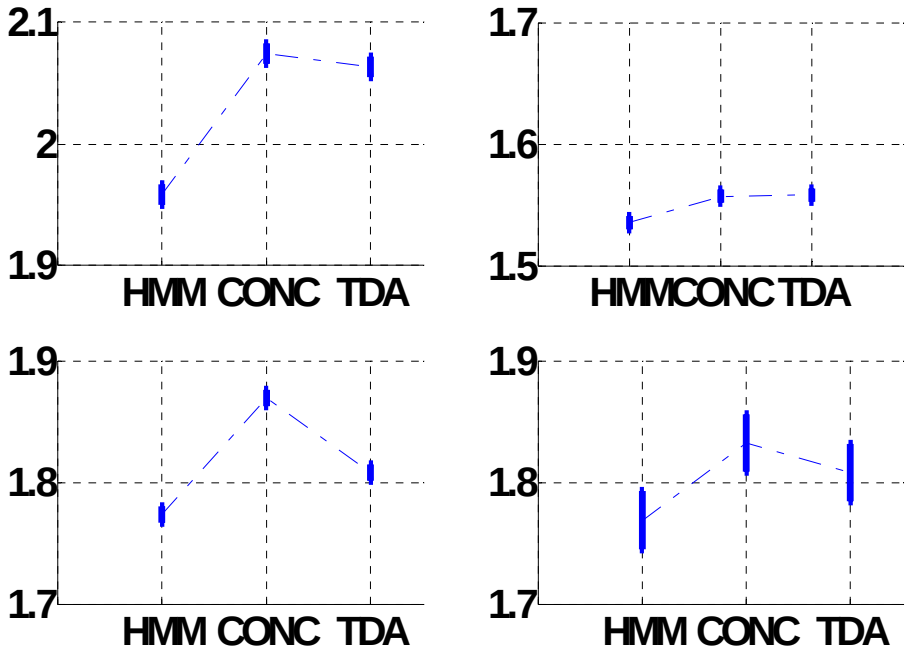


FIGURE 48: L'ERREUR MOYENNE (MM) POUR LES DIFFERENTS TYPES DE SYNTHESE POUR LES PARAMETRES GEOMETRIQUES ET LA MOYENNE DES PARAMETRES POUR LA CONCATENATION, LA TDA ET HMM. CORPUS II. DONNEES D'APPRENTISSAGE ET DE TEST.

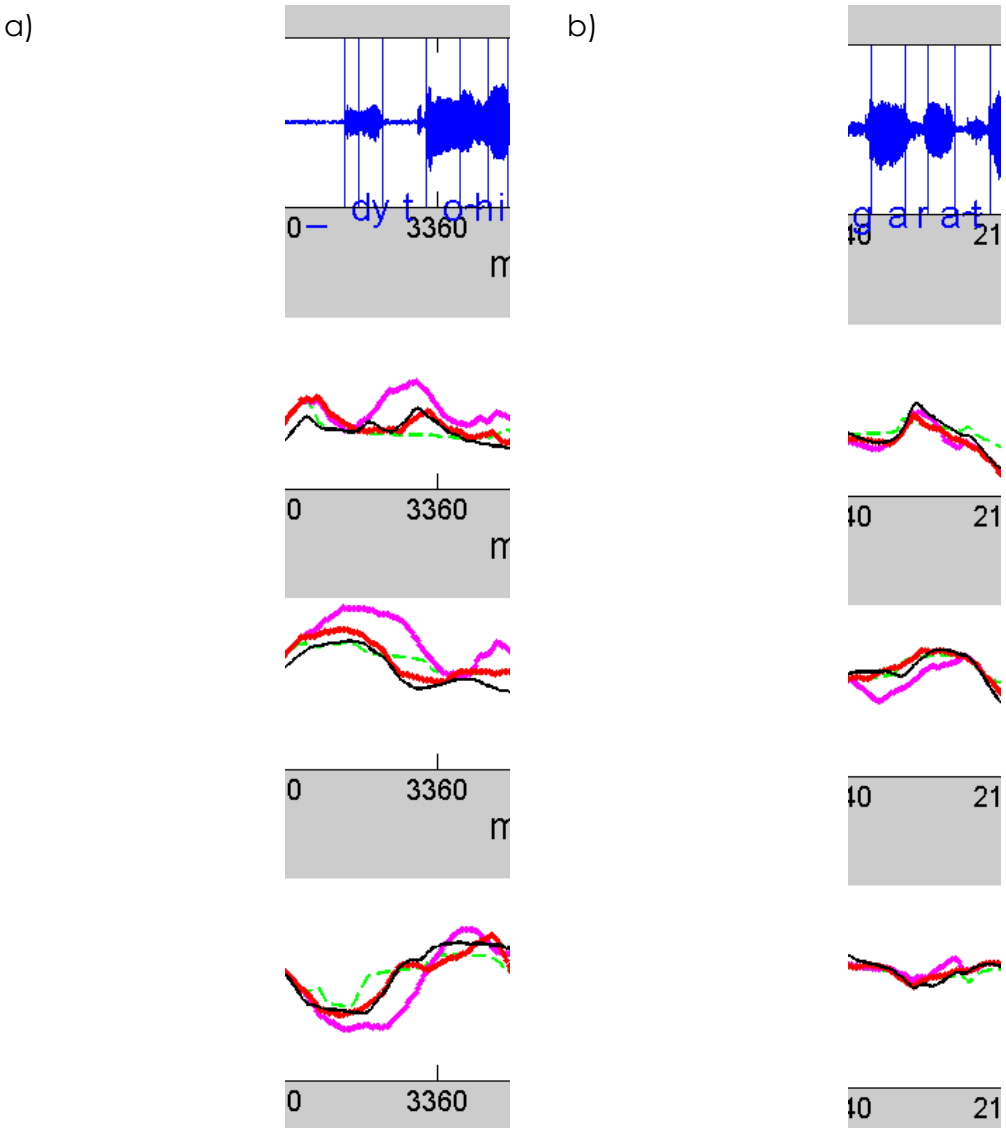


FIGURE 49: LES TRAJECTOIRES GEOMETRIQUES DE SYNTHESE POUR LES SEGMENTS DES PHRASES A) "DU THON HUILEUX" ET B) "IL GARANTIRA". EN NOIRE: DONNEES D'ORIGINE, EN ROUGE: TDA, EN VERT: HMM ET EN MAUVE: CONCATENATION.

3.5. RESUME

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 and

Tokuda 2007)). La synthèse par concaténation fournit une articulation détaillée car elle concatène des segments existants mais cette méthode a 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.

4. SYNTHÈSE PAR PHMM (*PHASED HIDDEN MARKOV MODEL*). ASPECT TEMPOREL

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

4.1. 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 50a. 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, Sullivan 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, Orliaguet et al. 1990), (Perkell and Matthies 1992). Nous sommes particulièrement sensibles aux mouvements labiaux et la non prise en compte de ce phénomène 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 très instables.

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 50b.

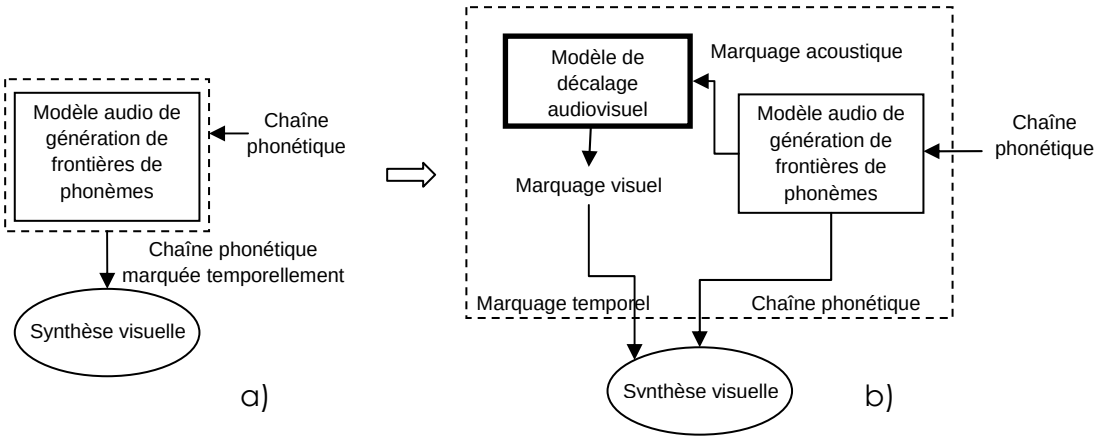


FIGURE 50: PRINCIPE DE GENERATION DES FRONTIERES TEMPORELLES DES PHONEMES A PARTIR D'UNE CHAINE PHONETIQUE POUR LA SYNTHESE AUDIOVISUELLE: A) MODELE DE MARQUAGE DE PHONEMES BASE AUDIO (ETAT DE L'ART EXISTANT); B) MODELE DE MARQUAGE DE PHONEME BASE AUDIO ET VISUEL (ALGORITHME PROPOSE).

4.2. SEGMENTATION TEMPORELLE EN GESTES VISUELS

4.2.1. DESCRIPTION DETAILLEE DE L'ALGORITHME DE REPOSITIONNEMENT DES FRONTIERES DES PHONEMES PAR L'ANALYSE PAR LA SYNTHESE

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 51:

1. La phase d'apprentissage du modèle de décalage des frontières (*off-line*) 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 (*on-line*)

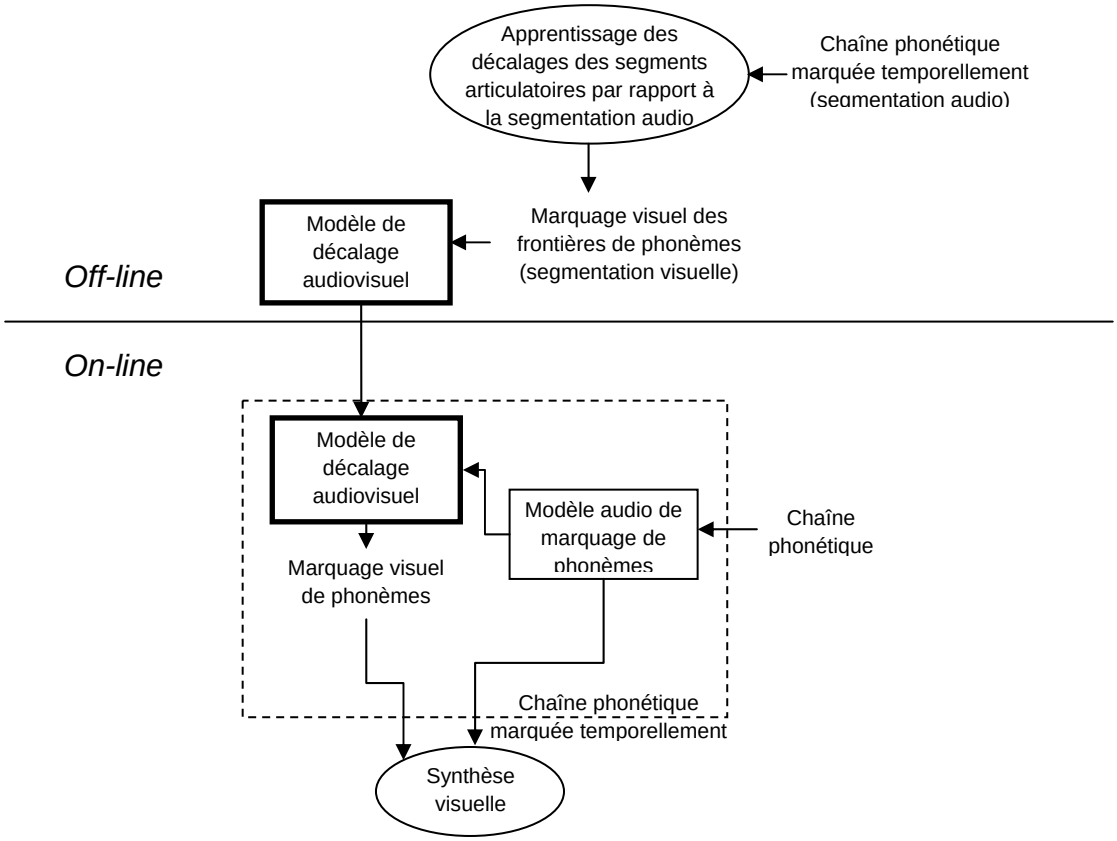


FIGURE 51: SCHEMA GLOBAL DE L'ALGORITHME DE REPOSITIONNEMENT DES FRONTIERES DE PHONEMES POUR LA SYNTHESE AUDIOVISUELLE. *OFF-LINE*: APPRENTISSAGE DU MODELE DE DECALAGE AUDIOVISUEL A PARTIR DE LA SEGMENTATION AUDIO ET DES PARAMETRES VISUELS. *ON-LINE*: UTILISATION DU MODELE DE DECALAGE AUDIOVISUEL DANS LA SYNTHESE AUDIOVISUELLE.

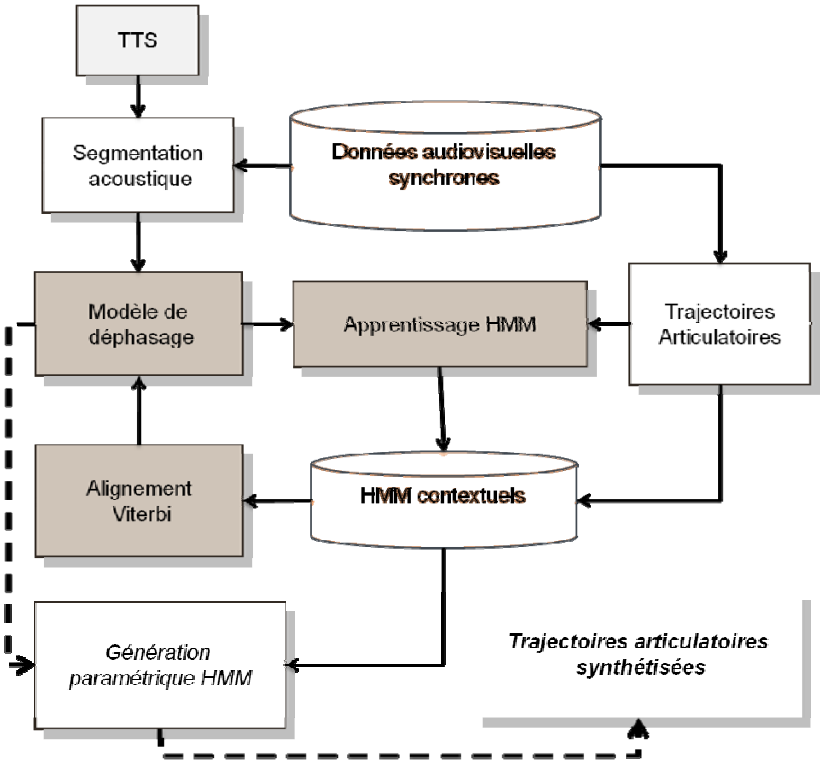


FIGURE 52: SCHEMA GLOBAL DE LA SYNTHESE PAR PHMM.

L'algorithme de repositionnement des frontières des phonèmes pour la synthèse visuelle par HMM (voir la Figure 52) comprend les étapes suivantes, Figure 53:

1. Apprentissage des modèles HMM des phonèmes en contexte viseme suivant sur les paramètres articulatoires à partir de la segmentation audio (SA, itération 0) ou visuelle (SV, itération > 0)
2. Réalignement par Viterbi de ces modèles HMM sur les paramètres articulatoires
3. Calcul des décalages audiovisuels
4. Calcul d'un modèle moyen de décalage par segment
5. Segmentation visuelle (SV) effectuée grâce aux modèles de décalage obtenus précédemment (4). Contrainte: durée minimale d'un phonème doit être au moins de 41 ms pour le corpus II ou de 50 ms pour le corpus I (Nombre d'états dans un HMM (ici 5) multiplié par la durée d'une trame visuelle (ici 8,33 ms ou 10 ms) : Dmin=Nst*Dframe).
    a. S'il y a une stabilisation de la segmentation visuelle entre celle obtenue pendant la boucle courante et celle de la boucle précédente aller à l'étape (6). Critère de stabilisation: coefficient de corrélation entre les frontières de phonèmes 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.

