INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° : ␣␣␣␣␣␣␣␣␣␣␣␣

THÈSE

pour obtenir le grade de

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

préparée au laboratoire
Grenoble Image Parole Signal Automatique, Département Parole et Cognition

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

présentée et soutenue publiquement par
Noureddine Aboutabit
le 11 décembre 2007

Titre :

Reconnaissance de la Langue Française Parlée Complétée (LPC) : Décodage phonétique des gestes main-lèvres

Directeurs de thèse :
Denis Beautemps, Laurent Besacier

JURY

M.	Pierre-Yves Coulon,	Président
Mme.	Régine André-Obrecht,	Rapporteur
M.	Paul Deléglise,	Rapporteur
M.	Denis Beautemps,	Directeur de thèse
M.	Laurent Besacier,	Co-encadrant
Mme.	Nadine Vigouroux,	Examinateur

Remerciements

Je tiens d'abord à remercier chaleureusement le directeur de cette thèse Denis Beautemps, chargé de recherche au département Parole et Cognition de Gipsa-lab, pour sa patience et sa générosité. Je tiens aussi à remercier avec la même chaleur Laurent Besacier, Maître de Conférences à l'université Joseph Fourier, le co-directeur de la thèse pour sa disponibilté et sa générosité. Aux deux, pour les conseils et les commentaires très utiles, je le redis MERCI !

Mr. Paul Deléglise, Professeur à l'université du Mans, et Mme. Régine André-Obrecht, Professeur à l'université Paul Sabatier à Toulouse, ont accepté d'être les rapporteurs de cette thèse, je les en remercie profondément. Leurs remarques et commentaires ont contribué à améliorer la qualité de ce mémoire, je ne peux ainsi qu'être reconnaissant.

Je voudrais remercier aussi Mme. Nadine Vigouroux, chargé de recherche au laboratoire ....., qui m'a fait l'honneur de participer au jury de soutenance.

Mr. Pierre-Yves Coulon, Professeur à l'Institut National Polytechnique de Grenoble, m'a fait l'honneur de présider le jury de soutenance, je l'en remercie vivement.

Je tiens à remercier aussi Marie-Agnès Cathiard pour toutes les réponses qui m'ont éclairci dans mon travail de recherche.

Je remercie ensuite les chercheurs et les non-chercheurs, les jeunes et les seniors, du département Parole et Cognition du laboratoire GIPSA-lab. Je remercie tout particulièrement Jean-Luc Schwartz et Gérard Bailly (ancien et actuel directeur du département), pour tous les conseils et les aides apportés à mon travail. Je remercie aussi entre autres Nino (le gaucher qui ne veut pas prendre sa retraite), Stephan (le grand allemend), Annemie (une belge sans blague), Nicolas (ah c'est le petit ! !), Pascal, Dalila, Nadine, Bertrand, Anahita, Viet Anh, Laurent, Emilie, Marion, les deux Frédéric, Pierre, Christian, Christophe, Coriandre ...ect, pour l'ambiance très sociable surtout au coin café.

Sans oublier ceux ou celles qui ont quitté le département et qui m'ont témoigné de leur sympathie Pauline, Virginie, Guillaume, Monique et Julie.

Je remercie tous mes amis à Grenoble ou qui l'on a déjà quitté : Grégoire, Nicolas, Jasmina, Ali, Anais, Mouloud et à Lamya.

Ma gratitude s'adresse à toute ma famille, témoin de mes joies, de mes enthousiasmes, de mes fatigues, de mes nuits blanches, de mes hauts et bas dans tout mon parcours scolaire et universitaire ! Vous avez longtemps attendu la fin, et maintenant c'est fini ! ! ! ! ! ! ! ! ! ! !

Merci à toi Sonya, ton soutien a été précieux comme d'habitude ! ! ! ...

« I would wake up at night dreaming about that awful problem - the tragedy that deaf kids don't read. » Dr. R. Orin Cornett

«Les savants des temps passés et des nations révolues n'ont cessé de composer des livres. Ils l'ont fait pour léguer leur savoir à ceux qui les suivent. Ainsi demeurera vive la quête de la vérité.» Al-Khwarizmi

Introduction

En 1965, le docteur Cornett devient vice-président de la première université aux USA pour les sourds, le Gallaudet Collège. Il remarque alors que les étudiants ont un faible niveau de langage et qu'ils ne s'intéressent pas à la lecture. Ces étudiants sourds, ne pratiquent en effet que la langue des signes, qui n'a que peu de rapport avec la parole. C'est pourquoi leur intérêt pour la pratique de la lecture labiale s'est déprécié d'autant plus que de nombreux sosies labiaux existent en parole. Dans un souci de rendre la lecture plus attractive, Cornett crée une méthode, nommée « Cued Speech », utilisant un augment manuel permettant de désambiguïser les lèvres facilement. Cette méthode permet ainsi aux personnes sourdes ou malentendantes de communiquer naturellement avec leur entourage normo-entendant.

Dans la pratique, le code LPC est souvent produit par la personne normo-entendante et perçue par la personne sourde en décodage. S'il est courant que la personne sourde sache aussi coder le LPC en production, il est très rare que des personnes normo-entendantes le décodent, même parmi celles qui le produisent, du fait qu'il est nécessaire d'avoir été exposé à cette méthode, souvent depuis le plus jeune âge. Par conséquent, la communication entre une personne sourde et une personne normo-entendante ne connaissant pas le code LPC est plus difficile. D'autre part, même pour des personnes maîtrisant le code en perception et en production, il est nécessaire d'être face à face et donc à proximité. L'utilisation d'interfaces permettrait de communiquer à distance, ce qui nécessite dans ce cas des modules d'interprétations automatiques des gestes intervenants dans le code LPC. C'est un des enjeux du projet TELMA (ANR/RNTS - 2005) de « TELéphonie à l'usage des Malentendants ».

Le projet TELMA vise à l'étude de fonctionnalités audio-visuelles tout à fait originales à l'usage des personnes malentendantes dans un cadre de télécommunication téléphonique. Il a pour objectif précis d'exploiter la modalité visuelle de la parole, d'une part pour améliorer les techniques du débruitage du son de parole (la minimisation du bruit environnemental permettant une meilleure exploitation des restes auditifs des malentendants), et d'autre part, en mettant en œuvre des techniques d'analyse/synthèse de lecture labiale et de gestes du code LPC. Le projet se propose de réaliser un système automatique de traduction lecture labiale + LPC vers parole acoustique et inversement, permettant à des utilisateurs malentendants de communiquer entre eux et avec des normo-entendants par l'intermédiaire d'un terminal autonome TELMA (voir figure 1).

Cette thèse se situe dans la chaîne de reconnaissance automatique des gestes main-lèvres du code LPC vers le code phonétique (voir figure 2). Comment extraire automatiquement l'information LPC de la main étant donné que celle-ci est « couplée » au visage et par conséquent

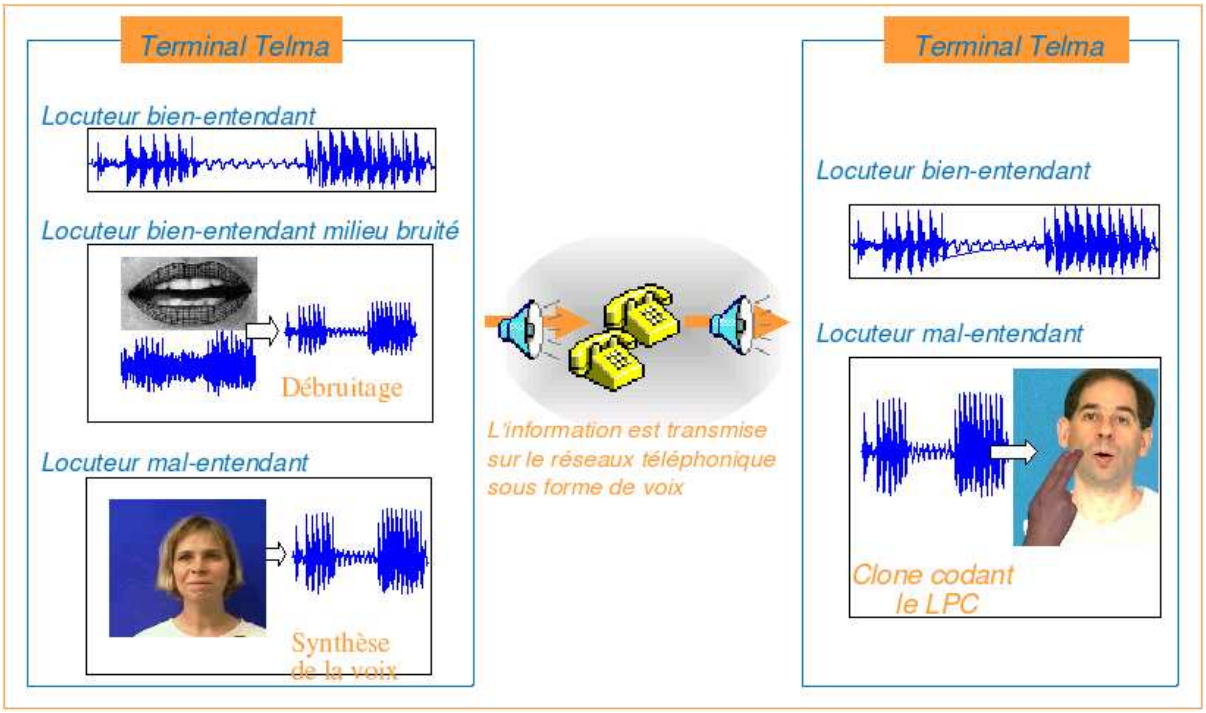

FIG. 1 – Schéma des fonctionnalités de TELMA.

souvent en contact direct ? Les travaux antérieurs de thèse de Virginie Attina (Attina (2005)) menés alors à l'ICP (devenu entre-temps Département Parole et Cognition de GIPSA-lab) ont montré en effet un certain nombre d'indices d'ancrage du mouvement de la main sur la parole visible et sonore. D'autre part, une fois extraite l'information LPC contenue dans les gestes de la main et des lèvres, peut-on modéliser l'information phonétique ? Les travaux d'Attina ont aussi mis en évidence sur plusieurs sujets l'anticipation de la main, avec une atteinte de la main en position cible en début de syllabe, c'est-à-dire dans la consonne et donc bien avant la réalisation de la voyelle aux lèvres. L'anticipation varie avec la durée de la syllabe. Cette anticipation de la main observée en production est utilisée en perception par les personnes sourdes lors de la phase de décodage. Quel modèle de fusion tenant compte de l'asynchronie entre ces flux, peut-on alors appliquer pour obtenir automatiquement le code phonétique et aussi rendre compte des données de perception ? Aucune étude, avant cette thèse, ne s'est confronté à l'ensemble de ces questions d'un point de vue modélisation et reconnaissance automatique du code LPC.

Ce travail de thèse se propose de fournir les premières réponses à ces importantes questions. Dans la première partie, nous présentons tout d'abord, une revue des questions autour de la lecture labiale qui constitue pour les sourds l'accès principal à la parole ; ce qui permet de bien comprendre l'intérêt de la Langue Française Parlée Complétée (LPC), définie comme une association complémentaire et coordonnée de la lecture labiale avec un système de gestes manuels, pour la perception complète de la parole (chapitre 1). Nous donnerons ensuite un état des connaissances sur les méthodes générales permettant d'extraire automatiquement l'information labiale et manuelle (chapitre 2). Cette revue nous permettra de situer notre contribution sur l'extraction de ce type d'information dans une approche orientée « modèle », où la main

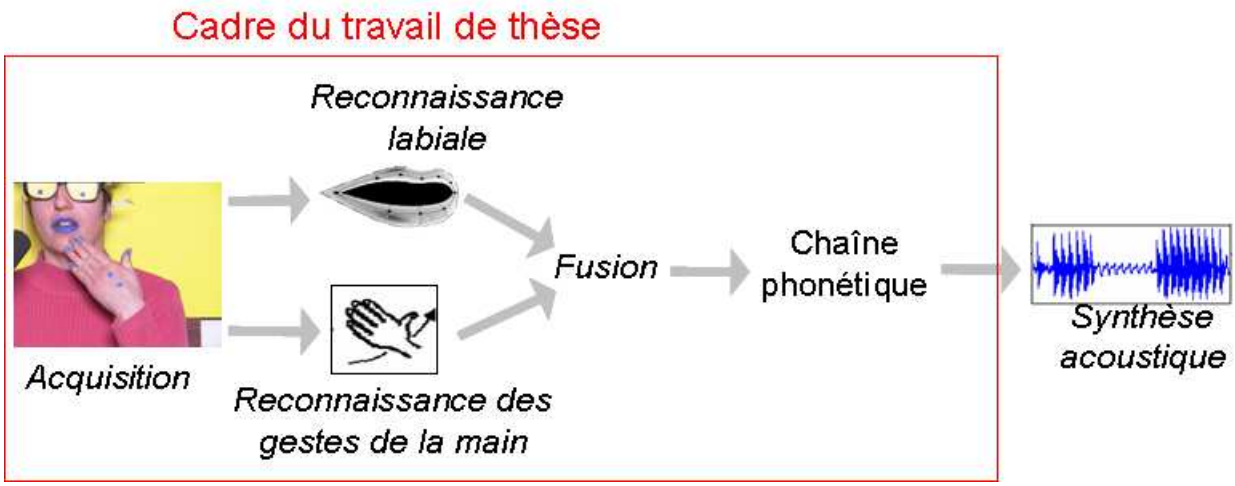

FIG. 2 – Reconnaissance des gestes main-lèvres du code LPC vers la chaîne phonétique. Une synthèse acoustique pourrait ensuite être envisageable.

doit être considérée en relation avec le visage dans les traitements automatiques. De plus, cela nous permettra de nous rendre compte qu'aucun travail, à notre connaissance, ne s'est consacré à l'étude d'un système fusionnant ces deux informations du code LPC (manuelle et labiale) dans le cadre de la reconnaissance automatique. Aussi, dans le chapitre 3, nous discuterons les différents modèles de fusion issus de l'intégration audio-visuelle en parole en vue d'être adaptés au cas du code LPC. Nous verrons que la fusion se ramène au problème général de combinaison de classifieurs. A ce propos, nous décrirons, dans le chapitre 4, des outils statistiques de classification couramment utilisés en reconnaissance automatique de la parole et que nous utiliserons par la suite dans les phases de modélisation de nos flux du code LPC.

Dans la seconde partie, nous aborderons nos études expérimentales en commençant par décrire nos données s'appuyant sur l'enregistrement audio-visuel d'un participant codant le LPC, ainsi que les signaux qui contiennent l'information de la main et des lèvres (chapitre 5). Le chapitre 6 sera dédié au codage automatique des gestes LPC réalisés par la main. Nous montrerons que cette technique, appliquée à l'analyse de l'organisation temporelle main-lèvres-son, fait apparaître un schéma similaire à celui obtenu par Attina *et al.* (2004) à partir de méthodes non automatiques. Le chapitre 7 est centré sur la modélisation du flux labial en traitant tout d'abord le cas des voyelles, puis celui des syllabes de type consonne-voyelle (CV). Nous montrerons que la classification des voyelles en fonction de la position LPC de la main est possible à partir d'un seul instant de mesure, défini à l'atteinte de la cible labiale. En revanche, pour les syllabes CV, la transition complète de la consonne à la voyelle intervient dans la modélisation pour tenir compte des effets de la coarticulation. Nous montrerons que cette modélisation peut être utilisée par concaténation pour la reconnaissance des formes labiales d'un vocabulaire de mots sans aucune phase d'apprentissage. Enfin, nous discuterons dans le chapitre 8 comment adapter les modèles d'intégration audio-visuelle au problème de fusion des flux manuel et labial dans le cas du code LPC. Nous proposerons ainsi un modèle, s'appuyant sur une intégration de décisions, piloté par la décision sur la main.

