INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° : ␣␣␣␣␣␣␣␣␣␣

THÈSE

pour obtenir le grade de

DOCTEUR de l'INPG
Spécialité : SIGNAL, IMAGE, PAROLE, TÉLÉCOMS

préparée au laboratoire
Institut de la Communication Parlée, UMR CNRS 5009

dans le cadre de l'Ecole Doctorale
« **Électonique, Électrotechnique, Automatique et Traitement du Signal** »

présentée et soutenue publiquement par
Guillaume Gibert
le 5 avril 2006

Titre :

Conception et évaluation d'un système de synthèse 3D de Langue française Parlée Complétée (LPC) à partir du texte

Directeur de thèse :
Gérard Bailly

JURY

M.	Jean-Marc Chassery,	Président
Mme	Sylvie Gibet,	Rapporteur
M.	Christophe d'Alessandro,	Rapporteur
M.	Gérard Bailly,	Directeur de thèse
Mme	Nadine Vigouroux,	Examinateur

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR de l'INPG
Spécialité : SIGNAL, IMAGE, PAROLE, TÉLÉCOMS

préparée au laboratoire
Institut de la Communication Parlée, UMR CNRS 5009

dans le cadre de l'Ecole Doctorale
« **Électonique, Électrotechnique, Automatique et Traitement du Signal** »

présentée et soutenue publiquement par
Guillaume Gibert
le 5 avril 2006

Titre :

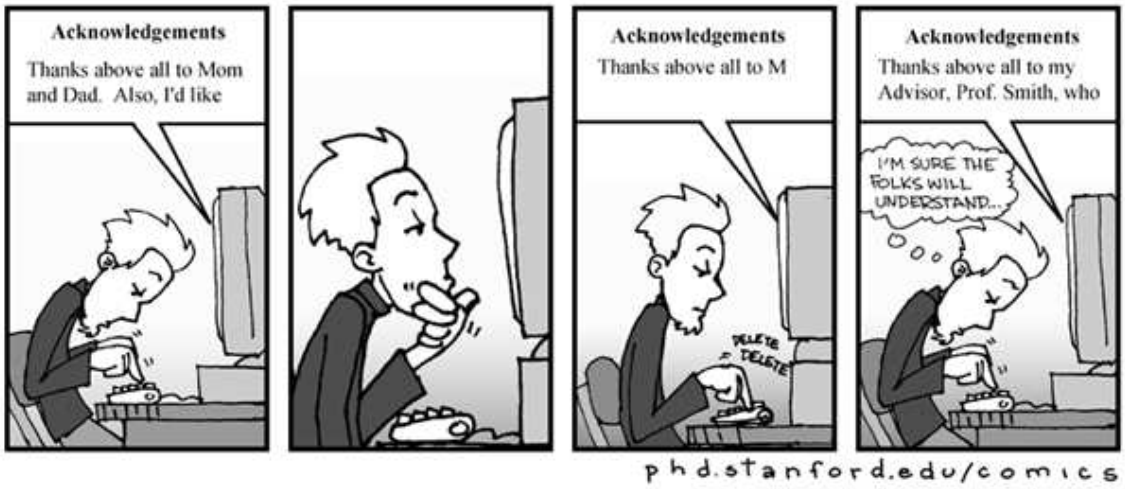
Conception et évaluation d'un système de synthèse 3D de Langue française Parlée Complétée (LPC) à partir du texte

Directeur de thèse :
Gérard Bailly

JURY

M.	Jean-Marc Chassery,	Président
Mme	Sylvie Gibet,	Rapporteur
M.	Christophe d'Alessandro,	Rapporteur
M.	Gérard Bailly,	Directeur de thèse
Mme	Nadine Vigouroux,	Examinateur

Remerciements



La tradition étant, je vais commencer comme Mike Slackenerny par les remerciements en direction de mon directeur de thèse, Gérard Bailly. Après ces quatre années de «vie commune», l'histoire s'achève finalement. L'expérience fut fort enrichissante, tant au niveau scientifique qu'au niveau relationnel. Merci d'avoir accepté de me prendre en stage de DEA et surtout d'avoir accepté de poursuivre l'expérience pour la thèse, savais-tu vraiment ce que tu faisais ? ! J'espère que tu ne te repens pas !

Je tiens également à remercier les directeurs qui se sont succédés au laboratoire : tout d'abord, Pierre Escudier qui m'a acceuilli au sein de l'ICP au début de ma thèse et avec qui je continue à avoir des discussions scientifiques de haut niveau (4-4-2 ou 4-2-3-1 ?), puis l'équipe dirigeante actuelle Jean-Luc Schwartz et Pierre Badin.

Je tiens à remercier Mme Sylvie Gibet et M. Christophe d'Alessandro d'avoir accepté d'être les rapporteurs de mes travaux de thèse. Je remercie également Mme Nadine Vigouroux d'en avoir été l'examinatrice et M. Jean-Marc Chassery d'avoir accepté de présider mon jury de thèse.

Je souhaite également remercier les personnes en relation avec le code LPC au labo et ailleurs : Marie-Agnès Cathiard, Denis Beautemps et Virginie Attina, Mme Martine Marthouret, les associations ADIDA38, AFIDEO, ALPC, ARDDS et tous les sujets qui ont bien voulu passer mes tests contre un *CaramBar* !

Bien sûr, je tiens à remercier Fabien sans qui je n'aurais jamais fini cette thèse. Tous ces déjeuners au RU d'Arsonval à refaire l'actualité sportive (ah ! les matchs de L1 et ses affiches **Nancy-Nice** (le 30 octobre 2005 à 20h au Stade Marcel Picot, un beau 0-0 à l'ancienne, quel

beau championnat quand même !), l'actualité tout court... qui se poursuivaient au Select chez Jacky et Jean-Pierre avec les habitués : Prof' et Crazu' entres autres ! Si les murs du Select pouvaient parler, non il ne vaut mieux pas... C'était un beau oai quand même ! Un bol d'air frais dans la journée ! Le truc qui te permet de tenir. Merci encore ! Je n'oublie pas Juliette sans qui Fabien ne serait pas celui qu'il est et qui a bien voulu partager son petit mari avec moi.

Fabien m'a introduit au groupe CIS du laboratoire TIMA dont je suis devenu un membre par alliance. Merci aux membres du groupe pour toutes ces soirées, ces Laser games, etc. Je ne me lance pas dans une liste de noms, je risque d'en oublier et ça serait dommage.

Revenons à l'ICP, avec les anciens et actuels doctorants et assimilés, merci à : Matthias Odisio (pour m'avoir montré la voie), Bleicke Holm (pour sa disponibilité, tu étais toujours là pour aider le pauvre stagiaire que j'étais), Marion Dohen (pour avoir passé 3 ans dans le même bureau que moi et de ne pas avoir craquée à cause de ma maniaquerie), Virginie Attina (pour m'avoir supporté quand je m'auto-layais et ça c'est énorme !), Annemie Van Hirtum (pour être tout le temps là et apporter de la vie et de l'humour à ce labo) et ses stagiaires de feu (Alexis, Philippe (quel escroc !)), Francesca, Bertrand Rivet dit RV (pour être mon frère nominatif), Julie Fontecave (pour ses gâteaux et sa gentillesse), Antoine Serrurier (pour être toujours plus bronzé qu'un aoûtien), Nicolas Ruty (parce qu'il est pire que moi en fait !), Mohammad Firouzmand, Stefan Raidt (pour avoir pris la relève et pour ses siestes mémorables aux séminaires), Oxana Govokhina (qui relève le défi de faire une thèse ICP-France Télécom !), Julien Cisonni et Lucie Bailly (vont-ils toujours à l'ènesergue ?), Amélie (parce que la parole est multimodale sinon ça ne rime pas), Pauline dit Popo (pour supporter de se faire charier à longueur de journée, c'est dur d'être américaine en France), Frédéric Elisei (pour être plus actif que moi), Iaroslav (toujours là pour aider quelqu'un surtout si c'est un problème de LaTeX), David Sodoyer (pour avoir partager les heures de TP LESTI et ce n'est pas rien), Noureddine Aboutabit (pour ses phrases du jour que je n'ai jamais reçues), Claire et Jérémy, Liang.

Un grand merci à l'équipe technique composée de Monique, Christian, Nino et du mail magique `support@icp.inpg.fr` ! J'ai souvent eu affaire à eux vu le nombre d'ordinateurs (6 !) qui sont passés entre mes mains durant ces 4 années.

Un grand merci également à l'équipe administrative constituée de Nadine et Mme Gaude, toujours disponibles et efficaces !

Enfin, un grand merci :
À Virginie, *per tutto che mi hai dato* et pour tout le reste !
À Nino, avec qui je partage plus que la passion du sport ! Heureusement que tu étais présent pendant tout ce temps au labo !
À Mme Gaude, pour son attention permanente et sa gentillesse.
Aux permanents de ce labo pour avoir été «présents» !
Au coin café de l'ICP ! Le meilleur spectacle d'improvisation en langue française que l'on peut trouver sur Terre avec son maître de cérémonie hors pair : Frédéric Berthommier.
À Nathalie qui n'aura pas eu à me supporter longtemps dans le bureau.
À M. Berthomier pour m'avoir donné goût au traitement du signal ; sans son cours je ne serai

pas ici aujourd'hui.
À Pascal Dubesset qui s'est retrouvé une nuit de dimanche à lundi à relire ma thèse alors que je pense que ce n'était pas sa priorité.
À Grégory dit «ma bécarette en tutu devant le Dia» qui a eu *los cojones* d'arrêter sa thèse, ce que je n'ai pas eu. Je l'ai fini un peu pour toi.
À Yves dit M. «bien au fond à gauche» pour avoir osé faire *ça* à sa soutenance de DEA !
À Cédric dit «Piche» pour tous ces bons moments en DEA et après.
Para ti, Isa, fuiste la primera victima de esta tesis, lo siento. Nunca te olvidaré.
For you, Sze Yee, *wo hen gaoxing* i met you and i'm very sorry it ended like this, you deserve better *yin wei ni zhen hao* !
À Mireille et Claude sans qui je n'en serai pas là, merci de m'avoir toujours soutenu.
À Bénédicte, ma petite soeur et Florent et Sébastien (et sa petite famille), mes «quasi» frères !
À Charly, pour me rappeler sans cesse que je dois commencer à cotiser pour ma retraite et pour tout le reste.
Au reste de ma famille, que je ne suis pas souvent allé voir depuis le début de cette thèse, je m'en excuse.
À mon ami, le silence !
Le seul ami qui ne trahit jamais. Confucius
À Jean-Claude, pour ta capacité à fournir des phrases du jour !
Selon les statistiques, il y a une personne sur cinq qui est déséquilibrée. S'il y a 4 personnes autour de toi et qu'elles te semblent normales, c'est pas bon ! J.-C. Van Damme
À Mike Slackenerny et toute la bande de Piled Higher and Deeper (`http://www.phdcomics.com/`) pour m'avoir aidé à ne plus culpabiliser lorsque j'étais atteint de procrastination.

Enfin, à toi qui n'a pas encore trouvé ton nom dans cette liste, je pense que c'est juste parce que je ne te connais pas ou alors parce que j'avais envie de te casser les deux genoux pendant ma thèse... sinon ça veut dire que ça va mal pour moi et qu'Aloïs n'est pas loin.

Introduction

La parole est par nature multimodale ; elle résulte d'un ensemble de gestes articulatoires rendus audibles mais aussi visibles. Avec les possibilités offertes par les nouvelles technologies, l'ambition des chercheurs s'est naturellement portée sur la faisabilité de systèmes informatiques permettant de générer de manière artificielle (ou virtuelle) cette *parole*, c'est-à-dire de pouvoir disposer de « *têtes parlantes audiovisuelles* » (re)produisant les différents gestes articulatoires permettant de transmettre les informations linguistiques et paralinguistiques.

Synthèse de parole audiovisuelle à partir du texte

Qu'est-ce donc ?

La synthèse de parole audiovisuelle à partir de texte (texte quelconque tapé au clavier) consiste à faire apparaître sur l'écran d'un (micro-)ordinateur, dans un temps *raisonnable*[1], un visage virtuel «parlant», délivrant la phrase synthétisée dans les modalités audio (signal acoustique) et visuelle (mouvements faciaux cohérents et réalistes). L'objectif de ces systèmes est d'expliquer et reproduire au mieux toute la variabilité des signaux acoustiques et des mouvements faciaux observés en parole spontanée. Pour être efficaces et utiles, ces systèmes doivent être capables de transmettre une information linguistique intelligible et compréhensible avec un minimum d'effort de la part de l'interlocuteur.

Attention ! À ne pas confondre...

Il est important de distinguer ce type de systèmes des visages parlants que l'on peut rencontrer dans le monde de l'animation, comme par exemple, dans les jeux vidéo ou dans les films tels que *Final Fantasy* ou *Shrek*. En effet, dans ces cas-là, il ne s'agit pas réellement de synthèse (dans le sens de «génération automatique») mais plutôt de vidéos. Dans l'industrie des jeux, la plupart du temps, les moteurs de jeux ne font que rejouer des trajectoires pré-enregistrées ; ainsi, ces sytèmes ne sont pas capables de générer et prononcer de nouvelles phrases. Dans l'industrie du cinéma, l'animation faciale est effectuée en grande partie «manuellement» par des animateurs, la notion de *temps raisonnable* ne pouvant donc pas s'appliquer dans ce cas, sans compter que les voix ne sont pas synthétisées, mais enregistrées puis rajoutées en post-production.

Plus précisément, en quoi cela consiste-t-il ?

Si l'on décompose le sujet, on voit apparaître quatre notions fondamentales : la *synthèse*, la *parole*, l'*audiovisuel* et enfin l'*entrée textuelle*. On peut donc remarquer d'ores et déjà que

[1] le terme «raisonnable» employé ici est très général et dépend en fait des applications visées ; définissons le toutefois dans un premier temps comme un temps de réponse de l'ordre de quelques secondes.

l'entrée de notre système sera du texte, qu'il faudra traiter afin d'en sortir une information utile. En sortie, nous avons un signal de parole audiovisuelle, soit un signal bimodal, audio et vidéo. Entre les deux, nous avons la *synthèse* : la phrase correspondant à l'entrée textuelle n'ayant pas été enregistrée (du moins pas dans sa totalité), le système de synthèse devra mettre en oeuvre un certain nombre d'algorithmes afin de pouvoir générer n'importe quelle nouvelle phrase.

Un peu d'histoire...

Avant de considérer la modalité visuelle de la parole en synthèse (c'est-à-dire le visage parlant), les recherches se sont portées sur *la synthèse de parole audio à partir du texte*, fournissant en sortie de manière synthétique la seule modalité audio pour la parole, soit le signal acoustique correspondant au texte.

L'idée première pour synthétiser de la parole audio consistait à modéliser les phénomènes à la base de la production de parole et d'en déduire des «règles». Le phénomène de production peut être modélisé par une source (les poumons), modulée ou non par les cordes vocales à l'entrée d'un système (le conduit vocal), à l'intérieur duquel des contraintes vont être réalisées pour mettre en forme le signal de parole en sortie. Cependant, le signal de parole a une forte variabilité inter- mais aussi intra-locuteur. La modélisation *par règles* ne permettant pas de capturer toute cette variabilité, une nouvelle voie de recherche est apparue avec le progrès technologique (notamment grâce aux capacités grandissantes de stockage en mémoire de données variées) : la synthèse *par concaténation* d'unités de signal. La variabilité du signal de parole peut alors être contenue dans les différents tronçons de signal stockés et reproduite/retransmise dans le produit de synthèse concaténé ; ceci permet en outre d'améliorer nettement le naturel et le réalisme du son.

Pour ce qui concerne la synthèse audiovisuelle de la parole, celle-ci a d'abord consisté à trouver des règles d'animation du visage en fonction du message véhiculé. De la même manière que pour la synthèse audio, les règles ne permettent pas de capturer de manière précise la variabilité des mouvements biologiques et en particulier celle des mouvements du visage. Tout naturellement, les systèmes basés sur la concaténation d'unités stockées se sont imposés progressivement pour leur efficacité à retransmettre le naturel des mouvements.

Cadre de ces travaux de thèse

Les données quantitatives concernant la surdité sont très difficiles à obtenir. En effet, les études généralement effectuées comptent ensemble différents types de handicaps. En outre, les études plus spécifiques à la surdité ne portent que sur une partie de la population sourde (ensemble de personnes scolarisées dans des établissements spécialisés, personnes touchant l'Allocation aux Adultes Handicapées (AAH), etc.). Enfin, il existe une variabilité importante dans la notion de surdité (légère, moyenne, sévère, profonde) et dans l'origine de la surdité (effet de l'âge, maladie, accident, etc.). Il en va de même avec les statistiques sur le mode de communication privilégié des personnes sourdes : le plus souvent, celles-ci utilisent l'oralité et la lecture labiale ; elles utilisent également la Langue française Parlée Complétée (LPC), la Langue des Signes Française (LSF), le Français Signé (FS)... Le chiffre le plus probant sur la population des

personnes sourdes est le nombre d'enfants nés sourds (à des degrés de surdité divers) par an en France : il est de 4000 pour un nombre de naissances de près de 800000 enfants, soit 0.5% des bébés.

Le choix de la Langue française Parlée Complétée par rapport à un autre mode de communication utilisé par les personnes sourdes se fonde sur plusieurs critères. Il s'agit de critères technologiques : le code LPC est directement relié à la langue française et donc les outils déjà développés dans le cadre de la transcription d'un texte en chaîne phonétique pourront être réutilisés (la Langue des Signes par exemple nécessitent des transcipteurs performants capables de passer d'un texte à une description gestuelle temporellement marquée [7, 4, 6]), de même au niveau de l'animation il s'agit d'animer un visage et une main dans un espace réduit près du visage (la génération de la Langue des Signes [5, 3, 2] nécessitent des gestes beaucoup plus complexes et amples que le code LPC), etc. Il existe également des raisons d'aide à la communication : ce système peut être rapidement appris par les parents d'enfants sourds (car basé sur le langage parlé) et permet de faciliter ainsi au quotidien la communication avec l'enfant et son insertion dans le monde des entendants. De plus, il est un augment à la lecture labiale qui est un autre vecteur de communication. C'est pour cela que nous allons nous attacher à produire du code LPC et non un autre de moyen de communication utilisé par les personnes sourdes.

Ce manuscrit présente la mise en oeuvre d'un système de synthèse de parole audiovisuelle *augmentée* d'une modalité manuelle codant la Langue française Parlée Complétée (code LPC) pour les personnes sourdes et malentendantes. Il s'agit d'un système de communication entièrement basé sur la parole (le langage parlé). Dans ce code, la main vient apporter une information complémentaire à celle fournie par la lecture labiale, qui est par nature ambiguë (la lecture sur les lèvres ne permet pas en effet de saisir la totalité du message oral).

Le but de nos travaux réside dans la synthèse d'un signal audio et de mouvements du visage et de la main cohérents et synchrones, c'est-à-dire le contrôle, la génération et la visualisation d'un visage virtuel parlant et codant en Langue française Parlée Complétée à partir d'un texte quelconque d'entrée. Dans la chaîne complète de synthèse, nous nous intéresserons plus particulièrement à la génération des paramètres d'animation du visage et de la main en fonction de la chaîne phonétique d'entrée. Ce système étant destiné à l'usage des sourds et malentendants décodant le code LPC, l'accent sera porté sur la génération de mouvements de visage et de main ancrée sur l'observation et l'analyse de codeurs humains, afin que celle-ci soit la plus proche possible d'un codage naturel. Dans ce but, nous analyserons les données réelles d'une codeuse LPC, puis implémenterons un système de génération de paramètres d'animation à partir de la chaîne phonétique et des informations prosodiques. Nous évaluerons ensuite ce système de génération en essayant de séparer les contributions de tous les modules qui composent un système de synthèse de parole audiovisuelle.

Le projet ARTUS

Ces travaux de thèse viennent s'insérer complètement dans le cadre du projet RNRT ARTUS (Animation Réaliste par Tatouage audiovisuel à l'Usage des Sourds) financé par le Réseau National de Recherche en Télécommunications (ce projet a été initié en même temps que nos

travaux de recherche). L'objectif de ce projet est de proposer aux personnes sourdes la possibilité d'afficher un codeur virtuel de Langue française Parlée Complétée en remplacement du télétexte dans les émissions télévisées. Le principe général repose sur l'insertion d'informations imperceptibles dans ces émissions TV, par des techniques de tatouage audio et vidéo (ces techniques permettent d'insérer à l'intérieur d'un signal vidéo ou audio une information additionnelle de manière indélébile et imperceptible, c'est-à-dire invisible en vidéo et inaudible en audio), qui seront interprétées à la réception et permettront ainsi d'animer le clone 3D codant en LPC [1].

L'intérêt général d'un tel projet réside clairement dans l'amélioration de l'accessibilité des systèmes d'information (en l'occurrence les émissions télévisées) proposée aux personnes souffrant d'un handicap sensoriel et permettant ainsi leur réhabilitation sociétale. A l'heure actuelle, ce sont en général des émissions sous-titrées qui leur sont proposées (on peut trouver également dans des émissions spécifiques des interprètes en Langue des Signes). La loi n° 2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées prévoit que les principales chaînes de télévision nationales rendent avant le 12 février 2010 la totalité de leurs programmes (hors écrans publicitaires) accessibles aux personnes sourdes et malentendantes. Or, la solution du télétexte, qui représente un bon moyen de retranscription, nécessite cependant d'avoir déjà un niveau de lecture moyen, que ne possèdent pas encore les enfants pré-lecteurs (voire même certains adultes sourds qui n'ont pas acquis la lecture de manière efficace). On peut donc espérer que l'application proposée dans le cadre de ce projet représente un réel bénéfice, améliorant dans le même temps la convivialité et l'usage de ce type de communication.

Ce projet implique de nombreuses collaborations nationales, tant dans le domaine de la recherche scientifique universitaire, que dans l'industrie et dans les télécommunications : LIS/INPG, ARTE, ATTITUDE STUDIO, HEUDIASYC/UTC, NEXTAMP, TSI/ENST et ICP/INPG. Chacun des partenaires apporte son expertise et son savoir-faire dans des domaines variés, allant entre autres du tatouage audio et vidéo à des problématiques d'animation et de rendu 3D en passant par de la synthèse de parole multimodale et par l'implémentation matérielle.

Ainsi, les travaux que nous présentons dans cette thèse constituent clairement la contribution de l'Institut de la Communication Parlée (ICP) en matière de synthèse multimodale de parole à partir d'information textuelle de type télétexte. Ils ont pour but principal de générer des paramètres d'animation d'un clone codant en LPC à partir de n'importe quel texte d'entrée. Il est important, cependant, de souligner qu'ils ne se limitent pas à cette application particulière, mais offre au contraire une multitude de possibilités (dans le cadre plus général des Interfaces Homme-Machine par exemple). En plus de lever un certain nombre de verrous technologiques et scientifiques, ces travaux constituent le premier synthétiseur 3D contrôlant le visage et la main d'un avatar virtuel codant en Langue française Parlée Complétée.

Plan du manuscrit

Ce manuscrit s'articule autour de trois parties. La première partie constitue un état de l'art non exhaustif de la synthèse de parole, qu'elle soit audio ou audiovisuelle, tant dans la façon de synthétiser que dans l'évaluation des systèmes de synthèse proposés. La deuxième partie

décompose les mécanismes nécessaires à l'obtention d'un synthétiseur : (1) une première phase consiste à déterminer et à enregistrer un corpus, (2) une deuxième phase consiste à le prétraiter et à l'analyser, (3) à l'aide des informations obtenues, la troisième phase met en oeuvre la synthèse proprement dite c'est-à-dire dans notre cas, la génération du signal acoustique et le contrôle du visage et de la main. Enfin, la troisième partie correspond à l'évaluation, tâche bien souvent sous-estimée et passant parfois en arrière-plan des systèmes de synthèse proposés dans la littérature. Cette évaluation donnera quelques points d'ancrage pour estimer les points forts et faibles du synthétiseur et ainsi proposer des perspectives d'amélioration du système.

Références bibliographiques

[1] G. Bailly, V. Attina, C. Baras, P. Bas, S. Baudry, D. Beautemps, R. Brun, J.-M. Chassery, F. Davoine, F. Elisei, G. Gibert, L. Girin, D. Grison, J.-P. Léoni, J. Liénard, N. Moreau, and P. Nguyen. ARTUS : calcul et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue française Parlée Complétée. In *Handicap 2006*, Paris, France, 2006.

[2] A. Héloir, S. Gibet, N. Courty, J. F. Kamp, N. Rezzoug, P. Gorce, F. Multon, and C. Pelachaud. Agent virtuel signeur aide à la communication pour personnes sourdes. In *Handicap 2006*, 2006.

[3] R. Kennaway. Synthetic animation of deaf signing gestures. In *Gesture Workshop*, pages 146–157, 2001.

[4] R. Kennaway. Experience with and requirements for a gesture description language for synthetic animation. In *Gesture Workshop*, pages 300–311, 2003.

[5] T. Lebourque and S. Gibet. A complete system for the specification and the generation of sign language gestures. In A. Braffort, R. Gherbi, S. Gibet, J. Richardson, and D. Teil, editors, *Lecture Notes in Artificial Intelligence, LNAI 1739, in Gesture-Based Communication in Human-Computer Interaction*, pages 227–238. Springer-Verlag, 1999.

[6] I. Marshall and E. Sáfár. Grammar development for sign language avatar-based synthesis. In *3rd International Conference on UA in HCI : Exploring New Dimensions of Diversity*, volume 8, Las Vegas, Nevada, 2005.

[7] E. Sáfár and I. Marshall. The Architecture of an English-Text-to-Sign-Languages Translation System. In G. Angelova et al., editor, *Recent Advances in Natural Language Processing (RANLP)*, pages 223–228. Tzigov Chark Bulgaria, 2001.

Première partie

Etat de l'art

Chapitre 1

Choix scientifiques

Le but de ces travaux de thèse est de mettre en oeuvre un système de synthèse 3D de Langue française Parlée Complétée à partir du texte. Plus particulièrement, nous allons nous intéresser à la génération des paramètres d'animation de la main et du visage à partir de la chaîne phonétique marquée en temps de la phrase d'entrée. Nous disposons pour cela d'un système support capable de générer un signal acoustique de parole à partir de n'importe quelle phrase d'entrée, le système COMPOST [5, 3, 2, 1]. Ce système sera décrit plus en détails dans le chapitre consacrée à la synthèse de code LPC de la deuxième partie. Toutefois, notons dès à présent qu'en plus de délivrer un signal acoustique correspondant à la phrase prononcée par une méthode de concaténation de polysons, il délivre la chaîne phonétique marquée en temps grâce aux modules de traitements linguistiques et prosodique [13]. À l'aide d'un test perceptif nous allons déterminer quel type de contrôle moteur est le plus adapté à la génération de code LPC. Puis, nous allons voir comment ajouter des modules à ce système existant afin qu'il puisse produire plus qu'un signal acoustique.

1.1 Paradigme d'animation

L'idée de base que nous avons retenue pour notre système de génération de code LPC est qu'il doit être capable d'atteindre les performances d'un être humain. En effet, le signal de référence reste le codeur humain. Afin de déterminer parmi un ensemble de systèmes de génération existants celui qui est capable de «cloner» au mieux l'humain, nous avons mis en oeuvre une évaluation de l'adéquation des mouvements faciaux avec le signal audio correspondant [4]. En effet, même si les réalisations de la communauté de l'animation sont remarquables, les lois du mouvement biologique vivant nous échappent encore. Afin de passer outre ces problèmes de modélisation du contrôle moteur, une solution est de capturer ce mouvement à l'aide de techniques de «MoCap»[1] afin de l'étudier et de le synthétiser. En outre, ces «données terrain» permettent une évaluation comparative probante entre des stimuli originaux et des stimuli des synthèse.

[1]Motion Capture : système de capture de mouvements capable de délivrer les trajectoires de marqueurs positionnés sur une personne par exemple.

1.2 Expérience préliminaire

1.2.1 Paradigme de test

Afin de déterminer le modèle de contrôle le plus à même de «cloner» les mouvements humains, nous avons mis en place un test perceptif visant à évaluer l'adéquation des mouvements du visage par rapport à un signal audio donné et ce pour un ensemble de modèles de contrôle de la littérature. Les systèmes de synthèse de parole audiovisuelle se compose en général de 3 modules :

- un module de contrôle moteur qui génère du mouvement par rapport à la chaîne phonétique ;
- un module de forme qui déclare la variation de la géométrie du visage par rapport aux mouvements ;
- un module d'apparence qui spécifie comment cette géométrie affecte l'apparence du visage.

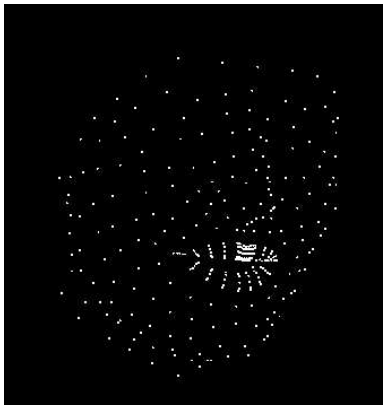


FIG. 1.1 – Distribution des points lumineux sur le visage de la locutrice.

Le but de notre expérience est de tester le premier module c'est-à-dire celui capable de fournir du mouvement à partir d'une information linguistique marquée en temps. Afin de séparer les contributions de chacun des modules et de tester uniquement le premier module nous utilisons le paradigme des points lumineux (voir figure 1.1). Il a en effet été montré que l'observation d'au moins 100ms de mouvements sous forme de points lumineux était suffisante pour identifier la tâche humaine sous-jacente accomplie [14]. Cette sensibilité au mouvement biologique semble quasi-innée ou tout du moins extrêmement précoce [6, 7]. Cette sensibilité a également été montrée pour la détermination du genre de la personne [8] et pour des actions complexes (saisie, jet d'objets mais aussi danse...) [9, 11, 10]. De plus, cette sensibilité est très fine puisque même un mime professionnel n'est pas capable de tromper des observateurs sur le poids réel d'un objet qu'il transporte [24, 25].

Des expériences utilisant ce paradigme des «points lumineux» ont été effectuées sur de la parole audiovisuelle. Des observateurs non entrainés sont capables de faire de la lecture labiale de voyelles, de syllabes et de mots courts sur un visage en points lumineux [22] avec des bénéfices en terme de rapport SNR (Signal to Noise Ratio) presqu'identiques à la même tâche sur un visage en vidéo complète [21]. L'effet Mc Gurk [15] a même été reproduit avec des stimuli en points luminuex [23]. Ainsi, même si cette présentation des stimuli ne contient aucune information sur

les critères faciaux tels que la peau, les dents ou les ombres produites par les mouvements de la bouche, elle produit des stimuli «purement de mouvement» dont l'information visuelle s'intègre parfaitement avec le signal acoustique correspondant.

1.2.2 Les stimuli

Le test perceptif compare les trajectoires articulatoires naturelles avec divers systèmes de génération de mouvement de la littérature. Toutes ces trajectoires articulatoires pilotent le même modèle de forme [20].

Enregistrement

Nous avons enregistré une locutrice française sur laquelle avaient été préalablement collées 245 billes colorées (voir figure 1.2) lors de la production de 120 visèmes (qui permettent de construire le modèle de forme), de 96 logatomes de type VCV (où C est une consonne parmi les 16 consonnes du français et V une des 6 voyelles /a,/i,/u,/e,/ø,/o/) et de 76 phrases phonétiquement équilibrées. Cet enregistrement est fait à l'aide d'un système vidéo multi-caméras qui délivre des séquences d'images non compressées à 50 Hz.

FIG. 1.2 – Position des 245 billes colorées placées sur le visage de la locutrice.

Analyse

À l'aide d'une procédure d'analyse par la synthèse [12]où la distance RMS (Root Mean Square) entre chaque image et un simple modèle d'apparence utilisant un mélange «blending/morphing» de 3 textures est utilisée dans une procédure bouclée d'estimation. Une technique d'optimization basée sur un simplex [18, 17] permet d'estimer le jeu de paramètres articulatoires qui minimise la distance RMS pour chaque trame. Nous avons donc à disposition un ensemble de paramètres articulatoires pilotant un modèle de forme pour chaque trame (à 50 HZ) de notre corpus (visèmes, logatomes et phrases).

Synthèse

Parmi l'ensemble des 76 phrases du corpus, nous en choisissons 10 pour lesquelles les diphones qu'elles contiennent se trouvent au moins une fois présents dans les 66 autres. Ces 10 phrases seront notre ensemble de test tandis que les 96 VCV et les 66 autres phrases seront notre ensemble d'apprentissage. Il est à noter que tous les stimuli (d'entraînement et de test) ont été segmentés et étiquettés à la main et que tous les systèmes de génération ont ces informations ainsi que les trajectoires articulatoires et le signal audio à leur disposition. Les 5 systèmes de génération que nous avons testés sont :

1. *Syn* est un système basé sur la concaténation de diphones audiovisuels. Les diphones sont multi-représentés et les diphones cibles sont selectionnés à l'aide d'une technique de programmation dynamique où la distance est la distance RMS entre les paramètres LSP (Line Spectrum Pairs) du signal audio aux frontières inter-diphones. Il est à noter qu'aucune distance articulatoire n'est prise en compte ici. La technique LPPSOLA [16] est utilisée pour manipuler le signal audio.
2. *Synl* est semblable au système précédent à l'exception de l'ajout d'un lissage anticipatoire (interpolation linéaire) sur les paramètres articulatoires pour éviter les sauts aux frontières inter-diphones.
3. *Reg* est un système qui génère des trajectoires articulatoires en utilisant le modèle de coarticulation d'Öhman [19]. Dans ce modèle, les mouvement rapides de fermeture des consonnes sont superposés aux mouvements articulatoires plus lents des voyelles.
4. *Mltst* est un système qui calculent les trajectoires articulatoires à partir des signaux acoustiques. Une régression linéaire relie les trajectoires LSP filtrées à 10 Hz avec les mouvements articulatoires [26] des 66 phrases du corpus d'apprentissage.
5. *Mlapp* est semblable au système précédent à l'exception du corpus d'apprentissage qui devient les phrases de test.

A ces 5 systèmes de génération, nous ajoutons les trajectoires originales *Org* et les trajectoires originales opposés *OrgInv* (inversion des signes des paramètres articulatoires originaux). Ces deux ajouts permettent d'étalonner notre évaluation puisque les stimuli *Org* sont parfaits (dans la limite où l'on suppose la capture du mouvement parfaite) et les stimuli *OrgInv* sont totalement incohérents.

1.2.3 Procédure

Protocole

Toutes les animations sont jouées à partir des fichiers de paramètres articulatoires pré-calculés et couplées au son original. La présentation du visage se fait sous forme de points lumineux blancs sur un fond noir (voir figure 1.1). La fenêtre a une taille de 576x768 pixels sur un écran de 17". L'interface du test a été créée sous Matlab®GUI. Le temps de réponse entre l'apparition du visage et la prise de décision des sujets est sauvegardé mais ceux-ci ne le savent pas.

Tâche

La tâche correspond à un test de type MOS (Mean Opinion Score). Il est demandé aux sujets de noter sur une échelle de 5 valeurs (Très bien , Bien, Moyen, Insuffisant, Très insuffisant) le degré de cohérence entre le son et les mouvements faciaux de 10 phrases prononcées suivant 7 systèmes de génération différents (*Org*, *OrgInv*, *Syn*, *Synl*, *Reg*, *Mlapp*, *Mltst*). Aucun mouvement de tête n'est ajouté et la tête est présenté de face.

Sujets

Les sujets sont au nombre de 23. Ils sont tous de langue maternelle française et «naïfs» par rapport à la tâche qui leur est demandée. Aucun problème visuel, ni auditif n'est rapporté par les sujets.

Déroulement

Avant le test, la tête parlante sous forme de points lumineux est présentée avec un déplacement de type rotation de $\pm$ 45 $^\circ$ par rapport à la vue de face autour d'un axe vertical situé derrière la tête. Le mouvement est lent (1 Hz) et pendant celui-ci le visage prononce la plus longue phrase du corpus avec des mouvements naturels. Tous les sujets rapportent voir un visage naturel en mouvement.

Ensuite, une phase d'adaptation aux stimuli est mis en oeuvre. Trois stimuli tirés aléatoirement pour chaque sujet dans les stimuli de test sont présentés et les sujets doivent noter l'adéquation de ceux-ci. Les résultats correspondant à ces stimuli ne sont pas conservés.

Enfin, c'est la phase de test à proprement parlé avec la présentation en ordre aléatoire des 10 phrases sous les 7 conditions soit 70 phrases.

1.2.4 Résultats

Les données enregistrées sur chaque stimuli sont la note mais également le temps de réponse (information inconnue des sujets). Cette seconde donnée nous permet d'avoir une idée de la charge cognitive mis en jeu par chacun des systèmes de génération. Les résultats pour les 2 types de données sont représentés sous forme de boîtes à moustaches sur la figure 1.3. On peut remarquer de façon qualitative que l'on retrouve nos systèmes de contrôle que sont *Org* et *OrgInv* aux extrémités de l'échelle de classification. Avant d'aller plus loin dans l'analyse statistique des données, nous devons définir notre seuil de signification : $\alpha = 0.01$.

Les notes

Les distributions des notes par systèmes ne suivent pas des lois normales ($\max(W) < 0.91$, $\max(p) = 10^{-10}$, *test de Shapiro-Wilk*). Nous allons utiliser par conséquent des tests non paramétriques basés sur le rang, notre échelle ordinale étant (Très bien > Bien > Moyen > Insuffisant > Très insuffisant).

L'hypothèse *H0* que nous posons est la suivante : *les notes sont issues de la même population.* Le test des rangs de Kruskal et Wallis ($H = 780.126$, 6 ddl, $p < 10^{-15}$) nous permet d'affirmer

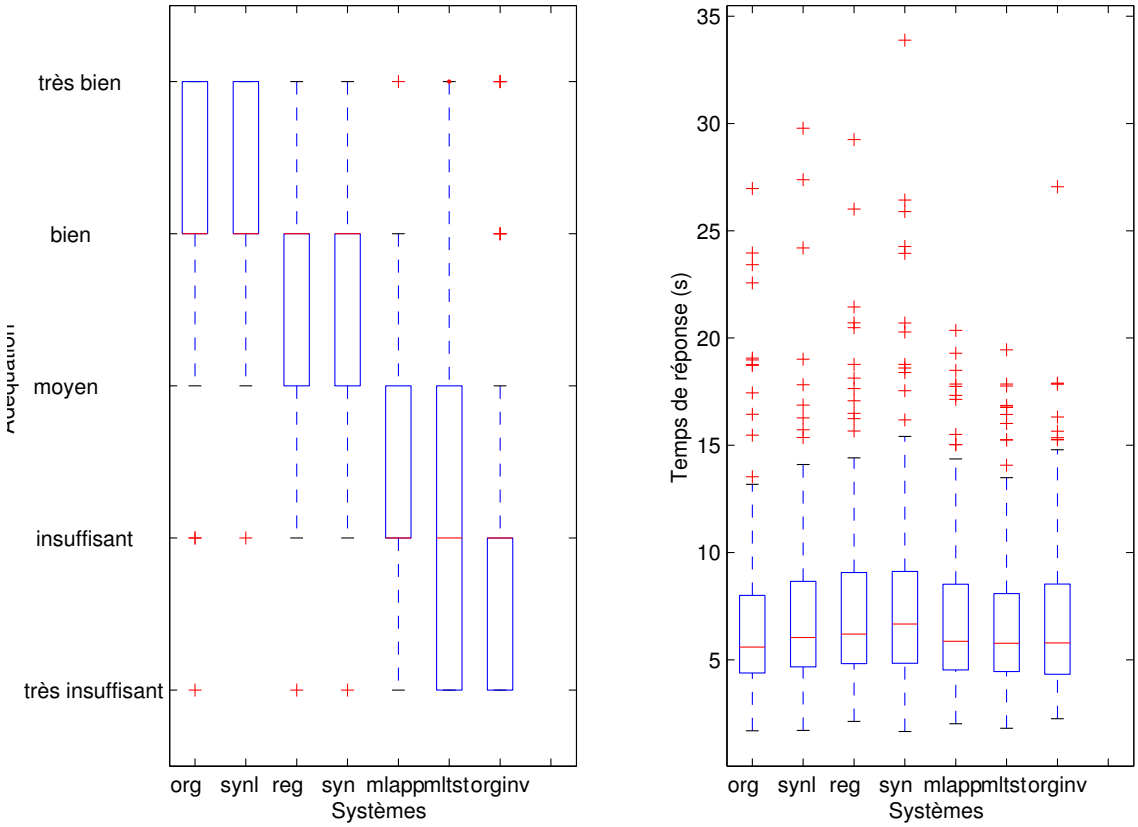


FIG. 1.3 – Résultas sous formes de boîtes à moustaches de l'expérience «points lumineux». À gauche : score MOS pour chacun des systèmes. À droite : temps de réponse pour chacun des systèmes.

le rejet de l'hypothèse nulle et donc par conséquent que les notes ne sont pas issues de la même population. Une analyse plus fine c'est-à-dire inter-systèmes à l'aide du test de Wilcoxon nous permet de regrouper les systèmes de la sorte :

1. groupe 1 : *Org*, *Synl*
2. groupe 2 : *Syn*, *Reg*
3. groupe 3 : *Mltst*
4. groupe 4 : *Mlapp*, *OrgInv*

Il n'y a pas de différence significative entre les systèmes d'un même groupe : le système de synthèse par concaténation avec lissage anticipatoire *Synl* est donc au même niveau que les trajectoires originales *Org*. Le système de génération par règles *Reg* délivre des trajectoires hyperarticulées du fait de la base d'apprentissage utilisée. Ce système a été bien noté mais les sujets ont rapporté avoir vu ces stimuli comme adéquats par rapport à la tâche phonétique mais peu naturels. Cette ambiguïté se retrouve d'ailleurs dans les temps de décision qui sont plus longs pour ce système. Il est intéressant de noter que des systèmes tels que *Mltst* et *Mlapp* sont très mal notées alors qu'en terme de corrélation avec les stimuli originaux ils sont les meilleurs.

Les temps de réponse

Si on observe qualitativement les temps de réponse par système, on se rend compte de légères différences. Comme précédemment, nous vérifions que les temps de réaction ne suivent pas une loi normale. Nous effectuons par conséquent un test non paramétrique : le test des rangs de Kruskal et Wallis. Les résultats ($H = 13.617$, 6 ddl, $p = 0.03422 > \alpha$) ne montrent pas de différence significative entre les temps de réponse par système.

1.3 Conclusions

Les résultats de cette expérience montrent que les sujets sont très sensibles à la cohérence des mouvements faciaux par rapport au signal acoustique correspondant. Les résultats montrent également qu'un système de concaténation d'unités audiovisuelles pré-stockées avec un simple lissage anticipatoire est capable de générer des mouvements de grande qualité. En résumé, la perception audiovisuelle de la parole est très sensible au phasage des évènements ce que la concaténation préserve, ce que les modèles de coarticulation simplifie trop et ce que l'inversion acoustique-articulatoire a peu de chance de récupérer.

Au vu de ces résultats, nous choisissons comme paradigme de génération de mouvements l'enregistrement de données réelles et la **synthèse par concaténation d'unités multimodales pré-stockées**. Cette méthode est la plus adéquate à accompagner le signal audio et à respecter les rendez-vous importants caractéristiques de la parole multimodale.

Nous allons adapter le système de synthèse existant COMPOST afin qu'il puisse générer des mouvements du visage et de la main en accord avec la Langue française Parlée Complétée. Dans la suite de cette partie, nous allons présenter la synthèse de parole audiovisuelle en décomposant le problème : nous présenterons la synthèse de parole audio, puis nous expliquerons comment se greffe un module visuel sur les systèmes précédents, enfin nous présenterons les systèmes de génération de code LPC existants et donc l'ajout d'une information visuelle sur un système de synthèse de parole audiovisuelle. Le choix de décomposer le problème ainsi permet de positionner dans un état de l'art non exhaustif les éléments essentiels de notre système de synthèse de Langue française Parlée Complétée.

Références bibliographiques

[1] M. Alissali. *Architecture logicielle pour la synthèse multilingue de la parole*. PhD thesis, INPG, Grenoble, France, 1993.

[2] M. Alissali and G. Bailly. Compost : a client-server model for applications using text-to-speech. In *European Conference on Speech Communication and Technology*, pages 2095–2098, Berlin, Germany, 1993.

[3] G. Bailly and M. Alissali. Compost : a server for multilingual text-to-speech system. *Traitement du Signal*, 9(4) :359–366, 1992.

[4] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

[5] G. Bailly and A. Tran. Compost : a rule-compiler for speech synthesis. In *European Conference on Speech Communication and Technology*, pages 136–139, 1989.

[6] B. I. Bertenthal, D. R. Proffitt, and J. E. Cutting. Infant sensitivity to figural coherence in biomechanical motions. *Journal of Experimental Child Psychology*, 37 :213–230, 1984.

[7] B. I. Bertenthal, D. R. Proffitt, and S. J. Kramer. Perception of biomechanical motions by infants : Implementation of various processing constraints. special issue : The ontogenesis of perception. *Journal of Experimental Psychology : Human Perception and Performance*, 13 :577–585, 1987.

[8] J. E. Cutting, D. R. Proffitt, and L. T. Kozlowski. A biomechanical invariant for gait perception. *Journal of Experimental Psychology : Human Perception and Performance*, 4, 1978.

[9] W. H. Dittrich. Actions categories and recognition of biological motion. *Perception*, 22 :15–23, 1993.

[10] W. H. Dittrich. *Lecture Notes in Artificial Intelligence : Gesture-Based Communication in Human-Computer Interaction*, chapter Seeing biological motion - Is there a role for cognitive strategies ?, pages 3–22. A. e. a. Braffort, Berlin, 1999.

[11] W. H. Dittrich, T. Troscianko, S. E. G. Lea, and D. Morgan. Perception of emotion from dynamic point-light displays represented in dance. *Perception*, 25 :727–738, 1996.

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

[13] B. Holm. *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application à l'énonciation de formules mathématiques.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2003.

[14] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14 :201–211, 1973.

[15] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 26 :746–748, 1976.

[16] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 :453–468, 1990.

[17] M. Odisio. *Estimation des mouvements du visage d'un locuteur dans une séquence audiovisuelle.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.

[18] M. Odisio, G. Bailly, and F. Elisei. Talking face tracking with shape and appearance models. *Speech Communication*, 44(1-4) :63–82, October 2004.

[19] S.E.G. Öhman. Numerical model of coarticulation. *JASA*, 41(2) :310–320, 1967.

[20] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[21] L. D. Rosenblum, J. A. Johnson, and H. M. Saldaña. Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research*, 39(6) :1159–1170, 1996.

[22] L. D. Rosenblum and H. M. Saldaña. Time-varying information for visual speech perception. In R. Campbell, B. Dodd, and D. Burnham, editors, *Hearing by eye : Part 2, The Psychology of Speechreading and Audiovisual Speech*, pages 61–81. Earlbaum : Hillsdale, 1998.

[23] L. D. Rosenblum, L. D. Schumuckler, and J. A. Johnson. The McGurk effet in infants. *Perception & Psychophysics*, 59(3) :347–357, 1997.

[24] S. Runeson and G. Frykholm. Visual perception of lifted weight. *Journal of Experimental Psychology : Human Perception and Performance*, 7 :733–740, 1981.

[25] S. Runeson and G. Frykholm. Kinematic specification of dynamics as an informational basis for person and action perception : Expectation, gender recognition and deceptive intention. *Journal of Experimental Psychology : General*, 112 :585–615, 1983.

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

Chapitre 2

Synthèse audio : du texte au signal acoustique

Les travaux présentés dans ce manuscrit traitent d'un système de synthèse de parole audio-visuelle augmentée. L'accent ne sera pas porté sur les spécificités des systèmes de synthèse de parole audio. Toutefois, notre système restant basé sur un système de synthèse de parole audio, nous décrirons brièvement les modules composant un tel système afin que le lecteur situe les options choisies par rapport à un système plus global.

La synthèse de parole à partir du texte vise à donner la possibilité aux (micro-)ordinateurs d'émettre des sons de parole à partir de n'importe quel texte tapé au clavier (ou reconnu par un système de reconnaissance de caractères...). L'entrée d'un tel système est un signal textuel (une chaîne de caractères alphanumériques) et la sortie un signal audio.

Ces systèmes peuvent se décomposer en plusieurs modules. De manière très générale, nous considérons une décomposition en trois modules : un module de traitements linguistiques qui va nous fournir une chaîne phonétique ainsi qu'une structure grammaticale, un module de traitements prosodiques qui va nous fournir des durées et des variations de fréquence fondamentale et enfin un module générateur du signal acoustique.

Cette description pouvant s'adapter à tous les types de systèmes de synthèse de parole, notre choix s'est porté sur un type de systèmes plus particuliers, les systèmes de synthèse par concaténation d'unités multimodales. La figure 2.1 représente le schéma synoptique d'un tel synthétiseur.

Nous allons maintenant détailler chacun de ces modules, en particulier leur structure et leur fonctionnement. Pour une revue complète de la synthèse de parole audio à partir du texte en français le lecteur pourra se référer à [10, 14, 48, 49, 6].

2.1 Les traitements linguistiques

Ce module prend en entrée une suite de caractères alphanumériques de la phrase à synthétiser et donne en sortie la chaîne de symboles phonétiques ainsi que la structure grammaticale.

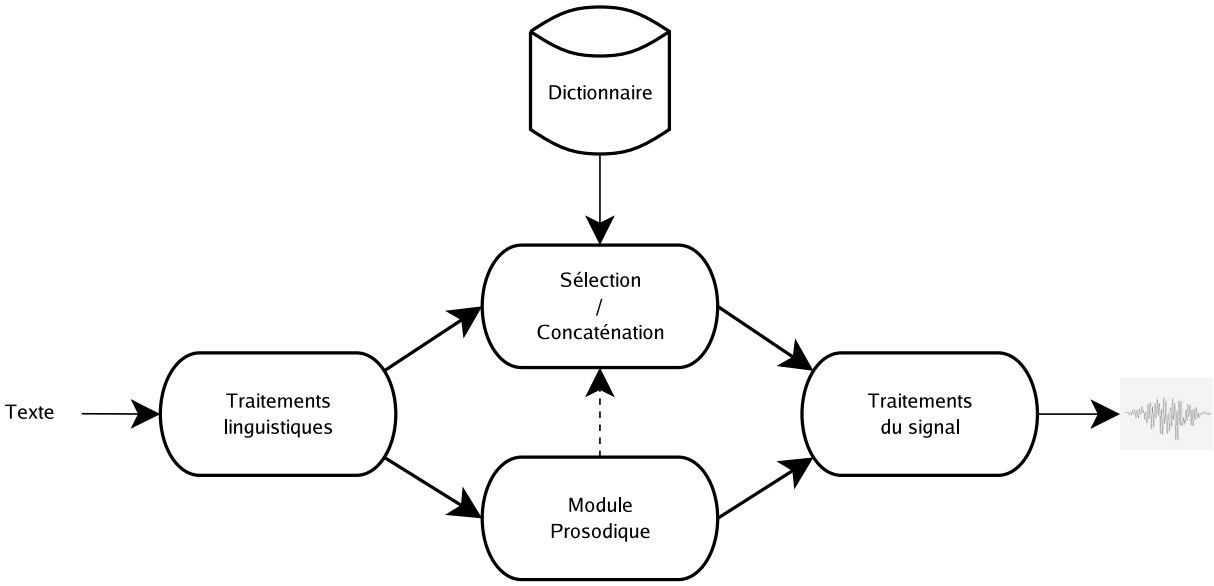


FIG. 2.1 – Schéma synoptique d'un système de synthèse de parole par concaténation d'unités.

Pourquoi ce module est-il nécessaire ?

La suite de caractères alphanumériques que l'on fournit en entrée du système de synthèse, bien que l'on suppose qu'elle soit exempte de toute erreur de frappe, de faute de grammaire, d'orthographe ou de conjugaison, présente des ambiguïtés. Ces ambiguïtés peuvent être de nature très diverse, citons par exemple celle liée aux homographes hétérophones : il s'agit de mots ayant la même orthographe mais une prononciation différente, comme par exemple la suite de lettres *couvent* peut être un nom masculin singulier se prononçant [k u v ɑ̃] ou bien le verbe couver conjugué à la 3e personne du pluriel du présent de l'indicatif se prononçant alors [k u v]. Il est donc nécessaire de traiter cette chaîne alphanumérique afin de connaître la structure de la phrase et les éléments qui la composent.

Les ambiguïtés étant nombreuses, ce module va se décomposer en un ensemble de sous-modules ayant une fonction propre. Nous allons rentrer plus en détails dans les traitements nécessaires à la désambiguïsation de l'entrée textuelle. Nous verrons les pré-traitements des éléments non-lexicaux, l'analyse morphologique, la recherche lexicale, l'analyse morpho-syntaxique et enfin la transcription orthographique-phonétique.

Pré-traitement des éléments non-lexicaux

Le premier problème rencontré par le système de transcription orthographique-phonétique est la transcription des éléments non-lexicaux. Ceux-ci peuvent être de nature différente : nombre, date, abréviation, acronyme... La fonction de ce sous-module est de transcrire en toutes lettres ces chaînes non-orthographiques ou non composées uniquement de caractères orthographiques. Les solutions sont l'utilisation de lexiques spécifiques, de règles heuristiques ou de grammaires régulières.

Analyse morphologique

Le but de ce sous-module est de décomposer les différents lexèmes (mots composant la phrase d'entrée) en composantes élémentaires. On détermine ainsi les morphèmes[1] qui correspondent aux préfixes, suffixes, désinences et racines des lexèmes [29]. Cette analyse s'effectue essentiellement sur les mots lexicaux en nombre *a priori* infini, alors que les mots grammaticaux (déterminants, pronoms, prépositions, conjonctions) en nombre fini seront traités comme des unités lexicales (cf. section suivante).

Les informations délivrées par ce sous-module vont permettre de déterminer des catégories grammaticales (par exemple grâce aux désinences). Dès lors, le lexique ne contient plus que des racines et des exceptions (mots grammaticaux par exemple).

Recherche lexicale

Lors de la recherche lexicale, on compare les différents lexèmes composant la phrase d'entrée avec ceux contenus dans le lexique. En général, le lexique contient des lexèmes et pour chacun d'eux, les catégories grammaticales possibles (nom, adjectif, verbe, etc.), les propriétés grammaticales (genre, nombre, conjugaison, etc.) et la transcription phonétique c'est-à-dire la suite de phonèmes correspondant au lexème.

Analyse morpho-syntaxique

Dès lors, l'information que l'on possède est une suite de lexèmes appartenant à une ou plusieurs catégories et/ou propriétés grammaticales. Il reste encore des ambiguïtés à lever avant d'obtenir la chaîne phonétique correspondant à la phrase. On va utiliser le contexte pour déterminer quelle catégorie et quelle propriété grammaticale correspond à chaque léxème. Il existe deux types d'analyseur contextuel, ceux basés sur des règles heuristiques, déduites de règles grammaticales et ceux basés sur des règles probabilistes. Les grammaires locales déterministes possèdent un avantage sur les approches probabilistes : les règles grammaticales sont explicites et peuvent donc être modifiées facilement en cas d'erreurs avérées. En revanche, les taux d'étiquetage correct sont en général inférieurs à ceux des analyseurs probabilistes.

On retrouve deux types d'analyseurs probabilistes : les n-grammes [28] et les réseaux de neurones [5]. Dans le premier cas, on calcule explicitement les probabilités de transitions entre catégories grammaticales en tenant compte de dépendances à plusieurs niveaux. En général, ce sont les tri-grammes qui sont le plus utilisés [25], c'est-à-dire ceux qui tiennent compte, étant donné un mot, de son prédécesseur et du prédécesseur de ce dernier. Dans le cas des réseaux de neurones, le calcul des probabilités entre catégories grammaticales est implicite ; il va se retrouver caché au niveau des poids du réseau de neurones. La taille des fenêtres contextuelles utilisées est en général plus grande que dans le cas des n-grammes.

[1] En linguistique, on définit généralement un morphème comme la plus petite unité de son, porteuse de sens qu'il est possible d'isoler dans un énoncé.

Transcription orthographique-phonétique

A l'issue des traitements précédents, nous avons une suite de morphèmes qui correspondent à la phrase prononcée. Pourtant le passage à la phonétique n'est pas trivial : la traduction phonétique associée à chacun des morphèmes correspond à une vision isolée de ceux-ci. Or, dans le cas d'une phrase, de nouvelles contraintes apparaissent. Ainsi, des contraintes de type articulatoire vont venir modifier certains phonèmes en fonction des phonèmes voisins, des liaisons vont venir modifier la prononciation de certaines terminaisons de mots et il subsiste encore des ambiguïtés quant au choix du morphème dans le cas d'homographes hétérophones de même catégorie grammaticale.

Pour phonétiser les morphèmes, il existe deux approches : l'approche basée dictionnaire et l'approche basée règles morphophonologiques. Dans le premier cas, on utilise un lexique contenant le maximum de morphèmes (formes canoniques et fléchies) avec leurs prononciations. Alors que dans le deuxième cas [4], on déduit la prononciation des formes fléchies à partir des morphèmes (racines morphologiques) du lexique.

Dans les deux cas, on applique par la suite un post-traitement phonologique, qui consiste en des règles de ré-écriture de la chaîne phonétique pour la gestion des liaisons et des contraintes articulatoires.

2.2 Le module prosodique

Le module précédent nous a fourni la chaîne phonétique correspondant à l'entrée textuelle, mais il ne nous a donné aucune information quant à la structure temporelle, spectrale et énergétique que doit avoir le signal audio de sortie. C'est le module prosodique qui va s'en charger.

Pourquoi ce module est-il nécessaire ?

Considérons par exemple la phrase «ça va». On peut modifier le sens de cette phrase (déclaratif, interrogatif...), en ne changeant que la prononciation. Ces différences de production vont s'effectuer au niveau spectral par une variation de la fréquence fondamentale, et/ou au niveau temporel par une modification de la longueur des syllabes, et/ou au niveau énergétique. Le module prosodique a donc pour mission de fournir les caractéristiques que doit posséder notre signal acoustique pour remplir des fonctions de segmentation, d'emphase... Nous avons vu que ces caractéristiques pouvaient modifier le sens de la phrase. Elles permettent également de rendre le signal de synthèse plus «naturel», en copiant ce que l'humain serait susceptible de faire en pareille situation.

Ce module va donc fournir la variation de la fréquence fondamentale, la longueur des syllabes et la variation de l'énergie du signal en fonction de la structure prosodique de la phrase. Nous allons voir comment est découpée la phrase en groupes prosodiques, puis comment sont générées les caractéristiques du signal en fonction de ce découpage. Pour une revue plus exhaustive, le lecteur pourra se référer à [50].

Segmentation prosodique

Une phrase peut être découpée au niveau syntaxique (détermination du syntagme sujet, syntagme verbal...) mais aussi prosodique (détermination des groupes prosodiques). L'évolution des paramètres prosodiques à l'intérieur de ces groupes va dépendre de plusieurs variables (rôle du groupe dans la phrase, dépendances par rapport aux autres groupes, nombre de syllabes du groupe, etc.). Il existe différentes méthodes de segmentation en groupes prosodiques d'une phrase donnée. Comme pour les traitements linguistiques vus dans la section précédente, s'opposent les approches «basées règles» (règles qui peuvent être de simples heuristiques ou déduites d'analyses morphosyntaxiques) aux approches «basées apprentissage». Les heuristiques sont organisées automatiquement en arbres de décision par des méthodes de classification et de régression [9].

Génération des caractéristiques prosodiques

Une fois la segmentation prosodique de la phrase déterminée, il reste à générer les caractéristiques prosodiques du signal de synthèse. Nous allons voir plus particulièrement comment sont générés les contours mélodiques et les durées des phonèmes.

Caractéristiques fréquentielles

Le contour mélodique est défini par l'évolution de la fréquence fondamentale F0. Pour générer cette évolution en fonction de la structure prosodique de la phrase, il existe trois grandes voies [3] : les modèles de commande mélodique, l'utilisation de contours pré-stockés et l'utilisation de modèles statistiques.

Les modèles de commande mélodique se déclinent sous trois formes : le modèle de commande de source vocale, le modèle par points cibles et le modèle type *école hollandaise*. Dans le cas des modèles à commande mélodique, ce sont des règles (déduites de l'analyse) qui seront utilisées pour synthétiser la variation de la fréquence fondamentale.
Le modèle de source vocale décompose les variations de F0 en deux types de composantes mélodiques : les composantes d'accents et les composantes de groupes. Ces commandes sont modélisées comme des réponses à des systèmes linéaires à différents types d'entrée puis sont superposées pour retrouver le contour mélodique [21, 20].
Dans le modèle par points cibles, on ne s'intéresse qu'aux extrema de la fréquence fondamentale puisqu'on suppose, dans ce modèle, que l'information mélodique est contenue essentiellement à ces endroits. On définit les transitions entre ces points cibles par des fonctions déduites de règles [38].
Le dernier modèle dit de *l'école hollandaise* (développé à l'Institut voor Perceptieonderzoek dans les années 1960) consiste à styliser la courbe mélodique par des segments de droites, puis à classer ces segments de droites parmi un ensemble fini de contours standards.

Les modèles utilisant des contours pré-stockés ne cherchent pas à décrire les variations mélodiques par des règles mais à utiliser des dictionnaires de formes issus de corpus de parole [47, 17, 1]. Pour cela, on construit un ensemble de classes reliant groupes syntaxiques (genre, longueur) et contours mélodiques (moyenne des contours mélodiques d'une même classe dans

le corpus de base). Dans la phase de synthèse, on génère, à l'aide de réseaux de neurones, des contours globaux [34] ou locaux [23] pour chaque groupe de la phrase que l'on superpose ensuite pour obtenir le contour mélodique final.

Les modèles statistiques réalisent quant à eux une cartographie implicite des relations entre caractéristiques linguistiques et caractéristiques prosodiques. Dans [42] par exemple, un réseau de neurones relie des paramètres linguistiques (type d'accent de la phrase, longueur de la phrase, etc.) à trois valeurs de F0, pour un autre exemple utilisant les réseaux de neurones voir [45]. Des modèles de Markov cachés (HMM) peuvent également être utilisés dans ce cas [19].

Caractéristiques temporelles

Les caractéristiques temporelles correspondent au rythme de la phrase incluant la longueur des durées des phonèmes la constituant mais aussi l'insertion des pauses, le débit de la phrase, etc. Nous allons nous limiter à la description de modèles de prédiction des durées des phonèmes.

En général, les modèles de durée se basent sur la durée intrinsèque du phonème considéré (calculée comme la moyenne de la durée de ce phonème sur un gros corpus) qu'ils modifient à l'aide de facteurs qui dépendent du contexte syntaxique et prosodique du phonème. Le modèle de Bartkova et Sorin [2] en est un exemple, lui-même dans la lignée des modèles de Klatt [26] et de O'Shaughnessy [37].

Et aussi...

Il existe également des systèmes de génération de la variation de l'énergie du signal en fonction de la structure prosodique [46].

Les systèmes précédemment cités séparent explicitement les paramètres prosodiques à générer mais l'apparition d'outils de *mapping*[2] performants a vu l'avènement de modèles prosodiques multi-paramétriques [33, 43].

2.3 La génération du signal acoustique

A ce niveau, nous avons en notre possession toutes les informations (obtenues à partir des modules linguistiques et prosodiques) nécessaires pour générer le signal acoustique, c'est-à-dire l'identité des phonèmes, leurs durées et l'évolution de la fréquence fondamentale F0. Il reste maintenant à utiliser ces informations afin de produire un signal de parole.

Il existe plusieurs types de système de synthèse, nous les séparerons en deux grandes familles, les systèmes de synthèse par règles et les systèmes de synthèse par concaténation. Les premiers ont pour but de modéliser et d'étudier le phénomène de production du signal acoustique, alors que les seconds se contentent de connaissances très limitées sur le phénomène de production. Nous allons voir plus en détails ces deux familles de systèmes leur mode de fonctionnement, leurs avantages et leurs inconvénients.

[2] Une carte de passage entre paramètres d'entrée et de sortie.

2.3.1 Synthèse par règles

Cette famille de système fut la première à voir le jour. En effet, il y a deux explications à cela : d'une part, les technologies de l'époque ne permettaient pas de stocker beaucoup de données et d'autre part, il y avait une envie de *comprendre* le phénomène de production de la parole. Le principe de base est de déterminer des règles de passage entre les informations phonético-prosodiques fournies par les modules en amont et les commandes pilotant un modèle de synthèse.
Pour cela, on utilise un modèle paramétrique de production de parole. Un modèle paramétrique couramment utilisé est le synthétiseur à formants [27]. Il s'agit de modéliser le phénomène de production comme un système source-filtre, la source correspondant au flux d'air modulé par les cordes vocales et le filtre correspondant au conduit vocal.
Puis, à partir d'un corpus (de logatomes VCV «Voyelle-Consonne-Voyelle» par exemple), on établit des règles de passage entre la représentation phonético-prosodique de l'entrée textuelle et les paramètres de contrôle du modèle paramétrique de production de parole. Ces règles sont déterminées historiquement par un processus d'essais-erreurs.
Dans ce type de système, on considère qu'à chaque phonème correspond une réalisation acoustique unique. Cependant, on sait que cette réalisation acoustique est dépendante du contexte dans lequel se trouve le phonème : c'est le phénomène de coarticulation. La difficulté principale du synthétiseur par règles consiste donc à déterminer des règles pertinentes en tenant compte de ce phénomène. On citera, par exemple, le modèle de coarticulation d'Öhman [36] dans lequel l'articulation perturbatrice des consonnes se superpose à l'articulation continue des voyelles.
Des modèles statistiques basés sur des HMM (Hidden Markov Models) [18] ou des réseaux de neurones [40] peuvent permettre aussi de décrire le phénomène de coarticulation et sont appliqués à la synthèse de parole. Il s'agit alors d'un compromis entre de la synthèse par règles et de la synthèse par concaténation.

2.3.2 Synthèse par concaténation

Dans ce type de synthèse, on ne s'intéresse plus à la modélisation du phénomène de production : en effet celui-ci va être capturé au sein d'unités acoustiques que l'on va concaténer. Ce type de système ne demande que très peu de connaissances sur le signal de parole, les principales difficultés vont intervenir au niveau du choix des unités (à concaténer) et au niveau du lissage acoustique post-concaténation.

Choix des unités élémentaires

Le choix des unités de base de notre système est la première étape à effectuer. Ces unités doivent avoir certaines propriétés, comme par exemple, contenir le phénomène de coarticulation (unité assez longue), être facilement concaténables (style de voix homogène). Le diphone est une unité qui se prête bien à ce type de système car il est court et il comporte le phénomène de coarticulation (puisqu'il est défini comme la portion de signal acoustique comprise entre les parties stables de deux phones consécutifs). Cependant, la qualité du signal de synthèse s'améliore avec la taille des unités, ainsi d'autres types d'unités telles que les dissyllabes et les

triphones ont été proposées. Le problème engendré par ces unités plus grandes est leur nombre qui devient vite excessif. Certains systèmes utilisent ainsi des polysons, c'est-à-dire des mélanges d'unités de longueur variable : diphones, triphones, dissyllabes, mots, etc. Lors de la sélection, ce sont les unités les plus grandes possibles qui seront conservées [41].

Méthodes de concaténation

Une fois l'unité de base de notre système choisie, il reste à implémenter notre module de concaténation. Ceci s'effectue en trois étapes : la sélection des unités en fonction de la chaîne phonétique d'entrée, l'ajustement des paramètres prosodiques déduits de l'analyse de la phrase à synthétiser et enfin la procédure de concaténation et de lissage des unités sélectionnées.

Sélection des unités

Les unités sélectionnées doivent vérifier des critères segmentaux, c'est-à-dire correspondant à la chaîne phonétique déduite après les traitements linguistiques sur la phrase d'entrée et aussi des critères supra-segmentaux, c'est-à-dire des caractéristiques de durée et de F0. Si les unités ne sont représentées qu'une seule fois dans la base de données, la sélection ne s'effectuera qu'au niveau segmental. En revanche, si les unités sont multi-représentées [24, 8], il faut définir un coût de sélection global sur toute la phrase à synthétiser. Le coût doit tenir compte du contexte dans lequel l'unité a été extraite (qui doit être le plus proche de celui à synthétiser), de la prosodie de cette unité et des discontinuités spectrales avec les futures unités voisines. On décompose généralement le coût global en un coût de sélection (contexte phonétique et prosodique) et en un coût de concaténation (discontinuités aux frontières de concaténation) [22]. Cependant des variantes sont possibles et même souhaitables [44]. Certains systèmes utilisent quatre critères de sélection/concaténation comme dans [39], alors que d'autres systèmes peuvent utiliser plusieurs dizaines de critères [24]. Une fois le coût défini, un algorithme de programmation dynamique ou de simplex [31] nous permet de choisir la meilleure chaîne possible (au sens du coût). Il existe également une autre voie pour la sélection qui se base sur des critères linguistiques : un arbre est construit où chaque unité est classée en fonction de critères linguistiques [15, 32].

Ajustement des paramètres prosodiques

Les unités à concaténer sont choisies ; il faut maintenant les modifier au niveau temporel (durée) et au niveau spectral (variation de F0), afin que la prosodie calculée soit appliquée. Contrairement aux systèmes de synthèse par règles, on ne cherche pas à créer un modèle de production de la parole, ce sont donc des modèles non paramétriques qui sont utilisés pour modifier les caractéristiques prosodiques des unités. Les techniques PSOLA (Pitch Synchronous Overlap-Add) [13] sont les plus utilisées dans ce cadre.
La version temporelle de cet algorithme s'appelle TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add) [35]. Le principe consiste à faire un découpage du signal de parole (correspondant à l'unité choisie) en une série de signaux *élémentaires*. Ce découpage est simple dans le cas d'un signal voisé car on considère chaque période de voisement. Par contre, s'il s'agit d'un signal non voisé on va considérer des tranches de durée équivalente à la période de voisement. Ensuite, on additionne les signaux *élémentaires* en augmentant ou diminuant la période de voisement et on supprime ou duplique certains de ces signaux pour correspondre aux contraintes prosodiques.

Cet algorithme est simple à mettre en oeuvre, peu coûteux en calculs et donne de très bons résultats. Cependant, il peut y avoir des cas de discontinuités lors de la concaténation (discontinuité de pitch, de phase et d'enveloppe spectrale [16]).
Afin de résoudre ces problèmes, de nouveaux systèmes sont apparus, tel LP-PSOLA [35] dans lequel un modèle paramétrique de type LPC (Linear Predictive Coding) est couplé à la méthode TD-PSOLA, ou encore MBROLA [16] dans lequel l'ensemble des unités de la base sont préalablement resynthétisées sur plusieurs bandes (il en va de même pour NU-MBROLA [7] qui est une extension de MBROLA pour des unités non uniformes). Ces méthodes offrent plus de fonctionnalités mais en contrepartie ont un coût de calcul plus important.