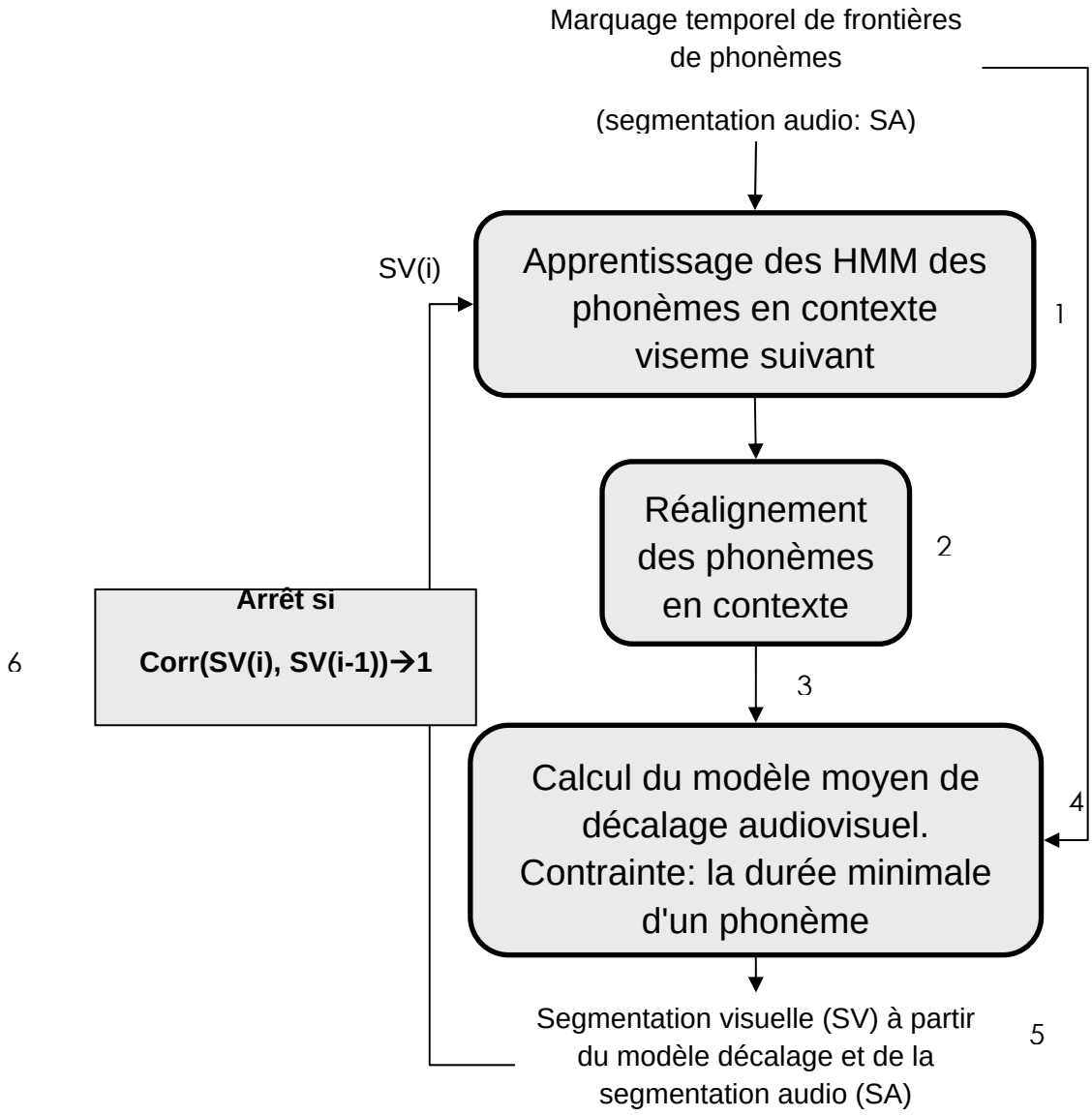


FIGURE 53: EXEMPLE DU PROCEDE D'APPRENTISSAGE DU MODELE DE DECALAGE AUDIOVISUEL BASE HMM.

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.

4.2.2. 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 54 l'erreur moyenne est représentée en

fonction du nombre d'itérations pour la synthèse par PHMM par monophone et par phonème contexte droit viseme. 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 ($p \leq 0.05$). De plus, grâce à l'algorithme proposé, la segmentation automatique en "allophones visuels" est effectuée. Sur la Figure 55 l'augmentation et la diminution des durées des phonèmes à partir de la segmentation visuelle par rapport à leurs durées issues d'une segmentation acoustique est présentée. 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 56. Ici, le geste préphonatoire du [œ̃] et le geste de la fin [o] sont bien présents. Le geste d'arrondissement pour le [□] est bien prédit par les PHMM par rapport à la génération par HMM classiques. Les durées de l'articulation des consonnes [k] et [l] sont réduites.

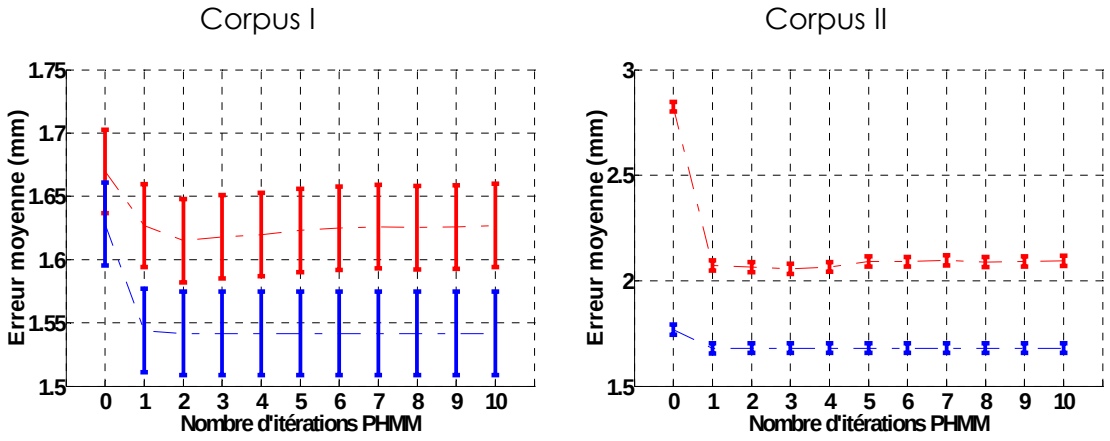


FIGURE 54: ERREUR MOYENNE (MM) (P<0.05) POUR LA SYNTHESE PAR HMM SANS CONTEXTE ET AVEC LE CONTEXTE DROIT VISEME EN FONCTION DU NOMBRE D'ITERATIONS DE L'ALGORITHME DE DECALAGE. CORPUS I (GAUCHE) ET II (DROIT). DONNEES D'APPRENTISSAGE ET DE TEST.

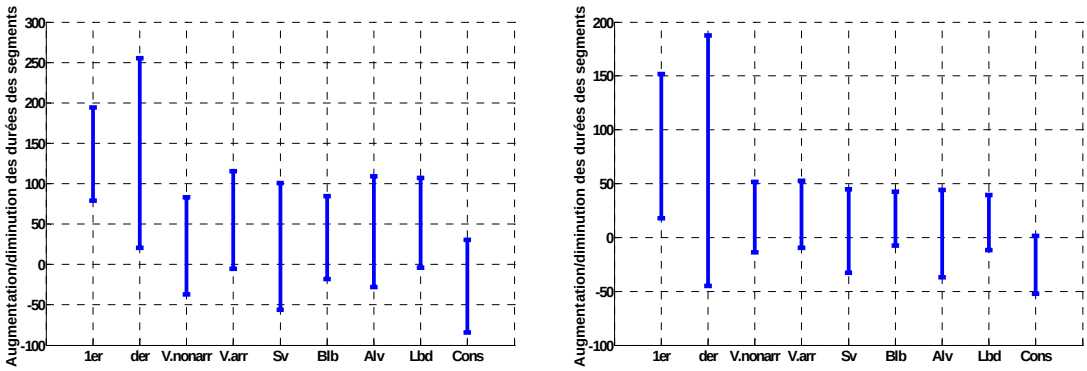


FIGURE 55: L'AUGMENTATION/DIMINUTION DES DUREES DES GESTES ARTICULATOIRES (MS) PAR RAPPORT A LEURS DUREES ACOUSTIQUES. CORPUS I ET II.

4.3. SYNTHESE PAR TDA AVEC LA PLANIFICATION PAR PHMM

La synthèse par TDA avec la planification par PHMM est présentée dans cette section. Grâce à la synthèse par PHMM et la segmentation visuelle trois autres types de synthèse sont obtenues : le TDA avec la planification par PHMM et avec la segmentation acoustique de départ, le TDA avec la planification PHMM et la segmentation visuelle et la concaténation avec la segmentation visuelle. Les phrases des deux corpus sont synthétisées pour tous ces types de synthèse et les résultats sont présentés dans la Figure 57. La synthèse par TDA/PHMM donne de meilleurs résultats que la synthèse par concaténation et la synthèse par TDA/HMM pour les deux corpus, ($p \leq 0.05$). Ce résultat confirme que le système TDA profite de l'étape de planification par HMM/PHMM et sa distorsion est moins importante que dans le cas de la synthèse par concaténation. La différence entre les modèles avec la segmentation phonétique ou visuelle n'est pas significative (TDA/PHMM et TDA/HMM), $p>0.05$.

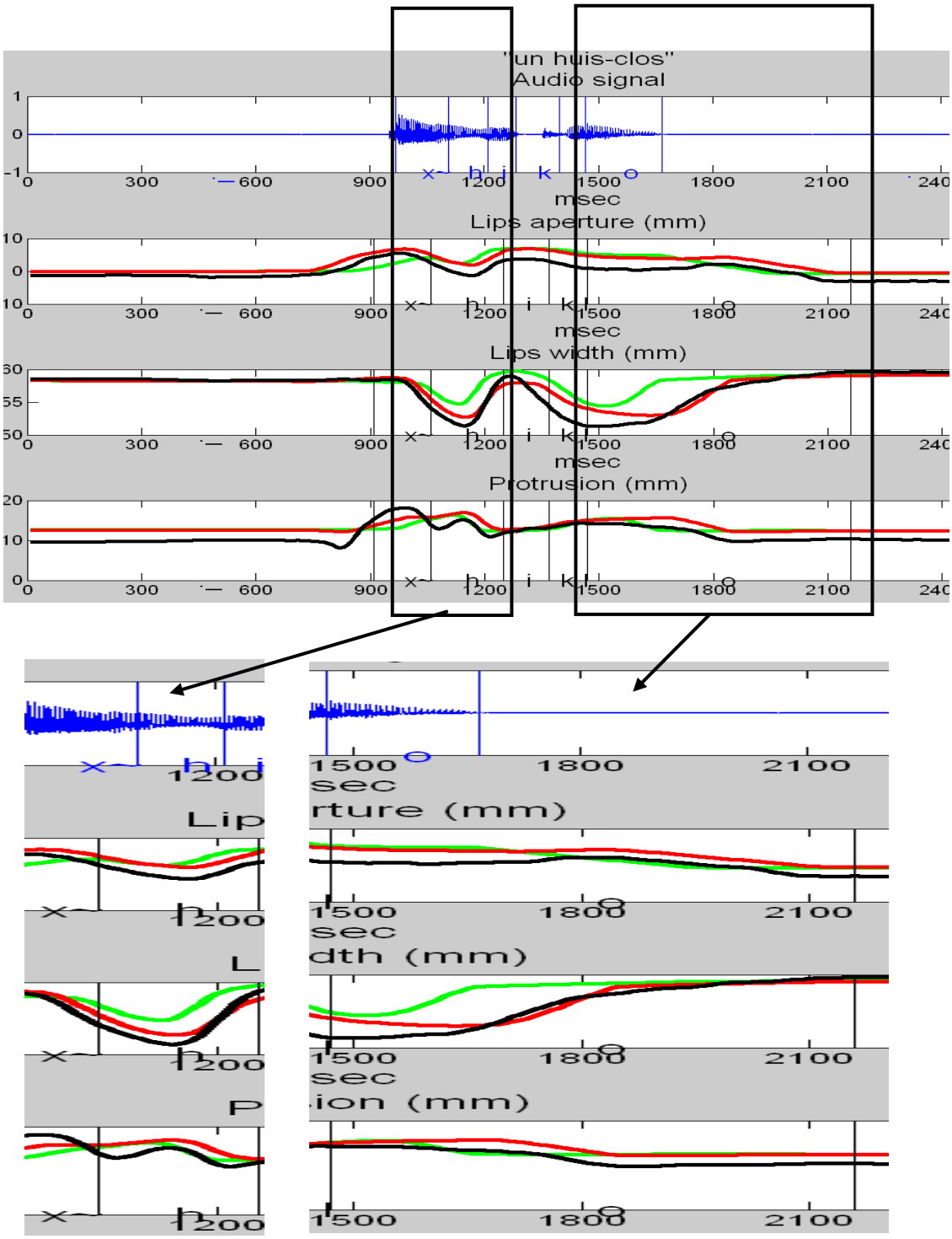
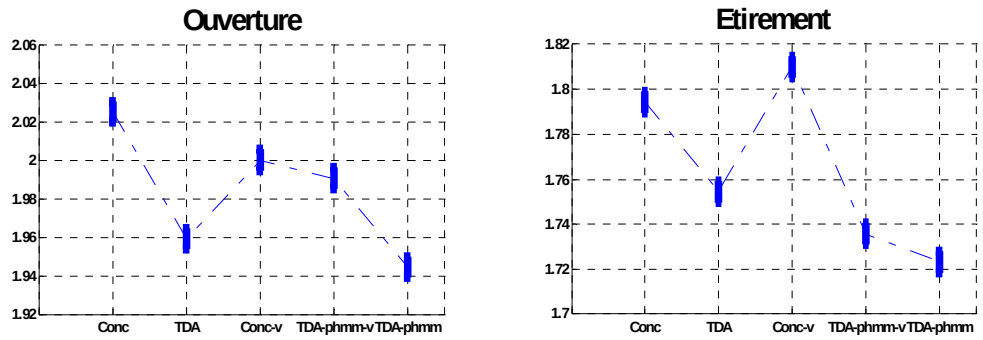
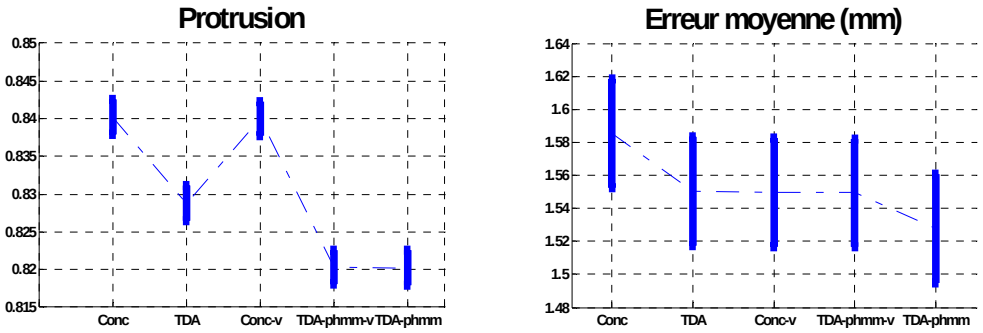


FIGURE 56: L'EXEMPLE E GENERATION DE LA PHRASE "UN HUIS-CLOS". EN NOIR: TRAJECTOIRES D'ORIGINE, EN VERT: HMM ET EN ROUGE: PHMM.

a)

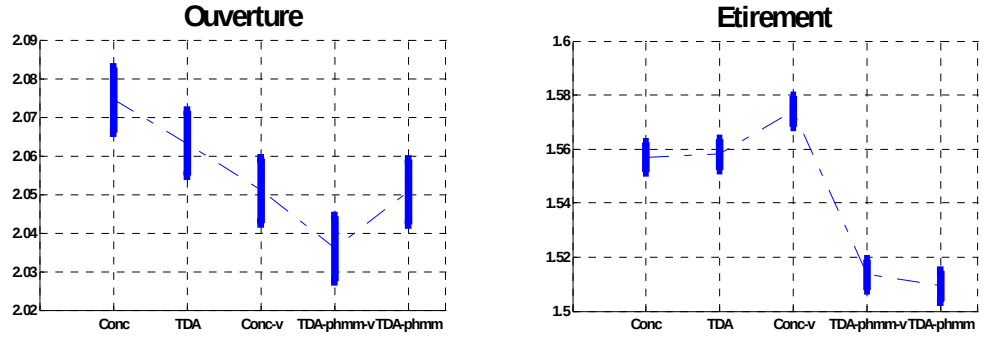
Corpus I





b)

Corpus II



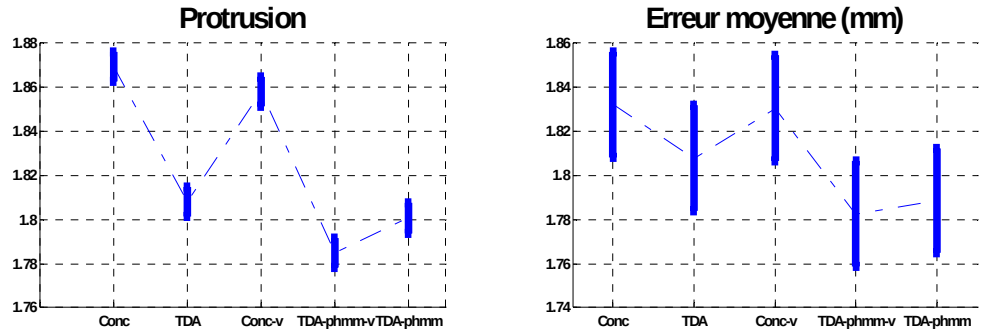


FIGURE 57 : L'ERREUR MOYENNE (MM) POUR LES SYSTEMES DE SYNTHESE DE GAUCHE A DROITE : CONCATENATION, TDA, CONCATENATION AVEC LA SEGMENTATION VISUELLE, TDA AVEC LA PLANIFICATION PHMM ET AVEC LA SEGMENTATION VISUELLE, TDA AVEC LA PLANIFICATION PHMM POUR LE CORPUS I A) ET POUR LE COPRUS II B).