Première partie

Etat de l'art

« There is an important element of visual hearing in all normal individus.» Cotton
«Si vous voulez que les hommes s'entendent, construisez leur un pont.» Saint-Exupéry
«If a person wishes to accomplish the greatest things that he is capable of accomplishing, he must form within himself a vision...» Cornett

Chapitre 1

De la lecture labiale à La Langue Française Parlée Complétée

La parole ne se réduit pas uniquement à du son transmis entre la bouche d'un locuteur et l'oreille de celui qui le reçoit. La chaîne de production de la parole est un système complexe mettant en œuvre un ensemble d'articulateurs dont certains sont peu visibles car placés à l'intérieur du conduit vocal, et d'autres visibles tels ceux engendrant les mouvements faciaux, principalement les lèvres. Des gestes de la main, peuvent aussi venir en appoint du mouvement des lèvres. C'est le cas de la Langue Française Parlée Complétée. La parole est donc multimodale et met en éveil les sens du système de perception de la parole qui sait recruter non seulement l'audition, mais aussi la vision, voire le toucher. Mais puisqu'après tout, et pour être provocateur, on peut parler et comprendre sans se voir, la multimodalité ne serait-elle pas un luxe pour la communication parlée, et même un plaisir gratuit que s'offre le chercheur aux marges de l'étude de la parole sonore et auditive ? Nous montrerons au contraire, par mouvements progressifs et convergents, que la multimodalité est bien au coeur du dispositif de la communication parlée. Ainsi, nous illustrerons comment l'audition et la vision sont par nature complémentaires, comment la vision peut parfois prendre le relais sur l'audition, comment des gestes de la main en appoint du mouvement des lèvres peuvent permettre la perception complète de la parole lorsque l'information auditive n'est pas accessible. Bref, nous tenterons de montrer quelle est la place de la vision dans la perception de la parole.

1.1 Lecture labiale

1.1.1 La vision pour bien comprendre la parole

Le bénéfice de l'information visuelle dans la perception de la parole est maintenant bien connu. Pour percevoir la parole, la vision est utile dans de nombreux cas. Plusieurs travaux de recherche ont ainsi mis en évidence le gain d'intelligibilité dû à la modalité visuelle et son influence sur la perception de la parole. L'apport de l'information visuelle se manifeste dans plusieurs cas de figure. Tout d'abord, quand la modalité auditive est perturbée par le bruit, des études montrent une amélioration de la perception de la parole. Ensuite, la vision est aussi

souhaitable dans d'autres cas ou elle devient même l'outil essentiel pour la perception de la parole notamment quand le canal auditif est déficient.

1.1.1.1 Bien voir dans le bruit

A travers les travaux précurseurs de Sumby et Pollack (1954), en passant par Summerfield (1979), Summerfield *et al.* (1989) jusqu'à Benoît *et al.* (1996) pour le Français, il est bien établi que l'information visuelle du visage du locuteur est utilisée afin d'améliorer la perception de la parole dans le contexte d'une dégradation de l'audio par du bruit.

Sumby et Pollack (1954) ont examiné l'apport en intelligibilité de la vision à la perception de la parole orale dans un milieu bruité. Ils ont conclu que plus le rapport signal-sur-bruit (RSB) diminue plus l'apport de la vision devient important, en fonction du RSB (variant de 0 à -30 dB) et de la taille du vocabulaire utilisé (mots bisyllabiques). En effet, dans des conditions où le signal acoustique est pratiquement noyé dans le bruit (RSB à -30 dB) la vue du locuteur permet de reconnaître 40% à 80% de mots de plus de ce qui est reconnu avec l'audio seul (la contribution visuelle estimée dans cette expérience varie de 5 à 23 dB). Ils proposent ainsi un protocole et un index qui deviennent une réference pour l'estimation de la contribution de la vision au réhaussement de la parole. Dans la suite, O'Neill (1954)), Neely (1956), Erber (1969), Ewertsen et Birk Nielsen (1971) ont mis en évidence eux aussi des gains en intelligibilité plus élevés en utilisant l'information visuelle en plus de l'audio. Par exemple, en utilisant le même index que Sumby et Pollack (1954), Erber (1969) a obtenu une contribution visuelle de 5 à 10 dB.

A partir de la fin des années 1970, MacLeod, McGrath et Summerfield apportent plus de détails sur la contribution de plusieurs composantes du visage à la compréhension de la parole dégradée. C'est en effet Summerfield qui étudia en 1979 (Summerfield, 1979) l'apport de la vision des lèvres à l'intelligibilité de la parole et donna une première évaluation. Un peu plus tard en 1987 avec MacLeod (MacLeod et Summerfield, 1987) ils estiment un gain moyen de 11 dB apporté par la vision des lèvres. Summerfield *et al.* (1989) évaluaient ensuite en 1989 la contribution des dents à la discrimination des voyelles. Ils concluent que le bénéfice apporté par les dents s'estime à 6% mais remarquent aussi que ce gain en intelligibilité est beaucoup plus important pour certaines voyelles. Dans cette expérience, ils utilisent un corpus de 11 syllabes différentes présentées visuellement dans 7 conditions selon l'apparition du visage entier, des lèvres et dents en utilisant un visage naturel ou synthétique.

S'appuyant sur les travaux précurseurs de Sumby et Pollack (1954) pour l'Anglais, Benoît *et al.* (1996) ont comparé pour le Français l'intelligibilité dans le bruit des consonnes C = [b, v, z, ʒ, ʁ, l] et des voyelles V = [a, i, u] placées à l'intérieur de logatomes de structure [VCVCVz]. Le signal audio des logatomes était dégradé par du bruit selon six niveaux de rapport signal sur bruit (S/N), pour constituer les stimuli de la condition audio seule A. Les stimuli audio de la condition A ont été associés l'image du visage vue de face du locuteur pour constituer les stimuli de la condition audio-visuelle AV. La figure 1.1 présente les réponses de 18 sujets bien entendants pour deux conditions d'expérience. Les deux courbes montrent une baisse de l'intelligibilité lorsque la proportion du bruit dans l'audio augmente (rapport signal-sur-bruit de plus en plus négatif)

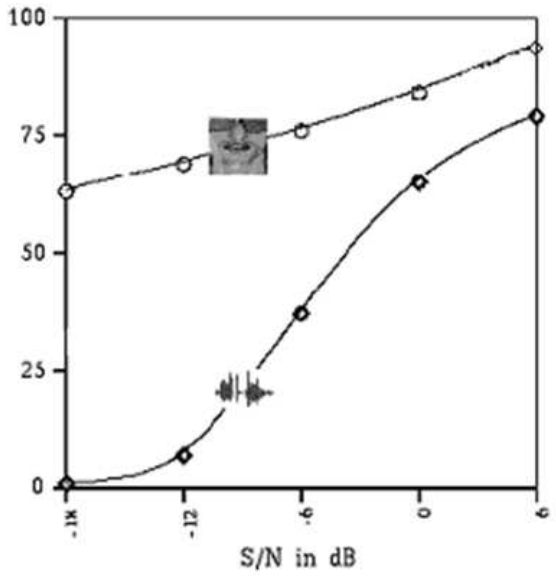

FIG. 1.1 – Identification moyenne (18 sujets normo-entendants) en pourcentage pour les stimuli de la condition audio seule A (courbe du bas) et audiovisuelle AV (courbe du haut), en fonction du rapport signal sur bruit (d'après (Benoît *et al.*, 1996)).

mais avec une meilleure résistance quand l'information visuelle est présente par rapport à la condition audio seule (A). Les performances de la condition AV atteignent un minimum de 60% correspondant à l'information visuelle seule (le niveau du RSB de -18 dB indiquant que le signal audio est totalement bruité et n'apporte aucune information). Les auteurs concluent ainsi que l'information visuelle, immédiatement disponible, récupérée par la lecture labiale sans entrainement spécifique, et combinée à l'information audio permet donc des performances en identification supérieures à la condition audio seule.

1.1.1.2 ... Et même sans bruit ...

La vision est bénéfique dans d'autres situations même lorsque le son est parfaitement audible (parole non bruitée), comme le montre par exemple les expériences dites de *Shadowing* ((Reisberg *et al.*, 1987)). Dans ses tâches de *Shadowing*, il est demandé aux sujets de répéter le plus rapidement possible un texte entendu dans lequel la compréhension du message est sémantiquement difficile ((Reisberg *et al.*, 1987) a utilisé un passage de la critique de la Raison pure de Kant) en condition audio seule et en condition audio-visuelle. L'analyse des temps de réaction de la tâche de répétition des deux conditions, montre que lorsque le visage du sujet est présenté en plus du son, les temps de réaction sont diminués en moyenne d'un facteur de 7,5%. Ce même phénomène est constaté pour d'autres langues étrangères avec un gain moyen de l'ordre de 20,5% pour la condition audiovisuelle ((Beautemps *et al.*, 2003)).

Par ailleurs, on ne peut pas se priver de citer le très célèbre effet McGurk qui démontre la capacité d'intégrer les deux informations auditive et visuelle même si elles ne sont pas congruentes ((McGurk et MacDonald, 1976) ; (MacDonald et McGurk, 1978)). Dans l'expérience pratiquée

par les auteurs, l'image vidéo d'une séquence naturelle [aga] est associée au son naturel de [aba]. En identification, cela conduit à reconnaître la séquence [ada], démontrant ainsi que le système de perception ne privilégie pas plus l'information audio que visuelle, même si celles-ci sont contradictoires, et qu'au contraire il sait les combiner. De plus, cette expérience montre que même si les sujets sont avertis, ces derniers restent sensibles à l'illusion.

En conclusion, l'information visuelle immédiatement disponible chez les normo-entendants ne nécessite pas d'apprentissage et contribue à la perception de la parole. Cette contribution est variable selon la difficulté de la situation. Et on ne peut pas s'empêcher de voir !

1.1.2 Lecture labiale : apprentissage et limite

La lecture labiale pour certains, lecture labio-faciale pour d'autres, est un moyen qui permet de percevoir la parole en regardant les lèvres et l'ensemble du visage ((Summerfield, 1979) ; (Summerfield, 1987) ; (Summerfield *et al.*, 1989)). Les personnes devenues sourdes ou malentendants à l'âge adulte utilisent fortement la lecture labiale. (Binnie *et al.*, 1974) rapportent que la vision est le moyen principal pour aider les personnes ayant une hypoacousie, et notamment les sourdes profonds, à accéder à l'information orale.

1.1.2.1 Variabilités des performances en lecture labiale

Peut-on apprendre à lire sur les lèvres ? On vient de le voir, un nombre important de travaux montre que les personnes normo-entendantes peuvent avoir des compétences en lecture labiale sans entraînement spécifique. Effectivement, lorsqu'une personne entend bien, sa faculté de lire sur les lèvres est inconsciente. Mais lorsque la personne devient sourde, ou lorsque la surdité survient partiellement, elle s'aperçoit qu'elle est capable de lire sur les lèvres. La lecture labiale devient alors une faculté compensatrice et naturelle. Cette faculté reste malgré tout insuffisante sauf pour certains individus devenus sourds très jeunes et qui s'adaptent très vite. Par contre, il est bien établi que les performances initiales en lecture labiale - sans apprentissage ni entraînement spécifique - varient considérablement d'un individu à l'autre. Ainsi, en 1990, (MacLeod et Summerfield, 1990) ont testé la capacité à lire sur les lèvres de 20 sujets normo-entendants qui ont obtenu des scores variant de 0 à 70% d'identification correcte de mots dans des phrases. Les auteurs ont montré qu'en situation d'écoute bruitée, ces sujets tiraient profit de la vue du visage du locuteur d'une manière également variable : les sujets bons lecteurs labiaux supportaient - à performance d'identification égale - une augmentation du rapport signal sur bruit de 11 dB en condition audiovisuelle par rapport à la condition auditive seule, tandis que les mauvais lecteurs labiaux ne supportaient que 2 dB supplémentaires. Plus récemment, (Bernstein *et al.*, 2000) ont effectué une étude sur 72 étudiants sourds et 96 étudiants bien entendants. Ils ont obtenu des scores de lecture labiale de syllabes sans sens, de mots isolés monosyllabiques et de phrases isolées. Les sujets mal-entendants sont des étudiants de l'univérsité Gallaudet et ont été séléctionnés suivant plusieurs critères dans le but de recruter des participants pour lesquels la lecture labiale est une compétence importante socialement et bien utilisée, et d'exclure des participants ayant comme langue première la langue des signes américaine ou autres systèmes manuels de communication autre que l'anglais. Les critères concernent : (i) l'age entre 18 et 45

ans ; (ii) le degré de surdité supérieur à 60 dB HL[1] ; (iii) aucun autre handicap que la surdité ; (iv) de langue maternelle anglaise ; (v) plus de 8 ans d'éducation dans un programme oral ; et (vi) une vision supérieure à 20/30 pour chaque oeil. Parmi les différents résultats obtenus, les scores de la lecture labiale des phrases varient entre les deux groupes de participants (de 0% à 57% de mots corrects pour les normo-entendants *vs* de 0% à 80% pour les mal-entendants). Nous observons alors, autre que la large variabilité dans chaque groupe, que les résultats sont plutôt favorables pour les mal-entendants. Les auteurs observent aussi que parmi ceux ayant les scores les plus importants, se trouvent les individus qui ont une surdité profonde et congénitale. Une explication de ces différences entre les deux groupes peut être, selon les auteurs, que les adultes normo-entendants ne sont pas habitués à essayer de comprendre la parole par la vision seule.

1.1.2.2 Effet de l'entraînement

On a souvent dit que ces bons lecteurs labiaux étaient nés ainsi car, en effet, les expériences d'entraînement à la lecture labiale ont souvent été moins concluantes que ce qui était attendu. Notons que dans les études où les sujets sont entraînés à décoder visuellement des segments de parole (des consonnes par exemple), une amélioration des performances est bien présente (Walden. *et al.*, 1977; Walden *et al.*, 1981). La question reste de savoir si ces effets positifs sont seulement présents à court terme plutôt qu'à long terme (ainsi qu'ont pu le suggérer des auteurs comme Heider et Heider (1940)). Néanmoins, l'étude de Bernstein *et al.* (2000) semble clairement indiquer que de bonnes performances globales en lecture labiale (sur des phrases par exemple) sont liées à de bons scores de perception phonétique visuelle (en identification de phonèmes ou de mots isolés). On peut donc inférer de cet ensemble d'informations qu'un entraînement régulier et répété à la lecture labiale doit permettre une amélioration des performances globales, même s'il ne faut peut-être pas espérer combler toutes les différences individuelles de départ. Ceci dit, même avec entraînement, les meilleurs lecteurs labiaux n'atteignent pas la perfection. En moyenne pour une langue donnée, seules 40 à 60 % des phonèmes sont appréhendés par la lecture labiale (Montgomery et Jackson, 1983), et seulement 10 à 30 % des mots (Nicholls et Ling, 1982; Bernstein *et al.*, 2000).