Les algorithmes préalablement cités sont généralement utilisés lorsque le nombre d'unités est faible dans le dictionnaire et/ou lorsque l'unité de base est petite (le diphone par exemple). Dans le cas de gros dictionnaires où les unités sont multi-représentées, d'autres méthodes, basées essentiellement sur la sélection et les coûts associés, sont utilisées. En effet, moins le signal sera modifié, plus il gardera un aspect naturel. Dans le cas de gros corpus, les unités étant présentes dans de nombreux contextes avec des caractéristiques prosodiques variées, il sera possible de trouver une chaîne d'unités qui, sans traitement du signal *a posteriori*, soit proche de la prosodie déterminée par l'analyse de la phrase d'entrée. Les méthodes de sélection et le choix des coûts deviennent alors primordiaux.

Concaténation et lissage

Bien que les unités choisies comportent en général des parties stables à leurs extrémités, il peut y avoir des discontinuités spectrales aux points de concaténation. Ces discontinuités sont d'ailleurs audibles dans les systèmes de synthèse utilisant la méthode TD-PSOLA [15]. Des méthodes de lissage sont donc mises en oeuvre pour éliminer ces artefacts aux frontières. Elles se distinguent par les paramètres spectraux utilisés, par la durée sur laquelle on les applique, par le genre d'interpolation. Une approche possible consiste à faire un lissage au niveau de ces frontières [11, 12].
Alors que la première approche ne tient pas compte du contexte dans lequel se trouvent les unités à concaténer, une deuxième voie consiste à adapter le lissage en fonction de l'information contextuelle [30].

2.4 Résumé

Nous avons vu que la synthèse de la parole (audio) à partir du texte était un mécanisme complexe faisant intervenir plusieurs modules, eux-mêmes composés de sous-modules ayant des tâches spécifiques. Ainsi à cause de la complexité des signaux d'entrée (entrée textuelle) et de sortie (signal audio), il n'existe pas de solution unique à la problématique de la synthèse de parole à partir de texte.

Dans le cadre de ces travaux de thèse, nous allons utiliser un système de synthèse par concaténation de polysons. Nous verrons plus en détails les choix adoptés par rapport à l'ensemble des modules composant notre système dans le chapitre consacré à la synthèse. Nous pouvons toutefois annoncer que le schéma synoptique de celui-ci correspond à celui de la figure 2.1. Un premier module de traitements linguistiques nous fournira la chaîne phonétique ainsi

que la structure grammaticale de la phrase proposée en entrée, puis le module prosodique nous donnera les informations temporelles, spectrales et énergétiques du signal à générer. Enfin, à partir de toutes ces informations, un système de sélection/concaténation d'unités de longueurs variables produira en sortie un signal acoustique de synthèse.

Le chapitre suivant est consacré à la partie visuelle de la parole qui peut être ajoutée à tout système de synthèse de parole audio ou directement inclus dans un système audiovisuel. Nous décrirons les modules nécessaires que nous illustrerons par quelques exemples de systèmes choisis dans le domaine.

Références bibliographiques

[1] V. Aubergé. *La synthèse de la parole : des règles aux lexiques*. PhD thesis, Université Pierre Mendès France, Grenoble, France, 1991.

[2] K. Bartkova and C. Sorin. A model of segmental duration for speech synthesis in French. *Speech Communication*, 6 :245–260, 1987.

[3] F. Beaugendre. Modèles de l'intonation pour la synthèse. In H. Méloni, editor, *Fondements et perspectives en traitement automatique de la parole*, chapter Production et synthèse de la parole, pages 97–107. AUPELF UREF, 1996.

[4] R. Belrhali, V. Aubergé, and L.-J. Boë. From lexicon to rules : towards a descriptive method of French text-to-phonetics transcription. In *ICSLP*, pages 1183–1186, 1989.

[5] J. Benello, A.W. Mackie, and J.A. Anderson. Syntactic category disambiguation with neural networks. *Computer Speech and Language*, (3) :203–217., 1989.

[6] R. Boite, T. Dutoit, J. Hancq, H. Leich, and H. Bourlard. *Traitement de la Parole*. Presses Polytechniques Universitaires Romandes, 2000.

[7] B. Bozkurt, T. Dutoit, and M. Bagein. From MBROLA to NU-MBROLA. In *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, pages 127–130, 2001.

[8] A. P. Breen and P. Jackson. Using f0 within a phonologically motivated method of unit selection. In *ICSLP*, 1998.

[9] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[10] Calliope. *La Parole et son Traitement Automatique*. 1989.

[11] D. T. Chappell and J H. L. Hansen. Spectral smoothing for concatenative speech synthesis. In *International Conference on Spoken Language Processing*, pages 1935–1938, Sydney, Australia, November 1998.

[12] D. T. Chappell and J. H. L. Hansen. Spectral smoothing for speech segment concatenation. *Speech Communication*, 36(3-4) :343–373, March 2002.

[13] F. J. Charpentier and M. G. Stella. Diphone synthesis using an overlap-add technique for speech wavforms concatenation. In *ICASSP*, pages 2015–2018, Tokyo, Japan, 1986.

[14] C. d'Alessandro, M. Garnier, and P. Boula de Mareüil. Synthèse de la parole à partir du texte. In H. Méloni, C. d'Alessandro, J.-P. Haton, G. Perennou, and J.-P. Tubach, editors, *Fondements et perspectives en traitement automatique de la parole*, pages 81–96. AUPELF UREF, 1996.

[15] R. Donovan. *Trainable Speech Synthesis.* PhD thesis, Cambridge University, 1996.

[16] T. Dutoit and H. Leich. MBR-PSOLA : Text-to-speech synthesis based on an MBE resynthesis of the segments database. *Speech Communication*, 13 :435–440, 1993.

[17] F. Emerard. *Synthèse par diphones et traitement de la prosodie.* PhD thesis, Université des langues et lettres de Grenoble, 1977.

[18] A. Falaschi, M Giustiniani, and M. Verola. A Hidden Markov Model approach to speech synthesis. In *EUROSPEECH*, pages 187–190, 1989.

[19] F. Fallside and A. Ljolje. Synthesis of natural sounding pitch contours in isolated utterance using hidden markov models. In *IEEE Trans. on ASSP*, volume 34, pages 1074–1080, 1986.

[20] H. Fujisaki. The role of quantitative modeling in the study of intonation. In *Proceedings of the International Symposium on Japanese Prosody*, pages 163–174, 1992.

[21] H. Fujisaki and H. Keikichi. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, 5(4) :233–241, 1984.

[22] J. H. L. Hansen and D. T. Chappell. An auditory-based distortion measure with application to concatrenative speech synthesis. *IEEE Transactions on speech and audio processing*, 6(5) :489–495, 1998.

[23] B. Holm and G. Bailly. Generating prosody by superposing multi-parametric overlapping contours. In *International Conference on Speech and Language Processing*, pages 203–206, Beijing, China, 2000.

[24] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, pages 373–376, Atlanta, GA, 1996.

[25] F. Jelinek. Up from trigrams ! In *Eurospeech*, pages 1037–1040, 1991.

[26] D. H. Klatt. Synthesis by rule of segmental durations in english sentences. In B. Lindblom and S. Ohlman, editors, *Frontiers of speech communication research*, pages 287–300. Academic Press, 1979.

[27] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67 :971–995, 1980.

[28] J. Kupiec. Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language*, (6) :225–242., 1992.

[29] D. Larreur, F. Emerard, and F. Marty. Linguistic and prosodic processing for a text-to-speech synthesis system. In *Eurospeech*, pages 510–513, Paris, France, 1989.

[30] K.-S. Lee and S.-R. Kim. Context-adaptative smoothing for concatenative speech synthesis. In *IEEE Signal Processing Letters*, volume 9, pages 422–425, 2002.

[31] M. Lee, D. P. Lopresti, and J. P. Olive. A text-to-speech platform for variable length optimal unit searching using perception based cost functions. *International Journal of Speech Technology*, 6 :347–356, 2003.

[32] M. W. Macon, A. E. Cronk, and J. Wouters. Generalization and discrimination in tree-structured unit selection. In *Proceedings of the 3rd ESCA/COCOSDA International Speech Synthesis Workshop*, 1998.

[33] H. Mixdorff and O. Jokisch. Building an integrated prosodic model of German. In *European Conference on Speech Communication and Technology*, pages 947–950, 2001.

[34] Y. Morlec, V. Aubergé, and G. Bailly. Evaluation of automatic generation of prosody with a superposition model. In *International Congress of Phonetic Sciences*, volume 4, pages 224–227, 1995.

[35] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 :453–468, 1990.

[36] S.E.G. Öhman. Coarticulation in VCV utterances : Spectrographic measurements. *J Acoust Soc Am.*, 39(1) :151–168, 1966.

[37] D. O'Shaughnessy. A study of French vowel and consonant durations. *Journal of phonetics*, 9 :385–406, 1981.

[38] J. Pierrehumbert. Synthesizing intonation. *Journal of the Acoustical Society of America*, 70 :985–995, 1981.

[39] R. Prudon and C. d'Alessandro. A selection/concatenation text-to-speech synthesis system : databases development, system design, comparative evaluation. In *4th ISCA ITRW on Speech Synthesis*, 2001.

[40] H.B. Richards and J.S. Bridle. The HDM : a segmental hidden dynamical model of coarticulation. In *ICASSP*, 1999.

[41] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform units. In *ICASSP*, pages 679–682, 1988.

[42] Y. Sagisaka. On the prediction of global F0 shpaes for Japanese text-to-speech. In *ICASSP*, volume 1, pages 325–328, 1990.

[43] F. Tesser, P. Cosi, C. Drioli, and G. Tisato. Prosodic data-driven modelling of narrative style in Festival TTS. In *ISCRA Workshop on Speech Synthesis*, 2004.

[44] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis. *Speech Communication*, 48 :45–56, 2006.

[45] C. Traber. F0 generation with a database of natural F0 patterns and with a neural network. In G. Bailly and C. Benoît, editors, *Talking Machines : Theories, Models and Designs*. North Holland, 1992.

[46] J. Trouvain, W. J. Barry, C. Nielsen, and O. Andersen. Implication of energy declination for speech synthesis. In *ETRW Workshop on speech synthesis*, pages 47–52, 1998.

[47] J. Vaissière. *Contribution à la synthèse par règles du français.* PhD thesis, Université des langues et lettres de Grenoble, 1971.

[48] J. van Santen. *Progress in Speech Synthesis.* SPRINGER VERLAG, 1996.

[49] J. van Santen. Segmental duration and speech timing. In Y. Sagisaka, W. N. Campbell, and N. Higuchi, editors, *Computing Prosody.* New York : Springer, 1996.

[50] J. van Santen. Prosodic modelling in text-to-speech synthesis. In *Proceedings of Eurospeech 1997*, Rhodos, Greece, 1997.

Chapitre 3

Synthèse de parole audiovisuelle : les visages parlants

Donner de la voix à un visage ou donner un visage à une voix ?

La synthèse de parole audiovisuelle consiste à délivrer, à partir d'un signal d'entrée de type *texte*, un signal de sortie à la fois acoustique et visuel. Au vu de la littérature, on se rend compte pourtant que la plupart des systèmes de synthèse de parole audiovisuelle délivre non pas un signal de sortie bimodal mais bien deux signaux de sortie unimodaux. S'agit-il alors de donner de la voix à un avatar ou de donner un visage à une voix ou encore les deux à la fois ?

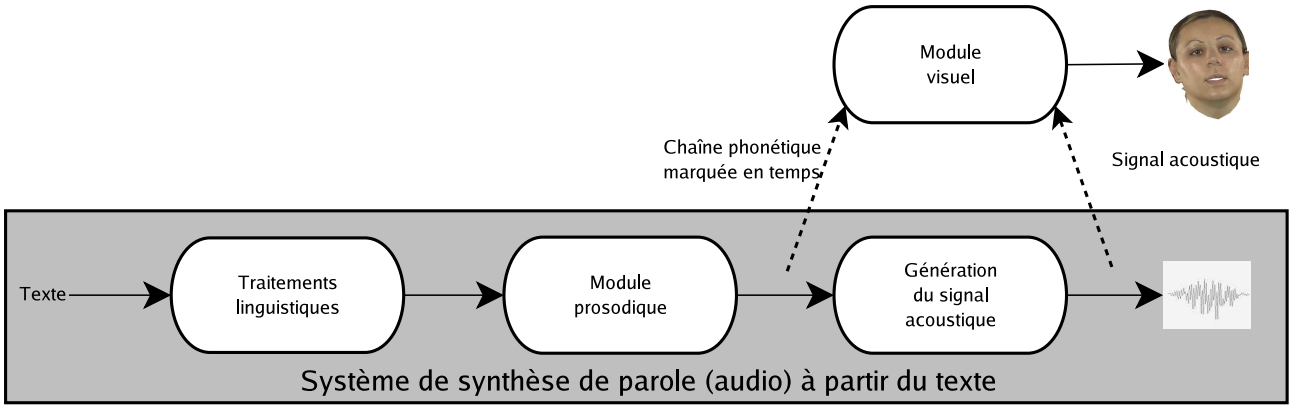


FIG. 3.1 – Le module audiovisuel dans la chaîne de synthèse de parole à partir du texte.

Si on se réfère au monde de l'animation et en particulier au monde du cinéma, l'animation de visage est toujours suivie d'une phase de doublage audio. Dans ce cas, il s'agit donc de donner une voix à un avatar. D'un autre côté, nous avons les systèmes de synthèse de parole audiovisuelle, généralement basés sur des systèmes de synthèse audio. Ces derniers (synthèse de parole audio à partir du texte, TTS : Text-to-Speech) ont un long historique derrière eux et avec les progrès technologiques (augmentation de la vitesse des processeurs, diminution du coût de la mémoire, etc.), ils ont vu l'avènement d'une nouvelle idée, celle de rajouter un visage sur une voix de synthèse. Il s'agit dans ce cas de donner un visage à une voix de synthèse. Pour preuve, la majorité des systèmes de synthèse de parole audiovisuelle consiste en une post-synchronisation

du signal vidéo sur le signal audio délivré par un système TTS *traditionnel*. Il existe cependant d'autres systèmes multimodaux qui gèrent à la fois l'audio et la vidéo [4]. Nous ne traiterons pas ici des avatars dans leur ensemble (une revue des techniques utilisées est disponible dans [58]) mais plus spécifiquement de l'articulation du visage dans l'acte de parole.

Le module visuel va venir se greffer sur un système TTS. Il nécessite en entrée soit une chaîne phonétique marquée en temps, c'est-à-dire le produit du travail du module **traitements linguistiques** et du **module prosodique**, soit l'information acoustique finale délivrée par le système TTS comme on peut le voir sur la figure 3.1.

3.1 Le signal vidéo : intérêts, risques

La synthèse audiovisuelle a pour résultat immédiat l'ajout d'une modalité au signal de sortie, ou plutôt l'ajout d'un signal vidéo à un signal audio existant. Cette opération n'est pas anodine. Va t-on retrouver en sortie la somme des bénéfices des deux signaux pris séparément ou au contraire va t-on apercevoir des interférences ?

3.1.1 Pourquoi rajouter un visage à la synthèse audio ?

Tout d'abord, quel est l'intérêt de rajouter un visage à un système de synthèse de parole à partir du texte ? Il n'existe pas une seule et unique réponse à cette question. C'est d'ailleurs ce qui nous posera des problèmes dans la partie évaluation, mais nous y reviendrons en temps voulu. On peut vouloir rajouter un visage à un système TTS pour des différentes raisons : raisons esthétiques (il est plus convivial d'avoir affaire à un visage parlant plutôt qu'à un système de synthèse de parole audio seul), raisons de falsifications (on pourrait vouloir créer une vidéo de synthèse où une personne prononce une phrase qu'elle n'a jamais dite en réalité), raisons de compréhension (la modalité visuelle étant loin d'être négligeable dans la parole)... Il existe donc de nombreuses raisons de rajouter un visage à un système TTS mais nous allons plus particulièrement nous focaliser sur l'aspect compréhension. Il ne faut pas oublier que la parole est multisensorielle par nature [78]. La parole est le résultat d'une série de mouvements rendus audibles et visibles.

Un exemple de signal vidéo, couplé au signal audio et utile à la compréhension, est le mécanisme de lecture labiale utilisé par les malentendants pour comprendre un message qui leur parvient dégradé dans l'une de ses modalités [79]. Ce mécanisme est également utilisé chez les personnes bien entendantes et c'est surtout apparent lorsqu'il s'agit de communication en milieu bruité. De nombreuses expériences d'intelligibilité, où l'on fait varier le rapport signal/bruit et les modalités de présentation (audio, visuelle, audiovisuelle), ont montré un gain d'intelligibilité lorsque le signal est bimodal (audiovisuel) par rapport à un signal unimodal (audio ou visuel seul), avec l'ajout de lèvres ou d'un visage complet [83, 10]. Cet apport d'intelligibilité a été montré lorsqu'il s'agissait de stimuli naturels mais aussi de synthèse tels que ceux issus d'un visage parlant [8] comme on peut le voir sur la figure 3.2. Une revue intéressante de la perception visuelle de la parole peut être consultée dans [82, 23, 77].

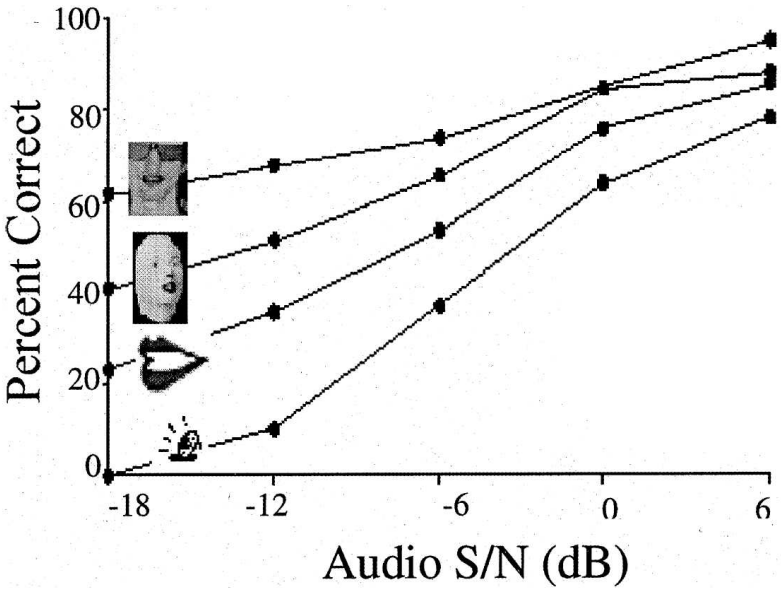


FIG. 3.2 – Score d'intelligibilité en fonction du rapport signal/bruit pour différents types de présentation : audio seul, audiovisuel original (vue de face), audiovisuel modèle de visage, audiovisuel modèle de lèvres [9].

3.1.2 Rajouter...d'accord, mais pas n'importe comment !

La modalité visuelle apporte une information *utile* en clair un gain d'intelligibilité avec un minimum d'effort cognitif, mais ceci est vrai si l'on vérifie deux critères :

- la synchronie ;
- la cohérence.

Il s'agit en effet de deux facteurs très importants car, dès le plus jeune âge, l'être humain est sensible à la synchronie [57] et à la cohérence entre les mouvements du visage et le son associé [49]. Un exemple concret de l'attention toute particulière qui doit être prise pour respecter ces deux critères est l'effet McGurk [61] : par exemple, un stimulus audio [ba] associé à un stimulus visuel différent [ga] est généralement perçu comme [da]. Donc, si l'information visuelle n'est pas cohérente et synchronisée avec l'information audio, l'interprétation peut être biaisée. Il sera donc nécessaire de vérifier ces critères dans le cadre de notre système de synthèse de parole audiovisuelle.

3.2 Le module visuel

Le module capable de générer la sortie visuelle du système de synthèse fonctionne séparément du module audio. Il est contrôlé par l'information phonétique que l'on trouve en sortie du module de traitements linguistiques et le timing délivré par le module prosodique et/ou le signal acoustique. On peut le décomposer en trois sous-modules comme représenté sur la figure 3.3 :

- un premier sous-module va générer des mouvements articulatoires à partir de la séquence de phonèmes marquée en temps ou à partir du signal acoustique : c'est le **modèle de contrôle** ;

– un deuxième sous-module va déterminer comment l'ensemble du visage va être affecté (d'un point de vue géométrique) par les mouvements articulatoires définis plus haut : c'est le **modèle de forme** ;
– un troisième et dernier sous-module va déterminer l'apparence du visage en fonction des déformations faciales : c'est le **modèle d'apparence**.

Dans certains cas (en particulier les systèmes basés sur des images), la séparation entre modules de forme et d'apparence n'est pas aussi tranchée. Cependant, cette représentation reste assez générale pour expliquer le fonctionnement interne du module visuel. Nous allons maintenant présenter chacun des sous-modules et analyser les méthodes utilisées dans chacun des cas. Pour une revue complète de la synthèse de parole audiovisuelle, le lecteur pourra se référer à [85, 5, 14].

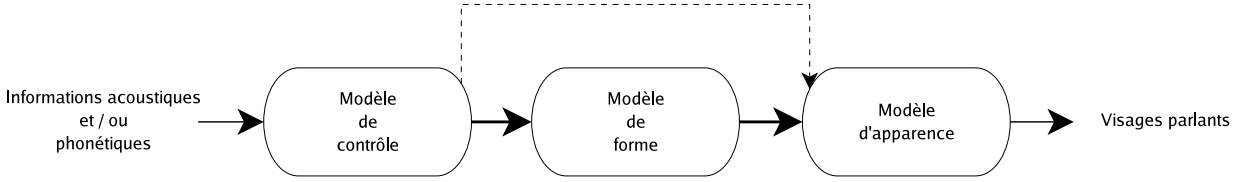


FIG. 3.3 – Décomposition du module visuel d'un système de synthèse de parole audiovisuelle à partir du texte.

3.2.1 Modèle de contrôle

Le modèle de contrôle transforme une information phonétique ou acoustique en mouvements. C'est la base sans laquelle le visage resterait statique. Il ne délivre pas en sortie des coordonnées ou des images mais des paramètres d'articulation qui sont de l'ordre de quelques dizaines.

Nous allons différencier les systèmes utilisant une information phonétique marquée en temps et les systèmes utilisant une information acoustique. Ensuite, dans chacune de ces voies, nous distinguerons les différentes méthodes employées.

Entrée phonétique

Les systèmes qui prennent en entrée un signal de type chaîne phonétique marquée en temps peuvent se décomposer (comme pour la génération du signal audio dans les systèmes TTS) en deux branches : ceux qui vont éditer des règles et ceux qui vont se baser sur de la concaténation.

Règles

Le principe de ce genre de modèles est de stocker uniquement des formes «cibles» et d'apprendre des règles de passage d'une cible à l'autre au cours du temps en fonction de la chaîne phonétique d'entrée. Ces formes cibles appelées visèmes correspondent aux réalisations visuelles des phonèmes [36]. Il peut ainsi exister un même visème pour différents phonèmes par exemple pour les consonnes [p], [b] et [m]. On en compte 21 en français [11]. La grande difficulté va consister à déterminer des règles de passage entre ces cibles. Il pourra s'agir d'interpolations [12, 13, 68, 63, 34, 74, 46, 45] ou de modélisations statistiques (par exemple, une chaîne de Markov peut être associée à la réalisation visuelle d'un phonème [22, 84, 19] ou encore de règles déduites de modèles de coarticulation [25, 51, 32]). Le problème des systèmes qui utilisent des

interpolations est de produire des mouvements de transition qui soient cohérents. En effet, ils ne tiennent pas compte des asynchronies (*i.e.* temps de montée plus court que temps de descente, etc.) des mouvements des articulateurs sous-jacents mis en jeu dans le processus de production de parole. De l'autre côté, les systèmes utilisant des modèles de coarticulation visuelle, tels que le modèle de production de Löfqvist [53] implémenté par Cohen et Massaro [25], étendu par Benoît et al. [9] et appliqué à des données articulographes [35], ou le modèle d'Öhman [65, 66] implémenté par Elisei et al. [32], décrivent mieux ces asynchronies au détriment de la simplicité de mise en place. Les modèles articulatoires de Cohen et Massaro, d'Öhman ainsi que deux modèles basés sur des réseaux de neurones sont présentés et évalués dans [15]. On peut noter que pour la parole audiovisuelle expressive, on utilise des modèles de coarticulation [17] ou des méthodes d'interpolations entre visèmes sur lesquelles sont rajoutées des vecteurs d'expression [48] .

Concaténation

Dans ce type d'approche, comme pour la synthèse de parole audio, la coarticulation va être inscrite dans des segments de signaux visuels qui seront concaténés en fonction de la phrase à prononcer. Les systèmes basés images (par exemple Video Rewrite [21]) ou les systèmes basés modèles [41] utilisent la concaténation d'unités visuelles. Les mêmes problèmes que pour les systèmes TTS apparaissent alors : Quel type d'unités concaténer ? Quels coût de sélection et/ou de concaténation choisir ? Quel lissage et sur quels paramètres l'appliquer ? Pour ce qui est du type d'unité, on trouve des demi-syllabes [41], des diphones [6], des divisèmes [20, 56], des visyllabes [47], des triphones [21, 67], des polysons [38], des séquences vidéo de taille variable [93], etc. En ce qui concerne la phase de sélection, un système de programmation dynamique choisit la meilleure chaîne d'unités à concaténer suivant des coûts qui peuvent être acoustiques seuls [6] ou visuels [21]. Dans la majeure partie des systèmes, même si l'on a recourt à la concaténation, on «post-synchronise» les mouvements articulatoires à l'audio, puisque le signal acoustique et le signal vidéo sont enregistrés séparément. Pourtant il existe des systèmes basés sur de la concaténation d'unités audiovisuelles [62, 6]. Une nouvelle piste d'études consiste à allier les systèmes basés concaténation et les systèmes basés HMM dans un paradigme de planification/exécution [39]. Ainsi, la séléction des unités s'effectuent grâce à un score gestuel sur la géométrie des lèvres déduit d'une synthèse par HMM ; ensuite, l'exécution de ce score est effectuée par un système de concaténation.

Entrée acoustique

Contrairement à la voie décrite précédemment, le signal d'entrée du module visuel n'est pas textuel mais acoustique. Le principe de ce type de module consiste en la création d'une carte de passage entre des données acoustiques et des données articulatoires. Pour la créer, on utilise soit des régressions linéaires comme dans [96, 50], soit des modèles de Markov cachés [44, 94, 84, 2, 42] ou encore plus généralement des réseaux de neurones [2, 60, 95, 40]. Les paramètres acoustiques utilisés sont très variés : il peut s'agir de coefficients LPC (Linear Prediction Coding) [67], de coefficients LSP (Line Spectral Pairs) [96], de coefficients cepstraux [60]. Il en va de même avec les paramètres articulatoires qui sont de nature et en nombre très divers suivant la modélisation

choisie. Pour une revue complète sur le sujet, le lecteur pourra se référer à [97].

3.2.2 Modèle de forme

A la sortie du module de contrôle, nous disposons de quelques dizaines de paramètres articulatoires. C'est le module de forme, qui pour chaque série de paramètres, va recomposer la géométrie du visage en jouant le rôle d'interface entre les paramètres articulatoires et la géométrie du visage correspondante. En sortie de ce module, nous avons les coordonnées d'au moins quelques centaines de points. Le modèle de forme peut être plus ou moins dense, plus ou moins précis, provenir d'une géométrie réelle (c'est-à-dire d'une personne existante) ou de l'imagination d'un infographiste. Nous verrons des exemples plus concrets de modèles de forme dans la section **Illustrons par quelques exemples**.

3.2.3 Modèle d'apparence

A la sortie du sous-module de modèle de forme, nous avons un nuage de points animés. Ceci est suffisant pour tester la cohérence des mouvements mais pas suffisant pour avoir un visage parlant vidéo-réaliste, c'est-à-dire un clone ressemblant (de près ou de loin) à un être humain. Il faut pour cela rajouter une texture à nos points isolés. C'est le but de ce module qui finalise le travail du module visuel. Nous allons voir dans la section suivante des exemples de modèles d'apparence, nous verrons en outre qu'il est parfois difficile de séparer ce module du module de forme dans certains systèmes de synthèse de parole audiovisuelle.

3.3 Illustrons par quelques exemples

Il existe deux grandes voies de recherche pour la création de visage de synthèse. La première est l'approche *basée modèles* dans laquelle on modélise la tête comme un objet 3D. Il s'agit de l'approche conventionnelle qui est apparue il y a 30 ans de cela avec les travaux de Parke [73]. La deuxième approche, *basée images*, est apparue plus récemment et traite des morceaux de vidéo. Il s'agit d'une approche 2D. Nous allons voir au travers de quelques exemples les ressemblances et les différences de chacune de ces deux approches dans l'implémentation des différents sous-modules constituant le module visuel.

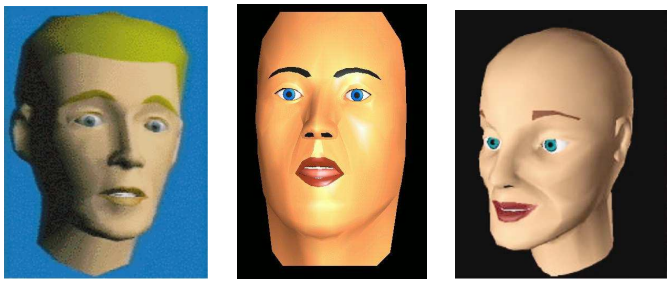
3.3.1 Les approches basées modèles (3D)

Dans ce type de modélisation, on construit un objet 3D représentant une tête que l'on cherche à animer avec des mouvements biologiques. Les différences dans ce type de modélisation vont se situer au niveau du contrôle des points du maillage 3D. Ce contrôle pourra alors se faire par des paramètres uniquement géométriques, des paramètres articulatoires ou encore biomécaniques. Nous allons faire une description des différents types de paramètres de contrôle ayant été implémentés. Les paramètres de contrôle correspondent au modèle de forme. Ils sont l'interface entre le modèle de contrôle et la géométrie 3D. Pour ce qui est de l'apparence, les

modèles 3D recourent généralement à des méthodes de plaquage de texture, de type *flat*[1] ou vidéo-réaliste.

Contrôle par paramètres géométriques

Les modèles contrôlés par des paramètres purement géométriques ne cherchent pas à comprendre les mécanismes physiologiques sous-jacents responsables des mouvements du visage lors de l'acte de parole, mais juste à les reproduire en termes géométriques. Ces paramètres peuvent agir sur un point en particulier ou sur des zones de points, pour imposer des expressions par exemple ; les déplacements sont en général des fonctions géométriques de base (rotation, translation). Les exemples les plus représentatifs sont Baldi [59], la tête parlante développée au Perceptual Science Laboratory (University of California) appliquée à plusieurs langues [71], Sven [12] et Olger [16] développées au Department of Speech, Music and Hearing (Royal Institute of Technology (KTH)) et enfin les têtes parlantes du Laboratory of Computational Engineering [68, 63, 37] (Helsinki University of Technology). Une représentation de certains de ces visages parlants se trouve sur la figure 3.4.



(a) Sven (b) Baldi (c) LCE

FIG. 3.4 – Exemples de descendants du modèle de Parke [73].

Le standard industriel ISO/IEC MPEG-4 [70, 69, 1, 72, 18] fait partie de cette catégorie de contrôle de visages parlants. Un ensemble de 84 points (3D) a été défini, les FPs (Feature Points) comme représentés sur la figure 3.5. Ils sont pilotés par 68 paramètres, les FAPs (Facial Action Parameters). Des extensions ont été proposées comme dans [43]. Ces paramètres ne sont pas que géométriques, ils sont aussi parfois articulatoires ce qui peut poser des problèmes de concurrence [5].

Contrôle par paramètres articulatoires

Les paramètres articulatoires sont un pas en avant dans la compréhension par rapport aux seuls paramètres géométriques. Ce n'est pas l'aboutissement d'une modélisation du visage mais une optimisation du choix des paramètres dans le but de produire des mouvements cohérents du visage. Il faut bien comprendre que ces paramètres ne pourraient *a priori* pas être utilisés pour des tâches pour lesquelles ils n'ont pas été déterminés. En effet, contrairement aux paramètres purement géométriques, ces paramètres portent en eux une explication du mouvement

[1] Il s'agit d'une modélisation non vidéo-réaliste correspondant à une marionnette.

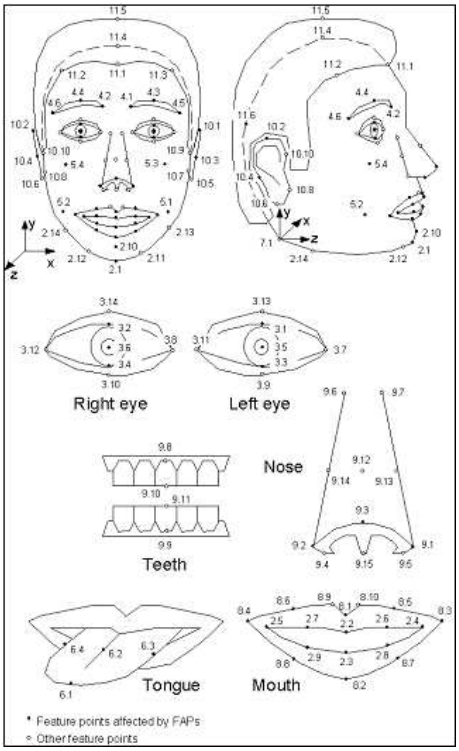


FIG. 3.5 – Position des FP de la norme MPEG-4 [64].

(de production de la parole dans notre cas). Un exemple de paramètres articulatoires est ceux introduits par [91] qui contrôlent les FAPs (de la norme MPEG-4) : ils correspondent au contrôle de la hauteur, de la largeur de la bouche, au contrôle de la protrusion des lèvres et au contrôle du mouvement de rotation de la mâchoire. Les chercheurs de l'Institut de la Communication Parlée (ICP) ont proposé un ensemble de six paramètres articulatoires (cf. figure 3.6) contrôlant la mâchoire, les lèvres et le larynx et une méthodologie d'extraction de ceux-ci par analyse en composantes principales guidée [76, 32]. Le modèle de contrôle utilisé est le modèle de coarticulation d'Öhman [32]. Le modèle de forme est un modèle linéaire permettant la conversion entre les six paramètres articulatoires et les coordonnées 3D de quelques centaines de points sur le visage. Le modèle d'apparence consiste en un plaquage d'un mélange de textures.

(a) Abaissement/Elévation de la mâchoire

(b) Avancée/Rétraction de la mâchoire

(c) Etirement/protrusion des lèvres

FIG. 3.6 – Exemples de paramètres articulatoires utilisés à l'ICP pour créer des clones [32].

Contrôle par paramètres biomécaniques

Ce dernier type de paramètres est *a priori* celui qui devrait décrire *le mieux* les mouvements du visage, autant dans l'acte de parole que dans les autres mouvements qu'il peut produire. En effet, dans ce cas ce sont les muscles et les articulateurs qui vont être modélisés. Il s'agit d'une modélisation *réelle* dans le sens où les paramètres ne vont pas engendrer des déformations mais des actions sur les muscles se propageant aux articulateurs qui vont engendrer des déformations faciales. Cependant, le contrôle musculaire est complexe : il y a près de 250 muscles dans le visage, auxquels il faut ajouter les interactions os/tissus très difficiles à modéliser.

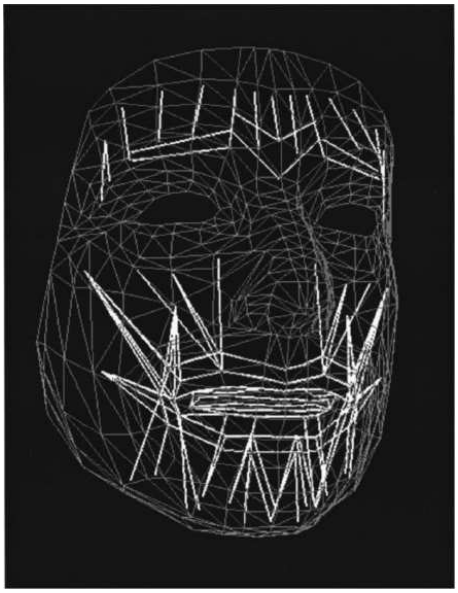


FIG. 3.7 – Lignes d'action des muscles du visage du modèle de Lucero et al. [55].

Les sommets des maillages 3D sont considérés dans ce type de méthodes comme des points de chair. On cherche donc à modéliser les propriétés biomécaniques de la peau et des systèmes musculo-squelettiques sous-jacents. 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 [31] 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 (un exemple [55, 54] est représenté sur la figure 3.7). 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 [75] et organisés en couches [92, 86, 87, 52, 3]. L'évolution de ce type de méthodes tend vers la modélisation par éléments finis des couches de tissus de peau [30, 7].

3.3.2 Les approches basées images (2D)

Alors que la voie pionnière en synthèse de visages parlants consistait à modéliser la tête comme un objet 3D, une nouvelle approche basée 2D a vu le jour depuis une dizaine d'années.

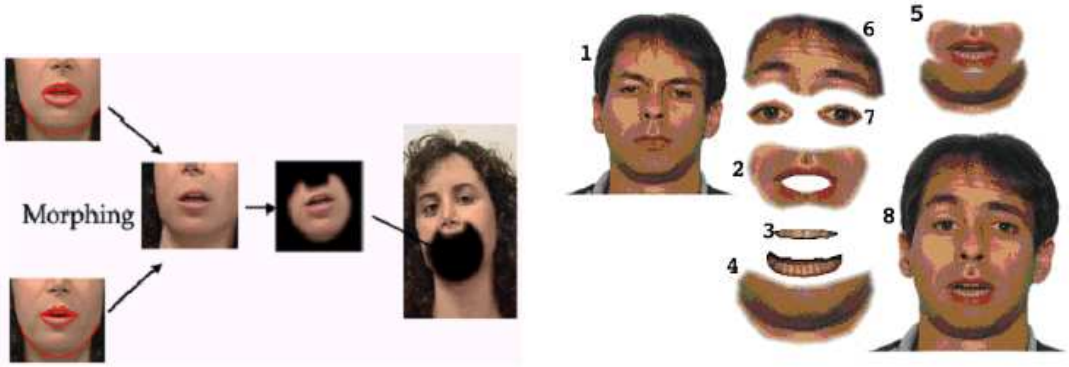
L'intérêt direct d'une telle approche est le vidéoréalisme. Nous allons décrire quelques-unes des méthodes existantes en essayant de séparer la contribution de chaque sous-module (contrôle moteur, contrôle de forme, contrôle d'apparence). On montrera que la distinction entre sous-modules n'est pas aussi simple que dans le cas de modèles 3D. Pour une description plus complète de ces modèles le lecteur pourra se référer à [81].

Superposition de segments vidéo sur une image de fond

Une première approche consiste à récupérer, pendant la phase d'apprentissage, les mouvements de certaines régions du visage en fonction des sons produits puis à les utiliser pour la synthèse, en choisissant et superposant les segments appropriés sur une image de fond.

Un système de synthèse caractéristique de cette approche est Video Rewrite élaboré par Bregler et al [21]. Ce système utilise une vidéo de fond qui sert de scène à la bouche synthétisée (cf. figure 3.8 (a)). Sur cette vidéo de fond, on superpose la plus longue séquence vidéo liée à la bouche de la base d'apprentissage qui correspond au bon visème, au bon phonème et à la bonne position de tête. Le modèle de contrôle est basé sur de 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 bouche.

Un autre système de synthèse audiovisuelle, développé dans les laboratoires de ATT par Cosatto & Graf [27, 28], 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 (cf. figure 3.8 (b)). 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. Dans ce type de système, il n'existe pas de distinction bien établie entre forme et rendu. En effet, il ne s'agit que d'un seul et même module puisque dans l'implémentation, le mouvement des points correspond à leur changement de couleur.



(a) VideoRewrite (b) ATT

FIG. 3.8 – Exemples de systèmes utilisant la superposition de segments vidéo.

Mouvements de pixel

Cette nouvelle approche consiste à considérer non plus des mouvements et des déformations de régions du visage mais le déplacement de chaque pixel de l'image en fonction du son prononcé : c'est le modèle de forme. Des algorithmes de type flux optique sont implémentés et permettent d'estimer le mouvement des pixels lors du passage d'une image à une autre. Le modèle de contrôle est basé généralement sur des visèmes.

Actors [80], MikeTalk [34] et Mary [33] par exemple, sont des systèmes qui synthétisent une tête parlante en utilisant des techniques de morphing sur des images représentant des visèmes. Pour chaque phonème en contexte, une image prototypique (le visème) est capturée et représente la cible à atteindre pour le visage quand il prononce un son particulier. Le flux optique est calculé pour chaque passage d'un visème à un autre dans le sens direct et indirect, ce qui permet de reconstruire les images intermédiaires. Dans le cas de Mary, on ne considère plus le passage entre deux visèmes mais on peut circuler dans une base de visèmes, le MMM (Multidimensional Morphable Models) (cf. figure 3.9). Le modèle d'apparence consiste ensuite à de l'interpolation entre les images-clé (les visèmes). Une méthode de pilotage d'une personne à partir du modèle d'une autre personne a été implémentée dans [24]. Contrairement à la première approche, on ne concatène pas des morceaux de vidéo mais on calcule les transitions à chaque fois entre les visèmes.

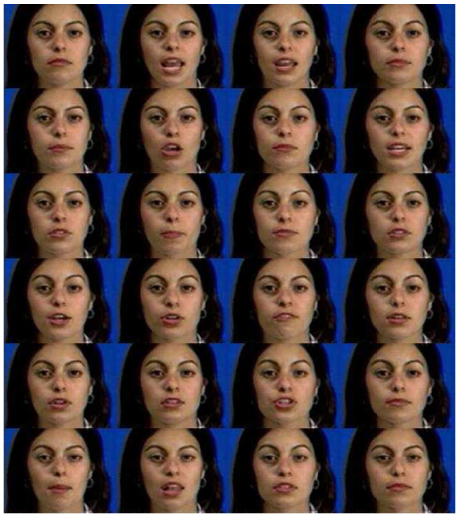
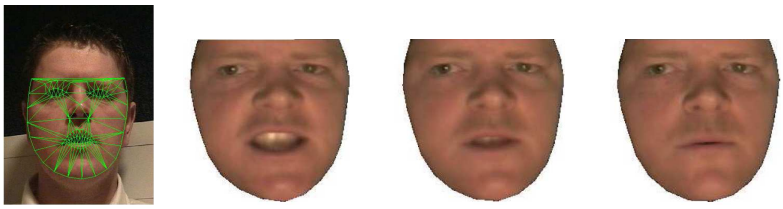


FIG. 3.9 – Mary : 24 des 46 images prototypiques constituant le MMM [33].

Modèle de forme et d'apparence

Une autre méthode consiste à créer un modèle statistique de forme et d'apparence du visage [26, 88, 89, 29]. Tout d'abord, on détermine le modèle statistique de forme : on étiquette à la main sur un ensemble d'images le modèle de distribution de points et on applique une ACP (Analyse en Composantes Principales) sur les coordonnées des points de repère. Puis un modèle d'apparence est calculé en normalisant toutes les images sur la forme moyenne comme représenté sur la figure 3.10. L'avantage est que chaque image est 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 : l'intérieur de la bouche et les yeux apparaissent flous. Une solution consiste à modéliser séparément ces différentes régions et créer des MAM (Multi Segment Appearance Models) [90]. Pour construire les MAM, les images sont segmentées en sous-régions perceptivement importantes : le visage entier, la bouche et chacun des yeux.



(a) (b)

FIG. 3.10 – Modèles de formes et d'apparence [88] : a) le modèle de forme est utilisé pour normaliser les images de la base d'apprentissage ; b) images de synthèse créées à partir du modèle de forme et d'apparence.

3.4 Résumé

Après avoir décrit les avantages et les dangers à rajouter un module visuel à un système de synthèse de la parole à partir du texte, nous en avons décrit le fonctionnement. Nous avons plus particulièrement mis l'accent sur le sous-module primordial : le modèle de contrôle. L'inventaire (non exhaustif) des systèmes de synthèses de parole audiovisuelle qu'ils soient basés modèles ou basés images nous a permis de voir les différentes implémentations possibles.

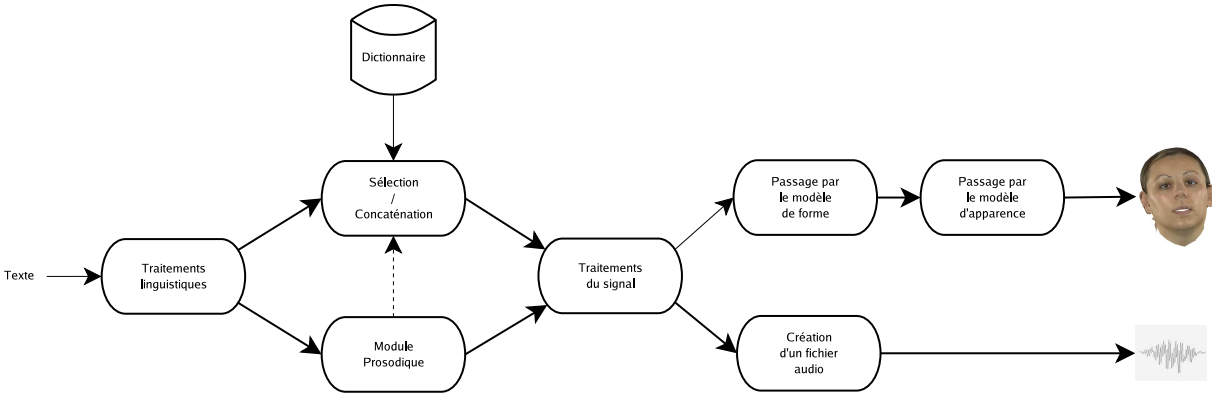


FIG. 3.11 – Décomposition du système AVTTS (AudioVisual Text-To-Speech) utilisé pour nos travaux de thèse.

Concernant nos travaux de thèse, nous allons utiliser un module visuel basé modèle c'est-à-dire 3D, dont le contrôle moteur consiste en de la concaténation de segments de paramètres audiovisuels. En effet, nous nous baserons sur un seul dictionnaire dans lequel seront stockés

en synchronie les polysons multimodaux (audio + paramètres articulatoires du visage). Ainsi, nous n'aurons pas de problème de post-synchronisation à régler. Le modèle de forme sera un modèle linéaire qui permettra le passage entre un ensemble de paramètres articulatoires et les coordonnées 3D des points du maillage. En ce qui concerne le modèle d'apparence il s'agira d'une méthode de plaquage de texture. Un schéma synoptique du système de synthèse de parole audiovisuelle que nous avons utilisé se trouve sur la figure 3.11.

Le chapitre suivant décrira l'originalité de ces travaux, c'est-à-dire la génération de parole multimodale *augmentée*. Nous décrirons plus particulièrement la Langue française Parlée Complétée (LPC), ses caractéristiques ainsi que les systèmes de synthèse de ce code existants dans la littérature.

Références bibliographiques

[1] G.A. Abrantes and F. Pereira. MPEG-4 facial animation technology : Survey, implementation, and results. *IEEE Transactions on circuits and systems for video technology*, 9(2) :290–305, 1999.

[2] E. Agelfors, J. Beskow, B. Granström, M. Lundeberg, G. Salvi, K.-E. Spens, and T. Öhman. Synthetic visual speech driven from auditory speech. In *AVSP*, 1999.

[3] I. Albrecht, J. Haber, and H.-P. Seidel. Speech synchronization for physics-based facial animation. In *WSCG*, pages 9–16, 2002.

[4] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

[5] G. Bailly, M. Bérar, F. Elisei, and M. Odisio. Audiovisual speech synthesis. *International Journal of Speech Technology*, 6 :331–346, 2003.

[6] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

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

[8] C. Benoît, A. Fuster-Duran, and B. Le Goff. An investigation of hypo- and hyper-speech in the visual modality. In *ETRW on Speech Production : from Control Strategies to Acoustics*, pages 237–240, Autrans, France, 1996.

[9] C. Benoît and B. Le Goff. Audio-visual speech synthesis from French text : Eight years of models, designs and evaluation at the ICP. *Speech Communication*, 26 :117–129, 1998.

[10] C. Benoît, T. Mohamadi, and S. Kandel. Audio-visual intelligibility of French speech in noise. *Journal of Speech and Hearing Research*, 37 :1195–1203, 1994.

[11] C. Benoît and L. C. W. Pols. *Talking Machines : Theories, Models and Designs.*, chapter On the assessment of synthetic speech, pages 435–441. Bailly, G. and Benoît, C. and Sawallis, T. R., Amsterdam, 1992.

[12] J. Beskow. Rule-based visual speech synthesis. In *EUROSPEECH'95*, volume 1, pages 299–302, Madrid, Spain, September 1995.

[13] J. Beskow. Talking heads - communication, articulation and animation. In *Proceedings of Fonetik '96, Swedish Phonetics Conference*, pages 53–56, 1996.

[14] J. Beskow. *Talking Heads Models and Applications for multimodal speech synthesis.* PhD thesis, Department of speech, Music and Hearing, KTH, 2003.

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

[16] J. Beskow, M. Dahlquist, B. Granstrom, M. Lundeberg, K.-E. Spens, and T. Ohman. The Teleface project - multimodal speech communication for the hearing impaired. In *Eurospeech*, pages 2003–2010, Rhodos, Greece, 1997.

[17] E. Bevacqua and C. Pelachaud. Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 15 :297–304, 2004.

[18] C. Bonamico, C. Braccini, M. Costa, F. Lavagetto, and R. Pockaj. Using MPEG-4 parameters for calibrating/animating talking heads. In *Tyrrhenian International Workshop on Digital Communications*, 2002.

[19] M. E. Brand. Voice puppetry. In *ACM SIGGRAPH*, 1999.

[20] A. P. Breen, E. Bowers, and W. Welsh. An investigation into the generation of mouth shapes for a talking head. In *ICSLP*, pages 2159–2162, 1996.

[21] C. Bregler, M. Cowell, and M. Slaney. Videorewrite : driving visual speech with audio. In *SIGGRAPH'97*, pages 353–360, Los Angeles, CA, 1997.

[22] N. Brooke and S. D. Scott. Two and three-dimensional audio-visual speech synthesis. In *AVSP*, 1998.

[23] M. A. Cathiard. La perception visuelle de la parole : aperçu des connaissances. In *Bulletin de l'Institut de Phonétique de Grenoble*, volume 18, pages 109–193. Institut de Phonétique de Grenoble, 1989.

[24] Y.-J. Chang and T. Ezzat. Transferable videorealistic speech animation. In *ACM Siggraph/Eurographics Symposium on Computer Animation*, 2005.

[25] M.M. Cohen and D.W. Massaro. Modeling coarticulation in synthetic visual speech. In Springer-Verlag, editor, *Models and Techniques in Computer Animation*, pages 139–156. N.M. Thalmann & D. Thalmann, Tokyo, Japan, 1993.

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

[27] E. Cosatto and H.-P. Graf. Sample-based synthesis of photo-realistic talking heads. *Computer Animation*, pages 103–110, 1998.

[28] E. Cosatto and H.-P. Graf. Photo-realistic talking heads from image samples. In *IEEE Transactions on Multimedia*, volume 2, pages 152–163, 2000.

[29] D. Cosker, Marshall D., Rosin P., and Hicks Y. Video realistic talking heads using hierarchical non-linear speech-appearance models. In *Proceedinds of Mirage*, INRIA Rocquencourt, France, March 2003.

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

[31] P. Ekman and W. V. Friesen. Facial action coding system (FACS) : a technique for the measurements of facial action. *Consulting Psycholigists Press*, 1978.

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

[33] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of ACM SIGGRAPH*, San Antonio, USA, 2002.

[34] T. Ezzat and T. Poggio. Miketalk : A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference*, Philadelphia, PA, June 1998.

[35] S. Fagel and C. Clemens. Two articulation models for audiovisual speech synthesis - description and determination. In *AVSP*, pages 215–220, 2003.

[36] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11 :796–804, 1968.

[37] M. Frydrych, J. Kätsyri, M. Dobšík, and M. Sams. Toolkit for animation of finnish talking head. In *AVSP*, St Jorioz, France, 2003.

[38] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun. Analysis and synthesis of the three-dimensional movements of the head, face and hand of a speaker using Cued Speech. *Journal of the Acoustical Society of America*, 118(2), August 2005.

[39] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw. TDA : A new trainable trajectory formation system for facial animation. In *Interspeech ICSLP*, Pittsburgh, PA, September 2006.

[40] R. Gutierrez-Osuna, P.K. Kakumanu, A. Esposito, O.N. Garcia, A. Bojorquez, J.L. Castillo, and I.J. Rudomin. Speech-driven facial animation with realistics dynamics. *IEEE Transaction on Multimedia*, 7(1) :33–42, 2005.

[41] A. Hallgren and B. Lyberg. Visual speech synthesis with concatenative speech. In *AVSP*, 1998.

[42] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. In *IEEE Transactions on Speech and Audio Processing*, 2004.

[43] P. Hong, Z. Wen, and T. Huang. Real-time speech-driven 3D face animation. In *International Symposium on 3D Data Processing Visualization Transmission*, 2002.

[44] F. J. Huang and T. Chen. Real-time lip-synch face animation driven by human voice. In *IEEE Mutlimedia Signal Processing Workshop*, 1998.

[45] G. A. Kalberer, P. Müller, and L. Van Gool. Visual speech, a trajectory in viseme space. *International Journal of Imaging Systems and Technology*, 13 :74–84, 2003.

[46] S. Kshirsagar and N. Magnenat-Thalmann. Viseme space for realistic speech animation. In *AVSP*, pages 30–35, Aalborg, Denmark, 2001.

[47] S. Kshirsagar and N. Magnenat-Thalmann. Visyllable based speech animation. In *Proceedings Eurographics*, 2003.

[48] S. Kshirsagar, T. Molet, and N. Magnenat-Thalmann. Principal components of expressive speech animation. In *Proc. Computer Graphics International*, pages 38–44, 2001.

[49] P. K. Kuhl and A. N. Meltzoff. The bimodal perception of speech in infancy. *Science*, 218 :1138–1141, 1982.

[50] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson. Kinematics-based synthesis of realistic talking faces. In *AVSP*, pages 185–190, 1998.

[51] B Le Goff and C. Benoît. A text-to-audiovisual-speech synthesizer for French. In *4th International Conference on Spoken Language Processing*, pages 2163–2166, Philadelphia, USA, 1996.

[52] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *SIGGRAPH*, pages 55–62, 1995.

[53] A. Löfqvist. Speech as audible gestures. *Speech Production and Speech Modeling*, pages 289–322, 1990.

[54] J. C. Lucero, S. T. R. Maciel, D. A. Johns, and K. G. Munhall. Empirical modeling of human face kinematics during speech using motion clustering. *Journal of the Acoustical Society of America*, 118(1) :405–409, 2005.

[55] J. C. Lucero and K. G. Munhall. A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America*, 106(5) :2834–2842, 1999.

[56] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise. Accurate automatic visible speech synthesis of abitrary 3D models based on concatenation of diviseme motion capture data. *Computer Animation and Virtual Worlds*, 15 :485–500, 2004.

[57] K. MacKain, M. Studdert-Kennedy, S. Spieker, and D. Stern. Infant intermodal perception speech perception is a left hemisphere function. *Science*, 219 :1347–1349, 1983.

[58] N. Magnenat-Thalmann and D. Thalmann. *Handbook of Virtual Humans*. John Wiley & Sons Ltd, 2004.

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

[60] D.W. Massaro, J. Beskow, M.M. Cohen, C.L. Fry, and T. Rodriguez. Picture my voice : audio to visual speech synthesis using artificial neural networks. *Proceedings of AVSP*, 1999.

[61] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 26 :746–748, 1976.

[62] S. Minnis and A. P. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *International Conference on Speech and Language Processing*, pages 759–762, Beijing, China, 2000.

[63] R. Möttönen, J.-L. Olivès, J. Kulju, and M. Sams. Parametrized visual speech synthesis and its evaluation. In *European Signal Processing Conference*, Tampere, Finland, 2000.

[64] MPEG Video and SNHC. Text of ISO/IEC FDIS 14496-2 : Visual. In *MPEG Meeting*, 1998.

[65] S.E.G. Öhman. Coarticulation in VCV utterances : Spectrographic measurements. *J Acoust Soc Am.*, 39(1) :151–168, 1966.

[66] S.E.G. Öhman. Numerical model of coarticulation. *JASA*, 41(2) :310–320, 1967.

[67] T. Okadome, T. Kaburagi, and M. Honda. Articulatory movement formation by kinematic triphone model. In *IEEE International Conference on System Man and Cybernetics*, pages 469–474, 1999.

[68] J.-L. Olivès, R. Möttönen, J. Kulju, and M. Sams. Audio-visual speech synthesis for finnish. In *Audio-Visual Speech Processing*, Santa Cruz, USA, August 1999.

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

[70] J. Ostermann, M. Beutnagel, A. Fischer, and Y. Wang. Integration of talking heads and text-to-speech synthesizers for visual TTS. In *ICSLP'98*, Sydney, Australia, December 1998.

[71] S. Ouni, M. M. Cohen, and D. W. Massaro. Training Baldi to be multilingual : A case study for an Arabic Badr. *Speech Communication*, 45 :115–137, 2005.

[72] I. S. Pandzic and R. Forchleimer. *MPEG-4 Facial Animation - the Standard, Implementation and Applications*. John Wiley & Sons, Chichster, England, 2002.

[73] F. I. Parke. A model for human faces that allows speech synchronised animation. *Journal of Computers and Graphics*, 1(1) :1–4, 1975.

[74] C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, and P. Cosi. Modelling an italian talking head. In *Proceedings of AVSP*, 2001.

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

[76] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[77] J. Robert-Ribes. *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles.* PhD thesis, Institut National Polytechique de Grenoble, 1995.

[78] J.-L. Schwartz. La parole multisensorielle : Plaidoyer, problèmes, perspective. In *Journées d'Etude sur la Parole*, Fès, Maroc, 2004.

[79] J.-L. Schwartz, F. Berthommier, and C. Savariaux. Seeing to hear better : evidence for early audio-visual interactions in speech identification. *Cognition*, 93 :B69–B78, 2004.

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

[81] M. Slaney. *Audiovisual Speech Processing*, chapter Image-based Facial Synthesis, pages 149–161. MIT Press, 2003.

[82] Q. Summerfield. Use of visual information for phonetic perception. *Phonetica*, 17 :3–46, 1979.

[83] Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke. Lips, teeth, and the benefits of lipreading. In A. W. Young and H. D. Ellis, editors, *Handbook of Research on Face Processing*, pages 223–233. Elsevier Science Publishers, 1989.

[84] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda. Visual speech synthesis based on parameter generation from HMM : speech-driven and text-and-speech-driven approaches. In *AVSP*, 1998.

[85] D. Terzopoulos, F. Parke, and K. Waters. Panel on facial animation : Past, present and future. In *Proceedings of SIGGRAPH'97*, 1997.

[86] D. Terzopoulos and K. Waters. Physicaly-based facial modeling, analysis and animation. *The Journal of Visual and Computer Animation*, 1 :73–80, 1990.

[87] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on pattern analysis and machine intelligence*, 15(6) :569–579, june 1993.

[88] B.-J. Theobald, Bangham J. A., Matthews I., and Cawley G. Towards video realistic synthetic visual speech. In *ICASSP*, pages 3892–3895, 2002.

[89] B.-J. Theobald, J. A. Bangham, I. Matthews, and G. Cawley. Evaluation of a talking head based on appearance models. In *Audio-Visual Speech Processing*, September 2003.

[90] B. J. Theobald, J. A. Bangham, I. Matthews, and G. C. Cawley. Visual speech synthesis using statistical models of shape and appearance. In *Proc. Auditory-Visual Speech Processing*, 2001.

[91] F. Vignoli and C. Braccini. A text-speech synchronization technique with applications to talking heads. In *AVSP'99*, Santa Cruz, CA, 1999.

[92] K. Waters. A muscle model for animating three-dimensional facial expression. *Computer Graphics*, 22(4) :17–24, 1987.

[93] C. Weiss. A framework for data-driven video-realistic audio-visual speech synthesis. In *LREC*, 2004.

[94] E. Yamamoto, S. Nakamura, and K. Shikano. Subjective evaluation for hmm-based speech-to-lip movement synthesis. In *AVSP*, pages 227–232, 1998.

[95] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. Facial animation and head motion driven by speech acoustics. In *The 5th International Seminar on Speech Production*, pages 265–268, Munich, Germany, 2000.

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

[97] G. Zorić. Real-time face animation driven by human voice. In *ConTEL*, 2003.

Chapitre 4

Synthèse audiovisuelle augmentée : synthèse de code LPC

Pourquoi rajouter une modalité à la synthèse audiovisuelle ?

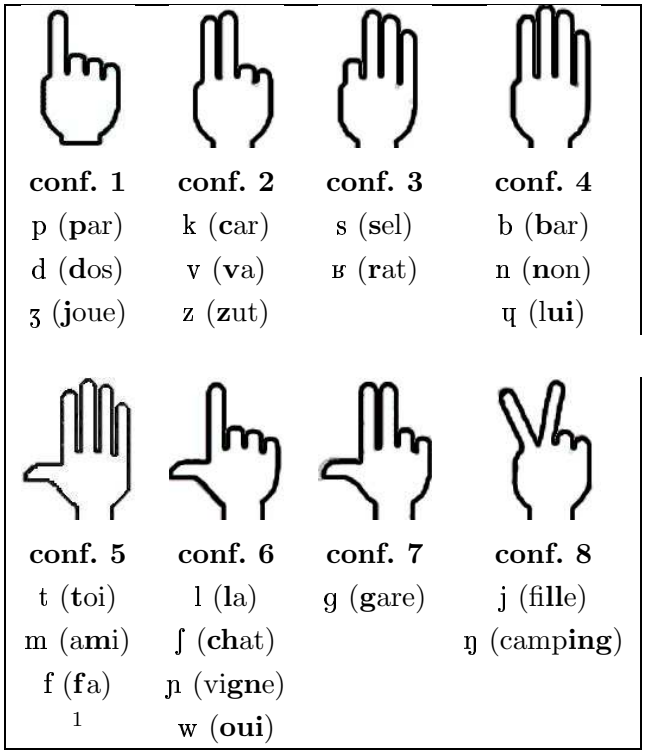
Dans le chapitre précédent, nous avons vu que si l'on rajoutait une modalité en respectant les conditions de cohérence et de synchronie, nous avions alors un apport d'information. Pourquoi donc vouloir rajouter une modalité puisque nous avons *a priori* déjà toute l'information et que celle-ci est même parfois redondante ? Les personnes sourdes et malentendantes reçoivent la modalité audio de façon appauvrie (voire quasi-nulle). Or, l'information vidéo seule est largement insuffisante pour avoir accès au langage parlé. En effet, une image ne correspond pas à un son unique du fait de l'existence des sosies labiaux (ou visèmes). Le phénomène de production de parole fait intervenir des mouvements de la mâchoire, des lèvres et des joues qui sont immédiatement visibles, mais également des mouvements d'organes sous-jacents tels que le larynx, le velum ou la langue. De plus, les mouvements de la langue sont faiblement corrélés avec les mouvements visibles du visage ($R \sim 0.7$) [32, 24] et cette corrélation est insuffisante pour retrouver les corrélats phonétiques importants comme le lieu d'articulation linguale par exemple [20, 8]. La lecture labiale seule est donc insuffisante dû à un manque d'information sur le point d'articulation de la langue, des modes d'articulation (nasalité, voisement) et à la similarité de certaines formes de lèvres pour certains phonèmes. Dans tous les cas, même le meilleur décodeur ne peut pas identifier plus de 50% de phonèmes dans des syllabes sans sens [28] ou dans des mots ou des phrases [9].

Il faut donc trouver une autre modalité capable de supplanter la perte de la première. Le canal utilisé pourrait être visuel et/ou tactile. Il existe des méthodes visant à augmenter l'intelligibilité de la parole par ajout d'une modalité, comme par exemple la méthode Tadoma [29] qui utilise le toucher (la main de l'auditeur est placée sur le visage du locuteur, le pouce se plaçant sur les lèvres et le reste des doigts sur la pommette et le cou) : c'est un système utilisé par des personnes sourdes et aveugles qui va leur permettre de récupérer par le toucher les différents mouvements du conduit vocal. Pour les personnes sourdes, nous citerons également la Langue française Parlée Complétée [12] qui utilise la vue de la main du locuteur en complément de ses mouvements labiaux pour améliorer la perception de la parole. C'est cette dernière méthode qui va être à la base de notre système de synthèse de parole audiovisuelle augmentée.

4.1 La Langue française Parlée Complétée

4.1.1 Description

Le Langage Parlé Complété, renommé récemment Langue française Parlée Complétée (ou code LPC), a été créé 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 [14, 16].



conf. 1	conf. 2	conf. 3	conf. 4
p (**p**ar)	k (**c**ar)	s (**s**el)	b (**b**ar)
d (**d**os)	v (**v**a)	ʁ (**r**at)	n (**n**on)
ʒ (**j**oue)	z (**z**ut)		ɥ (l**ui**)

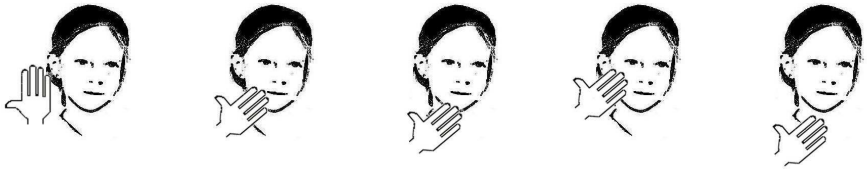
conf. 5	conf. 6	conf. 7	conf. 8
t (**t**oi)	l (**l**a)	g (**g**are)	j (fi**ll**e)
m (a**mi**)	ʃ (**ch**at)		ŋ (camp**ing**)
f (**f**a)	ɲ (vi**gn**e)		
[1]	w (**oui**)		

TAB. 4.1 – Formes de la main du code LPC pour le français.

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. tableau 4.1) pour indiquer une position sur le visage (déterminant un sous-ensemble de voyelles cf. tableau 4.2) 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.

[1] la configuration 5 est également utilisée pour coder toute voyelle non précédée d'une consonne.



Côté	**Bouche**	**Menton**	**Pommette**	**Gorge**
a (m**a**)	i (m**i**)	ɛ (m**ais**)	ɛ̃ (m**ain**)	œ̃ (**un**)
o (m**aux**)	ɔ̃ (**on**)	u (m**ou**)	ø (f**eu**)	y (t**u**)
œ (t**eu**f)	ɑ̃ (t**em**ps)	ɔ (f**o**rt)		e (f**é**e)
[2]				

TAB. 4.2 – Positions de la main par rapport au visage du code LPC pour le français.

4.1.2 Intérêts du code LPC

A l'origine, le Dr. Cornett a créé ce système pour permettre aux enfants sourds dont les parents étaient entendants de pouvoir communiquer avec eux facilement (l'apprentissage du code LPC étant rapide) et d'acquérir un modèle complet du langage parlé.

En effet, les sourds et malentendants utilisent, en général, la lecture labiale pour avoir accès à la parole. Cependant, cette méthode utilisant la modalité visuelle n'est pas suffisante pour avoir accès à l'intégralité de l'information fournie par le locuteur [22] et donc pour comprendre la parole. De plus, les enfants malentendants utilisant la lecture labiale seule développent des représentations phonologiques sous-spécifiées, ce qui entrave le développement normal du langage [25].

D'ailleurs, de nombreuses études ont montré l'accroissement de l'intelligibilité de la parole par ce codage comparé à la lecture labiale seule [27, 31] et l'apport en terme de facilité d'apprentissage de la langue (orale et écrite) [25, 26]. Pour une revue plus détaillée, le lecteur pourra se référer à [1].

4.2 L'organisation temporelle du code LPC

Beaucoup de travaux de recherche se sont focalisés sur la perception du code LPC, tant dans son intelligibilité qu'au niveau de son impact pour l'acquisition de la phonologie par les sourds (cf. références dans la section précédente). Il existe également des études sur la production du code LPC, qui ont été menées récemment à l'ICP. Attina et al. ont étudié les mouvements des différents articulateurs (main, doigts, lèvres et son) en jeu dans la production du code LPC, en se focalisant notamment sur leurs relations temporelles. Ils ont mis en évidence, à partir d'études sur des corpus de séquences syllabiques Consonne-Voyelle sans sens, un patron temporel de coordination très stable, retrouvé chez quatre codeuses différentes, montrant en particulier une avance des informations manuelles par rapport aux informations labiales et au signal acoustique correspondant [4, 1, 7]. Pour une syllabe CV, la main débute son geste (passage d'une position

[2] la position côté est également utilisée pour coder toute consonne non suivie d'une voyelle ou suivie d'un schwa.

sur le visage à une autre) avant l'émission du son correspondant (cette anticipation est de l'ordre d'une bonne demi-syllabe en moyenne, soit variant de 150 ms à 200 ms suivant le rythme de parole [1]) et arrive en position LPC (qui correspond à la voyelle) au début de la consonne acoustique et donc bien avant l'information labiale correspondante pour la voyelle (cf. figure 4.1). Notons par ailleurs que l'avance de la main par rapport au son avait également été proposée par Duchnowski et al. lorsqu'ils avaient établi des règles empiriques de transition pour leur système de génération automatique de Cued Speech [18]. Ainsi, il semble bien que l'anticipation manuelle soit une règle caractérisant la production du code LPC. Elle est d'ailleurs même maintenue dans des situations où les gestes labiaux anticipent fortement par rapport au signal acoustique, comme c'est le cas dans la coarticulation labiale anticipatoire [5, 10, 6, 1].

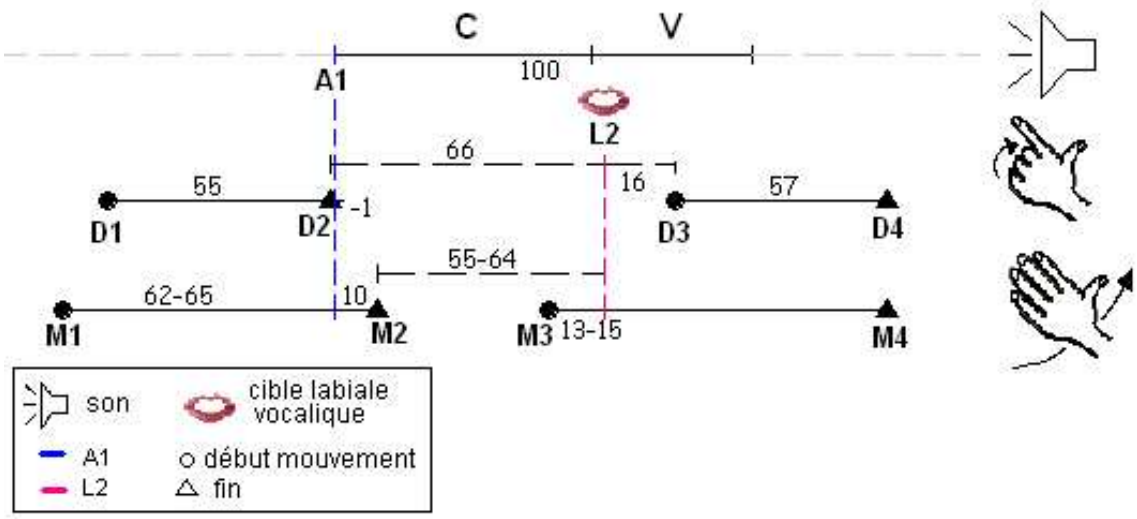


FIG. 4.1 – Schéma de coordination temporelle main-doigts-lèvres-son pour une syllabe CV [1] (les valeurs d'intervalles indiquées sont normalisées par la durée de la syllabe).