4.4. APPLICATION AU LANGAGE PARLE COMPLETE

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 d'être très important pour la synthèse du LPC. Dans les études précédentes les relations s'organisent autour d'un cadre général moyen où le geste de main se synchronise avec le début de la syllabe acoustique (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).

Dans notre étude il y a plusieurs types de données sur le LPC en entrée:

- Segmentation en gestes LPC
    - Segmentation manuelle en LPC (CONFIG)
    - Segmentation automatique en LPC à partir de la segmentation acoustique (CONFIG_SEG)
- Les paramètres de LPC
    - 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 :

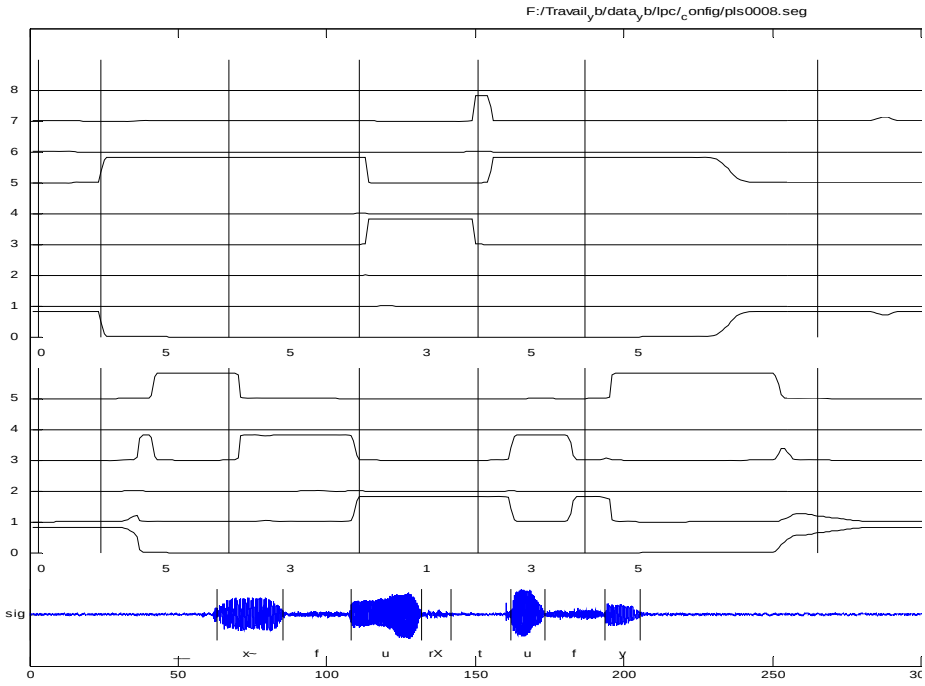
- La segmentation en gestes LPC obtenue grâce à l'algorithme de repositionnement par PHMM (CONFIG_SEG_DEC)
- Les modèles PHMM appris par diclé sur les paramètres LPC (HMM_CONFIG_SEG_DEC)

4.4.1. RECONNAISSANCE DES CIBLES DES GESTES LPC COMME MOYEN D'EVALUATION DE LA SYNTHESE 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 (CONFIG) aux instants

de constriction maximale en utilisant le système d'animation MOTHER de l'ICP (Revéret, Bailly 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 segments LPC sont obtenus et sur les cibles des positions et des formes de la main les modèles gaussiens sont calculés. Les taux de reconnaissance sont 98,36% et 95,89% pour les positions et les formes de la main respectivement, voir les Table 9 et Table 10. Un exemple de ces probabilités au cours du temps sur une phrase du corpus est représenté sur la Figure 58a avec le signal acoustique associé. 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.

a)



b)

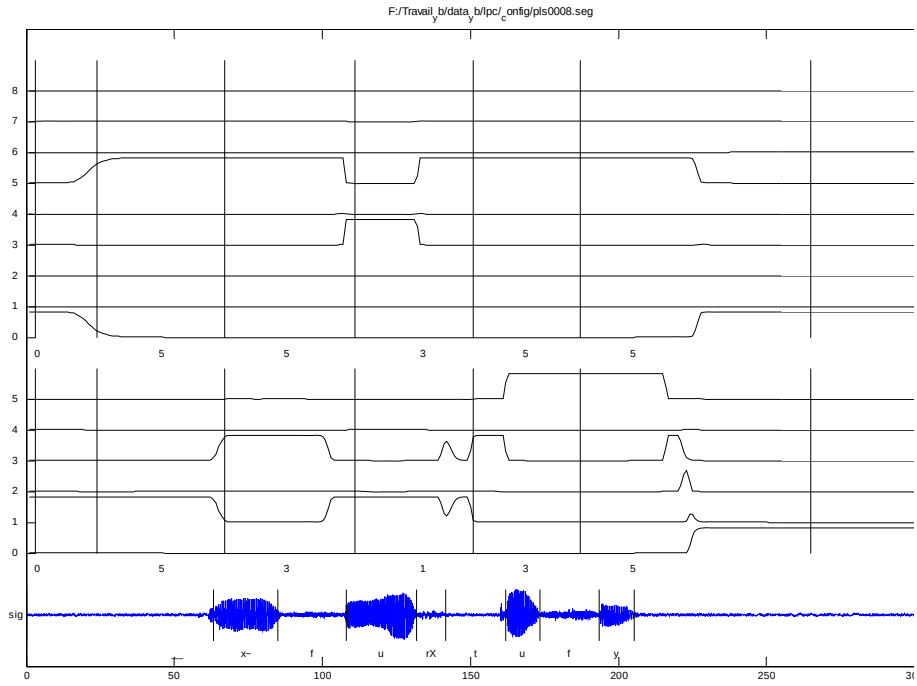


FIGURE 58: VARIATION DES PROBABILITES ISSUES DES MODELES GAUSSIENS POUR LA FORME (HAUT) ET LA POSITION (BAS) DE LA MAIN POUR LA PHRASE "UN FOUR TOUFFU" DU CORPUS II. A) DONNEES D'ORIGINE B) SYNTHESE LPC PAR PHMM. LES CIBLES SONT SUPPOSES ATTEINTES AU MILEU DE CHAQUE SEGMENT GESTUEL.

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

TABLE 9: LES TAUX DE RECONNAISSANCE DES FORMES DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (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

TABLE 10: LES TAUX DE RECONNAISSANCE DES POSITIONS DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (LIGNE DE HAUT).

4.4.2. RESULTATS DE LA SYNTHESE LPC PAR PHMM

Dans un premier temps, les modèles HMM sont appris pour les segments LPC avec la segmentation manuelle CONFIG. Les taux de reconnaissance obtenus avec HMM_CONFIG sont 95,08% et 98,27% pour les positions et les formes de la main respectivement, voir les Table 11 et Table 12. 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 « corriger » les erreurs de codage commises par la codeuse : on voit que les HMM effectuent une modélisation des données qui les nettoient en cohérence avec la partition des données.

Dans un deuxième temps, les modèles HMM sont appris pour les segments LPC avec la segmentation automatique CONFIG_SEG. Les taux de reconnaissance obtenus avec HMM_CONFIG_SEG sont 69,58% et 78,04% pour les positions et les formes de la main respectivement, voir les Table 13 et Table 14.

Enfin, l'algorithme PHMM est appliqué à aux HMM_CONFIG_SEG avec la segmentation automatique. L'algorithme de repositionnement PHMM donne une nouvelle segmentation en segments LPC CONFIG_SEG_DEC et les modèles HMM correspondants HMM_CONFIG_SEG_DEC. Les taux de reconnaissance obtenus avec HMM_CONFIG_SEG_DEC sont 75,81% et 85,91% pour les positions et les formes de la main respectivement, voir les Table 15 et Table 16. Ces taux sont plus grands que dans le cas des HMM_CONFIG_SEG. 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 59. La configuration HMM_CONFIG_SEG_DEC_MIX correspond aux modèles PHMM calculés avec l'initiation avec les HMM_CONFIG mais toujours avec la segmentation automatique CONFIG_SEG, voir les Table 17 et Table 18. Les résultats de synthèse en LPC par PHMM sont résumés dans les Figure 60 et Figure 61. La synthèse par PHMM permet de segmenter automatiquement en gestes LPC et fournit de la synthèse automatique en LPC à partir de la segmentation acoustique et les modèles HMM appris. La segmentation

automatique par PHMM fournit des meilleurs résultats que la segmentation basée seulement acoustique et fournit des les durées des gestes LPC en fonction des durées acoustiques.

- **HMM_CONFIG**

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

TABLE 11: LES TAUX DE RECONNAISSANCE DES FORMES DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (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

TABLE 12: LES TAUX DE RECONNAISSANCE DES POSITIONS DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (LIGNE DE HAUT).

- **HMM_CONFIG_SEG**

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

TABLE 13: LES TAUX DE RECONNAISSANCE DES FORMES DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG_SEG (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (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

TABLE 14: LES TAUX DE RECONNAISSANCE DES POSITIONS DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG_SEG (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (LIGNE DE HAUT).

- **HMM_CONFIG_SEG_DEC**

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

TABLE 15: LES TAUX DE RECONNAISSANCE DES FORMES DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG_SEG_DEC (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (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

TABLE 16: LES TAUX DE RECONNAISSANCE DES POSITIONS DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG_SEG_DEC (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (LIGNE DE HAUT).

- **HMM_CONFIG_SEG_DEC_MIX**

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

TABLE 17: LES TAUX DE RECONNAISSANCE DES FORMES DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG_SEG_DEC_MIX (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (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

TABLE 18: LES TAUX DE RECONNAISSANCE DES POSITIONS DE MAIN. POUR UNE CONFIGURATION SEGMENTEE CONFIG_SEG_DEC_MIX (COLONNE DE GAUCHE), LE NOMBRE DE REPRESENTANTS RECONNUS PAR CONFIGURATION EST REPRESENTE (LIGNE DE HAUT).

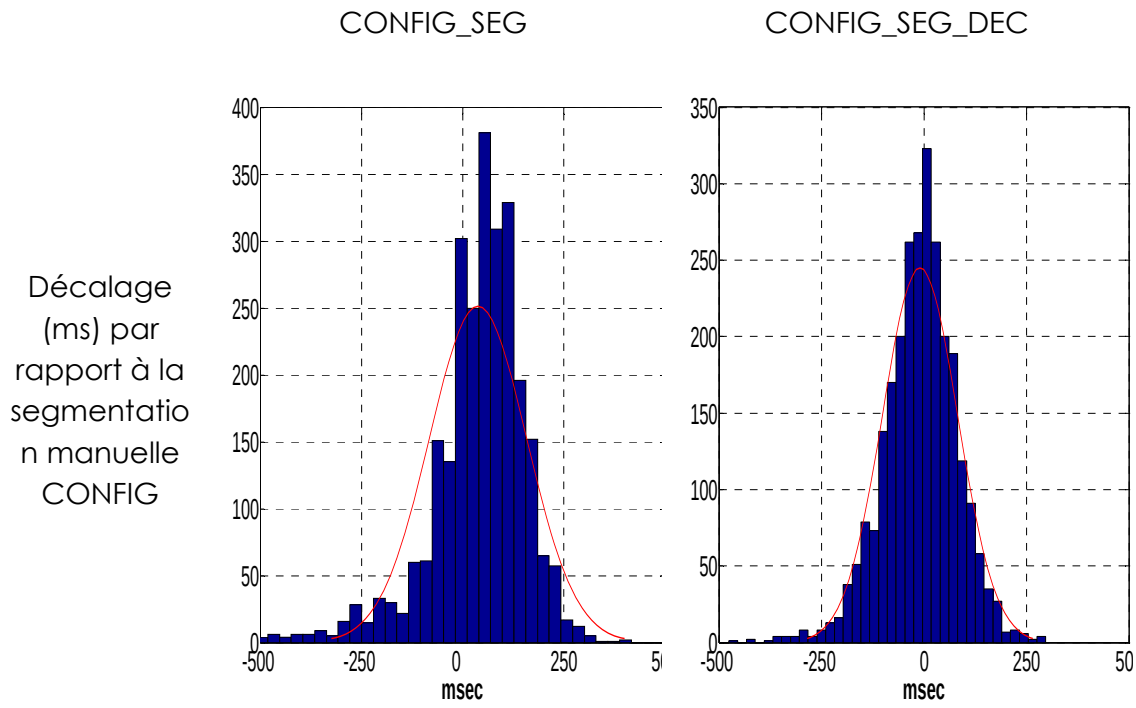


FIGURE 59: HISTOGRAMMES DE DECALAGE DES FRONTIERES DES GESTES LPC CONFIG_SEG ET CONFIG_SEG_DEC PAR RAPPORT A LA SEGMENTATION MANUELLE CONFIG. ON VOIT QUE LE SYSTEME PHMM REMET EN PHASE LES PREDICTIONS AVEC LES GESTES ETIQUETES A LA MAIN.

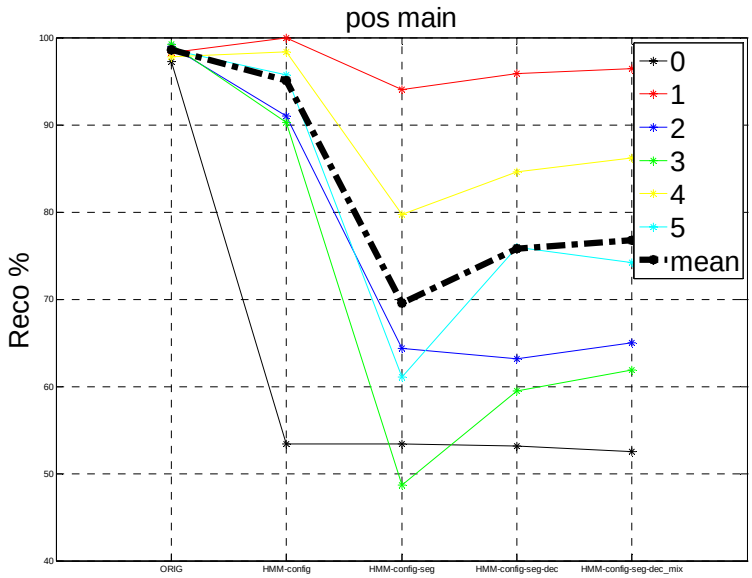


FIGURE 60: LES TAUX DE RECONNAISSANCE DES POSITIONS DE LA MAIN POUR LES DIFFERENTS MODELES ET LES DIFFERENTES SEGMENTATIONS. DE GAUCHE A DROITE: DONNEES D'ORIGINE, HMM_CONFIG, HMM_CONFIG_SEG, HMM_CONFIG_SEG_DEC, HMM_CONFIG_SEG_DEC_MIX.