1.1.2.3 La lecture labiale seule ne suffit pas : visèmes

Nous venons de voir que la lecture labiale ne permet pas toujours la compréhension complète de la parole même avec entraînement. La raison de cette limitation est simple. En effet, si nous voulons "voir" la parole produite par un locuteur, c'est au niveau des lèvres qu'il faut regarder. Or les phonèmes ne sont pas tous caractérisés par des formes distinctes aux lèvres. Par exemple, les consonnes occlusives [b], [m] et [p] sont produites avec des formes labiales similaires. Il est

[1] *dB HL = décibels HL (pour Hearing Level) : est une échelle de mesure de l'audition. Une valeur en décibels HL vaut la valeur en dB SPL pour laquelle le patient détecte (entend) un son, moins la valeur de référence (en dB SPL toujours). Cette échelle représente donc l'écart de l'audition de la personne testée par rapport à l'audition moyenne. Le dB SPL (Sound Pressure Level) est défini par le rapport de la puissance par unité de surface du son que l'on mesure et une puissance par unité de surface de référence.*

en effet difficile de les distinguer sans tenir compte du contexte et sans une différenciation du mode (voisement pour [b] et nasalité pour [m]).

Ces confusions visuelles des phonèmes ont conduit à définir la notion de "visèmes" ou "sosies labiaux". Un visème est l'unité de base de la parole dans le domaine visuel qui correspond au phonème (qui est l'unité de base de la parole dans le domaine acoustique) ((Fisher, 1968)). Le terme même "visème" (en anglais "viseme") a été défini par Fisher comme une contraction de "*vis*ual phon*emes*" : *« The phrase* ***visual phoneme*** *has been shortened to* ***viseme****, and will be used to refer to any individual and contrastive visually perceived unit »* (Fisher, 1968), p. 800. Le visème décrit les caractéristiques visuelles et/ou orales lors de la production des phonèmes. Ces unités sont perçues distinctement l'une de l'autre. Il n'y a pas cependant de correspondance biunivoque entre visème et phonème. En effet, plusieurs phonèmes partagent le même viseme. Il en est ainsi par exemple de [k], [ɡ] et [ŋ] (visème : [k]) dont le lieu d'articulation est au niveau du palais mou, ou [ʃ], [ʒ] (visème : [ch]) caractérisées par un geste d'éversion[2] aux lèvres. Réciproquement, certains sons difficiles à distinguer en acoustique sont clairement distingués avec l'image du visage (Chen, 2001). Ceci est illustré par exemple lors de certaines incompréhensions de mots au téléphone.

Benoit *et al.* (1992) reviennent sur le concept de visème. Les auteurs observent que les voyelles en contexte s'écartent du modèle établi à partir d'observations des voyelles isolées. L'étude de Montgomery *et al.* (1987), dans laquelle l'effet du contexte consonantique sur la lecture labiale de voyelles est évalué, confirme ce constat. Ainsi, le terme visème correspond davantage à une construction expérimentale et non théorique ; Ceci se rapproche du concept acoustique de l'allophone par rapport au phonème ((Benoit *et al.*, 1992)). Ce concept permet alors de dénombrer les formes visuelles qui décrivent les réalisations de la parole dans une langue précise.

1.1.3 Catégorisation visuelle des phonèmes

La catégorisation visuelle des phonèmes, en contexte ou non (phonèmes isolés), dépend manifestement de la production et de la perception de la parole. En effet, la description visuelle de la parole dépend de la géométrie du conduit vocal ce qui affecte l'acoustique mais aussi l'identification visuelle (Benoit *et al.*, 1992). Afin d'estimer les ressemblances visuelles entre les phonèmes, de nombreuses études ont été menées pour les voyelles et les consonnes.

1.1.3.1 Catégorisation des consonnes

Pour les consonnes, les résultats ne sont pas similaires d'une étude à l'autre. En effet, le nombre de visèmes et le nombre d'éléments d'un même visème varient d'un auteur à l'autre. Pour la langue anglaise, les tentatives de regroupement des consonnes en visèmes ont commencé dès les années soixante, l'époque où Fisher a défini le terme visème. Woodward et Barber (1960) ont défini quatre groupes de consonnes, tandis que Fisher (1968) trouvait cinq visèmes. Binnie *et al.* (1974) ont obtenu aussi cinq groupes de consonnes, et un peu plus tard Walden. *et al.* (1977) mentionnaient neuf visèmes. Si le nombre de visèmes diffère, la composition de ces visèmes

[2] *L'éversion est une saillie formée par une muqueuse qui s'est retournée vers l'extérieur.*

en est la cause. Ainsi, dans les quatre visèmes formés par Woodward et Barber (1960), les consonnes [p], [m] et [b] se retrouvent dans une même classe, tandis que Fisher (1968) ajoute la consonne [d] à ce groupe. Nous venons de voir que ces différents résultats montrent à quel point le regroupement sur des caractéristiques visuelles des consonnes n'est pas aussi simple qu'on le croit. Les recherches qui ont suivi, se sont notamment intéressées à la détection de facteurs qui peuvent expliquer de telles différences. Les facteurs que nous pouvons trouver dans la littérature sont relatifs aux participants, aux conditions expérimentales et aux analyses employées. Jackson (1988) rapporte que les conditions d'éclairage et d'observations ainsi que les critères utilisés par les chercheurs pour classifier les consonnes dans des catégories visuelles, peuvent avoir des effets sur la formation des catégories. La clarté de l'articulation du locuteur, un facteur relevé par Lesner (1988), influence la reconnaissance visuelle des consonnes même si l'intelligibilité de la modalité auditive du locuteur n'implique pas directement une bonne articulation et donc une modalité visuelle claire (Gagné, 1994; Gagné *et al.*, 1995). Enfin, chez un participant ayant eu un apprentissage, le nombre de visèmes identifiés augmente ; Walden. *et al.* (1977), illustre notamment cet effet de l'apprentissage sur la catégorisation.

Le regroupement le plus cité en général est celui de Summerfield (1987). Il présente des stimuli visuels de consonnes prononcées en contexte [ɑ] à des sujets expérimentés (donc ayant eu un entraînement) en lecture labiale pour un test perceptif d'identification. Il utilise ensuite une classification hiérarchique pour regrouper les consonnes qui ont été confondues par les interlocuteurs. Il obtient finalement, en considérant le seuil de 75% sur les présentations dans lesquelles les consonnes étaient bien identifiées dans leur classes de visèmes, neuf catégories.

Massaro *et al.* (1993) parviennent à faire ressortir sept catégories. Les auteurs utilisent 66 syllabes CV, chacune est répétée deux fois, combinant 22 consonnes initiales avec 3 voyelles ([a], [u], [i]). Les 132 syllabes sont présentées à des sujets normo-entendants de façon aléatoire en 3 conditions modales : audio seul, vidéo seul ou bimodale (audio+vidéo) et ceci selon que le débit est normal ou rapide. Les résultats d'identification obtenus dans la condition vidéo seule montrent que les consonnes peuvent se regrouper en sept catégories.

Pour bien décrire les travaux qui existent dans la littérature ou du moins ce que nous avons pu consulter, nous nous sommes inspirés d'une table dressée par Owens et Blazek (1985), et décrivant une compilation des études sur les visèmes des consonnes pour l'Anglais, pour reporter les visèmes obtenus lors de différents travaux qui ont été menés avant et après cette date. La table 1.1 présente les catégories de visèmes obtenues. Par ailleurs, nous avons trouvé nécessaire de reporter aussi tous les éléments concernant les expériences. Ainsi, nous présentons en annexe A les sujets, les locuteurs, les critères des regroupement en visèmes et les catégories obtenues.

En résumé, on peut dire que certaines classes de visèmes ressortent toujours, que ce soit avec entraînement en lecture labiale ou sans. Ces classes sont : [p, m, b], [f, v] et [ʃ, ʒ]. Les disctinctions de ces visèmes sont dérivées des traits visuels les plus faciles à percevoir tel la labiabilité, la protrusion ou l'éversion .... Le reste des consonnes est difficile à discriminer avec les lèvres seules, et leur disctinction nécessite un entraînement à la lecture labiale et/ou un éclairage particulier du locuteur. Ainsi, des groupes comme [s, z], [d, n, t], [k, g] ou [l] peuvent être distingués.

En ce qui concerne le Français, dans un test perceptif auprès de sujets ayant des déficiences

Auteurs	Catégories
Heider et Heider (1940)	- [p, b, m] ; [f,v] ; [r] ; [θ] ; [ʃ,tʃ,dʒ] ; [n,t,d] ; [l] ; [k,g]
Woodward et Barber (1960)	- [p,b,m] ; [f,v] ; [w,r,hw] ; [t,d,n,l,θ,ð,s,z,tʃ,dʒ,ʃ,ʒ,j,k,g,h]
Fisher (1968)	- Initiales : [p,b,m,d] ; [f,v] ; [w,hw,r]. - Finales : [p,b] ; [f,v] ; [ʃ,ʒ,dʒ,tʃ] ; [t,d,n,θ,ð,s,z,r,l].
Binnie *et al.* (1974)	- [p,b,m] ; [f,v] ; [ʃ,ʒ] ; [θ,ð] ; [n,d,t,s,z,k,g]
Walden. *et al.* (1977)	- Avant entraînement : [p,b,m] ; [f,v] ; [w] ; [θ,ð] ; [ʃ,ʒ,s,z] ; non classées : [t,d,n,k,g,r,l,j]. - Après entraînement : [p,b,m] ; [f,v] ; [w] ; [ʃ,ʒ] ; [θ,ð] ; [r] ; [s,z] ; [t,d,n,k,g,j] ; [l].
Walden *et al.* (1981)	- Avant entraînement : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ;[ʃ,ʒ,tʃ,dʒ] ; non classées : [t,d,n,s,k,g,l,j]. - Après entraînement : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ;[ʃ,ʒ,tʃ,dʒ] ; [t,d,n,s,k,g,l,j].
Kricos et Lesner (1982)	- Locuteur 1 : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,z,s] ; [l] ; [k,j,h,g,ŋ]. - Locuteur 2 : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,z,s] ; [l] ; [k,j,h,g,n,ŋ]. - Locuteur 3 : [p,b,m] ; [f,v,r] ; [w] ; [θ,ð] ; [ʃ,ʒ,tʃ,d] ; [k,g]. - Locuteur 4 : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,z,s]. - Locuteur 5 : [p,b,m] ; [f,v,z,s] ; [w,r] ; [ʃ,ʒ,tʃ,dʒ]. - Locuteur 6 : [p,b,m] ; [ʃ,ʒ,tʃ,dʒ] ; [w,r,θ,ð] ; [t,d,z,s] ; [l,j,h,n,l].
Owens et Blazek (1985)	- Context [ɑCɑ/ : [p,b,m] ; [f,v] ; [θ,ð] ; [w,r] ; [ʃ,ʒ,tʃ,dʒ] ; [n,k,g,l] ; [h]. - Context [ʌCʌ] et [iCi] : [p,b,m] ; [f,v] ; [θ,ð] ; [w,r] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,s,z] . - context [uCu] : [p,b,m] et [f,v].
Summerfield (1987)	- [p,b,m] ; [f,v] ; [θ] ; [ʃ,ʒ] ; [d,t] ; n,g,k] ; [s,z] ; [l] ; [r] ; [w] ; [y].
Massaro *et al.* (1993)	- [p,b,m] ; [s] ; [ʃ,tʃ,j] ; [d,t, n,g,k, h] ; [l] ; [r] ; [w].

TAB. 1.1 – Une compilation des études menées pour déterminer les visèmes des consonnes en Anglais.

auditives, Gentil (1981) aboutit à trois groupes de visèmes caractérisés par une articulation bien visible aux lèvres. Dans son étude sur la reconnaissance visuelle de 16 consonnes ([p, b, m, f, v, ʒ, ʃ, s, z, d, n, t, k, g, ŋ, r]) placées à l'initiale et en finale de mot, l'auteur regroupe les consonnes bilabiales ([p], [m] et [b]), les labiodentales ([f] et [v]) et les consonnes protruse avec éversion ([ʃ] et [ʒ]). Ces trois groupes ont été considérés par Jackson (1988) comme étant des visèmes de consonnes "universels" puisqu'on les retrouve dans plusieurs études et dans plusieurs langues.

Par contre, les autres consonnes sont moins articulées au niveau des lèvres, et leur catégorisation suivant un critère labial est donc plus compliquée et leur distinction visuelle se fonde sur le lieu d'articulation. Dans le même sens, Jutras *et al.* (1998), ont tenté d'établir des catégories de consonnes pour le Français québécois en se basant sur leur identification visuelle. Les auteurs présentent à 46 participants des stimuli linguistiques des 17 consonnes du français québécois (([p, b, m, f, v, ʒ, ʃ, s, z, d, n, t, k, g, ŋ, r, l]) dans le contexte de la voyelle [a] (syllabes de type VCV) provenant de deux locutrices. Les résultats démontrent que le nombre de catégories de consonnes trouvées varient d'une locutrice à l'autre, rejoignant ainsi Kricos et Lesner (1982, 1985) sur l'effet de la variabilité inter-locuteur de la production verbale sur la catégorisation des consonnes. Néanmoins, ils retrouvent les trois visèmes "universels" chez les deux locutrices. Nous présentons dans la table 1.2, les groupes de consonnes établis par Jutras et ses collègues pour deux locutrices en comparaison à ceux obtenus par Gentil.

Visème	Description	Gentil : finale de mot	Gentil : initiale de mot	Jutras : Loctrice 1	Jutras : Loctrice 2
p	bilabiales	[p,b,m]	[p,b,m]	[p,b,m]	[p,b,m]
f	labiodentales	[f,v]	[f,v]	[f,v]	[f,v]
ch	postalvéolaires protruses	[ʒ,ʃ]	[ʒ,ʃ]	[ʒ,ʃ]	[ʒ,ʃ]
s	alvéolaires	[s,z]		[s,z]	[s,z]
d	non labiales	[d,n,t]	[s,z,d, n,t,k,g,ŋ,r]	[d,n,t,k,g,ŋ,r]	[d,n,t,k,g,ŋ]
r	non labiales	[r]			[r]
l	latérales			[l]	[l]
k	vélaires	[ŋ,k,g]			

TAB. 1.2 – Visèmes de consonnes définis par Jutras *et al.* (1998) (2 locutrices) pour le Français québécois comparé à ceux établis par Gentil (1981) pour le Français.

La différence entre les visèmes retrouvés chez les deux locutrices de Jutras et al. est liée à la consonne [r] qui constitue une catégorie indépendante chez la locutrice 2 et non pas chez la locutrice 1. Par ailleurs, le fait de considérer la consonne [r] comme une catégorie à part a été aussi remarqué en anglais par Binnie *et al.* (1976) et la question se pose donc pour le français québécois.