En ce qui concerne les clés LPC codant les consonnes, une autre étude [2, 4] a permis d'observer que le mouvement pour atteindre la bonne forme de main commençait avant le début acoustique de la consonne et se terminait en synchronie avec celui-ci (cf. figure 4.1). Contrairement aux règles empiriques de Duchnowski et al. qui proposaient que le changement de clé se produise au milieu du geste de transition de main, Attina et al. ont observé que le geste de formation de la clé prenait en fait une grande partie du mouvement de transition et était complètement «superposé» sur celle-ci.

Lors d'une étude perceptive utilisant le paradigme du *gating*[1] [11], Cathiard et al. ont montré que les sujets sourds effectuent un décodage progressif des informations et exploitent ainsi l'anticipation manuelle mise en évidence dans la production : l'information manuelle (la position et la forme de main) leur permet d'identifier d'abord un groupe de consonnes et de voyelles possibles, puis l'information labiale leur permet de choisir une combinaison CV unique.

[1]Dévoilement progressif d'un signal [23]

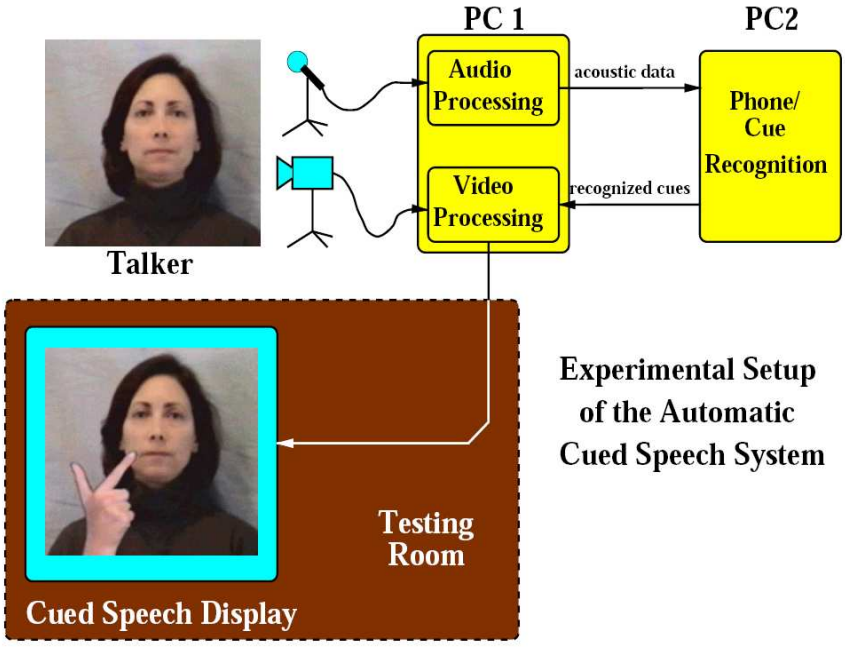


FIG. 4.2 – Système de génération automatique de Cued Speech développé par Duchnowski et al. [17] : un système de reconnaissance détermine les clés correspondant aux sons prononcés, puis les clés discrètes sont superposées sur la vidéo de la locutrice.

4.3 Systèmes de synthèse existants

(a) (b) (c) (d) (e)

FIG. 4.3 – Système de synthèse de LPC proposé par Attina et al. [4] : a) Image de fond utilisée pour les parties fixes de la tête parlante ; b) Maillage 3D de la tête parlante piloté par des paramètres articulatoires ; c) Texture appliquée au maillage défini en b ; d) Image de main pour une clé donnée (ici la clé 1) ; e) Superposition de l'image de main sur la tête parlante.

Le code LPC apporte un gain d'intelligibilité non négligeable par rapport à la lecture labiale seule. On comprend donc l'intérêt de proposer des systèmes de génération automatique de code LPC afin d'aider les sourds partout où ils en ont besoin même s'il n'y a aucun codeur présent. Quelques systèmes de génération ont été proposés :

1. en 1977, le Dr. Cornett propose l'AUTOCUER [15, 13, 14] : les limitations technologiques faisant, un équivalent du code LPC est généré (après un système de reconnaissance de

la parole) puis affiché sur une paire de sept segments de diodes lumineuses, ensuite cette image est projetée près du visage du locuteur. Ce système reste encore éloigné du code LPC manuel ;

2. Il faut attendre 1998, pour voir l'apparition d'un véritable système de génération de code Cued Speech. Ce système, développé par Duchnowski et al. [17, 17, 18], résulte du couplage d'un système de reconnaissance de parole audio et d'un système de génération de Cued Speech (superposition de photos de mains sur la vidéo d'un locuteur). Il nécessite deux micro-ordinateurs, le premier s'occupe de l'enregistrement des données audiovisuelles et de l'affichage post-traitements et le deuxième s'occupe de la transcription du signal audio en clés manuelles via un système de reconnaissance de la parole classique (un diagramme est représenté sur la figure 4.2). Deux types d'animation ont été implémentés : des clés discrètes qui changent instantanément dès que la clé suivante est détectée au niveau du signal acoustique et des changements de clés avec transition : le changement d'une clé à une autre reste discret mais des règles heuristiques (établies à partir de vidéos de codeur) permettent de répartir le temps d'affichage entre le temps en position cible et le temps de la transition. Plus précisément, ils ont observé que la clé était complètement formée avant le son correspondant, ils ont donc avancé l'affichage de la clé de 100ms par rapport à la réalisation acoustique et ils ont aussi remarqué que le changement de forme de la main se déroulait en moyenne au milieu de la transition de position ;

3. Attina et al. [3, 4] propose un système de synthèse LPC basé sur une tête parlante 3D contrôlée par des paramètres articulatoires [30] et doté d'un contrôle moteur de type modèle de coarticulation d'Öhman [19]. Un système de synthèse de parole audiovisuelle à partir du texte génère l'animation d'une tête parlante 3D (la partie basse du visage est animée par des paramètres articulatoires générés par le système TTS (cf. figure 4.3 b) sur laquelle on plaque une texture (cf. figure 4.3 c)). Le système TTS génère une chaîne phonétique marquée en temps, cette information ainsi que les règles déduites de l'analyse permettent de calculer les instants cibles et les périodes de transition des mouvements de la main : l'atteinte de la cible est imposée au début de la réalisation acoustique de la consonne de la CV correspondante et maintenue dans cette position jusqu'au début de la réalisation acoustique de la voyelle de la CV. Ensuite, des images de main (cf. figure 4.3 d) correspondant aux 8 formes de main sont déplacées suivant un modèle sinusoïdal (trajectoire spatiale linéaire et évolution sinusoïdale dans le temps) pour atteindre la position voulue puis superposées sur l'animation de la tête parlante (cf. figure 4.3 e) pour produire la série de clés LPC correspondant au texte d'entrée.

4. le projet LABIAO (Lecture Labiale Assistée par Ordinateur) [21] vise à la création de logiciels permettant d'augmenter l'autonomie des sourds et malentendants dans la vie courante. Ce projet qui est regroupé autour de 7 partenaires : EDF R&D, Audivimédia, INRIA/LORIA, l'équipe Artemis de l'INT, Le LINC, CNEFEI et l'association DATHA (Développement, d'Aides Technologiques pour les Personnes Handicapées) vise plus particulièrement à la mise en oeuvre d'un système de synthèse de code LPC dans un environnement MPEG-4.

Les systèmes présentés ci-dessus sont des systèmes de synthèse basés 2D voire mixte à l'exception du dernier. On y utilise des règles heuristiques, déduites de l'analyse de corpus, pour contrôler l'apparition des clés. Le mouvement de la main lors du passage d'une clé à une autre ne correspond pas à des mouvements naturels : soit il s'agit d'un basculement brutal d'une clé à une autre, soit il s'agit de mouvements sinusoïdaux.

4.4 Résumé

Le Cued Speech et son correspondant français, la Langue française Parlée Complétée, sont une aide à la lecture labiale, puisqu'ils permettent de désambiguïser les formes de lèvres identiques pouvant correspondre à plusieurs réalisations acoustiques différentes. De nombreuses études du code LPC ont été menées, tant dans la perception que dans la production, allant jusqu'à la mise en oeuvre de systèmes de génération automatique de LPC. C'est d'ailleurs lors de ces analyses que des règles de «synchronisation» des mouvements de la main par rapport au visage ont été déterminées et utilisées dans la phase de synthèse. Ces règles soulignent que le code manuel est en avance par rapport au signal audio et par rapport aux lèvres.

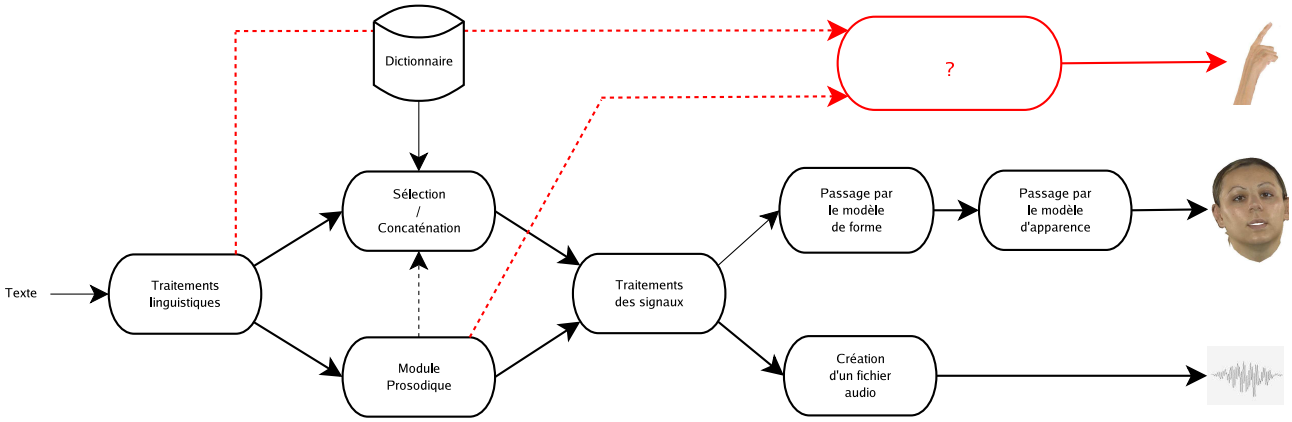


FIG. 4.4 – Décomposition du système de synthèse de parole audiovisuelle augmentée.

Notre tâche est de réaliser un système de synthèse de la parole audiovisuelle (basé modèles 3D) générant de plus le code LPC. La particularité d'un tel système réside dans l'animation de deux objets très différents en termes de degrés de liberté, de mouvements et de coordination. Nous nous sommes déjà fixé comme cadre de réaliser un système de synthèse de parole audiovisuelle par concaténation de polysons multimodaux (paramètres audio et articulatoires) pour nous affranchir de l'étape de post-synchronisation entre le signal audio et les mouvements du visage. Des analyses de coordinations des objets main et visage seront nécessaires pour déterminer s'il sera possible de rajouter aux polysons multimodaux une nouvelle modalité, les paramètres articulatoires de la main. Nous verrons comment nous allons effectuer la synthèse de parole audiovisuelle augmentée à partir du texte (cf figure 4.4).

Dans le chapitre suivant, nous aborderons l'évaluation des systèmes de synthèse en nous attachant plus particulièrement à l'évaluation de l'intelligibilité. En effet, n'oublions pas que notre système a pour but appliqué d'être un substitut au télétexte dans le cadre du projet

ARTUS, il doit par conséquent être aussi performant que lui dans la tâche qui lui incombe : transmettre de l'information.

Références bibliographiques

[1] V. Attina. *La Langue française Parlée Complétée (LPC) : Production et Perception*. PhD thesis, Institut National Polytechnique de Grenoble, 2005.

[2] V. Attina, D. Beautemps, and M.-A. Cathiard. Temporal motor organization of Cued Speech gestures in the French language. In *ICPHS*, pages 1935–1938, 2003.

[3] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. Toward an audiovisual synthesizer for Cued Speech : rules for CV French syllables. In *Audio Visual Speech Processing Workshop*, pages 227–232, St Jorioz, France, 2003.

[4] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables : rules for a Cued Speech synthesizer. *Speech Communication*, 44 :197–214, 2004.

[5] V. Attina, M.-A. Cathiard, and D. Beautemps. Contrôle de l'anticipation vocalique d'arrondissement en Langage Parlé Complété. In *Journées d'Etudes sur la Parole*, 2002.

[6] V. Attina, M.-A. Cathiard, and D. Beautemps. L'ancrage de la main sur les lèvres : Langue française Parlée Complétée et anticipation vocalique. In *Jounées d'Etude sur la Parole*, 2004.

[7] V. Attina, M.-A. Cathiard, and D. Beautemps. Temporal measures of hand and speech coordination during French cued speech production. *Lecture Notes in Artificial Intelligence*, 3881 :13–24, 2006.

[8] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

[9] L. E. Bernstein, M. E. Demorest, and P. E. Tucker. Speech perception without hearing. *Perception & Psychophysics*, 62 :233–252, 2000.

[10] M.-A. Cathiard, V. Attina, and D. Alloatti. Labial anticipation behavior during speech with and without Cued Speech. In *ICPHS*, pages 1939–1942, 2003.

[11] M.-A. Cathiard, F. Bouaouni, V. Attina, and D. Beautemps. Etude perceptive du décours de l'information manuo-faciale en langue française parlée complétée. In *Journées d'Etude sur la Parole*, 2004.

[12] R. O. Cornett. Cued Speech. *American Annals of the Deaf*, 112 :3–13, 1967.

[13] R. O. Cornett. Le Cued Speech. In *Aides manuelles à la lecture labiale et prespectives d'aides automatiques*, Centre scientifique IBM-France, Paris, France, 1982.

[14] R. O. Cornett. 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, 1988.

[15] R. O. Cornett, R. Beadles, and B. Wilson. Automatic Cued Speech. In *Research Conference on Speech Processing Aids for the Deaf*, pages 224–239, Washington, DC : Gallaudet College, 1977.

[16] R.O. Cornett. Adapting Cued Speech to additional languages. *Cued Speech Journal*, V :19–29, 1994.

[17] P. Duchnowski, L. Braida, M. Bratakos, D. Lum, M. Sexton, and J. Krause. Automatic generation of Cued Speech for the deaf : status and outlook. In *AVSP'98*, pages 161–166, Terrigal, Australia, 1998.

[18] P. Duchnowski, D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos, and L. D. Braida. Development of speechreading supplements based on automatic speech recognition. *IEEE Transactions on Biomedical Engineering*, 47(4) :487–496, 2000.

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

[20] O. Engwall and J. Beskow. Resynthesis of 3D tongue movements from facial data. In *EuroSpeech*, Geneva, 2003.

[21] J. Feldmar. Projet labiao (lecture labiale assistée par ordinateur) : Présentation de logiciels augmentant l'autonomie des sourds dans le milieu ordinaire. In *Liaison LPC*, volume 42, pages 159–163, 2005.

[22] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11 :796–804, 1968.

[23] F. Grosjean. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28 :267–283, 1980.

[24] J. Jiang, A. Alwan, L. Bernstein, P. Keating, and E. Auer. On the correlation between facial movements, tongue movements and speech acoustics. In *Proceedings of International Conference on Speech and Language Processing*, pages 42–45, Beijing, China, 2000.

[25] J. Leybaert. Phonology acquired through the eyes and spelling in deaf children. *Journal of Experimental Child Psychology*, 75 :291–318, 2000.

[26] J. Leybaert. The role of Cued Speech in language processing by deaf children : an overview. In *Auditory-Visual Speech Processing*, pages 179–186, St Jorioz, France, 2003.

[27] G. Nicholls and D. Ling. Cued Speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25 :262–269, 1982.

[28] E. Owens and B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28 :381–393, 1985.

[29] C. M. Reed. The implications of the Tadoma method of speechreading for spoken language processing. In *ICSLP*, 1996.

[30] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[31] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of Cued Speech. *Journal of Rehabilitation Research and Development*, 31 :20–41, 1994.

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

Chapitre 5

Evaluation des systèmes de synthèse

Le type d'évaluation de la parole de synthèse dépend avant tout de l'application visée. En effet, le but recherché par un système de synthèse qu'il soit audio (seul) ou audiovisuel, peut varier : la falsification, l'intelligibilité, la compréhension, le naturel... C'est pourquoi, il n'existe pas une seule méthode d'évaluation des systèmes de synthèse de parole. En outre, comme nous l'avons vu dans les chapitres précédents, les systèmes de synthèse se composent de sous-modules aux tâches spécifiques. L'évaluation complète et rigoureuse d'un système de synthèse consisterait à tester tous ces modules par des procédures indépendantes pour établir un diagnostic probant. L'évaluation peut se faire objectivement c'est-à-dire calculer des erreurs par rapport au signal réel sur des paramètres acoustiques (par exemple), ou de manière subjective, c'est-à-dire faire passer des tests perceptifs à des sujets *naïfs*. Cette deuxième méthode reste essentielle puisque l'évaluation objective ne peut rien nous certifier quant à la qualité du signal qui sera perçu, dès lors qu'elle révèle un signal imparfaitement reproduit (encore faut-il connaître le signal original).

Tout commence par la synthèse audio...
Un système TTS qui délivre un signal audio à partir d'une entrée textuelle se compose d'un module de traitements linguistiques, d'un module prosodique et d'un module de génération de signal acoustique. Tester de quelque façon que ce soit le signal acoustique de sortie ne donne qu'un diagnostic partiel de la qualité du synthétiseur. Il faudrait, pour bien faire, séparer tous les modules et les tester de manière indépendante.

Sur laquelle s'ajoute un visage...
Si on rajoute un module visuel sur un système de synthèse de parole audio capable de générer l'articulation d'un visage en fonction de la phrase à prononcer, on augmente le nombre de facteurs à évaluer. En effet, comme on l'a vu dans le chapitre **Synthèse de parole audiovisuelle : les visages parlants**, l'ajout d'une modalité apporte un gain d'intelligibilité si celle-ci est synchrone et cohérente avec la modalité de base. Il faudrait donc tester ces deux critères dans le cadre d'applications visant à apporter de l'intelligibilité. En revanche, si le but du système est le naturel, on oriente les tests en fonction. Il en va de même pour la convivialité.

Et même parfois une main !
Dans le cadre des systèmes de génération automatique de LPC à partir du texte, le but premier est forcément le gain d'intelligibilité et la compréhension. Il faut donc tester ces deux aspects sans oublier toutefois que des personnes sont en contact avec le système, il faut par conséquent

s'assurer que le système ne soit pas visuellement désagréable.

Nous allons voir plus en détails pour les trois types de synthèse (audio, audiovisuelle et audiovisuelle augmentée), les corpus, les paradigmes et les méthodes d'évaluation.

5.1 Synthèse de parole audio

Comme nous l'avons dit précédemment, les systèmes TTS sont composés de trois modules (traitements linguistiques, module prosodique, génération du signal). Le signal acoustique de sortie dépend des résultats et donc de l'efficacité de ces modules. Des méthodes d'évaluation ont été mises au point pour évaluer ces modules séparément les uns des autres [42]. Cependant, le résultat final d'un système TTS est le signal audio de sortie, c'est à lui que l'utilisateur a affaire. L'évaluation de ce dernier peut se faire sous plusieurs angles [35] : évaluation de l'intelligibilité, évaluation du naturel, évaluation de l'opinion (acceptabilité) [32]. Nous allons présenter plus particulièrement les tests d'évaluation de l'intelligibilité ; on distinguera l'intelligibilité au niveau segmental et au niveau des phrases (voire au niveau des paragraphes) (une revue est disponible dans [29] et dans [9]).

5.1.1 Intelligibilité au niveau segmental

L'intelligibilité au niveau segmental est la base de l'évaluation. La question suivante se pose : «Est-on capable de générer un signal audio de mots courts (monosyllabiques, bisyllabiques, etc.) intelligible ?

De nombreux tests d'intelligibilité segmentale existent : on citera en particulier la famille des tests de rime. A l'origine, le **Rhyme Test** a été introduit par Fairbanks [16] pour l'anglais américain. Ce test est composé de 250 mots monosyllabiques séparés en 50 groupes de 5 mots qui ne diffèrent que dans la consonne initiale. L'auditeur doit écrire la première lettre du mot qu'il croit reconnaître.

On trouve également le **Modified Rhyme Test** [21] : une extension du test de rime. Il est composé de 50 listes de six mots monosyllabiques (de type CVC) qui ne diffèrent, à l'intérieur d'une liste, que dans la consonne initiale ou finale. Les sujets doivent choisir, parmi la liste des six mots, celui qui a été prononcé. Les mots sont présentés à l'intérieur d'une phrase porteuse (du type Would you write *test word* now ?). Le mot n'est pas accentué et la phrase porteuse est la même tout au long du test. Ce test permet d'évaluer l'intelligibilité des consonnes initiales et finales, le score pouvant se calculer sur le nombre de mots reconnus, sur le nombre de mots non reconnus ou sur la fréquence de confusion d'une consonne particulière. On le retrouve par exemple dans l'évaluation de MITALK [30] ou plus récemment dans l'évaluation de quatre systèmes TTS [43]. Plusieurs défauts liés à ce test ont été rapportés [36]. En effet il ne contient pas tous les mots différents d'un trait. De plus, il ne contient que des consonnes seules, donc les résultats ne peuvent alors pas être étendus aux clusters de consonnes [1]. Enfin, le type de réponses à choix multiples ne correspond pas à la vie réelle.

Un test de la même famille plus utilisé est le **Diagnostic Rhyme Test** [44] : il se compose

[1] Ensemble de consonnes successives.

dans sa version pour l'anglais américain de 192 mots monosyllabiques de type CVC arrangés en 96 paires. Dans chaque paire, seule la consonne initiale est différente. Pour chaque stimulus, les auditeurs ont le choix entre une paire de mots. Aucune phrase support n'est utilisée. Les avantages de ce test sont la rapidité de passage et les résultats qui peuvent être analysés sous divers angles. Il existe une adaptation en français du test de rime, proposée en 1973 par J. P. Peckels et M. Rossi [28].
Parallèlement au Diagnostic Rhyme Test, qui teste la consonne initiale, le **Diagnostic Alliteration Test** teste de la même manière la consonne finale. Il utilise une liste de 96 paires de mots monosyllabiques qui ne diffèrent que par leur consonne finale. Ces différences sont organisées suivant six catégories et le score de chaque catégorie peut être utilisé pour identifier des problèmes spécifiques du système de synthèse. La moyenne des six scores donne une estimation de l'intelligibilité globale du système. Comme pour le DRT, on présente une paire de mots pour chaque stimulus et on n'utilise pas de phrase support.
De la même façon, le **Diagnostic Medial Consonant Test** teste la consonne médiane à l'aide d'une liste de 96 paires de mots bisyllabiques.

Il existe d'autres tests de l'intelligibilité segmentale, parmi eux le **Standard Segmental Test**. Il se compose d'une liste de mots sans sens de type CV, VC et VCV où les voyelles V appartiennent à [a, i, u] et permet de tester la consonne. Dans le même sens, T. Dutoit et H. Leich [15] comparent des algorithmes de synthèse à l'aide de logatomes de type CVC.
Le **CLuster Identification Test**, développé durant le projet européen ESPRIT, se compose de mots sans sens générés automatiquement en fonction de la probabilité d'apparition dans la langue. Les réponses y sont de type libre.
Le **Bellcore test corpus** [36] est un ensemble de mots monosyllabiques de type CVC qui sont pour la moitié avec sens et pour l'autre moitié sans sens.
Le **Phonetically Balanced Word List** se compose de 20 listes de mots phonétiquement équilibrés. Il s'agit de mots monosyllabiques choisis en fonction de la fréquence des phonèmes dans la langue. Une phrase support est utilisée et la réponse est de type libre. Il permet de tester la consonne initiale, finale et la voyelle médiane.

5.1.2 Intelligibilité au niveau supérieur

Si l'on se place au niveau supérieur, c'est-à-dire au niveau de la phrase voire du paragraphe, on trouve d'autres tests d'intelligibilité. Il s'agit en général de tests où l'on va s'intéresser au nombre de mots clés reconnus par phrase ou, dans le cas de paragraphes, à la compréhension de ceux-ci.

Le test basé sur les phrases **Harvard Psychoacoustics Sentences** se compose de 100 phrases développées pour tester l'intelligibilité des mots en contexte. Ces phrases ont été choisies afin de respecter la fréquence des phonèmes de la langue. Le problème de ce test est qu'il y a un effet d'apprentissage. Un autre test utilise les phrases du corpus **Haskins Sentences**. Il s'agit de phrases développées pour tester l'intelligibilité au niveau des mots ou des phrases. Elles ont été construites de telle sorte que les mots ne puissent pas se déduire du contexte mais comme pour le test précédent, il y a un effet d'apprentissage. Citons enfin, le test SUS **Seman-**

tically Unpredictable Sentences [5, 3, 4], développé pour évaluer la synthèse multilingue, il se compose de mots (monosyllabiques pour la plupart) choisis aléatoirement parmi une liste prédéfinie. Le test contient cinq structures grammaticales à partir desquelles sont construites 20 phrases sémantiquement imprédictibles. Les mots ne peuvent pas être déduits du contexte et leur identification permet de calculer un score d'intelligibilité global du système de synthèse. Il a été par exemple utilisé dans [37] pour évaluer l'impact de la variation du genre de la voix et de la qualité du signal sur l'intelligibilité. Il existe d'autres tests basés sur la reconnaissance de mots clés dans des phrases tels que le test **Bamford-Kowal-Bench** ou le test **IEEE Sentences**.

Il existe également des tests de compréhension où l'auditeur doit répondre à des questions sur le contenu d'un texte qu'il vient d'écouter [24, 31].

Pour une revue plus complète sur toutes ces méthodologies d'évaluation des systèmes TTS, le lecteur pourra se référer à [41, 20].

5.2 Synthèse de parole audiovisuelle

Lorsqu'on rajoute un visage parlant sur un système de synthèse de la parole, on ajoute une information qui peut être redondante ou complémentaire. Ce module visuel, qui a été présenté dans le deuxième chapitre, est composé de sous-modules. Il faudrait, comme pour la synthèse de parole audio, effectuer une évaluation séparée des sous-modules (modèle de contrôle, modèle de forme et modèle d'apparence) pour avoir un diagnostic complet du module visuel. Cependant, dans la majorité des évaluations, on ne s'occupe que du signal multimodal final obtenu en fin de chaîne de synthèse. On évalue l'intelligibilité, l'opinion [2] ou le naturel [18, 38, 39] de ce visage parlant. Comme nous l'avons déjà signalé dans les chapitres précédents, l'ajout de la modalité vidéo n'est pas sans risque, si les deux signaux (audio et vidéo) ne sont pas synchrones et cohérents, il se peut que l'information multimodale présente un biais. Les tests d'intelligibilité utilisés classiquement en synthèse acoustique de la parole sont aisément transposables à l'audiovisuel [19]. En général, les évaluations des têtes parlantes consistent en des tests d'intelligibilité segmentale, le but étant de vérifier l'adéquation des deux signaux ; ceci est validé quand on observe un gain d'intelligibilité par rapport au signal audio seul ou vidéo seul dans un environnement bruité. Nous allons voir plus en détails quelques méthodes d'évaluation de l'intelligibilité des têtes parlantes, d'abord au niveau segmental puis au niveau des phrases voire des paragraphes.

5.2.1 Intelligibilité au niveau segmental

Beaucoup de tests perceptifs visant à évaluer l'apport d'intelligibilité dans le bruit du signal vidéo se basent sur des stimuli de formes simples : on trouve par exemple dans [23, 8] des stimuli de type VCV en contexte vocalique symétrique où seule la consonne est testée, ou des stimuli de type VV avec voyelle répétée dans [23].

On trouve également des stimuli de type VCVCV dans [22] où la réponse porte sur la voyelle et la consonne, ou encore sur des mots monosyllabiques dans [10] où la réponse (de type ouverte) porte sur la consonne initiale, finale, les visèmes, les voyelles et les mots. On trouve encore des

stimuli de type **CV sans sens** dans [6, 17] ou plus surprenant des **mots monosyllabiques** jouant sur l'effet Mc Gurk dans [11]. Ces stimuli très variés ont tous le même but : vérifier que l'ajout d'une modalité est bénéfique au système TTS sous-jacent.

5.2.2 Intelligibilité au niveau supérieur

Comme pour la synthèse de signal de parole audio, des tests se placent au niveau supérieur. Des tests sur des phrases courtes de la vie quotidienne comme les phrases des tests Bamford-Kowal-Bench ou CUNY) sont utilisés dans [18, 7, 34, 33], l'évaluation étant quantifiée sur la reconnaissance au niveau des syllabes, des mots ou des phonèmes. Les problèmes liés à ce type de test comme la prédictibilité des mots ou l'effet d'apprentissage sont absents lorsqu'est utilisé le test SUS comme dans [5, 3].

5.2.3 Intelligibilité et aussi...

Peu d'évaluations systématiques c'est-à-dire des évaluations visant l'intelligibilité, la compré -hension, le naturel et l'acceptabilité ont été menées. On peut citer l'étude effectuée en 1999 au AT&T Labs par Pandzic et al. [27] qui proposait une évaluation de l'intelligibilité, de la compréhension et de l'acceptabilité de trois systèmes de synthèse différents (une tête parlante 3D, une tête parlante 3D texturée et un visage parlant basé images) sur 190 sujets. Cette étude montre que le vidéoréalisme n'est pas la solution idéale au problème de l'intelligibilité et de l'acceptabilité. De plus, les visages synthétiques demandent plus d'efforts cognitifs que la parole naturelle et certains visages synthétiques en demandent encore plus que d'autres (dans le cas de cette expérience, l'animation faciale basée image est plus «coûteuse» que l'animation basée modèle).

Une autre étude visant à évaluer le naturel et l'intelligibilité d'un système de synthèse audio-visuel basé image est présentée dans [18]. Cette étude se compose d'un ensemble de trois tests qui ont pour objectif d'évaluer le naturel des stimuli synthétiques et d'un test d'intelligibilité basé sur la lecture labiale. Bien que les sujets n'aient pu faire la différence entre les stimuli naturels et synthétiques lors de test de Turing, les auteurs ont noté une différence significative dans la reconnaissance des phonèmes, syllabes et mots en faveur des stimuli naturels. La complémentarité des évaluations est indéniable [26], la qualité générale (agrément, naturel) de l'avatar n'est pas un critère suffisant, il doit être complété par d'autres critères d'évaluation.

5.3 Synthèse de parole audiovisuelle augmentée

Dans le cas du Cued Speech ou pour le français du code LPC, le but premier est d'apporter un gain d'intelligibilité. Bien sûr, la tête parlante doit posséder une attractivité certaine pour être utilisée.

Comme on l'a vu dans le chapitre **Synthèse audiovisuelle augmentée**, il n'existe pas un grand nombre de systèmes automatiques capables de générer du code LPC (ou Cued Speech). Nous noterons toutefois que des expériences d'intelligibilité sur du Cued Speech *naturel* ont été

effectuées et qu'elles nous donneront une idée des résultats vers lesquels un système de synthèse devrait tendre.

Les tests de perception menés dans [25] ont montré par rapport à la lecture labiale seule un apport d'informations lorsque le Cued Speech était ajouté. Les stimuli présentés étaient des séries CV ou VC où les consonnes correspondent aux 28 consonnes de l'anglais américain et au sous-ensemble de voyelles à [a, i, u]. Le meilleur résultat d'identification des syllabes revenait à la présentation **lèvres + clés manuelles** avec 83.5% de reconnaissance et le pire pour l'audio seul avec seulement 2.3% de reconnaissance. Un autre test utilisant des mots monosyllabiques familiers dans des phrases a confirmé ces résultats avec une reconnaissance des mots clés de plus de 90% dans le cas des modalités **lèvres + clés manuelles** et **lèvres + clés manuelles + audio**. Avec d'autres types de stimuli (*i.e.* des phrases prédictibles ou peu prédictibles), Uchanski et al. [40] ont eu des résultats similaires avec des taux de reconnaissance compris entre 78% et 97% pour la modalité incluant les lèvres et le Cued Speech contre 21% à 62% pour la lecture labiale seule.

En ce qui concerne les systèmes de synthèse, Duchnowski et al. [12, 13, 14] proposent une évaluation de leur système en utilisant comme stimuli des phrases peu contextuelles et la détection de mots clés. Les taux de reconnaissance sont de 35% en lecture labiale seule contre 66% lorsqu'on lui additionne le Cued Speech. Ces résultats sont bien inférieurs au codage manuel, mais il est à noter que le système de Duchnowski et al. est tributaire des résultats du système de reconnaissance de la parole placé en amont.

Le système de génération automatique de code LPC développé par Attina et al. [1] a été soumis à une première évaluation sur un ensemble de 238 phrases phonétiquement équilibrées par un sujet. Le résultat de l'identification effectuée au niveau CV est de 96.6% de reconnaissance.

5.4 Résumé

L'évaluation des systèmes de synthèse n'est pas un problème résolu. Qu'il soit audio, audiovisuel ou audiovisuel augmenté, un système de synthèse peut et doit être évalué sous différents points de vue (intelligibilité, compréhension, naturel, acceptabilité) afin d'avoir un bon diagnostic de son efficacité et de ses lacunes. On peut toutefois signaler que la tâche d'évaluation est conditionnée par l'application visée. Ainsi dans le cas de ces travaux de thèse, l'application visée étant la substitution du télétexte par une tête parlante capable de générer des mouvements faciaux et du code LPC, l'évaluation se portera plus spécifiquement sur la capacité à apporter un gain d'intelligibilité. Il ne s'agit toutefois pas de la seule évaluation possible mais elle est nécessaire.

Références bibliographiques

[1] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables : rules for a Cued Speech synthesizer. *Speech Communication*, 44 :197–214, 2004.

[2] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

[3] C. Benoît. An intelligibility test using Semantically Unpredictable Sentences : towards the quantification of linguistic complexity. *Speech Communication*, 9 :293–304, 1990.

[4] C. Benoît, M. Grice, and V. Hazan. The SUS test : A method for the assessment of text-to-speechs synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18 :381–392, 1996.

[5] C. Benoît, A. van Erp, M. Grice, V. Hazan, and U. Jekosch. Multilingual synthesiser assessment using semantically unpredictable sentences. In *Proceedings of Eurospeech'89*, volume 2, pages 633–636, 1989.

[6] L. E. Bernstein, M. E. Demorest, and P. E. Tucker. Speech perception without hearing. *Perception & Psychophysics*, 62 :233–252, 2000.

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

[8] J. Beskow, M. Dahlquist, B. Granstrom, M. Lundeberg, K.-E. Spens, and T. Ohman. The Teleface project - multimodal speech communication for the hearing impaired. In *Eurospeech*, pages 2003–2010, Rhodos, Greece, 1997.

[9] Calliope. *La Parole et son Traitement Automatique*. 1989.

[10] M. M. Cohen, R. L. Walker, and D. W. Massaro. Perception of synthetic visual speech. In *Speechreading by Man and Machine : Models, Systems and Application, NATO Advanced Study Institute 940584*, Chateau de Bonas, France, August 1995.

[11] D. Cosker, S. Paddock, D. Marshall, P. L. Rosin, and S. Rushton. Towards perceptually realistic talking heads : Models, methods and McGurk. In *ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, California, USA, 2004.

[12] P. Duchnowski, L. Braida, M. Bratakos, D. Lum, M. Sexton, and J. Krause. Automatic generation of Cued Speech for the deaf : status and outlook. In *AVSP'98*, pages 161–166, Terrigal, Australia, 1998.

[13] P. Duchnowski, L. Braida, D. Lum, M. Sexton, J. Krause, and S. Banthia. A speechreading aid based on phonetic ASR. In *5th International Conference on Spoken Language Processing*, volume 7, pages 3289–3292, Sydney, Australia, 1998.

[14] P. Duchnowski, D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos, and L. D. Braida. Development of speechreading supplements based on automatic speech recognition. *IEEE Transactions on Biomedical Engineering*, 47(4) :487–496, 2000.

[15] T. Dutoit and H. Leich. Synthèse de parole de haute qualité à partir d'un texte : une comparaison de quatre algorithmes candidats. In *Journées d'Etudes sur la Parole*, 1994.

[16] G. Fairbanks. Test of phonemic differentiation : the rhyme test. *Journal of the Acoustical Society of America*, 30(7) :596–600, July 1958.

[17] A. Faulkner and S. Rosen. The contribution of temporally-coded acoustic speech patterns to audio-visual speech perception in normally hearing and profoundly hearing-impaired

listeners. In *Proceedings of the Workshop on the Auditory Basis of Speech Perception*, pages 261–264, 1996.

[18] G. Geiger, T. Ezzat, and T. Poggio. Perceptual evaluation of video-realistic speech. Technical Report CBCL Paper 224/ AI Memo, Massachusetts Institute of Technology, Cambridge, USA, February 2003.

[19] D. Gibbon, R. Moore, and R. Winski. *Handbook of standards and resources for spoken language systems*. Mouton de Gruyter, 1997.

[20] M. Goldstein. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication*, 16(3) :225–244, 1995.

[21] A. S. House, C. E. Williams, M. H. Hecker, and K. D. Kryter. Articulation testing methods : Consonantal differantiation with a closed-response set. *Journal of the Acoustical Society of America*, 37 :158–166, 1965.

[22] B. Le Goff, T. Guiard-Marigny, and C. Benoit. *Progress in speech synthesis*, chapter Analysis-synthesis and intelligibility of a talking face, pages 235–246. Van Santen, J.P.H. and Sproat, R.W. and Olive, J.P. and Hischberg, J., Berlin, 1996.

[23] R. Möttönen, J.-L. Olivès, J. Kulju, and M. Sams. Parametrized visual speech synthesis and its evaluation. In *European Signal Processing Conference*, Tampere, Finland, 2000.

[24] L. Neovius and P. Raghavendra. Comprehension of KTH text-to-speech with l̀istening speedṕaradigm. In *EUROSPEECH'93*, pages 1687–1690, Berlin, Germany, 1993.

[25] G. Nicholls and D. Ling. Cued Speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25 :262–269, 1982.

[26] M. Odisio. *Estimation des mouvements du visage d'un locuteur dans une séquence audiovisuelle*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.

[27] I. Pandzig, J. Ostermann, and D. Millen. Users evaluation : synthetic talking faces for interactive services. *The Visual Computer*, 15 :330–340, 1999.

[28] J. P. Peckels and M. Rossi. Le test de diagnostic par paires minimales. *Revue d'Acoustique*, 27 :245–262, 1973.

[29] D. B. Pisoni, B. G. Greene, and J. S. Logan. An overview of ten years of research on the perception of synthetic speech. In *ESCA Workshop on Speech Input/Output Assessment and Speech Databases*, 1989.

[30] D. B. Pisoni and S. Hunnicutt. Perceptual evaluation of MITALK : the MIT unrestricted text-to-speech system. *ICASSP*, pages 572–575, 1980.

[31] D. B. Pisoni, H. C. Nusbaum, and B. G. Greene. Perception of synthetic speech generated by rule. *IEEE*, 73 :1665–1676, 1985.

[32] R. Prudon and C. d'Alessandro. A selection/concatenation text-to-speech synthesis system : databases development, system design, comparative evaluation. In *4th ISCA ITRW on Speech Synthesis*, 2001.

[33] C. Siciliano, A. Faulkner, and G. Williams. Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners. In *ISCA Tutorial and Research Workshop on Audio Visual Speech Processing*, St Jorioz, France, 2003.

[34] C. Siciliano, G. Williams, J. Beskow, and A. Faulkner. Evaluation of a synthetic talking face as a communication aid for the hearing impaired. *Speech, Hearing and Language : Work in Progress*, 14 :51–61, 2002.

[35] C. Sorin and F. Emerard. Domaines d'application et évaluation de la synthèse de parole à partir du texte. In *Fondements et perspectives en traitement automatique de la parole.* AUPELF UREF, 1996.

[36] M. F. Spiegel, M. J. Altom, and M. J. Macchi. Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. *Speech Communication*, 9 :279–291, 1990.

[37] C. Stevens, N. Lees, J. Vonwiller, and D. Burnham. On-line experimental methods to evaluate text-to-speech (TTS) synthesis : effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language*, 19 :129–146, 2005.

[38] B.-J. Theobald, J. A. Bangham, I. Matthews, and G. Cawley. Evaluation of a talking head based on appearance models. In *Audio-Visual Speech Processing*, September 2003.

[39] B. J. Theobald, J. A. Bangham, I. A. Matthews, and G. C. Cawley. Near-videorealistic synthetic talking faces : implementation and evaluation. *Speech Communication*, 44 :127–140, 2004.

[40] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of Cued Speech. *Journal of Rehabilitation Research and Development*, 31 :20–41, 1994.

[41] R. Van Bezooijen and L. C. W. Pols. Evaluating text-to-speech systems : some methodological aspects. *Speech Communication*, 9 :263–270, 1990.

[42] J. Van Santen. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech & Language*, 7 :49–100, 1993.

[43] H. S. Venkatagiri. Segmental intelligibility of four currently used text-to-speech synthesis methods. *Journal of the Acoustical Society of America*, 113(4) :2095–2014, April 2003.

[44] D. W. Voiers. *The Diagnostic Rhyme Test.* PhD thesis, TRACOR, 1970.

Chapitre 6

Résumé de la partie

Cette partie nous a donné un état de l'art de la synthèse de parole, qu'elle soit audio seule ou multimodale. Cet état de l'art n'est pas exhaustif mais décrit les différentes voies abordées pour mettre en oeuvre des systèmes de synthèse de la parole.

Nous avons vu que la synthèse de parole se base en général sur un système de synthèse audio auquel on rajoute ou non un module visuel. Le système de base, celui capable de générer un signal audio à partir d'un texte en entrée, a été décomposé et l'activité des modules le constituant soulignée.

Ensuite, le module visuel qui va rajouter un visage au système TTS a été défini et décomposé en sous-modules. Nous avons vu que l'information nécessaire à la génération de mouvements faciaux pouvait être diverse (phonétique, audio) et que les méthodes de génération de mouvements ainsi que leur implémentation étaient variées, en séparant toutefois les approches basées 3D des approches basées 2D.

Le but de ces travaux de thèse étant d'implémenter un système de synthèse de code LPC à partir du texte, nous avons présenté la Langue française Parlée Complétée et les sujets de recherche rattachés, parmi eux l'implémentation de synthétiseurs.

Enfin, tout bon système de synthèse doit être associé à une critique. C'est pourquoi nous avons présenté les différents types d'évaluation tant au niveau audio, qu'audiovisuel et audiovisuel augmenté (cas du LPC).

Dans notre travail, le système de synthèse de la parole sera basé sur un système de type concaténation d'unités. Ces unités seront multimodales, c'est-à-dire à la fois audio et articulatoires, ce qui nous permettra de conserver une cohérence temporelle et ne nous obligera pas à effectuer de la post-synchronisation. Ces unités, en partie articulatoires, nous serviront à piloter un visage basé 3D. Enfin, l'apparence finale du visage parlant se basera sur un modèle de rendu de type plaquage de texture. Pour ce qui est de l'animation de la main, nous utiliserons un système de concaténation d'unités afin d'avoir un ensemble cohérent mais nous devrons attendre les analyses de la production du code LPC à partir de nos corpora de synthèse pour avoir une idée plus précise des verrous et des solutions que nous leur apporterons. En ce qui concerne l'évaluation, la tâche étant de remplacer une information textuelle par notre visage parlant codant le LPC, nous devrons nous assurer avant tout que l'intelligibilité de notre système est suffisante afin de pouvoir effectuer le changement. En outre, nous testerons la charge cognitive

nécessaire à la compréhension de cette nouvelle interface homme-machine. En effet, nous ne savons pas *a priori* quel temps est nécessaire au décodage LPC et quel temps est nécessaire à la lecture du même discours.

Deuxième partie

De l'analyse à la synthèse

Chapitre 7

L'ingrédient essentiel : le corpus

Définissons notre but !

La base essentielle de tout système de synthèse qu'il soit par règles ou par concaténation est le corpus. En effet, dans le premier cas, les règles sont dépendantes du corpus et vont être déterminées par la façon dont il a été conçu. Dans le second cas, c'est encore plus flagrant : sans corpus, il n'y a pas de synthèse possible puisque ce sont des éléments de ce corpus, des *briques* élémentaires, qui vont être concaténés. Il faut donc enregistrer, traiter et analyser avec le plus grand soin ce corpus sans quoi la qualité de la synthèse sera dégradée. Une fois l'enregistrement fini, il faudra enchaîner des phases de pré-traitements comme par exemple, la segmentation du signal audio ou le *nettoyage* des données, puis des phases d'analyses (on ne peut en effet pas concaténer des *briques* élémentaires sans un minimum de connaissances sur celles-ci). Le but de ces travaux est de synthétiser la Langue française Parlée Complétée dans ses trois modalités : l'audio, les mouvements du visage et les mouvements de la main. Pour cela, nous avons décidé de faire de la synthèse par concaténation d'unités audiovisuelles. Notre corpus, c'est-à-dire notre base, doit être (s'il n'y avait aucune limitations technologiques) un ensemble de phrases prononcées et codées par un(e) locuteur(rice). Il doit être constitué de telle sorte que l'on puisse déterminer les trajectoires du visage et de la main avec une grande précision temporelle et spatiale ainsi que le signal audio correspondant. Ce corpus doit, en outre, avoir pour spécificité intrinsèque de couvrir l'ensemble des unités élémentaires nécessaires à la production du code LPC.

Verrous technologiques principaux

Le premier verrou technologique vient de la couverture des unités minimales. En effet, les éléments constituant le corpus doivent les couvrir au mieux tout en respectant une limitation de taille (la locutrice ne pouvant subir un enregistrement de plusieurs jours).

Le second verrou correspond à la capacité à déterminer avec précision (temporelle et spatiale) les mouvements du visage et de la main. En effet, il est actuellement impossible d'avoir en même temps une résolution spatiale de l'ordre du mm sur plusieurs centaines de points et une résolution fréquentielle de plus de 100 Hz.

Solution envisagée

La solution que nous proposons consiste à décomposer le travail : il est en effet possible d'enregistrer un grand nombre de marqueurs à fréquence faible (de l'ordre de 50 Hz) et d'enregistrer

un faible nombre de marqueurs à vitesse élevée. Il ne reste plus qu'à combiner ces deux enregistrements. Nous avons dans un premier temps enregistré un corpus dynamique avec peu de marqueurs (de l'ordre de quelques dizaines sur la main et le visage). Ce corpus est la base de notre dictionnaire d'unités multimodales et il nous sert également à étudier le phénomène de production du code LPC. Dans un deuxième temps, nous avons enregistré un corpus *statique* avec de nombreux marqueurs sur les deux objets que sont la main et le visage afin de compléter nos données spatiales. Ce point sera abordé ultérieurement dans le chapitre **Passage à la haute définition et à l'apparence**.

Verrous technologiques de la solution

Même si lors de l'enregistrement le nombre de capteurs est peu élevé alors que le nombre de caméras est élevé, il se peut que la main et le visage se masquent partiellement et donc gêner l'obtention des coordonnées de certains marqueurs réfléchissants. C'est un nouveau problème qui est inhérent à toute capture mettant en jeu plusieurs objets évoluant au même instant. Nous verrons dans le chapitre suivant comment régler ce nouveau problème. Cependant, nous pouvons noter dès à présent qu'en plus du corpus dynamique, nous avons enregistré lors de la même session deux corpora supplémentaires qui nous seront utiles par la suite pour le passage de la basse définition à la haute définition.

7.1 Description des corpora dynamiques

Lors de cette session d'enregistrement, nous allons avoir affaire à trois corpora, les deux premiers sont des corpora d'«appoint» où le visage et la main seront enregistrés séparément, le troisième quant à lui est le corpus dynamique qui nous intéresse plus particulièrement. Nous distinguerons donc les corpora **visage seul** et **main seule** du corpus **main+visage**.

Corpus visage seul

Ce corpus se compose de 34 stimuli : 10 voyelles isolées et 8 consonnes prononcées suivant 3 contextes vocaliques symétriques (VCV). Il s'agit d'un corpus largement éprouvé à l'Institut de la Communication Parlée (ICP) pour la construction des clones [1].

Les 10 voyelles isolées font partie du groupe suivant : [a], [e], [ɛ], [i], [u], [y], [ø], [œ], [o], [ɔ] . Pour les transitions VCV, la consonne C appartient à l'ensemble [p], [t], [k], [f], [s], [ʃ], [l], [ʁ] et la voyelle V à l'ensemble [a], [i], [u] .

Corpus main seule

Ce corpus se compose de l'ensemble des transitions possibles entre 2 clés consonantiques. Pour cela, nous avons construit des logatomes de type CV**CV**CV, où le groupe **CV** central est en contexte symétrique. La voyelle est identique pour les trois séries CV, il n'y a donc aucun mouvement de transition de main. Le tableau 7.1 décrit l'ensemble des logatomes utilisés. Il nous permet d'avoir accès à tous les changements de forme de main possibles avec la coarticulation associée.

de / vers	clé 1	clé 2	clé 3	clé 4	clé 5	clé 6	clé 7	clé 8
clé 1	pøpøpø	køpøkø	søpøsø	bøpøbø	tøpøtø	løpøløø	gøpøgø	jøpøjø
clé 2	pøkøpø	køkøkø	søpøsø	bøkøbø	tøkøtø	løkøløø	gøkøgø	jøkøjø
clé 3	pøsøpø	køsøkø	søkøsø	bøsøbø	tøsøtø	løsøløø	gøsøgø	jøsøjø
clé 4	pøbøpø	købøkø	søbøsø	bøbøbø	tøbøtø	løbøløø	gøbøgø	jøbøjø
clé 5	pøtøpø	køtøkø	søtøsø	bøtøbø	tøtøtø	løtøløø	gøtøgø	jøtøjø
clé 6	pøløpø	køløkø	søløsø	bøløbø	tøløtø	løløløø	gøløgø	jøløjø
clé 7	pøgøpø	køgøkø	søgøsø	bøgøbø	tøgøtø	løgøløø	gøgøgø	jøgøjø
clé 8	pøjøpø	køkøkø	søjøsø	bøjøbø	tøjøtø	løjøløø	gøjøgø	jøjøjø

TAB. 7.1 – Ensemble des logatomes du corpus **main seule**.

Corpus main + visage

Le corpus **main+visage** est composé d'un ensemble de 238 phrases phonétiquement équilibrées qui couvrent l'ensemble des diphones du français. La liste des phrases utilisées est disponible en annexe A. Si l'on rentre plus en détails, on se rend compte que la moitié des diphones ne sont présents qu'une fois. Ce corpus dynamique va nous permettre la construction de dictionnaires multimodaux qui seront à la base du système d'animation. De plus, il nous permet d'étudier la prosodie de la locutrice-codeuse contrairement à un corpus composé de logatomes.

7.2 L'enregistrement

Nous allons maintenant décrire le déroulement et les caractéristiques des enregistrements de chacun des corpora.

La locutrice - codeuse

La codeuse (20 ans au moment de l'enregistrement) pratique quotidiennement la Langue française Parlée Complétée depuis 7 ans avec sa jeune soeur sourde. Elle effectue également du codage en lycée pour d'autres sourds. Il s'agit d'une personne entendante et oralisante. Elle n'a pas encore le diplôme de codeuse professionnelle pour des raisons de disponibilité mais nous a été recommandée par le service d'orthophonie du service ORL du CHU de Grenoble. Elle suit une formation de linguistique à Grenoble, a de bonnes connaissances en phonétique et souhaite avoir le diplôme de codeur très prochainement.

Le matériel

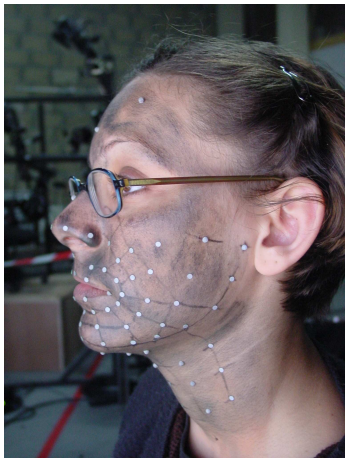
La phase d'enregistrement s'est déroulée dans les locaux d'Attitude Studio[1]. Une première phase a consisté à valider et calibrer le principe d'enregistrement par capture du mouvement

[1]Attitude Studio est une entreprise leader dans le domaine des agents virtuels et de l'animation par capture de mouvements.
Attitude Studio SA, Bât. 126 - 50 avenue du Président Wilson 93 214 St Denis-La Plaine CEDEX, France
`http://www.attitude-studio.com`

optique des gestes de la Langue française Parlée Complétée. Pour effectuer la capture de mouvement optique, nous avons utilisé des capteurs rétro-réfléchissants (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 ont été placés sur le visage et la main de la codeuse comme représenté sur la figure 7.1. 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 côté 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



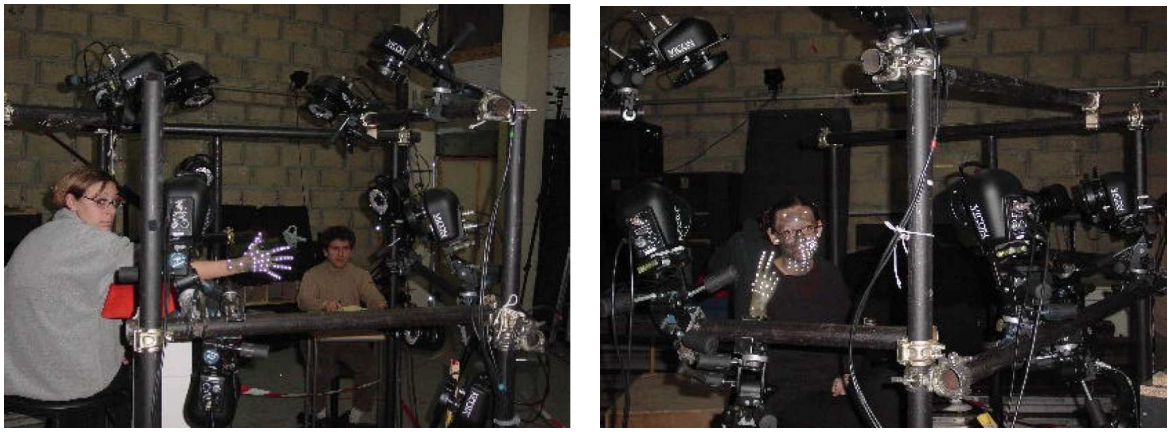
(b) visage : vue de profil

(c) main : vue du dessus

FIG. 7.1 – Position des marqueurs sur la codeuse lors de l'enregistrement.

Outre les marqueurs, le système de caméras a été disposé selon deux configurations différentes en fonction des corpora à enregistrer afin d'éviter les occlusions. Ainsi, une première disposition des caméras a été mise en place pour l'enregistrement du corpus **main seule** et une deuxième pour l'enregistrement des corpora **visage seul** et **main + visage** comme représentées sur la figure 7.2. Dans le cas du corpus **main seule**, on a pu imposer un axe principal à la main :

elle était positionnée de telle sorte qu'en position poing fermé pouce ouvert, celui-ci se trouve selon la verticale. Ainsi, les mouvements de rotation des doigts se trouvent alors dans un plan horizontal. Les caméras ont ainsi été disposées suivant deux arcs de cercles horizontaux et des caméras supplémentaires ont été rajoutées pour pouvoir suivre le pouce. Dans le cas des corpora **visage seul** et **main + visage**, une configuration dissymétrique a été utilisée pour tenir compte du mouvement de la main droite lors du codage.



(a) configuration **main seule** (b) configurations **visage seul** et **main + visage**

FIG. 7.2 – Configurations des caméras pour les enregistrements.

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

Le protocole d'enregistrement

Les phrases du corpus étaient d'abord présentées sur un écran placé en face de la codeuse. Puis une personne énonçait la phrase à haute voix à un rythme normal d'élocution. La locutrice-codeuse prononçait et codait cette phrase. Après chaque phrase, on passait immédiatement à la phrase suivante. En cas d'erreur (évaluée par la codeuse uniquement) la phrase était mise de côté 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 a été utilisée alors que pour la main, la codeuse a conservé les marqueurs sur la main à l'aide d'un gant entre l'enregistrement du corpus **main seule** et du corpus **main + visage**.

7.3 Avantages et inconvénients de ces corpora

Les corpora **main seule** et **visage seul** sont des corpora annexes qui nous serviront par la suite. Le corpus **main + visage**, quant à lui, nous sert à synthétiser toute nouvelle entrée textuelle. Il doit donc posséder certaines caractéristiques que nous allons étudier plus en détails.

Répartition des diphones

Ce corpus se compose de 238 phrases phonétiquement équilibrées (référencées en annexe A). Il a été construit pour mettre en oeuvre des systèmes de synthèse de parole par concaténation de polysons[2] (multimodaux). Les polysons minimaux, les diphones, sont donc présents au moins une fois dans le corpus. Ce corpus se compose de 1814 polysons distincts et de 7279 polysons au total. La moitié des polysons sont multi-représentés. Il s'agit d'un bon compromis compte tenu de la taille réduite du corpus.

Répartition des diclés

En ce qui concerne la main, nous avons au moins une fois toutes les transitions de main, tant au niveau de la forme que de la position (*cf.* tableau 7.2 et 7.3, les formes de main sont au nombre de 8 plus une forme «repos» *cf.* Tab. 4.1 ; les positions de la main par rapport au visage sont au nombre de 5 plus une position «repos» *cf.* Tab. 4.2). En revanche, toutes les transitions $(forme + position)_1$ vers $(forme + position)_2$ ne sont pas présentes (cf. tableau 7.4). Elles sont en effet en très grand nombre (1680 transitions) et il serait incongru de vouloir toutes ces transitions dans un corpus de taille raisonnable. Nous verrons par la suite comment nous proposerons de compléter les transitions CV_1 vers CV_2 manquantes. Nous nommerons ces transitions, dès à présent, des diclés[3] par analogie avec les diphones, afin de pouvoir générer toutes les transitions possibles du code LPC.

de/vers	0	1	2	3	4	5	6	7	8
0	0	36	12	26	11	80	69	1	3
1	27	38	55	99	49	118	62	14	22
2	27	44	45	65	41	95	69	7	29
3	47	89	68	74	61	157	67	11	30
4	28	43	38	47	23	74	75	13	20
5	47	123	94	183	105	244	130	15	31
6	37	83	87	71	48	138	41	17	24
7	7	5	9	20	10	13	16	3	1
8	18	23	14	19	13	53	17	3	5

TAB. 7.2 – Nombre de représentants lors des transitions de forme à forme. La forme 0 correspond à la forme de la main en début et fin de phrase (position «repos»).

[2]Portion de signal comprise entre deux allophones stables successifs c'est-à-dire similaires aux diphones mais en excluant les glides comme allophones stables.

[3]Portion de signal (paramètres articulatoires de la main et paramètres de roto-translation de la tête et de la main) comprise entre deux clés LPC successives.

de/vers	0	1	2	3	4	5
0	0	70	34	27	51	56
1	127	689	305	172	168	246
2	32	306	77	47	39	91
3	14	288	22	21	10	25
4	19	142	78	46	42	47
5	46	212	76	67	64	120

TAB. 7.3 – Nombre de représentants lors des transitions de position à position. La position 0 correspond à la position de la main en début et fin de phrase (position «repos»).

7.4 Résumé

Le corpus, à la base de tout système de synthèse, nécessite de nombreuses précautions tant dans l'élaboration que dans l'enregistrement. Par rapport à notre problématique, nous avons construit non pas un seul et unique corpus mais un ensemble de corpora complémentaires. Il est composé de corpora dynamiques à haute résolution fréquentielle (120 Hz). Le corpus «idéal» pour créer un dictionnaire d'unités pour la synthèse de code LPC doit contenir plus de 1680 unités («diclés»). Il s'agit là d'un nombre prohibitif en terme de temps d'enregistrement et de pré-traitements. Par conséquent, nous avons choisi de nous baser sur un corpus construit à l'origine pour de la synthèse de parole audiovisuelle par concaténation. Ce corpus, plus léger, comporte des «trous» dans les unités nécessaires à la synthèse des mouvements du code LPC. Nous verrons par la suite comment résoudre ce problème.

Dans le chapitre suivant, nous étudierons les pré-traitements appliqués aux différents corpora afin d'obtenir des données «utiles» c'est-à-dire des données prêtes à être utilisées par le système de synthèse.

Références bibliographiques

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

	0	11	12	13	14	15	21	22	23	24	25	31	32	33	34	35	41	42	43	44	45	51	52	53	54	55	61	62	63	64	65	71	72	73	74	75	81	82	83	84	85
0	-	7	6	2	9	12	6	2	4	-	-	3	1	6	10	6	6	1	4	-	-	17	22	8	2	31	29	2	1	30	7	1	-	-	-	-	1	-	2	-	-
11	13	9	4	4	3	2	8	6	1	-	4	26	6	4	1	7	9	3	1	2	3	8	8	1	3	11	10	4	2	7	1	3	-	-	-	1	3	1	1	3	3
12	4	1	2	-	-	-	4	-	-	-	8	4	4	1	1	2	5	1	1	-	1	3	4	1	1	2	8	-	1	5	2	-	-	-	-	-	11	-	-	-	-
13	3	1	1	-	-	-	3	-	-	-	-	19	-	-	-	2	5	-	-	-	1	4	1	-	-	2	4	-	-	-	-	1	-	1	-	-	-	-	-	-	-
14	1	3	-	-	1	1	5	1	-	-	1	4	2	2	1	-	1	1	1	-	-	7	15	3	2	9	6	-	2	-	-	2	-	-	-	-	-	-	-	-	-
15	6	2	1	1	1	1	8	-	3	1	2	2	5	-	4	2	4	5	3	-	2	9	7	5	5	7	6	1	-	-	3	2	1	1	1	1	-	-	-	-	-
21	18	5	6	3	3	5	6	5	3	-	2	19	6	5	2	10	8	7	3	-	2	16	6	7	3	16	29	6	1	5	5	5	-	-	-	-	7	2	1	6	2
22	2	5	3	1	2	3	5	1	-	-	-	-	-	-	1	-	2	-	1	-	-	8	6	1	1	8	2	1	1	-	-	1	-	-	-	-	9	-	-	-	-
23	4	-	-	2	-	-	15	-	1	1	1	12	-	-	-	-	2	-	1	-	-	4	-	-	2	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24	-	-	1	-	1	2	1	-	-	-	-	1	-	-	-	-	1	-	-	1	1	1	4	-	-	1	2	1	-	2	-	1	-	-	-	-	1	-	-	-	-
25	3	-	-	-	1	1	2	1	-	-	1	3	1	2	1	2	7	-	-	2	3	1	4	1	1	4	6	-	1	1	2	-	-	-	-	-	1	-	-	-	-
31	29	18	8	5	9	16	19	5	2	2	5	9	7	6	4	5	10	11	2	2	2	32	17	7	3	17	16	7	3	9	5	3	1	-	-	1	4	1	-	10	2
32	7	7	3	1	1	-	12	1	-	1	-	2	1	1	1	2	6	1	-	-	4	15	4	4	1	3	5	-	-	1	1	-	-	-	-	-	8	-	-	1	-
33	-	5	2	-	-	-	4	1	-	-	2	2	2	-	-	-	7	1	-	-	-	16	1	1	-	2	5	-	1	1	1	-	-	-	-	-	1	-	-	-	-
34	2	4	-	-	2	1	4	-	3	2	1	2	1	2	-	3	1	3	1	1	1	5	3	5	-	1	4	1	1	-	-	2	-	-	-	-	-	-	1	-	-
35	9	-	2	1	1	3	2	1	-	-	1	15	1	-	1	7	2	4	1	-	1	3	2	5	5	5	4	-	-	2	-	2	-	-	-	2	-	2	-	-	-
41	12	10	4	2	4	2	10	2	3	-	3	12	6	1	2	3	2	6	4	1	-	9	11	2	2	7	23	16	3	7	3	5	-	-	1	-	2	-	-	6	-
42	3	3	3	3	-	2	3	3	-	2	1	4	3	1	-	-	3	-	1	-	-	11	3	3	2	-	6	1	-	4	3	2	-	-	-	1	7	-	-	-	2
43	5	2	1	1	-	2	3	-	-	-	-	10	-	-	1	-	3	-	1	-	-	3	-	2	1	1	5	-	-	-	-	3	-	-	-	-	2	-	-	-	-
44	1	3	-	-	-	-	3	-	-	1	-	1	-	-	1	-	1	1	-	-	-	2	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
45	7	-	-	-	1	-	4	-	-	-	-	1	-	-	-	1	-	-	-	-	-	2	1	-	2	6	2	-	-	1	1	1	-	-	-	-	-	-	-	-	-
51	24	11	8	4	8	8	15	2	13	3	3	47	20	6	7	5	9	5	1	4	-	26	14	3	6	15	33	6	6	7	7	7	-	1	-	-	6	1	4	1	3
52	8	14	4	1	-	8	15	2	4	-	3	8	3	1	1	5	13	3	4	1	3	24	5	4	6	15	22	4	3	1	4	2	1	-	-	-	10	-	-	-	-
53	-	5	1	1	-	1	4	-	1	-	-	34	1	1	1	1	14	-	-	-	1	16	5	3	1	4	13	1	2	-	-	-	-	-	-	-	3	-	-	-	-
54	7	7	-	4	1	2	6	3	-	-	-	4	2	1	7	1	1	-	1	-	-	8	6	3	3	4	1	1	1	-	-	-	-	-	-	-	-	-	-	-	1
55	8	15	2	4	9	5	7	3	4	2	4	11	3	6	-	7	35	4	3	2	1	21	10	12	9	21	1	1	7	3	6	4	-	-	-	-	1	1	-	-	-
61	23	13	6	2	6	8	27	8	3	5	5	21	1	3	5	3	10	9	7	-	3	27	12	13	2	11	7	5	4	4	5	6	-	3	-	-	4	-	3	1	2
62	4	9	1	1	1	-	4	2	-	-	1	4	1	1	-	1	4	1	1	-	1	5	4	1	1	1	2	-	-	-	1	1	-	-	-	-	4	-	1	1	2
63	1	2	2	-	-	-	8	-	-	-	-	13	-	-	-	1	3	1	-	-	-	6	-	-	-	2	2	-	2	-	-	2	-	-	-	-	3	-	-	-	-
64	2	10	-	3	3	5	6	5	2	1	2	3	5	1	1	2	4	2	-	2	-	10	9	5	4	5	1	1	1	-	-	3	-	-	-	-	1	-	-	-	-
65	7	4	2	1	2	2	5	3	-	-	-	1	-	-	2	2	-	-	-	-	-	9	2	3	2	4	2	1	1	-	2	1	1	-	-	-	1	-	-	-	1
71	2	3	-	-	-	-	5	2	1	-	1	8	6	2	-	1	3	4	-	-	-	4	2	2	1	2	9	-	3	1	1	2	-	-	-	-	-	1	-	-	-
72	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-
73	1	1	-	-	-	-	-	-	-	-	-	2	-	-	-	-	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
74	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
75	2	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
81	6	5	2	-	1	5	4	3	1	-	1	3	3	1	1	1	1	4	3	1	-	8	8	4	2	10	6	1	1	2	1	2	-	-	-	-	2	-	1	-	-
82	2	-	-	-	-	1	-	1	-	-	-	-	-	1	1	-	1	-	-	-	-	1	-	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-
83	0	1	-	-	-	-	-	-	-	-	-	4	1	-	-	-	2	-	-	-	-	2	-	-	-	1	1	-	-	1	-	1	-	-	-	-	1	-	-	-	-
84	6	1	2	-	1	-	2	1	-	1	-	-	1	-	-	1	-	-	-	-	-	5	1	1	1	2	-	-	-	2	-	-	-	-	-	-	1	-	-	-	-
85	4	-	-	-	-	4	-	-	-	-	-	-	-	-	-	1	-	-	1	-	-	1	2	1	-	2	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-

TAB. 7.4 – Nombre de représentants lors des transitions de forme + position vers une autre forme + position. Remarque : lorsqu'une diclé est absente, elle est représentée sous forme de tiret «-»

Chapitre 8

Pré-traitements et modélisations statistiques

Des pré-traitements divers et variés

La phase de pré-traitement débute par la segmentation du signal audio. Il s'agit d'une segmentation «semi-automatique». Comme nous connaissons la suite de phonèmes contenue dans la phrase et le signal audio, nous appliquons tout d'abord un système de reconnaissance forcée (basé sur HTK [12]). Nous obtenons ainsi en sortie une première segmentation grossière, qu'il s'agira dans un second temps d'affiner à la main (en ajustant les frontières des consonnes et des voyelles).

Une fois cette phase de segmentation accomplie, nous devons *nettoyer* les données afin de connaître à tout moment 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.

Pour quelles raisons allons-nous construire des modèles statistiques ?

Lors de l'enregistrement, malgré le nombre élevé de caméras, il se peut que la main et/ou le visage se masquent partiellement, ce qui gêne l'obtention des coordonnées de certains marqueurs réfléchissants et aboutit à des données manquantes. Cependant, les positions des marqueurs les uns par rapport aux autres sont intrinsèquement corrélées, du fait de la géométrie de la main et du visage et de leurs mouvements lors de l'exécution du code LPC. On peut ainsi créer des modèles statistiques des deux objets et inférer la position de marqueurs absents connaissant la position des marqueurs voisins.

Un second intérêt lié à la création des modèles statistiques du visage et de la main concerne l'application finale dans le cadre du projet ARTUS (même s'il ne s'agit pas de la seule application possible) : remplacer le télétexte par un clone capable de synthétiser la Langue française Parlée Complétée à partir de n'importe quel texte. Le canal de transmission ayant une bande limitée, les modèles statistiques permettent de réduire la quantité d'information à envoyer via ce canal.

Les modèles statistiques en quoi cela consiste-t-il ?

Pour la modélisation, nous allons considérer le visage et la main comme deux objets bien distincts et créer ainsi deux modèles indépendants car nous n'avons aucun a priori en ce qui concerne le

phasage des mouvements des deux objets l'un par rapport à l'autre :

- un modèle articulatoire de la main : il va contenir les corrélations entre les coordonnées des différents marqueurs placés sur la main lors de la production du code LPC pour la codeuse étudiée et pour ce corpus donné ;
- un modèle articulatoire du visage : il va capturer les corrélations entre les coordonnées des différents marqueurs placés sur le visage lors de l'activité de parole.

Il faut remarquer que ces deux modèles articulatoires sont à portée limitée : ils ne correspondent qu'aux mouvements de production de la parole (le visage) et de production du code LPC (la main). On ne peut pas *a priori* animer un visage et une main avec ces modèles pour des tâches plus génériques (gestualité prosodique, grimaces, ...) que celles dont elles sont issues. Ceci souligne ainsi, le choix et la construction du corpus. Pour créer ces deux modèles, nous utiliserons les données issues des deux corpora indépendants (**visage seul** et **main seule**), enregistrés, rappelons-le, sur la même locutrice que pour le corpus conjoint **main + visage** et pour lesquels les marqueurs avaient été (re-)placés sur les mêmes points de chair.

8.1 Méthodologie

Nous allons décrire les méthodes statistiques et les algorithmes utilisés dans la création de modèle pour le visage et pour la main.