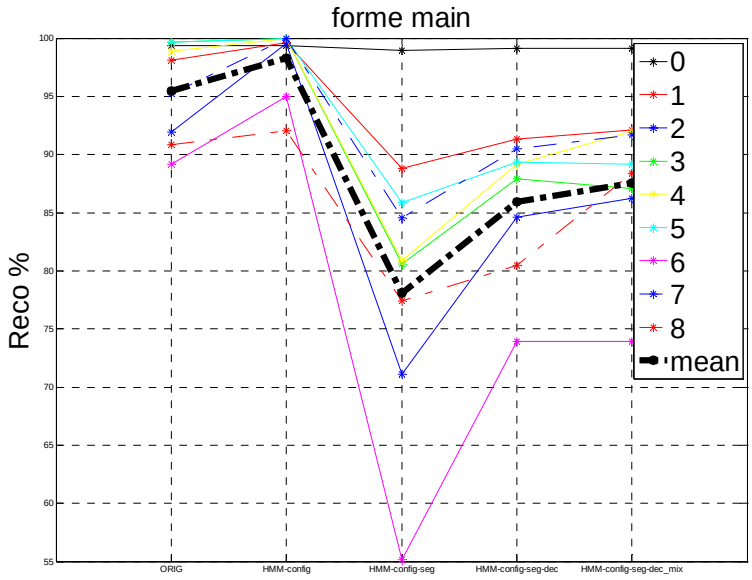


FIGURE 61: LES TAUX DE RECONNAISSANCE DES FORMES DE LA MAIN POUR LES DIFFERENTS MODELES ET LES DIFFERENTES SEGMENTATIONS. DE GAUCHE A DROITE: DONNEES D'ORIGINE, HMM_CONFIG, HMM_CONFIG_SEG, HMM_CONFIG_SEG_DEC, HMM_CONFIG_SEG_DEC_MIX.

4.5. RESUME

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 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 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.

5. 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.

5.1. MODELE 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 §1.3.1. 1 ci-dessusModèles d'apparence et de forme) des cibles gestuelles des allophones de toutes les phrases originales.

5.2. MODELES DE CONTROLE UTILISES

5 modèles de génération de trajectoires articulatoires ont été évalués subjectivement. Les modèles sont : Naturel, HMM, PHMM, concaténation et TDA. 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 3 et 4.

5.3. DEROULEMENT 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 62. 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 les yeux modélisés.

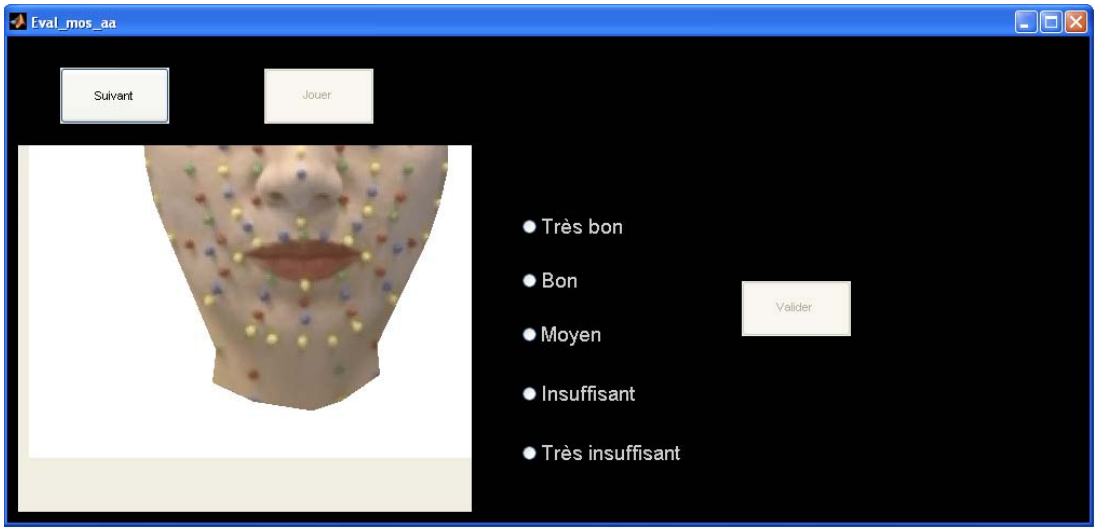


FIGURE 62 : UNE CAPTURE D'ECRAN DE L'INTERFACE DU TEST MOS DE L'EVALUATION SUBJECTIVE DES DIFFERENTS MODELES DE CONTROLE : NAT, HMM, PHMM, CONCATENATION ET TDA.

5.4. RESULTATS

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 63 et la Figure 64. Les résultats sont les suivants : le modèle naturel donne les meilleurs résultats, suivi par les modèles PHMM, TDA, HMM et concaténation. La différence est significative entre le modèle PHMM et concaténation. Les résultats obtenus du test MOS confirment les résultats des tests objectifs, notamment l'intérêt de la gestion du déphasage entre signal acoustique et mouvements faciaux.

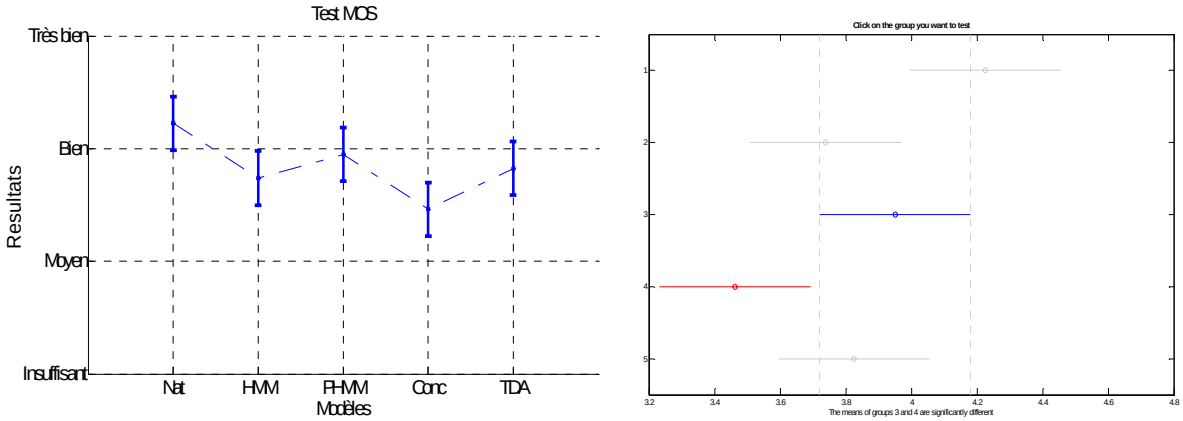


FIGURE 63 : RESULATS DU TEST MOS DU CORPUS II. A GAUCHE : MOYENNES ET ECARTS-TYPES DES NOTES DES SUJETS POUR LES DIFFERENTS MODELES DE GENERATION ; A DROITE : RESULATS DU TEST ANOVA DU TEST MOS.

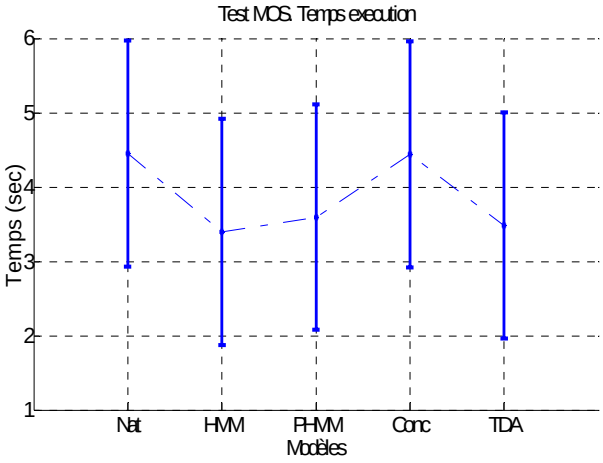


FIGURE 64 : RESULTATS DU TEST MOS DU CORPUS II (TEMPS D'EXECUTION EN SECONDES).

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 la tendance de moyenner, de 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 la théorie de planification et d'exécution issue 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 La 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 ce résultat est confirmé 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 and 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 bilabiaux que les paramètres d'étirement ou de protrusion des lèvres, que le paramètre articulatoire des mouvements de la mâchoire sera important pour les labiodentaux, etc. Ainsi, on pourrait affiner les modèles de synthèse des différents paramètres en fonction du phonème à prononcer (en introduisant, par exemple, les poids de pondération ou autre).

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, …).

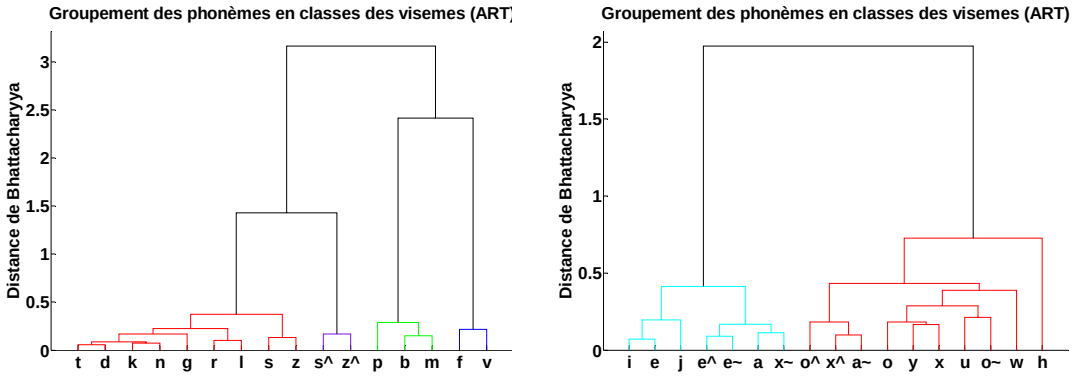
A long terme, le principal axe serait très certainement d'adapter les modèles de synthèse (notamment HMM) aux différents locuteurs, afin de s'approcher de plus en plus d'un modèle multi-locuteur paramétrable suivant des paramètres anatomiques et idiosyncratiques.

6. ANNEXES

6.1. ANNEXE A – RESULTATS

II Consonnes Voyelles

ABC



ART

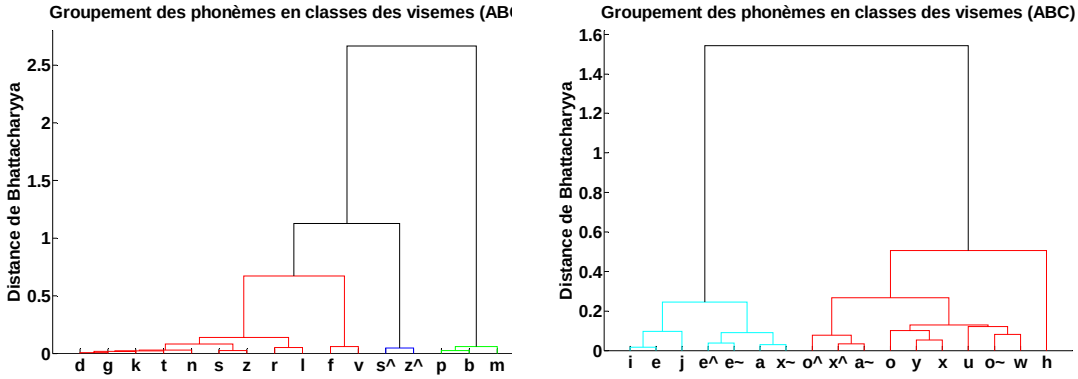
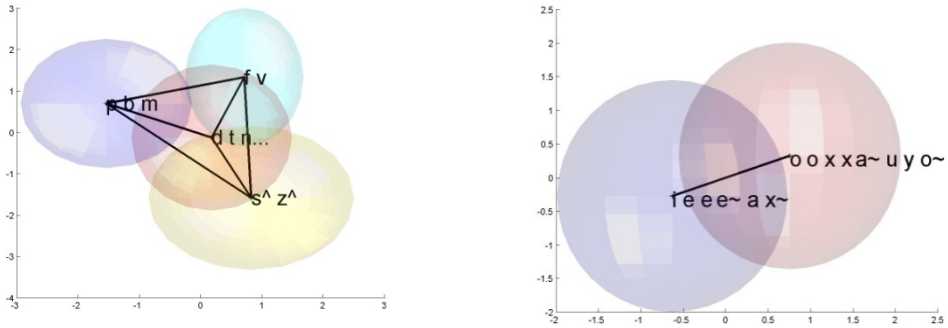


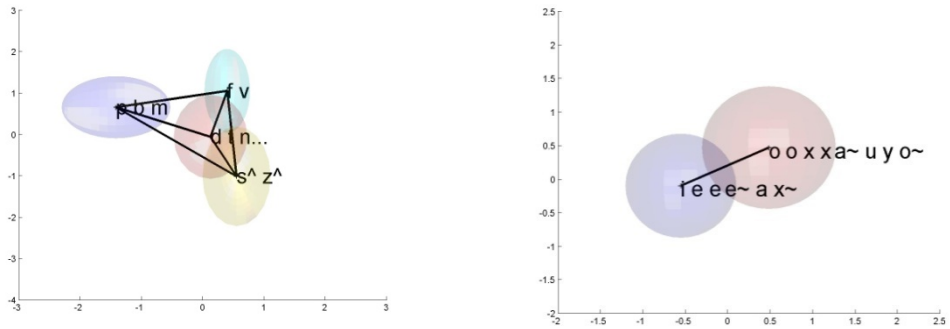
FIGURE 65: GROUPEMENT DES CONSONNES ET VOYELLES EN CLASSES DES VISEMES GRACE A LA DISTANCE DE BHATTACHARYYA POUR LES PARAMETRES ARTICULATOIRES ET GEOMETRIQUES. CORPUS II.