Finalement, il convient de préciser que toutes ces expériences de catégorisation visuelle des consonnes, que ce soit pour l'anglais ou pour le français, ont été effectuées par un test perceptif et non par un test de reconnaissance à partir de mesures de paramètres visuels. Certaines classes de visèmes sont plus facilement perçues par l'œil que d'autres classes ; notamment, les segments produits à l'avant de la bouche sont plus facilement lus sur les lèvres que ceux produits en arrière.

1.1.3.2 Catégorisation des voyelles

Pour les voyelles, les premiers tests perceptifs visuels effectués depuis Heider et Heider (1940) n'ont pas abouti à un vrai regroupement visuel pour l'Anglais. La plupart de ces études trouvaient que les voyelles n'ont pas les mêmes similarités visuelles que les consonnes. En 1983, Montgomery et Jackson (1983) arrivent plus au moins à dresser des visèmes pour les voyelles de l'anglais dans une étude qui avait pour objectif d'évaluer la relation entre les caractéristiques physiques des lèvres pendant la production des voyelles et la confusion qu'entraîne la lecture labiale de ces voyelles. Dans cette étude pour l'anglais américain, dix sujets normo-entendants identifient les stimuli provenant de vidéos enregistrant quatre locutrices prononçant 15 voyelles et diphtongues dans le contexte [h-**V**-g]. Les visèmes qu'on obtient en se référant à l'analyse perceptive de cette étude, peuvent globalement se résumer en quatre catégories : visème [i]= [i, ɪ], visème [ɑ]= [ɑ, aɪ, eɪ, ʌ, æ, ɛ], visème [ɔ]= [ɔ, aʊ, əʊ] et visème [u]= [u, ʊ, ɔɪ, ɜ ː].

En Français, les premiers tests ont été effectués par Mourand-Dornier (1980) pour des sujets normo-entendants et Gentil (1981) pour des sujets mal-entendants. Tseva (1989), s'appuyant sur les résultats et les données enregistrés par ces deux derniers et en employant une analyse factorielle des correspondances, éprouve les mêmes difficultés pour construire des groupes de visèmes pour les voyelles du Français. Toutefois, l'auteur arrive à distinguer nettement deux groupes pour les normo-entendants ainsi que pour les mal-entendants : les voyelles arrondies et les voyelles non arrondies. Il a été démontré dans cette étude que le seul trait robuste pour classifier les voyelles en Français est l'arrondissement rejoignant ainsi Jackson *et al.* (1976) qui ont démontré auparavant la même robustesse de l'arrondissement pour l'anglais. Il est à noter que les données de Mourand-Dornier (1980) et de Gentil (1981) étaient plus au moins similaires. Ils ont utilisé tous les deux 13 voyelles avec quelques petites différences sur la composition de chaque groupe ([a, o, ɛ, i, u, e, ɔ, œ, ɑ̃, ø, y, ɔ̃, ɛ̃] pour Mourand-Dornier (1980) et [a, o, ɛ, i, u, e, ɔ, ɑ, ɑ̃, œ̃, y, ɔ̃, ɛ̃] pour Gentil (1981)). Les voyelles de Mourand-Dornier (1980) ont été extraites de mots composés de 3 phonèmes tandis que les voyelles de Gentil (1981) sont en contexte syllabique de type consonne-voyelle (CV) dont C est la consonne [l]. D'un autre coté, les angles de vue des locuteurs dans les deux cas sont différents (vue de face pour Mourand-Dornier (1980) et vue de profil avec angle de 30° pour Gentil (1981)). Par contre, tous les deux ont optimisé les conditions d'éclairage de sorte que les mouvements articulatoires de parole des locuteurs soient facilement visibles sur les vidéos.

Un autre trait peut s'avérer utile pour percevoir visuellement les voyelles : la hauteur (ou la séparation inter-labiale). Robert-Ribès (1995) montre que ce trait est moins récupéré que l'arrondissement ; notamment pour les voyelles antérieures non-arrondies. Cependant, l'auteur a montré aussi que la hauteur semble créer des confusions entre certaines voyelles (ouvert ou fermé). De ce fait, il peut être ainsi utilisé avec l'arrondissement pour un éventuel groupement visuel précis des voyelles. En ce sens, Cathiard (1994) fournit une revue de la littérature bien détaillée sur la perception visuelle des voyelles.

Parmi les voyelles des deux corpus précédents, il y avait des voyelles nasales qui semblent très difficiles à distinguer au niveau des lèvres. Ce type de voyelles n'a pas été considéré par Robert-Ribès (1995) dans son étude sur les modèles d'intégration audiovisuelle. Il a travaillé sur

un corpus enregistré de dix voyelles orales ([a, ɛ, i, e, oe, ɔ,y, ø, u, o]) prononcées isolément et tenues sur une durée d'une seconde. Chaque voyelle est répétée plusieurs fois de façon à obtenir 340 réalisations (34 réalisation pour chaque voyelle). Des paramètres géométriques (étirement, hauteur et aire aux lèvres) ont été extraits du contour des lèvres en maquillant les lèvres en bleu avec un éclairage optimum pour faciliter l'extraction. Une partie de ces données (les stimuli visuels) a servi à illustrer la dispersion des dix voyelles dans les plans des différents paramètres géométriques du contour des lèvres. La figure 1.2 montre un exemple tiré de Robert-Ribès (1995) de la dispersion de ces voyelles dans le plan des paramètres (hauteur B en fonction de l'étirement A). Sur ces dispersions, trois grands groupes de voyelles sont bien distincts : les voyelles étirées [a, ɛ, i, e], les voyelles arrondies [y, ø, u, o] et les voyelles semi-arrondies [oe, ɔ]. Cette étude reste la base jusqu'à maintenant pour la classification des voyelles en visèmes. L'avantage de cette classification est qu'elle est faite à partir de mesures physiques sur les lèvres et non pas avec un test perceptif comme on l'a vu précédemment. Les données de Robert-Ribès (1995) sont une référence dans la modélisation de nos données labiales et principalement le classement de nos visèmes de voyelles.

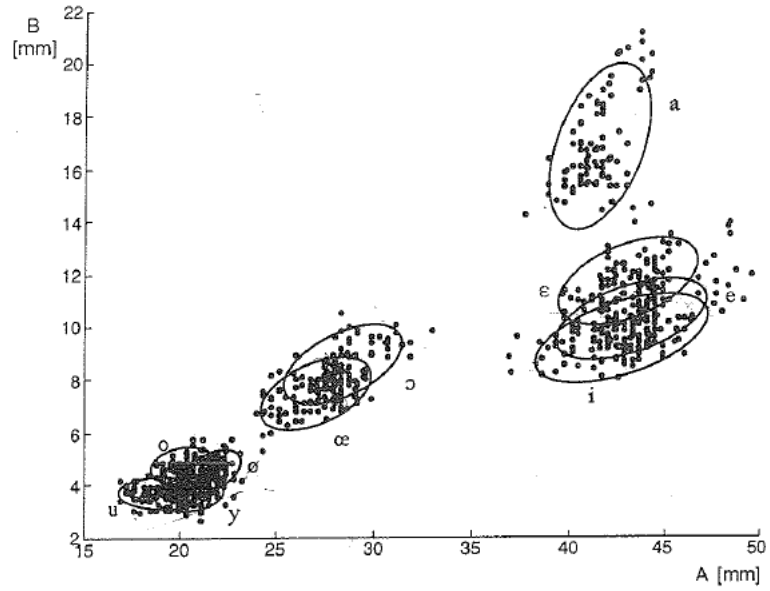

FIG. 1.2 – Stimuli visuels des voyelles dans le plan des paramètres géométriques (étirement A, hauteur B) (extrait de Robert-Ribès (1995)).

1.1.4 Effets contextuels et coarticulatoires

Les regroupements en visèmes, que ce soit pour l'Anglais ou pour le Français, dépendent de plusieurs facteurs, aussi bien pour les voyelles que pour les consonnes. Nous avons décrit auparavant les principaux facteurs agissant sur les résultats finaux des expériences décrites précédemment pour le cas des consonnes. Pour les voyelles, ces facteurs ont aussi la même influence. Les facteurs décrits concernent notamment la variabilité inter-locuteurs (différences

entre les locuteurs en production de la parole et donc différences entre leurs mouvements articulatoires), la variabilité inter-sujets (et dans le cas des tests perceptifs, par exemple savoir si les participants ont eu une expérience ou non en lecture labiale), la procédure employée (et surtout les critères utilisés pour classifier les phonèmes en visèmes) et le type de stimuli utilisés. Dans les différents regroupements, les auteurs n'ont pas utilisé le même type de données. Certains ont testé des phonèmes extraits de syllabes CV (Heider et Heider, 1940; Binnie *et al.*, 1974, 1976; Walden. *et al.*, 1977; Gentil, 1981), d'autres utilisent des syllabes VCV (Walden *et al.*, 1981; Kricos et Lesner, 1982; Owens et Blazek, 1985), CVC (Mourand-Dornier, 1980) ou plus complexes de type $VC_1VC_1VC_2$ (Benoit *et al.*, 1994) ou encore des mots (Fisher, 1968). Cette variabilité des données trouve son explication dans la présence de la coarticulation dans la production de la parole. En effet, dans ces études, les contextes dans lesquels sont extraits les phonèmes (consonnes ou voyelles) sont globalement contôlés.

La coarticulation est un phénomène qui fait que selon les phonèmes adjacents dans une phrase, un phonème n'est pas prononcé de la même façon. Ceci est dû au fait que la transition entre les phonèmes ne se fait pas par une modification instantanée de la configuration du conduit vocal mais d'une façon progressive.

«The term of coarticulation refers to the altering of the set of articulatory movements made in the production of one phoneme by those made in the production of an adjacent or nearby phoneme» (Benguerel et Pichora-Fuller, 1982).

Evidement, ce phénomène entraîne des conséquences sur les caractéristiques acoustiques d'un phonème particulier quand celui-ci est prononcé dans des contextes phonétiques différents. Il implique aussi tout le système articulatoire en jeu dans la production de la parole, y compris les lèvres. Ainsi, d'un contexte à l'autre, un phonème pourrait ne pas avoir les mêmes formes aux lèvres. La perception visuelle des phonèmes est alors modifiée selon le contexte.

La coarticulation a fait l'objet de plusieurs recherches afin de déterminer ses racines et ses effets sur la production et la perception de la parole. Pour bien comprendre la coarticulation, sa nature, ses domaines et la structure coarticulée de la parole nous renvoyons les lecteurs vers l'analyse bien détaillée de Cathiard (1994). Quant aux effets perceptifs du contexte et principalement sur la lecture labiale, nous retrouvons l'investigation de la nature de la coarticulation dans des recherches portant sur l'arrondissement labial. Par ailleurs, de nombreuses études ont montré l'influence mutuelle du contexte entre les consonnes et les voyelles (Gentil, 1981; Benguerel et Pichora-Fuller, 1982; Owens et Blazek, 1985; Montgomery *et al.*, 1987; Cathiard, 1988; Benoit *et al.*, 1992, 1994). Toutes ces études sont détaillées en annexe C.

Ce que nous pouvons retenir de toutes ces études est que les configurations visuelles des phonèmes sont, comme dans le cas de l'audio, différentes selon le contexte phonétique avoisinant. Certains phonèmes, qui peuvent être décelables grâce à des traits visuels pertinents et bien robustes comme l'arrondissement par exemple, sont moins influencés par l'environnement phonétique adjacent. Dans certains cas, certains phonèmes peuvent marquer les phonèmes adjacents. C'est le cas par exemple du contexte consonantique protru de [ʃ] et [ʒ] qui domine son entourage vocalique même si celui-ci présente moins les mêmes traits visuels (protrusion).

1.1.5 La lecture labiale chez les personnes sourdes ou mal-entendantes

Le bénéfice de la lecture labiale est bien établi pour les personnes normo-entendantes. Pour les déficients auditifs, la lecture labiale prend le relais de l'audition. La question qui se pose est alors la suivante : normo-entendants et mal-entendants ont-ils les mêmes performances en lecture labiale ? Il est difficile de répondre précisément car les différences individuelles au sein de chaque groupe sont assez larges même si les labiolecteurs mal-entendants ou sourds sont relativement meilleurs que les normo-entendants. Comme nous l'avons décrit précédemment, Bernstein *et al.* (2000) ont comparé les performances de 96 étudiants normo-entendants et 72 étudiants sourds profonds. Ces auteurs retrouvent des performances extrêmement variables selon les individus, dans les deux groupes, mais démontrent clairement que les meilleurs lecteurs labiaux sont sourds. Dans une autre étude, Bernstein *et al.* (2001) tentent d'explorer aussi cette possibilité sur des sujets adultes sourds et entendants participant à une étude d'apprentissage de la lecture labiale. Les résultats montrent que, malgré le maigre gain apporté par l'apprentissage, les individus sourds gardent leur supériorité sur les individus bien entendants. Cependant, avec un tel nombre d'individus, peut-on généraliser ces conclusions à toutes les études ? Plus récemment, Auer et Bernstein (A paraître) ont examiné les scores de la lecture labiale d'un ensemble plus large de participants adultes sourds et bien entendants : 112 sourds et 220 normo-entendants. Il en résultait un score moyen de 43,6% (écart type 17,5) de mots corrects pour les participants sourds, et de 18,6% (écart type 13,2) pour les participants bien entendants. Ces scores montrent un grand écart entre les deux groupes de participants et vient confirmer la supériorité des sourds à lire sur les lèvres ; résultat aussi confirmé en Anglais britanique par (Mohammed *et al.*, 2005).

Bien que les personnes sourdes sont des labiolecteurs avec des performances relativement meilleures, la lecture labiale reste une tâche difficile. L'identification de tous les phonèmes d'une langue par l'information visuelle, la seule disponible pour les sourds, n'est généralement pas complète. La lecture labiale est, comme nous l'avons vu précédemment, pleine d'ambiguïté et ne permet pas de différencier tous les phonèmes. Ceci cause évidement un problème majeur pour les enfants sourds, sachant que ces derniers dépendent en grande partie de la lecture labiale pour acquérir le langage. Plusieurs recherches ont démontré en effet que les enfants acquièrent et développent des codes phonologiques lors de leur apprentissage du langage. Ces codes sont dérivés au moins partiellement de la lecture labiale. Les enfants sourds ont aussi besoin de codes phonologiques (Dodd, 1977, 1987; Alegria *et al.*, 1992; Leybaert et Alegria, 1995). Cependant, reposant sur une lecture labiale ambiguë, ces codes sont, pour la grande majorité des enfants sourds, incomplets. Ces enfants développent alors un langage oral déviant et retardé par rapport aux enfants entendants (Dodd, 1976, 1987; Alegria *et al.*, 1992). La lecture labiale seule est donc un support certes important mais insuffisant pour une communication complète et pour le développement du langage chez les enfants sourds d'où la nécessité d'un système complémentaire. L'utilisation de la main va permettre ce complément dans le cadre de la Langue Française Parlée Complétée (LPC).