8.1.1 Le visage

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 appliquées aux mouvements de différents sous-ensembles de points de peau [11, 6, 1, 2]. 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, celui des coins de lèvres et au mouvement de la gorge.

Cette méthodologie est normalement appliquée à des têtes quasi-statiques. Or, dans les corpora **visage seul** et **main + visage**, le mouvement de la tête est libre. Nous avons donc dû résoudre le problème de la répartition de la variance des positions des 18 marqueurs placés sur la gorge entre les mouvements de tête et les mouvements faciaux. Ce problème a été résolu en 4 étapes :

1. Estimation du centre de rotation de la tête le mieux adapté : nous avons minimisé la dispersion des coordonnées des points du visage au niveau des cibles en retranchant la translation moyenne de chaque phrase.
2. Estimation d'un mouvement de tête utilisant l'hypothèse d'un mouvement rigide des marqueurs placés sur les oreilles, le nez et le front. Une analyse en composantes principales sur les 6 paramètres de roto-translation extraits du corpus **main + visage** est calculée et les **nmF** premières composantes sont retenues comme paramètres de contrôle de la tête.
3. Le clonage des mouvements articulatoires du visage est effectué en inversant le mouvement rigide sur toutes les données. Nous retenons **naF** composantes comme paramètres

de contrôle des mouvements articulatoires du visage.

4. Les mouvements de la gorge sont considérés comme égaux aux mouvements de tête à un facteur près inférieur à 1. Une optimisation des poids et des déformations du visage est ensuite calculée en gardant la même valeur pour les prédicteurs **nmF** et **naF**.

L'algorithme final (décrit sous forme de programme Matlab) nous permettant de calculer les positions 3D *P3DF* des 63 marqueurs placés sur le visage est le suivant :

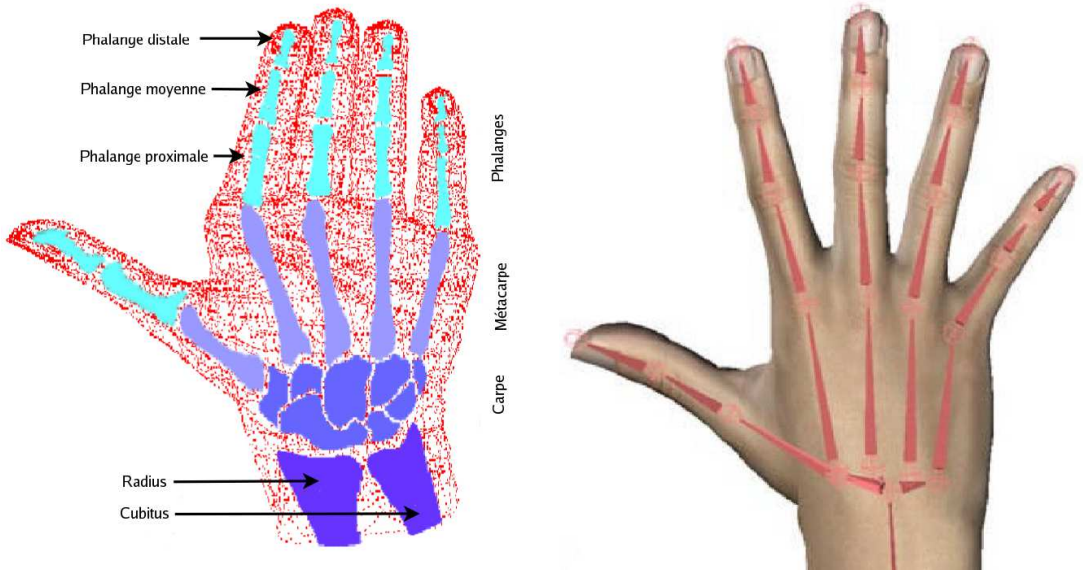
```
mvt = mean_mF + pmF * eigv_mF ;
P3D = reshape((mean_F+paF*eigv_F),3,63) ;
for i := 1 to 63
    M = mvt .* wmF( :,i) ;
    P3DF( :,i) = Rigid_Motion(P3D( :,i),M) ;
end
```

où **mvt** sont les mouvements de tête contrôlés par les **nmF** paramètres **pmF**, **M** est le mouvement pondéré de chaque marqueur (poids égal à 1 pour les marqueurs situés sur le visage et inférieur à 1 pour ceux situés sur le cou) et **P3D** sont les positions 3D des marqueurs sans mouvement de tête contrôlés par les **naF** paramètres **paF**.

8.1.2 La main

Les modèles de main de la littérature sont, en général, des modèles génériques utilisés pour capturer les mouvements de la main lors de séquences vidéo. On retrouve des modèles surfaciques [13, 5], des modèles volumiques [9] et des modèles squelettiques [4]. Il existe également des modèles de main pour l'animation d'avatars 3D. Il peut s'agir de modèles volumiques [8] composés de cylindres, de sphères et de parallélépipèdes contrôlés par des méthodes de cinématique inverse. Il existe également dans la norme MPEG-4 un modèle de main contrôlé par des BAPs (Body Animation Parameters) : celui-ci peut-être animé par des méthodes d'interpolation entre des formes clés [10]. Comme pour la capture de mouvements, des modèles de main musculo-squelettiques ont été implémentés [7] et l'animation de ceux-ci peut se faire par activation de segments de mouvements prédéfinis. La méthode la plus répandue en animation par ordinateur pour piloter un tel objet 3D reste de créer un squelette virtuel sur lequel on connecte une enveloppe censée recréer les tissus de peau. Cette méthode appelée «skinning» permet de faire bouger les points de la surface de la main en fonction de la forme du squelette sous-jacent (voir figure 8.1 (b)). Chaque point du maillage de l'enveloppe bouge avec la même matrice de transformation que l'os auquel il est rattaché. Pour les points situés sur les zones d'articulation et qui sont donc rattachés à 2 os on passe par des méthodes de pondération pour pouvoir assurer une zone de transition flexible.

Si l'on pouvait garder la même méthodologie de modélisation pour les objets visage et main on s'assurerait d'une certaine cohérence. On pourrait donc être tenté de ré-appliquer aux données de la main les paradigmes qui ont fait leurs preuves sur le visage. Mais, comparés aux mouvements du visage, les mouvements de la main présentent de grandes amplitudes. Ces mouve-



(a) Représentation anatomique

(b) Squelette sous-jacent de la méthode «skinning»

FIG. 8.1 – Représentations de la main sous forme anatomique et dans la méthode de «skinning».

ments ne peuvent donc pas être approchés avec un modèle linéaire additif. Le squelette de la main (voir figure 8.1 (a)) se compose des os du poignet (radius et cubitus), du carpe, de la métacarpe et des phalanges. La main est un organe possédant 28 degrés de liberté [9]. Il y a 6 degrés de liberté au niveau du poignet (considéré séparé du bras), deux degrés de liberté (abduction/adduction, extension/flexion) pour les articulations métacarpo-phalangiennes et un degré de liberté (flexion/extension) pour chaque articulation phalangienne. Le pouce quant à lui possède 3 degrés de liberté (abduction/adduction, extension/flexion et pseudo-roation due à l'incongruité entre les os du carpe et la base du métacarpe). Il s'agit donc d'un objet 3D difficile à modéliser et à animer. Cependant, des contraintes biomécaniques liées à la structure et à l'anatomie de cet organe réduisent les possibilités : certains mouvements sont en effet impossibles à réaliser. Cet argument permet ainsi d'espérer une réduction de l'espace des paramètres d'animation de cet objet. Nous n'allons pas proposer un modèle anthropométriques de la main avec une structure sous-jacente pour deux raisons :

- nous possédons des données *MoCap* et il serait dommage de passer par un modèle structurel qui de par sa modélisation appauvrirait les données et ferait perdre la caractéristique des mouvements biologiques ;
- nous devons dans le cadre de l'application souhaitée fournir un minimum de paramètres de contrôle du à la faible bande passante du canal.

Nous proposons par conséquent d'utiliser des modèles statistiques non-linéaires [3] pour pouvoir modéliser au mieux ces mouvements.

Nous n'avons accès qu'à la position de marqueurs placés sur la peau et non pas aux mouvements des structures rigides sous-jacentes (os) que l'on pourrait approcher par des rotations. Ainsi, la déformation de la peau engendrée par les tissus musculaires, les tendons, etc. produit

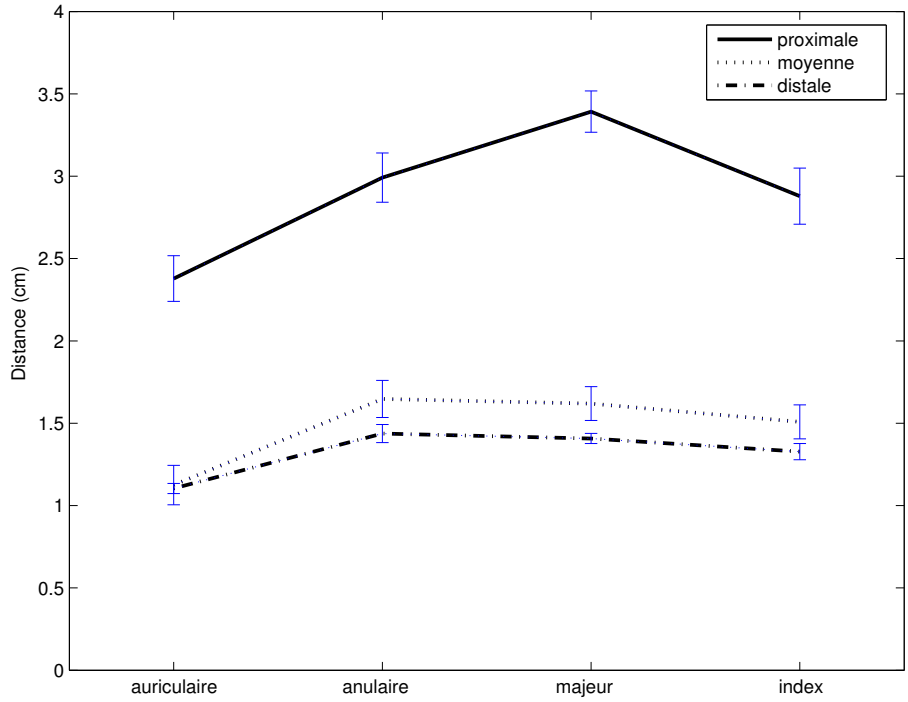


FIG. 8.2 – Distance (moyenne et écart-types) entre les marqueurs placés sur une même phalange.

d'importantes variations de distances entre les marqueurs collés sur une même phalange (ex. : variation de 3mm sur une distance de 1.6cm, pour les points situés sur la deuxième phalange du majeur). La figure 8.2 permet de voir que les points des marqueurs placés sur une même phalange ne subissent pas un mouvement simple (naïvement, la même rotation que celle de l'os de la phalange) puisque leur espacement varie selon l'articulation en cours : l'écart-type des 2 premières phalanges est particulièrement important.

Nous faisons l'hypothèse que des mouvements elliptiques existent et peuvent avoir une influence à distance sur les phalanges. Les paramètres de ce modèle génératif ne sont pas imposés mais calculés à partir des données mesurées. On peut donc évaluer la pertinence non pas biomécanique mais fonctionnelle de ce modèle simple de la même façon que pour le visage.

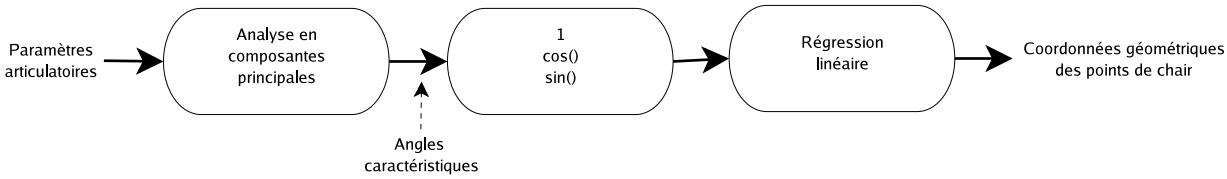


FIG. 8.3 – Diagramme du modèle non-linéaire de contrôle de la géométrie de la main.

La construction du modèle de déformation de la main peut se résumer en quatre étapes :

1. Estimation des mouvements de la main en utilisant l'hypothèse d'un mouvement rigide des marqueurs placés sur le dos de la main. Une analyse en composantes principales est ensuite calculée sur les 6 paramètres de mouvement de la main et nous conservons les

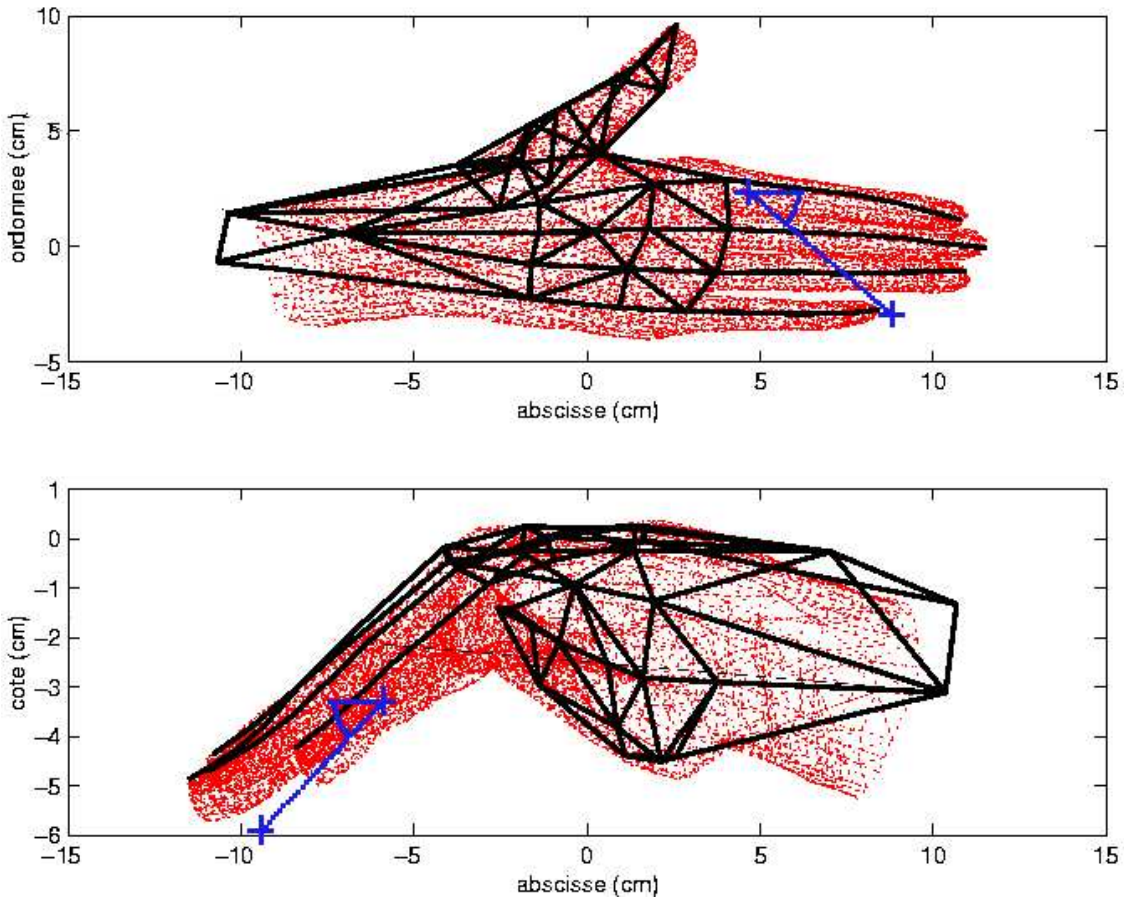


FIG. 8.4 – Détermination des valeurs d'angles lors des mouvements d'abduction/adduction (en haut) et lors des mouvements d'extension/flexion (en bas).

nmH premières composantes comme paramètres de contrôle des mouvements de la main.

2. Tous les angles entre les différents segments composant la main et le dos de la main ainsi qu'entre les phalanges successives sont calculés (abduction/adduction, extension/flexion). Dans le plan *abscisse-ordonnée* (voir figure 8.4 haut), sont calculés les angles correspondant à l'abduction/adduction pour la phalange proximale de l'index, du majeur, de l'anulaire et de l'auriculaire soit 4 angles. Dans le même plan est calculé l'écartement du poignet par rapport au dos de la main soit un angle supplémentaire. Dans le plan *abscisse-cote* (voir figure 8.4 bas), sont calculés les angles d'extension/flexion des phalanges proximales, moyennes et distales de l'index, du majeur, de l'anulaire et de l'auriculaire par rapport au dos de la main soit 12 angles supplémentaires. Dans le même plan, est calculé l'angle de rotation du poignet par rapport au dos de la main soit 1 angle supplémentaire. Enfin, vu la mobilité plus importante du pouce, sont calculés pour toutes les phalanges et dans les deux plans les angles correspondant à l'abduction/adduction, à l'extension/flexion et à la pseudo-rotation soit 6 angles supplémentaires. L'ensemble des angles calculés se portent à 24.

3. Une analyse en composantes principales est ensuite calculée sur tous ces angles et les **naH**

premières composantes sont retenues comme paramètres de contrôle de la forme de la main.

4. Nous calculons ensuite les sinus et cosinus de toutes ces valeurs prédites et nous faisons une régression linéaire entre les 2***naH+1** valeurs et les coordonnées 3D des marqueurs collés sur la main.

Remarque : l'étape 4 fait l'hypothèse que le déplacement induit par une rotation pure au niveau d'une articulation produit un mouvement elliptique de la surface de la peau.

L'algorithme final qui permet de calculer les positions 3D **P3DH** des 50 marqueurs de la main est le suivant (sa représentation sous forme de diagramme se trouve sur la figure 8.3) :

```
mvt = mean_mH + pmH * eigv_mH ;
ang = mean_A + paH * eigv_A ;
P = [1 cos(ang) sin(ang)] ;
P3DH = Rigid_Motion(reshape(P*Xang,3,50),mvt) ;
```

où **mvt** est le mouvement du dos de la main contrôlé par les **nmH** paramètres **pmH** et **ang** est l'ensemble des angles de la main contrôlés par les **naH** paramètres **paH**.

8.2 Implémentation

Après avoir présenté les deux méthodologies que nous avons utilisées pour modéliser nos objets 3D, nous allons voir plus en détails l'implémentation dans le cadre spécifique de nos données d'abord sur le visage puis sur la main.

8.2.1 Le visage

Toutes les opérations nécessaires au calcul du modèle sont réalisées sur les mouvements du visage des corpora **visage seul** et **main + visage** où tous les marqueurs sont visibles. Une quantification vectorielle nous assurant un minimum de distance 3D entre les trames sélectionnées (égal ici à 2 mm), est mis en oeuvre avant la modélisation. Nous retenons 4938 trames comme base d'apprentissage de notre modèle.

À partir des mouvements de ces 63 points et plus particulièrement ceux des lèvres et de la mâchoire (leurs mouvements étant supposés prépondérants), on calcule le modèle linéaire composé des 7 degrés de liberté de la parole visuelle qui nous intéressent :

1. montée/descente de la mâchoire (paramètre **Jaw1**) ;
2. étirement/protrusion des lèvres (paramètre **Lips1**) ;
3. montée/descente de la lèvre inférieure (paramètre **Lips2**) ;
4. montée/descente de la lèvre supérieure (paramètre **Lips3**) ;
5. montée/descente des commissures (paramètre **Lips 4**) ;
6. avancée/rétraction de la mâchoire (paramètre **Jaw2**) ;
7. montée/descente du larynx (paramètre **Lar1**).

Les 6 premiers degrés de liberté sont représentés sous forme de nomogrammes sur la figure 8.5. Nous obtenons pour chaque paramètre la variance du mouvement total expliqué par celui-ci, comme référencée dans le tableau 8.1. Puis cette variance est réduite à l'aide du modèle comme illustré sur la figure 8.6.

Nom du paramètre	Variance expliquée	Variance cumulée
jaw1	0.462	0.462
lips1	0.187	0.649
lips2	0.038	0.687
lips3	0.032	0.719
lips4	0.016	0.735
jaw2	0.046	0.781
lar1	0.013	0.794
mvtV1	0.480	0.480
mvtV2	0.340	0.820
mvtV3	0.079	0.899
mvtV4	0.064	0.963
mvtV5	0.029	0.992
mvtV6	0.008	1

TAB. 8.1 – Variance expliquée et cumulée des paramètres articulatoires et de roto-translation pilotant le modèle de visage.

8.2.2 La main

Toutes ces opérations sont faites sur les mouvements de la main des corpora **main seule** et **main + visage** 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). Nous conservons 8446 trames comme base d'apprentissage.

À partir de ces points, on calcule 24 angles : 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.

L'analyse en composantes principales (ACP) sur les angles nous permet de ne conserver que $nmH = 9$ paramètres expliquant 99% (seuil que nous nous sommes fixé) du mouvement articulatoire de la main. Les variances du mouvement expliqué par chacun de ces paramètres sont référencées dans le tableau 8.2. Cette variance est réduite à l'aide du modèle comme illustré sur la figure 8.8. Les 6 premiers degés de liberté sont représentés sous forme de nomogrammes sur la figure 8.7.

Nom du paramètre	Variance expliquée	Variance cumulée
ang01	0.648	0.648
ang02	0.172	0.820
ang03	0.093	0.913
ang04	0.032	0.945
ang05	0.018	0.963
ang06	0.013	0.976
ang07	0.007	0.983
ang08	0.005	0.988
ang09	0.003	0.991
mvtM1	0.464	0.464
mvtM2	0.333	0.797
mvtM3	0.143	0.940
mvtM4	0.052	0.992
mvtM5	0.007	0.999
mvtM6	0.001	1

TAB. 8.2 – Variance expliquée et cumulée des paramètres articulatoires et de roto-translation pilotant le modèle de main.

8.3 Résultats de la modélisation

Le seuil d'explication du mouvement pour la main que nous nous sommes fixés est de 99%. Nous supposons qu'il s'agit d'un seuil *convenable*. Comme les trajectoires enregistrées comportent des erreurs, il est inutile de vouloir expliquer 100% du mouvement. En ce qui concerne le visage, la méthodologie utilisée impose le seuil d'explication puisque il s'agit d'une analyse guidée. Selon ces critères, nous avons retenu $naH = 9$ paramètres de contrôle pour la forme de la main et $naF = 7$ paramètres de contrôle articulatoire du visage. Quant aux mouvements de roto-translation, nous considérons qu'il s'agit de mouvements plus grossiers et donc moins sensibles aux erreurs et aux bruits. Nous choisissons un seuil dans le cas des deux objets de 100% du mouvement expliqué. En effet, la conséquence directe de l'ablation du dernier paramètre de roto-translation est la reconstruction erronée de certaines phrases à cause d'une position de départ de la codeuse très eloignée de la position de codage. Ainsi, nous conservons $nmF = 6$ et $nmH = 6$ paramètres de contrôle du mouvement de la tête et de la main. L'erreur absolue moyenne de modélisation pour la position d'un marqueur visible est de 1.5 mm pour la main et 1 mm pour le visage. Cette erreur représentée sur la figure 8.9 est calculée par rapport aux données terrain. Cet ordre de grandeur de l'erreur commise par la modélisation est à conserver en vue de la partie évaluation du synthétiseur.

8.4 Résumé

La modélisation statistique de nos objets 3D a deux objectifs : *nettoyer* les données de capture de mouvement dans lesquelles apparaissent des *trous*, des inversions de points, etc. (voir figure 8.10) et réduire l'information à transmettre.

Nous avons donc créé deux modèles statistiques que l'on pilote à l'aide de 7 paramètres articulatoires en ce qui concerne le modèle du visage et à l'aide de 9 paramètres articulatoires pour celui de la main. À ces paramètres, il faut rajouter les 6 paramètres de roto-translation pour chacun des objets.

A ce stade, nous devons inverser pour chaque phrase les coordonnées 3D des deux objets pour chaque trame. Nous avons donc un ensemble de paramètres articulatoires et de roto-translation pour chaque trame de chaque phrase qui nous permet via les deux modèles de reconstruire les coordonnées 3D des points de ces deux objets.

L'information nécessaire pour coder la géométrie des deux objets a ainsi été réduite d'un ordre de grandeur. En effet, pour chaque trame on passe de 63*3 coordonnées pour le visage et 50*3 coordonnées pour la main soit 339 valeurs flottantes à 7+6 et 9+6 paramètres, soit 28 valeurs flottantes.

Références bibliographiques

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

[2] G. Bailly, M. Bérar, F. Elisei, and M. Odisio. Audiovisual speech synthesis. *International Journal of Speech Technology*, 6 :331–346, 2003.

[3] R. Bowden. *Learning non-linear Models of Shape and Motion.* PhD thesis, Dept Systems Engineering, Brunel University, Uxbridge, Middlesex, UK, 2000.

[4] M. Bray, E. Koller-Meier, P Müller, L. Van Gool, and N. N. Schraudolph. 3D hand tracking by rapid stochastic gradient descent using a skinning model. In *First European Conference on Visual Media Production*, pages 59–68, 2004.

[5] R. Cipolla, B. Stenger, A. Thayananthan, and P. H. S. Torr. Hand tracking using a quadric surface model and bayesian filtering. In *The British Machine Vision Conference*, 2001.

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

[7] P. Kalra, N. Magnenat-Thalmann, L. Moccozet, G. Sannier, A. Aubel, and D. Thalmann. Real-time animation of realistic virtual humans. *IEEE Computer Graphics and Applications*, 18(5) :42–55, 1998.

[8] R. Mas Sanso and D. Thalmann. A hand control and automatic grasping system for synthetic actors. In *Proc. Eurographics '94*, 1994.

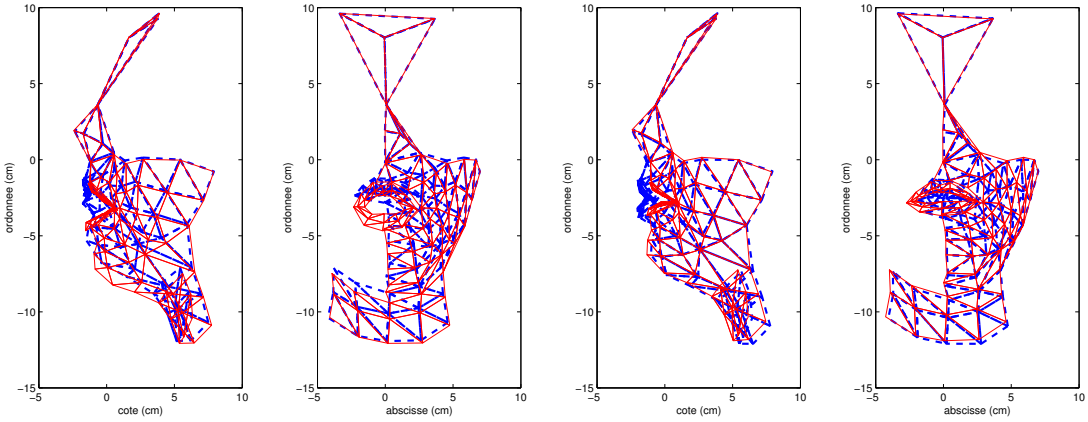
[9] H. Ouhaddi and P. Horain. Conception et ajustement d'un modèle 3D articulé de la main. In *Actes des 6èmes Journées du Groupe de Travail Réalité Virtuelle*, 1998.

[10] M Preda, T. Zaharia, and F. Preteux. 3D body animation and coding within a MPEG-4 compliant framework. In *Proceedings International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging*, 1999.

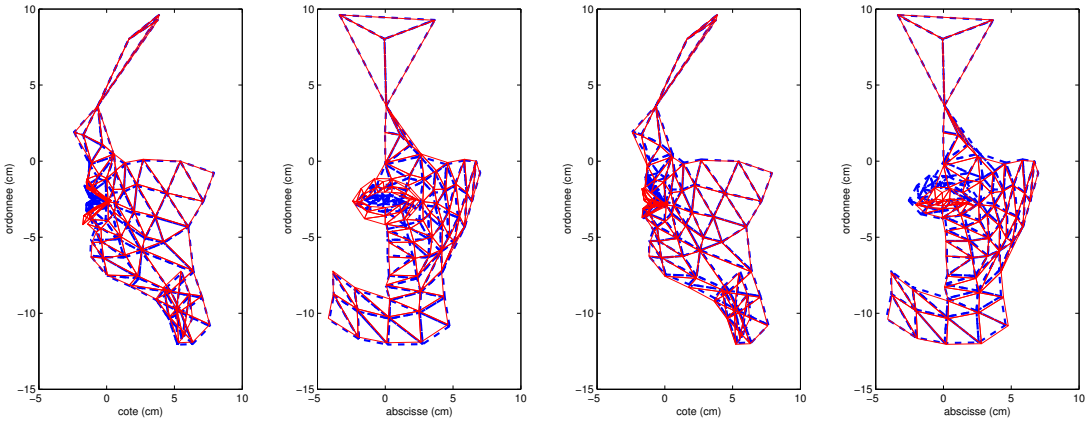
[11] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[12] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. Large vocabulary continuous speech recognition using HTK. In *Proceedings ICASSP'94*, 1994.

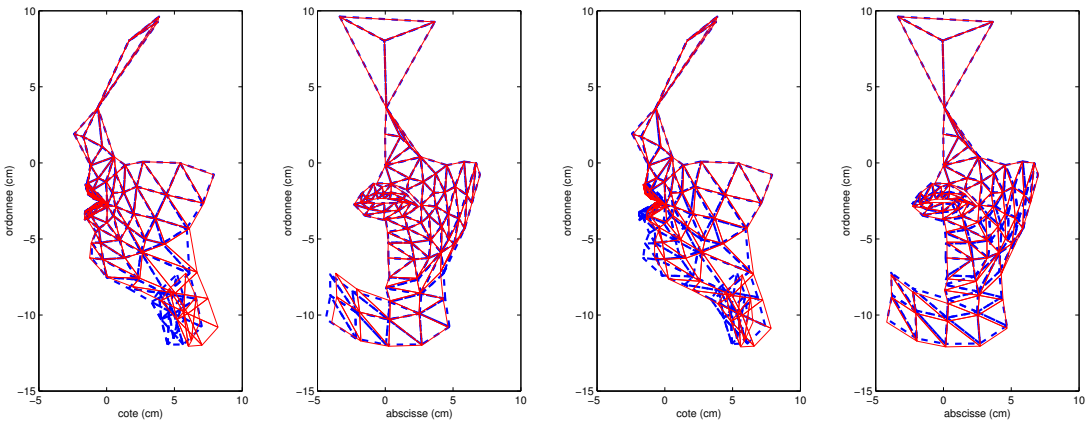
[13] Y. Wu, J. L. Lin, and T. S. Huang. Capturing natural hand articulation. In *Proceedings of IEEE International Conference on Computer Vision*, Canada, 2001.



(a) Paramètre Jaw1 (b) Paramètre Lips1



(c) Paramètre Lips2 (d) Paramètre Lips3



(e) Paramètre Lar1 (f) Paramètre Jaw2

FIG. 8.5 – Nomogrammes représentant les 6 premiers degrés de liberté recherchés dans le visage.

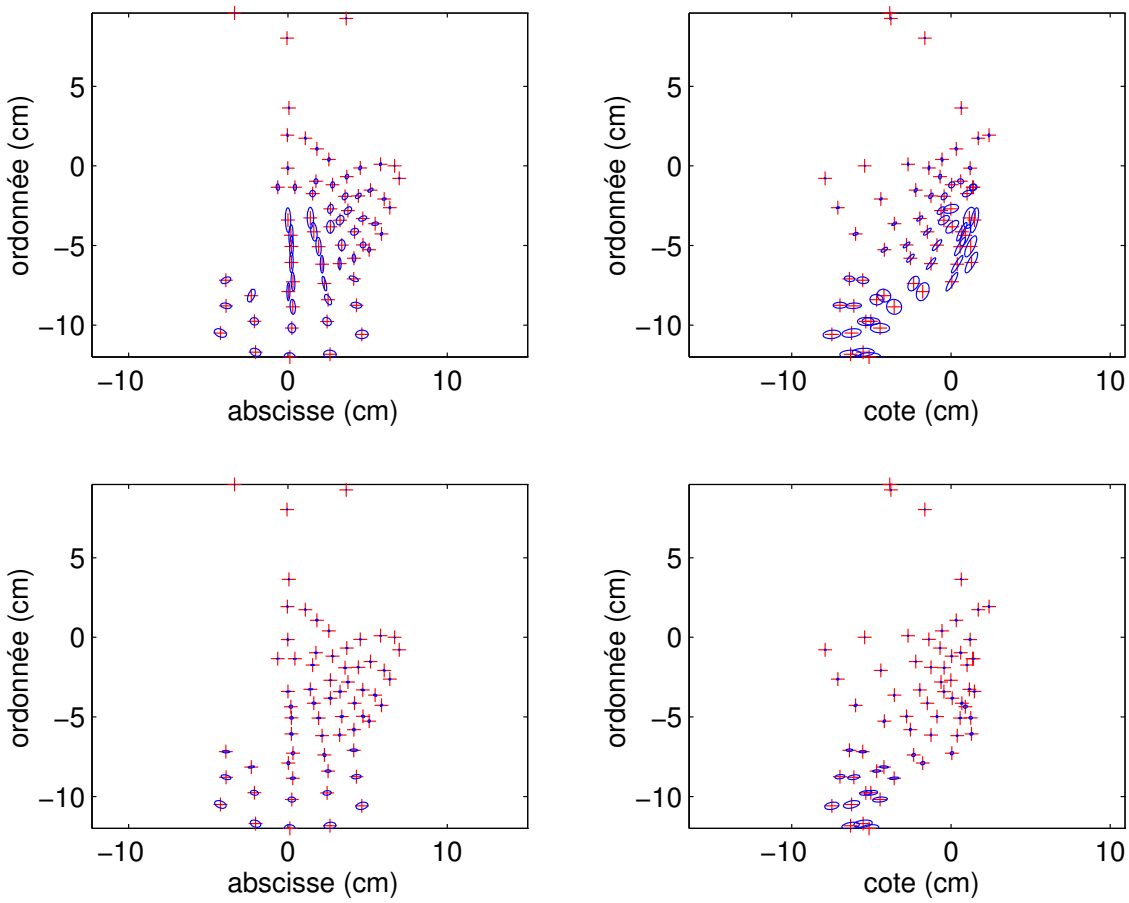
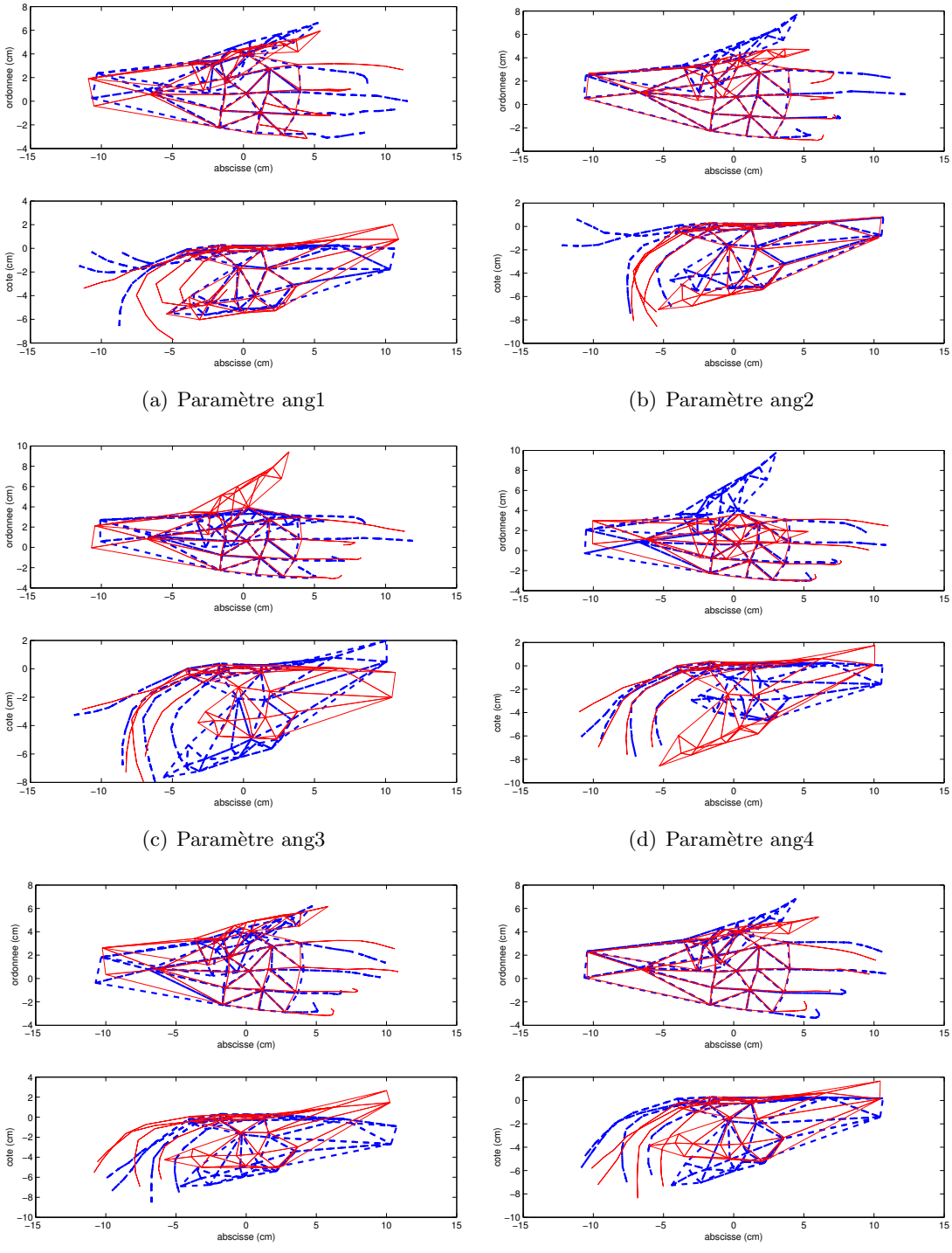


FIG. 8.6 – Ellipses de dispersion (à 1 σ) des données du visage par rapport à la configuration moyenne, en haut données originales, en bas données (résiduelles) après passage par le modèle articulatoire (de gauche à droite : vue de face, vue de profil).



(a) Paramètre ang1

(b) Paramètre ang2

(c) Paramètre ang3

(d) Paramètre ang4

(e) Paramètre ang5

(f) Paramètre ang6

FIG. 8.7 – Nomogrammes représentant les 6 premiers degrés de liberté recherchés dans la main.

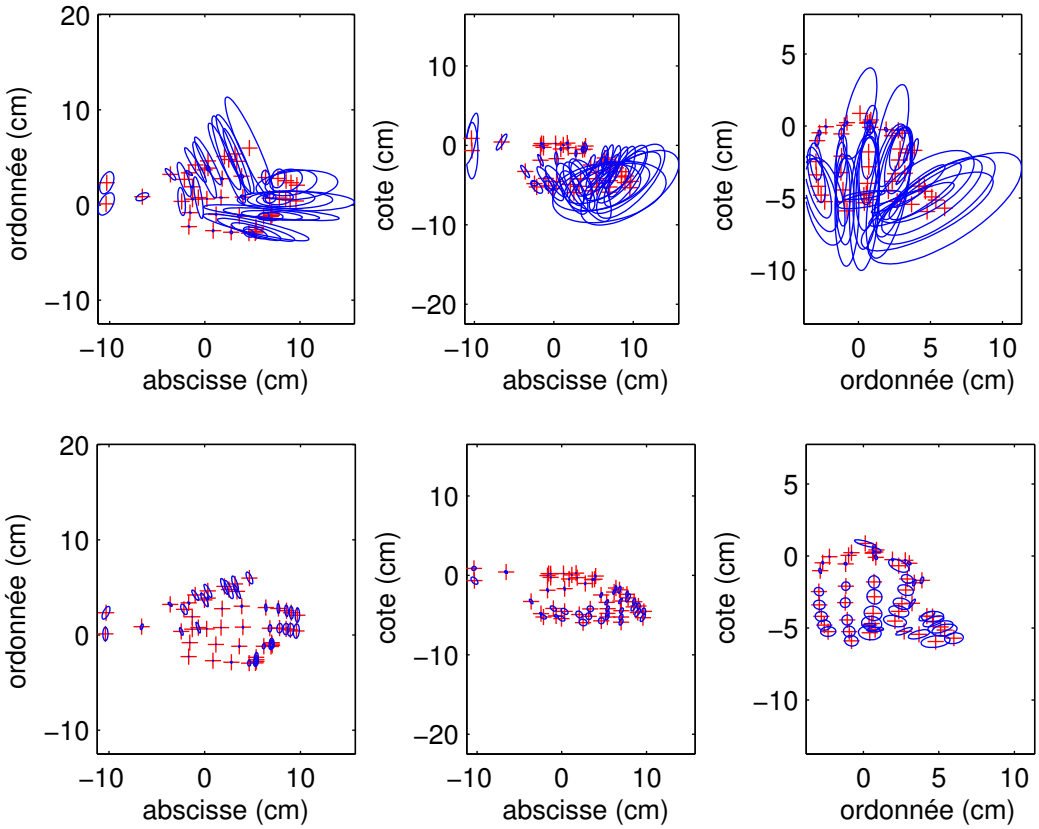


FIG. 8.8 – Ellipses de dispersion (à 1 σ) des données de la main par rapport à la configuration moyenne. En haut, les données originales, en bas les données (résiduelles) après passage par le modèle articulatoire (de gauche à droite : vue de dessus, vue de côté, vue de face).

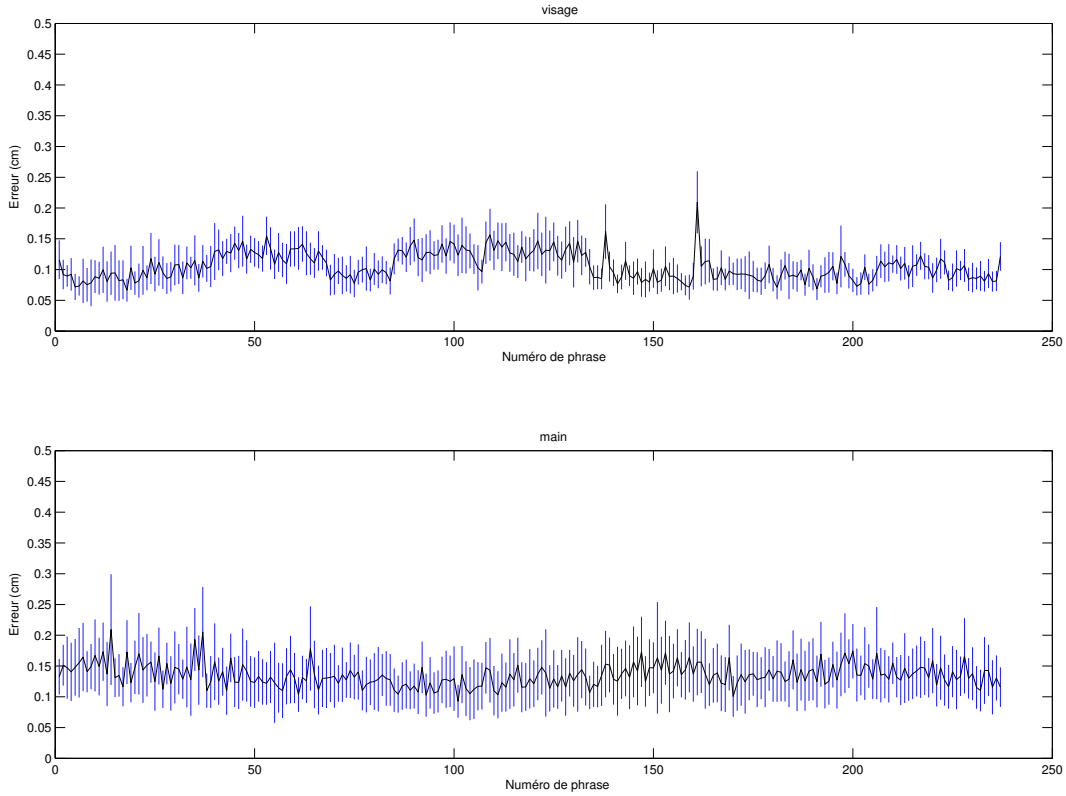
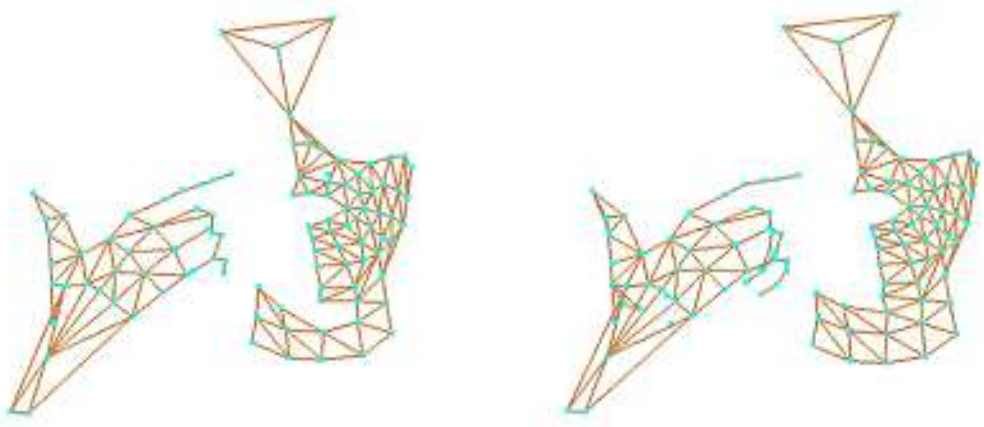


FIG. 8.9 – Erreurs (moyenne et écart-type) de modélisation pour chaque phrase du corpus pour la main et le visage.



(a) Données issues de la capture de mouvement (b) Reconstruction des données par les modèles

FIG. 8.10 – Reconstruction des données de capture de mouvements à partir des modèles statistiques de la main et du visage : les points du cou et de la main qui n'ont pas été capturés par le système (cf. a) de capture sont reconstruits par les modèles (cf. b).

Chapitre 9

Analyse de la production du code LPC

Pourquoi analyser quand on veut faire de la synthèse ?
Le corpus enregistré est maintenant pré-traité ; nous sommes donc en possession d'un ensemble de paramètres articulatoires à 120 Hz et cela pour 238 phrases. Nous possédons également le signal acoustique correspondant que nous avons segmenté : nous connaissons ainsi les chaînes phonétiques marquées en temps pour chaque phrase. Alors pourquoi ne pas passer directement à la synthèse ?

Car nous devons faire des analyses sur l'ensemble de ces 238 phrases avant de passer à la synthèse. Il faut, en effet, vérifier que la codeuse a réalisé les bonnes transitions de forme et de position de main en fonction de la phrase prononcée.

Ensuite, nous devons tenir compte du phasage des différents signaux lors de la production du code LPC par notre codeuse. On pourra se baser dans un premier temps sur les résultats de Attina et al. [2] qui montrent que la main est en avance par rapport au signal acoustique et aux mouvements de lèvres, mais il faudra vérifier ce schéma pour notre codeuse dans la tâche qui lui incombait.

Cette phase d'analyse apparaît donc comme indispensable à la phase de synthèse. Elle débute par la vérification des données, puis se poursuit par l'étude de la coordination entre les divers articulateurs et s'achève par l'étude de la prosodie spécifique de la Langue française Parlée Complétée.

9.1 Vérification des données : Code LPC / chaîne phonétique

La base de notre système de synthèse est le dictionnaire contenant les unités à concaténer. Nous avons déjà segmenté et vérifié le signal audio. Par contre, en ce qui concerne la partie *vidéo* (les trajectoires des marqueurs) nous ne savons rien : nous ne connaissons pas le déroulement et nous ne savons pas si des erreurs se sont immiscées (une erreur pouvant être une série CV non codée ou une erreur de codage). La première tâche consiste donc à vérifier le code produit et à en déduire une segmentation.

Le code LPC fonctionne comme un modèle de constriction : la main avec une certaine forme

et le visage produisent, la plupart du temps (à l'exception de la position côté), une occlusion, un contact. La place de cette constriction correspond à la voyelle de la série CV et la forme de la main correspond à la consonne. Nous vérifions que ces deux informations, forme et position de la main, correspondent à ce qui a été prononcé. Pour ce faire, nous mettons en oeuvre un système de reconnaissance capable de délivrer la probabilité à chaque instant d'être dans une configuration (forme et position de main) donnée. Nous allons voir plus en détails son implémentation.

9.1.1 Reconnaissance de la forme de la main

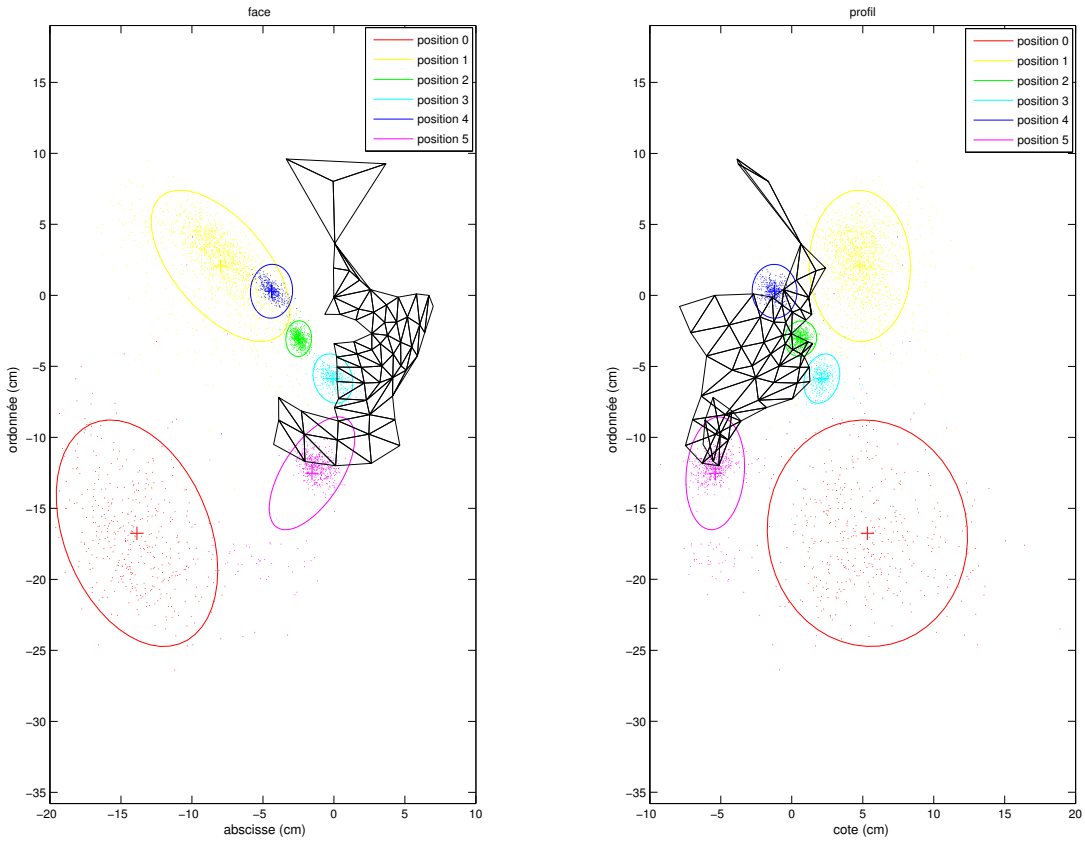


FIG. 9.1 – Données et ellipses de dispersion de la position du bout du doigt le plus long pour chaque réalisation atteinte de cible (vue de face à gauche et vue de profil à droite).

Dans un premier temps, nous avons utilisé les résultats de production des études de Attina et al. pour faire une segmentation automatique des trajectoires de la main. Nous avons supposé que la cible (c'est-à-dire le moment durant lequel la constriction a lieu) était atteinte au début acoustique de toute série CV. Nous nous sommes basés sur la chaîne phonétique marquée en temps (déduite de la segmentation audio) et l'information de phasage précédemment décrite pour initialiser la segmentation *vidéo*.

Nous avons choisi 7 paramètres caractéristiques :

– Pour chaque doigt (hormis le pouce), la distance entre le marqueur positionné sur la phalange proximale le plus proche du dos de la main et celui placé sur la phalange distale le plus proche du bout du doigt est calculée : une valeur maximale correspond à une extension du doigt alors qu'une valeur minimale correspond à une rétraction.
– La distance entre les marqueurs placés sur les bouts des doigts index et majeur est déterminée pour éviter toute confusion entre les formes 2 et 8.
– La distance entre le bout du pouce et le dos de la main est déterminée pour différencier les formes 1 et 6, 2 et 7.

Ces 7 paramètres associés aux formes de main correspondantes permettent d'estimer des modèles gaussiens pour chaque forme de main. La probabilité *a posteriori* de chaque nouvelle trame d'appartenir à une des 8 formes de main peut être calculée. Nous avons utilisé, pour calculer ces modèles, les trames correspondant au début de séries CV acoustiques.

Cette première segmentation nous a permis de déduire un certain nombre d'erreurs, dues aux modèles mais aussi dues à des erreurs de codage. Une analyse plus précise de ces erreurs nous a aussi montré qu'il fallait segmenter plus précisément les cibles LPC. Nous avons donc vérifié et segmenté manuellement les 238 phrases, en s'aidant de la première segmentation automatique, aux instants de constriction maximale en utilisant le système d'animation MOTHER de l'ICP [13] et étiqueté la valeur appropriée de la clé, c'est-à-dire un chiffre entre 0 et 8 : 0 correspondant à la position de repos choisie par la codeuse (poing fermé à l'écart du visage) et les chiffres allant de 1 à 8 correspondant à une des 8 formes du code LPC. Nous avons ainsi identifié et segmenté 3831 réalisations de formes de main.

Nous avons ensuite recalculé nos modèles gaussiens avec cette nouvelle segmentation. Le taux de reconnaissance est de 98.78%. Les erreurs (cf. tableau 9.1) sont en général dues à des problèmes de réductions consonantiques voire à des omissions (notamment des *glides* dans des séquences complexes CCCV).

Un exemple de ces probabilités au cours du temps sur la première phrase du corpus est représenté sur la figure 9.2 avec le signal acoustique associé.

seg. / reco.	0	1	2	3	4	5	6	7	8	Total
0	395	0	1	2	3	5	1	3	0	410
1	1	440	0	0	0	0	2	0	0	443
2	5	0	392	0	1	0	0	0	1	399
3	1	0	0	585	0	1	0	0	0	587
4	2	0	0	0	350	2	0	0	0	354
5	10	0	0	0	0	956	0	2	0	968
6	1	0	0	0	0	0	433	0	0	434
7	1	0	0	0	0	0	0	78	0	79
8	0	0	1	0	0	0	0	0	156	157

TAB. 9.1 – Matrice de confusion du système de reconnaissance des formes de main. Pour une configuration segmentée (colonne de gauche), on présente le nombre de représentants reconnus par configuration (ligne du haut).

9.1.2 Reconnaissance de la position de la main

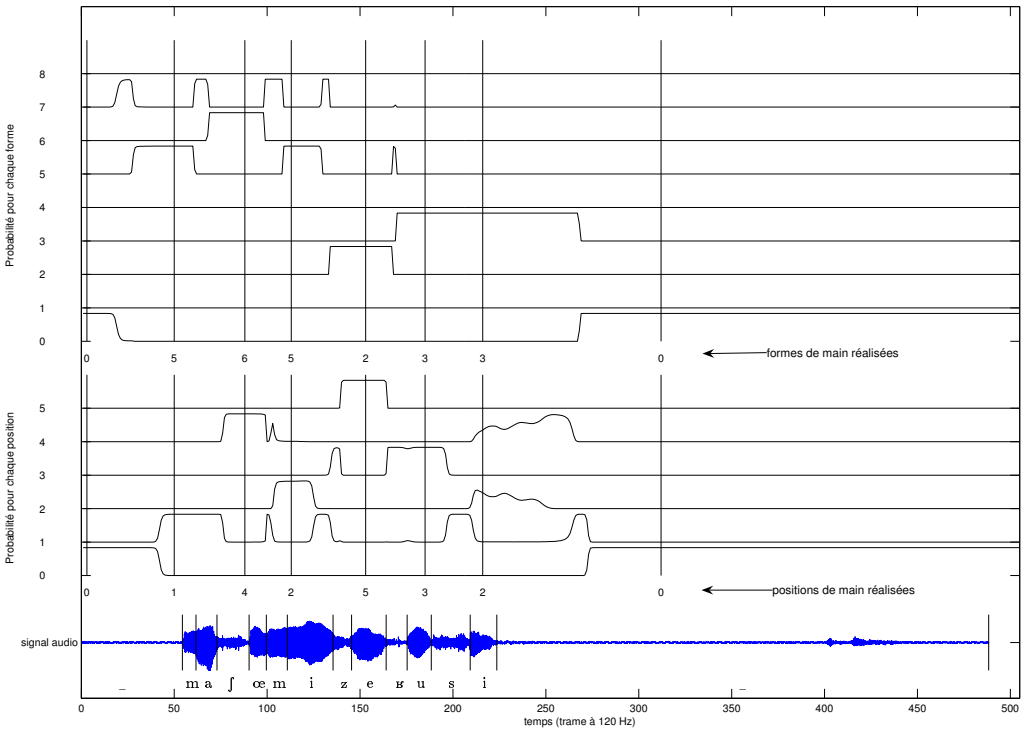


FIG. 9.2 – Variation des probabilités issues des modèles gaussiens pour la forme (haut) et la position (bas) de la main pour la première phrase du corpus «ma chemise est roussie».

Nous avons ajouté à l'étiquetage précédent 6 valeurs pour la position de la main : la position 0 correspondant à la position de repos et les chiffres de 1 à 5 correspondant à l'une des 5 positions de la main du code LPC. Nous avons caractérisé la position de la main pour chaque cible dans un référentiel 3D rattaché à la tête : la position 3D du doigt le plus long (l'index pour les configurations 1 et 5 et le majeur dans les autres cas) (cf. figure 9.1) a été enregistrée et des modèles gaussiens ont été estimés comme précédemment. Sur les 3831 trames, 98.56% des réalisations des positions ont été identifiées pour un total de 133 erreurs de reconnaissance (cf. tableau 9.2).

Il y a 3 sources d'erreurs possibles :

- la plus importante source d'erreur vient de la position 1 (coté) : cette position est aussi utilisée pour coder des consonnes précédées d'une consonne et pour des schwas : la codeuse pointe la position coté mais ne l'atteint pas.
- La position de repos 0 a une grande variance et les positions 1 et 4 réalisées trop loin du visage sont parfois capturées par le modèle gaussien de la position 0.
- des confusions de codage des voyelles intermédiaires (/e/ *vs* /ɛ/ par exemple).

Ces deux systèmes de reconnaissance nous ont permis dans un premier temps de vérifier notre segmentation *vidéo* puis de connaître les erreurs de la codeuse afin d'en tenir compte dans

seg. / reco.	0	1	2	3	4	5	Total
0	407	2	0	0	0	1	410
1	13	1602	2	6	6	1	1630
2	1	3	562	0	1	0	567
3	1	5	0	352	0	0	358
4	2	4	1	0	337	0	344
5	5	0	0	1	0	516	522

TAB. 9.2 – Matrice de confusion du système de reconnaissance des positions de main. Pour une configuration segmentée (colonne de gauche), on présente le nombre de représentants reconnus par configuration (ligne du haut).

la constitution de notre dictionnaire. Cependant, l'utilité de ces deux systèmes ne s'arrête pas là, nous allons voir dans la prochaine section qu'ils sont un moyen de caractériser les relations temporelles des mouvements des articulateurs.

9.2 Synchronisation entre les mouvements des articulateurs et l'acoustique

Ce système de reconnaissance des formes et positions de la main a un but essentiel : nous donner une probabilité d'être dans une configuration (forme et position) donnée à tout instant (c'est-à-dire à toute trame). Il nous fournit également une partition gestuelle de n'importe quelle phrase (voir figure 9.2) : on peut déterminer un début et une fin de mouvement pour chaque geste de main. En effet, nous définissons le début d'un geste par l'instant à partir duquel la probabilité d'être dans la configuration à coder est et reste supérieure aux autres jusqu'à l'atteinte de la cible. Nous définissons de cette manière le début de mouvement (respectivement la fin de mouvement). Il s'agit d'une définition qui permet de quantifier facilement le début et la fin d'un geste. Nous pouvons remarquer que cette définition se base sur des données «statiques» de position (coordonnées 3D de points de la main) pour déduire des relations dynamiques. Ce choix est avantageux à cause de sa simplicité d'implémentation.

		début du geste	atteinte de cible	fin du geste
Forme de la main	moyenne (ms)	5.01	81.57	153.06
	écart-type (ms)	89.49	81.37	82.46
Position de la main	moyenne (ms)	16.63	81.57	152.94
	écart-type (ms)	81.89	82.19	81.37

TAB. 9.3 – Temps moyens et écart-types des caractéristiques du geste par rapport à l'instant acoustique initial de la consonne C pour les séries CV (cf. figure 9.3).

L'étude de la synchronisation des mouvements de la main par rapport au signal audio s'inscrit dans le projet de synthèse. Pour des études poussées sur la production du code LPC, le lecteur

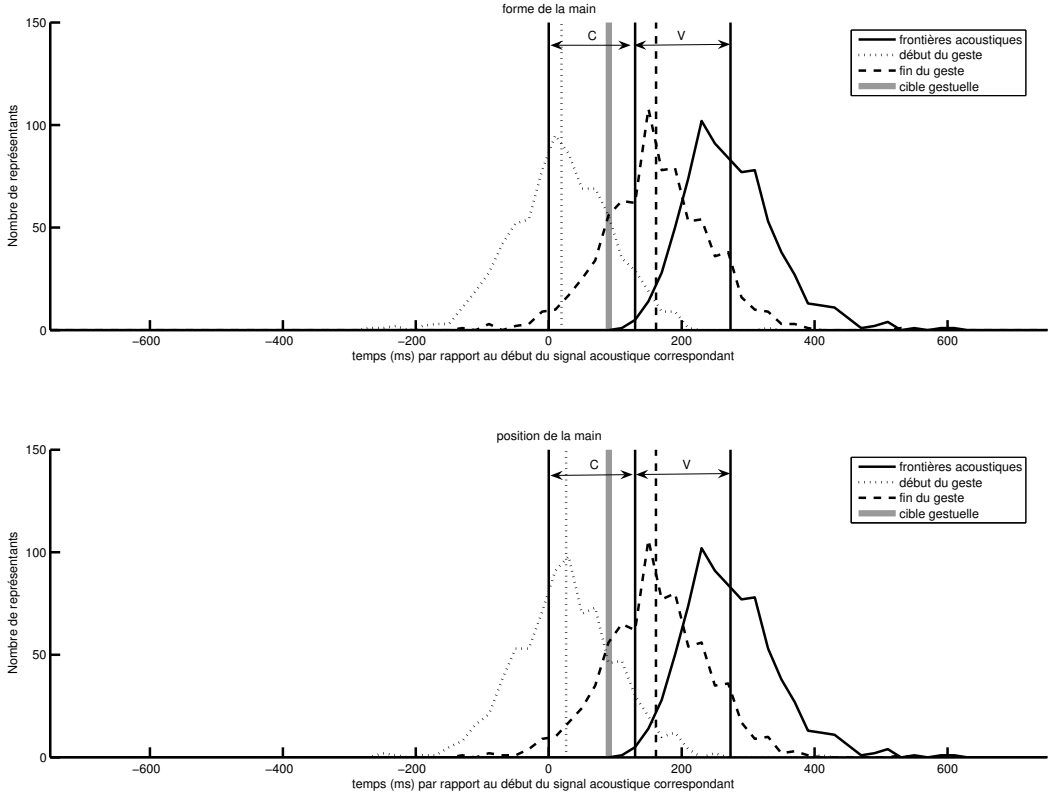


FIG. 9.3 – Phasage des mouvements de main par rapport au segment CV acoustique correspondant : distributions des différences de temps pour le début, la cible et la fin du mouvement de la main par rapport au début acoustique de la série CV.

pourra se référer à [1].

Nous obtenons les résultats suivants (description qualitative sur les figures 9.3, 9.4 et 9.5, description quantitative sur les tableaux 9.3, 9.4 et 9.5) :

- Dans le cas des segments CV où les résultats sont déduits de 840 trames du corpus sur lesquelles le code LPC et l'information acoustique correspondent, le début du mouvement se situe, en moyenne, au niveau du début acoustique correspondant ; la cible est alors atteinte dans la deuxième partie de la consonne et se termine dans la première partie de la voyelle.
- Dans le cas des segments C où les résultats sont déduits de 440 trames du corpus sur lesquelles le code LPC et l'information acoustique correspondent, le début du geste est, en moyenne, en avance par rapport au début acoustique correspondant ; la cible se situe dans la première partie de la consonne acoustique et la fin du geste se trouve dans la consonne.
- Dans le cas des segments V où les résultats sont déduits de 135 trames du corpus sur lesquelles le code LPC et l'information acoustique correspondent, le début du geste a lieu, en moyenne, avant le début acoustique, la cible est quasi-synchrone avec le début acoustique et la fin du geste a lieu dans la première partie de la voyelle.

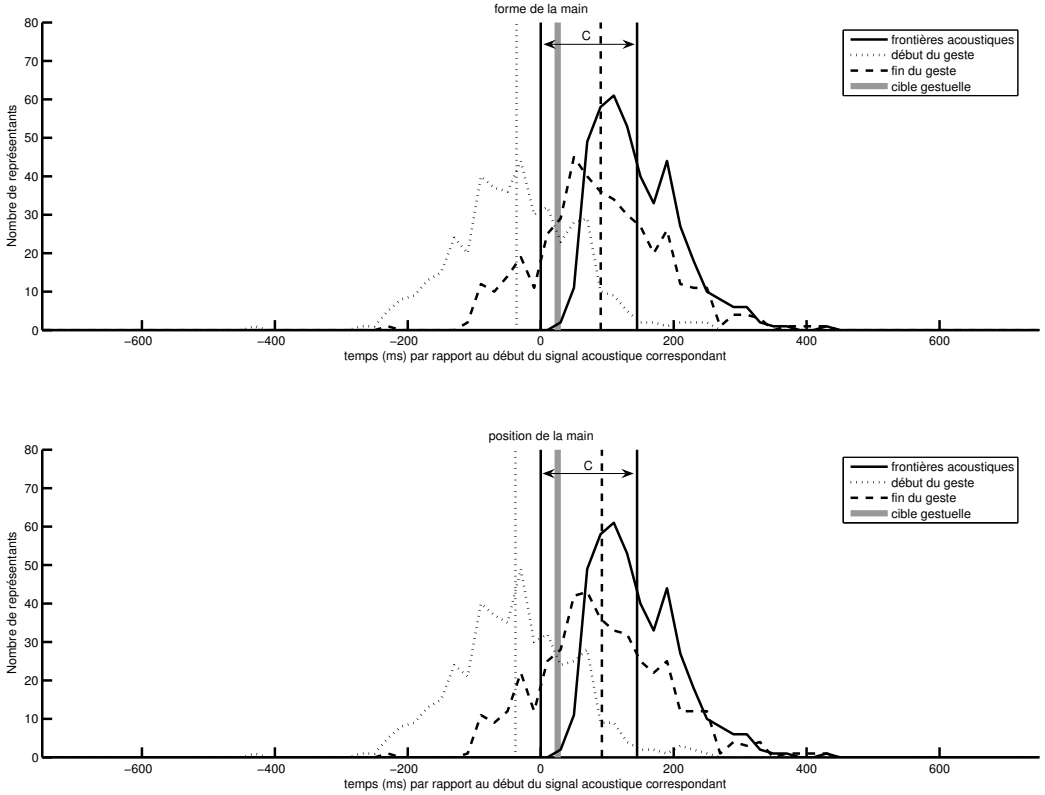


FIG. 9.4 – Phasage des mouvements de main par rapport au segment C isolée acoustique correspondant : distributions des différences de temps pour le début, la cible et la fin du mouvement de la main par rapport au début acoustique de la série C.

Ces conclusions sont valables pour le geste relatif à la forme de la main et le geste relatif à la position de la main par rapport au visage. Il s'agit là de conclusions qualitatives basées sur la moyenne des coordinations temporelles. Comme on peut le noter sur les figures représentant les distributions des différences de temps, les données sont dispersées.

Existe-il des corrélations qui permettraient d'expliquer cette dispersion ?
Si l'on considère tout d'abord les gestes de la main (forme et position), on note une forte corrélation linéaire ($\rho > 0.9$) entre les instants liés au début et à la fin du mouvement d'une part, au début du mouvement et à la cible d'autre part ainsi qu'à la cible et à la fin du mouvement. Ceci se vérifie à la fois pour la forme et pour la position de la main et dans les trois cas de segments : CV, C et V. En revanche, on remarque également une forte corrélation linéaire ($\rho > 0.9$) entre les instants liés aux débuts de geste relatif à la forme et celui relatif à la position, idem pour la fin des mouvements. La conclusion que l'on peut faire à partir de ces résultats est qu'il existe un modèle temporel de geste de la main bien défini et qui varie très peu. Par contre, ce geste n'est pas positionné tout le temps de la même façon par rapport à la réalisation acoustique du segment à coder.

Si l'on considère maintenant les durées des gestes par rapport aux durées acoustiques, on

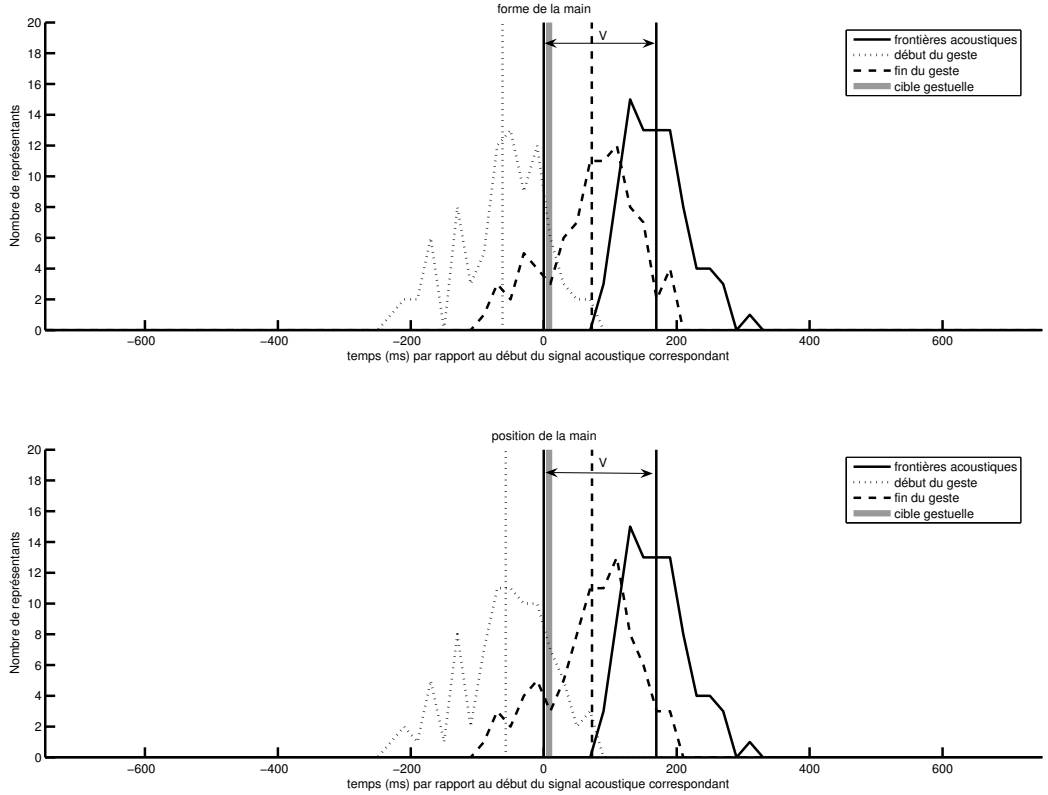


FIG. 9.5 – Phasage des mouvements de main par rapport au segment V isolée acoustique correspondant : distributions des différences de temps pour le début, la cible et la fin du mouvement de la main par rapport au début acoustique de la série V.

remarque qu'il n'y a pas de corrélation linéaire flagrante ($\rho \sim 0.3$) entre les durées des gestes de la main (forme ou position) et la durée du segment acoustique. Il n'y a pas non plus de corrélation linéaire probante ($\rho \sim 0.3$) entre le début du mouvement de la main (forme ou position) et la durée du segment acoustique, ni entre l'instant lié à la cible et la durée du segment acoustique ($\rho \sim 0.3$).