ART Consonnes Voyelles

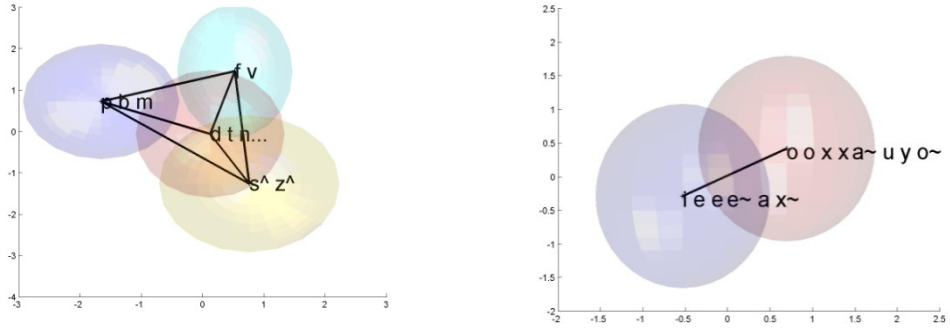
NAT



HMM context e viseme droit



Concat énation



TDA

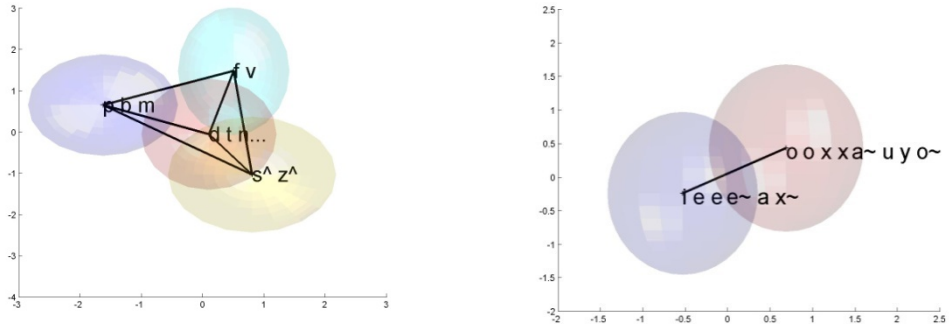


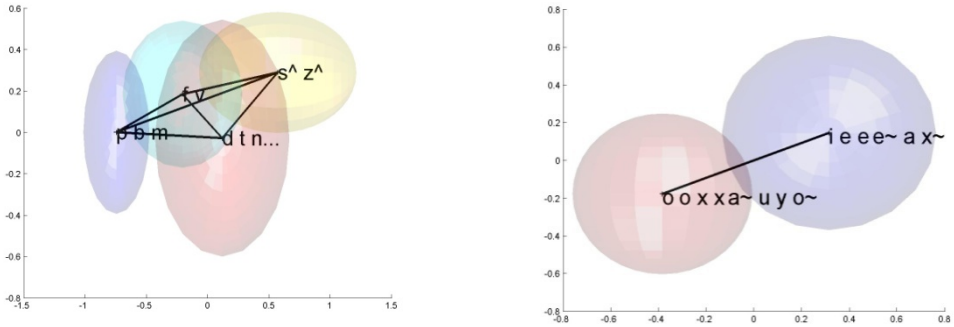
FIGURE 66: ELLIPSES DE DISPERSION DES CIBLES ARTICULATOIRES POUR LES PRINCIPALES CLASSES DES CONSONNES ET DES VOYELLES AVEC LA ADL POUR LES DONNEES NATURELLES, LA SYNTHESE PAR HMM, LA SYNTHESE PAR LA CONCATENATION ET LA SYNTHESE PAR TDA. CORPUS I

ABC

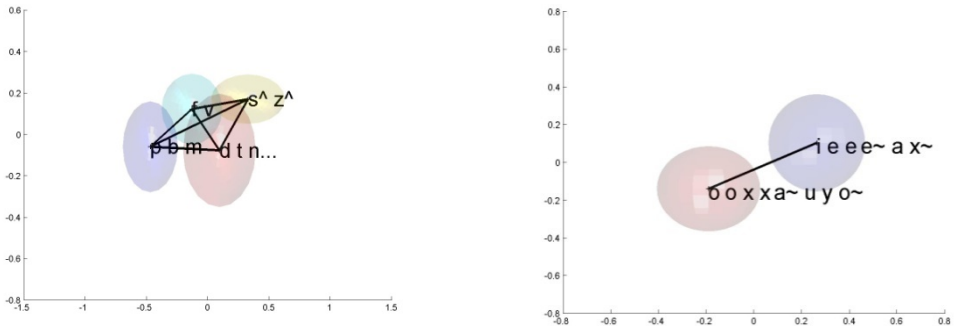
Consonnes

Voyelles

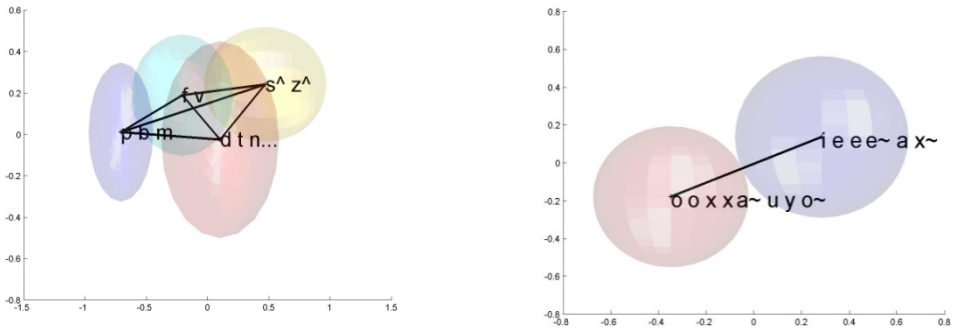
Nat



HMM context e viseme droit



Concat énation



TDA

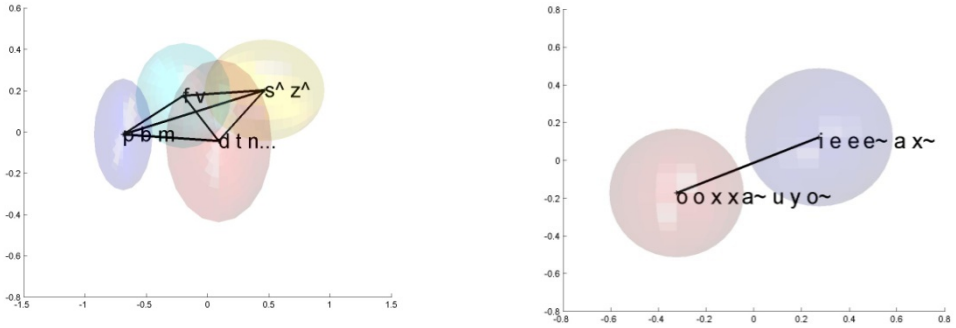
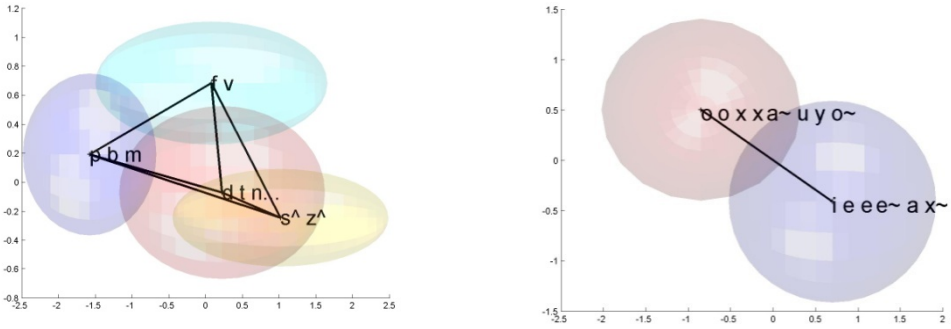


FIGURE 67: ELLIPSES DE DISPERSION DES CIBLES GEOMETRIQUES POUR LES PRINCIPALES CLASSES DES CONSONNES ET DES VOYELLES AVEC LA ADL POUR LES DONNEES NATURELLES, LA SYNTHESE PAR HMM, LA SYNTHESE PAR LA CONCATENATION ET LA SYNTHESE PAR TDA. CORPUS II

ART Consonnes Voyelles

NAT



HMM context e viseme droit viseme

Concat énation

TDA

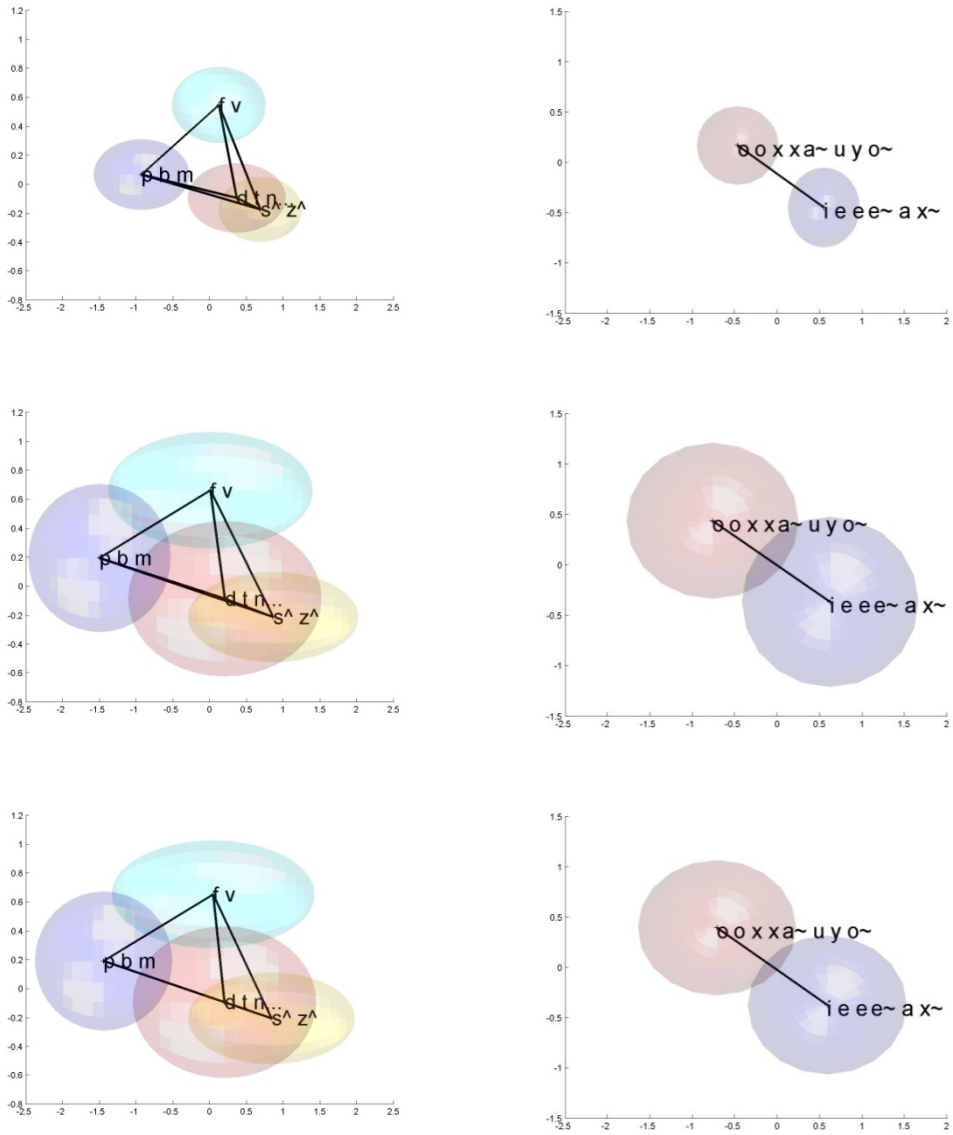


FIGURE 68: ELLIPSES DE DISPERSION DES CIBLES ARTICULATOIRES POUR LES PRINCIPALES CLASSES DES CONSONNES ET DES VOYELLES AVEC LA ADL POUR LES DONNEES NATURELLES, LA SYNTHESE PAR HMM, LA SYNTHESE PAR LA CONCATENATION ET LA SYNTHESE PAR TDA. CORPUS II

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

FIGURE 69: LES CARACTERISTIQUES DE LA ADL (INTER-DISTANCE, INTRA-DISTANCE ET LEUR RAPPORT) DES CONSONNES ET VOYELLES DANS LES ESPACES GEOMETRIQUE ET ARTICULATOIRE POUR LES DONNEES NATURELLES, LA SYNTHESE PAR HMM, LA SYNTHESE PAR LA CONCATENATION ET LA SYNTHESE PAR TDA. CORPUS II. DONNEES D'APPRENTISSAGE ET DE TEST.

6.2. ANNEXE B – LES ALGORITHMES D'APPRENTISSAGE ET DE SYNTHESE PAR HMM

6.2.1. LES MODELES DE MARKOV

Ce chapitre est rédigé en se basant sur l'article de Rabiner (Rabiner 1989) sur l'utilisation des HMM dans la reconnaissance vocale et sur les travaux de (Tamura, Masuko et al. 1998), (Zen, Tokuda et al. 2004), (Yoshimura, Tokuda et al. 1998), (Tokuda, Yoshimura et al. 2000).

NOTATIONS

n	Le nombre d'états du modèle de Markov caché
$S = \{s_1, s_2, \dots, s_n\}$	Les états du HMM
A	La matrice des probabilités de transitions entre les états
$a_{i,j}, i,j \in [1,n]$	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
$O = O_1 \dots O_t \dots O_T$	Une séquence observée
$q_1 \dots q_t \dots q_T$ avec $q_t \in S$	Une suite des états qui a émis une séquence

MODELES DE MARKOV OBSERVABLES

Un ***modèle stochastique observable*** est un processus aléatoire qui peut changer d'état s_i, $i = 1,\dots,n$ au hasard, aux instants $t = 1,2,\dots,T$. Le résultat observé est la suite des états dans lesquels il est passé. Chaque séquence est émise avec une probabilité $P(S) = P(s_1, s_2, \dots, s_T)$. 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

$$\forall t, P(q_t = s_i | q_{t-1} = s_j, q_{t-2} = s_k \dots) = P(q_t = s_i | q_{t-1} = s_j) \quad (1)$$

d'où:

$$P(q_1 \dots q_T) = P(q_1) \times P(q_2|q_1) \times \dots \times P(q_T|q_{T-1}) \quad (2)$$

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 :

$$a_{ij} = P(q_t = s_j | q_{t-1} = s_i), 1 \leq i \leq n, 1 \leq j \leq n \quad (3)$$

avec:

$$\forall i, j, a_{ij} \geq 0, \forall i, \sum_{j=1}^{j=n} a_{ij} = 1 \quad (4)$$

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

$$\lambda = (A, \pi) \quad (5)$$

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 $q_1, q_2, \dots, q_T$, notée $Q(1:T)$, où les q_i prennent leur valeur parmi l'ensemble des n états du modèle $s_1, s_2, \dots, s_n$.

• la suite observable correspond à la ***séquence des observations*** $O_1, O_2, \dots, O_T$, 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 $S = \{s_1, s_2, \ldots, s_n\}$. 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 $V = V_1, V_2, \ldots, V_M$, ces lois sont décrites par une matrice B de taille $(n \times M)$:

$$B = b_i(k) \qquad (6)$$

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:

$$b_j(k) = P(O_t = v_k | q_t = s_j), 1 \leq j \leq n, 1 \leq k \leq M \qquad (7)$$

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 $p(O_t | q_t = s_i) = f_i(O_t)$, qui est appelée par analogie ***vraisemblance d'émission***.

L'hypothèse d'indépendance s'écrit maintenant:

$$p(O_1, O_2, \ldots, O_T | q_1 = s_i, q_2 = s_j, \ldots q_T = s_n) = p(O_1 | q_1 = s_i) p(O_2 | q_2 = s_j) \ldots p(O_T | q_T = s_n) \qquad (8)$$

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

$$f_i(O_t) = \frac{1}{(2\pi)^{N/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(O_t - \mu_i)^T \Sigma_i^{-1}(O_t - \mu_i)\right) \quad (9)$$

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:

$$\lambda_c = \{\pi, A, \mu_i, \Sigma_i | i = 1, \dots, n\} \quad (10)$$

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.

MODELE DE MELANGES DE GAUSSIENNES GMM

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

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

• 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 (9).

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

$$g(O) = \sum_{k=1}^{K} P(\eta = k) p(O|\eta = k) = \sum_{k=1}^{K} c_k g_k(O) \quad (11)$$

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

$$f_i(O_t) = \sum_{k=1}^{M} \frac{c_{ik}}{(2\pi)^{N/2}|\Sigma_{ik}|^{1/2}} \exp\left(-\frac{1}{2}(O_t - \mu_{ik})^T \Sigma_{ik}^{-1}(O_t - \mu_{ik})\right) \quad (12)$$

avec les contraintes:

$$\begin{cases} c_{ik} \geq 0, & \forall i,k \\ \sum_{k=1}^{M} c_{ik} = 1, & \forall i \end{cases} \quad (13)$$

Un HMM à densités continues est alors paramétrisé comme:

$$\Lambda_G = \{\pi, A, \mu_{ik}, \Sigma_{ik}, c_{ik} | i = 1, \ldots, n, k = 1, \ldots, K\} \quad (14)$$

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 et synthèse de la parole.

6.2.2. LA THEORIE DE LA SYNTHESE DE LA PAROLE PAR HMMS

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.

PROBLEMATIQUES DE LA SYNTHESE PAR HMM

Les problématiques de la synthèse par HMM sont les suivantes:

CHOIX DES UNITES DE MODELISATION. 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.