1.2 La Langue Française Parlée Complétée

La lecture labiale, bénéfique pour une meilleure compréhension de la parole, est incomplète et sans le son ou sans information sémantique, il est impossible de récupérer les traits phonétiques de nasalité (distinction [p] vs [m] par exemple) et de voisement (distinction [p] vs [b] par exemple). Ceci, comme nous l'avons vu précédemment, résulte de l'ambiguïté de la lecture labiale due à la présence de sosies labiaux. Ainsi, en raison des informations insuffisantes apportées par la lecture labiale, l'enfant sourd ne peut acquérir et maîtriser la langue en s'appuyant seulement sur l'éducation oraliste classique. Plusieurs techniques ont vu le jour pour pallier ce manque d'information. En général, ces techniques se basent sur des indices visuels, souvent codés avec la main, permettant d'apporter l'information complémentaire. Le *Cued Speech* (CS) est l'une de ces techniques. Dans cette section nous présenterons d'une part le CS avec un historique rapide de son développement, et d'autre part, nous exposerons la description des principes de sa construction. Dans un autre volet, nous nous focaliserons sur l'adaptation du CS à la langue Française. Finalement, nous rapporterons quelques études menées dans le domaine de la production et la perception du CS et montrant l'efficacité de ce système dans les communications face à face pour les personnes mal-entendantes, ainsi que son rôle dans le développement de la langue chez les enfants sourds.

1.2.1 Le *Cued Speech* : Définition et historique

Il n'y a pas mieux que les mots de l'auteur pour définir son invention :"*A system for support of speechreading ...*" (Cornett, 1967), p.6. Le *Cued Speech* est un système utilisant la main en complément de la lecture labiale, inventé par le Dr. Cornett et qui a pour objectif de permettre aux personnes déficientes auditives l'accès au langage parlé. Les raisons de la création du *Cued Speech* viennent de deux grands souhaits du Dr. Cornett. En effet, en étant le vice-président de la première université pour les déficients auditifs aux USA, le Gallaudet collège à Washington, Cornett découvre que les enfants ayant des déficiences auditives profondes et prélinguistiques ont une faible compréhension de la lecture. Il a compris que l'accès au langage parlé est entravé par leur handicap auditif et que la re-éducation purement oraliste est insuffisante et donne des résultats peu satisfaisants. Par ailleurs, les systèmes gestuels qui existaient tel la langue des signes ne font qu'éloigner les personnes sourdes de la langue utilisée. Il a donc eu l'idée de créer le CS, un compromis entre les communications oralistes et gestuelles. Ce système permet aux personnes sourdes de disposer d'une représentation phonologique complète de la langue leur permettant ainsi de développer des compétences en lecture comme en écriture comparable à celles d'une personne normo entendante (Leybaert et Charlier, 1996). Ainsi, Cornett espère donner la possibilité à l'enfant sourd "*d'acquérir un modèle complet et précis du langage parlé, par le canal visuel*" (Destombes, 1982), p.6. Le deuxième souhait du docteur Cornett relève des aspects de la vie sociale. Il souhaite réduire les obstacles qui compliquent la communication initiale entre les enfants sourds et leurs parents sachant que 90% de ces enfants ont des parents normo entendants (Périer, 1987).

Le *Cued Speech* défini pour la langue anglaise utilise douze clés (ou "*Cues*" en Anglais) de façon à ce qu'elles fournissent suffisamment d'information complémentaire à la lecture labiale

pour permettre une identification précise des phonèmes. Ces clés utilisées seules sans la lecture labiale sont confuses.

Dans le système du *Cued Speech*, le locuteur ou l'utilisateur pointe des positions précises autour du visage en présentant des formes (ou configuration) de main bien définies. Les positions de la main sont utilisées pour coder les voyelles tandis que les formes servent à coder les consonnes. Cornett avait construit le système en se fixant deux critères majeurs : le contraste visuel fourni doit être maximum tout en codant avec un effort minimum.

Ainsi, pour économiser l'effort, le nombre de configurations et de positions est limité de sorte que chaque position de la main identifie un groupe de voyelles, et que chaque configuration de la main identifie un groupe de consonnes. La version finale définie par Cornett pour l'Anglais américain s'appuie sur quatre positions de main et huit configurations (quatre groupes de voyelles et huit groupes de consonnes, voir figure 1.3). Les phonèmes de chacun de ces groupes peuvent être différenciés par les lèvres. De même, les phonèmes qui seraient confondus entre eux au niveau labial seront distingués par l'utilisation des clés différentes. Par exemple, les consonnes occlusives [p], [b] et [m] (qui ont la même forme aux lèvres) seront codées respectivement par les configurations 1, 4 et 5. Ainsi, l'information de la main (position ou configuration) et la forme labiale permettent l'identification d'un percept unique.

D'un autre coté, la maximisation du contraste dans chaque groupe de phonèmes impose un regroupement adéquat. Cornett a utilisé les groupes de visèmes établis au préalable par Woodward et Barber (1960) pour grouper les consonnes en ensembles visuellement contrastés. Les cas non traités par Woodward et Barber (1960) ont été classées par des choix empiriques. Les tables de fréquence établies par Denes (1963) ont également servi au regroupement des phonèmes. L'utilisation de ces tables a pour objectif de minimiser l'énergie pendant le codage et de faciliter les mouvements de la main pour les combinaisons des consonnes les plus fréquentes dans la langue. Ainsi, les consonnes les plus fréquentes comme [m], [t] et [f] sont codées par les configurations les plus faciles à exécuter et qui nécessitent le moins d'énergie, ici la configuration n° 5 pour notre exemple. Concernant les voyelles, vu que le groupement en visèmes est délicat (notamment à l'époque de la création du *Cued Speech*), Cornett s'est appuyé sur des traits visuels comme l'ouverture, l'arrondissement et l'étirement pour bien répartir les voyelles au sein de chaque position. Reste le cas des *diphtongues*, ils sont codés par des *glides* (glissements) de la main entre deux positions de voyelles. Dans certains cas comme pour coder le mot "papa" où le code se répète, l'ensemble main-bras effectue un mouvement avant-arrière pour indiquer une répétition du code. Finalement, les mouvements impliquant le codage du *Cued Speech* peuvent être résumés en 4 mouvements :

- Le déplacement de la main d'une position à l'autre
- Le changement de configuration de la main
- Le glissement entre deux positions pour coder les *diphtongues*
- Un mouvement léger d'avant arrière de l'ensemble bras-main pour indiquer une répétition du code

Dans le but de produire et transmettre les clés du *Cued Speech* à un débit proche de celui de la parole non codée (Attina, 2005) relève que le débit d'une parole non codée, est un peu moins de 5Hz), le *Cued Speech* est par définition un système syllabique où la syllabe CV est

considérée comme l'unité fondamentale. Ainsi, la main dans un mouvement simultané pointe une position et présente une forme précise pour fournir le code de la consonne C et celui de la voyelle V pour la syllabe CV. L'association de la position et de la configuration de la main permet à l'interlocuteur (récepteur), par exemple, de différencier la syllabe [ma] de [pa] ou de [mi]. Dans les cas particuliers d'une consonne isolée ainsi qu'une voyelle isolée, une position et une configuration neutres ont été prévues. En effet, une consonne isolée est codée par une main présentant la forme correspondante et pointant la position coté, la position "neutre". De même, une voyelle isolée est codée par une main en position correspondante avec la configuration n° 5, la configuration "neutre".

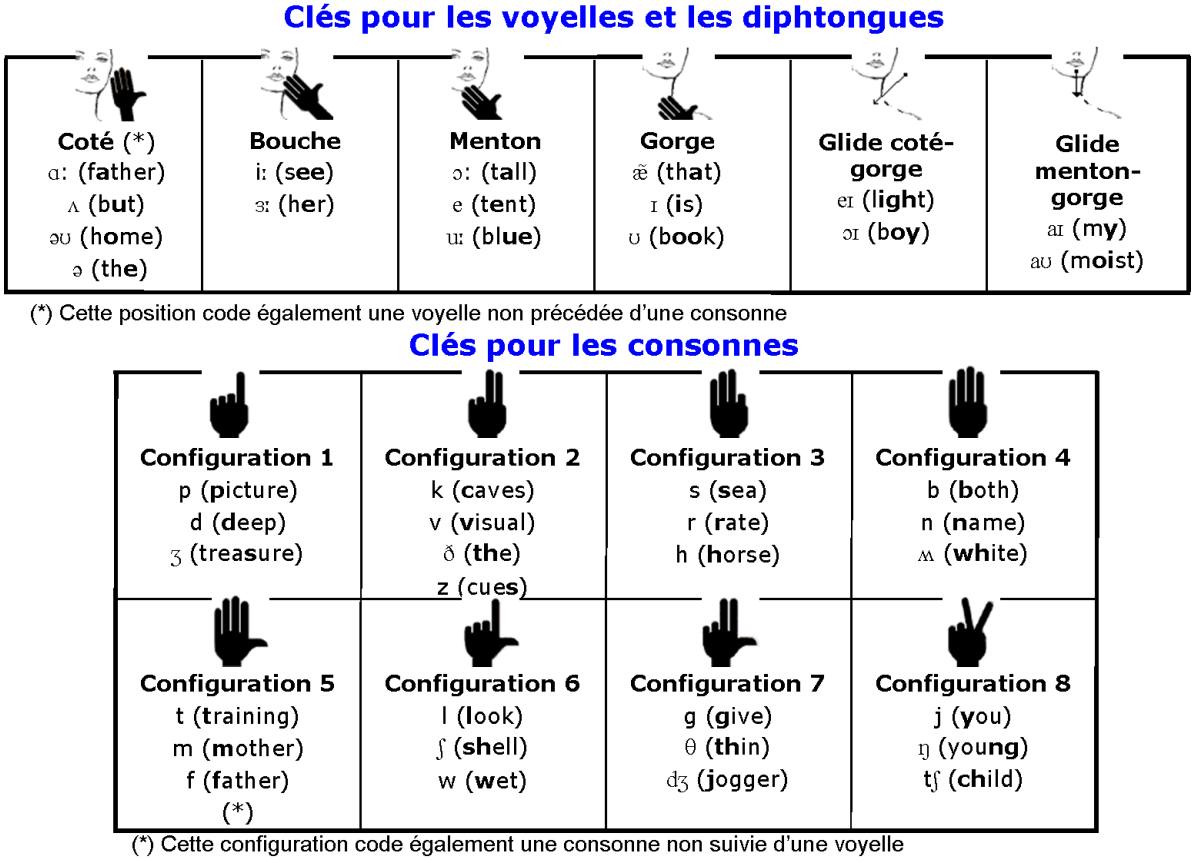

FIG. 1.3 – Clés manuelles du *Cued Speech* conçu par Cornett pour l'anglais américain.

Rappelons que pour résoudre la problématique de l'ambiguïté de la lecture labiale, plusieurs autres systèmes manuels complémentaires ont été développés. Par exemple, Wouts (1982) a crée l'Alphabet des Kinèmes Assistés (AKA). C'est un système gestuel syllabique qui traduit chacun des mouvements bucco-faciaux identifiables (on les appelle des kinèmes) en une série de gestes facilement reconnaissables qui sont codés à coté de la bouche, mais qui de plus traduisent en même temps les mouvements de la phrase et de l'intonation. Chaque kinème comprend souvent plusieurs phonèmes. Il s'agit donc d'un codage des caractéristiques articulatoires des phonèmes (chaque phonème a son code). S'appuyant sur une approche phonétique, l'AKA favorise donc l'apprentissage des personnes sourdes au langage parlé. L'inconvénient de ce système est la

complexité de son acquisition. En effet, l'apprentissage du AKA est plus complexe et nécessite beaucoup plus de temps. Ceci rend compliqué la tâche des parents normo-entendants souhaitant communiquer avec leurs enfants sourds. A cet effet, une dizaine d'heures peut être suffisante pour apprendre le *Cued Speech*. La simplicité du *Cued Speech* est la principale raison qui a poussé Cornett à abandonner l'approche basée sur la phonétique pour développer le *Cued Speech*, malgré l'efficacité avérée de l'approche "phonétique" chez les enfants sourds.

1.2.2 LPC : la version française du *Cued Speech*

Le *Cued Speech* a été adapté à plus de 60 langues et dialectes à travers le monde entier, dont la langue française. Le *Cued Speech* a été importé en France vers 1977. Son adaptation a été dénommée la première fois le Langage Codé Cornett (L.C.C.). Puis, l'Association pour la promotion et le développement du LPC (l'A.L.P.C.) l'a évolué vers Langage Parlé Complété (le L.P.C.). Et récemment le nom a changé en Langue française Parlée Complétée (L.P.C.) afin d'insister sur le fait que le code LPC est basé complètement sur la langue française.

Hérité du *Cued Speech*, le code LPC a maintenu les critères principaux à sa construction, à savoir la maximisation du contraste visuel dans chaque groupe de clés LPC et la minimisation de l'effort. Le principe de fonctionnement reste ainsi le même. L'unité de codage est toujours la syllabe CV. Cependant, l'adaptation à la langue française suscite quelques changements par rapport au regroupement des phonèmes et au nombre de clés (voir figure 1.4). Précisément, cinq positions de la main sont utilisées pour coder les voyelles et huit configurations de la main sont utilisées pour coder les consonnes. Il n'y a par ailleurs aucun changement en ce qui concerne les règles pour les phonèmes isolés : la position "'côté"' reste toujours la position "neutre" et la configuration n°5 (main complètement déployée) reste aussi la configuration "neutre". L'ensemble des mouvements mentionnés pour le *Cued Speech* restent donc les mêmes à l'exception du glissement entre deux positions pour coder les *diphtongues* inexistants en Français.

Enfin, il est très important de mentionner que le code LPC s'appuie sur la transcription phonétique de ce qui est prononcé et non sur l'orthographe du mot. Ainsi, un locuteur communicant avec des personnes sourdes code tout ce qu'il prononce (de la syllabe à la phrase en passant par le mot), tout en tenant compte des liaisons entre les mots. Ceci permet donc de transmettre sans ambiguïté ni confusion tous les contrastes phonologiques.

1.2.3 Efficacité du code LPC

1.2.3.1 Efficacité perceptive

L'intérêt du code réside dans son efficacité à améliorer la perception de la parole. D'ailleurs, c'est la raison principale pour laquelle ce système a été inventé. L'expansion du *Cued Speech* dans le monde entier par ses adaptations aux différentes langues et son utilisation croissante dans plusieurs milieux (en famille ou à l'école), témoignent de l'efficacité de ce système pour une bonne réception de la parole. Le site internet de l'ALPC présente des témoignages concrets de parents utilisant le code LPC pour communiquer avec leurs enfants sourds et qui atteste de l'apport perceptif du code LPC.