Cette étude des relations temporelles entre les différents articulateurs mis en jeu lors de la production du code LPC nous a permis de dégager des conclusions à propos de la production : la cible est, en moyenne, atteinte dans la première partie de la réalisation acoustique, le geste commençant avant ou au début de cette réalisation. Cependant, nous n'avons pu déduire aucune relation plus fine liée par exemple à la durée des segments acoustiques. Nous tiendrons compte de ces conclusions dans le chapitre suivant relatif à la synthèse. Il apparaît d'ores et déjà difficile d'imaginer un système de synthèse se composant d'un seul dictionnaire d'unités multimodales.

		début du geste	atteinte de cible	fin du geste
Forme de la main	moyenne (ms)	-43.76	18.83	84.15
	écart-type (ms)	108.50	109.44	106.03
Position de la main	moyenne (ms)	-44.53	18.83	85.97
	écart-type (ms)	103.73	109.44	105.61

TAB. 9.4 – Temps moyens et écart-types des caractéristiques du geste par rapport à l'instant acoustique initial de la consonne C pour les séries C isolée (cf. figure 9.4).

		début du geste	cible	fin du geste
Forme de la main	moyenne (ms)	-142.33	-46.41	20.56
	écart-type (ms)	130.05	101.92	100.02
Position de la main	moyenne (ms)	-119.06	-46.41	21.24
	écart-type (ms)	110.15	101.92	99.74

TAB. 9.5 – Temps moyens et écart-types des caractéristiques du geste par rapport à l'instant acoustique initial de la voyelle V pour les séries V isolée (cf. figure 9.5).

9.3 Prosodie de la Langue française Parlée Complétée

La prosodie correspond en général à l'étude de la variation de la fréquence fondamentale, de la durée des syllabes voire de l'énergie du signal audio en fonction de la structure de la phrase. C'est l'étude des ces différents paramètres qui va nous donner un aperçu du rythme spécifique du codage LPC. Il faudra rajouter à ces paramètres couramment étudiés, ceux liés aux mouvements de la tête et de la main. En effet, ceux-ci sont directement liés au discours ; nous allons dans un premier temps nous attacher à décrire notre système d'analyse de la prosodie acoustique, puis nous verrons comment nous l'avons appliqué à l'étude de la prosodie gestuelle.

9.3.1 Prosodie acoustique

L'outil d'analyse utilisé pour étudier les caractéristiques prosodiques de notre codeuse dans l'acte de production de la Langue française Parlée Complétée est le modèle SFC (Superposition of Functional Contours) [10]. Il permet d'effectuer un apprentissage automatique de la représentation du substrat prosodique en décomposant le contour prosodique en une superposition de contours prototypiques encodant diverses fonctions de communication. Il est appliqué à 2 paramètres : la fréquence fondamentale F0 et le coefficient d'allongement C. Une représentation de la décomposition effectuée par cet outil d'analyse peut être visualisée sur la figure 9.6 pour le paramètre F0 et sur la figure 9.7 pour le paramètre C. Une phrase longue de notre corpus a été sélectionnée pour cette représentation afin de visualiser plus particulièrement les contributions de chaque fonction de communication. Comme on peut le voir, il existe différents types de générateurs de contour : ceux responsables de l'encodage de la modalité (dans notre exemple, la modalité est affirmative), ceux responsables de la hiérarchisation des constituants syntaxiques

(relation groupe nominal sujet/groupe verbal par exemple)... Cet outil d'analyse va créer, après apprentissage sur notre corpus de 238 phrases, des générateurs de contours pour toutes les fonctions de communication présentes dans le corpus. Nous les utiliserons dans la phase de synthèse pour déduire la variation de la fréquence fondamentale et du coefficient d'allongement en fonction de la phrase à prononcer. Cette phase d'analyse de la prosodie acoustique n'est pas inutile *a priori* car on a pu vérifier sur 65 phrases qu'un codeur oraliste de Langue française Parlée Complétée a un débit ralentit d'un facteur compris entre 1 et 2 par rapport à un oraliste non codeur.

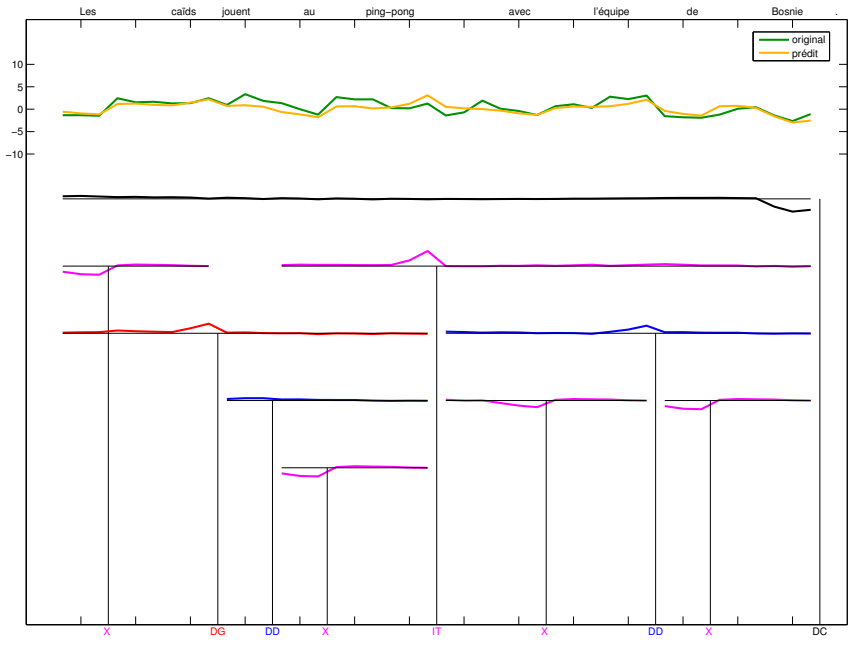


FIG. 9.6 – Variation de la fréquence fondamentale (les valeurs de F0 sont données en demi-tons par rapport à une fréquence de référence proche du registre moyen de la locutrice : 230 Hz) durant la phrase n° 89 du corpus «Les caïds jouent au ping-pong avec l'équipe de Bosnie» avec sa décomposition en contours chevauchants et sa reconstruction. Les traits sur l'abscisse indiquent les GIPCs (Group Inter Perception Center : unité rythmique qui correspond à l'intervalle entre deux centres perceptifs [9]).

9.3.2 Prosodie et mouvements de tête

Dans le cas de la production de la Langue française Parlée Complétée encore plus que pour la production du français, les mouvements de la tête et de la main vont concourir à la communication du message. Ainsi de nombreuses études traitent des relations entre mouvements de tête et prosodie, autant dans l'aspect perception [8, 11, 6, 12, 5] que dans l'aspect production [4, 7]. Dans notre cas, les mouvements de tête sont partie intégrante du discours puisqu'ils sont impliqués dans le mouvement de constriction avec la main pour former les clés et ils possèdent

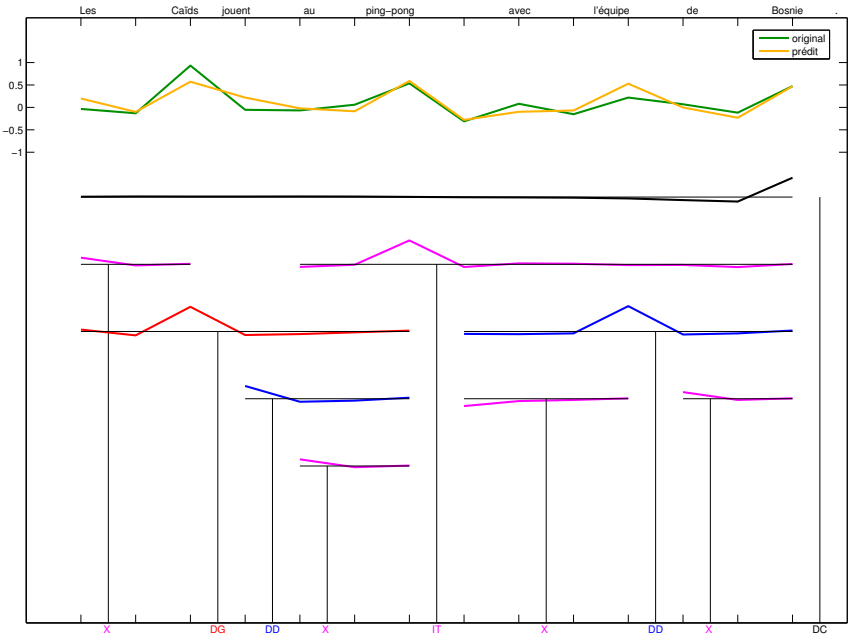


FIG. 9.7 – Variation du coefficient d'allongement (variation de la durée d'un GIPC par rapport à une durée attendue [9]) durant la phrase n° 89 du corpus «Les caïds jouent au ping-pong avec l'équipe de Bosnie» avec sa décomposition en contours chevauchants et sa reconstruction.

également des aspects prosodiques.

Contribution de la tête dans le mouvement de constriction

Nous allons tout d'abord étudier ce que l'on pourrait nommer le mouvement segmental du code LPC. Cette dénomination se rapporte aux mouvements de tête effectués dans le seul but de produire le contact main/visage (dans le cas des clés positions 2, 3, 4 et 5) ou l'éloignement de la main et du visage dans le cas de la clé position 1. Ce sont ces mouvements qui devront être contenus dans le dictionnaire multimodal qui sera à la base de notre système de synthèse.

Dans la définition originale du code LPC, la main vient effectuer une clé pour chaque série CV prononcée (CV peut aussi être une consonne isolée ou une voyelle isolée). La clé est définie par une forme de main et une position sur le visage. Dans le cas spécifique de notre codeuse, le mouvement consistant à pratiquer une constriction entre le visage et la main est dû, au premier ordre, au mouvement de la main (comme dans la définition initiale) mais aussi au deuxième ordre, au mouvement de la tête. Ainsi, la tête effectue des mouvements de roto-translation «segmentaux» : mouvements nécessaires à l'exécution du contact qui correspondent en moyenne à 16.43 % du mouvement total à parcourir pour qu'il y ait contact entre les 2 effecteurs. L'histogramme des proportions de mouvements effectués par la tête par rapport au mouvement total est représenté sur la figure 9.8.

Outre ces mouvements segmentaux, la tête effectue des mouvements que l'on nommera supra-

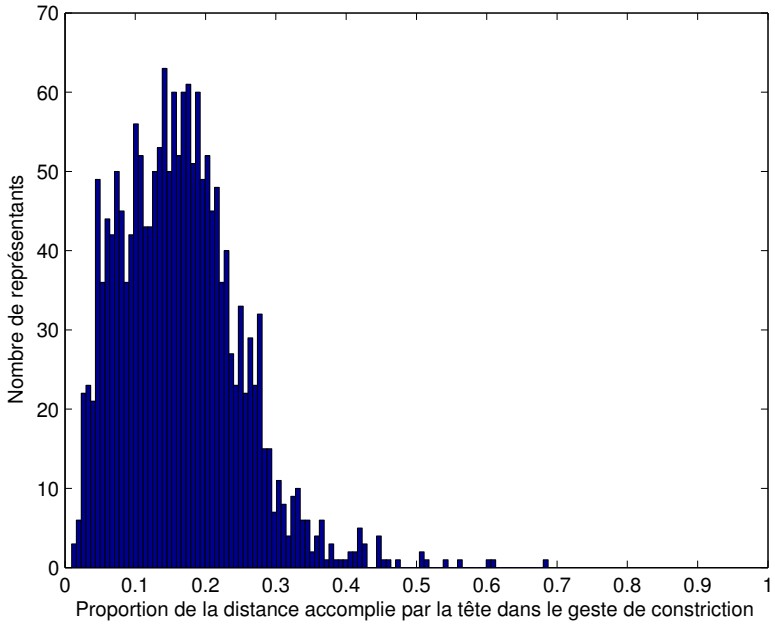


FIG. 9.8 – Histogramme des proportions de distance accomplie par la tête dans le mouvement de constriction pour les 1769 cibles (position 2, 3, 4 ou 5) du corpus.

segmentaux en directe relation avec les fonctions de communication composant la phrase à prononcer. Ces mouvements de tête ne sont pas anodins, ils sont en effet à la vidéo ce que le coefficient d'allongement et la fréquence fondamentale sont à l'audio. Leur étude nous permettra de mettre en place une prosodie à la fois acoustique et vidéo.

La prosodie des mouvements de tête

Pour étudier la partie prosodique des mouvements de tête, nous devons séparer les mouvements en une composante segmentale qui correspond à la constriction et en une composante supra-segmentale qui dépend cette fois-ci de la phrase et de sa structure. Notons que cette étape est nécessaire à la fois pour étudier la prosodie des mouvements de tête mais aussi pour construire un dictionnaire d'unités multimodales le plus cohérent possible. En effet, ces unités ne doivent comporter dans l'absolu que des mouvements de type segmental. Afin de répondre à ces deux impératifs, nous allons adopter la méthodologie suivante sur les 238 phrases de notre corpus dynamique :

1. nous retranchons, pour chaque phrase, la moyenne des mouvements de roto-translation de la tête sur la phrase aux mouvements de roto-translation de la tête et de la main ;
2. nous créons ensuite une position de repos global codée comme la clé 00 (clé de repos), ainsi le début et la fin de chaque phrase se termine au même endroit, ce qui pourrait être utile si l'on veut synthétiser un paragraphe par exemple ;
3. pour chaque transition d'une clé (position initiale vers position finale), nous calculons un mouvement moyen (normalisé en temps) sur toutes les occurrences de ce type dans le

corpus en séparant les cibles finales faisant intervenir l'index et le majeur ;

4. pour chaque phrase du corpus, nous retranchons aux mouvements de roto-translations originaux un mouvement de roto-translation issu des mouvements segmentaux moyens correspondant à la succession des cibles de la phrase ;
5. sur ce résidu, nous calculons une analyse en composantes principales, pour faire ressortir les mouvements principaux : on peut visualiser les deux premières composantes qui expliquent respectivement 47,64% et 28,52% des mouvements sur les figures 9.9 et 9.10. Le premier mouvement correspond à un mouvement de rotation autour de l'axe des abscisses avec une rotation autour de l'axe des cotes alors que le deuxième se compose presque uniquement d'une rotation autour de l'axe des cotes ;
6. nous utilisons ensuite l'outil d'analyse SFC comme pour les variables *classiques* de la prosodie mais cette fois-ci sur les deux mouvements principaux calculés sur chaque phrase ;
7. après apprentissage, nous retirons aux mouvements de roto-translation de chaque phrase celui calculé à partir de la structure prosodique de la phrase ;
8. nous sommes alors en possession d'un ensemble de 238 phrases auxquelles nous avons ôté la partie supra-segmentale des mouvements de tête. Nous pouvons donc construire notre dictionnaire d'unités multimodales. De plus, grâce aux SFC [9], nous pouvons prédire le mouvement de roto-translation supra-segmental pour chaque nouvelle phrase à synthétiser.

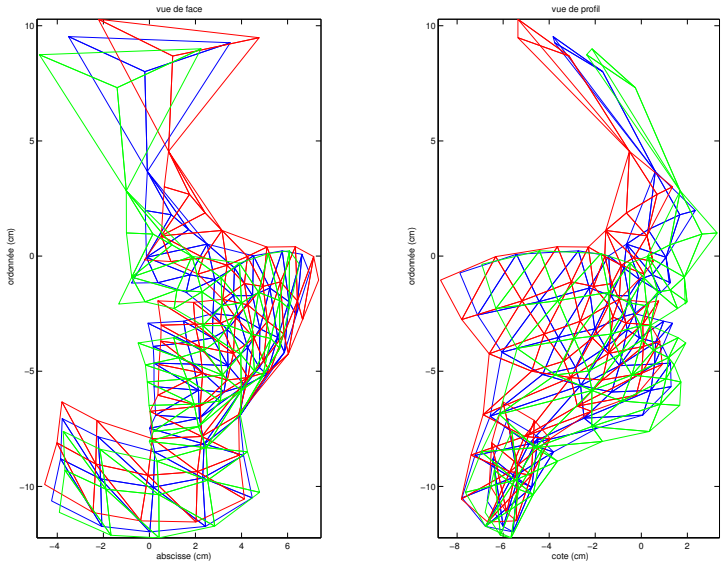


FIG. 9.9 – Variation de la première composante de l'analyse en composantes principales du résidu des mouvements de roto-translation de la tête. Les valeurs appliquées à ce paramètre vont de -3 (en vert, trait pointillé long) à +3 (en rouge, trait pointillé) en passant par 0 (en bleu, trait plein).

Nous avons représenté sur la figure 9.11 la variation et la décomposition en contours prototypiques du premier mouvement déduit par ACP sur les gestes supra-segmentaux pour la phrase

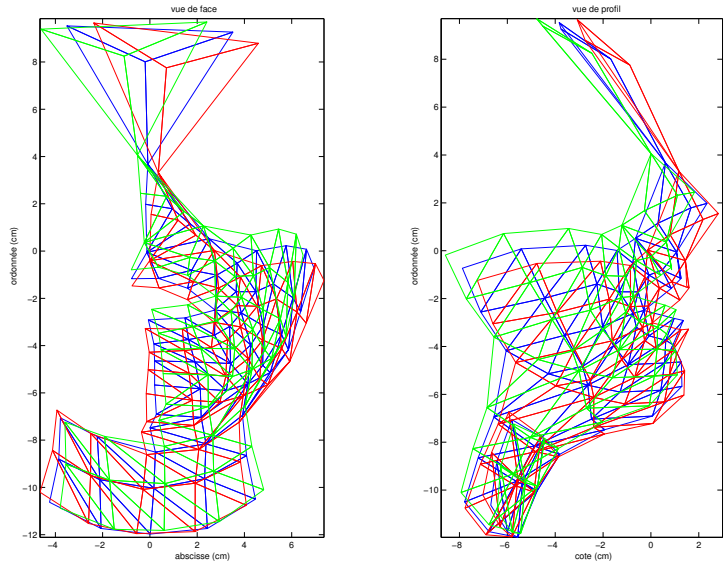


FIG. 9.10 – Variation de la deuxième composante de l'analyse en composantes principales du résidu des mouvements de roto-translation de la tête. Les valeurs appliquées à ce paramètre vont de -3 (en vert, trait pointillé long) à +3 (en rouge, trait pointillé) en passant par 0 (en bleu, trait plein).

«Les caïds jouent au ping-pong avec l'équipe de Bosnie» . Nous pouvons remarquer que l'on peut décomposer la variation du premier mouvement en 2 grands groupes, la séparation s'effectuant à la frontière de deux groupes prosodiques (d'un côté «Les caïds jouent au ping-pong» et de l'autre «avec l'équipe de Bosnie»).

Il serait intéressant dans une étude de perception de quantifier l'apport de ces mouvements de tête, directement relié à la structure prosodique de la phrase, dans l'intelligibilité de notre système de synthèse. En effet, il a déjà été montré que dans le cas de la langue des signes (suisse allemande) que les signeurs ayant appris jeunes cette langue étaient plus intelligible que les signeurs tardifs parce leur langue est plus «rythmée» [3].

9.4 Résumé

Après avoir pré-traité l'ensemble du corpus de 238 phrases, nous avons analysé la production de la Langue française Parlée Complétée délivrée par notre codeuse. Comme notre objectif est de construire un système de synthèse par concaténation, les modules nécessitant une analyse sont le module prosodique et le dictionnaire. C'est pourquoi, nous avons dans un premier temps vérifié la validité du code délivré puis segmenté le corpus afin de construire un dictionnaire de polysons multimodaux exempt de toute erreur (enfin, nous espérons !). Ensuite, afin de savoir si l'on pouvait espérer créer un seul et même dictionnaire contenant à la fois les segments de signal audio, de paramètres articulatoires liés au visage, de paramètres articulatoires liés à la main et de paramètres de roto-translation de la main et du visage, nous avons étudié le phasage

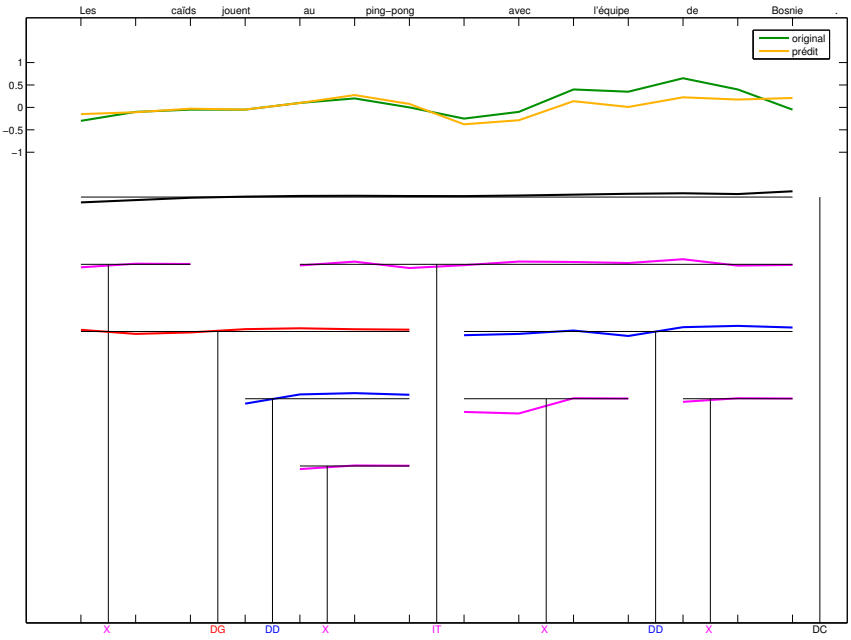


FIG. 9.11 – Variation du coefficient relatif au 1er mouvement déduit de l'ACP sur les gestes supra-segmentaux durant la phrase n° 89 du corpus «Les caïds jouent au ping-pong avec l'équipe de Bosnie» avec sa décomposition en contours chevauchants.

entre les gestes de la main et le signal acoustique. Nous en avons déduit que la main n'était pas synchronisée de la même façon que le visage : le geste de la main correspondant à la série CV en cours commence avant le début acoustique tandis que la cible est atteinte dans la première partie de la réalisation acoustique. Il nous faudra tenir compte de ces règles de production pour déterminer la meilleure façon de générer le code LPC. Enfin, nous avons modélisé et étudié la prosodie spécifique de notre locutrice codant le LPC tant au niveau acoustique qu'au niveau gestuel. Cette étude nous a permis de calculer les caractéristiques prosodiques et de *nettoyer* les mouvements de roto-translation de la main et du visage afin que notre dictionnaire de polysons multimodaux ne contiennent que des composantes segmentales.

La phase d'analyse effectuée, la phase de synthèse peut commencer. En se basant sur les conclusions des différentes études nous allons construire un système capable de générer un signal audio, des mouvements articulatoires de la main et du visage ainsi que des mouvements de roto-translation cohérents et respectant la synchronie du code LPC.

Références bibliographiques

[1] V. Attina. *La Langue française Parlée Complétée (LPC) : Production et Perception*. PhD thesis, Institut National Polytechnique de Grenoble, 2005.

[2] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables : rules for a Cued Speech synthesizer.

Speech Communication, 44 :197–214, 2004.

[3] P. Boyes Braem. Rhythmic Temporal Patterns in the Signing of Deaf Early and Mate Learners of Swiss German Sign Language. *Language and Speech*, 42(2–3) :177–208, 1999.

[4] M. Costa, T. Chen, and F. Lavagetto. Visual prosody analysis for realistic motion synthesis of 3D head models. In *International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging*, 2001.

[5] M. Dohen. *Deixis prosodique multisensorielle : production et perception audiovisuelle de la focalisation contrastive en français.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, December 2005.

[6] H.P. Graf, E. Cosatto, V. Strom, and F.J. Huang. Visual prosody : facial movements accompanying speech. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[7] B. Granström and D. House. Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46 :473–484, 2005.

[8] U. Hadar, T.J. Steiner, E.C. Grant, and F. Clifford Rose. Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26 :117–129, 1983.

[9] B. Holm. *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application à l'énonciation de formules mathématiques.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2003.

[10] B. Holm and G. Bailly. SFC : a trainable prosodic model. *Speech Communication*, 46(3–4) :348–364, 2005.

[11] E. Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32 :855–878, 2000.

[12] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. Head movements improves auditory speech perception. *Psychological Science*, 15(2) :133–137, 2004.

[13] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

Chapitre 10

La synthèse par concaténation de parole audiovisuelle

Enfin de la synthèse...

Le but essentiel de ces travaux de recherche, ne l'oublions pas, est de mettre en oeuvre un système de synthèse de parole audiovisuelle augmentée, c'est-à-dire capable de générer à partir d'un texte quelconque un signal audio, des mouvements du visage et de la main correspondants. Les signaux de sortie doivent être synchrones et cohérents afin que la transcription en Langue française Parlée Complétée de la phrase d'entrée soit la plus intelligible possible (le but à atteindre étant l'intelligibilité du codage humain de la même phrase). Nous avons décrit dans les paragraphes de la partie **Etat de l'art** les modules constituant un système conventionnel de synthèse de parole, qu'elle soit audio ou audiovisuelle. L'approche qui va être utilisée pour synthétiser le code LPC est la synthèse par concaténation ; elle a été largement utilisée pour la synthèse de parole audio [10, 7] et aussi audiovisuelle [11]. Cependant, c'est la première fois qu'elle est utilisée pour synthétiser du code LPC.

Quels éléments devons-nous concaténer ?

Une fois définie l'approche utilisée, il reste à définir les ingrédients nécessaires à la synthèse. Dans le cas de la concaténation, il s'agit des unités, des briques élémentaires ; nous allons considérer deux types d'unités :

- les diphones (partie du signal comprise entre deux allophones successifs) pour la génération du signal de synthèse audio et les mouvements (articulatoires) du visage ;
- les diclés (partie du mouvement comprise entre deux clés successives) pour la génération des mouvements de tête, des mouvements de main et de l'articulation de la main.

Comment devons-nous concaténer ces éléments ?

Les unités élémentaires ayant été choisies, il reste encore à déterminer la méthode de concaténation. Du fait de la synchronisation particulière de la main par rapport au signal acoustique (cf. chapitre **Analyse de la production du code LPC**) et du modèle de «rendez-vous» entre la main et le visage, notre système procéde en deux étapes :

1. un premier système de synthèse par concaténation génère le signal audio et les mouvements articulatoires du visage en utilisant la méthode TDPSOLA avec des polysons (des diphones dans le pire des cas) ;

2. l'articulation de la main et les mouvements de la tête et de la main sont générés par un second système de synthèse par concaténation utilisant la même méthode que précédemment avec des diclés comme unités.

Alors que le premier système de synthèse utilise les caractéristiques prosodiques délivrées par le module de même nom, le second système se base sur une transcription de la chaîne phonétique marquée en temps et traduite en une chaîne de clés LPC avec un marquage en temps correspondant aux contraintes déduites de l'analyse. Le diagramme de notre système est représenté sur la figure 12.1.

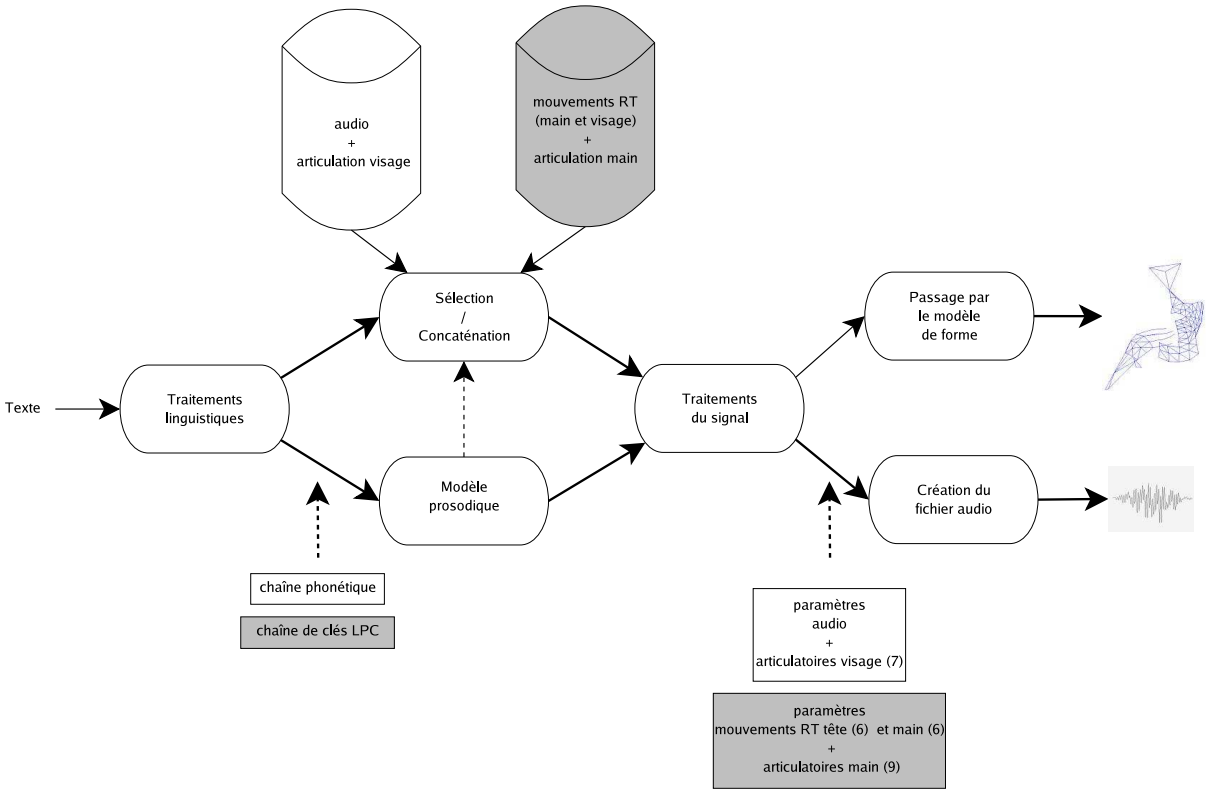


FIG. 10.1 – Diagramme du système de synthèse de Langue française Parlée Complétée à partir du texte : un premier sous-système de synthèse par concaténation de diphones multimodaux (paramètres audio et paramètres articulatoires du visage) est couplé à un second sous-système de synthèse par concaténation de diclés (paramètres de roto-translation de la tête et de la main et paramètres articulatoires de la main).

Nous allons voir en détails comment chacun des modules est implémenté et comment ils s'enchaînent afin de fournir en sortie le code LPC audiovisuel de synthèse.

10.1 Les traitements linguistiques

La première étape dans la synthèse de parole (audiovisuelle) à partir du texte consiste à traduire le texte fourni en entrée en une chaîne phonétique. Le système que nous avons utilisé pour effectuer cette tâche est le système COMPOST [6, 3, 2, 1] développé à l'Institut de la Communication Parlée.

Ce système effectue une série de traitements (linguistique, phonologique, phonétique) sur une structure linguistique multi-niveaux (SLM) c'est-à-dire un arbre dont les noeuds sont des instances de divers objets permettant de décrire la structure linguistique, phonologique ou phonétique de la phrase d'entrée. La série de traitements se décompose en trois phases :

1. le pré-traitement du texte permet de désambiguïser tous les sigles, dates, nombres, etc. et de les transcrire en suites de caractères alphanumériques que le module suivant pourra traiter ;
2. les traitements linguistiques permettent de déterminer les mots (et leurs classes lexicales) composant la phrase ; pour cela, une analyse morphologique (qui exploite un lexique de 65000 morphèmes) délivre un treillis de possibilités qui est exploité par un modèle de trigrams fournissant la classe lexicale la plus probable de chaque mot parmi les 26 classes prédéfinies ;
3. les traitements phonologiques permettent de faire la transcription orthographique-phonétique et de délivrer ainsi la chaîne phonétique correspondant à la phrase à synthétiser.

Cette suite de traitements fournit la chaîne phonétique mais pas encore d'information quant au *timing*. C'est le module suivant, le modèle prosodique qui délivre les caractéristiques temporelles et spectrales que le signal de synthèse doit vérifier.

10.2 Le module prosodique

La prosodie est un élément essentiel en synthèse de la parole : elle permet de faire véhiculer une émotion, un sentiment et elle est clairement un support essentiel du message. Pour la modéliser, nous nous basons sur les travaux de Holm et al. [5, 9] et leur modèle SFC (Superposition of Functional Contours).

Dans une phase d'apprentissage (cf. chapitre **Analyse de la production du code LPC**), le découpage des relations hiérarchiques de dépendances sur chaque phrase est utilisé pour paramétrer automatiquement des générateurs de contours. Dans la phase de synthèse, ceux-ci sont capables de générer la variation de la fréquence fondamentale (F0) et du coefficient d'allongement (C) pour toute nouvelle phrase. Pour plus de détails sur le fonctionnement de ce système SFC, vous pouvez vous référer à la thèse de B. Holm [8]. Pour résumer, des contours chevauchants, appris durant la phase d'apprentissage sur le corpus contenant 238 phrases, sont utilisés pour générer la variation de la fréquence fondamentale et du coefficient d'allongement pour la phrase d'entrée. Ces contours sont spécifiques à la locutrice qui plus est à la locutrice codant LPC pour le corpus donné.

Ainsi, pour chaque nouvelle phrase, on utilise ce module pour générer la variation de F0 et de C, ce qui permet d'avoir une chaîne phonétique marquée en temps. C'est elle que nous allons fournir au module sélection/concaténation.

A partir de cette chaîne phonétique marquée en temps, nous déduisons également la suite de clés à former pour être conforme à la phrase à prononcer. Un simple traducteur permet de passer de la chaîne phonétique à la suite de clés : en effet, à chaque série CV correspond une seule clé codée par une forme et une position. Le traducteur découpe la chaîne phonétique en

séries CV (voire C isolées et V isolées si tel est le cas) puis génère la suite de clés. En ce qui concerne le timing, nous nous basons sur le marquage en temps de la chaîne phonétique et nous appliquons les relations (moyennes) de phasage que nous avons déduites du corpus (cf. chapitre **Analyse de la production du code LPC**). C'est ainsi qu'à la sortie de ce module nous avons deux chaînes : une chaîne phonétique marquée en temps et un chaîne de clés LPC *synchronisée*.

Et les mouvements de tête ?

Dans le chapitre correspondant à l'analyse des données, nous avons modélisé les mouvements de roto-translations de la tête qui correspondent à des mouvements supra-segmentaux. Les générateurs de contours existent puisqu'ils ont été appris à l'aide du modèle SFC. Toutefois, ils n'ont pas encore été appliqués en synthèse. Il serait intéressant d'analyser et de quantifier l'apport de ces mouvements dans la synthèse de code LPC.

10.3 Synthèse par concaténation : la sélection

Notre système de synthèse de code LPC fonctionne en 2 étapes : la première étape génére le signal audio et la série de paramètres articulatoires du visage, alors que la seconde étape génére la série de paramètres articulatoires de la main et de paramètres de roto-translation de la main et du visage. Nous avons défini précédemment les unités utilisées dans les deux cas ; la sélection de ces unités s'opére de la façon suivante : un algorithme de programmation dynamique recherche dans le treillis de candidats possibles, ceux réalisant la distance cumulée minimale. Cette distance est composée de 2 coûts [12] : un coût de sélection et un coût de concaténation. Nous allons voir dans chacune des 2 étapes à quoi ils correspondent exactement.

10.3.1 Son et mouvements faciaux

Cette première étape se base sur la chaîne phonétique marquée en temps délivrée par le module prosodique.

Complétude des unités

Le premier système consiste en un système de concaténation basé sur des polysons multimodaux. Or, le corpus de 238 phrases, à la base de notre système, a été construit afin d'avoir une couverture quasi-optimale des polysons du français : sur les 7279 polysons présents, la moitié sont au moins présents en 2 exemplaires. C'est donc la construction de ce corpus qui nous assure la complétude des unités.

Sélection des unités

Comme certains polysons sont multi-représentés, nous devons déterminer la meilleure (au sens d'un coût à déterminer) chaîne de polysons à concaténer. Nous définissons 2 types de coûts :

1. le coût de sélection : il quantifie l'adéquation de l'unité sélectionnée à la tâche phonologique. Dans notre cas, il permet de trouver le candidat le plus proche des spécifications prosodiques (délivrées par le module prosodique) ; plus précisément, il s'agit d'une métrique

basée sur la différence de durée entre segments concaténés et spécification rythmique déterminée par le module prosodique ;

2. le coût de concaténation : il quantifie la gêne perceptive engendrée par la juxtaposition de l'unité avec l'unité précédente. Dans notre cas, il est proportionnel à la distance RMS (Root Mean Square) entre les paramètres spectraux LSP (Line Spectrum Pairs) aux points de concaténation et à la distance RMS entre les paramètres articulatoires (pondérée par la variance du mouvement expliqué par chaque paramètre) aux points de concaténation.

10.3.2 Mouvements de main et de tête

Cette deuxième étape se base sur la chaîne de diclés déduite de la chaîne phonétique marquée en temps délivrée par le module prosodique et des relations de phasage déduites de la phase d'analyse.

Complétude des unités

Le second système concatène des diclés : une clé (l'association de la forme et de la position de la main) est référencée par deux chiffres, le premier correspondant à la forme, le deuxième à la position par rapport au visage. Dans ces diclés, sont inclus les mouvements articulatoires de la main et les mouvements de roto-translation de la main et du visage. En effet, comme nous l'avons vu dans le chapitre précédent, les mouvements de tête font partie intégrante du geste de constriction, il est donc logique de les coupler aux mouvements de roto-translation de la main. Quant aux mouvements articulatoires de la main, ils sont intimement liés aux mouvements de roto-translation, pour preuve les corrélations (linéaires) importantes entre les débuts et fins de mouvements respectifs.

Nous avons vu dans le chapitre consacré au corpus dynamique que la complétude était assurée pour les transitions de type position-position et forme-forme, en revanche pour ce qui est des transitions $(forme + position)_1$ vers $(forme + position)_2$ la complétude n'est pas assurée. En effet, le corpus n'a pas été construit initialement pour la Langue française Parlée Complétée et ne contient pas toutes les transitions CV_1-CV_2 possibles (où CV peut être une série consonne-voyelle ou consonne isolée ou voyelle isolée).

Afin de pouvoir synthétiser n'importe quelle phrase en entrée, nous avons complété notre corpus de diclés par des diclés reconstruites. Plusieurs possibilités se présentent à nous : soit on essaie d'extraire des règles de transition entre formes et/ou entre positions afin de synthétiser par règles les diclés manquantes, soit on se base sur des transitions «réelles» de formes et/ou de positions que l'on modifie afin de créer des diclés de synthèse pour les représentants manquants. C'est la deuxième solution que nous avons privilégiée afin de conserver le naturel des mouvements. L'hypothèse utilisée (qui n'est pas vérifiée dans l'absolu) est que les mouvements articulatoires de la main et les gestes de la main sont indépendants. Pour toutes les diclés absentes du corpus, nous avons procéder comme suit :

1. dans toutes les transitions position-position du corpus, nous déterminons et stockons la plus longue (en nombre de trames) ;
2. de la même manière, nous stockons la plus longue transition forme-forme ;

3. à l'aide du système de reconnaissance de clés (forme et position), nous déterminons pour chacune des transitions précédentes l'instant de transition entre la clé de départ et la clé de fin ;
4. nous concaténons chacune des sous-parties (début-transition) et (transition-fin) en rééchantillonant par rapport au segment le plus court ;
5. on obtient une diclé de synthèse.

Sélection des unités

Comme certaines diclés sont multi-représentées, nous devons déterminer la meilleure chaîne de diclés à concaténer (au sens d'un coût à déterminer). Nous définissons 2 types de coûts comme précédemment :

1. un coût de sélection correspondant à la différence de durée entre les diclés concaténées et la spécification de la chaîne de diclés déduite de la chaîne phonétique marquée en temps ;
2. un coût de concaténation proportionnel à la distance RMS entre les paramètres articulatoires et de roto-translations (pondérée par la variance du mouvement expliqué) aux points de concaténation.

10.4 Synthèse par concaténation : la concaténation et le lissage

Une fois les unités sélectionnées, qu'il s'agisse d'une part des polysons multimodaux ou d'autre part des diclés, un algorithme de type TDPSOLA nous sert à concaténer ces unités et à vérifier les contraintes prosodiques imposées par le modèle SFC. En effet, bien qu'il y ait un coût *prosodique* lors de notre phase de sélection, les unités ne correspondent que très rarement de façon parfaite aux contraintes.

De plus, bien que les unités aient été sélectionnées pour correspondre au mieux (au sens du coût de concaténation) avec leurs voisines, il reste encore des artefacts liés à la concaténation. Pour éviter les sauts liés à cette méthodologie, nous implémentons une procédure de lissage anticipatoire [4] sur les mêmes paramètres qui nous ont servi lors de la phase de sélection dans le calcul du coût de concaténation. Cette procédure compense les sauts aux frontières inter-polysons et inter-diclés : une interpolation linéaire est calculée sur le saut observé durant le polyson (ou diclé) précédent. Un exemple de ce lissage anticipatoire peut être visualisé sur la figure 10.2 qui représente la variation du premier paramètre articulatoire de la main (ang1) pour la phrase de synthèse : «Bonjour !».

10.5 Le modèle de forme

A ce point, nous sommes capables de générer une suite de paramètres (articulatoires et de mouvements de roto-translation) ainsi qu'un fichier audio pour n'importe quelle nouvelle phrase. Afin de visualiser le résultat (en termes géométriques) de notre synthèse, nous devons utiliser le modèle de forme ou plutôt les modèles de forme (un pour le visage, un pour la main) calculés dans le chapitre «Pré-traitements : modélisations statistiques». Ainsi, à chaque trame (à 120 Hz)

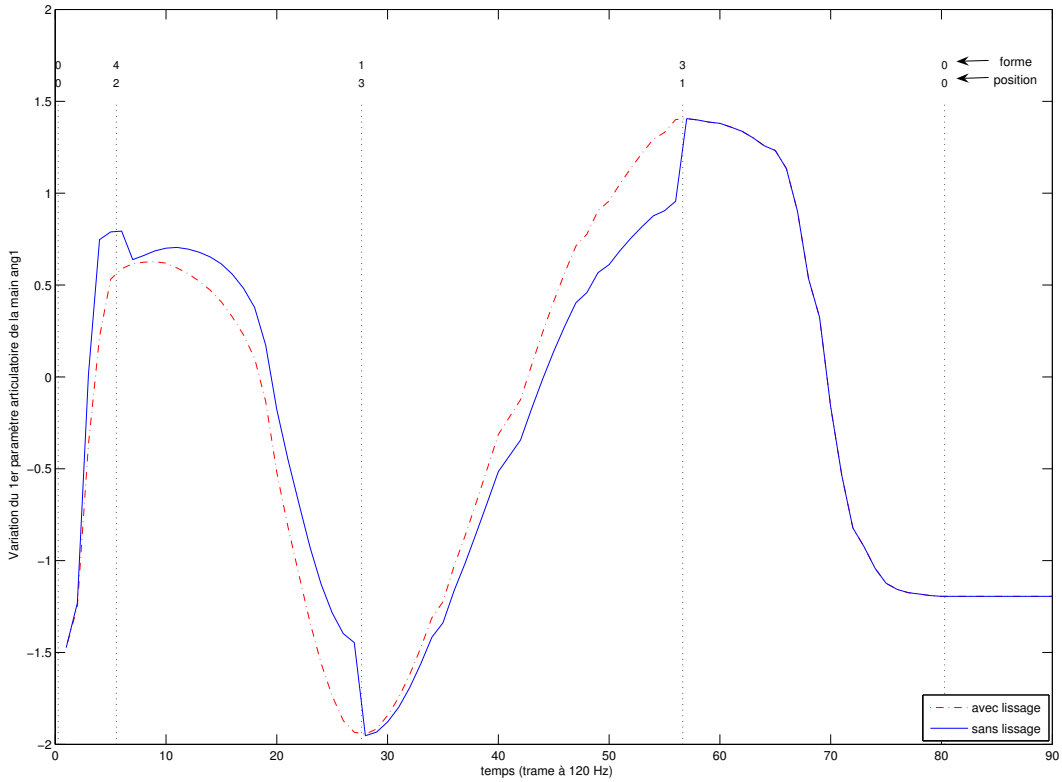
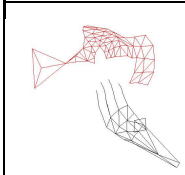
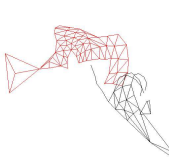
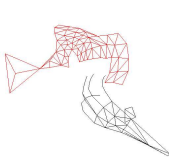
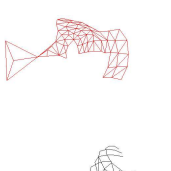
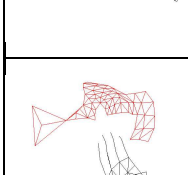
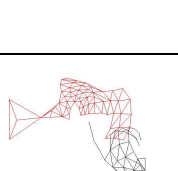
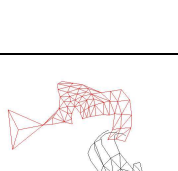
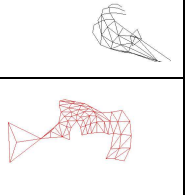
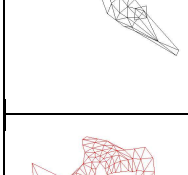
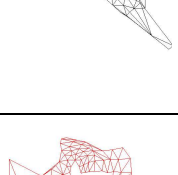
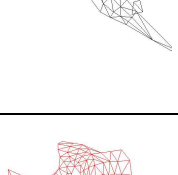
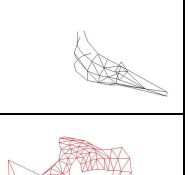
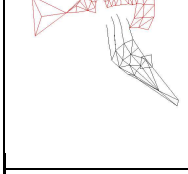
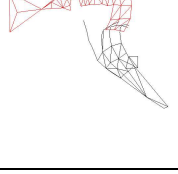
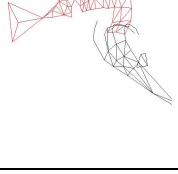
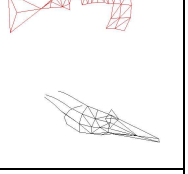
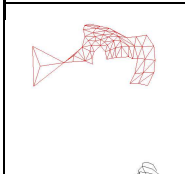
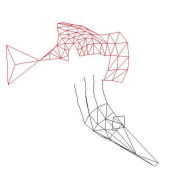
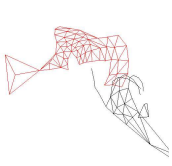
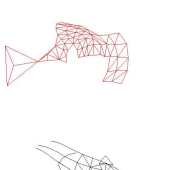
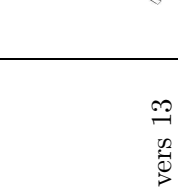
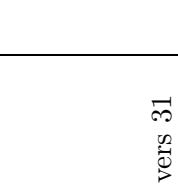
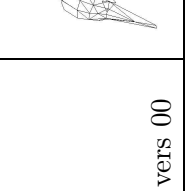
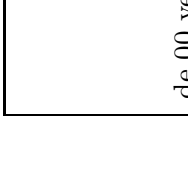
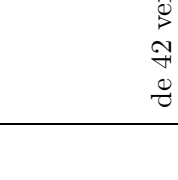
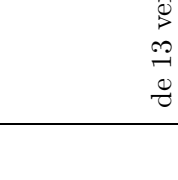
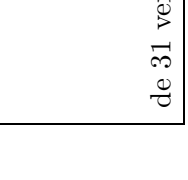
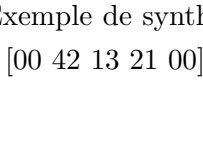
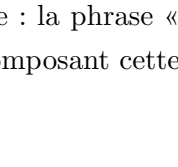
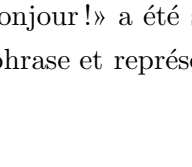
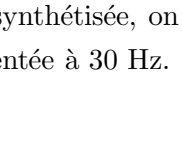


FIG. 10.2 – Variation du premier paramètre articulatoire de la main (ang1) au cours du temps pour la phrase de synthèse «Bonjour !» : en bleu (trait plein) concaténation sans lissage et en rouge (trait point) concaténation avec lissage anticipatoire.

correspond un ensemble de 7 paramètres articulatoires et 6 de roto-translation pour le visage et un ensemble de 9 paramètres articulatoires et 6 de roto-translation pour la main. A cet ensemble de paramètres nous faisons correspondre une géométrie, grâce aux modèles de forme, à chaque objet (main et visage).

Un exemple de phrase de synthèse est représenté sur le tableau 10.1. Nous avons synthétisé la phrase «Bonjour !» et nous avons représenté les différentes postures du visage et de la main à 30 Hz. À ce point de l'étude, les modèles de forme sont des modèles basse définition et nous ne sommes donc capables de présenter que 63 points sur le visage et 50 points sur la main.

Afin de montrer la validité de notre synthèse, nous avons représenté sur la figure 10.3 la variation des probabilités issues des modèles gaussiens (pour la forme et la position) pour la synthèse par concaténation avec et sans lissage. Les clés correspondantes à la phrase d'entrée sont correctement reconnues par le système de reconnaissance développé lors de la phase d'analyse.

de 00 vers 42	de 42 vers 13	de 13 vers 31	de 31 vers 00
			
			
			
			
			
			
			
			

TAB. 10.1 – Exemple de synthèse : la phrase «Bonjour !» a été synthétisée, on a décomposé la série de diclés [00 42 13 21 00] composant cette phrase et représentée à 30 Hz.

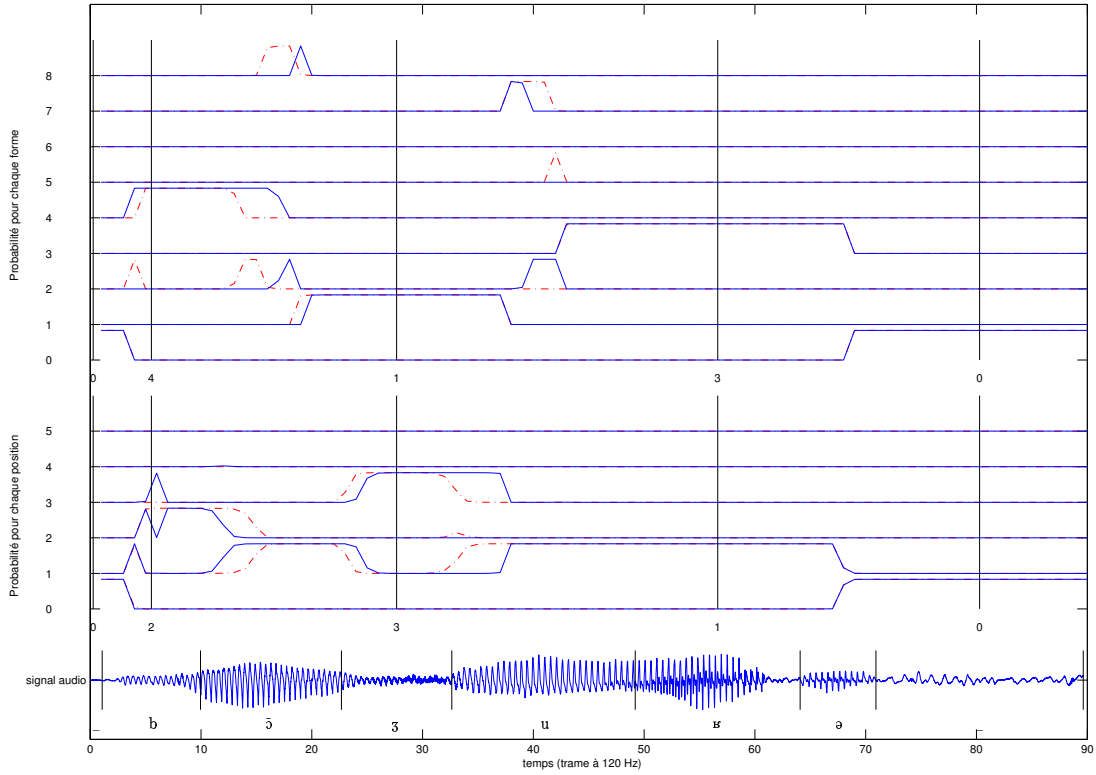


FIG. 10.3 – Variation des probabilités issues des modèles gaussiens pour la forme (haut) et la position (bas) de la main pour la phrase de synthèse «Bonjour !» : en bleu (trait plein) concaténation sans lissage, en rouge (trait point) concaténation avec lissage.

10.6 Résumé

Le système de synthèse implémenté est basé sur le paradigme de concaténation d'unités multimodales pré-stockées. Or, nous avons vu dans le chapitre consacré à l'analyse des données que des contraintes temporelles existaient entre les différents articulateurs et le signal acoustique correspondant. Ainsi, il ne nous était pas possible de créer un seul et unique dictionnaire contenant des unités multimodales. La résolution de ce problème passe par la création de 2 dictionnaires et par une synthèse en 2 étapes. Le premier dictionnaire contient des polysons multimodaux (signal audio et paramètres articulatoires du visage) tandis que le second contient des diclés (paramètres articulatoires de la main et mouvements de roto-translation de la tête et de la main). Cette solution permet de respecter les synchronisations des différents mouvements observées dans la phase d'analyse. Toutefois, la synthèse ainsi générée ne délivre que les coordonnées de 63 points sur le visage et 50 points sur la main (via les paramètres articulatoires et de roto-translation) à la fréquence de 120 Hz. La définition spatiale est encore insuffisante pour pouvoir annoncer que notre système de synthèse est abouti. Nous allons voir dans le chapitre suivant comment nous allons compléter nos données afin de pouvoir générer pour chaque trame plusieurs centaines de

points sur la main et le visage et même donner une apparence vidéo-réaliste au visage parlant.

Références bibliographiques

[1] M. Alissali. *Architecture logicielle pour la synthèse multilingue de la parole.* PhD thesis, INPG, Grenoble, France, 1993.

[2] M. Alissali and G. Bailly. Compost : a client-server model for applications using text-to-speech. In *European Conference on Speech Communication and Technology*, pages 2095–2098, Berlin, Germany, 1993.

[3] G. Bailly and M. Alissali. Compost : a server for multilingual text-to-speech system. *Traitement du Signal*, 9(4) :359–366, 1992.

[4] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

[5] G. Bailly, B. Holm, and V. Aubergé. A trainable prosodic model : learning the contours implementing communicative functions within a superpositional model of intonation. In *International Conference on Speech and Language Processing*, pages 1425–1428, Jeju, Korea, 2004.

[6] G. Bailly and A. Tran. Compost : a rule-compiler for speech synthesis. In *European Conference on Speech Communication and Technology*, pages 136–139, 1989.

[7] N. Campbell. Computing prosody : Computational models for processing spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Synthesizing Spontaneous Speech*, pages 165–186. Springer-Verlag, 1997.

[8] B. Holm. *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application à l'énonciation de formules mathématiques.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2003.

[9] B. Holm and G. Bailly. SFC : a trainable prosodic model. *Speech Communication*, 46(3–4) :348–364, 2005.

[10] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, pages 373–376, Atlanta, GA, 1996.

[11] S. Minnis and A. P. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *International Conference on Speech and Language Processing*, pages 759–762, Beijing, China, 2000.

[12] K. Takeda, K. Abe, and Y. Sagisaka. On the basic scheme and algorithms in non-uniform units speech synthesis. In G. Bailly and C. Benoît, editors, *Talking machines : Theories, Models and Designs*, pages 93–105. Elsevier B.V., 1992.

Chapitre 11

Passage à la haute définition et à l'apparence

Le synthétiseur génère du mouvement mais qu'en est-il de la restitution ? Les corpora enregistrés précédemment ne nous permettent que d'afficher un nombre de points limité sur le visage et la main de la codeuse. Nous sommes donc dans l'incapacité de fournir en sortie de notre système une tête parlante vidéo-réaliste. Pour pallier à ce problème nous allons voir les méthodologies utilisées (différentes pour le visage et la main) pour compléter le nombre de points de nos deux objets et appliquer une texture sur chacun d'eux. Ces méthodologies nécessitent l'enregistrement d'autres corpora et l'utilisation de moulages. Elles permettent de construire un modèle de forme (Haute Définition spatiale HD) qui supplante le modèle de forme précédent et d'y associer un modèle d'apparence. Après enregistrement des différents corpora, il s'agit de déterminer des règles de passage entre le modèle de contrôle et le nouveau modèle de forme afin que les paramètres d'animation que fournit le synthétiseur pilotent une tête et une main en haute définition spatiale. Puis, nous détaillerons l'enregistrement des corpora nécessaires au modèle d'apparence et nous montrerons comment ce dernier «habillera» le modèle de forme HD.

11.1 Modèles de forme Haute Définition

Les méthodologies utilisées pour le visage et la main étant différentes, nous séparerons l'étude de ces deux objets : alors que dans le premier cas de nouveaux corpora vidéo sont utilisés, le deuxième cas nécessite un moulage de la main.

11.1.1 Le visage

Nous avons enregistré 2 corpora pour passer d'un modèle BD (Basse Définition) à un modèle HD (Haute Définition) du visage. Le premier corpus nous permet de construire le modèle de forme du visage et le second nous permet de construire le modèle de forme de la tête.

Description des corpora supplémentaires

Le premier corpus appelé **visage+billes** est un corpus composé de 46 visèmes statiques : ils sont issus de la réalisation de voyelles isolées et de consonnes en contexte symétrique VCV, où V appartient à une des voyelles suivantes [a], [i] ou [u]. Ce corpus privilégie le visage de face et de profil (*cf.* figure 11.1). La codeuse a pour objectif de tenir l'articulation de chacun de ces visèmes pendant l'enregistrement.

Le deuxième corpus appelé **tête en rotation** est un corpus de visèmes tenus dans le temps et choisis pour leurs caractéristiques à faire apparaître ou disparaître certains plis et détails sur le visage. Ce corpus privilégie les angles de vue au détriment de certaines articulations (voir tableau 11.1).

Protocole d'enregistrement

FIG. 11.1 – Corpus **visage+bille** : les 5 vues capturées par les 3 caméras pour le visème [utu].

Le corpus **visage+billes** a été conçu et enregistré à l'ICP. Des billes colorées au nombre de 247 ont été collées sur le visage de la codeuse. Lors de l'enregistrement, 3 caméras (50 Hz, PAL) enregistrent l'image de la codeuse et celle envoyée par les deux miroirs latéraux placés derrière elle (ce qui donne cinq vues comme représenté sur la figure 11.1). Il est à noter que ce corpus est suffisant à lui seul pour la création d'un modèle statistique articulatoire du visage mais il est insuffisant pour pouvoir piloter ce modèle en vue de la création d'un système synthèse. Une phase préliminaire de calibration (avec un objet de référence) permet de connaître les coordonnées précises de chaque bille dans les 5 vues après cliquage manuel. Ce nouveau corpus offre une définition spatiale bien plus importante que lors de l'enregistrement des corpora dynamiques. Cependant, un nombre limité de billes pouvant être placées sur les lèvres sans gêner l'articulation, nous avons eu recours à un modèle générique 3D de lèvres [4] ajusté manuellement sur chacun des visèmes.

Le corpus **visage en rotation** a lui aussi été conçu et enregistré à l'ICP. Un sous-ensemble des 247 billes du corpus précédent est conservé sur le visage soit environ 40 billes. Des pastilles

TAB. 11.1 – Corpus **visage en rotation** : les 16 vues pour le visème [afa].

placées sur les cheveux de la codeuse ont été rajoutées. Pour chaque réalisation de visème, on enregistre 16 vues comme représenté sur le tableau 11.1, elles permettent de disposer d'un tour complet de la tête de la locutrice.

Méthodologie de passage de la BD à la HD

La première partie du passage d'un modèle de forme basse définition à haute définition spatiale concerne le visage, c'est-à-dire les parties de notre tête parlante contrôlées par des paramètres articulatoires. Pour cela, nous allons utiliser le corpus **visage+billes** : les 247 billes de ce corpus couvrent l'ensemble du visage, certaines étant même replacées dans des zones proches de celles de certains marqueurs du corpus dynamique basse définition spatiale. Une première méthodologie pourrait consister à mettre en correspondance ces points communs entre les deux corpora mais malheureusement ces correspondances ne sont pas suffisantes pour pouvoir calculer automatiquement un modèle de forme HD. Nous avons donc choisi une autre méthodologie : pour chaque visème du corpus **visage+billes**, nous estimons, dans un premier temps, un mouvement rigide à partir des positions de points placés sur la partie supérieure de la tête (il s'agit de parties fixes au sens où elles ne sont pas modifiées par les paramètres articulatoires) puis nous calculons par optimisation (moindres carrés) la série de paramètres articulatoires correspondante à ce visème. Pour le calcul d'optimisation, nous utilisons des points de correspondance placés sur les lèvres et sur la mâchoire (un exemple de cette correspondance est représentée sur la figure 11.2). Une fois cette opération faite pour tous les visèmes du corpus, nous effectuons une simple régression linéaire entre les paramètres articulatoires et les coordonnées 3D des 247 points. Le nouveau modèle de forme créé est compatible avec les paramètres de contrôle du modèle de forme précédent.

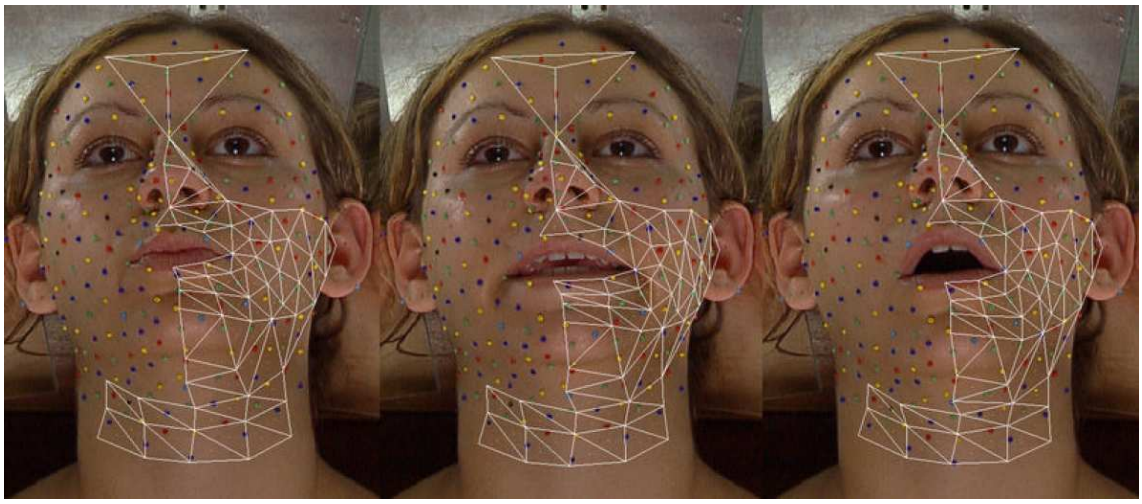


FIG. 11.2 – Correspondance entre le modèle de forme basse définition et 3 visèmes (vue de face) du corpus **visage+billes** : après avoir déterminé un mouvement rigide optimal, une optimisation permet de trouver les paramètres articulatoires correspondants à la géométrie 3D.

La deuxième partie du passage de la basse définition à la haute définition correspond à la modélisation du volume de la tête. Cette partie est considérée comme rigide et n'admet des transformations que de type roto-translation. Pour cela, nous allons utiliser le corpus **visage en rotation** : les 16 vues pour chaque visème permettent, entre autre, de calculer les coordonnées des pastilles placées sur les cheveux de la codeuse ; ces points vont servir de référence pour déformer un maillage générique de tête [1] (représenté sur la figure 11.3). Une fois l'adaptation du modèle générique sur les caractéristiques volumiques de la tête de notre codeuse réalisée, nous ne conservons que la partie du maillage arrière pour compléter notre modèle de forme.

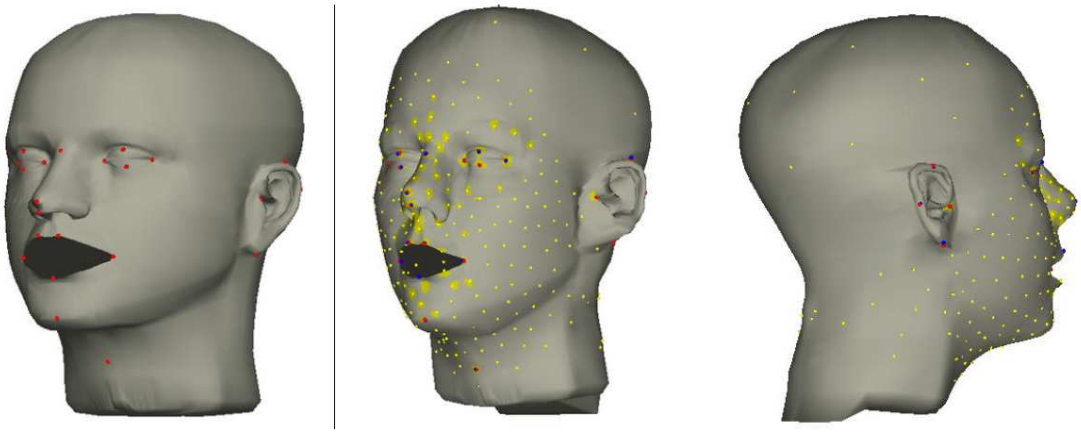


FIG. 11.3 – Tête générique : à gauche avant déformation, au centre et à droite après déformation grâce aux points de référence du corpus **visage en rotation**.

A ce stade, nous possédons un modèle de forme Haute définition de la tête qui est contrôlé par les mêmes paramètres articulatoires et de roto-translation que le modèle basse définition (calculé dans la partie analyse). Mais cette fois-ci, nous contrôlons 1934 points contre 63 pour

le précédent.

11.1.2 Les dents

Un moulage de la dentition de la locutrice permet de créer un modèle de dents 3D. Or, la position des incisives supérieures et inférieures est prédite par le modèle linéaire. Le modèle de dents est attaché à ces points. Le mouvement des dents inférieures est controllé géométriquement par la combinaison d'une rotation autour de l'axe défini par les condyles (les éminences articulaires, arrondies par un de leurs côtés, aplaties dans le reste de leur étendue) et d'une translation dans le plan medio-sagital (voir figure 11.4) [3].

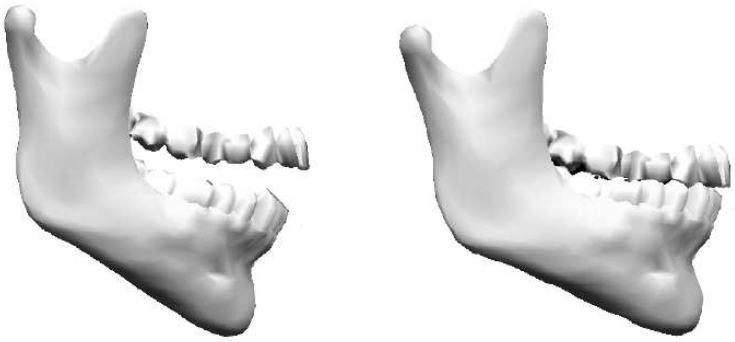


FIG. 11.4 – Illustration d'un modèle 3D de mâchoire et de dents pour le geste d'ouverture tirée de [3]

11.1.3 La main

La méthodologie utilisée pour la construction du modèle HD de la main est différente de celle utilisée pour le visage : le moulage de la main est une solution au problème posé par le grand nombre de degrés de liberté de cet objet 3D.

Description des données supplémentaires

La main de la codeuse a été moulée dans deux positions différentes (cf. figure 11.5 (a)) lors de l'enregistrement des corpora dynamiques chez Attitude Studio. Ces moulages ont été réalisés avec la présence des marqueurs réfléchissants utilisés pour l'enregistrement des corpora. On connaît donc parfaitement les coordonnées 3D de ces 50 points sur la main moulée.

Technique de création et d'animation d'une main HD

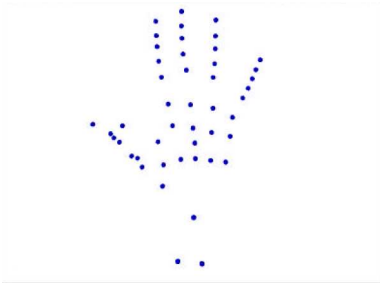
Toutes les opérations suivantes se sont déroulées chez Attitude Studio. Les moulages de la main sont scannés (cf. figure 11.5 (b)), ce qui permet d'obtenir un maillage 3D haute définition (2641 points) de la main. Ensuite, on adapte (redimensionnement) le maillage de notre main à un maillage générique qui possède une structure squelettique qui permet de l'animer. Ainsi, par des algorithmes de «skinning» (voir figure11.6), on est capable de générer, pour n'importe

(a) Moules de la main suivant deux positions

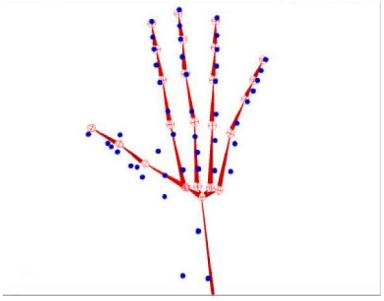
(b) Le maillage HD défini et dessiné sur un des 2 moules

FIG. 11.5 – Moulages de la main suivant deux positions (a) , le maillage HD est dessiné sur l'une des 2 positions de la main (b). La position des marqueurs réfléchissants (modèle BD) est visible sur les moulages.

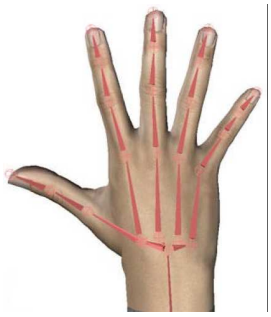
quelle forme de main BD, la forme de main HD et ainsi avoir accès aux coordonnées 3D de tous les points du maillage HD.



(a) 50 points tirés de la MoCap



(b) les os du squelette sont attachés aux points de MoCap



(c) le squelette pilote le modèle HD

FIG. 11.6 – Passage d'une configuration BD tirée de la MoCap à une configuration HD par la méthode de «skinning».