TROIS PROBLEMES POUR HMMS. Les tâches associées aux HMMs sont généralement formulées sous la forme de trois problèmes.

PROBLEME 1 (CALCUL DE LA VRAISEMBLANCE D'UNE SEQUENCE OBSERVEE). Étant donné la modèle λ, défini par(10), *comment calcule-t-on $p(O|\lambda)$, la vraisemblance de la séquence d'observations $O = O_1, O_2, \ldots, O_T$?*

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

$$p(O, \lambda) = \sum_q p(O|q, \lambda) P(q, \lambda) \quad (15)$$

où la somme est faite sur toutes les séquences d'états q possibles. En utilisant l'hypothèse d'indépendance (8):

$$p(O|q,\lambda) = \prod_{t=1}^{T} p(O_t|s_t = q_t,\lambda) = \prod_{t=1}^{T} f_{q_t}(O_t) \quad (16)$$

Ensuite, en utilisant la définition de la probabilité conditionnelle et l'hypothèse (1), le deuxième terme de (15) peut être réécrit comme:

$$\begin{aligned} P(q|\lambda) &= P(s_1 = q_1|\lambda) \prod_{t=2}^{T} P(s_t = q_t|s_{t-1} = q_{t-1},\dots,\lambda) \\ &= P(s_1 = q_1|\lambda) \prod_{t=2}^{T} P(s_t = q_t|s_{t-1} = q_{t-1},\lambda) \\ &= \pi_{q_1} \prod_{t=2}^{T} a_{q_t q_{t-1}} \end{aligned} \quad (17)$$

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

$$p(O|\lambda) = \sum_q \left[\pi_{q_1} f_{q_1}(O_1) \prod_{t=2}^{T} a_{q_t q_{t-1}} f_{q_t}(O_t)\right] \quad (18)$$

Pour calculer cette vraisemblance en utilisant directement l'équation (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.

PROBLEME 2 (RECHERCHE DE LA SEQUENCE D'ETATS OPTIMALE). *Étant donné le modèle λ, comment choisir la séquence d'états $q = q_1, q_2, \dots, q_T$ maximisant $p(O,q|\lambda)$, la vraisemblance conjointe de la séquence d'observations $O = O_1, O_2, \dots, O_T$ et de la séquence d'états? Donc on cherche q^*, tel que*

$$q^* = \arg\max_q p(O,q|\lambda) \quad (19)$$

Selon (16) et (17) $p(O,q|\lambda)$ s'écrit comme:

$$\begin{aligned} p(O,q|\lambda) &= p(O|q,\lambda)P(q|\lambda) \\ &= \pi_{q_1} f_{q_1}(O_1) \prod_{t=2}^{T} a_{q_t q_{t-1}} f_{q_t}(O_t) \end{aligned} \quad (20)$$

Si on fait la recherche de q^* en utilisant directement l'expression (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.

PROBLEME 3 (ESTIMATION DES PARAMETRES DU MODELE). *Comment ajuster les paramètres λ du modèle HMM d'une façon telle que la vraisemblance $p(O|\lambda)$ soit maximale? On cherche alors λ^*, satisfaisant*

$$\lambda^* = \arg\max_\lambda p(O|\lambda) \quad (21)$$

GENERATION DES PARAMETRES VISUELS AVEC LA SYNTHESE PAR HMM. L'objectif de la phase de synthèse est de générer les paramètres visuels à partir de la suite phonétique étiquettée 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 SYNTHESE PAR HMM

Le système de synthèse par HMM comprend deux étapes principales : l'étape d'apprentissage de paramètres des modèles HMM et l'étape de synthèse de paramètres à partir d'une séquence de HMMs.

APPRENTISSAGE DES HMM

En entrée de la phase d'apprentissage il y a une suite de vecteurs de paramètres, chaque paire de vecteurs correspond à une trame. Les vecteurs sont classés en groupes. Chaque groupe de vecteurs correspond à une unité phonétique. Ensuite, un HMM est construit pour chaque unité. Un ensemble de séquences de vecteurs O^k(acoustiques, visuels ou autres) $O = \{O^1, O^2, \ldots, O^m\}$ est utilisé dans l'apprentissage d'un HMM λ, Figure 71. 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**). *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.

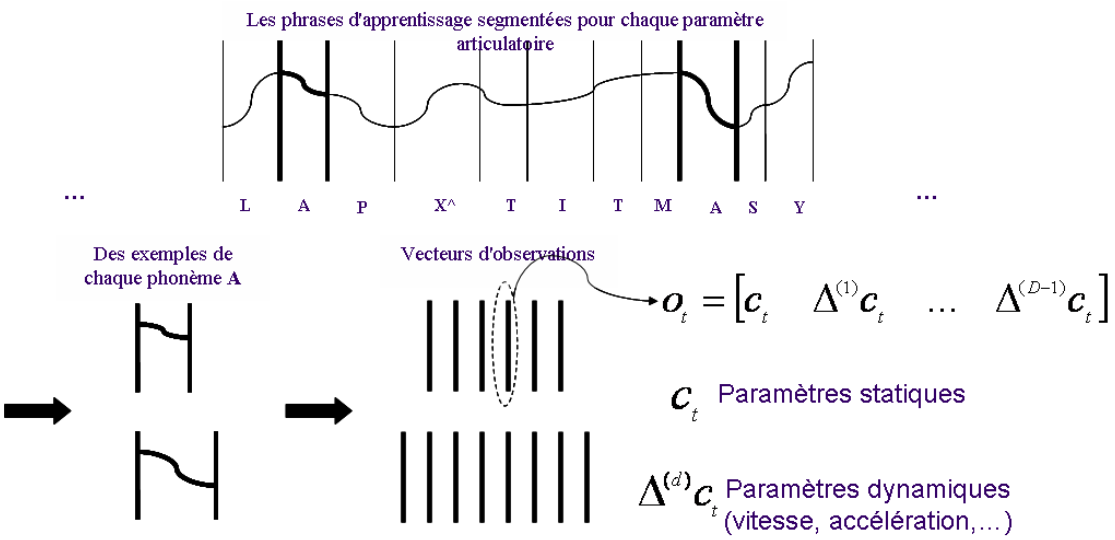


FIGURE 70. EXEMPLE DE LA CONSTRUCTION D'UN VECTEUR D'OBSERVATION POUR UN HMM.

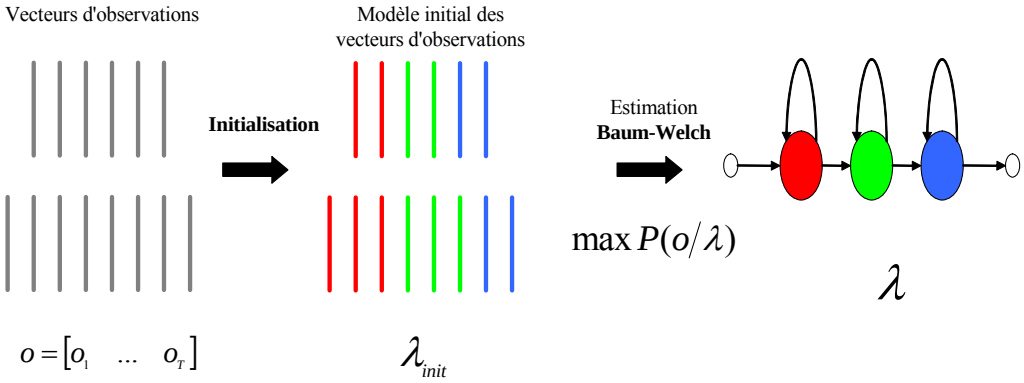


FIGURE 71 : ILLUSTRATION DE L'APPRENTISSAGE D'UN HMM PAR BAUM-WELCH.

Introduisons les notations suivantes:

- $\lambda = \{\pi, A, \mu_i, \Sigma_i | i = 1, \dots, n\}$ les paramètres du modèle estimés à l'itération précédente,
- $\hat{\lambda} = \{\hat{\pi}, \hat{A}, \hat{\mu}_i, \hat{\Sigma}_i | i = 1, \dots, n\}$ les paramètres du modèle estimés à l'itération courante,
- $\gamma_t(i) = P(s_t = q_i | O, \lambda)$ la probabilité d'être dans l'état q_i à l'instant t, étant donnés la séquence d'observations O et le modèle λ,
- $\xi_t(i,j) = P(s_t = q_i, s_{t+1} = q_j | O, \lambda)$ 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 PROBLEME 1: PROCEDURE "AVANT-ARRIERE"

Pour résoudre le **problème 1**, on introduit d'abord deux quantités:

$$\alpha_t(i) = p(O_1, O_2, \dots, O_t, s_t = q_i|\lambda) \quad (22)$$

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

$$\beta_t(i) = p(O_{t+1}, O_{t+2}, \dots, O_T|s_t = q_i, \lambda) \quad (23)$$

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

Procédure "avant"

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

$$\alpha_1(i) = \pi_i f_i(O_1) \quad (24)$$

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

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{n} \alpha_t(i) a_{ij}\right] f_j(O_{t+1}) \quad (25)$$

- Calcul de vraisemblance:

$$p(O|\lambda) = \sum_{i=1}^{n} \alpha_T(i) \quad (26)$$

Procédure "arrière"

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

$$\beta_T(i) = 1 \quad (27)$$

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

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij} f_j(O_{t+1}) \beta_{t+1}(j) \quad (28)$$

- Calcul de vraisemblance:

$$p(O|\lambda) = \sum_{i=1}^{n} \pi_i f_i(O_1) \beta_1(i) \quad (29)$$

Dans l'équation (25) de la procédure "avant" on calcule tout d'abord $p(O_1, \dots, O_T, s_{t+1} = q_j|\lambda)$ la vraisemblance d'émettre la séquence d'observations partielle $O_1, O_2, \dots, O_t$ 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 (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 (30) et (31) on a:

$$\alpha_t(i) = p(O_1, O_2, \dots, O_t, s_t = q_i|\lambda) \tag{30}$$

$$\beta_t(i) = p(O_{t+1}, O_{t+2}, \dots, O_T|s_t = q_i, \lambda) \tag{31}$$

$$\gamma_t(i) = \frac{p(s_t=q_i, O|\lambda)}{p(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{p(O|\lambda)} \tag{32}$$

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

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}f_j(o_{t+1})\beta_{t+1}(j)}{p(O|\lambda)} \tag{33}$$

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:

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

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.

• Initialisation: choisir une approximation initiale $\lambda = \lambda^0$,

• Estimation des probabilités (l'étape d'expectation de l'algorithme EM): calculer $\gamma_t(i)$ et $\xi_t(i,j)$ en utilisant les expressions (32) et (33) avec la procédure "avant-arrière" (en annexe).

• Réestimation des paramètres (l'étape de maximisation de l'algorithme EM):

$$\hat{\pi}_i = \gamma_1(i) \tag{34}$$

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T}\gamma_t(i,j)}{\sum_{t=2}^{T}\gamma_t(i)} \tag{35}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T}\gamma_t(i)o_t}{\sum_{t=1}^{T}\gamma_t(i)} \tag{36}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{T}\gamma_t(i)(o_t-\hat{\mu}_i)(o_t-\hat{\mu}_i)^T}{\sum_{t=1}^{T}\gamma_t(i)} \tag{37}$$

• 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

$$p(O|\hat{\lambda}) \geq p(O|\lambda) \tag{38}$$

SYNTHESE DE SEQUENCES D'OBSERVATION

CONSTRUCTION DE LA SEQUENCE D'ETATS

La séquence d'états peut être obtenue soit à partir des durées des unités (les durées des états sont souvent modélisées par des distributions monogaussiennes correspondantes à chaque état d'un HMM), soit à partir des paramètres acoustiques dans le cas de la synthèse à partir de l'audio. 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. La séquence est obtenue en maximisant la probabilité de sortie **(Problème 2)**:

$$q^* = \arg\max_q p(O,q|\lambda) \tag{39}$$

Tout d'abord, en prenant le logarithme de (20), on définit

$$U(O,q|\lambda) \triangleq \log p(O,q|\lambda) = \log\left(\pi_{q_1} f_{q_1}(O_1)\right) + \sum_{t=2}^{T}\log\left(a_{q_t q_{t-1}} f_{q_t}(O_t)\right) \tag{40}$$

Puisque le logarithme est une fonction croissante, le problème (39) est équivalent au problème suivant:

$$q^* = \arg\max_q U(O,q|\lambda) \tag{41}$$

Ce passage au logarithme sert seulement à simplifier les calculs. Effectivement, dans l'expression (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 72. On voit bien que la solution du problème (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

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

$$\begin{aligned} \delta_1(i) &= \log(\pi_i) + \log(f_i(O_1)) \\ \psi_1(i) &= 0 \end{aligned} \tag{42}$$

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

$$\begin{aligned} \delta_t(j) &= \max_{1 \leq i \leq n} [\delta_{t-1}(i) + \log(a_{ij})] + \log(f_j(O_t)) \\ \psi_t(j) &= \operatorname*{argmax}_{1 \leq i \leq n} [\delta_{t-1}(i) + \log(a_{ij})] \end{aligned} \tag{43}$$

• Terminaison:

$$\begin{aligned} U^* &= \max_{1 \leq i \leq n} [\delta_T(i)] \\ q_{i_T}{}^* &= \operatorname*{argmax}_{1 \leq i \leq n} [\delta_T(i)] \end{aligned} \tag{44}$$

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

$$q_{i_t}{}^* = \psi_{t+1}(q_{i_{t+1}}{}^*) \tag{45}$$

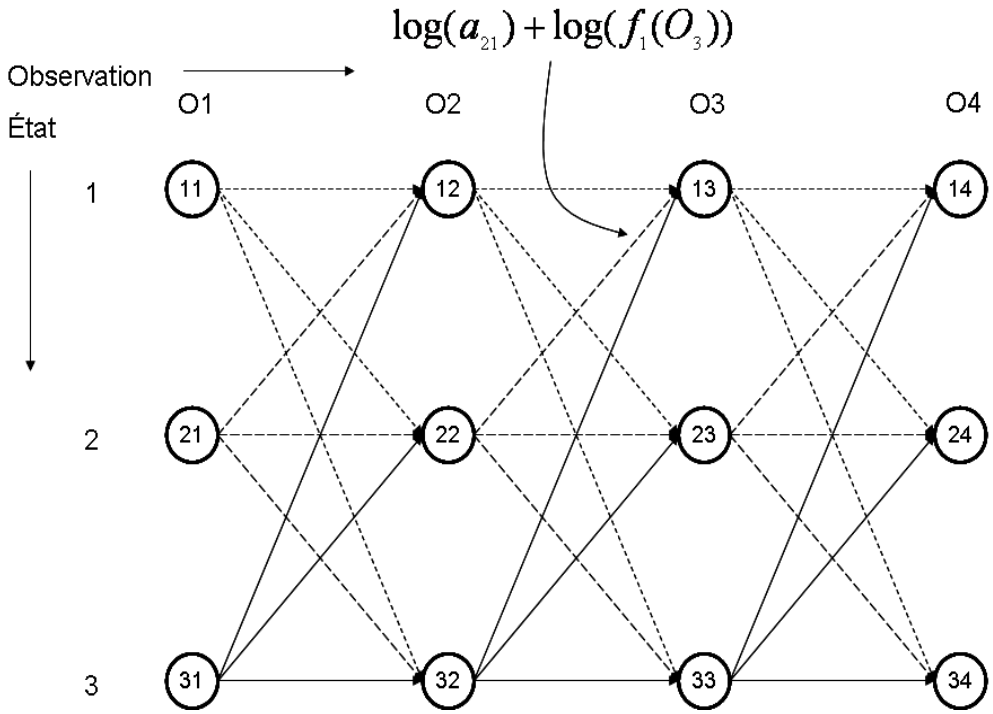


FIGURE 72 : GRAPHE POUR LA RECHERCHE DE VITERBI (N=3, T=4).

Donc l'algorithme de Viterbi donne la séquence d'états optimale $q^* = q_1^*, q_2^*, \ldots, q_T^*$. La vraisemblance maximisant (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.