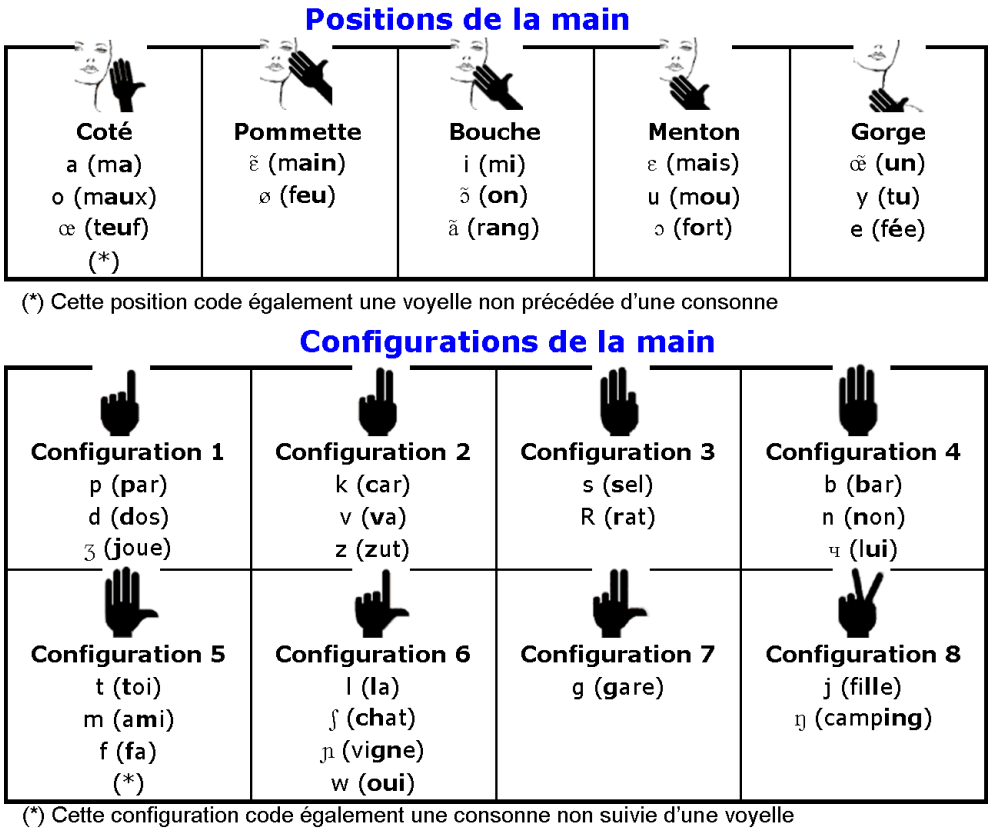

FIG. 1.4 – Clés manuelles du code LPC.

Sur le plan expérimental, plusieurs études ont été menées sur la réception des différentes versions du *Cued Speech* dans le monde. Pour le *Cued Speech*, dans sa version originale nous pouvons citer les travaux de Ling et Clarke (1975), Clarke et Ling (1976), Nicholls et Ling (1982), et Uchanski *et al.* (1994). Pour la version française du code nous trouvons par exemple les études de Charlier *et al.* (1990) et Alegria *et al.* (1999). Nous nous contentons par la suite de décrire deux de ces travaux ((Nicholls et Ling, 1982) ; (Alegria *et al.*, 1999)). La description des autres travaux est présentée en annexe B.

Dans une étude considérée comme la première véritable étude systématique relative à la perception du *Cued Speech*, Nicholls et Ling (1982), testaient la réception dans sept conditions différentes de plusieurs messages par 18 enfants sourds âgés de 9 à 16 ans. Les enfants participant, choisis parmi des élèves sourds d'une école australienne où le *Cued Speech* était pratiqué depuis une dizaine d'années, étaient exposés au moins 4 ans au *Cued Speech* avant cette expérience. Les messages perçus par ces sujets étaient de deux types. Il s'agissait soit de syllabes de type consonne-voyelle (CV) ou voyelle-consonne (VC) soit de mots clés dans des phrases. Les stimuli syllabiques étaient construits par la combinaison de 28 consonnes et les 3 voyelles [a], [i] et [u]. Les mots clés étaient insérés à la fin de phrases simples dans deux contextes sémantiques : soit dans un contexte pouvant faciliter la prédiction du mot clé (*High-Predictability*, HP) ou non (*Low-Predictability*, LP). Par exemple, la prédiction du mot "purse" peut être considérée comme facile dans le contexte "Mum's money is in her purse". En revanche, la tâche semble moins facile

Conditions	A	L	C	AL	AC	LC	ALC
Résultats sur les syllabes en %	2.3	30	36	35	39	83,5	80,4
Résultats sur les mots dans les phrases LP en %	2,5	32	50	47,8	68,8	96,2	96
Résultats sur les mots dans les phrases HP en %	0,9	25,5	42,9	42	59,2	96,6	95

TAB. 1.3 – Pourcentages de bonne réception des syllabes et des mots clés obtenus par Nicholls et Ling (1982) dans chacune des conditions de présentation

pour le mot "room" dans le contexte "Go in that room". Les auteurs ont présenté les deux types de stimuli dans sept conditions expérimentales différentes : soit avec l'audio, la lecture labiale ou les clés du *Cued Speech* seules (conditions respectivement A, L et C) soit en combinant deux des trois conditions précédentes (AL, LC ou AC) soit les trois conditions ensemble (ALC). La table 1.3 résume les scores moyens de perception obtenus dans cette expérience.

D'abord, en ce qui concerne la réception des syllabes, la première chose à noter, est qu'en condition "audio seul" les scores sont très faibles. Ce qui est tout à fait attendu puisque les sujets testés dans cette expérience ont une surdité profonde. Deuxièmement, on retrouve un résultat déjà évoqué précédemment dans ce chapitre en section précédente, qui concerne l'effet du contexte vocalique sur l'identification labiale des consonnes. On remarque en effet, qu'en condition L, AL, LC et ALC (conditions où la lecture labiale est utilisée) les résultats en contexte vocalique [a] et [i] sont meilleurs qu'en contexte arrondi [u]. Enfin, et c'est le résultat le plus important à tirer de cette expérience, les performances de perception sont significativement meilleures quand la lecture labiale est associée au code manuel du *Cued Speech*. Effectivement, on remarque que le pourcentage des syllabes correctement identifiées est nettement supérieur (de plus de 40% au moins) en conditions LC et ALC que dans les autres conditions.

Par ailleurs, les mêmes remarques peuvent être formulées dans le cas de la réception des mots clés. Des exceptions sont tout de même à remarquer. Les performances sous la condition AL sont significativement supérieures à celles des conditions A et L seules. De même, les performances sont significativement supérieures en condition AC qu'en condition A et C seules. Par contre, les scores moyens des conditions LC et ALC ne sont pas différents et restent supérieurs à tous

les autres scores. L'ambiguïté ou non de la réception de la modalité visuelle peut expliquer les différences entre ces résultats. Si les stimuli sont présentés en condition visuelle dans laquelle la réception des phonèmes est ambiguë (condition L ou C), le son est intégré au geste dans la limite de sa perception par les sourds profonds. En revanche, si le message visuel ne présente aucune ambiguïté (en condition LC) le son n'apporte aucun bénéfice à la perception.

D'un autre coté, les auteurs relèvent une interaction entre les niveaux de prédictibilité contextuelle (HP ou LP) et les conditions L, AL, C et AC (conditions où l'information visuelle est ambiguë). Dans ces conditions, les scores sont en effet supérieurs en niveau HP par rapport au niveau LP. Les différences de prédictibilité sont en revanche non significatives dans les conditions LC et ALC. Ceci laisse supposer que l'ambiguïté visuelle du message est atténuée par le contexte sémantique.

Finalement, on peut conclure de cette étude que la parole peut être reçue de manière claire et exacte uniquement par la voie visuelle et sans émettre aucun son. Le contexte sémantique peut contribuer à la compréhension du message notamment quand celui-ci comporte des ambiguïtés visuelles. Dans cette étude l'efficacité du *Cued Speech* à améliorer la perception de la parole pour les sourds est belle et bien établie. On peut donc affirmer qu'une personne sourde profonde perçoit la parole codée aussi bien qu'une personne bien entendante perçoit la parole orale.

Les travaux de Alegria *et al.* (1999) avaient deux objectifs : approfondir la notion que le LPC améliore la lecture labiale et explorer la façon selon laquelle les informations provenant de ces deux flux se combinent. Dans cette partie, seul le premier objectif nous intéresse. Pour cet objectif, des mots et pseudo-mots étaient présentés à 31 enfants sourds avec des caractéristiques différentes selon l'age, la durée et le début d'exposition au code LPC. Les participants étaient répartis en deux groupes. Le groupe "LPC-précoce" était constitué de 7 enfants âgés de 8 à 12 ans et qui étaient exposés au code LPC avant l'âge de 2 ans durant une durée moyenne de 9 ans et 5 mois. Le reste des enfants (24 enfants) constituaient le second groupe appelé "LPC-tardif". Les enfants de ce groupe, âgés de 11 à 19 ans, étaient exposés au code LPC durant une période moyenne de 6 ans et 5 mois après l'âge de 2 ans. Les mots et pseudo-mots présentés étaient composés de 4 phonèmes suivant 4 structures différentes : CV-CV, VC-CV, V-CVC, V-CCV. Les auteurs avaient utilisé 8 mots et 8 pseudo-mots par structure ; ce qui constitue 64 combinaisons. Chaque combinaison était présentée deux fois avec ou sans l'aide du code LPC. Les résultats obtenus dans cette expérience sont présentés dans la figure 1.5.

A première vue, on remarque que l'utilisation du code LPC dans les deux conditions (exposition précoce ou tardive) améliore significativement les performances de réception des mots et des pseudo-mots. Toutefois, on remarquera aussi quelques différences significatives entre la réception des deux types de combinaisons. La réception des pseudo-mots est en effet inférieure à celle des mots pour les deux groupes de participants. Ceci peut s'expliquer par le fait que pour identifier un pseudo-mot les sujets s'appuient seulement sur l'information phonologique issue de la lecture labiale et le code LPC, tandis que pour déterminer un mot ils font appel aussi à leurs connaissances lexicales. L'effet "lexical" peut expliquer en partie l'infériorité des pourcentages des pseudo-mots obtenus par le groupe "LPC-tardif" par rapport à ceux du groupe "LPC-précoce". En général, les performances des enfants exposés précocement au code LPC sont supérieures à celles des enfants tardivement exposés. Ce résultat est confirmé par d'autres

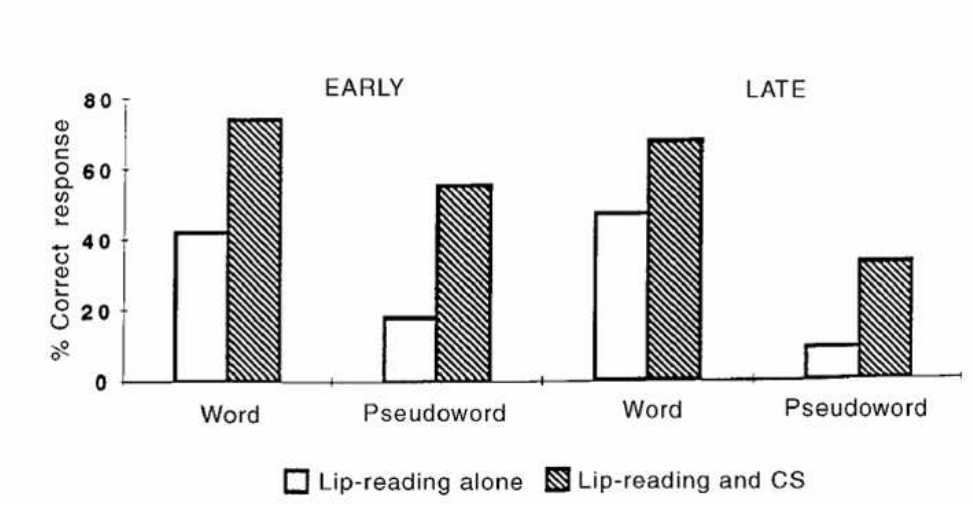

FIG. 1.5 – Pourcentages moyens obtenus par Alegria *et al.* (1999) dans son expérience en perception de mots et de pseudo-mots en condition de lecture labiale seule ou avec les clés du LPC.

études montrant l'importance de la durée d'exposition au code (Clarke et Ling, 1976; Périer, 1987)

En conclusion, la lecture labiale ne transmet qu'une partie de l'information. L'autre partie peut être transmise aussi visuellement par les clés du code LPC. L'ambiguïté de la lecture labiale peut donc être réduite par le code LPC. Ainsi, la perception d'un message de parole codé par les deux modalités visuelles ne diffère guère de la perception de la parole orale non codée. Ceci indique que les enfants sourds peuvent bénéficier de ce code pour développer leur langage oral avec des performances similaires aux personnes entendantes.

1.2.4 Efficacité sur le développement du langage parlé

En plus de son efficacité dans la perception d'un message oral, le code LPC permet l'accès à une représentation complète du système phonologique pour les malentendants exposés à cette méthode depuis leur plus jeune âge, avec un impact positif sur le développement du langage. C'est ce qui a été montré dans plusieurs études (notamment : (Kipila, 1985; Cornett, 1990; Hage *et al.*, 1991; Leybaert et Charlier, 1996; Leybaert, 2000; Charlier et Leybaert, 2000; Leybaert et Lechat, 2001)) nous présenterons brièvement les conclusions en annexe B (pour une revue détaillée voir Attina (2005) p. 28-30 ; Alegria et Leybaert (2005)). Ces conclusions concernent spécialement le rôle positif du *Cued Speech* pour :

- Le développement du langage
- L'apprentissage de l'écriture avec le développement de la phonique et l'orthographe
- L'apprentissage de la lecture
- Le développement de la mémoire de travail

Sans aller plus loin dans la description de ces expériences, nous pouvons dire que le code LPC est avantageux pour les enfants sourds dans la mesure où il permet une acquisition de connaissances

des mots (surtout ceux ayant une signification) et aussi de la morpho-phonologie de la langue parlée. Avec l'aide du code LPC, les enfants développent des représentations phonologiques précises de la parole qui leur permettent de juger des rimes correctes et de les produire. Enfin, les enfants exposés au code LPC peuvent développer des compétences en lecture et en orthographe similaires à celles des enfants entendants.

1.3 Coordination temporelle main-lèvres en Langue Française Parlée Complétée

Le fait que les clés manuelles (positions et formes de main) doivent être associées à la forme des lèvres pour que le code LPC soit efficace nécessite probablement une véritable coordination entre la main et les lèvres. Depuis son invention et hormis quelques indications aidant la pratique du code, aucune véritable étude n'a traité la production de ce système jusqu'aux travaux précurseurs de Attina *et al.* (2002). Sur le plan perceptif, il existait quelques études menées spécialement par Alegria *et al.* (1999) qui se sont intéressés à la combinaison des informations labiale et manuelle chez un décodeur de code LPC pour l'identification d'un percept unique.