Méthodologie de passage de la BD à la HD

Afin de créer le modèle de forme HD pour l'objet 3D main, nous procédons en 3 étapes :

1. on détermine un sous-ensemble de paramètres articulatoires (de la main) par quantification vectorielle sur l'ensemble d'apprentissage. On ne conserve alors que 128 séries de paramètres ; on calcule ainsi à partir de celui-ci et du modèle de forme BD les coordonnées 3D des 50 points placés sur la main ;

2. ces postures BD sont envoyées à Attitude Studio où sont générées les postures HD correspondantes par la méthode précédemment décrite ;
3. une régression linéaire entre les paramètres articulatoires et les coordonnées 3D des postures HD est calculée, ce qui nous donne le modèle HD de la main.

A cet instant, nous sommes en possession de deux modèles de forme HD qui sont pilotés avec les paramètres articulatoires et de roto-translation déduits des corpora dynamiques.

11.2 Modèle d'apparence

Ayant un modèle de forme HD, il nous est maintenant possible d'ajouter un modèle d'apparence à nos objets. Dans les deux cas, nous utilisons une méthode de plaquage de texture sur le maillage 3D.

11.2.1 Le visage

En utilisant les différentes vues du corpus **tête en rotation** pour un même visème et une technique de projection-inverse [2], on crée une texture cylindrique du visage de notre codeuse (représentation sur la figure 11.7 en haut). Même si pour ce corpus, le nombre de billes et de pastilles est beaucoup moins important que pour le corpus **visage+billes**, il est nécessaire d'effacer ou plutôt de maquiller manuellement les traces laissées par ces marqueurs (représentation sur la figure 11.7 en bas).

FIG. 11.7 – Textures cylindriques de notre codeuse pour le visème [i] : en haut, texture directement obtenue après projection-inverse des 16 vues ; en bas, retouche manuelle pour effacer les billes et pastilles collées sur le visage.

11.2.2 Les dents

Des prises de vue haute résolution des dents (voir figure 11.8 (a)) permettent d'habiller le modèle 3D de dents calculé à partir du moulage des dents.

(a) (b)

FIG. 11.8 – Images utilisées comme textures pour les dents (a) et pour la main (b).

11.2.3 La main

Pour ce qui est de la main, c'est Attitude Studio qui nous a fourni, avec les coordonnées 3D du maillage pour les 128 configurations, une image (voir figure 11.8 (b)) ainsi qu'une table de conversion (image vers maillage).

11.3 Résumé

Le chapitre précédent décrivait la mise en oeuvre de la pièce clé de notre système de synthèse, le modèle de contrôle. Toutefois s'il n'est pas associé à des modèles de forme et d'apparence performants, la qualité globale du système sera dégradée.

C'est donc de ces deux modules qui viennent se greffer au bout de la chaîne dont il est question dans ce chapitre. L'analyse des données dynamiques nous avait fourni un modèle de forme BD inadapté avec un modèle d'apparence. Nous avons donc construit un modèle de forme HD pour chacun des objets : le visage et la main suivant des méthodologies différentes. Dans le premier cas, nous nous sommes basés sur une série de photographies de plusieurs visèmes et dans le second cas, nous nous sommes basés sur le moulage de la main. Après avoir fait concorder nos données sur un ensemble pertinent (l'erreur engendrée sera traitée dans la partie suivante), nous avons créé des modèles HD. Cette densité de points importante dans ces modèles nous a permis de rajouter des modèles d'apparence basés sur le plaquage de texture. Nous avons finalement en bout de chaîne un clone vidéo-réaliste capable de synthétiser du code LPC à partir de n'importe quel nouveau texte (comme représenté sur la figure 11.9).

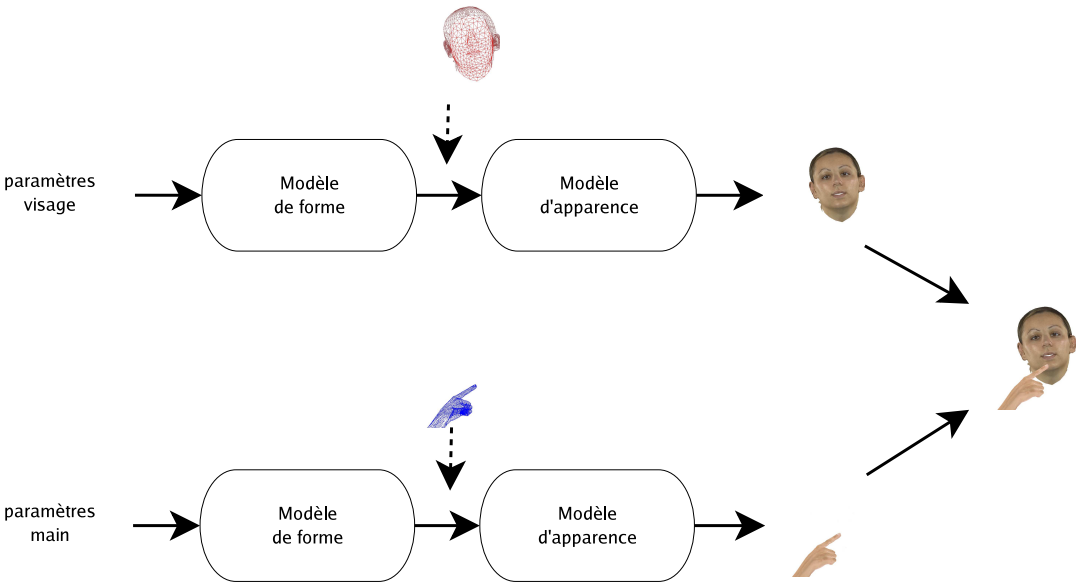


FIG. 11.9 – Passage des paramètres articulatoires et de roto-translation délivrés par le système de synthèse par concaténation à un visage et une main vidéoréalistes.

Références bibliographiques

[1] M. Bérar, G. Bailly, M. Chabanas, M. Desvignes, F. Elisei, M. Odisio, and Y. Pahan. *Towards a better understanding of speech production processes*, chapter Towards a generic talking head, pages 341–362. Psychology Press, New York, 2006.

[2] M. Odisio and F. Elisei. Clonage 3D et animation articulatoire du visage d'une personne réelle pour la communication parlée audiovisuelle. In *Journées de l'AFIG*, pages 225–232, Grenoble, France, 2000.

[3] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[4] L. Revéret and C. Benoît. A new 3D lip model for analysis and synthesis of lip motion in speech production. In *Auditory-Visual Speech Processing Workshop*, 1998.

Chapitre 12

Résumé de la partie

Après avoir défini notre stratégie face aux verrous technologiques qui se présentaient à nous, nous décrivons la phase d'enregistrement puis de pré-traitement des données. Ces deux phases préliminaires mènent bien évidemment à la phase d'analyse des données car bien que des études aient déjà été faites sur la production de la Langue française Parlée Complétée, nos données sont de nature différente et nous tenions à être sûr de ce qu'elles contenaient. Les conclusions des différentes analyses effectuées sur les divers corpora nous donnent des voies de réflexion pour résoudre les verrous technologiques.

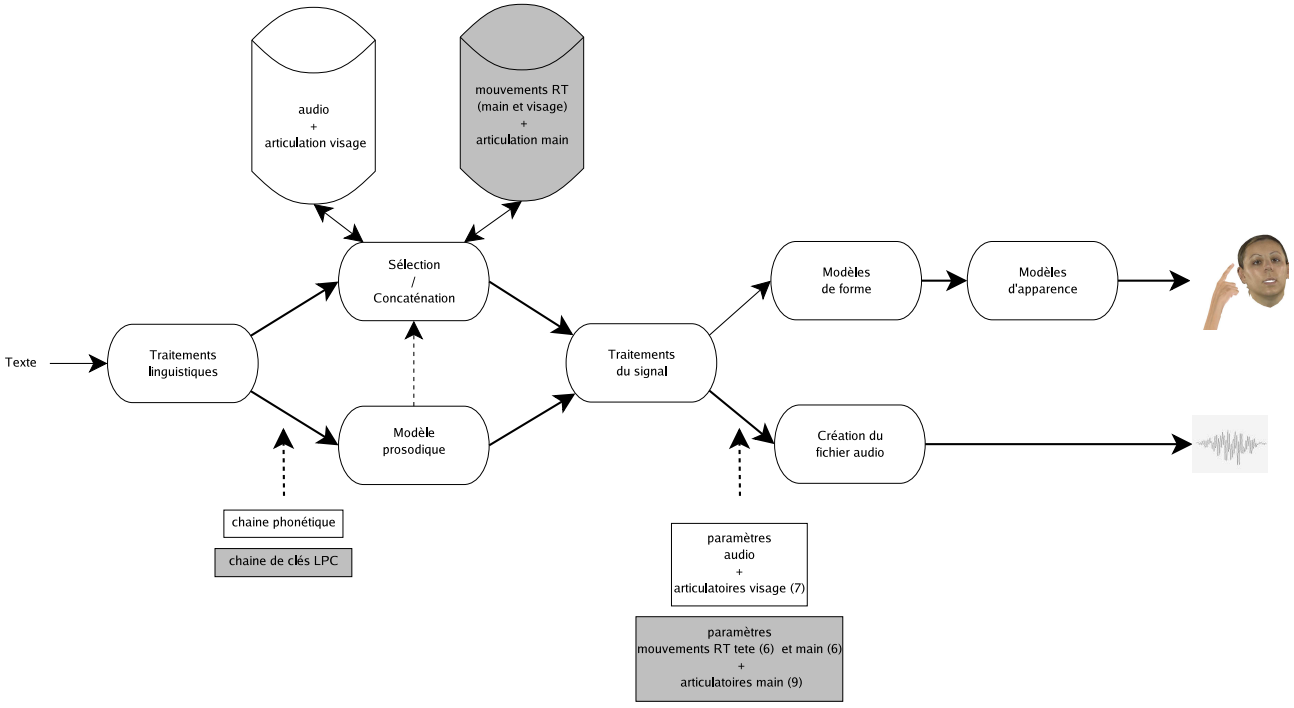


FIG. 12.1 – Diagramme du système de synthèse de Langue française Parlée Complétée à partir du texte.

Nous avons opté pour une solution qui consiste à synthétiser en 2 temps les différents paramètres pour résoudre les contraintes temporelles des différents articulateurs.

Cependant, les premiers corpora enregistrés ne sont pas suffisants pour générer une géométrie

haute définition de la main et du visage. De nouveaux corpora sont donc utilisés pour compléter nos données et pouvoir ainsi générer des objets 3D vidéo-réalistes.

Finalement, le diagramme complet de notre système de synthèse est celui représenté sur la figure 12.1.

Dans la partie suivante, nous allons nous attacher à évaluer notre système de synthèse tant de façon objective que subjective.

Troisième partie

Sans oublier l'évaluation

Chapitre 13

Evaluer, oui mais...

13.1 Pourquoi évalue-t-on ?

Un système est en général créé dans le but d'atteindre un objectif précis, de répondre à un besoin. En ce sens il faut vérifier, une fois la réalisation du système achevée, qu'il répond bien aux attentes de la tâche qui lui incombe. Il est donc primordial de définir avec soin et précision, dès le départ *i.e.* avant même la réalisation, les contraintes liées à l'application finale. Ainsi, à partir du cahier des charges à remplir, nous pouvons déterminer les solutions susceptibles de résoudre les verrous inhérents à la tâche. Dans le même temps, nous sommes capables de déterminer et de proposer un ensemble de tests permettant l'évaluation effective du système.

Le système, que nous proposons, a été créé dans le but d'aider la communication des personnes sourdes et malentendantes. Si nous nous référons au projet ARTUS plus particulièrement, il s'agit de substituer un clone 3D codant LPC à la place du sous-titrage télétexte. Cette application va nous donner des pistes de réflexion quant aux caractéristiques que doit posséder notre système. En effet, il doit être capable de remplacer le télétexte *i.e.* posséder au moins tous ses avantages voire plus, sans cumuler d'autres inconvénients. Les avantages du télétexte sont qu'il est intelligible et qu'il ne demande qu'une infrastructure légère (rappelons que le sous-titrage est encodée dans les images transmises). D'un autre côté, il y a des inconvénients à l'utilisation du sous-titrage télétexte : les personnes ne sachant pas lire ne peuvent l'utiliser (les jeunes enfants par exemple) et la charge cognitve nécessaire à son utilisation n'a jamais été mesurée.

Les points importants sur lesquels notre système de synthèse doit être performants sont donc l'intelligibilité et la charge cognitive associée, ainsi que la légèreté de l'infrastructure. Nous ne devons pas oublier que ce système peut être utilisé sur d'autres applications où des contraintes supplémentaires pourraient s'ajouter. Nous citerons, par exemple, les applications de téléphonie où le clone pourrait se substituer à la parole et accomplir ainsi sa tâche d'aide à la communication. Des contraintes nouvelles apparaissent dans ce genre d'applications : les temps de réponse sont contraints et il faut donc que le système soit suffisament optimisé pour que ses temps de réponse soient en adéquation avec ceux de l'application.

13.2 Qu'évalue-t-on ?

Nous avons vu dans la section précédente que pour pouvoir évaluer efficacement un système il fallait définir un cadre précis des buts à atteindre. Cependant, même si la définition de la tâche finale est précise et exhaustive, le choix des critères reste un problème fort complexe. En effet, ll n'est pas évident de déterminer quel est l'ensemble des critères nécessaires et suffisants à l'évaluation du système.

Nous allons décomposer les critères d'évaluation en deux types : les critères objectifs et les critères subjectifs. Les premiers peuvent être estimés sur le système seul alors que les seconds nécessitent une expertise humaine. Les critères objectifs même s'ils fournissent une évaluation pertinente du système ne sont pas suffisants pour émettre un avis tranché sur la qualité d'un système. En effet, n'oublions pas que le système accomplit sa tâche auprès d'humains et que c'est leur expertise qui finalisera l'évaluation. Il est donc nécessaire d'effectuer les deux types d'évaluation pour avoir une idée précise des points faibles et forts du système.

Dans l'objectif d'évaluer notre système, nous allons nous préoccuper plus particulièrement des critères suivants :

- le temps de réponse du système ;
- l'attractivité du système ;
- l'intelligibilité et la charge cognitive associée.

Le premier critère se place dans la catégorie objectives contrairement aux deux autres qui se situe dans la catégorie subjective.

13.2.1 Temps de réponse du système

Dans l'application visée (la substitution du sous-titrage télétexte par un clone 3D codant le LPC), le temps de réponse du système n'est pas un critère primordial. Toutefois, le champ d'application de ce système ne se restreint pas à cette application. Il pourrait très bien être utilisé dans des applications de téléphonie ou de traduction en ligne qui sont des tâches imposant des contraintes temporelles fortes.

Cela nous amène donc à nous poser la question suivante : «est-ce que ce système[1] fonctionne ou est capable de fonctionner en temps-réel ?». Pour y répondre, nous devons revenir à la définition de l'expression *temps-réel*. Si nous prenons la définition de *temps-réel* suivante : «La correction d'un système ne dépend pas seulement des résultats logiques des traitements, mais dépend en plus de la date à laquelle ces résultats sont produits» [4] comme le suggère [2], alors les contraintes temporelles à respecter sont relatives à un temps physique mesurable et font partie de la spécification du système. En d'autres termes, il ne sert à rien que le système ait une rapidité d'exécution moyenne élevée s'il n'arrive pas à respecter ne serait-ce qu'une seule contrainte temporelle (le pire cas). Pour aller plus loin dans les définitions, nous pouvons noter que les systèmes temps-réels sont séparés en deux classes : les systèmes *temps-réel stricts* et les systèmes *temps-réel souples*. Alors que les premiers respectent à la lettre la définition de système temps-réel et qu'ils ont donc été conçus avec une connaissance *a priori* de tous les

[1] la notion de système correspond à l'implémentation logicielle de synthétiseur.

scénarii d'exécution possibles, les seconds peuvent avoir un taux de non respect des contraintes temporelles «acceptables» (ces systèmes acceptent des variations dans le traitement des données de l'ordre de 500 ms dans le cadre des systèmes multimédias).

Qu'en est-il de notre système ? La synthèse de parole, qu'elle soit audio ou audiovisuelle fait intervenir un module prosodique qui a un rôle essentiel dans la qualité du signal de sortie. Or, ce module a besoin d'une structure de phrase pour pouvoir prédire les variations de fréquence fondamentale et de longueurs d'unités à concaténer ainsi que d'énergie. Notre système a donc un fonctionnement «phrase par phrase». Ce mode de fonctionnement peut sembler *a priori* un frein à la vitesse d'éxecution du système. Par exemple, si l'on synthétise une phrase d'une durée de 4s, le temps de réponse du système correspond à environ 80% de la durée de la phrase *i.e.* environ 3s. Ainsi, dans des applications «half-duplex»[2] (deux interlocuteurs ne se coupant pas la parole par exemple), il serait envisageable de faire du «temps-réel» souple (le système «intelligent» générant le dialogue en fonction de celui perçu reste l'aspect limitant). En revanche, si nous visons des applications plus contraignantes temporellement, nous pourrions envisager un autre type de solution qui consisterait à générer de la parole de synthèse à partir de quelques mots et non plus de phrases complètes, la prosodie serait dégradée et la qualité du signal de sortie en pâtirait.

13.2.2 Attractivité

Ce critère est plus important que le précédent dans l'application qui nous intéresse. En effet, un visage parlant de synthèse reste une interface homme-machine (IHM). En ce sens, il ne faut pas négliger l'attractivité d'un tel système. Si l'on imagine une interface homme-machine fonctionnant parfaitement (en termes de décision) mais désagréable à l'usage, alors les utilisateurs préféreront peut-être ne plus l'utiliser. Dans notre cas précis, il ne faut pas oublier qu'une autre interface IHM existe déjà et est utilisée : le sous-titrage télétexte. Ce type de critère d'évaluation est subjectif et qualitatif c'est-à-dire qu'il ne peut passer que par des expériences de perception avec des tests de type MOS (Mean Opinion Score) [1].

Nous n'effectuerons pas explicitement un test visant à quantifier l'attractivité de notre système par rapport à d'autres. Toutefois, nous tiendrons compte des commentaires des sujets passant le test d'intelligibilité par rapport à leurs attentes en terme d'attractivité et de l'utilisation d'un tel système dans leur vie courante.

13.2.3 Efficacité : ratio Intelligibilité-Compréhension/Ressources cognitives

Ce critère est sans conteste le plus important aux vues de notre objectif. Nous allons donc évaluer l'efficacité *i.e.* l'intelligibilité du signal délivré en sortie de notre système et la charge cognitive nécessaire à celle-ci. Les visages parlants de synthèse demandent plus de charge cognitive que la parole naturelle [3]. Il est important de regarder ces deux critères en même temps.

Dans un premier temps, nous allons évaluer la qualité segmentale de nos signaux de synthèse *i.e.* la compatibilité segmentale des mouvements du visage et de la main. Ce critère englobe à

[2]Transmission à deux sens dans un même canal, mais qui ne s'effectue que dans un sens en même temps.

la fois la qualité de nos approximations heuristiques temporelles de coordination main/visage et la qualité d'animation de la main.

Dans un deuxième temps, si la qualité segmentale est satisfaisante, nous devrons passer au critère «capacité à comprendre un discours». Alors que la qualité segmentale peut se mesurer sur des mots, la qualité de compréhension devra s'effectuer au niveau de la phrase ou du paragraphe.

13.3 Comment évalue-t-on ?

Dans le cadre de ces travaux, nous allons supposer que la chaîne phonétique marquée en temps est parfaite. Nous n'évaluerons pas le module «Traitements linguistiques» ni le module «prosodique». Nous allons évaluer le module de «contrôle moteur» *i.e.* le module qui à partir d'une chaîne phonétique marquée en temps génère, les paramètres d'animation du visage et de la main.

Nous allons effectuer cela en deux étapes : une évaluation objective puis une évaluation subjective. Lors de l'évaluation objective, nous quantifierons l'erreur moyenne 3D commise par le système en synthétisant les phrases du corpus d'apprentissage et nous proposerons un autre critère grâce au système de reconnaissance développé dans le chapitre «Analyse des données». Lors de la phase d'évaluation subjective, nous quantifierons dans un premier temps la qualité segmentale de notre système puis, si les résultats sont probants, nous pourrons déterminer la qualité de compréhension.

Plutôt qu'une note globale que l'on donnerait au système, il semble plus logique de représenter les résultats suivant un tableau de synthèse. Cette méthode nous évite de déterminer de manière trop subjective des poids à appliquer à chaque critère et donne un résultat plus lisble.

13.4 Résumé

Dans ce chapitre, introductif à la partie **Sans oublier l'évaluation**, nous avons expliqué et justifié pourquoi cette partie est nécessaire. Pour cela, nous avons répondu aux questions «Pourquoi évalue-t-on ?», «Qu'évalue-t-on ?» et «Comment évalue-t-on ?». Il en est ressorti que nous devions évaluer notre système par rapport à la tâche pour laquelle il a été créé. Par conséquent nous devons évaluer en priorité l'intelligibilité et la compréhension des signaux de sortie délivrés par le système sans oublier la charge cognitive correspondante.

Les deux prochains chapitres vont présenter les résultats des évaluations objectives et subjectives de notre synthétiseur.

Références bibliographiques

[1] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

[2] D. Decotigny. *Une infrastructure de simulation modulaire pour l'évaluation de performances de systèmes temps-réel.* PhD thesis, Université de Rennes 1, 2003.

[3] I. Pandzig, J. Ostermann, and D. Millen. Users evaluation : synthetic talking faces for interactive services. *The Visual Computer*, 15 :330–340, 1999.

[4] J. A. Stankovic. Misconceptions about real-time computing. *IEEE Computer*, 21(10) :10–19, 1988.

Chapitre 14

Evaluations objectives

Afin d'évaluer objectivement notre système, nous définissons une référence par rapport à laquelle nous comparons nos signaux de synthèse. Nous utilisons, pour cela, le corpus dynamique **main + visage** composé de 238 phrases comme référence.

Dans un permier temps, nous imposons au synthétiseur les contraintes temporelles et fréquentielles des stimuli de référence. Nous synthétisons de deux manières différentes : un cas parfait où tous les polysons du dictionnaire sont disponibles et un cas un peu plus réaliste où les polysons du stimulus d'origine sont écartés du dictionnaire. Nous calculons les erreurs RMS (Root Mean Square) au niveau des positions 3D des points du visage et de ceux de la main entre les stimuli référence et leurs équivalents de synthèse. Nous déterminons également l'erreur RMS commise sur les paramètres d'animation entre les stimuli références et leurs équivalents de synthèse. Dans un deuxième temps, nous n'imposons au synthétiseur que l'entrée textuelle du stimulus : cas le plus réaliste. Dans ce cas, une erreur 3D RMS n'a plus beaucoup de sens puisque la partition temporelle du mouvement est différente entre les deux stimuli. Nous utilisons un paradigme original qui consiste à comparer les taux de reconnaissance d'un système automatique pour les deux types de stimuli (originaux et synthétiques).

14.1 Synthèse contrainte

14.1.1 Cas parfait

Ce cas, le plus favorable, consiste à comparer les stimuli enregistrés et les stimuli synthétisés avec un dictionnaire complet de polysons et de diclés. *A priori*, le signal de synthèse doit être identique au signal d'origine ou tout du moins avoir subi des distorsions mineures. En effet, nous nous donnons tous les moyens pour que ce signal de sortie soit «parfait» en imposant la segmentation temporelle et la présence des polysons et diclés d'origine. Nous étudions quel est l'effet de la synthèse dans ce cas très favorable sur les paramètres d'animation et sur la géométrie du visage et de la main.

Erreur RMS des paramètres d'animation

Les paramètres d'animation ne sont pas des paramètres visibles ni interprétables en tant que tels. Cependant, ils sont à la base de l'animation des objets 3D que sont le visage et la main. Par conséquent, une erreur importante sur l'un de ces paramètres peut induire une grande erreur géométrique voire conduire à une forme de visage et/ou de main erronée. L'erreur RMS sur ces paramètres (articulatoires et de mouvements) est calculée et permet de diagnostiquer les stimuli qui présentent des erreurs importantes.

La figure 14.1 représente l'erreur en terme de paramètres d'animation pour toutes les phrases du corpus synthétisées. Nous remarquons que cette erreur est stable de l'ordre de 0.5 unités (de paramètres) pour le visage et pour la main hormis pour un nombre limité de phrases. Ce critère est un outil nous indiquant les stimuli à vérifier en priorité *i.e.* les phrases dont l'erreur est supérieure à l'erreur moyenne.

Après vérification, l'erreur plus importante sur certaines phrases nous a permis de mettre en évidence un problème de sélection d'unités. La sélection étant erronnée, les unités choisies ne provenaient pas des phrases d'origine d'où une erreur plus importante. Il s'agit d'un très bon moyen de diagnostiquer de possibles erreurs.

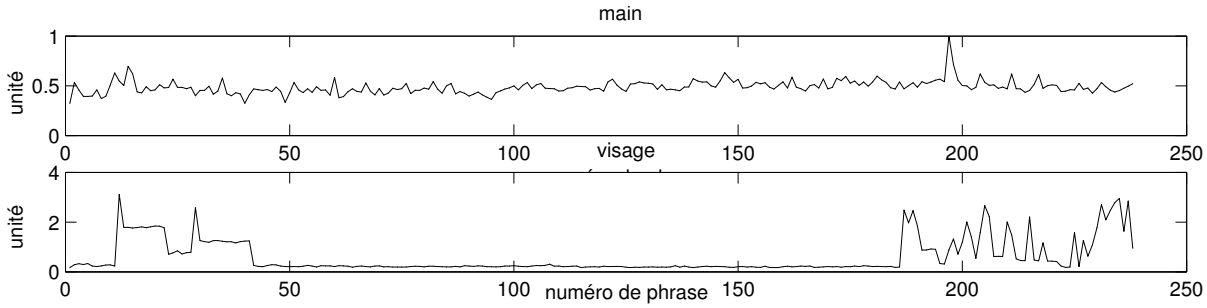


FIG. 14.1 – Erreur RMS (moyenne) des paramètres d'animation pour toutes les phrases du corpus dans le cas parfait. (N.B. : les paramètres d'animation sont des paramètres sans unité)

Erreur RMS des positions 3D des points de peau

Contrairement aux paramètres d'animation, les points de peau sont visibles. Les mouvements de ceux-ci ont un impact important sur la qualité du système de synthèse. Par conséquent, l'erreur RMS des positions 3D des points de chair est un critère essentiel de la qualité de synthèse même s'il peut être difficile de l'interpréter. La figure 14.2 représente l'erreur RMS des positions des points 3D pour les phrases synthétisées par rapport aux phrases originales. Nous remarquons que l'erreur est de l'ordre de 5 mm pour la main et inférieure au mm pour le visage. Nous retrouvons des erreurs plus importantes sur certaines phrases comme lors de l'étude de l'erreur sur les paramètres d'animation. Cette représentation est comme précédemment un outil pratique pour la vérification des stimuli présentant une erreur plus importante que la moyenne. La relation entre les paramètres articulatoires du visage et la géométrie 3D de celui-ci étant linéaire, les deux critères sont identiques.

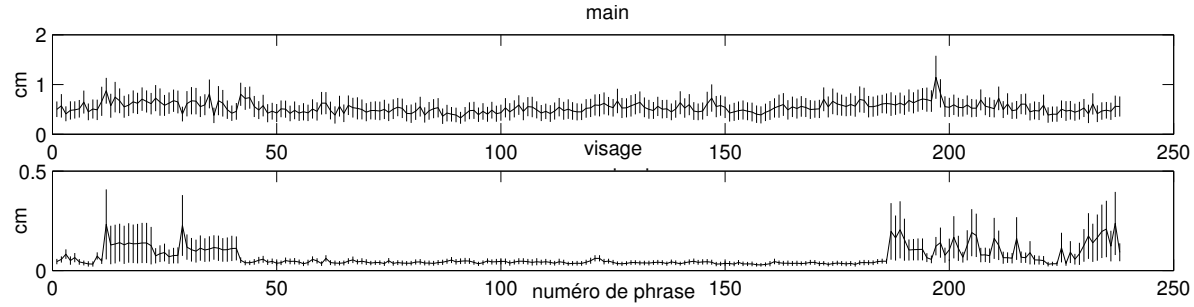


FIG. 14.2 – Erreur RMS (moyenne et écart-type) des positions 3D des points du visage et de la main pour toutes les phrases du corpus dans le cas parfait.

14.1.2 Cas réel

Ce cas correspond à la comparaison des phrases enregistrées avec les phrases synthétiques sans les diphones de la phrase correspondante. Nous nous attendons à ce que les erreurs, tant en termes de paramètres d'animation que de positions 3D des points de chair, soient d'un ordre de grandeur plus grand que précédemment. En effet, ce cas est moins favorable puisque les polysons et diclés utilisés sont forcément différents de ceux d'origine.

Erreur RMS paramètres d'animation

La figure 14.3 représente l'erreur en terme de paramètres d'animation pour toutes les phrases du corpus synthétisées. Nous remarquons que cette erreur est stable de l'ordre de 2 unités (de paramètres) pour le visage et pour la main. Cette valeur moyenne est nettement supérieure à celle obtenue dans le cas parfait et pourrait conduire à des formes du visage et de la main erronées. Il est toutefois difficile d'interpréter ce résultat sans informations supplémentaires. S'il intervenait dans le cadre d'une comparaison de systèmes, il serait un critère d'évaluation.

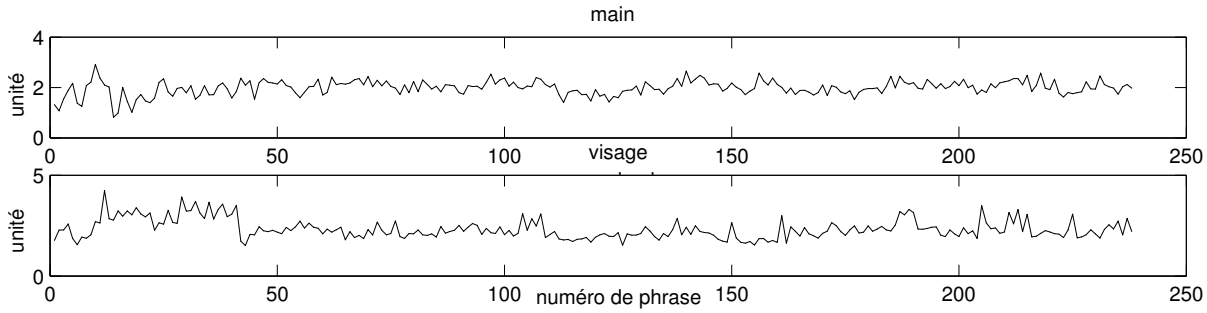


FIG. 14.3 – Erreur RMS (moyenne) des paramètres d'animation pour toutes les phrases du corpus dans le cas réel. (N.B. : les paramètres d'animation sont des paramètres sans unité)

Erreur 3D RMS

La figure 14.4 représente l'erreur RMS des positions des points 3D pour les phrases synthétisées par rapport aux phrases originales. Nous remarquons que l'erreur est de l'ordre de 2 cm pour la main et de 5 mm pour le visage. Nous retrouvons des erreurs d'un ordre de grandeur supérieur

par rapport à la synthèse «parfaite». Comme pour l'erreur RMS sur les paramètres d'animation, ce résultat n'est pas facilement interprétable. En revanche, dans le cas d'une comparaison de différents synthétiseurs, il s'agirait d'un critère pertinent. On remarque toutefois que l'erreur est plus stable que précédemment. Le choix des unités étant contraints à des unités forcément différentes, il est logique que l'erreur soit plus importante et répartie.

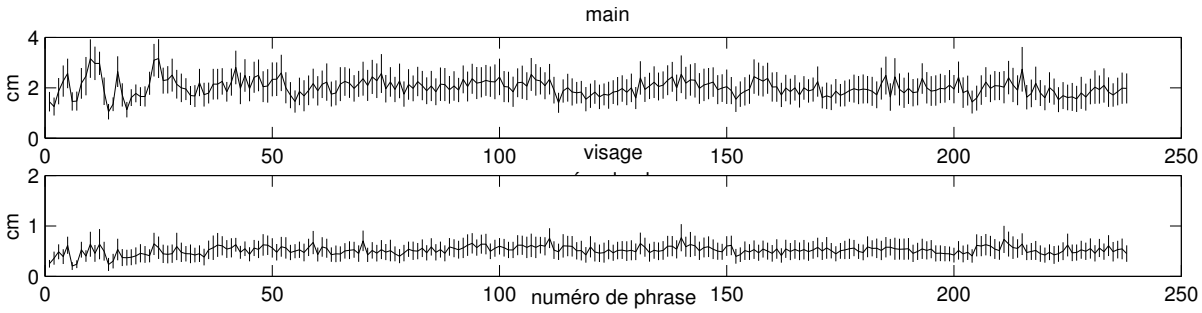


FIG. 14.4 – Erreur RMS (moyenne et écart-type) des positions 3D des points du visage et de la main pour toutes les phrases du corpus dans le cas réel.

14.2 Synthèse libre

Dans le cas d'une synthèse «libre»[1], il est difficile d'appliquer les outils précedemment définis. En effet, il n'est plus possible de faire une comparaison trame à trame que ce soit en terme de position 3D des points ou des valeurs des paramètres d'animation. Pour résoudre ce problème, nous proposons d'utiliser le système de reconnaissance des mouvements de la main défini dans le chapitre **Vérification des données : Code LPC/chaîne phonétique** et de créer son pendant pour le visage.

14.2.1 Système de reconnaissance des formes de lèvres

Le système de reconnaissance des formes labiales se base sur les valeurs des paramètres A (écartement des lèvres), B (séparation des lèvres), S (aire intérolabiale) et L (profondeur du pavillon) [1]. Ces paramètres ont déjà été utlisés dans le cadre de l'étude de Robert-Ribes et al. [6] sur la complémentarité et la synergie dans la parole multimodale. Le paramètre A permet entre autres de bien séparer les voyelles arrondies des voyelles non arrondies et on peut noter qu'il existe une forte corrélation entre les paramètres B et S [6] et entre AxB et S [1, 2]. Nous allons apprendre des modèles gaussiens pour chacune des classes phonétiques à partir des données originales et comparer les résultats du système de reconnaissance sur les stimuli originaux et les stimuli de synthèse.

Consonnes

Ce système fonctionne sur le même principe que celui utilisé pour la reconnaissance des clés et des positions de main. La principale différence réside dans le choix des paramètres d'apprentis-

[1] aucune contrainte hormis le texte d'entrée.

sage. Ainsi, pour estimer les modèles gaussiens de chaque consonne, nous avons utilisé les valeurs des paramètres labiaux (précédemment cités) pour la trame centrale de chaque phonème. Afin d'optimiser le taux de reconnaissance, nous avons regroupé les différents sosies labiaux et créé un modèle gaussien par groupe.

	p b m	t d n	k g ŋ	ʒ ʃ	v f	z s	ʁ l	j w ɥ	Total
p b m	307	23	3	29	88	7	9	2	468
t d n	187	199	44	28	102	72	14	9	655
k g ŋ	67	44	86	5	3	24	39	11	279
ʒ ʃ	65	6	2	118	0	10	3	0	204
v f	20	4	0	4	179	5	0	0	212
z s	81	39	10	22	24	144	12	16	348
ʁ l	128	206	116	18	18	25	216	30	757
j w ɥ	60	19	25	7	3	28	22	12	176

TAB. 14.1 – Matrice de confusion du système de reconnaissance basé sur les paramètres labiaux sur tous les groupes de consonnes (sosies labiaux) du corpus **main + visage**.

Le taux de reconnaissance sur les phrases originales est de 40.69%. La matrice de confusion (représentée dans le tableau 14.1) révèle que certains groupes de consonnes sont très mal discriminés. On retrouve cette conclusion en se référant à la figure 14.5 qui représente les ellipses de dispersion sur l'espace des deux premiers paramètres de l'ACP calculée à partir des paramètres labiaux. Par exemple, les consonnes du groupe [t d n] sont reconnues dans la moitié des cas comme des consonnes du groupe [p b m]. En revanche, les consonnes du groupe [p b m] sont bien discriminées. En effet, lorsque l'on produit des bilabiales il y a fermeture complète des lèvres : c'est une caractéristique visuellement importante.

Voyelles

De la même façon que pour les consonnes, nous avons implémenté un système de reconnaissance pour les voyelles. Afin d'estimer les modèles gaussiens de chaque voyelle, nous avons utilisé les valeurs des paramètres labiaux (précédemment cités) pour la trame centrale de chaque phonème. Afin de maximiser le taux de reconnaissance, nous avons regroupé les différents sosies labiaux et créé un modèle gaussien par groupe.

Le taux de reconnaissance sur les phrases originales est de 46.11%. La matrice de confusion (représentée dans la table 14.2) révèle que certains groupes de voyelles sont très mal discriminés. On retrouve cette conclusion en se référant à la figure 14.6 qui représente les ellipses de dispersion sur l'espace des deux premiers paramètres de l'ACP calculée à partir des paramètres labiaux. Par exemple, les voyelles du groupe [ɔ̃ ɔ o] sont reconnues dans la moitié des cas comme les voyelles du groupe [u y]. En revanche, les voyelles du groupe [u y] sont très bien discriminées. En effet, lorsque l'on produit ces voyelles, on effectue une protrusion qui est bien capturée par le paramètre L (la profondeur du pavillon) calculé par le système de reconnaissance.

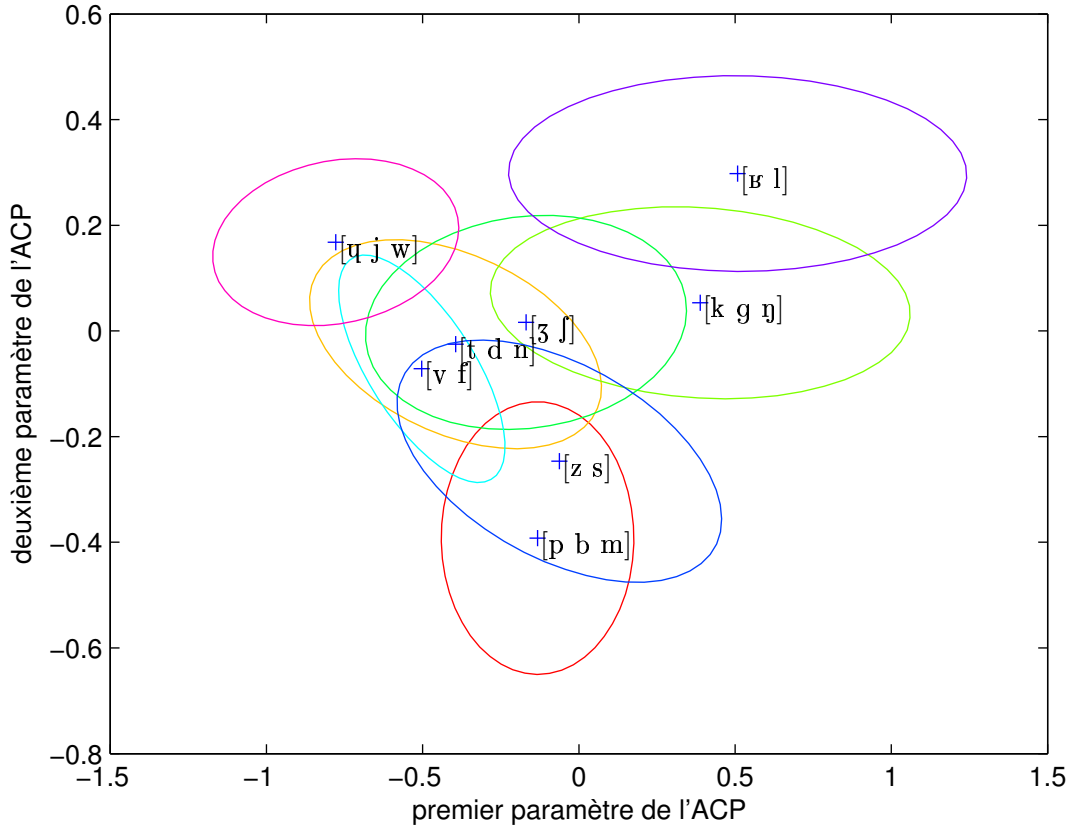


FIG. 14.5 – Ellipses de dispersion pour les groupes de sosies labiaux pour les consonnes dans l'espace des deux premières composantes de l'ACP effectuée sur les paramètres labiaux.

14.2.2 Erreur probabiliste

Bien que les taux de reconnaissance pour les phrases originales ne soient pas très élevés, nous pouvons les comparer à ceux délivrés par le système dans le cas des phrases de synthèse. Nous allons faire cette comparaison pour les consonnes, les voyelles mais également pour les formes et positions de la main.

Consonnes

Le taux de reconnaissance sur les phrases synthétisées est de 37.42% c'est-à-dire inférieur à celui des phrases originales (pour rappel : 40.69%). Nous notons toutefois qu'il s'agit du même ordre de grandeur. Les conclusions de ces résultats doivent être nuancées, le système de reconnaissance n'ayant pas des taux élevés sur les phrases originales.

Un autre moyen de mettre en relief les ambiguïtés visuelles est de créer un arbre de classification basé sur une distance entre modèles gaussiens. Afin de déterminer cette classification, nous avons calculé la distance de Bhattacharyya (eq. 14.1) qui est une mesure de similarité entre deux distributions gaussiennes pour toutes les classes phonétiques. On peut visualiser cette classification sous forme d'un dendrogramme (voir la figure 14.7 (a) pour les stimuli originaux et la

	a	ã	ɛ̃ ɛ e	õe	ɔ̃ ɔ o	œ õe	u y	i	Total
a	168	2	69	166	1	5	0	54	465
ã	2	109	0	0	24	10	17	1	163
ɛ̃ ɛ e	67	3	113	92	0	5	0	45	325
õe	6	1	2	72	0	3	0	6	90
ɔ̃ ɔ o	0	50	0	7	48	55	103	3	266
œ õe	6	19	0	21	42	90	178	6	362
u y	0	1	0	1	18	23	281	2	326
i	34	11	23	34	4	24	8	192	330

TAB. 14.2 – Matrice de confusion du système de reconnaissance basé sur les paramètres labiaux sur tous les groupes de consonnes (sosies labiaux) des phrases du corpus **main + visage**.

figure 14.7 (b) pour les stimuli de synthèse). Cette représentation et cette distance sont souvent utilisés pour séparer les classes phonétiques tant au niveau acoustique [5] qu'au niveau visuel [3, 4].

$$d_{Bhattacharyya} = \frac{1}{8}(\mu_1 - \mu_2)^t(\frac{\Sigma_1 + \Sigma_2}{2})^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\frac{|(\Sigma_1 + \Sigma_2)/2|}{(|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}})} \qquad (14.1)$$

On remarque que dans le cas des stimuli originaux on retrouve les classes de consonnes habituelles : les bilabiales [p, b, m], les dentales [t, d, n], les vélaires [k, g], les fricatives [v, f] et [s, z], etc. Pour les stimuli de synthèse, certaines classes sont conservées telles que les bilabiales mais d'autres comme les vélaires et les fricatives ont été dégradées.

Voyelles

Le taux de reconnaissance sur les phrases synthétisées est de 44.82% c'est-à-dire inférieur à celui des phrases originales (pour rappel : 46.11%). Nous notons toutefois qu'il s'agit du même ordre de grandeur. Comme précedemment, nous devons nuancer nos conclusions par rapport à ces valeurs de taux de reconnaissance.

De la même manière que précedemment, nous représentons le dendrogramme de la classification calculée sur les modèles gaussiens (voir figure 14.8).

Comme pour les consonnes, on retrouve les classes de voyelles habituelles : les voyelles arrondies *vs* les voyelles non arrondies dans les deux types de stimuli (originaux et synthétiques). Dans la classe des voyelles non arrondies, on retrouve bien la distinction ouverte [a] et fermée [i] ces deux voyelles étant le plus éloigné dans cette partie de l'arbre pour les stimuli originaux. Pour les stimuli de synthèse, il y a une légère dégradation puisque dans la partie de l'arbre correspondant aux voyelles non arrondies, [a] et [i] ne se retrouvent pas les plus éloignées. Ces informations sont intéressantes pour discriminer les voyelles entre elles mais on remarque que l'information sur la nasalité par exemple qui n'est pas fournie par les lèvres manquent pour pouvoir discriminer plus finement les voyelles entre elles. Il en va de même à propos de la position de la langue et donc de la détermination de la caractéristique «antérieure» ou «postérieure» de

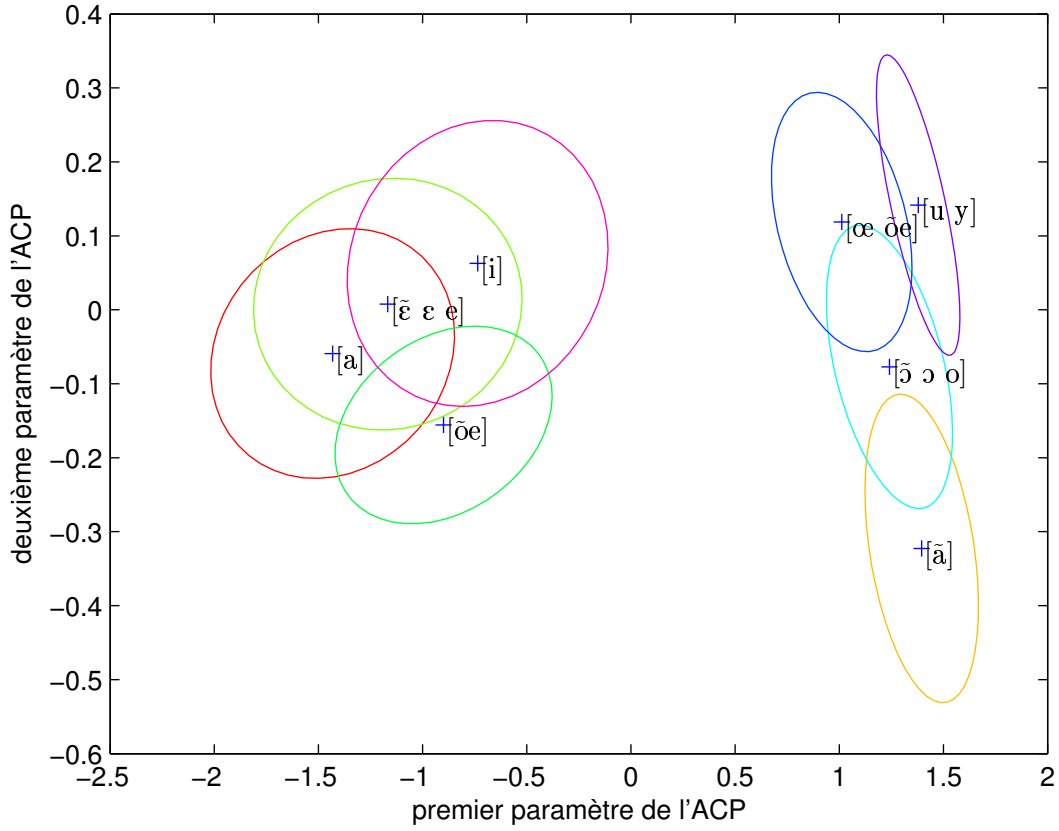


FIG. 14.6 – Ellipses de dispersion pour les groupes de sosies labiaux pour les voyelles dans l'espace des deux premières composantes de l'ACP effectuée sur les paramètres labiaux.

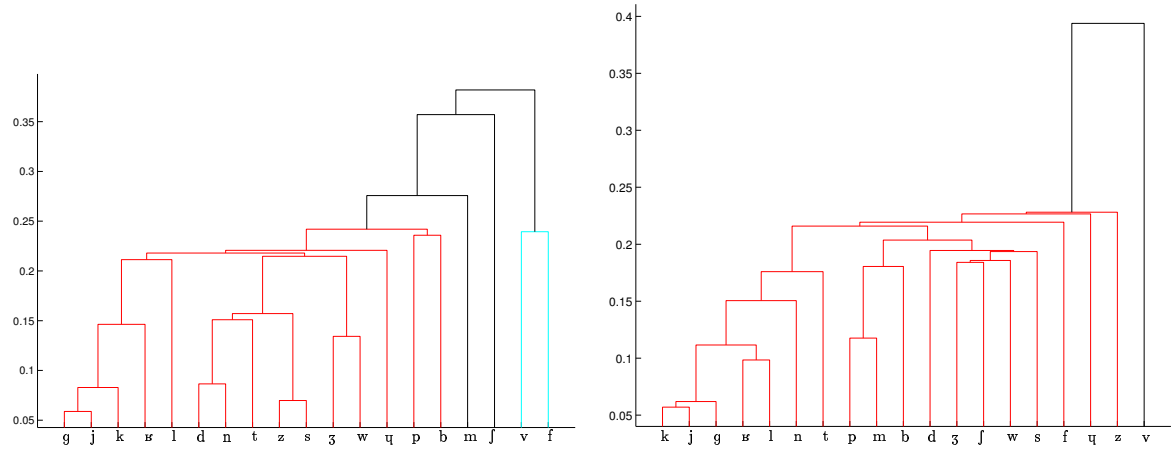
la voyelle.

Forme de main

Le taux de reconnaissance sur les phrases synthétisées est de 49.44% c'est-à-dire très inférieur à celui des phrases originales (pour rappel : 98.78%). Cette différence significative de reconnaissance semble signifier que la forme de main synthétisée est erronée. Après vérification, il apparaît que l'interpolation linéaire anticipatoire, calculée sur les paramètres d'animation lors de la phase de synthèse, impose une avance temporelle.

De la même manière que pour les consonnes et les voyelles, nous déterminons l'arbre de classification pour les formes de main et le représentons sous forme d'un dendrogramme (voir figure 14.9).

Ce qu'il faut noter de prime abord lorsqu'on visualise ce dendrogramme par rapport aux précédents c'est que les distances inter-classes ne sont pas du même ordre de grandeur pour la partie stimuli originaux *vs* stimuli synthétiques. La discrimination est donc plus aisée sur les originaux. On retrouve les formes 2 et 8 très proches en effet un seul paramètre les séparent : l'écartement de l'index et du majeur. Le groupe formé par les formes 1 et 6 se trouve proche du



(a) (b)

FIG. 14.7 – dendrogramme basé sur la distance de Bhattacharyya entre modèles gaussiens de chaque classe phonétique pour les consonnes du français (a) stimuli originaux, (b) stimuli de synthèse.

précédent groupe ce qui semble logique puisque seul la forme du majeur les différencie. Cette caractéristique ne se retrouve pas dans les stimuli de synthèse.

Position de la main

Le taux de reconnaissance sur les phrases synthétisées est de 35.90% c'est-à-dire très inférieur à celui des phrases originales (pour rappel : 98.56%). L'explication donnée pour les formes de main peut s'appliquer aux positions de la main par rapport au visage (*i.e.* l'interpolation linéaire anticipatoire, calculée sur les paramètres d'animation lors de la phase de synthèse, impose une avance temporelle).

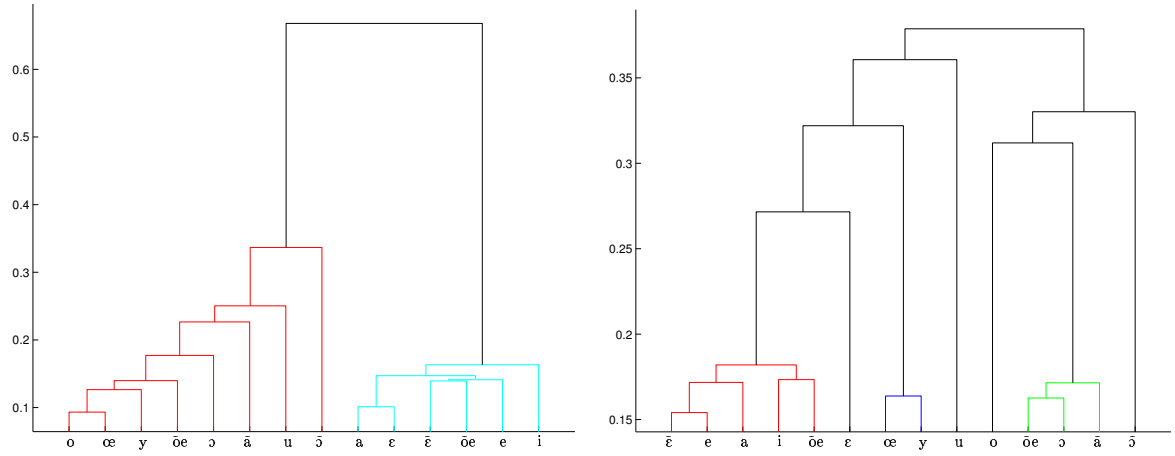
Comme précédemment, nous calculons l'arbre de classification pour les positions de main par rapport au visage et nous le représentons sous forme d'un dendrogramme (voir figure 14.10).

Comme pour les formes de main, la première chose à noter est la différence d'échelle entre les distances inter-classes. Il est plus simple de discriminer les stimuli originaux que les stimuli de synthèse. Les classes sont beaucoup plus rapprochées sur les stimuli de synthèse ce qui pose des problèmes de discrimination. Il faut remarquer toutefois que dans les deux types de stimuli, la position «repos» codée 0 est à part.

14.2.3 Synchronisation

Dans le cas d'une synthèse «libre», le module prosodique engendre des erreurs par rapport aux stimuli originaux. Ces erreurs se situent au niveau de la variation de la fréquence fondamentale, de la longueur des unités à concaténer et de la variation de l'énergie du signal.

En outre, dans le cas de la synthèse de code LPC, nous effectuons une synthèse en deux passes. Ainsi, les erreurs commises sur la longueur des unités phonétiques se propagent à la chaîne de clés LPC à concaténer.



(a) (b)

FIG. 14.8 – dendrogramme basé sur la distance de Bhattacharyya entre modèles gaussiens de chaque classe phonétique pour les voyelles du français (a) stimuli originaux, (b) stimuli de synthèse.

C'est au moyen de l'évaluation subjective que nous émettrons un diagnostic sur la qualité de synchronisation des objets en mouvement.

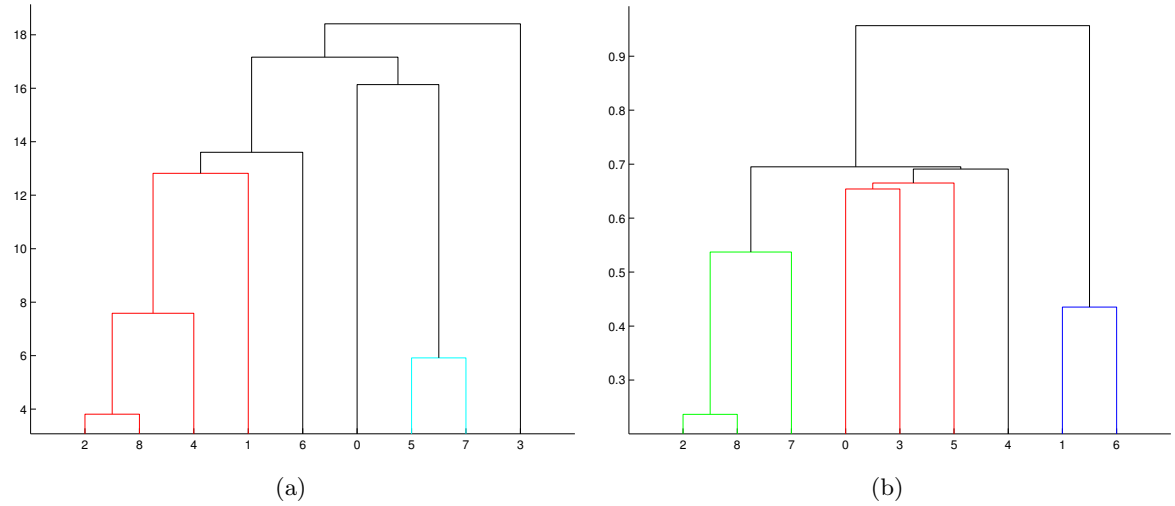
En effet, le choix ayant été fait de choisir une synchronisation moyenne entre les deux effecteurs, on peut en déduire une erreur moyenne mais seule l'évaluation perceptive peut nous signifier la pertinence d'un tel choix. Pour bien faire, il faudrait compléter par une évaluation perceptive supplémentaire où la synchronisation serait modifiée manuellement et graduellement afin de quantifier son effet.

14.3 Résumé

L'evaluation objective nous a amené à quantifier différents types d'erreur en prenant comme référence les phrases du corpus dynamique. Ces erreurs sont de trois types : une erreur RMS sur les paramètres d'animation, une erreur RMS sur la position 3D des points de chair et enfin une erreur de reconnaissance. Ces critères peuvent être utilisés soit comme un diagnostic *i.e.* souligner des erreurs probables sur certains stimuli soit pour comparer plusieurs synthétiseurs. Cependant, comme nous l'avons souligné dans l'introduction de ce chapitre, une analyse quantitative seule n'est pas suffisante à l'établissement d'un diagnostic des points forts et faibles de notre système. Le chapitre suivant sera par conséquent consacré à l'évaluation qualitative en terme d'intelligibilité, compréhension et charge cognitive.

Références bibliographiques

[1] C. Abry and L.-J. Boë. "Laws" for lips. *Speech Communication*, 5 :97–104, 1986.



(a) (b)

FIG. 14.9 – dendrogramme basé sur la distance de Bhattacharyya entre modèles gaussiens de chaque classe de forme de main (a) stimuli originaux, (b) stimuli de synthèse.

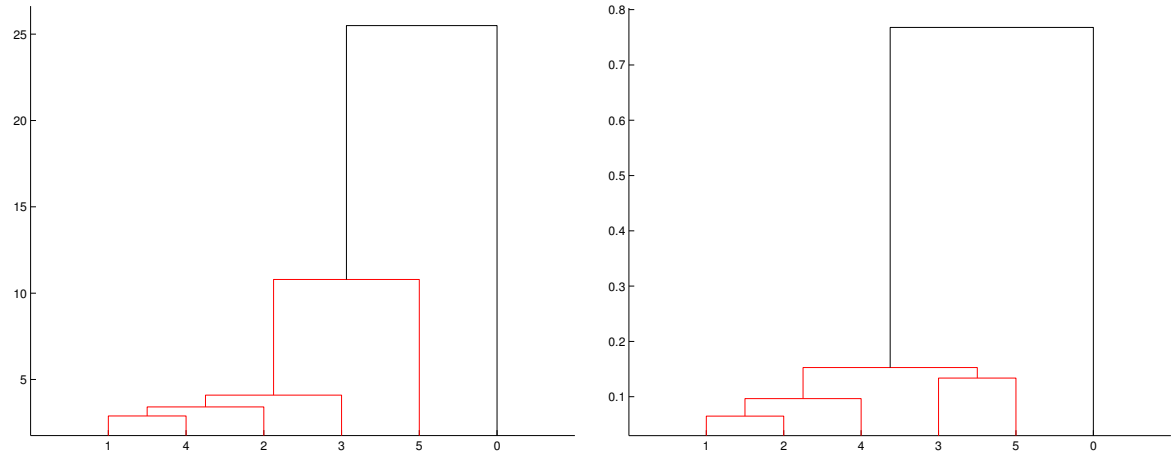
[2] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry. A set of visual French visemes for visual speech synthesis. In G. Bailly, Ch. Benoît, and T. R. Sawallis, editors, *Talking Machines : theories, models and designs*, pages 485–504. Elsevier Science, Amsterdam, NL, 1992.

[3] J. Hazen, T., K. Saenko, C.-H. La, and J. R. Glass. A segment-based audio-visual speech recognizer : Data collection, development and initial experiments. In *Proceedings of the International Conference on Multimodal Interfaces*, State College, Pennsylvania, October 2004.

[4] T. J. Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(3) :1082–1089, May 2006.

[5] B. Mak and E. Barnard. Phone clustering using the Bhattacharyya distance. In *Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 2005–2008, Philadelphia, 1996.

[6] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, and P. Escudier. Complementarity and synergy in bimodal speech : Auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America*, 103(6) :3677–3689, June 1998.



(a) (b)

FIG. 14.10 – dendrogramme basé sur la distance de Bhattacharyya entre modèles gaussiens de chaque classe de position de main (a) stimuli originaux, (b) stimuli de synthèse.

Chapitre 15

Evaluations subjectives : tests de perception

L'évaluation objective de notre système doit être complétée par une analyse subjective basée avant tout sur la quantification de la capacité de notre système à fournir une information linguistique. Comme nous voulons que le test soit adapté à notre problématique de synthèse, nous n'avons pas utilisé les tests à destination des professionnels de la surdité, tels que le test TERMO (Tests d'Evaluation de la Réception du Message Oral) [1] (construit avec un objectif différent). Nous avons par conséquent redéfini un test complet. Ce test se décompose en deux parties : la première vise à mesurer la qualité segmentale de celui-ci tandis que la seconde tend à quantifier la qualité de compréhension globale. Dans le même temps, nous enregistrons les temps de réponse afin de posséder des indices sur la charge cognitive nécessaire à la compréhension des stimuli de synthèse.

15.1 Intelligibilité segmentale

15.1.1 Description du test

Le test que nous proposons est basé sur le test de rime de Fairbanks [2] et plus particulièrement sur son adaptation à la langue française [4].

Le *test de rime* proposé par Fairbanks vise à tester la qualité des synthétiseurs de parole audio en proposant pour chaque stimulus un choix par paires minimales. Dans l'adaptation au français de ce test, on se focalise sur six traits du signal de parole :

1. voisé / non voisé ;
2. nasal / non nasal ;
3. interrompu / non interrompu ;
4. compact / diffus ;
5. grave / aigu ;
6. vocalique / non vocalique.

Nous proposons une adaptation de ce test qui permet de tester les synthétiseurs de code LPC. L'idée de base est que nous comparons sur les mêmes stimuli les taux de reconnaissance en «lecture labiale» et en «lecture labiale + code LPC». Le test doit par conséquent se dérouler en deux étapes : dans la première, sont proposés des stimuli avec la modalité «visage» tandis que dans la seconde les mêmes stimuli sont présentés avec la modalité «visage + main». Après chaque stimuli, nous demanderons aux sujets de choisir parmi deux mots (dont seule la consonne initiale diffère), celui qui a été prononcé. Comme pour le *test de rime*, nous allons tester la capacité de notre système à transmettre la consonne initiale dans des mots de type CVC.

15.1.2 Description des stimuli

La construction de la base des stimuli repose sur la volonté de pouvoir différencier des sosies labiaux par le codage manuel. Nous avons donc décidé de déterminer des opposition de mots qui ne varient que dans leur consonne initiale (avec pour priorité les sosies labiaux) et qui s'opposent au niveau de la forme de la main. Nous avons donc chercher ces mots pour toutes les oppositions configuration i / configuration j, avec $i \in$ [1..7] et $j \in$ [i+1..8]. Pour exemple, l'opposition configuration 1 (forme) / configuration 4 (forme) correspond à l'opposition [p] *vs* [b]. Ceci a été réalisé pour les cinq positions de la main par rapport au visage. Toutes les oppositions n'existent pas dans tous les contextes vocaliques. Le nombre de stimuli final est de 196. Le choix de la consonne a donc été effectuée dans le but de mettre à l'épreuve la lecture labiale, tandis que le choix des voyelles visait à tester la coarticulation.

configuration 1

C / V	a	i	u	ɛ̃	y
d / z	-	digue/zig	doute/zoute	daim/zinc	duppe/ZUP
d / s	dalle/salle	deal/cil	doute/soute	daim/saint	dupe/sup
p / b	pâte/batte	pile/bile	poule/boule	pain/bain	pure/bure
d / t	dard/tard	dire/tire	douche/touche	daim/teint	dure/turent
ʒ / ʃ	jatte/chatte	gîte/chite	joue/choux	-	jurent/churent
ʒ / g	jase/gaze	gisent/guise	joute/goutte	geint/gain	jute/gutte
ʒ / j	-	-	joule/ioule	-	jules/iule

configuration 2

C / V	a	i	u	ɛ̃	y
z / s	zappe/sape	zil/cil	zouc/souc	zende/scinde	ZUP/sup
z / n	zappe/nappe	zil/Nil	-	-	-
v / f	Var/phare	vil/fil	voute/foot	feinte/vîntes	-
k / ʃ	cale/châle	quiche/chiche	coule/choule	coute/shoot	cure/churent
k / g	case/gaze	quiche/guiche	court/gourd	-	-
k / j	-	-	coule/ioule	-	-

configuration 3

C / V	a	i	u	ɛ̃	y
s / n	sache/nage	cil/Nil	nouille/souille	saint/nain	-
s / t	sache/tache	cire/tire	soute/toute	sainte/teinte	suc/tuque
ʁ / l	ratte/latte	rire/lyre	Rhur/lourd	Rhin/lin	rut/luth
ʁ / g	râle/gale	rise/guise	route/goutte	rein/gain	rut/gutte
ʁ / j	-	-	roule/ioule	-	-

configuration 4

C / V	a	i	u	ɛ̃	y
b / m	bache/mâche	biche/miche	boule/moule	bain/main	bulle/mulle
n / l	nappe/lappe	niche/liche	nous/loup	nimbe/limbes	nuque/Lucques
n / g	nappe/Gap	nib/guib	nouille/gouille	-	-
n / j	-	-	-	-	nulle/iule

configuration 5

C / V	a	i	u	ɛ̃	y
t / l	tard/lard	tire/lyre	tour/lourd	teint/lin	tûmes/lûmes
t / g	tag/gag	-	tour/gourd	teint/gain	tut/gutte
t / j	-	-	-	-	tulle/iule

configuration 6

C / V	a	i	u	ɛ̃	y
ʃ / g	châle/gale	chiche/guiche	shoot/goutte	-	chute/gutte
ʃ / j	-	-	-	-	-

configuration 7

C / V	a	i	u	ɛ̃	y
g / j	-	-	goule/ioule	-	-

15.1.3 Protocole

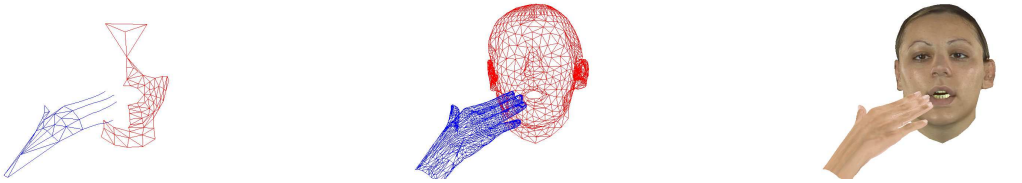
Format des stimuli

Les stimuli sont des mots de type CVC prononcés par la tête parlante de l'ICP sous son aspect vidéo-réaliste comme représentée figure 15.1 (c).

Les modalités de présentation sont de deux types : «visage», «visage + main». Il est à noter que le son n'est jamais présent.

Matériel

Les stimuli sont présentés sur l'écran d'un micro-ordinateur portable (taille de l'écran : 17", carte vidéo : NVIDIA GeForce FX 128 Mo, CPU : Intel Pentium M Centrino 1,6 GHz, RAM : 1 Go). Ils sont joués par le moteur d'animation 3D de l'ICP.



(a) BD (b) HD (c) Texturée

FIG. 15.1 – Tête parlante de l'ICP capable de coder le LPC à partir de n'importe quel texte tapé au clavier.

Sujets

Les sujets, au nombre de 8, sont des personnes sourdes profondes ayant appris la Langue française Parlée Complétée dès 3 ans. Leur âge est compris entre 17 et 26 ans. La moitié des sujets a déjà travaillé sur des axes de recherche liés à la synthèse de parole audiovisuelle. Ils n'ont reçu aucune rémunération pour passer ce test.

Déroulement

Pendant le test, les sujets sont assis à une distance confortable de l'écran du micro-ordinateur portable. Pour le sujet, la tâche consiste, à déterminer pour chaque stimulus ce que le visage parlant a dit parmi un choix imposé de deux réponses possibles. Les stimuli sont présentés dans un ordre aléatoire pour chaque sujet et pour chaque modalité. L'expérience dure environ 20 minutes pour la modalité «lecture labiale» et une douzaine de minutes pour la modalité «lecture labiale + code LPC». Une phase de présentation débute le test où est expliqué le but de celui-ci. Puis, trois stimuli d'entraînement sont proposés. Enfin, vient la phase de test : chaque stimulus est présenté (1 seule répétition, pas de possibilité de rejouer) puis à la fin du stimulus l'ensemble des deux choix possibles est proposé. Lorsque le sujet clique sur une des réponses (possibles), le stimulus suivant est lancé.

15.1.4 Résultats

Les résultats se composent des taux d'intelligibilité pour les deux modalités mais également des temps de réponse qui est un facteur de la charge cognitive nécessaire à la tâche proposée.

Taux d'intelligibilité

L'intelligibilité moyenne pour la condition «lecture labiale» est de 52.36% (voir figure 15.2 à gauche). Ce taux n'est pas significativement différent du hasard, ce qui montre que les sosies labiaux établis dans la liste des noms étaient si proches dans leur représentation que les sujets ne pouvaient que répondre au hasard. Pour ce qui est de la condition «lecture labiale + code LPC», l'intelligibilité moyenne est de 94.26%. La différence entre ces 2 taux d'intelligibilité est

significativement différente et prouve que le clone codeur est capable de fournir une information linguistique significative en termes de mouvements de main. On retrouve les ordres de grandeur mis en évidence sur le codage manuel *vs* lecture labiale par [3, 5]. On peut voir sur les matrices de confusion pour les 2 conditions (voir tableaux 15.1 et 15.2) les erreurs commises par les sujets. Par exemple, pour le groupe des consonnes bilabiales, la consonne [p] n'est pas reconnue dans la modalité «lecture labiale» (25 fois sur 40 elle est reconnue comme un [b]) alors qu'elle est toujours reconnue dans la modalité «lecture labiale + code LPC». L'accroissement du taux d'intelligibilité entre les deux modalités se traduit par la «diagonalisation» de la matrice de confusion pour la modalité «lecture labiale + code LPC». Il est à noter que les stratégies de réponses pour la modalité «lecture labiale» sont différentes suivant les sujets : si l'on reprend le cas des consonnes bilabiales, certains sujets vont répondre systématiquement [p] lorsqu'ils ont le choix entre [p] *vs* [b] comme si le trait voisé ne pouvait être présent alors que d'autres sujets vont répondre au hasard avec 50% des réponses [p] et 50% [b]. On peut remarquer également qu'il y a plus d'erreurs dans l'opposition [b] *vs* [m] que dans l'opposition [b] *vs* [p] pour la modalité «lecture labiale + code LPC». Ceci peut s'expliquer par la forme de main qui bien que distincte est assez proche pour coder [b] et [m] : seul le pouce différencie les 2 configurations contrairement à l'opposition [b] *vs* [p] où les formes sont bien différentes puisque 3 doigts rentrent en compte dans cette différence. Cette différence a été mis en évidence dans le chapitre précédent : la classfification sous forme d'arbre montrait que les formes 4 et 5 sont plus proches que les formes 1 et 4.