GENERATION DE PARAMETRES

Cette phase de synthèse a pour but de générer les paramètres en ayant la suite des HMMs et les séquences d'états optimales pour chaque HMM. La séquence des paramètres visuels $O = [O_1^T, \ldots, O_T^T]$ est obtenue en maximisant la vraisemblance $P(O|\lambda)$ par rapport à O.

$$P(O|\lambda) = \sum_q P(O|q,\lambda)P(q,\lambda) \quad (46)$$

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

$$O_t = \left[c_t^T, \Delta^{(1)}c_t^T, \ldots, \Delta^{(2-1)}c_t^T\right]^T \quad (47)$$

$$c_t = [c_t(1), c_t(2), \ldots, c_t(M)]^T \quad (48)$$

$$\Delta^{(d)}c_t = \sum_{\tau=-L_-^{(d)}}^{L_+^{(d)}} w^{(d)}(\tau)c_{t+\tau} \quad (49)$$

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:

$$P(O|q,\lambda) = \prod_{t=1}^{T} N(O_t|\mu_{q_t},\Sigma_{q_t}) = N(O|\mu_q,\Sigma_q) \quad (50)$$

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:

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

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

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

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

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

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

La condition (49) peut être exprimée comme suit:

$$O = Wc \quad (51)$$

où

$$c = [c_1^T, c_2^T, \ldots, c_T^T]^T \quad (52)$$

$$W = [W_1, W_2, \ldots, W_T]^T \otimes I_{M\times M} \quad (53)$$

$$W_t = \left[w_t^{(0)}, w_t^{(1)}, \ldots, w_t^{(D-1)}\right] \quad (54)$$

$$w_t^{(d)} = \left[\underbrace{0,\ldots,0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), w^{(d)}(L_+^{(d)}), \underbrace{0,\ldots,0}_{T-(t+L_+^{(d)})}\right]^T, \quad (55)$$

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

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

Il est evident que la vraisemblance (50) 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 (51) qui existe entre ces paramètres est prise en compte, Figure 73. C'est pour cette raison que la maximisation de la (50) par rapport à O revient à la même chose si l'on fait par rapport à c:

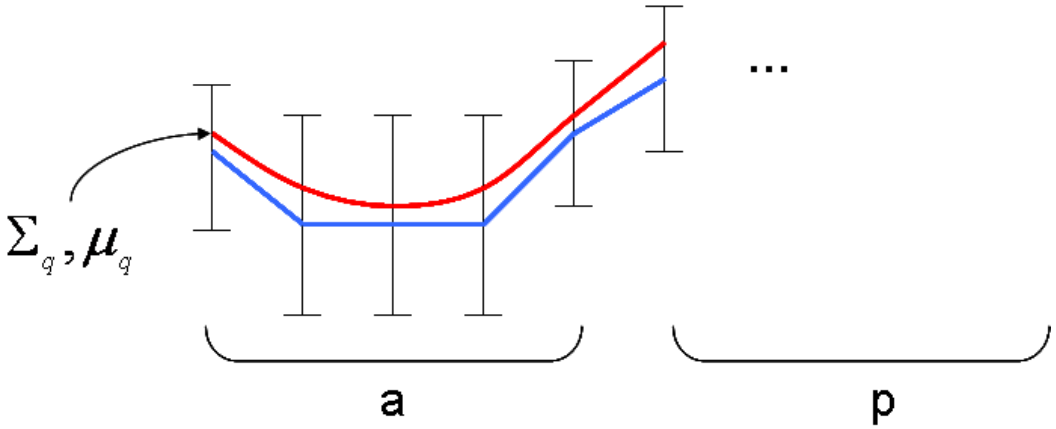


FIGURE 73 : ILLUSTRATION DE L'ALGORITHME DE « LISSAGE ».

$$\frac{\delta \log P(Wc|q,\lambda)}{\delta c} = 0 \tag{56}$$

$$\log P(O|Q,\lambda) = -\frac{1}{2} O^T \Sigma^{-1} O + O^T \Sigma^{-1} M + K \tag{57}$$

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

$$W^T \Sigma^{-1} W c = W^T \Sigma^{-1} M \tag{58}$$

d'où

$$c = (W^T \Sigma^{-1} W)^{-1} W^T \Sigma^{-1} M \tag{59}$$

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

6.3. ANNEXE C

6.3.1. CORPUS I. VISAGE

	'i	'e'	'e^'	'a'	'o^'	'o'	'u'	'y'	'x'	'x^'	'a~'	'e~'	'o~'	'x~'	'p'	't'	'k'	'b'	'd'	'g'	'f'	'v'	's'	'z'	's^'	'z^'	'r'	'l'	'm	'n'	'h'	'j'	'w'	'sil	'silf	'q'	'sil'	'n^	'nt'	'#'	Total
'i'	0	1	0	1	0	1	1	1	0	0	0	0	1	2	1	2	2	1	2	4	9	6	1	1	8	9	6	3	7	1	1	1	1	0	1	0	2	3	3	0	330
'e'	1	0	2	3	2	2	1	4	2	3	1	3	2	7	1	2	1	1	1	1	8	6	1	9	4	3	1	2	3	7	2	3	5	0	2	1	4	0	0	0	298
'e^'	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	1	0	3	3	2	7	6	4	4	2	4	1	8	1	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	2	2	2	2	1	1	1	3	2	1	1	1	5	4	2	1	1	1	1	0	1	0	1	0	2	0	465
'o^'	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	2	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	1	7	5	9	4	2	3	1	5	2	3	7	1	6	8	2	2	2	0	7	0	2	0	0	0	154
'u'	0	3	1	1	0	1	0	0	2	1	2	0	2	1	5	9	2	1	7	1	2	7	8	7	2	5	2	6	2	1	1	2	1	0	1	0	7	2	0	0	140
'y'	1	6	2	3	0	1	0	1	5	2	1	0	2	2	1	6	4	1	2	5	3	2	6	4	2	4	2	1	5	3	1	1	1	0	1	0	1	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	1	2	4	6	2	4	1	0	2	1	0	1	0	2	0	0	0	106
'x^'	0	0	1	1	0	1	1	1	0	1	1	0	0	0	1	2	1	1	4	6	8	1	2	2	1	1	4	1	2	6	1	7	1	0	0	0	2	1	0	0	259
'a~'	1	1	0	2	0	2	0	0	0	0	0	0	1	0	4	2	3	8	1	6	1	2	1	8	1	7	1	1	3	4	1	1	1	0	1	0	2	1	0	0	163
'e~'	0	1	0	0	0	1	0	0	0	0	0	0	0	0	6	1	1	2	8	1	1	2	6	1	0	1	2	1	2	1	1	1	1	0	8	0	2	1	0	0	84
'o~'	1	1	0	0	0	0	0	1	0	0	0	0	0	1	3	1	6	8	5	1	2	1	7	3	3	2	2	2	3	7	1	1	1	0	6	0	1	2	0	0	98
'x~'	0	0	0	0	1	0	0	1	1	0	0	0	1	0	1	8	5	8	4	1	2	1	2	2	1	3	1	3	4	6	1	1	2	0	1	0	1	0	0	0	89
'p'	1	1	1	3	5	7	1	3	9	7	5	3	2	1	0	0	0	1	1	0	0	1	1	1	1	0	1	8	1	2	1	4	2	0	1	0	6	0	0	0	164
't'	3	1	1	2	1	6	1	2	5	2	9	2	1	4	2	0	2	1	2	0	1	0	4	1	4	1	3	2	2	1	3	4	4	0	2	2	1	0	0	0	281
'k'	1	3	5	3	3	1	9	1	2	1	6	1	5	5	0	7	0	1	0	1	0	1	1	1	0	1	7	9	2	1	1	1	4	0	2	7	1	0	0	1	193
'b'	1	6	8	1	3	9	1	8	2	7	3	2	4	1	1	2	1	0	1	0	1	1	0	1	0	1	1	2	1	1	1	7	1	0	1	0	2	0	0	0	154
'd'	1	2	2	9	4	4	2	3	7	3	2	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	1	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	1	0	0	1	86
'f'	1	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	1	1	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'	1	1	1	1	2	5	5	2	3	3	1	4	1	0	6	1	5	2	2	1	1	0	1	1	1	1	1	2	1	1	5	1	4	0	4	0	4	0	0	0	235
'z'	1	1	4	1	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	1	0	0	0	113
's^'	5	6	4	1	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
'z^'	9	2	3	1	1	6	1	2	6	1	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'	3	2	1	3	1	1	8	9	8	1	1	4	2	7	7	3	5	8	1	3	3	4	1	1	5	7	0	1	8	6	1	4	2	0	2	1	2	0	0	0	372
'l'	3	4	1	7	9	1	7	7	1	7	9	3	3	1	4	1	5	1	3	2	3	1	4	3	1	2	1	5	1	3	4	6	7	0	1	2	9	0	0	0	386
'm'	1	1	3	2	1	8	2	5	2	4	3	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'	1	7	6	1	5	1	7	6	3	1	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	1	0	0	0	166
'h'	4	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	1	1	8	0	1	3	0	1	3	1	1	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	4	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	1	2	3	5	1	2	2	1	1	1	2	1	1	1	5	1	2	3	9	1	3	9	2	1	6	2	3	6	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'																																			2		1				
'sil'																																									
	1	2	9	4	8	1	1	1	1	1	1	1	8	2	2	1	2	5	3	8	1	6	2	2	4	4	8	2	5	6	1	2	2	0	0	0	0	0	0	0	442

'n^	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	2	0	0	0	38
'nt'	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

	'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

	'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

6.3.2. CORPUS I. MAIN.

	0	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	7	7	7	7	8	8	8	8	8
0	-	7	6	2	7	1	6	2	4	0	0	2	2	6	1	6	6	1	4	1	0	1	1	6	2	3	2	0	1	2	7	1	0	0	0	0	1	0	2	0	0
1	1	1	4	4	3	1	8	6	1	0	4	2	6	4	0	7	8	3	1	2	3	7	8	1	3	1	9	4	2	7	1	3	0	1	0	1	3	1	1	3	3
1	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	1	0	0	0	0
1	3	1	1	0	0	0	3	0	0	0	0	1	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
1	1	3	0	0	2	1	5	1	0	0	1	4	2	2	1	0	1	1	1	0	0	8	1	3	2	9	6	0	2	0	0	2	0	0	0	0	0	0	0	0	0
1	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
2	1	5	6	3	3	5	9	5	3	0	2	1	6	5	2	1	8	7	3	0	2	1	6	7	2	1	2	6	1	5	5	5	0	0	0	0	7	2	1	6	2
2	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
2	3	0	0	2	0	0	1	0	1	1	1	1	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
2	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
2	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
3	2	1	8	5	9	1	2	5	2	2	5	1	7	6	3	4	1	1	2	2	2	3	1	7	4	1	1	7	3	9	5	3	1	0	0	1	4	1	0	1	2
3	7	7	3	1	1	0	1	1	0	1	0	3	0	1	0	2	6	1	0	0	4	1	4	4	1	3	5	0	0	1	1	0	0	0	0	0	8	0	0	1	0
3	0	5	2	0	0	0	4	1	0	0	1	2	1	1	0	0	7	1	0	0	0	1	1	1	0	2	5	0	1	1	1	0	0	0	0	0	1	0	0	0	0
3	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
3	9	0	2	1	1	3	2	1	0	0	1	1	1	0	1	9	2	4	1	0	1	3	2	5	5	4	4	0	0	2	0	2	0	0	0	2	0	2	0	0	0
4	1	1	4	2	4	2	1	2	2	0	3	1	6	1	2	3	5	6	5	1	0	9	1	2	2	7	2	1	3	7	3	5	0	0	1	0	2	0	0	6	0
4	3	3	3	3	0	2	3	3	0	2	1	4	3	0	0	0	3	0	1	0	0	1	3	3	2	0	6	1	0	4	3	2	0	0	0	1	7	0	0	0	2
4	5	2	1	1	0	2	3	0	0	0	0	1	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
4	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
4	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
5	2	1	8	4	8	8	1	3	1	3	3	4	2	6	7	5	9	5	1	4	0	3	1	3	5	1	3	6	6	7	7	7	0	1	0	0	6	1	4	1	3
5	8	1	4	1	0	8	1	2	4	0	3	8	3	1	1	5	1	3	4	1	3	2	5	3	5	1	2	4	3	1	4	2	1	0	0	0	1	0	0	0	0
5	1	5	1	1	0	1	4	0	1	0	0	3	1	1	1	1	1	0	0	0	1	1	5	3	1	4	1	1	2	0	0	0	0	0	0	0	3	0	0	0	0
5	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
5	8	1	2	4	9	5	7	3	4	2	4	1	4	5	0	7	3	4	3	2	1	2	1	1	9	2	1	1	6	4	6	4	0	0	0	0	1	1	0	0	0
6	2	1	6	1	6	7	2	9	3	5	4	2	1	3	6	3	1	7	7	0	3	2	1	1	2	1	1	4	5	4	5	6	0	3	0	0	4	0	3	1	1
6	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
6	1	2	2	0	0	0	8	0	0	0	0	1	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
6	2	1	0	3	3	5	4	4	2	1	2	3	5	1	1	2	3	3	0	1	0	1	9	4	4	5	1	1	1	0	0	3	0	0	0	0	1	0	0	0	0
6	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
7	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
7	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
7	1	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	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
7	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
8	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	1	6	1	1	2	1	2	0	0	0	0	3	0	1	0	0
8	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
8	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
8	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
8	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	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

CORPUS I. FORME

	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

6.3.1. CORPUS II. VISAGE

	'i'	'e'	'e^'	'a'	'o^'	'o'	'u'	'y'	'x'	'x^'	'a~'	'e~'	'o~'	'x~'	'p'	't'	'k'	'b'	'd'	'g'	'f'	'v'	's'	'z'	's^'	'z^'	'r'	'l'	'm'	'n'	'h'	'j'	'w'	'sil'	'silf'	'q'	'sil'	Total
'i'	3	1	1	4	2	2	1	2	3	0	1	0	3	1	6	3	2	1	1	4	9	9	3	1	3	2	1	3	9	2	1	6	0	0	1	0	1	295
'e'	7	3	4	3	1	9	0	2	4	2	6	9	4	6	1	1	9	8	1	7	5	3	1	1	1	5	1	2	9	7	3	0	1	0	2	0	1	248
'e^'	1	1	1	3	1	1	3	2	0	0	2	2	0	1	5	3	2	2	3	0	1	1	1	8	2	2	4	2	8	1	2	3	1	0	6	0	1	205
'a'	2	2	2	0	1	1	2	5	1	3	3	0	1	1	1	2	1	1	9	7	5	2	2	1	2	1	3	3	1	1	2	6	1	0	1	0	0	295
'o^'	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	2	4	1	8	0	1	0	0	0	0	0	72
'o'	2	1	1	2	0	2	2	2	0	0	2	1	1	0	1	1	6	3	1	2	1	1	7	8	2	5	8	1	6	7	1	2	2	0	1	0	0	139
'u'	1	0	0	2	0	1	1	2	3	0	1	0	2	0	5	1	3	2	6	2	4	6	4	2	0	1	1	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	1	3	2	3	2	8	5	3	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	2	4	5	2	1	1	0	0	8	0	0	98
'x^'	0	0	1	0	0	0	1	0	0	0	1	0	0	1	1	4	8	4	1	6	2	1	1	1	3	4	4	2	1	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	2	1	1	1	5	7	4	1	2	5	9	3	9	1	4	3	1	1	0	2	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	1	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	1	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	1	5	2	8	6	9	1	5	3	2	2	0	0	0	6	0	1	0	0	0	0	4	0	1	0	2	1	0	0	9	2	1	0	4	0	0	161
't'	2	2	1	1	4	6	4	1	5	1	1	4	1	9	2	2	1	0	4	0	0	2	1	0	7	0	4	3	3	3	4	4	5	0	2	1	0	264
'k'	1	0	9	1	1	9	1	6	1	1	4	2	1	3	3	9	0	0	2	0	0	0	1	0	0	0	5	8	1	0	0	0	2	0	1	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	2	0	0	0	4	1	0	3	0	0	77
'd'	2	3	5	9	1	3	1	1	1	3	1	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	1	1	5	2	2	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	2	0	0	1	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	0	1	0	0	0	7	3	0	0	0	1	3	0	2	0	0	67
'v'	1	0	1	4	2	6	3	2	1	0	5	3	1	0	0	0	1	0	2	0	0	0	0	0	0	0	1	0	0	5	0	1	7	0	2	0	0	84
's'	1	2	2	6	2	3	5	1	3	1	1	5	8	1	6	3	1	0	5	0	3	0	1	0	0	0	0	0	7	2	1	2	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
's^'	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
'z^'	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'	2	2	2	2	7	2	4	1	1	2	1	3	3	2	8	1	4	3	2	4	6	2	1	0	3	5	2	1	9	6	1	9	1	0	3	5	0	382
'l'	1	3	1	4	5	1	4	1	1	4	1	4	5	4	2	6	4	1	6	0	2	1	4	1	0	1	0	1	7	1	3	3	1	0	2	1	0	295
'm'	1	1	1	1	1	3	1	5	4	4	2	2	2	1	1	2	0	0	2	0	0	0	1	0	0	0	0	1	0	1	1	4	2	0	1	1	0	121
'n'	1	1	8	1	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	1	2	0	168
'h'	4	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	1	1	5	2	8	0	0	4	2	2	1	1	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	3	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	1	2	1	2	1	7	1	7	1	0	1	2	3	1	1	3	1	1	2	5	4	9	2	0	2	1	3	4	5	1	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

	'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
'e^'	7	11	15	2	4	156	3	7
'a'	7	17	40	25	14	175	3	14
'o^'	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
'x^'	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
's^'	17	9	0	1	0	7	1	5
'z^'	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

	'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

7. REFERENCES BIBLIOGRAPHIQUES

Abry, C., J.-P. Orliaguet, et al. (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. and D. Talkin (1998). 3-D Face Point Trajectory Synthesis using an Automatically derived Visual Phoneme Similarity matrix. AVSP Proceedings, Terrigal, Australia.