L'étude de Alegria *et al.* (1999) visait, comme nous l'avons évoqué un peu plus haut dans ce chapitre, à étudier l'apport déterminant du code LPC à la lecture labiale et le processus d'intégration des informations provenant de la main et des lèvres dans une tâche perceptive d'identification des mots et pseudo-mots par des sujets sourds. Pour étudier cette intégration, les auteurs analysaient les erreurs sur les pseudo-mots, qui sont relatives aux caractéristiques du code LPC. Précisément, deux types d'erreurs les intéressaient. Le premier type concernait les erreurs de substitution des phonèmes dans chaque clé LPC et qui sont liées à la structure du code. Par exemple, le sourd peut percevoir la consonne [m] alors que c'est la consonne [t] qui a été codée et ceci car les deux consonnes partagent la même configuration de main (voir figure 1.4). Ceci signifie alors que le sourd perçoit l'information issue de la main sans intégrer celle des lèvres. Les auteurs ont analysé une sorte de matrice de confusion stimuli-réponses des pseudo-mots. Ils ont déterminé ainsi pour chaque phonème la fréquence des substitutions relative au nombre total des erreurs tout en séparant les voyelles et les consonnes. Les résultats pour des sourds exposés précocement et tardivement au code LPC montrent une tendance de substitutions LPC plus visibles dans le cas des consonnes (c'est-à-dire au sein des configurations de la main) que dans le cas des voyelles (i.e. les positions de la main). Le second type d'erreurs concernait la structure syllabique du code LPC. En effet, comme le code LPC est basé sur une organisation syllabique de type CV (unité du code LPC), les structures syllabiques "non canoniques" (en anglais : *non canonical*) comme (VC-CV, V-CVC et V-CCV qui contiennent 3 unités LPC) sont codées avec un nombre de clés supérieur à celui nécessaire pour coder une structure "canonique" de type CV-CV (composé seulement des unités LPC). Ainsi, dans son test de perception, le sourd peut décoder une voyelle ou une consonne supplémentaire (qui n'était pas réellement produite) et l'intégrer dans les structures non canoniques. Par exemple, la structure V-CCV peut être perçue avec une voyelle supplémentaire de type V-C[V]CV. Les résultats indiquent que le code LPC aide à déterminer un nombre exact de syllabes dans une suite de syllabes canoniques (donc

de type CV) tandis que dans le cas non canonique il est entravé par la tendance à interpréter les clés supplémentaires comme des syllabes supplémentaires. Pour expliquer ces erreurs, les auteurs font l'hypothèse que lorsque la visibilité labiale de certains segments est basse (au niveau de l'intelligibilité), le sourd s'appuie seulement sur les clés LPC indépendamment de la lecture labiale. A partir de ces résultats, les auteurs font l'hypothèse que les principes de l'intégration main-lèvres en perception sont similaires à ceux observés en traitement de la parole audio visuelle. Deux modèles contrastant avec les modèles de traitement de la parole sont ainsi proposés. Dans le premier modèle, de type hiérarchique, l'information portée par les clés LPC vient tardivement enlever les ambiguïtés de la lecture labiale. Un modèle qui traduit en effet la définition initiale du code LPC. Dans le second modèle, les deux informations sont intégrées avec des poids équivalents de la même manière que sont traitées les informations visuelles et auditives chez les personnes normo-entendantes. En d'autres termes, le code LPC serait considéré dans ce modèle comme une deuxième entrée de même poids que la lecture labiale dans un système de traitement automatique de la parole. Dans sa conception, ce modèle est similaire à la métrique commune de Summerfield (1987) ("common metric"). Pour résumer, suivant les deux modèles, les clés manuelles sont soit considérées comme des signaux "artificiels" qui viennent en seconde passe après la lecture labiale, soit elles sont intégrées avec l'information labiale de façon équivalente. Cependant, peut-on exclure la troisième option qui consiste à dire que la lecture labiale vient désambiguïser la première information phonologique apportée par les clés LPC ? La réponse est donnée par Attina et ses collègues (Attina *et al.*, 2002, 2004; Attina, 2005) avec l'analyse de la production du code LPC.

Les travaux pionniers dans ce domaine et menés à l'Institut de la Communication Parlée (ICP) consistent à déterminer comment le mouvement de la main co-produit l'information sur la consonne et la voyelle en LPC. Pour une première réponse à cette question, Attina *et al.* (2004) (voir aussi Attina *et al.* (2002)) se sont focalisés sur l'organisation temporelle des clés manuelles en relation avec le mouvement des lèvres et le signal acoustique correspondant. Dans ces études, à partir de l'analyse d'un codeur LPC diplômé en LPC, Attina *et al.* (2004) montrent une avance du début du mouvement de la main d'une valeur moyenne de 200 ms sur la réalisation acoustique de la syllabe CV. Par ailleurs, les données montrent une superposition complète du geste de formation de la configuration digitale sur le geste propre de la main. Dans leurs expériences, les auteurs suivent la position bi-dimensionnelle (2D) d'un point (marqué par une pastille facilement repérable) placé au milieu du dos de la main (voir figure 1.6) à partir des images vidéo d'un enregistrement d'un codeur LPC ainsi que l'aire intérolabiale aux lèvres. Les mouvements de la main sont caractérisés par des transitions lentes entre plateaux selon les trajectoires x et y au cours du temps de la position 2D. Les auteurs délimitent les plateaux par la position du maximum de décélération ($M2$) pour le début et le maximum d'accélération ($M3$) du début de la transition suivante qui marque aussi l'extrémité du plateau le long des axes x et y (figure 1.7). Sur cette figure, on peut remarquer certains événements retenus par les auteurs pour décrire les différents signaux. Nous présentons la nomenclature suivante utilisée par les auteurs dans ces expériences, et que nous reprendrons dans nos propres travaux :

sur le signal acoustique

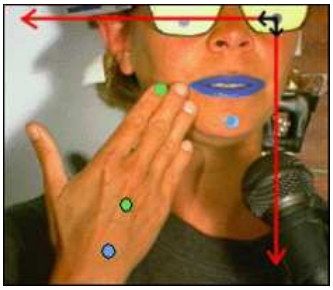

FIG. 1.6 – Image du codeur LPC avec les axes x et y en superposition (Attina *et al.*, 2004).

- $A1$: début de la réalisation acoustique de la consonne
- $A2$: début de la réalisation acoustique de la voyelle
- $A3$: fin de la réalisation acoustique de la voyelle

sur les signaux caractérisant les mouvements de la main

- $M1$: début de la transition de la main d'une position à l'autre
- $M2$: début de la tenue de la main en position
- $M3$: fin de la tenue de la main

sur les signaux des paramètres labiaux

- $L2$: l'instant de la réalisation vocalique sur les lèvres

La cible de la main est ainsi supposée atteinte lorsque x et y atteignent simultanément leur plateau. Utilisant cette analyse cinématique du geste de la main, les auteurs montrent que la main arrive en position cible de manière quasi synchrone avec le début de la réalisation acoustique de la syllabe CV, donc en début de consonne et ainsi bien avant la réalisation acoustique de la voyelle ; la main quitte sa position vers une nouvelle cible avant le climax vocalique des lèvres (figure 1.8).

Les résultats sur l'avance de la main ont été confirmés par l'analyse de la production de trois codeurs supplémentaires (Attina, 2005). Le patron temporel de coordination main-lèvres de chacune de ces trois codeuses est similaire à celui trouvé pour la première. Mais la question qui se pose, en considérant l'hypothèse d'une liaison entre la perception et la production de la parole (Rizzolatti *et al.*, 1996) est la suivante : la désynchronisation entre les gestes labiaux et manuels est-elle perçue ? Pour répondre à cette question Attina (2005) a étudié à l'aide d'une expérience de *Gating*, l'avance de la main en perception du code LPC par 16 sujets sourds profonds. L'expérience consistait à découper temporellement chaque stimulus en plusieurs points clés et à les présenter progressivement du début jusqu'aux différents points de troncature. Si un stimulus a p points, il est découpé en une série de p+1 présentations du début (P0) à chaque points clés (Pi, i de 1 à p) : de P0 à P1, P0 à P2, ..., P0 à Pp et P0 à la fin. Dans cette expérience, les participants pratiquaient le code LPC quotidiennement sauf deux d'entre eux. Les stimuli testés étaient des structures sans signification composées de 5 syllabes (des logatomes donc) de type [mytymaCVma]. La quatrième syllabe (notée CV) de la structure peut varier et c'est précisément ce que les sujets devaient identifier. Chaque logatome a été tronqué aux alentours de la syllabe CV à identifier en 6 points : le 1er point correspond à l'instant M1, c'est-à-dire à

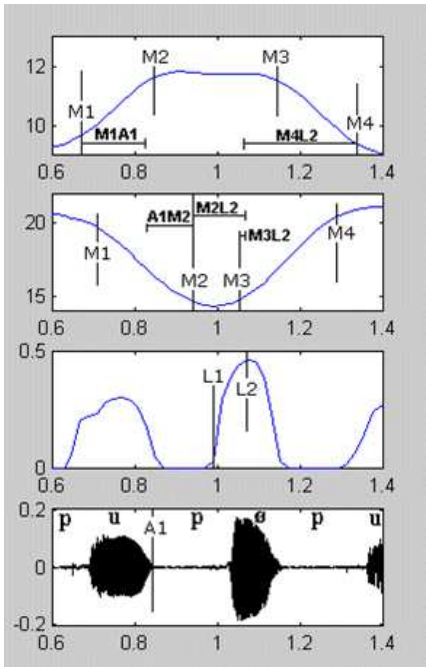

FIG. 1.7 – De haut en bas : (1) trajectoires x (cm) et (2) y (cm) de la main pour une séquence [pupøpu], (3) décours temporel de l'aire intérolabiale S (cm^2) ; (4) signal acoustique correspondant (Attina *et al.*, 2004).

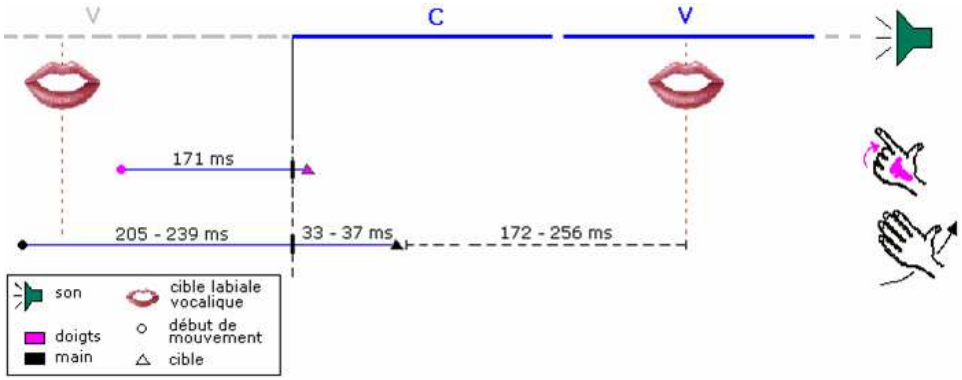

FIG. 1.8 – Schéma général de coordination de la main et des lèvres en relation avec le son de parole pour le code LPC (Attina *et al.*, 2004).

l'instant où la main quitte sa position coté du [ma] et transite vers la position de la syllabe [CV] ; le 2ème point correspond à la configuration de la main pour la consonne C ; le 3ème correspond à l'instant M2, c'est-à-dire au début de l'atteinte de la position cible de la syllabe CV ; le 4ème correspond à l'instant où la configuration et la position de la main sont identifiables ; le 5ème au début du mouvement des lèvres pour atteindre la cible vocalique ; et le 6ème à la réalisation labiale de la voyelle (L2). La tâche des participants était d'identifier la syllabe codée à chaque point de troncature. Il s'avère d'après les résultats que le point n° 4 est le point où les participants identifiaient correctement plus de 80% des groupes manuels des clés LPC. Ce résultat est en effet prédictible du fait qu'à cet instant les informations manuelles (sur la configuration et la position) sont presque toutes visibles et en même temps l'information labiale ne l'est pas encore d'après le patron temporel observé dans la production. Ceci semble confirmer que l'avance de la main sur les lèvres observée en production est récupérée en perception. Ce qui permet aux auteurs ((Attina *et al.*, 2004)) de proposer l'hypothèse suivante : la position de la main pourrait donner tout d'abord un sous-ensemble de voyelles, les lèvres dans un second temps fournissant la solution et ainsi ce serait les lèvres qui désambiguïseraient l'information issue de la main. Ces résultats ont servi de référence pour l'élaboration d'un synthétiseur audio-visuel 2D d'abord (Attina *et al.*, 2004) puis 3D du code LPC (Gibert *et al.*, 2006).

Notons enfin que les résultats d'Attina ont été obtenus sur un corpus de logatomes et par des moyens techniques manuels. Nous tenterons dans les travaux que nous présenterons ensuite de voir si ce patron temporel de l'organisation main-lèvres reste confirmé sur un corpus de phrases. La détection des différents événements relatifs au mouvements de la main se fera de façon automatique.

1.4 Conclusion

En perception de la parole, observer les mouvements des lèvres n'est pas un phénomène accessoire qui accompagne l'audio. Bien au contraire, dans diverses situations, cette observation augmente considérablement l'information perçue de la parole produite. La parole est donc par nature multimodale.

Le dispositif de la communication parlée des personnes normo-entendantes s'appuie ainsi sur la multimodalité audio-visuelle de la parole. Pour les personnes mal-entendantes, la modalité auditive peut ne pas être disponible. Dans ce cas, la modalité visuelle, représentée par la lecture labiale, est donc la seule accessible. Cependant, la lecture labiale, qui est par nature incomplète et ambiguë, ne peut transmettre l'information phonétique complète nécessaire pour la compréhension du message de parole. Pour compléter cette insuffisance de la lecture labiale pour les personnes sourdes ou mal-entendantes, la Langue Française Parlée Complétée (LPC) héritée du *Cued Speech* a été conçue. Dans ce système, le locuteur pointe avec la main des positions précises près du visage en présentant de dos des formes de main bien définies. La main et les lèvres portent chacune une partie complémentaire de l'information phonétique. L'association de ces deux composantes permet à un mal-entendant (ou un sourd) de récupérer un percept unique.

Enfin, il a été démontré, en production et en perception du code LPC, que le geste de la main suit une organisation temporelle spécifique en coordination étroite avec la parole audio-visuelle.

Cette organisation a été obtenue sur un matériel de parole s'appuyant sur des logatomes, avec des techniques qui extraient les informations main-lèvres d'une manière manuelle. Dans nos propres travaux, nous tenterons d'automatiser ces techniques. Dans ce sens, le chapitre suivant présente une revue des approches d'extraction automatique des gestes main-lèvres.

Chapitre 2

Description de l'information labiale et manuelle

Dans la Langue Française Parlée Complétée, l'information phonétique complète que perçoit le décodeur de ce système provient de deux flux visuels complémentaires. Dans un système de reconnaissance automatique, l'extraction de l'information visuelle pertinente est primordiale. Mais comment extraire cette information ? Et qu'elles sont les contraintes (d'angle de vue, de paramétrisation, de complexité ...) à imposer pour une extraction optimale et appropriée ? Il n'y a cependant dans la littérature très peu étude concernant la reconnaissance automatique des gestes manuels et labiaux du code LPC (ou du Cued Speech). Toutefois, les études réalisées sur les systèmes de reconnaissance automatique audiovisuelle de la parole nous offrent de nombreuses pistes pour l'extraction des caractéristiques visuelles de la lecture labiale. La synthèse de la parole visuelle utilise aussi d'autres techniques. En revanche, les travaux concernant la reconnaissance automatique des gestes de la main en LPC sont très rares. De ce fait, les techniques automatiques de reconnaissance des gestes que nous pouvons trouver dans le domaine de la vision par ordinateur ou le domaine des langues des signes peuvent être intéressantes. Dans ce chapitre, nous présenterons les différentes techniques d'extraction des caractéristiques visuelles de la lecture labiale (mesure labiale ou labiométrie), tout en décrivant quelques expériences analysant l'influence de certains facteurs de visibilité du visage du locuteur sur la pertinence de l'information extraite. Ensuite, nous nous focaliserons sur les différentes approches de reconnaissance automatique des gestes.