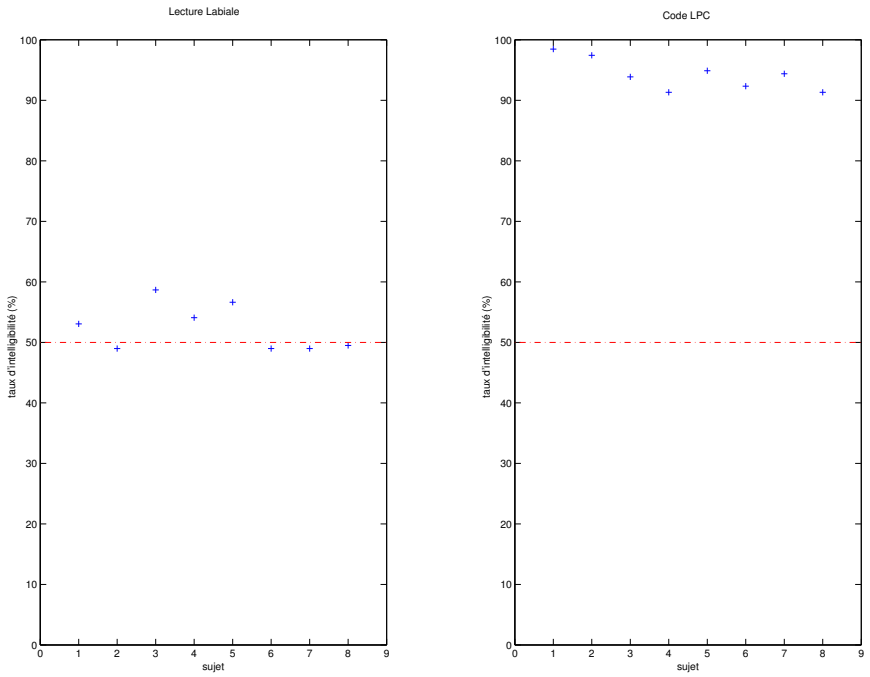


FIG. 15.2 – Taux d'intelligibilité pour nos 8 sujets pour les deux modalités de présentation. À gauche, «lecture labiale». À droite, «lecture labiale + code LPC».

	d	t	n	z	s	p	b	m	ʒ	ʃ	g	k	v	f	ʁ	j	l
d	48	18	-	18	20	-	-	-	-	-	-	-	-	-	-	-	-
t	29	71	-	-	23	-	-	-	-	-	16	-	-	-	-	19	2
n	-	-	69	10	17	-	-	-	-	-	7	-	-	-	-	11	6
z	11	-	4	51	14	-	-	-	-	-	-	-	-	-	-	-	-
s	22	15	16	22	77	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	15	25	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	-	15	41	24	-	-	-	-	-	-	-	-	-
m	-	-	-	-	-	-	24	16	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	-	-	57	17	9	-	-	-	-	-	5
ʃ	-	-	-	-	-	-	-	-	17	61	12	14	-	-	-	-	-
g	-	14	7	-	-	-	-	-	15	19	115	9	-	-	17	-	4
k	-	-	-	-	-	-	-	-	-	13	11	45	-	-	-	-	3
v	-	-	-	-	-	-	-	-	-	-	-	-	17	15	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	18	14	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	17	-	-	-	51	17	3
l	-	22	27	-	-	-	-	-	-	-	-	-	-	-	24	47	-
j	-	4	3	-	-	-	-	-	13	-	1	2	-	-	7	-	26

TAB. 15.1 – Matrice de confusion globale (ensemble des sujets) pour la consonne initiale pour la modalité «lecture labiale».

	d	t	n	z	s	p	b	m	ʒ	ʃ	g	k	v	f	ʁ	j	l
d	102	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-
t	2	155	-	-	-	-	-	-	-	-	1	-	-	-	-	2	-
n	-	-	99	-	17	-	-	-	-	-	1	-	-	-	-	3	-
z	2	-	-	74	4	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	1	-	151	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	40	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	-	1	74	5	-	-	-	-	-	-	-	-	-
m	-	-	-	-	-	-	5	35	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	-	-	85	2	1	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	-	3	100	1	-	-	-	-	-	-
g	-	3	1	-	-	-	-	-	3	6	180	2	-	-	5	-	-
k	-	-	-	-	-	-	-	-	-	4	6	62	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	-	32	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	3	29	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88	-	-
l	-	1	1	-	-	-	-	-	-	-	-	-	-	-	1	117	-
j	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55

TAB. 15.2 – Matrice de confusion globale (ensemble des sujets) pour la consonne initiale pour la modalité «lecture labiale + code LPC».

Temps de réponse

Au niveau de l'effort cognitif, la modalité «lecture labiale + code LPC» est plus aisée : le temps de réponse (voir figure 15.3) est significativement différent (ANOVA à mesures répétées ($F(1, 3134) = 7.5$, $p < 0.01$) et plus faible que celui de la modalité «lecture labiale». C'est d'ailleurs un résultat qualitatif relevé auprès des sujets de cette expérience qui trouvaient dans leur majorité que la tâche «lecture labiale» était pénible. Le gain est donc double, en termes d'intelligibilité et en termes de charge cognitive nécessaire.

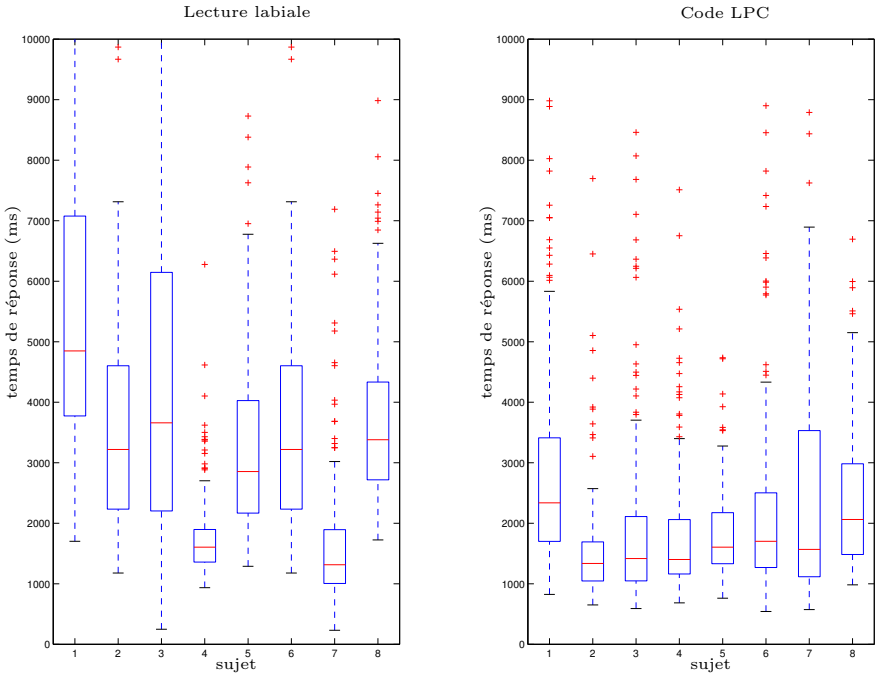


FIG. 15.3 – Temps de réponse pour nos 8 sujets pour les deux modalités de présentation. À gauche, «lecture labiale». À droite, «lecture labiale + code LPC»

15.2 Compréhension

15.2.1 Protocole

Cette deuxième partie du test correspond à l'évaluation de la compréhension de notre système de synthèse. Pour évaluer cette compréhension générale, nous demandons au sujet de visualiser une vidéo dans laquelle est incrustée un clone 3D de synthèse codant LPC (*cf.* la figure 15.4). Cette vidéo est issue de l'émission d'ARTE nommée *Karambolage*. Il s'agit d'un documentaire d'une durée de 3'24 sur l'histoire du *CaramBar*. Après avoir visualisé l'émission, il est demandé aux sujets de répondre à une série de dix questions (listées en **annexe B**) portant à la fois sur le contenu de la vidéo (image de fond) et sur le discours (codé par le clone). Il est à noter que seul 7 sujets sur les 8 ayant passé le précédent test ont visualisé cette vidéo, un des sujets n'ayant pas pu faire l'expérience pour des raisons de disponibilités.

FIG. 15.4 – Impression écran (taille réelle : 720x576 pixels) lors de l'émission *Karambolage* avec le clone LPC incrusté.

15.2.2 Résultats

Le nombre de réponse moyen est de 3 réponses pour les 3 sujets qui n'ont pas été dérouté par le clone de synthèse en incrustation. Les résultats par sujet se trouvent en **Annexe C**. Pour la majorité des sujets cette incrustation n'est pas habituelle et pose des problèmes d'adaptation. Certains sujets n'ont pas souhaité regarder l'émission avec incrustation dans sa totalité. Une seconde vidéo où cette fois-ci la vidéo d'une vraie codeuse était incrustée, était présentée aux sujets. Les personnes ayant visualisé ces 2 vidéos en entier ont pu répondre à 3,3 questions de plus en moyenne que précedemment. Même s'il y a un effet d'apprentissage évident, cet accroissement se traduit qualitativement par des commentaires positifs envers le naturel *vs* la synthèse. Il est à noter que quelque soit l'incrustation aucun sujet n'est capable de percevoir le discours en entier mais seulement des mots isolés. Les commentaires des sujets après l'expérience nous permettent d'émettre des hypothèses pour expliquer le peu de réponses correctes fournies : le rythme (imposé par le sous-titrage télétexte) trop élevé, la présence de beaucoup de noms propres, le manque de marqueurs «prosodiques» de segmentation et d'emphase...

15.3 Conclusions

15.3.1 Intelligibilité

Pour les 8 sujets, il y a un gain d'intelligibilité significatif entre les deux modalités. En effet, pour la modalité «lecture labiale», les taux de reconnaissance ne sont pas significativement différents du hasard alors que pour la modalité «lecture labiale + code LPC», les taux de reconnaissance sont supérieurs à 90%.

Les stimuli ont été choisis de telle sorte qu'il y ait un maximum d'oppositions de sosies

labiaux. Ainsi, les taux de reconnaissance non significativement différents du hasard pour la modalité «lecture labiale» suggèrent que les mouvements labiaux sont assez réalistes pour tromper les sujets. Afin de confirmer cette conclusion, Il nous faut créer un test supplémentaire où les oppositions impliqueraient des formes de lèvres différentes pour des configurations de clés identiques.

Dans le cadre du test qui est construit pour vérifier l'apport d'information de la main, la différence significative des taux de reconnaissance en fonction de la modalité, nous permet d'affirmer que cet apport est avéré.

Les temps de réponse qui sont significativement supérieurs pour la session lecture labiale comparée à la session lecture labiale augmentée du code LPC, viennent confirmer la conclusion précédente : l'apport de l'information de la main est significatif et se complète d'un confort d'utilisation. Cette remarque est reprise par tous les sujets qui trouvent la première tâche très difficile.

15.3.2 Compréhension

Pour le test de compréhension, les résultats sont moins probants. En effet, la moitié des sujets ne révèle qu'une différence minime entre le clone de synthèse et la vidéo naturelle incrustée (si l'on tient compte de l'effet d'apprentissage). L'autre moitié au contraire trouve la tâche trop difficile dans le cas du clone LPC mais réussit à capter une partie du message dans le cas de la vidéo naturelle incrustée. Afin de pouvoir fournir des conclusions plus tranchées, nous devons poursuivre les tests auprès d'un plus grand nombre de sujets.

Toutefois, nous pouvons noter que la tâche demandée est difficile puisqu'à la compréhension de la vidéo de fond s'ajoute la compréhension du code LPC de l'émission. Il semble que des «marqueurs» de début de phrases pourraient aider les sujets. Par marqueur, nous entendons un signal qui souligne le début de la phrase. Il pourrait s'agir simplement d'une lumière ou de mouvements prosodiques effectués par le clone. Il serait intéressant d'ajouter des marqueurs prosodiques gestuels ou autres capables de fournir une segmentation du message et de mettre en emphase des parties difficiles du discours.

Pour aller plus loin...

Afin de comprendre les raisons des résultats du test de compréhension, nous avons conduit une expérience supplémentaire à l'aide d'un système oculométrique Tobii® non invasif. Nous avons demandé à 4 sujets sourds profonds ayant appris le LPC de façon précoce de visualiser la vidéo du test de compréhension sous-titré dans sa première partie et avec incrustation d'une vidéo de codeuse dans la seconde partie (voir figure 15.5). Les résultats préliminaires montrent que les sujets passent 56.36% en moyenne sur le cadre télétexte contre 80.70% sur le cadre de l'incrustation vidéo avec une différence significative ($F(1,6) = 9.06$, $p < 0.05$). Ce résultat indique que l'incrustation d'une vidéo à la place du télétexte n'est pas bénéfique en terme de charge cognitive dans notre cas, puisque le temps passé à décoder est significativement supérieur à celui de la lecture. Ceci peut être dû au rythme trop élevé du télétexte dans cette émission (un documentaire). Une série de 18 questions (les 10 questions précédentes auxquelles on a rajouté

8 questions supplémentaires afin d'avoir une répartition homogène des questions sur toute la vidéo) sont posées aux sujets. Le nombre de réponses correctes est de 10, les erreurs se situant équitablement dans les deux parties de la vidéo. Il est à noter que les erreurs de la deuxième partie sont plus des erreurs de compréhension du discours prononcé dans la vidéo incrustée que sur la trame de fond, alors que les sujets passent significativement beaucoup plus de temps sur l'incrustation que sur le reste de la vidéo. Dans la première partie, il y a 3 questions relatives à des noms propres et 1 relative à une date. Afin de déterminer si ces questions ou d'autres sont trop difficiles nous avons demandé à 16 sujets entendants de visualiser la même vidéo entièrement sous-titrée avec le même dispositif oculométrique. Les sujets entendants passent 40.14% du temps en moyenne sur le cadre télétexte soit légèrement moins que les sujets sourds mais dans le même ordre de grandeur. Le nombre moyen de réponses correctes est de 13 : les erreurs sont commises pour les questions impliquant des noms propres ou des dates et sur des détails de la trame de vidéo de fond. Ainsi, dans la deuxième partie de la vidéo où des questions sur des noms propres ou des dates ne sont pas présentes les erreurs sont essentiellement des erreurs par rapport à la vidéo. L'incrustation de la vidéo de la codeuse **pour cette émission** n'est pas bénéfique par rapport au télétexte tout du moins pour les personnes ayant accès à la lecture. Il faudrait poursuivre l'expérience pour déterminer quel type d'émission est le plus approprié en terme de débit pour cette application.

(a) Télétexte (b) Incrustation vidéo

FIG. 15.5 – Port du regard pour un sujet visualisant une émission de télévision dans sa première partie avec sous titrage télétexte (a) puis dans sa seconde partie avec incrustation d'une vidéo d'une codeuse.

15.4 Résumé

Nous avons présenté dans ce chapitre le test perceptif que nous utilisons pour mesurer l'intelligibilité et la compréhension de notre synthétiseur. Ce test se déroule en deux parties : la première consiste à vérifier l'intelligibilité segmentale et la seconde à quantifier la compréhension. Nous avons relevé les résultats sur quatre sujets sourds ou malentendants ayant une pratique régulière du code LPC. Au niveau segmental, nous montrons qu'il y a une différence significative

entre la lecture labiale seule et la lecture labiale augmentée de code LPC de synthèse. La main apporte par conséquent une information linguistique utile qui est perçue par les sujets pour comprendre les signaux délivrés par le système de synthèse. En outre, nous vérifions ce résultat en terme de charge cognitive : les temps de réponse sont significativement plus faibles lorsque la main est présente. Au niveau supérieur, les résultats sont moins tranchés, il semblerait que le débit soit trop élevé pour que les sujets puissent à la fois regarder l'émission et comprendre l'énoncé. De plus, il semblerait qu'une information prosodique pourrait aider à augmenter la compréhension globale de la vidéo.

Références bibliographiques

[1] D. Busquet and C. Descourtieux. *T.E.R.M.O. Tests d'Evaluation de la Réception du message Oral par l'enfant sourd à destination des professionnels de la surdité.* 2003.

[2] G. Fairbanks. Test of phonemic differentiation : the rhyme test. *Journal of the Acoustical Society of America*, 30(7) :596–600, July 1958.

[3] G. Nicholls and D. Ling. Cued Speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25 :262–269, 1982.

[4] J. P. Peckels and M. Rossi. Le test de diagnostic par paires minimales. *Revue d'Acoustique*, 27 :245–262, 1973.

[5] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of Cued Speech. *Journal of Rehabilitation Research and Development*, 31 :20–41, 1994.

Chapitre 16

Résumé de la partie

Cette partie consacrée à l'évaluation nous a amené à nous poser des questions sur la nécessité et la façon d'évaluer notre synthétiseur. Pour cela, nous avons décider de suivre deux voies différentes : une évaluation objective servant de diagnostic et une évaluation subjective servant de quantificateur.

Le but final de notre système étant d'être une aide à la communication, nous nous sommes focalisés sur des critères d'intelligibilité. Tout d'abord, d'un point de vue objectif, nous avons choisi des critères liés aux lèvres et aux formes de main pour quantifier l'apport d'information. Nous avons calculé trois types d'erreurs entre les phrases du corpus **main + visage** (notre référence) et les mêmes phrases synthétisées : une erreur RMS sur les positions 3D des points du visage et de la main, une erreur RMS sur les valeurs des paramètres d'animation et une erreur probabiliste. Cependant, il est difficile d'interpréter les résultats liés à ces critères. C'est pourquoi, nous avons poursuivi par des évaluations subjectives.

Nous avons présenté un test perceptif visant à mesurer l'intelligibilté segmentale de notre synthétiseur puis plus globalement sa capacité à transmettre de l'information dans une tâche complexe (substitution au sous-titrage télétexte). Les résultats obtenus nous permettent d'affirmer que notre main de synthèse apporte une information significative tant au niveau de l'intelligibilité qu'au niveau du confort de compréhension. Cependant, dans la tâche plus complexe visant à quantifier la compréhension, les résultats sont moins tranchés. Il apparaît toutefois que le code LPC délivré est trop rapide (limitations liées à la tâche de substitution du télétexte). En outre, des «marqueurs» temporelles pourraient être d'une aide significative pour segmenter le discours. Ces «marqueurs» pourraient être des mouvements prosodiques ajoutés au code LPC de synthèse.

Conclusion

La Langue française Parlée Complétée est une technique d'aide à la lecture labiale, permettant aux sourds de percevoir, par la vision uniquement, la parole dans sa totalité. Ce code, dont l'efficacité perceptive est maintenant avérée, permet aux enfants sourds d'acquérir une phonologie de la langue comparable à celle des enfants entendants. Le code LPC constitue donc un bon moyen pour remédier aux problèmes liés au handicap auditif.

Or, dans le cadre de l'assistance aux personnes handicapées, les nouvelles technologies et notamment les interfaces homme-machine ont leur rôle à jouer. Le projet ARTUS [2] s'inscrit parfaitement dans cette logique. Il vise en effet à donner la possibilité aux sourds de bénéficier d'un codeur virtuel de Langue française Parlée Complétée remplaçant le télétexte dans les émissions télévisuelles, un cas d'applications de la transmodalité [7].

C'est dans ce cadre que nous avons effectué ces travaux de thèse qui font partie intégrante de ce projet. Ils sont en effet consacrés à la conception et à l'évaluation d'un système de synthèse 3D de code LPC à partir de texte, technologie tout à fait novatrice. Ce système de synthèse de code LPC s'inscrit dans la continuité des apports de la technologie pour l'accés aux communications des personnes déficientes auditives [6]. Une boucle complète «*expérimentation, modélisation, analyse, synthèse, évaluation*» a permis de mettre en oeuvre et de valider **le premier système de synthèse 3D de Langue française Parlée Complétée à partir du texte**. Celle-ci s'inscrit dans un paradigme de clonage de mouvements naturels biologiques dans le but de se rapprocher au plus près du système humain. La phase d'*expérimentation* a consisté à enregistrer par une technique de *MoCap* la dynamique caractéristique des articulateurs (visage, main) mis en jeu dans la production du code LPC. Dans la phase de «modélisation», nous avons déterminé les paramètres articulatoires d'animation les plus appropriés pour les articulateurs du code LPC afin de préserver les données originales. La phase d'«analyse» de ces gestes a permis de confirmer pour des phrases l'anticipation de la main par rapport au visage qui avait été montrée par Attina et al. [1] et de déterminer la méthodologie la plus adaptée pour la génération automatique des mouvements. Dans la phase de «synthèse», nous avons proposé pour tenir compte du phasage spécifique une synthèse en 2 étapes : tout d'abord la génération de polysons multimodaux (paramètres acoustiques et paramètres articulatoires du visage), puis la génération de diclés (paramètres articulatoires de la main et paramètres de mouvements de la tête et de la main) respectant le schéma temporel de production. Enfin, dans la phase d'«évaluation», nous avons montré un apport significatif, en terme d'information linguistique fournie et en terme d'effort cognitif nécessaire de notre code LPC de synthèse .

Ces travaux de thèse aboutissent, d'un point de vue technologique, sur un démonstrateur de

synthèse 3D de code LPC efficient. D'un point de vue plus théorique, les méthodes développées lors de ces travaux de thèse sont portables à d'autres types de recherche et d'applications. Elles ne se restreignent pas à la modélisation, l'analyse et la synthèse des mouvements du code LPC. Elle pourront être réutilisées pour d'autres types de gestes (gestes communicatifs, gestes sportifs, gestes artistiques, etc.) et pour d'autres applications (synthèse de LSF, synthèse de gestes coverbaux [8], etc.).

Perspectives

D'un point de vue théorique, les perspectives sont diverses : mise en place d'évaluations perceptives complémentaires, amélioration du système de synthèse de parole audiovisuelle, ajout d'expressivité au clone, étude et ajout de traits prosodiques gestuels...

Les **évaluations perceptives complémentaires** du système de synthèse se situent à plusieurs niveaux. Tout d'abord nous avons testé l'apport de l'information manuelle par rapport à la seule lecture labiale ; il serait intéressant de proposer un test similaire où les paires minmales ne se distinguent non pas par les formes de main mais par les formes de lèvres. Les deux résultats permettraient d'évaluer la qualité du matriçage de l'indice phonétique fourni par notre système de synthèse. Ensuite, l'utilisation d'un dispositif oculométrique de façon plus intensive permettrait d'évaluer plus précisément la charge cognitve mis en jeu par les 2 systèmes d'aide à la communication que sont le télétexte et l'incrustation d'un clone de synthèse. Plus particulièrement, la différence entre le décodage et la scrutation de la scène ; des résultats préliminaires ont été présentés dans ce rapport, les analyses se doivent d'être poursuivies. En outre, il serait intéressant d'étudier les différences de décodage entre une vidéo d'une codeuse réelle et la vidéo de son clone : Quelles sont les zones parcourues par le regard ? Quel temps les sujets passent-ils sur ces zones ? Au niveau de l'évaluation de la compréhension globale, nous avons mis en oeuvre des tests perceptifs avec des contraintes (débit élevé, taille de l'image, etc.) liées à une application donnée . Afin d'évaluer plus précisément la qualité de notre système à fournir un texte compréhensible, nous pourrions proposer une série de tests basée sur la vidéo «Karambolage» fournie par Arte en faisant varier le débit de l'émission, la taille de l'image d'incrustation...

L'**amélioration du système de synthèse de parole audiovisuelle** peut passer par l'implémentation de nouveaux modèles de contrôle moteur. Ainsi, de nouveaux modèles tels que le système TDA (Task Dynamics for Animation) [5] qui met en oeuvre un système de planification/exécution du mouvement à base de HMM et de concaténation pourrait améliorer la qualité globale des mouvements de synthèse. Une autre piste d'amélioration se situe au niveau du modèle de déphasage entre les mouvements de la main et ceux du visage. L'enregistrement de plus grands corpora permettrait une analyse plus fine de ce déphasage et l'implémentation d'un modèle plus complexe que celui mis en oeuvre durant ces travaux. Ces données supplémentaires pourraiet permettre également de mettre en place une sélection d'unités beaucoup plus grandes (série CV-CV par exemple) contenant tous les mouvements de tous les articulateurs du code LPC.

L'**ajout d'expressivité** peut intervenir au niveau des mouvements faciaux, leurs analyses dans le cadre d'expérience de parole face à face [3] facilitera l'intégration de ces paramètres aux

paramètres articulatoires déjà existants. De l'expressivité peut également se voir rajouter en modélisant finement les yeux et les paupières et en déterminant un modèle de contrôle de la gestion du regard dans une conversation face à face. Comme précédemment, des expériences de dialogues face à face avec des dispositifs oculométriques [4] permettront d'étudier le mouvement spécifique des yeux dans ce type de tâche.

L'**étude de traits prosodiques gestuels** a été abordé au cours de ces travaux de thèse, toutefois il semble important, vu les résultats des tests de compréhension, d'approfondir ce point particulier d'étude. En effet, il semble que pour des tâches complexes c'est-à-dire la compréhension d'un discours long (plus long qu'une phrase) des marqueurs prosodiques que l'on retrouve dans la parole audio sont absents dans la synthèse de code LPC. L'étude et l'ajout de prosodie gestuelle, c'est-à-dire un ensemble de gestes ajoutés au code LPC pour fournir des informations prosodiques sur le discours, permettraient de segmenter le discours et de mettre en emphase des mots difficiles. La première caractéristique est utile pour pouvoir se raccrocher à un début de phrase par exemple. La deuxième caractéristique est fort profitable pour la compréhension des noms propres qui, on l'a vu, sont difficilement perçus s'ils ne sont pas signalés.

D'un point de vue technologique, les méthodologies développées et le système mis en oeuvre dévoilent des perspectives intéressantes pour d'autres applications technologiques et pour des études sur des sujets connexes.

Ce système peut s'inscrire dans une chaîne complète de substitution du télétexte comme dans le projet ARTUS mais pas seulement. Via des modifications mineures, il pourra s'intégrer dans d'**autres applications technologiques** ambitieuses ayant pour objectif l'aide à la communication des personnes sourdes. Il sera d'ailleurs utilisé dans le cadre du projet RNTS TELMA (Téléphonie à l'usage des Malentendants) qui a pour but de proposer des fonctionnalités audiovisuelles originales dans le cadre des télécommunications mobiles. Ce projet vise une véritable transduction de la parole audio vers le code LPC et inversement afin que les communications entre entendants et sourds soient facilitées. Notre système pourra également s'appliquer à des situations quotidiennes de codage (en classe par exemple), aux technologies éducatives (apprentissage du code LPC, apprentissage de la lecture labiale, situations de classe virtuelle à distance) ou à d'autres situations (interprétation temps-réel, codage temps-réel d'émissions...).

L'**étude de sujets connexes** tels que l'étude de la perception du code LPC ou la validation du télétexte par exemple devient plus aisée avec un tel système. En ce qui concerne l'étude de la perception du code LPC, on pourrait étudier plus particulièrement le rôle du déphasage entre les deux effecteurs, que sont la main et le visage, dans l'intégration audiovisuelle de ce code. Cet aspect pourrait être réalisé adéquatement à l'aide de notre outil de synthèse par la génération de stimuli présentant divers degrés de phasage. Quant à la validation du télétexte, il faut savoir qu'à l'heure actuelle aucune vérification n'est faite sur la capacité de réception du sous-titrage télétexte par les personnes sourdes. La production du sous-titrage télétexte est faite sans savoir si on a le temps matériel de le lire et/ou de le comprendre. Comme il est difficile de passer la «barrière du son» - la lecture à voix basse -, ce système permettrait de vérifier si on a le temps matériel de produire le texte proposé par le sous-titrage.

Références bibliographiques

[1] V. Attina. *La Langue française Parlée Complétée (LPC) : Production et Perception.* PhD thesis, Institut National Polytechnique de Grenoble, 2005.

[2] G. Bailly, V. Attina, C. Baras, P. Bas, S. Baudry, D. Beautemps, R. Brun, J.-M. Chassery, F. Davoine, F. Elisei, G. Gibert, L. Girin, D. Grison, J.-P. Léoni, J. Liénard, N. Moreau, and P. Nguyen. ARTUS : calcul et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue française Parlée Complétée. In *Handicap 2006*, Paris, France, 2006.

[3] G. Bailly, F. Elisei, P. Badin, and C. Savariaux. Degrees of freedom of facial movements in face-to-face conversational speech. In *International Workshop on Multimodal Corpora*, pages 33–36, Genoa, Italy, 2006.

[4] G. Bailly, F. Elisei, and S. Raidt. Virtual talking heads and ambiant face-to-face communication. In A. Esposito, E. Keller, M. Marinaro, and M. Bratanic, editors, *The fundamentals of verbal and non-verbal communication and the biometrical issue.* IOS Press BV, Amsterdam, NL, 2006.

[5] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw. TDA : A new trainable trajectory formation system for facial animation. In *Interspeech ICSLP*, Pittsburgh, PA, September 2006.

[6] J. E. Harkins and M. Bakke. Technologies for communication : Status and trends. In M. Marschark and P. E. Spencer, editors, *Deaf studies, Language, and Education*, chapter Hearing and Speech Perception, pages 406–419. Oxford University Press, New York, 2003.

[7] F. Maurel, N. Vigouroux, M. Raynal, and B. Oriola. Contribution of the transmodality concept to improve web accessibility. In M. Mokhtari, editor, *Independent Living For Persons With Disabilities and Elderly People*, volume 12 of *Assistive Technology Research Series*, pages 186–193. IOS Press, 2003.

[8] D. McNeill, E. T. Levy, and L. L. Pedelty. Speech and gesture. In G. R. Hammond, editor, *Advances in psychology : cerebral control of speech and limb movements*, pages 203–256. Elsevier/North Holland Publishers, Amsterdam, 1990.

Quatrième partie

Annexes

Chapitre 17

Annexe A - Corpus dynamique

1. Ma chemise est roussie.
2. Voila des bougies !
3. Donne un petit coup !
4. Il a du gout.
5. Elle m'étripa.
6. Une réponse ambigüe.
7. Louis pense à ça.
8. Un four touffu.
9. Un tour de magie.
10. Voilà du filet cru.
11. La force du coup.
12. Prête lui seize ecus.
13. Il fait des achats.
14. Chevalier du gué.
15. Le jeune hibou.
16. Il fume son tabac.
17. Un piège à poux.
18. L'examen du cas.
19. Je suis à bout.
20. Elle a chu.
21. Je vais chez l'abbé.
22. Deux jolis boubous.
23. Une belle rascasse.
24. Il part pour Vichy.
25. Faire la nouba.
26. C'est Louis qui joue.
27. C'est ma tribu.
28. Gilles m'attaqua.
29. Une rocaille moussue.
30. Un pied fourchu.
31. La chaise du bout.
32. Trop d'abus.
33. J'en ai assez.
34. Jean est faché.
35. Le pied du gars.
36. Vous avez réussi.
37. Ils n'ont pas pu.
38. Le vent mugit.
39. Une autre roupie.
40. Deux beaux bijoux.
41. Tu ris beaucoup.
42. Dès que le tambour bat les gens accourent.
43. Annie s'ennuie loin de mes parents.
44. Vous poussez des cris de colère.
45. Mon père m'a donné l'autorisation.
46. Un loup s'est jeté immédiatement sur la petite chèvre.
47. J'ai un scorpion sec dans mon talon aiguille.
48. Nos dalmatiens campaient au camping à la montagne.
49. Vend on un cake intact à Hong-kong.
50. Noam Chomsky balaie encore le club ce soir.
51. L'avoué a besoin d'un joint sous huitaine.
52. La sueur suinte du thon huileux.
53. Le beau ouistiti suit le riche huissier à Waterloo.
54. Tout Winipeg attend Wendy sur le parking ouest.
55. Bud et Buck font un bon whist à Maubeuge.

56. Youri fouette l'ail ionique de Kohoutek.
57. Beung j'ai heurté le puits dans la lueur.
58. Vuitton fait cuire dix wapitis goîtreux.
59. David Bowie s'est rué sur le quai où j'ai organisé ce must.
60. Young fait un petit huit avec un joueur noueux.
61. Jean Nohain a chargé Watson de louer le huitième buisson.
62. Li-peng met du nuoc-mam dans son amuse-gueule.
63. J'ai Eugène au téléphone qui cueille joliment du gui.
64. Les keums du wharf rament évidemment dans le paysage.
65. Ivanhoé a fait un bug au huitième essai.
66. Tu huiles l'étui du buzzer de deux watts.
67. J'ai étudié le parking huit à Plancoët.
68. Eh oui les forums de l'accueil sont chouettes.
69. Des Ewoks habitent la maison en paille du centre spatial.
70. La famille ouistiti a éternué sous les dolmens.
71. J'ai identifié un mohican dans un western pyrénéen.
72. Le balai a fait un looping sur la toundra.
73. Ce tuyau a voyagé très haut chez les martiens.
74. Les caïds jouent au ping-pong avec l'équipe de Bosnie.
75. Je souhaite que sa peau usée ne reçoive jamais cette greffe ridicule.
76. Ce fou ordinaire fiche le turban indien dans le bain optionnel.
77. Une agraphe géante a pu heurter son beau hors-bord.
78. De mauvaises gens privent Victor de sa coiffe bretonne.
79. La grive perchée sur l'if noir couve toujours ce canif chinois.
80. Le vase zen a perdu aussi un anneau en roche grise.
81. La houle lave les hublots d'une case déserte.
82. Il abrase chaque jour un pneu ancien avec ses griffes pointues.
83. Le photographe garantit un gag tordu au goût incertain.
84. Le bateau heurta les housses du hublot un peu humides.
85. La feuille fut sertie avec une dent usée de la biche docile.
86. Le géologue trouve finalement la houille en vrac dans le gave de Pau.
87. Le loup oublie son plan astucieux dans une poche chinoise.
88. Le prof mielleux triche souvent à ce jeu idiot.
89. Ce jazz rythmé est un cadeau inespéré.
90. Le veau heureux attend Eudes dans le hameau indien.
91. L'âne bègue voit que la vache de Joseph se vexe.
92. Tu houspilles ton amant onctueux qui louche réellement.
93. Lagaffe fabrique une ruche carrée si tu y coopères.
94. Cette phrase particulière étouffe toute une strophe vertueuse.
95. Ce chant hideux rase son héros venu en hâte.
96. Le camp hostile coordonne le putsch dans la cohue.
97. Cette pêche fameuse a vu onduler l'endive blanche.
98. Il se lève chaque jour et attend Hercule qui oublie.
99. Il n'arrive nullement qu'une vague surgisse du hors-d'oeuvre.
100. Un zébu heureux ne touche jamais au houblon.
101. Son gant entoure la valise trouvée sur la digue droite.
102. Jean heurta une cuve large pleine de gouache verte.
103. Le vent établi sèche bien le houx où crèche mon hibou.
104. Tchang ôte sa toge cintrée d'une main innocente.

105. Dom Juan drague finalement une jeune fille mal faite.
106. à eux la soif zoologique du bourgeon ouvert.
107. Au yen la tâche pénible de ce prêt embarassant.
108. En haut la guêpe pense aux fleurs.
109. L'anglaise lui offre ce qu'elle a au doigt ou à l'oreille.
110. Elle joue uniquement avec la neige chantante.
111. On tua onze ou douze torchons archaïques.
112. Oudini ignore le train où doit se produire le spectacle.
113. Il est parti illico en avion ou en gondole.
114. Il gobe douze fèves et bêche tout mon jardin.
115. La caisse seule a enflé sur le ring en bois.
116. Votre crêpe chaude vise bien le haut du feu.
117. Tailles-en un bien haut et travaille chaque nom.
118. Fernand oublie de moudre son café.
119. L'abeille n'engrange pas de miel sur un chemin.
120. éole aide sa robe fendue à se soulever.
121. Bashung oublie aussi qu'il lègue quelque chose.
122. Je passe chercher ce que j'ai lu avec vous.
123. Un zoom ferait ce que neuf demis pensent faire.
124. Le fou immerge son aiguille et brode finement.
125. Chaque bout du rail carré est une tige ténue.
126. Un argument élogieux échappe bien au rosbif.
127. Le malade guéri attrape mon solide microbe.
128. Zola demande notamment du bon lait à un mage zürichois.
129. Cette dame veut galber un tube vertical.
130. Nous traquions bien Euler pendant son footing urbain.
131. J'ai vu un holding important sur un terre-plein escarpé.
132. Pain et pudding gallois aident le petit hussard oublieux.
133. Une bouteille de Riesling heurta le balcon humide.
134. Ce jeu invite un type joueur et une dame riche.
135. Miss Zazie effectue un travelling heureux sur un machin imposant.
136. Une vache normande dirige rarement un jumping zélé.
137. Le viking honteux a mal chuté sur cette petite nappe.
138. Le pape vient en Yamaha dans une bourgade curieuse.
139. Le lapin utilise son yoyo et a besoin d'aide.
140. Le dumping l'incite à jeter les prunes tombées.
141. Les yétis mal rasés ont la bouille pâteuse.
142. Ils oublièrent Chuck dans un tube carré.
143. Le king charmeur porte une chemise rouge foncé.
144. Yasmine aime ton standing japonais.
145. Gaspard blague mollement sur le leasing omniprésent.
146. Eux aussi aiment la tripe glorieuse un peu euphorique.
147. Oeuvrez pour l'ove du globe bleu des yeux.
148. Mes juges vont manger ce fichu yaourt à la truelle.
149. Cet oeil globuleux porte une lentille luisante.
150. La sage baleine zoophile n'a aucune patte valide.
151. Un pâle zébu agnostique mange normalement une solide pizza.
152. Le prieur brade tout centime gagné.
153. La caille revient sans eux dans l'herbage gourmand.
154. Une guenon heureuse a vu un balcon ombragé.
155. Chaque garçon aime que le soleil brille.
156. Il y a un truc qui ondule dans la cage murale.
157. Tapes-en au noir sur une petite zone.
158. La fausse reine en tailleur agace Guy.
159. Nous tuons chaque chiot qui a été heureux.
160. Flambes-y une crêpe bretonne de gamme moyenne.

161. Chaque zéro est un looping tordu.
162. La meilleure omelette du Larzac peut rivaliser avec le yachting normand.
163. Un nain heurta une bogue charnue un onze janvier.
164. Une tombe Ming ne passe jamais pour un karting belge.
165. Un homme jeune ne tombe pas pendant cette java.
166. Des rides charmantes aèrent cette robe choisie dans les pages jaunes.
167. La foule a afflué quand mon neveu heurta le rer.
168. Le thon heurta un bleuet.
169. Ceux des gueux bigleux veulent libérer Bob Taylor.
170. Où était Oxymel.
171. Le jeu otait illico au parfum oublié un fin bouquet d'embrun.
172. Le cousin chinois du tribun évalue au jugé autrement le tissu invendu.
173. Moreau étale immanquablement un déficit commun à la queue de l'ue.
174. Aladin élève chacun en symbiose avec le vieux ouzbek.
175. Chacun ignore son c.e. un peu un moment.
176. Avec un aplomb imparable nous avons chacun un c.e. énergique.
177. Cette énergie insensée grève un quinzième de Ugines.
178. Sa tape un peu impolie heurta Bernache un peu trop violemment.
179. Sylvain ne suit pas le parfum imprévu.
180. Ce cabot ombrageux fête son accession au pouvoir.
181. Un noir de jais évoque le front eurasien.
182. Ce suspect heurta le bibelot ancien un peu lourdement.
183. Le bedeau euphorique secoue l'anneau un jour par an.
184. Aux lilas violet européens Corot Eugène préfère vingt-et-un oeillets.
185. Jojo heurta le défunt et le tua.
186. Le l.p.e. insiste et les p.m.e. ont signé.
187. Regardes il zigzague un peu vite.
188. Un huit dans l'eau a huilé l'un des tiroirs.
189. Railles un bourrin oisif.
190. Prends- le Euclide.
191. Tailles huit brins ouatés.
192. Je m'huile le corps dans ce lieu iodé.
193. Jourdain rajoute un pneu huileux.
194. Il se ouate le tein rebelle.
195. J'ai reçu ton dessin hier.
196. Quantum suédois ou rituel wolof.
197. La secoueuse fait des percings linguaux.
198. J'ai oublié ton message.
199. Tiens toi assis !
200. Vous êtes exclue.
201. Pas plus de quatre rubis.
202. C'est lui qui me poussa.
203. Il se garantira du froid avec ce bon capuchon.
204. Les deux camions se sont heurtés de face.
205. La vaisselle propre est mise sur l'évier.
206. Les gangs infligent des bings et des bangs périlleux sur une île.
207. Huit jésuites très huileux se font un brushing yougoslave.
208. J'avais honte car la fille huait les Who.
209. L'africa song s'emballe en juillet sur un walkman muet.
210. C'est Hervé qui fuit dans un yacht en leasing.
211. Walid a hué les Pink Floyd à Rouen.
212. Nous jouons aux billes dans les ruines muettes.
213. J'ai huilé un rayon du train huit à l'équinoxe.
214. J'ai eu les symptômes de la presbytie en huit jours.
215. Pose calmement ta dague pointue sur cette étoffe carrée.
216. Va dans une cave quelconque et caches-y ce drapeau honteux.
217. Rêves-y car l'extase vient de cette bague gracieuse.

218. Il élague curieusement la houpe qui est récalcitrante.
219. Quand je soulève ma hache le banc ondule.
220. La horde de hors-la-loi alpague bientôt l'épave galloise.
221. Un très bon vin en bouteille exige un planning idoine.
222. Objectez à Neuilly contre le gaz nocif des hommes.
223. La pin-up feind de tomber chez toi mais ne blague jamais.
224. Cherche où est le thon obtu que je trouve sot.
225. Ce buveur balte augmente sa masse veineuse à heure régulière.
226. Le moteur du Boeing ronronne dans la brouette.
227. Le rotring exige une page carrée dans une feuille verte.
228. Léon range le parking vendéen où on aime zoner.
229. Nous draguions le torrent pour trouver des crabes noirs.
230. Ce soldat un peu honteux fait un job glorieux.
231. Il a été heurté par un pêcheur.
232. Intonnes un u ou un euh à intervalles réguliers.
233. Le c.e. isole les engins communs aux deux charlots.
234. Une québécoise pleurnicheuse brandit Euclide lors des réunions.
235. Un coup heureux et impétueux modifie un vulgaire pain onctueux en gnome.
236. Sur le zing chacun interprète l'atlas humblement posé sur l'ancien jabot.
237. à jeun Antoine le heurte et cet accident le hantera.
238. Antoine avait ouint son numéro huit.

Chapitre 18

Annexe B - Liste de questions du test de compréhension

Vous avez regardé une émission accessible aux sourds et mal-entendants produite par ARTE. Répondez s'il vous plait à quelques questions sur le contenu de cette émission qui nous permettront de savoir si le clone ARTUS apporte effectivement l'information nécessaire à la compréhension du film.

1. Dans quelle ville vit l'inventeur du CaramBar ?
   A. Mayence B. Fribourg C. Coblence D. Cologne E. Ne sait pas

2. Quel est le nom de l'inventeur du CaramBar ?
   A. Niederhoff B. Nigel C. Niebling D. Nilsen E. Ne sait pas

3. Quel pays a un lac Karambar ?
   A. Afghanistan B. Kazakhstan C. Pakistan D. Tadjikistan E. Ne sait pas

4. Où fut inventé le CaramBar ?
   A. Marcheseuil B. Marcq-en-Bareuil C. Marcigny-sur-seine D. Marcheville-en-woevre E. Ne sait pas

5. Combien y a-t-il de vignettes dans la carte postale de la vallée Karambar ?
   A. 1 B. 2 C. 3 D. 4 E. Ne sait pas

6. Quelle est la longueur actuelle d'un CaramBar ?
   A. 7,5 cm B. 6,2 cm C. 6,5 cm D. 7,2 cm E. Ne sait pas

7. Quel est la couleur du costume du gentleman dans la poche duquel le CaramBar a été glissé ?
   A. marron B. noir C. gris D. bleu E. Ne sait pas

8. Avec quoi chauffe-t-on le CaramBar dans la poche du gentleman ?
   A. un sèche-cheveux B. un radiateur C. un chalumeau à gaz D. un dragon E. Ne sait pas

9. Qui est sous l'armoire dans la blague CaramBar ?

A. l'amant de la femme	B. le mari	C. le fils	D. le chat	E. Ne sait pas

10. Qui ne rient pas beaucoup des blagues CaramBar ?

A. les suédois	B. les belges	C. les allemands	D. les hollandais	E. Ne sait pas

Chapitre 19

Annexe C - Résultats de l'évaluation par sujet

Nous présentons dans cette annexe les résultats par sujet. Ce chapitre sera donc découpé en 8 sections chacune relative aux résultats d'un sujet.
Remarque : Lorsque nous utilisons l'adjectif «codée» pour désigner une question c'est qu'elle se réfère à ce que le clone a codé, tandis que lorsque nous utilisons l'adjectif «visuelle» il s'agit d'une question relative à la vidéo de fond.

19.1 Sujet 1

19.1.1 Renseignements

La fiche de renseignements du sujet n°1 est la suivante :
- Nombre d'années de pratique de la LPC : plus de 20 ans
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : déjà vu plusieurs fois
- Surdité : sourd profond

19.1.2 Modalité «lecture labiale»

Lors du passage des stimuli «lecture labiale», le taux de reconnaissance obtenu par le sujet est de 53.61 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.1. Les erreurs correpondent en général à des ambiguïtés liées aux sosies labiaux. Ainsi, le phonème [b] est perçu pour 1/3 des cas comme un [p], pour un autre 1/3 comme un [m] et enfin comme un [b]. Pour ces oppositions, le sujet n'a pas eu de préférence, les trois étant équiprobablement représentés. En revanche, si l'on s'attache à l'opposition [s] / [z], on se rend compte que le sujet privilégie [s].

19.1.3 Modalité «lecture labiale+code LPC»

Lors du passage des stimuli «lecture labiale+code LPC», le taux de reconnaissance obtenu par le sujet est de 98.47 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.2. Le sujet n'a fait que trois erreurs sur les 196 stimuli. Deux erreurs correspondent à des erreurs d'inattention (*sic* le sujet), tandis que la troisième [ʃ] vs [g] correspond à une ambiguïté due à deux formes de doigts proches.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	7	1	3	-	-	2	-	-	-	-	-	-	-	-	-	-	-
z	1	5	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	2	4	11	-	-	1	-	-	-	-	1	-	-	-	-	-	-
p	-	-	-	2	3	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	3	3	-	-	-	-	-	-	-	-	-	-	-	4
t	5	-	3	-	-	6	-	-	2	-	-	-	-	-	-	4	-
ʒ	-	-	-	-	-	-	7	3	-	1	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	3	8	1	-	-	-	-	1	-	-	-
g	-	-	-	-	-	2	2	1	15	-	-	-	-	2	3	-	-
j	-	-	-	-	-	-	1	-	-	4	1	-	-	-	1	-	-
n	-	1	2	-	-	-	-	-	1	1	10	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	3	1	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	-	-
k	-	-	-	-	-	-	-	-	-	-	-	-	-	9	-	-	-
ʁ	-	-	-	-	-	-	-	-	4	1	-	-	-	-	3	3	-
l	-	-	-	-	-	3	-	-	-	-	1	-	-	-	4	7	-
m	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	2

TAB. 19.1 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale» sujet n°1.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	-	9	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-	-
t	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	11	-	-	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	12	1	-	-	-	-	-	-	-	-
g	-	-	-	-	-	-	-	-	25	-	-	-	-	-	-	-	-
j	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-
n	-	-	-	-	-	-	-	-	1	-	14	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-
k	-	-	-	-	-	-	-	-	-	-	-	-	-	9	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-
m	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5

Tab. 19.2 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°1.

19.1.4 Charge cognitive

Les temps de réponse sont représentés sous formes de boîtes à moustaches (voir figure 19.1). Ils sont en moyenne de 5.95s pour la modalité «lecture labiale» contre 3.59s pour la modalité «lecture labiale+code LPC». Les temps de réponse entre les deux modalités sont significativement différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 25.8$, $p < 0.001$)). Le sujet nous a fait part de la difficulté de cette tâche de compréhension par rapport à la tache «lecture labiale+code LPC».

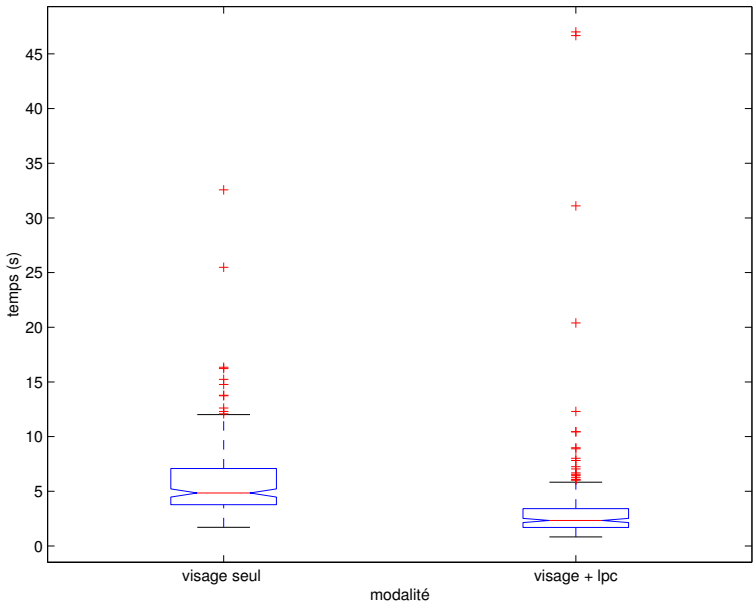


FIG. 19.1 – Temps de réponse pour les deux modalités de présentation pour le sujet n°1.

19.1.5 Compréhension

Lors du test de compréhension, le sujet n'a compris que quelques mots. Il n'a répondu qu'à 2 questions sur les 10 : la question 8 qui est une question essentiellement visuelle et la question 6 qui est visuelle et codée par le clone.

Afin d'obtenir un «étalonnage», le sujet a visualisé à nouveau la vidéo mais cette fois-ci avec la vidéo de la codeuse incrustée. Le sujet rapporte qu'il ne comprend que certains mots comme précédemment mais a pu répondre à d'autres questions : les questions 3, 7, 9 et 10. Il faut tenir compte de l'effet d'apprentissage. Toutefois, cette tâche semble difficile pour le sujet.

19.2 Sujet 2

19.2.1 Renseignements

La fiche de renseignements du sujet n°2 est la suivante :

- Nombre d'années de pratique de la LPC : bébé Lpciste
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : oui
- Surdité : restes auditifs

n° question	réponse attendue	réponse synthèse	réponse naturel	type de réponse
1	C	E	E	codée
2	C	E	E	codée
3	C	E	**C**	codée
4	B	E	E	codée
5	C	E	E	visuelle
6	A	**A**	-	visuelle et codée
7	D	E	**D**	visuelle
8	C	**C**	-	visuelle
9	B	E	**B**	codée
10	C	E	**C**	codée et visuelle

TAB. 19.3 – Réponses au test de compréhension de la part du sujet n°1

19.2.2 Modalité «lecture labiale»

Le taux de reconnaissance global pour les stimuli en lecture labiale seule est de 48.98%. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.4. De même que pour le sujet n°1, les erreurs correspondent en général à des ambiguïtés liées aux sosies labiaux. En revanche, ce sujet privilégie certains phonèmes par rapport à d'autres. Ainsi, lorsque l'opposition [b] / [p] ou l'opposition [b] / [m] se présentent, le sujet répond dans la majorité des cas [b]. Le sujet adopte une stratégie de réponse par rapport aux sosies labiaux.

19.2.3 Modalité «lecture labiale+code LPC»

Le taux de reconnaissance global pour les stimuli en «lecture labiale+code LPC» est de 97.45%. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.5. Le sujet n'a fait que quatre erreurs sur les 196 stimuli. Trois erreurs correspondent à des erreurs d'inattention (*sic* le sujet), tandis que la quatrième [g] vs [ʃ] correspond à une ambigüité dues à deux formes de doigts proches.

19.2.4 Charge cognitive

Les temps de réponse sont représentés sous formes de boîtes à moustaches (voir figure 19.2). Ils sont en moyenne de 3.95 s pour la modalité «lecture labiale» contre 1.55 s pour la modalité «lecture labiale+code LPC». Les temps de réponse entre les deux modalités sont significativement différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 61.15$, $p < 0.001$)). Ce sujet a des temps de réponse beaucoup plus faibles que le sujet n°1. Cependant, la différence entre les deux modalités reste significative.

19.2.5 Compréhension

Lors du test de compréhension, le sujet n'a compris que quelques mots et n'a pas pu comprendre l'intégralité du programme. Il a répondu à 5 questions sur les 10 : les questions 6, 7, 8, 9, 10 avec une erreur à la question 7 et à la question 9. On notera que la réponse à la question 6 est à la fois visuelle et codée, les questions 7 et 8 ont des réponses visuelles et enfin les questions 9 et 10 ont des réponses codées seulement.

Afin d'obtenir un «étalonnage», le sujet a visualisé à nouveau la vidéo mais cette fois-ci avec la vidéo de la codeuse incrustée. Le sujet rapporte qu'il ne comprend que certains mots comme précédemment

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	6	2	3	-	-	2	-	-	-	-	-	-	-	-	-	-	-
z	2	5	2	-	-	-	-	-	-	-	1	-	-	-	-	-	-
s	4	3	9	-	-	2	-	-	-	-	1	-	-	-	-	-	-
p	-	-	-	1	4	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	2	6	-	-	-	-	-	-	-	-	-	-	-	2
t	5	-	2	-	-	11	-	-	1	-	-	-	-	-	-	1	-
ʒ	-	-	-	-	-	-	8	2	1	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	2	7	2	-	-	-	-	2	-	-	-
g	-	-	-	-	-	3	1	3	12	1	1	-	-	1	3	-	-
j	-	-	-	-	-	1	2	-	-	2	-	-	-	1	1	-	-
n	-	1	3	-	-	-	-	-	-	1	10	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	1	3	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	3	1	-	-	-	-
k	-	-	-	-	-	-	-	2	2	-	-	-	-	5	-	-	-
ʁ	-	-	-	-	-	-	-	-	1	-	-	-	-	-	7	3	-
l	-	-	-	-	-	3	-	-	-	-	5	-	-	-	3	4	-
m	-	-	-	-	4	-	-	-	-	-	-	-	-	-	-	-	1

TAB. 19.4 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale» sujet n°2.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	12	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	-	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	9	-	-	-	-	-	-	-	-	-	-	-	1
t	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	11	-	-	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	13	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	-	-	1	24	-	-	-	-	-	-	-	-
j	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-
n	-	-	-	-	-	-	-	-	-	-	15	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-
k	-	-	-	-	-	-	-	-	2	-	-	-	-	7	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-
m	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5

Tab. 19.5 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°2.

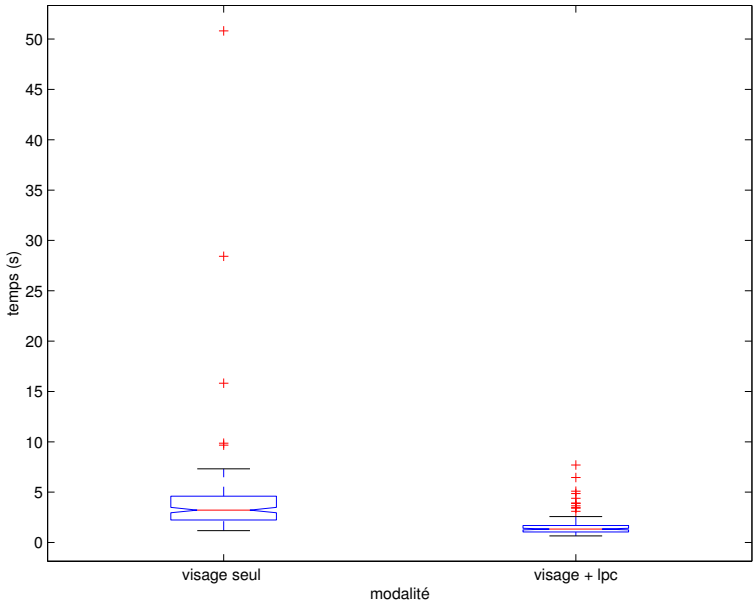


FIG. 19.2 – Temps de réponse pour les deux modalités de présentation pour le sujet n°2.

mais a pu répondre à des questions supplémentaires ou corriger des erreurs : l'erreur corrigée correspond à la question 7 et les questions supplémentaires auxquelles le sujet a répondu sont les questions 2, 3, 4 qui ont des réponses codées uniquement. Il faut tenir compte de l'effet d'apprentissage. Toutefois, cette tâche de compréhension semble difficile pour le sujet.

19.2.6 Compréhension

n° question	réponse attendue	réponse synthèse	réponse naturel	type de réponse
1	C	E	E	codée
2	C	E	**C**	codée
3	C	E	**C**	codée
4	B	E	**B**	codée
5	C	E	E	visuelle
6	A	**A**	-	visuelle et codée
7	D	C	**D**	visuelle
8	C	**C**	-	visuelle
9	B	C	C	codée
10	C	**C**	-	codée et visuelle

TAB. 19.6 – Réponses au test de compréhension de la part du sujet n°2

19.3 Sujet 3

19.3.1 Renseignements

La fiche de renseignements du sujet n°3 est la suivante :
- Nombre d'années de pratique de la LPC : bébé Lpciste (depuis l'âge de 3 ans)
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : oui
- Surdité : restes auditifs

19.3.2 Modalité «lecture labiale»

Le taux de reconnaissance global pour les stimuli en lecture labiale seule est de 58.67 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.7.

De même que pour les sujets n°1 et n°2, les erreurs correpondent en général à des ambigüités liées aux sosies labiaux. En revanche, ce sujet privilégie certains phonèmes par rapport à d'autres. Ainsi, lorsque l'opposition [ʁ] / [l] se présente, le sujet répond dans la majorité des cas [ʁ]. Le sujet adopte, comme le sujet n°2, une stratégie de réponse par rapport aux sosies labiaux.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	5	3	3	-	-	2	-	-	-	-	-	-	-	-	-	-	-
z	-	9	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	2	4	10	-	-	1	-	-	-	-	2	-	-	-	-	-	-
p	-	-	-	2	3	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	1	5	-	-	-	-	-	-	-	-	-	-	-	4
t	2	-	5	-	-	6	-	-	3	1	-	-	-	-	-	3	-
ʒ	-	-	-	-	-	-	7	3	1	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	1	10	1	-	-	-	-	1	-	-	-
g	-	-	-	-	-	-	3	1	17	-	2	-	-	-	2	-	-
j	-	-	-	-	-	-	2	-	-	3	1	-	-	-	1	-	-
n	-	2	1	-	-	-	-	-	-	-	10	-	-	-	-	2	-
v	-	-	-	-	-	-	-	-	-	-	-	3	1	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	-	-
k	-	-	-	-	-	-	-	1	1	-	-	-	-	7	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	1	-
l	-	-	-	-	-	1	-	-	-	-	4	-	-	-	4	6	-
m	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	3

Tab. 19.7 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale» sujet n°3.

19.3.3 Modalité «lecture labiale+code LPC»

Le taux de reconnaissance global pour les stimuli «lecture labiale+code LPC» est de 93.88%. Ce taux de reconnaissance est plus faible que pour les deux sujets précédents. Les conditions de passation du test

était différente, ce qui peut avoir engendrer une dissipation. Il s'agit d'une explication possible de cette baisse du taux de reconnaissance. On note toutefois qu'il est significativement supérieur à celui obtenu pour la modalité «lecture labiale».

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	12	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	-	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	9	-	-	-	-	-	-	-	-	-	-	-	1
t	-	-	-	-	-	19	-	-	1	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	9	2	-	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	13	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	-	-	-	24	-	-	-	-	-	1	-	-
j	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-
n	-	-	3	-	-	-	-	-	-	-	12	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-
k	-	-	-	-	-	-	-	-	2	-	-	-	-	7	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-
m	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	4

TAB. 19.8 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°3.

19.3.4 Charge cognitive

Les temps de réponse sont représentés sous formes de boîtes à moustaches (voir figure 19.3). Ils sont en moyenne de 6.01 s pour la modalité «lecture labiale» contre 2.36 s pour la modalité «lecture labiale+code LPC».

Les temps de réponse entre les deux modalités sont significativement différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 34.89$, $p < 0.001$)).

On remarque toutefois qu'il y a beaucoup de valeurs à l'extérieur de l'espace interquartile, ce qui justifie l'hypothèse de dissipation évoquée précedemment.

19.3.5 Compréhension

Lors du test de compréhension, le sujet n'a compris que quelques mots et a préféré arrêter au bout de quelques secondes. Cette tâche lui paraît trop difficile parce que le codage lui semble trop rapide.

Afin d'obtenir un «étalonnage», le sujet a visualisé à nouveau la vidéo mais cette fois-ci avec la vidéo de la codeuse incrustée. Le sujet rapporte qu'il comprend mieux et arrive à saisir les subtilités du reportage. Il souligne que la codeuse articule très peu. Finalement, il ne répond qu'à 2 questions : la 8 et la 9, précisant qu'il n'était pas très attentif au début de l'émission.

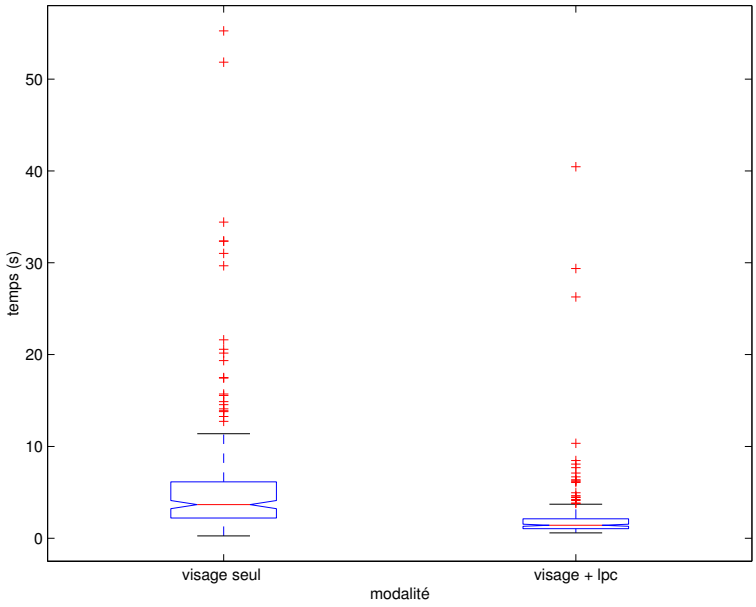


FIG. 19.3 – Temps de réponse pour les deux modalités de présentation pour le sujet n°3.

19.4 Sujet 4

19.4.1 Renseignements

La fiche de renseignements du sujet n°4 est la suivante :

- Nombre d'années de pratique de la LPC : bébé Lpciste (depuis l'âge de 3 ans)
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : oui
- Surdité : restes auditifs

19.4.2 Modalité «lecture labiale»

Le taux de reconnaissance global pour les stimuli «lecture labiale» est de 54.08 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.10.

19.4.3 Modalité «lecture labiale+code LPC»

Le taux de reconnaissance global pour les stimuli «lecture labiale+code LPC» est de 91.33 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.11.

19.4.4 Charge cognitive

Les temps de réponse sont représentés sous formes de boîtes à moustaches (voir figure 19.4). Ils sont en moyenne de 1.90 s pour la modalité «lecture labiale» contre 3.42 s pour la modalité «lecture labiale+code LPC».

Les temps de réponse entre les deux modalités sont significativement différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 3.96$, $p < 0.05$)). Ce sujet a répondu dans des temps très courts pour les deux modalités. Ce sujet semble ne pas être géné par la modalité «lecture labiale». En effet, les résultats en termes de taux de reconnaissance sont similaires aux autres sujets mais dans des temps de réponse

n° question	réponse attendue	réponse synthèse	réponse naturel	type de réponse
1	C	-	-	codée
2	C	-	-	codée
3	C	-	-	codée
4	B	-	-	codée
5	C	-	-	visuelle
6	A	-	-	visuelle et codée
7	D	-	-	visuelle
8	C	-	**C**	visuelle
9	B	-	**B**	codée
10	C	-	-	codée et visuelle

TAB. 19.9 – Réponses au test de compréhension de la part du sujet n°3

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	7	2	2	-	-	2	-	-	-	-	-	-	-	-	-	-	-
z	1	6	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	1	3	11	-	-	2	-	-	-	-	2	-	-	-	-	-	-
p	-	-	-	1	4	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	3	5	-	-	-	-	-	-	-	-	-	-	-	2
t	4	-	3	-	-	10	-	-	1	-	-	-	-	-	-	2	-
ʒ	-	-	-	-	-	-	7	2	1	1	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	2	6	3	-	-	-	-	2	-	-	-
g	-	-	-	-	-	2	1	3	15	1	1	-	-	1	1	-	-
j	-	-	-	-	-	1	2	-	1	1	1	-	-	-	1	-	-
n	-	1	2	-	-	-	-	-	1	1	9	-	-	-	-	1	-
v	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	-	-
k	-	-	-	-	-	-	-	2	1	1	-	-	-	5	-	-	-
ʁ	-	-	-	-	-	-	-	-	2	1	-	-	-	-	8	-	-
l	-	-	-	-	-	3	-	-	-	-	1	-	-	-	2	9	-
m	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	2

TAB. 19.10 – Matrice de confusion pour la consonne initiale pour le test «Lecture labiale» sujet n°4.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	-	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	1	7	-	-	-	-	-	-	-	-	-	-	-	2
t	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	11	-	-	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	13	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	1	1	1	19	-	1	-	-	1	1	-	-
j	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-
n	-	-	3	-	-	-	-	-	-	-	12	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-
k	-	-	-	-	-	-	-	2	-	-	-	-	-	7	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	1	-	-	-	-	1	-	-	-	-	13	-
m	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	4

Tab. 19.11 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°4.

comparables. La poursuite du test sur un plus grand nombre de sujets nous permettra de déterminer s'il existe vraiment une différence significative de charge cognitive.

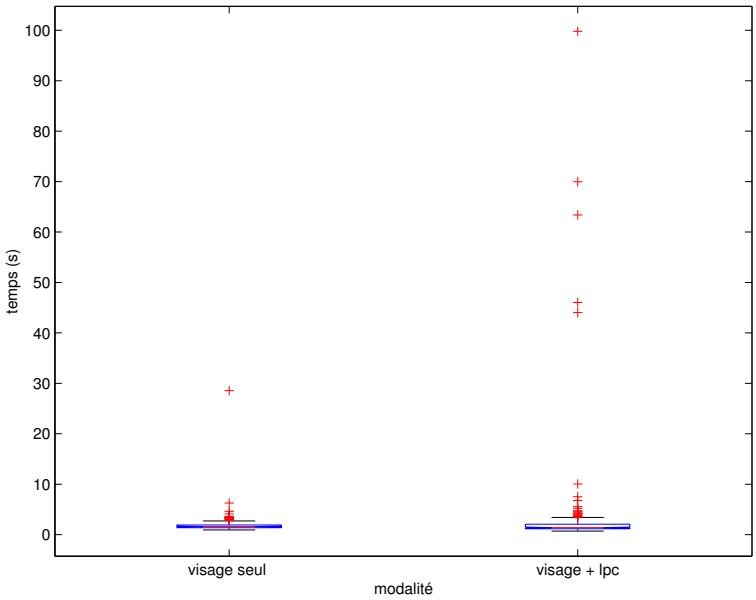


FIG. 19.4 – Temps de réponse pour les deux modalités de présentation pour le sujet n°4.

19.4.5 Compréhension

Lors du test de compréhension, le sujet ne comprenait pas la totalité du message mais quelques mots. Comme précedemment, le sujet n'a pas souhaité continuer car la tâche lui semblait trop difficile.

Afin d'obtenir un «étalonnage», le sujet a visualisé à nouveau la vidéo mais cette fois-ci avec la vidéo de la codeuse incrustée. Le sujet rapporte qu'il comprend mieux et arrive à comprendre les détails du reportage comme précedemment. Il souligne que la codeuse articule très peu.

n° question	réponse attendue	réponse synthèse	réponse naturel	type de réponse
1	C	-	-	codée
2	C	-	-	codée
3	C	-	**C**	codée
4	B	-	**B**	codée
5	C	-	B	visuelle
6	A	-	D	visuelle et codée
7	D	-	C	visuelle
8	C	-	**C**	visuelle
9	B	-	**B**	codée
10	C	-	**C**	codée et visuelle

TAB. 19.12 – Réponses au test de compréhension de la part du sujet n°4

19.5 Sujet 5

19.5.1 Renseignements

La fiche de renseignements du sujet n°5 est la suivante :
- Nombre d'années de pratique de la LPC : depuis l'âge de 4 ans
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : oui (en démonstration)
- Surdité : restes auditifs

19.5.2 Modalité «lecture labiale»

Le taux de reconnaissance global pour les stimuli «lecture labiale» est de 56.63 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.13.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	8	2	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-
z	1	8	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
s	3	2	9	-	-	3	-	-	-	-	2	-	-	-	-	-	-
p	-	-	-	4	1	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	1	6	-	-	-	-	-	-	-	-	-	-	-	3
t	3	-	2	-	-	11	-	-	3	-	-	-	-	-	-	1	-
ʒ	-	-	-	-	-	-	6	2	2	1	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	1	10	1	-	-	-	-	1	-	-	-
g	-	-	-	-	-	1	2	3	17	-	-	-	-	1	1	-	-
j	-	-	-	-	-	1	1	-	-	5	-	-	-	-	-	-	-
n	-	1	3	-	-	-	-	-	2	1	4	-	-	-	-	4	-
v	-	-	-	-	-	-	-	-	-	-	-	1	3	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	1	3	-	-	-	-
k	-	-	-	-	-	-	-	1	1	1	-	-	-	6	-	-	-
ʁ	-	-	-	-	-	-	-	-	3	-	-	-	-	-	6	2	-
l	-	-	-	-	-	4	-	-	-	-	3	-	-	-	3	5	-
m	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	2

TAB. 19.13 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale» sujet n°5.

19.5.3 Modalité «lecture labiale+code LPC»

Le taux de reconnaissance global pour les stimuli «lecture labiale+code LPC» est de 94.90 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.14.

19.5.4 Charge cognitive

Le temps de réponse est en moyenne de 3.52s pour la modalité «lecture labiale» contre 1.86s pour la modalité «lecture labiale+code LPC». Les temps de réponse entre les deux modalités sont significative-

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	-	9	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-	-
t	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	11	-	-	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	2	11	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	-	-	1	22	-	-	-	-	1	1	-	-
j	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-
n	-	-	2	-	-	-	-	-	-	-	13	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-
k	-	-	-	-	-	-	-	1	-	-	-	-	-	8	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-
m	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	4

TAB. 19.14 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°5.

ment différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 98.91$, $p < 0.05$)).

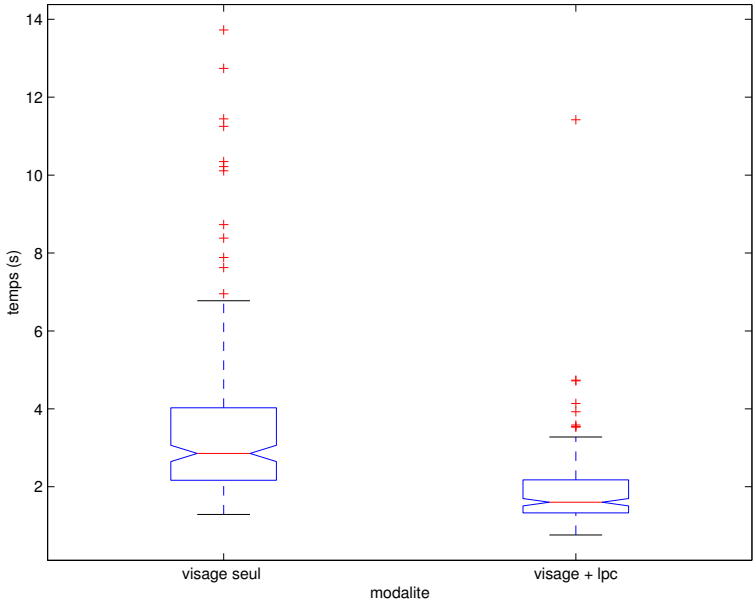


FIG. 19.5 – Temps de réponse pour les deux modalités de présentation pour le sujet n°5.