Attina, V. (2005). La Langue française Parlée Complétée (LPC) : production et perception. Institut de la Communication Parlée. Grenoble, INPG.

Badin, P., G. Bailly, et al. (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. ETRW on Speech Synthesis, Perthshire - Scotland.

Bailly, G., M. Bérar, et al. (2003). "Audiovisual speech synthesis." International Journal of Speech Technology **6**: 331-346.

Bailly, G., F. Elisei, et al. (2006). Degrees of freedom of facial movements in face-to-face conversational speech. International Workshop on Multimodal Corpora. Genoa - Italy**:** 33-36.

Bailly, G., G. Gibert, et al. (2002). Evaluation of movement generation systems using the point-light technique. IEEE Workshop on Speech Synthesis, Santa Monica, CA.

Basu, S., N. Oliver, et al. (1998). "3D lip shapes from video: a combined physical-statistical model." Speech Communication **26**: 131-148.

Beier, T. and S. Neely (1992). "Feature-based image metamorphosis." Computer graphics.

Berthommier, F. (2003). Direct Synthesis of Video from Speech Sounds for New Telecommunication Applications. Smart Object Conference.

Beskow, J. (1995). Rule-based Visual Speech Synthesis. Proceedings of Eurospeech '95, Madrid, Spain.

Beskow, J. (2003). Talking Heads - Models and Applications for Multimodal Speech Synthesis. KTH. Stockholm.

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

Boite, R., H. Boulard, et al. (2000). Traitement de la Parole. Lausanne, Presses Polytechniques et Universitaires Romandes.

Bowden, R. (2000). Learning non-linear Models of Shape and Motion. Systems Engineering. Uxbridge, UK, Brunel University. **PhD**.

Bregler, C. (1997). Video rewrite: driving visual speech with audio. Proceedings of Computer Graphics.

Breton, G., C. Bouville, et al. (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. and S. D. Scott (1998). Two and Three-Dimensional Audio-Visual Speech Synthesis. AVSP'98, Australia.

Browman, C. P. and L. Goldstein (1990a). "Gestural specification using dynamically-defined articulatory structures." Journal of Phonetics **18**: 299-320.

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

Browman, C. P. and L. M. Goldstein (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. Eurospeech'95, Madrid/Spain.

Campbell, W. N. and S. D. Isard (1991). "Segment durations in a syllable frame." Journal of Phonetics **19**: 37-47.

Caplier A., S. Stillittano, et al. (2007). "Image and video for hearing impaired people." EURASIP Journal on Image and Video Processing: 14.

Chang, Y.-J. and T. Ezzat (2005). Transferable videorealistic speech animation. ACM Siggraph/Eurographics Symposium on Computer Animation.

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

Cohen, M. M., D. W. Massaro, et al. (2002). Training a talking head. Fourth International Conference on Multimodal Interfaces, Pittsburgh Pennsylvania.

Cootes, T. F., G. J. Edwards, et al. (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. and L. Miclet (2002). Apprentissage Artificiel.

Cosatto, E. and H. P. Graf (2000). Photo-realistic talking-heads from image samples. Trans. on Multimedia.

Cosi, P., E. M. Caldognetto, et al. (2002). Labial Coarticulation Modeling for Realistic Facial Animation. ICMI Pittsburgh, PA, USA.

Cosker, D., D. Marshall, et al. (2003). Video realistic talking heads using hierarchical non-linear speech-appearance models. Mirage, INRIA Rocquencourt, France.

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

Curinga, S., F. Lavagetto, et al. (1996). Lips Movements Synthesis using Time-Delay Neural Networks. Signal Processing VIII Theory and Applications, Trieste - Italy.

d'Alessandro, C. and E. Tzoukermann, Eds. (2001). Synthèse de la parole à partir du texte. Traitement automatique des langues. Paris, Hermès.

Deng, Z., J. P. Lewis, et al. (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. and L. Spitz (1980). "The detection of audiovisual desynchrony." Perception **9**: 719-721.

Donovan, R. (1996). Trainable Speech Synthesis, University of Cambridge. **Phd**.

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

Elisei, F., M. Odisio, et al. (2001). Creating and controlling video-realistic talking heads. Auditory-Visual Speech Processing Workshop, Scheelsminde, Denmark.

Engwall, O. (2000). A 3D tongue model based on MRI data. International Conference on Speech and Language Processing, Beijing - China.

Engwall, O. (2000). Are statistical MRI data representative of dynamic speech? Results from a comparative study using MRI, EMA and EPG. International Conference on Speech and Language Processing, Beijin - China.

Engwall, O. (2002). Evaluation of a system for concatenative articulatory visual speech synthesis. Proc of ICSLP'2002.

Eriksson, E. J., K. P. H. Sullivan, et al. (2002). The importance of anticipatory coarticulation in the perception of ±round in Swedish front vowels: an investigation comparing natural speech with diphone synthesis. Cognitive Science Conference, Perth, Australia.

Ezzat, T., G. Geiger, et al. (2002). MARY101: A TRAINABLE VIDEOREALISTIC SPEECH ANIMATION SYSTEM. Audiovisual Speech Processing. E. Vatikiotis-Bateson, MIT Press.

Ezzat, T. and T. Poggio (1998). MikeTalk: a talking facial display based on morphing visemes. Computer Animation, Philadelphia, PA.

Fagel, S. (2006). Joint Audio-Visual Unit Selection - The JAVUS Speech Synthesizer. International Conference on Speech and Computer, St. Petersburg.

Fagel, S. and C. Clemens (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., T. Ezzat, et al. (2003). Perceptual evaluation of videorealistic speech. Cambridge, USA, Massachusetts Institute of Technology.

Gibert, G. (2006). Conception et évaluation d un système de synthèse 3D de Langue française Parlée Complétée (LPC) à partir du texte. Institut de la Communication Parlée. Grenoble, INPG. **Phd:** 251.

Girosi, F., M. Jones, et al. (1995). "Regularization theory and neural networks architectures." Neural Computation **7**: 219--269.

Groleau J., Chabanas M., et al. (2007). A biomechanical model of the face including muscles for the prediction of deformations during speech production. Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA'2007.

Hallgren, A. and B. Lyberg (1998). Visual speech synthesis with concatenative speech. AVSP 1998.

Hazen, T. J. (2006). Visual model structures and synchrony constraints for audio-visual speech recognition. 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. IEEE Transactions on Speech and Audio Processing.

Hong, P., Z. Wen, et al. (2002). Real-Time Speech-Driven Face Animation. MPEG-4 Facial Animation, John Wiley & Sons, Ltd**:** 115-124.

J.Hardcastle, W. and N. Hewlett (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, Computer Science Department, WSU.

Kakumanu, P., R. Gutierrez-Osuna, et al. (2001). Speech-driven Facial Animation. Proceedings of the ACM conference on Perceptual User Interfaces (PUI 2001).

Kakumanu, P., R. Gutierrez-Osuna, et al. (2002). "Comparing Different Acoustic Data-Encoding for Speech-Driven Facial Animation." Speech Communication.

Klaus, H., H. Klix, et al. (1993). An evaluation system for ascertaining the quality of synthetic speech based on subjective category rating tests. Proceedings of the Third European Conference on Speech Communication and Technology, Berlin.

Kozhevnikov, V. and L. Chistovich (1965). "Speech: Articulation and Perception." Joint Publications Research Service.

Lanitis, A., C. J. Taylor, et al. (1995). A unified approach for coding and interpreting face images. Proc. Int. Conf. Computer Vision, Cambridge.

Le Goff, B., T. Guiard-Marigny, et al. (1994). Real-Time Analysis-Synthesis and Intelligibility of Talking Faces. Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA.

Lee, Y., D. Terzopoulos, et al. (1995). Realistic modeling for facial animation. SIGGRAPH, Los Angeles, CA.

LeGoff, B. and C. Benoit (1996). A text-to-audiovisual-speech synthesizer for french. International Conference on Spoken Language Processing (ICSLP), Philadelphia, USA.

Lofqvist, A. (1990). "Speech as audible gestures." Speech Production and Speech Modeling: 289–322.

Lucero, J. C. and K. G. Munhall (1999). "A model of facial biomechanics for speech production." Journal of the Acoustical Society of America **106**: 2834-2848.

Lucero, J. C., K. G. Munhall, et al. (1997). "Muscle-based modeling of facial dynamics during speech." The Journal of the Acoustical Society of America **101**(5): 3175-3176.

Mak, B. and E. Banard (1996). "Phone clustering using Bhattacharyya distance." Proceedings of the International Conference on Spoken Language Processing **4**: 2005{2008.

Marschark, M., D. LePoutre, et al. (1998). Mouth movement and signed communication. Hearing by Eye II. R. Campbell, B. Dodd and D. Burnham. Hove, United Kingdom, Psychology Press Ltd. Publishers**:** 245-266.

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

Massaro, D. W., Beskow, J., Cohen, M.M., Fry C.L., Rodriquez, T. (1999). Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. Proceedings from AVSP'99, Santa Cruz, USA.

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

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

Minnis, S. and A. P. Breen (1998). Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. International Conference on Speech and Language Processing, Beijing, China.

Odisio, M. and G. Bailly (2004). Audiovisual perceptual evaluation of resynthesised speech movements. Proceedings of the International Conference on Spoken Language Processing, Jeju Island, Korea.

Odisio, M., G. Bailly, et al. (2004). "Tracking talking faces with shape and appearance models." Speech Communication

Öhman, S. E. G. (1967). "Numerical model of coarticulation." Journal of the Acoustical Society of America **41**: 310-320.

Öhman, T. (1998). An audio-visual speech database and automatic measurements of visual speech. Stockholm - Sweden, Quaterly Progress and Status Report, Department of Speech, Music and Hearing - KTH**:** 61-76.

Okadome, T., T. Kaburagi, et al. (1999). Articulatory movement formation by kinematic triphone model. SMC '99 Tokyo, Japan.

Okadome, T., S. Suzuki, et al. (2000). Recovery of articulatory movements from acoustics with phonemic information. Proceedings of the 5th Seminar on Speech Production, Kloster Seeon.

Olives, J.-L., R. Möttönen, et al. (1999). Audio-Visual Speech Synthesis for Finnish. Auditory-visual Speech Processing Workshop, Santa Cruz, CA.

Pandzic, I., J. Ostermann, et al. (1999). "Users evaluation: synthetic talking faces for interactive services." The Visual Computer **15**: 330-340.

Pandzic, I. and F. R. (2002). MPEG4 - Facial Animation.

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., N. Badler, et al. (1996). "Generating Facial Expressions for Speech." Cognitive Science **20**(1): 1-46.

Perkell, J. S. and C.-M. Chiang (1986). Preliminary support for a "hybrid model" of anticipatory coarticulation. XII International Congress of Acoustics, Toronto, Canada.

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

Pighin, F., J. Hecker, et al. (1998). Synthesizing Realistic Facial Expressions from Photographs. Proceedings of Siggraph, Orlando, FL, USA.

Pitermann, M. (2004). Chaos dans la modélisation des tissus mous. Journées d'Etude sur la Parole (JEP) Fès, MOROCCO.

Platt, S. M. and N. I. Badler (1981). "Animating facial expressions." Computer Graphics **15**(3): 245-252.

Potamianos, G., C. Neti, et al. Audiovisual automatic speech recognition: an overview. Audiovisual speech processing, MIT Press.

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

Revéret, L., G. Bailly, et al. (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. International Conference on Speech and Language Processing, Beijing - China.

Saltzman, E. L. and K. G. Munhall (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., D. S. Kagels, et al. (1994). Synthesis of speaker facial movement to match selected speech sequences. Australian Conference on Speech Science and Technology, Perth, Australia.

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. and I. Pollack (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. Hearing by eye: the pyschology of lipreading. B. Dodd and R. Campbell. Hillsdale, NJ - USA, Lawrence Erlbaum Associates**:** 3-51.

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

Tachibana, M., J. Yamagishi, et al. (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., S. Kondo, et al. (1999). Text-to-audiovisual speech synthesis based on parameter generation from HMM. European Conference on Speech Communication and Technology, Budapest, Hungary.

Tamura, M., T. Masuko, et al. (1998). Visual speech synthesis based on parameter generation from HMM: speech-driven and text-and-speech-driven approaches. AVSP 1998.

Taylor, P. and A. W. Black (1999). Speech synthesis by phonological structure matching. EuroSpeech, Budapest, Hungary.

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

Theobald, B.-J., J. A. Bangham, et al. (2003). Evaluation of a talking head based on appearance models. Auditory-visual Speech Processing Workshop, St Jorioz, France.

Theobald, B. J., J. A. Bangham, et al. (2001). Visual speech synthesis using statistical models of shape and appearance. Auditory-Visual Speech Processing Workshop, Scheelsminde - Denmark.

Toda, T., A. W. Black, et al. (2008). "Statistical Mapping between Articulatory Movements and Acoustic Spectrum with a Gaussian Mixture Model." Speech Communication **50**(3): 215-227.

Toda, T. and K. Tokuda (2007). "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis." IEICE Transactions on Information and Systems(E90-D(5)): 816-824.

Tokuda, K., T. Masuko, et al. (1995). An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. Proc. EUROSPEECH.

Tokuda, K., T. Yoshimura, et al. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. ICASSP.

van Santen, J. P. H. (1997). Segmental duration and speech timing. Computing prosody: Computational models for processing spontaneous speech. Y. Sagisaka, N. Campbell and N. Higuchi, Springer Verlag**:** 225-249.

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. INTERSPEECH.

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

Woodland, P. C., J. J. Odell, et al. (1994). Large vocabulary continuous speech recognition using HTK. ICASSP'94.

Yamamoto, E., S. Nakamura, et al. (1998). Subjective evaluation for HMM-based speech-to-lip movement synthesis. AVSP-1998.

Yehia, H. C., P. E. Rubin, et al. (1998). "Quantitative association of vocal-tract and facial behavior." Speech Communication **26**: 23-43.

Yoshimura, T., K. Tokuda, et al. (1998). Duration modeling for HMM-based speech synthesis. ICSLP.

Zen, H., K. Tokuda, et al. (2004). An introduction of trajectory model into HMM-based speech synthesis. ISCA Speech Synthesis Workshop.