19.5.5 Compréhension

Lors du test de compréhension, le sujet ne comprenait pas la totalité du message mais quelques mots. La tâche avec la vdéo de la codeuse incrustée lui paraissait difficile également, le sujet n'arrivait pas récupérer uneinformation globale, seulement des mots isolés.

n° question	réponse attendue	réponse synthèse	réponse naturel	type de réponse
1	C	E	-	codée
2	C	E	-	codée
3	C	E	-	codée
4	B	E	-	codée
5	C	E	-	visuelle
6	A	B	-	visuelle et codée
7	D	**D**	-	visuelle
8	C	**C**	-	visuelle
9	B	**B**	-	codée
10	C	**C**	-	codée et visuelle

TAB. 19.15 – Réponses au test de compréhension de la part du sujet n°5

19.6 Sujet 6

19.6.1 Renseignements

La fiche de renseignements du sujet n°6 est la suivante :
- Nombre d'années de pratique de la LPC : depuis l'âge de 3 ans
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : oui (test de l'application Labiao)
- Surdité : restes auditifs

19.6.2 Modalité «lecture labiale»

Le taux de reconnaissance global pour les stimuli «lecture labiale» est de 48.98 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.16.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	6	2	3	-	-	2	-	-	-	-	-	-	-	-	-	-	-
z	2	5	2	-	-	-	-	-	-	-	1	-	-	-	-	-	-
s	4	3	9	-	-	2	-	-	-	-	1	-	-	-	-	-	-
p	-	-	-	1	4	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	2	6	-	-	-	-	-	-	-	-	-	-	-	2
t	5	-	2	-	-	11	-	-	1	-	-	-	-	-	-	1	-
ʒ	-	-	-	-	-	-	8	2	1	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	2	7	2	-	-	-	-	2	-	-	-
g	-	-	-	-	-	3	1	3	12	1	1	-	-	1	3	-	-
j	-	-	-	-	-	1	2	-	-	2	-	-	-	1	1	-	-
n	-	1	3	-	-	-	-	-	-	1	10	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	1	3	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	3	1	-	-	-	-
k	-	-	-	-	-	-	-	2	2	-	-	-	-	5	-	-	-
ʁ	-	-	-	-	-	-	-	-	1	-	-	-	-	-	7	3	-
l	-	-	-	-	-	3	-	-	-	-	5	-	-	-	3	4	-
m	-	-	-	-	4	-	-	-	-	-	-	-	-	-	-	-	1

TAB. 19.16 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale» sujet n°6.

19.6.3 Modalité «lecture labiale+code LPC»

Le taux de reconnaissance global pour les stimuli «lecture labiale+code LPC» est de 92.35 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.17.

19.6.4 Charge cognitive

Le temps de réponse moyen pour la modalité «lecture labiale» est de 3.96 s contre 2.37 s pour la modalité «lecture labiale+code LPC». -avec un ecart-type de 4.208172 Les temps de réponse sont

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	1	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-	-
t	1	-	-	-	-	18	-	-	-	-	-	-	-	-	-	1	-
ʒ	-	-	-	-	-	-	11	-	-	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	13	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	2	1	1	21	-	-	-	-	-	-	-	-
j	-	-	-	-	-	1	-	-	-	6	-	-	-	-	-	-	-
n	-	-	3	-	-	-	-	-	-	-	10	-	-	-	-	2	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	1	3	-	-	-	-
k	-	-	-	-	-	-	-	-	-	-	-	-	-	9	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-
m	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	4

TAB. 19.17 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°6.

représentés sous formes de boîtes à moustaches (voir figure 19.6).

Les temps de réponse entre les deux modalités sont significativement différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 21.97$, $p < 0.05$)).

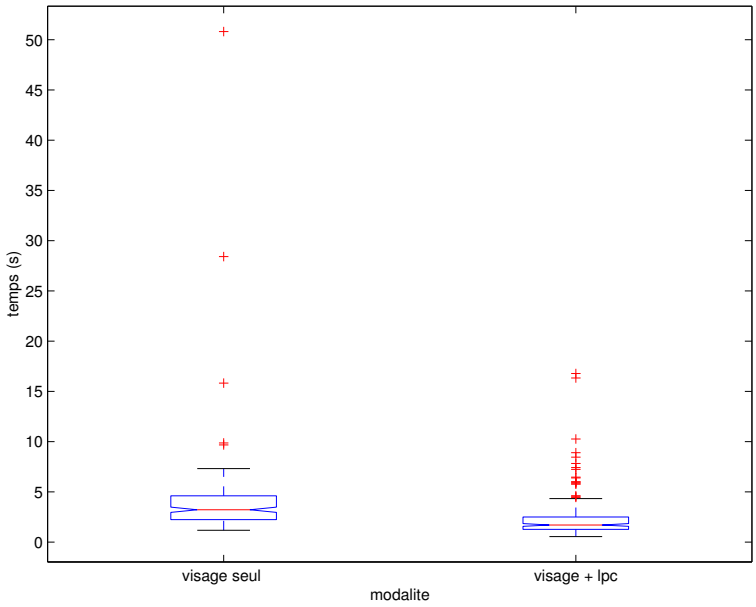


FIG. 19.6 – Temps de réponse pour les deux modalités de présentation pour le sujet n°6.

19.6.5 Compréhension

Le sujet n'a pas pris part à la phase d'évaluation «compréhension» pour des raisons de disponibilités.

19.7 Sujet 7

19.7.1 Renseignements

La fiche de renseignements du sujet n°7 est la suivante :

- Nombre d'années de pratique de la LPC : depuis l'âge de 4 ans
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : non
- Surdité : restes auditifs

19.7.2 Modalité «lecture labiale»

Le taux de reconnaissance global pour les stimuli «lecture labiale» est de 48.98 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.18.

19.7.3 Modalité «lecture labiale+code LPC»

Le taux de reconnaissance global pour les stimuli «lecture labiale+code LPC» est de 94.38 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.19.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	2	3	4	-	-	4	-	-	-	-	-	-	-	-	-	-	-
z	1	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	2	2	10	-	-	2	-	-	-	-	3	-	-	-	-	-	-
p	-	-	-	3	2	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	3	5	-	-	-	-	-	-	-	-	-	-	-	2
t	3	-	3	-	-	7	-	-	3	-	-	-	-	-	-	4	-
ʒ	-	-	-	-	-	-	8	-	2	1	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	3	5	2	-	-	-	-	3	-	-	-
g	-	-	-	-	-	2	2	3	14	-	-	-	-	2	2	-	-
j	-	-	-	-	-	-	2	-	-	4	-	-	-	-	1	-	-
n	-	1	1	-	-	-	-	-	2	-	8	-	-	-	-	3	-
v	-	-	-	-	-	-	-	-	-	-	-	3	1	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
k	-	-	-	-	-	-	-	2	2	-	-	-	-	5	-	-	-
ʁ	-	-	-	-	-	-	-	-	4	-	-	-	-	-	4	3	-
l	-	-	-	-	-	2	-	-	-	-	4	-	-	-	3	6	-
m	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	3

Tab. 19.18 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale» sujet n°7.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	1	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-	-
t	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	10	-	1	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	1	12	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	-	1	1	23	-	-	-	-	-	-	-	-
j	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-
n	-	-	4	-	-	-	-	-	-	-	11	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-
k	-	-	-	-	-	-	-	1	-	-	-	-	-	8	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-
m	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	4

TAB. 19.19 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°7.

19.7.4 Charge cognitive

Le temps de réponse pour la modalité «lecture labiale» est de 2.16 s contre 8.74 pour la modalité «lecture labiale+code LPC». Les temps de réponse sont représentés sous formes de boîtes à moustaches (voir figure 19.7).

Les temps de réponse entre les deux modalités sont significativement différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 14.56$, $p < 0.05$)). Ce sujet a étonnament des temps de réponse plus courts pour la modalité «lecture labiale» que pour la modalité «lecture labiale+code LPC». Ceci s'explique par le fait que le sujet a été dérouté par la tâche et a décidé de répondre au hasard très rapidement.

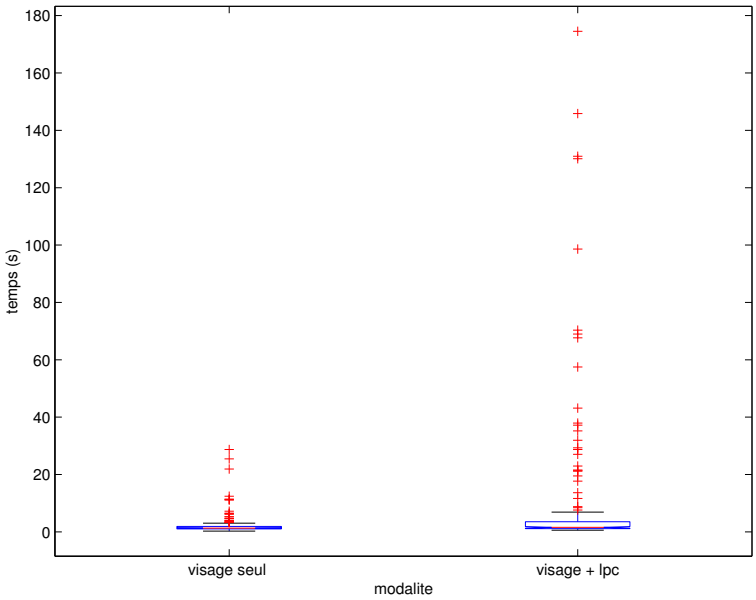


FIG. 19.7 – Temps de réponse pour les deux modalités de présentation pour le sujet n°7.

19.7.5 Compréhension

Le sujet rapporte que l'incrustation d'un clone de synthèse ou d'une vidéo réelle ne l'aide pas et qu'il ne comprend pas le discours.

19.8 Sujet 8

19.8.1 Renseignements

La fiche de renseignements du sujet n°8 est la suivante :

- Nombre d'années de pratique de la LPC : depuis l'âge de 2 ans
- Fréquence de pratique : quotidienne
- Familiarité avec la synthèse de la parole : oui (test de Labiao)
- Surdité : restes auditifs

19.8.2 Modalité «lecture labiale»

Le taux de reconnaissance global pour les stimuli en lecture labiale seule est de 49.49 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.21.

n° question	réponse attendue	réponse synthèse	réponse naturel	type de réponse
1	C	-	-	codée
2	C	-	-	codée
3	C	-	-	codée
4	B	-	-	codée
5	C	-	-	visuelle
6	A	C	-	visuelle et codée
7	D	C	-	visuelle
8	C	**C**	-	visuelle
9	B	-	-	codée
10	C	-	-	codée et visuelle

TAB. 19.20 – Réponses au test de compréhension de la part du sujet n°7

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	7	3	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-
z	3	4	2	-	-	-	-	-	-	-	1	-	-	-	-	-	-
s	4	1	8	-	-	2	-	-	-	-	4	-	-	-	-	-	-
p	-	-	-	1	4	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	5
t	2	-	3	-	-	9	-	-	2	1	-	-	-	-	-	3	-
ʒ	-	-	-	-	-	-	6	3	1	1	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	3	8	-	-	-	-	-	2	-	-	-
g	-	-	-	-	-	1	3	2	13	1	2	-	-	1	2	-	-
j	-	-	-	-	-	-	1	-	-	5	-	-	-	-	1	-	-
n	-	2	2	-	-	-	-	-	1	1	8	-	-	-	-	1	-
v	-	-	-	-	-	-	-	-	-	-	-	3	1	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	1	3	-	-	-	-
k	-	-	-	-	-	-	-	3	2	1	-	-	-	3	-	-	-
ʁ	-	-	-	-	-	-	-	-	2	1	-	-	-	-	6	2	-
l	-	-	-	-	-	3	-	-	-	-	4	-	-	-	2	6	-
m	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	2

TAB. 19.21 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale» sujet n°8.

19.8.3 Modalité «lecture labiale+code LPC»

Le taux de reconnaissance global pour les stimuli «lecture labiale+code LPC» est de 91.33 %. Les résultats sont résumés sous forme d'une matrice de confusion *cf.* table 19.22.

	d	z	s	p	b	t	ʒ	ʃ	g	j	n	v	f	k	ʁ	l	m
d	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	-	8	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	18	-	-	-	-	-	-	-	1	-	-	-	-	-	-
p	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	9	-	-	-	-	-	-	-	-	-	-	-	1
t	1	-	-	-	-	18	-	-	-	-	-	-	-	-	-	1	-
ʒ	-	-	-	-	-	-	11	-	-	-	-	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	13	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	-	-	1	22	-	-	-	-	-	2	-	-
j	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-
n	-	-	2	-	-	-	-	-	-	-	12	-	-	-	-	1	-
v	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	-	-
k	-	-	-	-	-	-	-	-	2	-	-	-	-	7	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	14	-
m	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5

TAB. 19.22 – Matrice de confusion pour la consonne initiale pour le test «lecture labiale+code LPC» sujet n°8.

19.8.4 Charge cognitive

Le temps de réponse est en moyenne pour la modalité «lecture labiale» de 5.81 s contre 2.83 s pour la modalité «lecture labiale+code LPC». Les temps de réponse sont représentés sous formes de boîtes à moustaches (voir figure 19.8). Les temps de réponse entre les deux modalités sont significativement différents (ANOVA à mesures répétées à un facteur ($F(1, 390) = 8.63$, $p < 0.05$)).

19.8.5 Compréhension

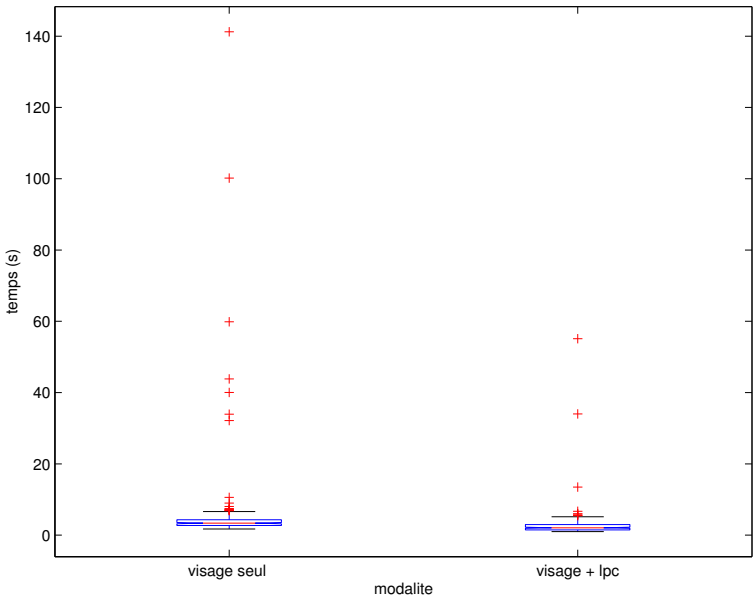


FIG. 19.8 – Temps de réponse pour les deux modalités de présentation pour le sujet n°8.

n° question	réponse attendue	réponse synthèse	réponse naturel	type de réponse
1	C	E	-	codée
2	C	E	-	codée
3	C	E	-	codée
4	B	E	-	codée
5	C	E	-	visuelle
6	A	E	-	visuelle et codée
7	D	C	-	visuelle
8	C	**C**	-	visuelle
9	B	D	-	codée
10	C	E	-	codée et visuelle

TAB. 19.23 – Réponses au test de compréhension de la part du sujet n°8

Références bibliographiques

Chapitre 0

[1] G. Bailly, V. Attina, C. Baras, P. Bas, S. Baudry, D. Beautemps, R. Brun, J.-M. Chassery, F. Davoine, F. Elisei, G. Gibert, L. Girin, D. Grison, J.-P. Léoni, J. Liénard, N. Moreau, and P. Nguyen. ARTUS : calcul et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue française Parlée Complétée. In *Handicap 2006*, Paris, France, 2006.

[2] A. Héloir, S. Gibet, N. Courty, J. F. Kamp, N. Rezzoug, P. Gorce, F. Multon, and C. Pelachaud. Agent virtuel signeur aide à la communication pour personnes sourdes. In *Handicap 2006*, 2006.

[3] R. Kennaway. Synthetic animation of deaf signing gestures. In *Gesture Workshop*, pages 146–157, 2001.

[4] R. Kennaway. Experience with and requirements for a gesture description language for synthetic animation. In *Gesture Workshop*, pages 300–311, 2003.

[5] T. Lebourque and S. Gibet. A complete system for the specification and the generation of sign language gestures. In A. Braffort, R. Gherbi, S. Gibet, J. Richardson, and D. Teil, editors, *Lecture Notes in Artificial Intelligence, LNAI 1739, in Gesture-Based Communication in Human-Computer Interaction*, pages 227–238. Springer-Verlag, 1999.

[6] I. Marshall and E. Sáfár. Grammar development for sign language avatar-based synthesis. In *3rd International Conference on UA in HCI : Exploring New Dimensions of Diversity*, volume 8, Las Vegas, Nevada, 2005.

[7] E. Sáfár and I. Marshall. The Architecture of an English-Text-to-Sign-Languages Translation System. In G. Angelova et al., editor, *Recent Advances in Natural Language Processing (RANLP)*, pages 223–228. Tzigov Chark Bulgaria, 2001.

Chapitre 1

[1] M. Alissali. *Architecture logicielle pour la synthèse multilingue de la parole.* PhD thesis, INPG, Grenoble, France, 1993.

[2] M. Alissali and G. Bailly. Compost : a client-server model for applications using text-to-speech. In *European Conference on Speech Communication and Technology*, pages 2095–2098, Berlin, Germany, 1993.

[3] G. Bailly and M. Alissali. Compost : a server for multilingual text-to-speech system. *Traitement du Signal*, 9(4) :359–366, 1992.

[4] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

[5] G. Bailly and A. Tran. Compost : a rule-compiler for speech synthesis. In *European Conference on Speech Communication and Technology*, pages 136–139, 1989.

[6] B. I. Bertenthal, D. R. Proffitt, and J. E. Cutting. Infant sensitivity to figural coherence in biomechanical motions. *Journal of Experimental Child Psychology*, 37 :213–230, 1984.

[7] B. I. Bertenthal, D. R. Proffitt, and S. J. Kramer. Perception of biomechanical motions by infants : Implementation of various processing constraints. special issue : The ontogenesis of perception. *Journal of Experimental Psychology : Human Perception and Performance*, 13 :577–585, 1987.

[8] J. E. Cutting, D. R. Proffitt, and L. T. Kozlowski. A biomechanical invariant for gait perception. *Journal of Experimental Psychology : Human Perception and Performance*, 4, 1978.

[9] W. H. Dittrich. Actions categories and recognition of biological motion. *Perception*, 22 :15–23, 1993.

[10] W. H. Dittrich. *Lecture Notes in Artificial Intelligence : Gesture-Based Communication in Human-Computer Interaction*, chapter Seeing biological motion - Is there a role for cognitive strategies ?, pages 3–22. A. e. a. Braffort, Berlin, 1999.

[11] W. H. Dittrich, T. Troscianko, S. E. G. Lea, and D. Morgan. Perception of emotion from dynamic point-light displays represented in dance. *Perception*, 25 :727–738, 1996.

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

[13] B. Holm. *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application à l'énonciation de formules mathématiques.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2003.

[14] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14 :201–211, 1973.

[15] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 26 :746–748, 1976.

[16] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 :453–468, 1990.

[17] M. Odisio. *Estimation des mouvements du visage d'un locuteur dans une séquence audiovisuelle.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.

[18] M. Odisio, G. Bailly, and F. Elisei. Talking face tracking with shape and appearance models. *Speech Communication*, 44(1-4) :63–82, October 2004.

[19] S.E.G. Öhman. Numerical model of coarticulation. *JASA*, 41(2) :310–320, 1967.

[20] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[21] L. D. Rosenblum, J. A. Johnson, and H. M. Saldaña. Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research*, 39(6) :1159–1170, 1996.

[22] L. D. Rosenblum and H. M. Saldaña. Time-varying information for visual speech perception. In R. Campbell, B. Dodd, and D. Burnham, editors, *Hearing by eye : Part 2, The Psychology of Speechreading and Audiovisual Speech*, pages 61–81. Earlbaum : Hillsdale, 1998.

[23] L. D. Rosenblum, L. D. Schumuckler, and J. A. Johnson. The McGurk effet in infants. *Perception & Psychophysics*, 59(3) :347–357, 1997.

[24] S. Runeson and G. Frykholm. Visual perception of lifted weight. *Journal of Experimental Psychology : Human Perception and Performance*, 7 :733–740, 1981.

[25] S. Runeson and G. Frykholm. Kinematic specification of dynamics as an informational basis for person and action perception : Expectation, gender recognition and deceptive intention. *Journal of Experimental Psychology : General*, 112 :585–615, 1983.

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

Chapitre 2

[1] V. Aubergé. *La synthèse de la parole : des règles aux lexiques*. PhD thesis, Université Pierre Mendès France, Grenoble, France, 1991.

[2] K. Bartkova and C. Sorin. A model of segmental duration for speech synthesis in French. *Speech Communication*, 6 :245–260, 1987.

[3] F. Beaugendre. Modèles de l'intonation pour la synthèse. In H. Méloni, editor, *Fondements et perspectives en traitement automatique de la parole*, chapter Production et synthèse de la parole, pages 97–107. AUPELF UREF, 1996.

[4] R. Belrhali, V. Aubergé, and L.-J. Boë. From lexicon to rules : towards a descriptive method of French text-to-phonetics transcription. In *ICSLP*, pages 1183–1186, 1989.

[5] J. Benello, A.W. Mackie, and J.A. Anderson. Syntactic category disambiguation with neural networks. *Computer Speech and Language*, (3) :203–217., 1989.

[6] R. Boite, T. Dutoit, J. Hancq, H. Leich, and H. Bourlard. *Traitement de la Parole*. Presses Polytechniques Universitaires Romandes, 2000.

[7] B. Bozkurt, T. Dutoit, and M. Bagein. From MBROLA to NU-MBROLA. In *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, pages 127–130, 2001.

[8] A. P. Breen and P. Jackson. Using f0 within a phonologically motivated method of unit selection. In *ICSLP*, 1998.

[9] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[10] Calliope. *La Parole et son Traitement Automatique*. 1989.

[11] D. T. Chappell and J H. L. Hansen. Spectral smoothing for concatenative speech synthesis. In *International Conference on Spoken Language Processing*, pages 1935–1938, Sydney, Australia, November 1998.

[12] D. T. Chappell and J. H. L. Hansen. Spectral smoothing for speech segment concatenation. *Speech Communication*, 36(3-4) :343–373, March 2002.

[13] F. J. Charpentier and M. G. Stella. Diphone synthesis using an overlap-add technique for speech wavforms concatenation. In *ICASSP*, pages 2015–2018, Tokyo, Japan, 1986.

[14] C. d'Alessandro, M. Garnier, and P. Boula de Mareüil. Synthèse de la parole à partir du texte. In H. Méloni, C. d'Alessandro, J.-P. Haton, G. Perennou, and J.-P. Tubach, editors, *Fondements et perspectives en traitement automatique de la parole*, pages 81–96. AUPELF UREF, 1996.

[15] R. Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge University, 1996.

[16] T. Dutoit and H. Leich. MBR-PSOLA : Text-to-speech synthesis based on an MBE resynthesis of the segments database. *Speech Communication*, 13 :435–440, 1993.

[17] F. Emerard. *Synthèse par diphones et traitement de la prosodie*. PhD thesis, Université des langues et lettres de Grenoble, 1977.

[18] A. Falaschi, M Giustiniani, and M. Verola. A Hidden Markov Model approach to speech synthesis. In *EUROSPEECH*, pages 187–190, 1989.

[19] F. Fallside and A. Ljolje. Synthesis of natural sounding pitch contours in isolated utterance using hidden markov models. In *IEEE Trans. on ASSP*, volume 34, pages 1074–1080, 1986.

[20] H. Fujisaki. The role of quantitative modeling in the study of intonation. In *Proceedings of the International Symposium on Japanese Prosody*, pages 163–174, 1992.

[21] H. Fujisaki and H. Keikichi. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, 5(4) :233–241, 1984.

[22] J. H. L. Hansen and D. T. Chappell. An auditory-based distortion measure with application to concatrenative speech synthesis. *IEEE Transactions on speech and audio processing*, 6(5) :489–495, 1998.

[23] B. Holm and G. Bailly. Generating prosody by superposing multi-parametric overlapping contours. In *International Conference on Speech and Language Processing*, pages 203–206, Beijing, China, 2000.

[24] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, pages 373–376, Atlanta, GA, 1996.

[25] F. Jelinek. Up from trigrams ! In *Eurospeech*, pages 1037–1040, 1991.

[26] D. H. Klatt. Synthesis by rule of segmental durations in english sentences. In B. Lindblom and S. Ohlman, editors, *Frontiers of speech communication research*, pages 287–300. Academic Press, 1979.

[27] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67 :971–995, 1980.

[28] J. Kupiec. Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language*, (6) :225–242., 1992.

[29] D. Larreur, F. Emerard, and F. Marty. Linguistic and prosodic processing for a text-to-speech synthesis system. In *Eurospeech*, pages 510–513, Paris, France, 1989.

[30] K.-S. Lee and S.-R. Kim. Context-adaptative smoothing for concatenative speech synthesis. In *IEEE Signal Processing Letters*, volume 9, pages 422–425, 2002.

[31] M. Lee, D. P. Lopresti, and J. P. Olive. A text-to-speech platform for variable length optimal unit searching using perception based cost functions. *International Journal of Speech Technology*, 6 :347–356, 2003.

[32] M. W. Macon, A. E. Cronk, and J. Wouters. Generalization and discrimination in tree-structured unit selection. In *Proceedings of the 3rd ESCA/COCOSDA International Speech Synthesis Workshop*, 1998.

[33] H. Mixdorff and O. Jokisch. Building an integrated prosodic model of German. In *European Conference on Speech Communication and Technology*, pages 947–950, 2001.

[34] Y. Morlec, V. Aubergé, and G. Bailly. Evaluation of automatic generation of prosody with a superposition model. In *International Congress of Phonetic Sciences*, volume 4, pages 224–227, 1995.

[35] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 :453–468, 1990.

[36] S.E.G. Öhman. Coarticulation in VCV utterances : Spectrographic measurements. *J Acoust Soc Am.*, 39(1) :151–168, 1966.

[37] D. O'Shaughnessy. A study of French vowel and consonant durations. *Journal of phonetics*, 9 :385–406, 1981.

[38] J. Pierrehumbert. Synthesizing intonation. *Journal of the Acoustical Society of America*, 70 :985–995, 1981.

[39] R. Prudon and C. d'Alessandro. A selection/concatenation text-to-speech synthesis system : databases development, system design, comparative evaluation. In *4th ISCA ITRW on Speech Synthesis*, 2001.

[40] H.B. Richards and J.S. Bridle. The HDM : a segmental hidden dynamical model of coarticulation. In *ICASSP*, 1999.

[41] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform units. In *ICASSP*, pages 679–682, 1988.

[42] Y. Sagisaka. On the prediction of global F0 shpaes for Japanese text-to-speech. In *ICASSP*, volume 1, pages 325–328, 1990.

[43] F. Tesser, P. Cosi, C. Drioli, and G. Tisato. Prosodic data-driven modelling of narrative style in Festival TTS. In *ISCRA Workshop on Speech Synthesis*, 2004.

[44] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis. *Speech Communication*, 48 :45–56, 2006.

[45] C. Traber. F0 generation with a database of natural F0 patterns and with a neural network. In G. Bailly and C. Benoît, editors, *Talking Machines : Theories, Models and Designs*. North Holland, 1992.

[46] J. Trouvain, W. J. Barry, C. Nielsen, and O. Andersen. Implication of energy declination for speech synthesis. In *ETRW Workshop on speech synthesis*, pages 47–52, 1998.

[47] J. Vaissière. *Contribution à la synthèse par règles du français*. PhD thesis, Université des langues et lettres de Grenoble, 1971.

[48] J. van Santen. *Progress in Speech Synthesis*. SPRINGER VERLAG, 1996.

[49] J. van Santen. Segmental duration and speech timing. In Y. Sagisaka, W. N. Campbell, and N. Higuchi, editors, *Computing Prosody*. New York : Springer, 1996.

[50] J. van Santen. Prosodic modelling in text-to-speech synthesis. In *Proceedings of Eurospeech 1997*, Rhodos, Greece, 1997.

Chapitre 3

[1] G.A. Abrantes and F. Pereira. MPEG-4 facial animation technology : Survey, implementation, and results. *IEEE Transactions on circuits and systems for video technology*, 9(2) :290–305, 1999.

[2] E. Agelfors, J. Beskow, B. Granström, M. Lundeberg, G. Salvi, K.-E. Spens, and T. Öhman. Synthetic visual speech driven from auditory speech. In *AVSP*, 1999.

[3] I. Albrecht, J. Haber, and H.-P. Seidel. Speech synchronization for physics-based facial animation. In *WSCG*, pages 9–16, 2002.

[4] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

[5] G. Bailly, M. Bérar, F. Elisei, and M. Odisio. Audiovisual speech synthesis. *International Journal of Speech Technology*, 6 :331–346, 2003.

[6] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

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

[8] C. Benoît, A. Fuster-Duran, and B. Le Goff. An investigation of hypo- and hyper-speech in the visual modality. In *ETRW on Speech Production : from Control Strategies to Acoustics*, pages 237–240, Autrans, France, 1996.

[9] C. Benoît and B. Le Goff. Audio-visual speech synthesis from French text : Eight years of models, designs and evaluation at the ICP. *Speech Communication*, 26 :117–129, 1998.

[10] C. Benoît, T. Mohamadi, and S. Kandel. Audio-visual intelligibility of French speech in noise. *Journal of Speech and Hearing Research*, 37 :1195–1203, 1994.

[11] C. Benoît and L. C. W. Pols. *Talking Machines : Theories, Models and Designs.*, chapter On the assessment of synthetic speech, pages 435–441. Bailly, G. and Benoît, C. and Sawallis, T. R., Amsterdam, 1992.

[12] J. Beskow. Rule-based visual speech synthesis. In *EUROSPEECH'95*, volume 1, pages 299–302, Madrid, Spain, September 1995.

[13] J. Beskow. Talking heads - communication, articulation and animation. In *Proceedings of Fonetik '96, Swedish Phonetics Conference*, pages 53–56, 1996.

[14] J. Beskow. *Talking Heads Models and Applications for multimodal speech synthesis.* PhD thesis, Department of speech, Music and Hearing, KTH, 2003.

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

[16] J. Beskow, M. Dahlquist, B. Granstrom, M. Lundeberg, K.-E. Spens, and T. Ohman. The Teleface project - multimodal speech communication for the hearing impaired. In *Eurospeech*, pages 2003–2010, Rhodos, Greece, 1997.

[17] E. Bevacqua and C. Pelachaud. Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 15 :297–304, 2004.

[18] C. Bonamico, C. Braccini, M. Costa, F. Lavagetto, and R. Pockaj. Using MPEG-4 parameters for calibrating/animating talking heads. In *Tyrrhenian International Workshop on Digital Communications*, 2002.

[19] M. E. Brand. Voice puppetry. In *ACM SIGGRAPH*, 1999.

[20] A. P. Breen, E. Bowers, and W. Welsh. An investigation into the generation of mouth shapes for a talking head. In *ICSLP*, pages 2159–2162, 1996.

[21] C. Bregler, M. Cowell, and M. Slaney. Videorewrite : driving visual speech with audio. In *SIGGRAPH'97*, pages 353–360, Los Angeles, CA, 1997.

[22] N. Brooke and S. D. Scott. Two and three-dimensional audio-visual speech synthesis. In *AVSP*, 1998.

[23] M. A. Cathiard. La perception visuelle de la parole : aperçu des connaissances. In *Bulletin de l'Institut de Phonétique de Grenoble*, volume 18, pages 109–193. Institut de Phonétique de Grenoble, 1989.

[24] Y.-J. Chang and T. Ezzat. Transferable videorealistic speech animation. In *ACM Siggraph/Eurographics Symposium on Computer Animation*, 2005.

[25] M.M. Cohen and D.W. Massaro. Modeling coarticulation in synthetic visual speech. In Springer-Verlag, editor, *Models and Techniques in Computer Animation*, pages 139–156. N.M. Thalmann & D. Thalmann, Tokyo, Japan, 1993.

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

[27] E. Cosatto and H.-P. Graf. Sample-based synthesis of photo-realistic talking heads. *Computer Animation*, pages 103–110, 1998.

[28] E. Cosatto and H.-P. Graf. Photo-realistic talking heads from image samples. In *IEEE Transactions on Multimedia*, volume 2, pages 152–163, 2000.

[29] D. Cosker, Marshall D., Rosin P., and Hicks Y. Video realistic talking heads using hierarchical non-linear speech-appearance models. In *Proceedinds of Mirage*, INRIA Rocquencourt, France, March 2003.

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

[31] P. Ekman and W. V. Friesen. Facial action coding system (FACS) : a technique for the measurements of facial action. *Consulting Psycholigists Press*, 1978.

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

[33] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of ACM SIGGRAPH*, San Antonio, USA, 2002.

[34] T. Ezzat and T. Poggio. Miketalk : A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference*, Philadelphia, PA, June 1998.

[35] S. Fagel and C. Clemens. Two articulation models for audiovisual speech synthesis - description and determination. In *AVSP*, pages 215–220, 2003.

[36] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11 :796–804, 1968.

[37] M. Frydrych, J. Kätsyri, M. Dobšík, and M. Sams. Toolkit for animation of finnish talking head. In *AVSP*, St Jorioz, France, 2003.

[38] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun. Analysis and synthesis of the three-dimensional movements of the head, face and hand of a speaker using Cued Speech. *Journal of the Acoustical Society of America*, 118(2), August 2005.

[39] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw. TDA : A new trainable trajectory formation system for facial animation. In *Interspeech ICSLP*, Pittsburgh, PA, September 2006.

[40] R. Gutierrez-Osuna, P.K. Kakumanu, A. Esposito, O.N. Garcia, A. Bojorquez, J.L. Castillo, and I.J. Rudomin. Speech-driven facial animation with realistics dynamics. *IEEE Transaction on Multimedia*, 7(1) :33–42, 2005.

[41] A. Hallgren and B. Lyberg. Visual speech synthesis with concatenative speech. In *AVSP*, 1998.

[42] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. In *IEEE Transactions on Speech and Audio Processing*, 2004.

[43] P. Hong, Z. Wen, and T. Huang. Real-time speech-driven 3D face animation. In *International Symposium on 3D Data Processing Visualization Transmission*, 2002.

[44] F. J. Huang and T. Chen. Real-time lip-synch face animation driven by human voice. In *IEEE Mutlimedia Signal Processing Workshop*, 1998.

[45] G. A. Kalberer, P. Müller, and L. Van Gool. Visual speech, a trajectory in viseme space. *International Journal of Imaging Systems and Technology*, 13 :74–84, 2003.

[46] S. Kshirsagar and N. Magnenat-Thalmann. Viseme space for realistic speech animation. In *AVSP*, pages 30–35, Aalborg, Denmark, 2001.

[47] S. Kshirsagar and N. Magnenat-Thalmann. Visyllable based speech animation. In *Proceedings Eurographics*, 2003.

[48] S. Kshirsagar, T. Molet, and N. Magnenat-Thalmann. Principal components of expressive speech animation. In *Proc. Computer Graphics International*, pages 38–44, 2001.

[49] P. K. Kuhl and A. N. Meltzoff. The bimodal perception of speech in infancy. *Science*, 218 :1138–1141, 1982.

[50] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson. Kinematics-based synthesis of realistic talking faces. In *AVSP*, pages 185–190, 1998.

[51] B Le Goff and C. Benoît. A text-to-audiovisual-speech synthesizer for French. In *4th International Conference on Spoken Language Processing*, pages 2163–2166, Philadelphia, USA, 1996.

[52] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *SIGGRAPH*, pages 55–62, 1995.

[53] A. Löfqvist. Speech as audible gestures. *Speech Production and Speech Modeling*, pages 289–322, 1990.

[54] J. C. Lucero, S. T. R. Maciel, D. A. Johns, and K. G. Munhall. Empirical modeling of human face kinematics during speech using motion clustering. *Journal of the Acoustical Society of America*, 118(1) :405–409, 2005.

[55] J. C. Lucero and K. G. Munhall. A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America*, 106(5) :2834–2842, 1999.

[56] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise. Accurate automatic visible speech synthesis of abitrary 3D models based on concatenation of diviseme motion capture data. *Computer Animation and Virtual Worlds*, 15 :485–500, 2004.

[57] K. MacKain, M. Studdert-Kennedy, S. Spieker, and D. Stern. Infant intermodal perception speech perception is a left hemisphere function. *Science*, 219 :1347–1349, 1983.

[58] N. Magnenat-Thalmann and D. Thalmann. *Handbook of Virtual Humans*. John Wiley & Sons Ltd, 2004.

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

[60] D.W. Massaro, J. Beskow, M.M. Cohen, C.L. Fry, and T. Rodriguez. Picture my voice : audio to visual speech synthesis using artificial neural networks. *Proceedings of AVSP*, 1999.

[61] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 26 :746–748, 1976.

[62] S. Minnis and A. P. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *International Conference on Speech and Language Processing*, pages 759–762, Beijing, China, 2000.

[63] R. Möttönen, J.-L. Olivès, J. Kulju, and M. Sams. Parametrized visual speech synthesis and its evaluation. In *European Signal Processing Conference*, Tampere, Finland, 2000.

[64] MPEG Video and SNHC. Text of ISO/IEC FDIS 14496-2 : Visual. In *MPEG Meeting*, 1998.

[65] S.E.G. Öhman. Coarticulation in VCV utterances : Spectrographic measurements. *J Acoust Soc Am.*, 39(1) :151–168, 1966.

[66] S.E.G. Öhman. Numerical model of coarticulation. *JASA*, 41(2) :310–320, 1967.

[67] T. Okadome, T. Kaburagi, and M. Honda. Articulatory movement formation by kinematic triphone model. In *IEEE International Conference on System Man and Cybernetics*, pages 469–474, 1999.

[68] J.-L. Olivès, R. Möttönen, J. Kulju, and M. Sams. Audio-visual speech synthesis for finnish. In *Audio-Visual Speech Processing*, Santa Cruz, USA, August 1999.

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

[70] J. Ostermann, M. Beutnagel, A. Fischer, and Y. Wang. Integration of talking heads and text-to-speech synthesizers for visual TTS. In *ICSLP'98*, Sydney, Australia, December 1998.

[71] S. Ouni, M. M. Cohen, and D. W. Massaro. Training Baldi to be multilingual : A case study for an Arabic Badr. *Speech Communication*, 45 :115–137, 2005.

[72] I. S. Pandzic and R. Forchleimer. *MPEG-4 Facial Animation - the Standard, Implementation and Applications*. John Wiley & Sons, Chichster, England, 2002.

[73] F. I. Parke. A model for human faces that allows speech synchronised animation. *Journal of Computers and Graphics*, 1(1) :1–4, 1975.

[74] C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, and P. Cosi. Modelling an italian talking head. In *Proceedings of AVSP*, 2001.

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

[76] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[77] J. Robert-Ribes. *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles*. PhD thesis, Institut National Polytechique de Grenoble, 1995.

[78] J.-L. Schwartz. La parole multisensorielle : Plaidoyer, problèmes, perspective. In *Journées d'Etude sur la Parole*, Fès, Maroc, 2004.

[79] J.-L. Schwartz, F. Berthommier, and C. Savariaux. Seeing to hear better : evidence for early audio-visual interactions in speech identification. *Cognition*, 93 :B69–B78, 2004.

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

[81] M. Slaney. *Audiovisual Speech Processing*, chapter Image-based Facial Synthesis, pages 149–161. MIT Press, 2003.

[82] Q. Summerfield. Use of visual information for phonetic perception. *Phonetica*, 17 :3–46, 1979.

[83] Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke. Lips, teeth, and the benefits of lipreading. In A. W. Young and H. D. Ellis, editors, *Handbook of Research on Face Processing*, pages 223–233. Elsevier Science Publishers, 1989.

[84] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda. Visual speech synthesis based on parameter generation from HMM : speech-driven and text-and-speech-driven approaches. In *AVSP*, 1998.

[85] D. Terzopoulos, F. Parke, and K. Waters. Panel on facial animation : Past, present and future. In *Proceedings of SIGGRAPH'97*, 1997.

[86] D. Terzopoulos and K. Waters. Physicaly-based facial modeling, analysis and animation. *The Journal of Visual and Computer Animation*, 1 :73–80, 1990.

[87] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on pattern analysis and machine intelligence*, 15(6) :569–579, june 1993.

[88] B.-J. Theobald, Bangham J. A., Matthews I., and Cawley G. Towards video realistic synthetic visual speech. In *ICASSP*, pages 3892–3895, 2002.

[89] B.-J. Theobald, J. A. Bangham, I. Matthews, and G. Cawley. Evaluation of a talking head based on appearance models. In *Audio-Visual Speech Processing*, September 2003.

[90] B. J. Theobald, J. A. Bangham, I. Matthews, and G. C. Cawley. Visual speech synthesis using statistical models of shape and appearance. In *Proc. Auditory-Visual Speech Processing*, 2001.

[91] F. Vignoli and C. Braccini. A text-speech synchronization technique with applications to talking heads. In *AVSP'99*, Santa Cruz, CA, 1999.

[92] K. Waters. A muscle model for animating three-dimensional facial expression. *Computer Graphics*, 22(4) :17–24, 1987.

[93] C. Weiss. A framework for data-driven video-realistic audio-visual speech synthesis. In *LREC*, 2004.

[94] E. Yamamoto, S. Nakamura, and K. Shikano. Subjective evaluation for hmm-based speech-to-lip movement synthesis. In *AVSP*, pages 227–232, 1998.

[95] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. Facial animation and head motion driven by speech acoustics. In *The 5th International Seminar on Speech Production*, pages 265–268, Munich, Germany, 2000.

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

[97] G. Zorić. Real-time face animation driven by human voice. In *ConTEL*, 2003.

Chapitre 4

[1] V. Attina. *La Langue française Parlée Complétée (LPC) : Production et Perception.* PhD thesis, Institut National Polytechnique de Grenoble, 2005.

[2] V. Attina, D. Beautemps, and M.-A. Cathiard. Temporal motor organization of Cued Speech gestures in the French language. In *ICPHS*, pages 1935–1938, 2003.

[3] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. Toward an audiovisual synthesizer for Cued Speech : rules for CV French syllables. In *Audio Visual Speech Processing Workshop*, pages 227–232, St Jorioz, France, 2003.

[4] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables : rules for a Cued Speech synthesizer. *Speech Communication*, 44 :197–214, 2004.

[5] V. Attina, M.-A. Cathiard, and D. Beautemps. Contrôle de l'anticipation vocalique d'arrondissement en Langage Parlé Complété. In *Journées d'Etudes sur la Parole*, 2002.

[6] V. Attina, M.-A. Cathiard, and D. Beautemps. L'ancrage de la main sur les lèvres : Langue française Parlée Complétée et anticipation vocalique. In *Jounées d'Etude sur la Parole*, 2004.

[7] V. Attina, M.-A. Cathiard, and D. Beautemps. Temporal measures of hand and speech coordination during French cued speech production. *Lecture Notes in Artificial Intelligence*, 3881 :13–24, 2006.

[8] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

[9] L. E. Bernstein, M. E. Demorest, and P. E. Tucker. Speech perception without hearing. *Perception & Psychophysics*, 62 :233–252, 2000.

[10] M.-A. Cathiard, V. Attina, and D. Alloatti. Labial anticipation behavior during speech with and without Cued Speech. In *ICPHS*, pages 1939–1942, 2003.

[11] M.-A. Cathiard, F. Bouaouni, V. Attina, and D. Beautemps. Etude perceptive du décours de l'information manuo-faciale en langue française parlée complétée. In *Journées d'Etude sur la Parole*, 2004.

[12] R. O. Cornett. Cued Speech. *American Annals of the Deaf*, 112 :3–13, 1967.

[13] R. O. Cornett. Le Cued Speech. In *Aides manuelles à la lecture labiale et prespectives d'aides automatiques*, Centre scientifique IBM-France, Paris, France, 1982.

[14] R. O. Cornett. 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, 1988.

[15] R. O. Cornett, R. Beadles, and B. Wilson. Automatic Cued Speech. In *Research Conference on Speech Processing Aids for the Deaf*, pages 224–239, Washington, DC : Gallaudet College, 1977.

[16] R.O. Cornett. Adapting Cued Speech to additional languages. *Cued Speech Journal*, V :19–29, 1994.

[17] P. Duchnowski, L. Braida, M. Bratakos, D. Lum, M. Sexton, and J. Krause. Automatic generation of Cued Speech for the deaf : status and outlook. In *AVSP'98*, pages 161–166, Terrigal, Australia, 1998.

[18] P. Duchnowski, D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos, and L. D. Braida. Development of speechreading supplements based on automatic speech recognition. *IEEE Transactions on Biomedical Engineering*, 47(4) :487–496, 2000.

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

[20] O. Engwall and J. Beskow. Resynthesis of 3D tongue movements from facial data. In *EuroSpeech*, Geneva, 2003.

[21] J. Feldmar. Projet labiao (lecture labiale assistée par ordinateur) : Présentation de logiciels augmentant l'autonomie des sourds dans le milieu ordinaire. In *Liaison LPC*, volume 42, pages 159–163, 2005.

[22] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11 :796–804, 1968.

[23] F. Grosjean. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28 :267–283, 1980.

[24] J. Jiang, A. Alwan, L. Bernstein, P. Keating, and E. Auer. On the correlation between facial movements, tongue movements and speech acoustics. In *Proceedings of International Conference on Speech and Language Processing*, pages 42–45, Beijing, China, 2000.

[25] J. Leybaert. Phonology acquired through the eyes and spelling in deaf children. *Journal of Experimental Child Psychology*, 75 :291–318, 2000.

[26] J. Leybaert. The role of Cued Speech in language processing by deaf children : an overview. In *Auditory-Visual Speech Processing*, pages 179–186, St Jorioz, France, 2003.

[27] G. Nicholls and D. Ling. Cued Speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25 :262–269, 1982.

[28] E. Owens and B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28 :381–393, 1985.

[29] C. M. Reed. The implications of the Tadoma method of speechreading for spoken language processing. In *ICSLP*, 1996.

[30] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[31] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of Cued Speech. *Journal of Rehabilitation Research and Development*, 31 :20–41, 1994.

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

Chapitre 5

[1] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables : rules for a Cued Speech synthesizer. *Speech Communication*, 44 :197–214, 2004.

[2] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

[3] C. Benoît. An intelligibility test using Semantically Unpredictable Sentences : towards the quantification of linguistic complexity. *Speech Communication*, 9 :293–304, 1990.

[4] C. Benoît, M. Grice, and V. Hazan. The SUS test : A method for the assessment of text-to-speechs synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18 :381–392, 1996.

[5] C. Benoît, A. van Erp, M. Grice, V. Hazan, and U. Jekosch. Multilingual synthesiser assessment using semantically unpredictable sentences. In *Proceedings of Eurospeech'89*, volume 2, pages 633–636, 1989.

[6] L. E. Bernstein, M. E. Demorest, and P. E. Tucker. Speech perception without hearing. *Perception & Psychophysics*, 62 :233–252, 2000.

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

[8] J. Beskow, M. Dahlquist, B. Granstrom, M. Lundeberg, K.-E. Spens, and T. Ohman. The Teleface project - multimodal speech communication for the hearing impaired. In *Eurospeech*, pages 2003–2010, Rhodos, Greece, 1997.

[9] Calliope. *La Parole et son Traitement Automatique*. 1989.

[10] M. M. Cohen, R. L. Walker, and D. W. Massaro. Perception of synthetic visual speech. In *Speechreading by Man and Machine : Models, Systems and Application, NATO Advanced Study Institute 940584*, Chateau de Bonas, France, August 1995.

[11] D. Cosker, S. Paddock, D. Marshall, P. L. Rosin, and S. Rushton. Towards perceptually realistic talking heads : Models, methods and McGurk. In *ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, California, USA, 2004.

[12] P. Duchnowski, L. Braida, M. Bratakos, D. Lum, M. Sexton, and J. Krause. Automatic generation of Cued Speech for the deaf : status and outlook. In *AVSP'98*, pages 161–166, Terrigal, Australia, 1998.

[13] P. Duchnowski, L. Braida, D. Lum, M. Sexton, J. Krause, and S. Banthia. A speechreading aid based on phonetic ASR. In *5th International Conference on Spoken Language Processing*, volume 7, pages 3289–3292, Sydney, Australia, 1998.

[14] P. Duchnowski, D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos, and L. D. Braida. Development of speechreading supplements based on automatic speech recognition. *IEEE Transactions on Biomedical Engineering*, 47(4) :487–496, 2000.

[15] T. Dutoit and H. Leich. Synthèse de parole de haute qualité à partir d'un texte : une comparaison de quatre algorithmes candidats. In *Journées d'Etudes sur la Parole*, 1994.

[16] G. Fairbanks. Test of phonemic differentiation : the rhyme test. *Journal of the Acoustical Society of America*, 30(7) :596–600, July 1958.

[17] A. Faulkner and S. Rosen. The contribution of temporally-coded acoustic speech patterns to audio-visual speech perception in normally hearing and profoundly hearing-impaired listeners. In *Proceedings of the Workshop on the Auditory Basis of Speech Perception*, pages 261–264, 1996.

[18] G. Geiger, T. Ezzat, and T. Poggio. Perceptual evaluation of video-realistic speech. Technical Report CBCL Paper 224/ AI Memo, Massachusetts Institute of Technology, Cambridge, USA, February 2003.

[19] D. Gibbon, R. Moore, and R. Winski. *Handbook of standards and resources for spoken language systems*. Mouton de Gruyter, 1997.

[20] M. Goldstein. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication*, 16(3) :225–244, 1995.

[21] A. S. House, C. E. Williams, M. H. Hecker, and K. D. Kryter. Articulation testing methods : Consonantal differantiation with a closed-response set. *Journal of the Acoustical Society of America*, 37 :158–166, 1965.

[22] B. Le Goff, T. Guiard-Marigny, and C. Benoit. *Progress in speech synthesis*, chapter Analysis-synthesis and intelligibility of a talking face, pages 235–246. Van Santen, J.P.H. and Sproat, R.W. and Olive, J.P. and Hischberg, J., Berlin, 1996.

[23] R. Möttönen, J.-L. Olivès, J. Kulju, and M. Sams. Parametrized visual speech synthesis and its evaluation. In *European Signal Processing Conference*, Tampere, Finland, 2000.

[24] L. Neovius and P. Raghavendra. Comprehension of KTH text-to-speech with l̀istening speedṕaradigm. In *EUROSPEECH'93*, pages 1687–1690, Berlin, Germany, 1993.

[25] G. Nicholls and D. Ling. Cued Speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25 :262–269, 1982.

[26] M. Odisio. *Estimation des mouvements du visage d'un locuteur dans une séquence audiovisuelle*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.

[27] I. Pandzig, J. Ostermann, and D. Millen. Users evaluation : synthetic talking faces for interactive services. *The Visual Computer*, 15 :330–340, 1999.

[28] J. P. Peckels and M. Rossi. Le test de diagnostic par paires minimales. *Revue d'Acoustique*, 27 :245–262, 1973.

[29] D. B. Pisoni, B. G. Greene, and J. S. Logan. An overview of ten years of research on the perception of synthetic speech. In *ESCA Workshop on Speech Input/Output Assessment and Speech Databases*, 1989.

[30] D. B. Pisoni and S. Hunnicutt. Perceptual evaluation of MITALK : the MIT unrestricted text-to-speech system. *ICASSP*, pages 572–575, 1980.

[31] D. B. Pisoni, H. C. Nusbaum, and B. G. Greene. Perception of synthetic speech generated by rule. *IEEE*, 73 :1665–1676, 1985.

[32] R. Prudon and C. d'Alessandro. A selection/concatenation text-to-speech synthesis system : databases development, system design, comparative evaluation. In *4th ISCA ITRW on Speech Synthesis*, 2001.

[33] C. Siciliano, A. Faulkner, and G. Williams. Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners. In *ISCA Tutorial and Research Workshop on Audio Visual Speech Processing*, St Jorioz, France, 2003.

[34] C. Siciliano, G. Williams, J. Beskow, and A. Faulkner. Evaluation of a synthetic talking face as a communication aid for the hearing impaired. *Speech, Hearing and Language : Work in Progress*, 14 :51–61, 2002.

[35] C. Sorin and F. Emerard. Domaines d'application et évaluation de la synthèse de parole à partir du texte. In *Fondements et perspectives en traitement automatique de la parole*. AUPELF UREF, 1996.

[36] M. F. Spiegel, M. J. Altom, and M. J. Macchi. Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. *Speech Communication*, 9 :279–291, 1990.

[37] C. Stevens, N. Lees, J. Vonwiller, and D. Burnham. On-line experimental methods to evaluate text-to-speech (TTS) synthesis : effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language*, 19 :129–146, 2005.

[38] B.-J. Theobald, J. A. Bangham, I. Matthews, and G. Cawley. Evaluation of a talking head based on appearance models. In *Audio-Visual Speech Processing*, September 2003.

[39] B. J. Theobald, J. A. Bangham, I. A. Matthews, and G. C. Cawley. Near-videorealistic synthetic talking faces : implementation and evaluation. *Speech Communication*, 44 :127–140, 2004.

[40] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of Cued Speech. *Journal of Rehabilitation Research and Development*, 31 :20–41, 1994.

[41] R. Van Bezooijen and L. C. W. Pols. Evaluating text-to-speech systems : some methodological aspects. *Speech Communication*, 9 :263–270, 1990.

[42] J. Van Santen. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech & Language*, 7 :49–100, 1993.

[43] H. S. Venkatagiri. Segmental intelligibility of four currently used text-to-speech synthesis methods. *Journal of the Acoustical Society of America*, 113(4) :2095–2014, April 2003.

[44] D. W. Voiers. *The Diagnostic Rhyme Test*. PhD thesis, TRACOR, 1970.

Chapitre 7

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

Chapitre 8

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

[2] G. Bailly, M. Bérar, F. Elisei, and M. Odisio. Audiovisual speech synthesis. *International Journal of Speech Technology*, 6 :331–346, 2003.

[3] R. Bowden. *Learning non-linear Models of Shape and Motion*. PhD thesis, Dept Systems Engineering, Brunel University, Uxbridge, Middlesex, UK, 2000.

[4] M. Bray, E. Koller-Meier, P Müller, L. Van Gool, and N. N. Schraudolph. 3D hand tracking by rapid stochastic gradient descent using a skinning model. In *First European Conference on Visual Media Production*, pages 59–68, 2004.

[5] R. Cipolla, B. Stenger, A. Thayananthan, and P. H. S. Torr. Hand tracking using a quadric surface model and bayesian filtering. In *The British Machine Vision Conference*, 2001.

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

[7] P. Kalra, N. Magnenat-Thalmann, L. Moccozet, G. Sannier, A. Aubel, and D. Thalmann. Real-time animation of realistic virtual humans. *IEEE Computer Graphics and Applications*, 18(5) :42–55, 1998.

[8] R. Mas Sanso and D. Thalmann. A hand control and automatic grasping system for synthetic actors. In *Proc. Eurographics '94*, 1994.

[9] H. Ouhaddi and P. Horain. Conception et ajustement d'un modèle 3D articulé de la main. In *Actes des 6èmes Journées du Groupe de Travail Réalité Virtuelle*, 1998.

[10] M Preda, T. Zaharia, and F. Preteux. 3D body animation and coding within a MPEG-4 compliant framework. In *Proceedings International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging*, 1999.

[11] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[12] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. Large vocabulary continuous speech recognition using HTK. In *Proceedings ICASSP'94*, 1994.

[13] Y. Wu, J. L. Lin, and T. S. Huang. Capturing natural hand articulation. In *Proceedings of IEEE International Conference on Computer Vision*, Canada, 2001.

Chapitre 9

[1] V. Attina. *La Langue française Parlée Complétée (LPC) : Production et Perception.* PhD thesis, Institut National Polytechnique de Grenoble, 2005.

[2] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables : rules for a Cued Speech synthesizer. *Speech Communication*, 44 :197–214, 2004.

[3] P. Boyes Braem. Rhythmic Temporal Patterns in the Signing of Deaf Early and Mate Learners of Swiss German Sign Language. *Language and Speech*, 42(2–3) :177–208, 1999.

[4] M. Costa, T. Chen, and F. Lavagetto. Visual prosody analysis for realistic motion synthesis of 3D head models. In *International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging*, 2001.

[5] M. Dohen. *Deixis prosodique multisensorielle : production et perception audiovisuelle de la focalisation contrastive en français.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, December 2005.

[6] H.P. Graf, E. Cosatto, V. Strom, and F.J. Huang. Visual prosody : facial movements accompanying speech. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[7] B. Granström and D. House. Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46 :473–484, 2005.

[8] U. Hadar, T.J. Steiner, E.C. Grant, and F. Clifford Rose. Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26 :117–129, 1983.

[9] B. Holm. *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application à l'énonciation de formules mathématiques.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2003.

[10] B. Holm and G. Bailly. SFC : a trainable prosodic model. *Speech Communication*, 46(3–4) :348–364, 2005.

[11] E. Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32 :855–878, 2000.

[12] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. Head movements improves auditory speech perception. *Psychological Science*, 15(2) :133–137, 2004.

[13] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

Chapitre 10

[1] M. Alissali. *Architecture logicielle pour la synthèse multilingue de la parole.* PhD thesis, INPG, Grenoble, France, 1993.

[2] M. Alissali and G. Bailly. Compost : a client-server model for applications using text-to-speech. In *European Conference on Speech Communication and Technology*, pages 2095–2098, Berlin, Germany, 1993.

[3] G. Bailly and M. Alissali. Compost : a server for multilingual text-to-speech system. *Traitement du Signal*, 9(4) :359–366, 1992.

[4] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

[5] G. Bailly, B. Holm, and V. Aubergé. A trainable prosodic model : learning the contours implementing communicative functions within a superpositional model of intonation. In *International Conference on Speech and Language Processing*, pages 1425–1428, Jeju, Korea, 2004.

[6] G. Bailly and A. Tran. Compost : a rule-compiler for speech synthesis. In *European Conference on Speech Communication and Technology*, pages 136–139, 1989.

[7] N. Campbell. Computing prosody : Computational models for processing spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Synthesizing Spontaneous Speech*, pages 165–186. Springer-Verlag, 1997.

[8] B. Holm. *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application à l'énonciation de formules mathématiques.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2003.

[9] B. Holm and G. Bailly. SFC : a trainable prosodic model. *Speech Communication*, 46(3–4) :348–364, 2005.

[10] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, pages 373–376, Atlanta, GA, 1996.

[11] S. Minnis and A. P. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *International Conference on Speech and Language Processing*, pages 759–762, Beijing, China, 2000.

[12] K. Takeda, K. Abe, and Y. Sagisaka. On the basic scheme and algorithms in non-uniform units speech synthesis. In G. Bailly and C. Benoît, editors, *Talking machines : Theories, Models and Designs*, pages 93–105. Elsevier B.V., 1992.

Chapitre 11

[1] M. Bérar, G. Bailly, M. Chabanas, M. Desvignes, F. Elisei, M. Odisio, and Y. Pahan. *Towards a better understanding of speech production processes*, chapter Towards a generic talking head, pages 341–362. Psychology Press, New York, 2006.

[2] M. Odisio and F. Elisei. Clonage 3D et animation articulatoire du visage d'une personne réelle pour la communication parlée audiovisuelle. In *Journées de l'AFIG*, pages 225–232, Grenoble, France, 2000.

[3] L. Revéret, G. Bailly, and P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[4] L. Revéret and C. Benoît. A new 3D lip model for analysis and synthesis of lip motion in speech production. In *Auditory-Visual Speech Processing Workshop*, 1998.

Chapitre 13

[1] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

[2] D. Decotigny. *Une infrastructure de simulation modulaire pour l'évaluation de performances de systèmes temps-réel*. PhD thesis, Université de Rennes 1, 2003.

[3] I. Pandzig, J. Ostermann, and D. Millen. Users evaluation : synthetic talking faces for interactive services. *The Visual Computer*, 15 :330–340, 1999.

[4] J. A. Stankovic. Misconceptions about real-time computing. *IEEE Computer*, 21(10) :10–19, 1988.

Chapitre 14

[1] C. Abry and L.-J. Boë. "Laws" for lips. *Speech Communication*, 5 :97–104, 1986.

[2] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry. A set of visual French visemes for visual speech synthesis. In G. Bailly, Ch. Benoît, and T. R. Sawallis, editors, *Talking Machines : theories, models and designs*, pages 485–504. Elsevier Science, Amsterdam, NL, 1992.

[3] J. Hazen, T., K. Saenko, C.-H. La, and J. R. Glass. A segment-based audio-visual speech recognizer : Data collection, development and initial experiments. In *Proceedings of the International Conference on Multimodal Interfaces*, State College, Pennsylvania, October 2004.

[4] T. J. Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(3) :1082–1089, May 2006.

[5] B. Mak and E. Barnard. Phone clustering using the Bhattacharyya distance. In *Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 2005–2008, Philadelphia, 1996.

[6] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, and P. Escudier. Complementarity and synergy in bimodal speech : Auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America*, 103(6) :3677–3689, June 1998.

Chapitre 15

[1] D. Busquet and C. Descourtieux. *T.E.R.M.O. Tests d'Evaluation de la Réception du message Oral par l'enfant sourd à destination des professionnels de la surdité.* 2003.

[2] G. Fairbanks. Test of phonemic differentiation : the rhyme test. *Journal of the Acoustical Society of America*, 30(7) :596–600, July 1958.

[3] G. Nicholls and D. Ling. Cued Speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25 :262–269, 1982.

[4] J. P. Peckels and M. Rossi. Le test de diagnostic par paires minimales. *Revue d'Acoustique*, 27 :245–262, 1973.

[5] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of Cued Speech. *Journal of Rehabilitation Research and Development*, 31 :20–41, 1994.

Chapitre 16

[1] V. Attina. *La Langue française Parlée Complétée (LPC) : Production et Perception.* PhD thesis, Institut National Polytechnique de Grenoble, 2005.

[2] G. Bailly, V. Attina, C. Baras, P. Bas, S. Baudry, D. Beautemps, R. Brun, J.-M. Chassery, F. Davoine, F. Elisei, G. Gibert, L. Girin, D. Grison, J.-P. Léoni, J. Liénard, N. Moreau, and P. Nguyen. ARTUS : calcul et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue française Parlée Complétée. In *Handicap 2006*, Paris, France, 2006.

[3] G. Bailly, F. Elisei, P. Badin, and C. Savariaux. Degrees of freedom of facial movements in face-to-face conversational speech. In *International Workshop on Multimodal Corpora*, pages 33–36, Genoa, Italy, 2006.

[4] G. Bailly, F. Elisei, and S. Raidt. Virtual talking heads and ambiant face-to-face communication. In A. Esposito, E. Keller, M. Marinaro, and M. Bratanic, editors, *The fundamentals of verbal and non-verbal communication and the biometrical issue.* IOS Press BV, Amsterdam, NL, 2006.

[5] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw. TDA : A new trainable trajectory formation system for facial animation. In *Interspeech ICSLP*, Pittsburgh, PA, September 2006.

[6] J. E. Harkins and M. Bakke. Technologies for communication : Status and trends. In M. Marschark and P. E. Spencer, editors, *Deaf studies, Language, and Education*, chapter Hearing and Speech Perception, pages 406–419. Oxford University Press, New York, 2003.

[7] F. Maurel, N. Vigouroux, M. Raynal, and B. Oriola. Contribution of the transmodality concept to improve web accessibility. In M. Mokhtari, editor, *Independent Living For Persons With Disabilities and Elderly People*, volume 12 of *Assistive Technology Research Series*, pages 186–193. IOS Press, 2003.

[8] D. McNeill, E. T. Levy, and L. L. Pedelty. Speech and gesture. In G. R. Hammond, editor, *Advances in psychology : cerebral control of speech and limb movements*, pages 203–256. Elsevier/North Holland Publishers, Amsterdam, 1990.

Liste des publications

Références

[1] G. Bailly, **Gibert, G.**, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, pages 27–30, Santa Monica, CA, 2002.

[2] **Gibert, G.**, G. Bailly, F. Elisei, D. Beautemps, and R. Brun. Evaluation of a speech cuer : from motion capture to a concatenative text-to-Cued Speech system. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2123–2126, Lisboa, Portugal, 2004.

[3] **Gibert, G.**, G. Bailly, F. Elisei, D. Beautemps, and R. Brun. Mise en oeuvre d'un synthétiseur 3D de Langage Parlé Complété. In *Journées d'Etude sur la Parole*, pages 245–248, Fès, Maroc, 2004.

[4] **Gibert, G.**, G. Bailly, F. Elisei, D. Beautemps, and R. Brun. Audiovisual Text-to-Cued Speech Synthesis. In *EUSIPCO*, pages 1007–1010, Vienna, Austria, 2004.

[5] **Gibert, G.**, G. Bailly, and F. Elisei. Audiovisual Text-to-Cued speech synthesis. In *5th ISCA Speech Synthesis Workshop*, pages 85–90, Pittsburgh, USA, 2004.

[6] **Gibert, G.**, G. Bailly, D. Beautemps, F. Elisei, and R. Brun. Analysis and synthesis of the three-dimensional movements of the head, face and hand of a speaker using Cued Speech. *Journal of the Acoustical Society of America*, 118(2) :1144–1153, August 2005.

[7] F. Elisei, G. Bailly, **Gibert, G.**, and R. Brun. Capturing data and realistic 3D models for Cued Speech analysis and audiovisual synthesis. In *Auditory-Visual Speech Processing Workshop*, British Columbia, Canada, 2005.

[8] **Gibert, G.**, G. Bailly, and F. Elisei. Evaluation d'un système de synthèse 3D de Langue française Parlée Complétée. In *Journées d'Etude sur la Parole*, pages 495–498, Dinard, France, 2006.

[9] G. Bailly, V. Attina, C. Baras, P. Bas, S. Baudry, D. Beautemps, R. Brun, J.-M. Chassery, F. Davoine, F. Elisei, **Gibert, G.**, L. Girin, D. Grison, J.-P. Léoni, J. Liénard, N. Moreau, and P. Nguyen. ARTUS : calcul et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue française Parlée Complétée. In *HANDICAP*, Paris, France, 2006.

[10] **Gibert, G.**, G. Bailly, and F Elisei. Evaluation of a virtual speech cuer. In *Workshop on Experimental Linguistics*, Athens, Greece, 2006.

[11] **Gibert, G.** *Conception et évaluation d'un système de synthèse 3D de Langue française Parlèe (LPC) Complétée à partir du texte.* PhD thesis, Institut National Polytechnique de Grenoble, April 2006.

[12] **Gibert, G.**, G. Bailly, and F Elisei. Evaluating a virtual speech cuer. In *International Conference on Speech and Language Processing*, Pittsburgh, PA, September 2006.

Table des figures

Liste des tableaux

Table des matières

RÉSUMÉ

Cette thèse traite de la mise en oeuvre d'un système de synthèse 3D de parole audiovisuelle capable, à partir d'une simple chaîne phonétique, de générer un signal audio synthétique, les mouvements du visage correspondant ainsi que les mouvements de la main reproduisant les gestes de la Langue française Parlée Complétée (LPC). Nous avons enregistré les mouvements faciaux et manuels d'une codeuse LPC par une technique de *motion capture*, ainsi que le signal audio correspondant, lors de la production d'un corpus de 238 phrases couvrant l'ensemble des diphones du français. Après traitements et analyses des données, nous avons implémenté un système de synthèse par concaténation d'unités en deux étapes capable de générer de la parole codée. Enfin, nous avons évalué notre système tant au niveau de l'intelligibilité segmentale qu'au niveau de la compréhension. Les résultats sont prometteurs et montrent clairement un apport d'information du code de synthèse.

MOTS CLÉS

Synthèse de parole audiovisuelle, Langue française Parlée Complétée (LPC), Capture de mouvements, Animation 3D, Évaluation.

TITLE

Text-to-Cued Speech synthesizer : from implementing to evaluating.

ABSTRACT

This thesis deals with the implementation of a complete 3D text-to-Cued Speech synthesizer : from a text input, we generate the facial and manual movements and the audio corresponding to the French Cued Speech transcription. We recorded the trajectories of flesh points on the face and the hand during the production of a corpus of sentences designed to cover all French diphones by a cuer. After processing and analysis, we implement a system able to generate French Cued Speech from a text input. Finally, we evaluate our system according to different methods. Evaluations show promising results in terms of intelligibility and understanding.

KEYWORDS

Text-to-Audiovisual Speech, Cued Speech, Motion capture, 3D Animation, Evaluation.

Thèse préparée à l'Institut de la Communication Parlée
Institut National Polytechnique de Grenoble
46, avenue Félix Viallet, 38031 Grenoble CEDEX 1, France

RÉSUMÉ

Cette thèse traite de la mise en oeuvre d'un système de synthèse 3D de parole audiovisuelle capable, à partir d'une simple chaîne phonétique, de générer un signal audio synthétique, les mouvements du visage correspondant ainsi que les mouvements de la main reproduisant les gestes de la Langue française Parlée Complétée (LPC). Nous avons enregistré les mouvements faciaux et manuels d'une codeuse LPC par une technique de *motion capture*, ainsi que le signal audio correspondant, lors de la production d'un corpus de 238 phrases couvrant l'ensemble des diphones du français. Après traitements et analyses des données, nous avons implémenté un système de synthèse par concaténation d'unités en deux étapes capable de générer de la parole codée. Enfin, nous avons évalué notre système tant au niveau de l'intelligibilité segmentale qu'au niveau de la compréhension. Les résultats sont prometteurs et montrent clairement un apport d'information du code de synthèse.

MOTS CLÉS

Synthèse de parole audiovisuelle, Langue française Parlée Complétée (LPC), Capture de mouvements, Animation 3D, Évaluation.

TITLE

Text-to-Cued Speech synthesizer : from implementing to evaluating.

ABSTRACT

This thesis deals with the implementation of a complete 3D text-to-Cued Speech synthesizer : from a text input, we generate the facial and manual movements and the audio corresponding to the French Cued Speech transcription. We recorded the trajectories of flesh points on the face and the hand during the production of a corpus of sentences designed to cover all French diphones by a cuer. After processing and analysis, we implement a system able to generate French Cued Speech from a text input. Finally, we evaluate our system according to different methods. Evaluations show promising results in terms of intelligibility and understanding.

KEYWORDS

Text-to-Audiovisual Speech, Cued Speech, Motion capture, 3D Animation, Evaluation.

Thèse préparée à l'Institut de la Communication Parlée
Institut National Polytechnique de Grenoble
46, avenue Félix Viallet, 38031 Grenoble CEDEX 1, France