2.1 Extraction de l'information visuelle

Nous rappelons que l'information visuelle est d'un bénéfice important dans le domaine de la reconnaissance audio-visuelle de la parole. Elle est un vecteur d'information nécessaire et essentiel dans la compréhension, même partielle, de la parole chez les personnes sourdes. Associée aux clés du code LPC, elle porte une partie complémentaire de l'information de parole perçue par les utilisateurs de ce code. La présentation des informations visuelles doit être optimale pour une reconnaissance maximale des gestes visuels. En d'autres termes, dans quelles conditions de présentation et de visibilité du visage, un sujet (ou un système de reconnaissance) peut-il

percevoir (reconnaître) un maximum d'information de parole ?

2.1.1 Influence de l'angle de vue

Dans les tests de perception visuelle de la parole, que nous avons présenté dans le chapitre 1, les auteurs choisissent de présenter leurs stimuli visuels sous des angles de vue différents. Ceci prouve en quelque sorte que l'information visuelle perçue dépend en partie de ce facteur de visibilité. Ce dernier a été l'objet de plusieurs études (parmi lesquels Neely (1956); Larr (1959); Nakano (1961); Berger *et al.* (1971); Erber (1974); Cathiard (1988, 1994); Adjoudani (1998)). A l'exception de l'étude de Adjoudani (1998), utilisant des paramètres extraits des contours des lèvres, toutes ces études, que nous décrirons en annexe D, s'appuient sur des tests perceptifs.

Dans ces études, trois vues ont été comparées : la vue de face, la vue de profil et la vue de 3/4. De ces comparaisons, nous pouvons conclure que :

- la vue de face apporte plus d'information que la vue de profil, à l'exception de certains cas spécifiques concernant la classification des traits labiaux de protrusion et d'étirement (Cathiard, 1988, 1994), où la vue de profil peut être plus efficace que la vue de face.
- La vue de 3/4 est globalement équivalente à la vue de face.

Dans le cas du code LPC, où la main et les lèvres doivent être simultanément visibles, la vue de 3/4 poserait des problèmes de visibilité notamment pour la forme de la main. De même, la vue de profil ne peut permettre la visibilité complète des positions de la main ni des formes. De plus, elle est, en général, moins efficace que les deux autres vues. Il reste donc la vue de face qui, a priori, semble la plus appropriée au cas du code LPC.

2.1.2 Visage complet ou indices visuels ?

Percevoir le visage d'un locuteur apporte bien un gain d'intelligibilité en perception de la parole. Mais quelles sont les parties qui contribuent le plus à ce gain ? Pour répondre à cette question, rappelons d'une part que dans la majorité des expériences décrites au chapitre 1, notamment celles sur la perception visuelle de la parole, le visage complet (et dans certains cas les épaules et la tête) était présenté aux sujets testés. D'autre part, des études ont montré que la région de la bouche transmettait la plus grande partie de l'information visuelle de parole. D'autres études allaient jusqu'à suggérer de se contenter seulement des lèvres. Dans cette section, nous présentons les résultats de quelques études comparant différentes conditions de présentation des stimuli visuels. Summerfield (1979) a comparé les gains d'intelligibilité de différents types d'information visuelle. Il a présenté à 10 sujets (âgés de 15 à 27 ans) des stimuli audiovisuels produits par un locuteur anglais sous forme de phrases, mélangés avec d'autres signaux de parole, dans cinq conditions différentes :(i) signal acoustique seul, (ii) signal acoustique + le visage du front à la mandibule, (iii) signal acoustique + les lèvres seules, (iv) signal acoustique + 4 points lumineux placés autour des lèvres sur les coins et sur les intersections de l'axe de symétrie avec les lèvres supérieure et inférieure,(v) et signal acoustique + un cercle dont le diamètre varie selon l'amplitude du signal acoustique non bruité. Sous ces différentes conditions les sujets devaient identifier les phrases testées et les noter sur papier. Les résultats obtenus dans cette expérience sont présentées par la table 2.1.

Condition	*Audio seul*	*Audio + visage complet*	*Audio + lèvres*	*Audio + 4 points*	*Audio + cercle*
Pourcentage moyen (%)	22,7	65,3	54	30,7	20,8
Ecart type	8,59	19,7	14,5	16,2	10

TAB. 2.1 – Scores d'identification obtenus par Summerfield (1979) dans cinq conditions de présentation des stimuli.

De ces résultats nous pouvons tirer quelques constats intéressants. Tout d'abord, les deux informations visuelles dans les conditions (iv) et (v) ne semblent apporter aucune information aidant à comprendre les phrases bruitées. Les différences entre ces deux conditions et la condition (i) sont en effet, selon l'auteur, non significatives. Ensuite, il est évident que la présentation de l'image complète ou de l'image des lèvres est bénéfique pour la compréhension du message. Dans les deux conditions, les scores d'identification augmentent en moyenne de plus de 31% par rapport au scores dans la condition audio seul. Et enfin, les lèvres seules portent une information importante mais restent encore inférieures à celle portée par le visage complet. Ces deux derniers constats ont été confirmés par d'autres études (Le Goff *et al.*, 1995, 1996; Adjoudani *et al.*, 1994). Globalement, le visage complet est l'indice visuel qui apporte le plus d'information visuelle. Les lèvres portent une grande partie de l'information visuelle équivalente en quantité à peu près aux deux tiers de celle transmise par le visage complet. L'étude de Summerfield (1983) a porté sur les conditions de présentation des indices visuels pour que l'information visuelle contribue plus pertinemment à la perception audiovisuelle de la parole. Ainsi, il suggérait les conditions suivantes :

- une distance de 1,5m,
- une luminance suffisante,
- le corps et les bras visibles aussi,
- pas de moustache ni de barbe sur le visage,
- et un maquillage des lèvres pour augmenter le contraste.

2.1.3 Approches pour la mesure labiale

La mesure labiale est la mesure des paramètres labiaux d'un locuteur. Il existe de nombreuses techniques en traitement des images permettant d'extraire ces paramètres. Généralement, ces techniques peuvent être classifiées en deux grandes approches : l'approche "image" et l'approche "modèle".

2.1.3.1 Approche "image"

L'approche image consiste en des transformations appliquées sur les pixels d'une fenêtre d'analyse contenant la bouche. Souvent, l'image de la bouche en niveau de gris est utilisée directement (ou après certains prétraitements) comme vecteur de l'information. Cette approche

a des avantages et des inconvénients. Le grand avantage de cette méthode est qu'elle garantit une dégradation minime des données et donc peu d'informations sont perdues. Les informations labiales les plus importantes tels que l'arrondissement, la protrusion, la présence des dents ou de la langue, sont en effet déterminées par le classifieur. Inversement, deux grands inconvénients sont à signaler. D'un coté, les systèmes fondés sur cette approche nécessitent d'apprendre des modèles pour l'ultime classification et utilisent pour ceci des techniques d'apprentissage tels que les réseaux de neurones ou la quantification vectorielle. Or, si le nombre de paramètres du système de reconnaissance augmente, ce type d'approche nécessite une quantité importante de données pour apprendre les modèles. De l'autre coté, cette approche n'est pas très robuste à la variation de l'éclairage. Si par exemple la couleur ou la direction de l'éclairage changent, les valeurs des pixels changent aussi. Toute translation, rotation ou changement de graduation change aussi les valeurs des pixels et par conséquence dégrade fortement la reconnaissance. Sans oublier que ce type de méthode est aussi peu robuste face à la variabilité inter ou intra locuteur.

Parmi les techniques d'extraction qui peuvent être affiliées à cette approche, nous trouvons entre autres celles développées par Yuhas *et al.* (1989); Mase et Pentland (1991); Prasad *et al.* (1997); Bregler et Konig (1994); Petajan et Graf (1996); Matthews *et al.* (1996); Gray *et al.* (1997); Potamianos *et al.* (2001). Dans la suite nous décrivons quelques unes de ces techniques.

Prasad *et al.* (1997) ont développé un système pour calculer des paramètres géométriques. Pour une image donnée dans une séquence vidéo, le système consistait d'abord à calculer la différence entre cette image et la suivante. Les deux images étaient auparavant filtrées par un filtre passe bas et seuillées. Ensuite, la région de la bouche (région d'intérêt, *ROI"' "Region Of Interest*) était localisée sur la première image grâce à un seuillage simple qui détectait le centroïde de la bouche. Enfin, sur l'image de cette région, des mesures permettaient d'extraire les caractéristiques géométriques requises.

Petajan et Graf (1996) ont conçu une des techniques les plus connues avec la motivation qu'elle soit une technique applicable en pratique sans aucune contrainte de maquillage des lèvres. Elle exploite deux études faites précédemment par Petajan (1984, 1985) dans lesquelles ce dernier a montré que les narines peuvent être considérées comme deux points faciles à détecter. L'intérêt de détecter ces deux points est de réduire le calcul informatique et d'assurer une certaine robustesse. Cependant, la détection de ces deux points contraint la position de la caméra qui doit être placée légèrement au dessous du visage pour que les narines soient dans le champ de vision de la caméra. Ce système est conçu pour être robuste face aux variations de l'éclairage et aux mouvements de la tête. Ces performances ne sont pas dégradées par la présence de poils sur le visage (barbe ou moustache) ou de lunettes. Dans une première version de ce système, un algorithme de reconnaissance utilisant un filtrage morphologique est appliqué sur chaque image et permet de localiser le visage ainsi que les positions relatives des yeux, du nez et la région de la bouche. Un seuillage chromatique est appliqué ensuite pour détecter les narines (représentées par deux points). Une table LUT (*Look Up Table*) contient des seuils pré-stockés et qui sont nécessaires à ce seuillage. Les positions des narines servent ensuite à former une fenêtre d'analyse autour la région de la bouche. Après avoir équilibré l'image de la bouche pour la rotation, des tests consécutifs sur les seuils du contour interne des lèvres sont appliqués à chaque fois que la bouche est fermée. Une valeur finale de seuil est donc retenue. Enfin, chaque pixel qui est sous

le seuil est étiqueté comme appartenant au contour interne. De cette méthode résultant ainsi des détails sur la bouche incluant des informations sur le contour interne des lèvres, les dents et la langue.

Potamianos *et al.* (2001) ont proposé un algorithme d'extraction de l'information visuelle dans un objectif final de reconnaissance automatique audio-visuelle de la parole. L'algorithme consiste en 3 transformations en cascade s'appliquant sur une vidéo 3D de la région d'intérêt (ROI) qui contient la bouche du locuteur. L'image à traiter passe premièrement par une transformation traditionnelle de l'image à partir des pixels (de type par exemple transformée en cosinus discrète (DCT[1]) ou transformée en ondelette discrète (DWT[2]), etc.) pour compresser les données. Ensuite, une analyse discriminante linéaire est appliquée (*Linear discriminant analysis LDA*) pour optimiser les performances de la classification en réduisant la dimension des données. Et enfin, les données résultantes subissent une rotation en utilisant une transformée linéaire qui optimise la fonction de vraisemblance (sans prendre de décision) construite à partir des données observées. A l'issue de cette dernière transformée, un vecteur de paramètres est obtenu. La figure 2.1 résume le principe de l'algorithme proposé par Potamianos *et al.* (2001).

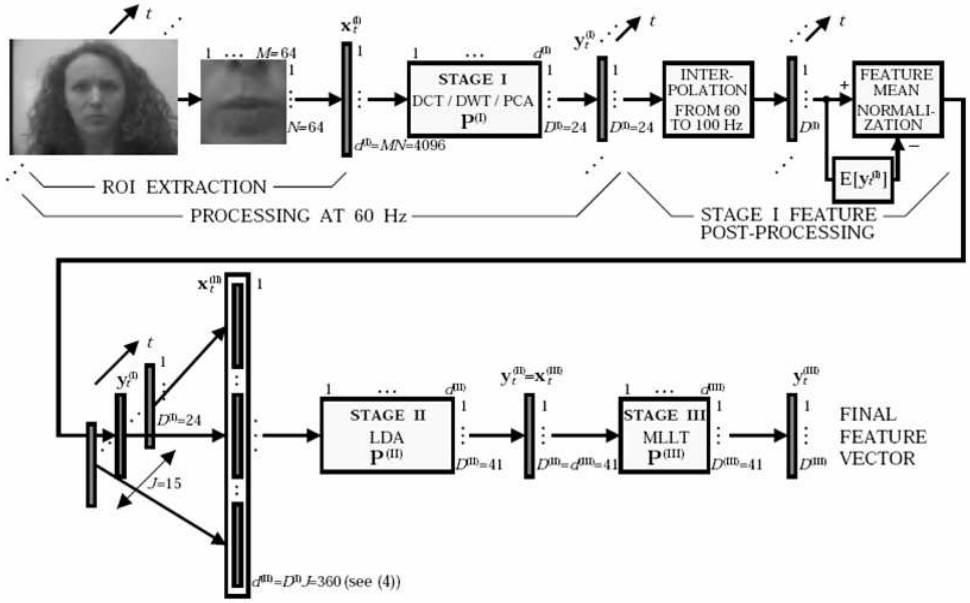

FIG. 2.1 – Diagramme en blocs décrivant l'algorithme en cascade proposé par Potamianos *et al.* (2001).

2.1.3.2 Approche "modèle"

Les méthodes d'extraction automatique des paramètres visuels fondées sur l'approche "modèle" se concentrent sur les articulateurs visibles de la parole que sont les lèvres. En général, dans

[1] *Discrete cosine transform*
[2] *Discrete wavelet transform*

ces techniques, des paramètres décrivant les contours des lèvres sont considérés comme les caractéristiques qui mesurent l'information labiale. En effet, il est bien connu que les lèvres sont à la fois la partie complètement visible et l'extrémité du conduit vocal. Il semble donc que les contours des lèvres portent de précieuses informations sur la parole et sur le locuteur. Summerfield (1979) a démontré que l'information visuelle pertinente sur la parole est contenue dans les lèvres. Techniquement, l'approche "modèle" utilise certaines connaissances a priori sur les lèvres. Un modèle de lèvres est appliqué sur l'image contenant les lèvres ; puis il est ajusté pour correspondre au mieux aux lèvres réelles. Finalement, à partir du modèle ajusté, les contours interne et externe des lèvres sont obtenus permettant d'extraire différentes informations.

L'avantage principal de cette approche est qu'elle permet de représenter le locuteur avec une dimensionnalité inférieure à celle de l'approche "image". Ceci a pour conséquence de réduire quantitativement la redondance du système. Cependant, le fait de se focaliser seulement sur les lèvres représente un inconvénient majeur. Les paramètres labiaux ainsi extraits par ces techniques ne captent pas toute l'information pertinente sur la parole. La visibilité des dents ou de la langue peut en effet ajouter une information supplémentaire qui ne peut être utilisée par le modèle des contours des lèvres seules. Rappelons que Summerfield *et al.* (1989) a montré que les dents et la langue améliore l'information visuelle de 7%.

Dans cette approche, les techniques peuvent se répartir en deux catégories suivant le type des paramètres labiaux à extraire : les paramètres géométriques et les paramètres fondés sur la forme du modèle des lèvres (Potamianos *et al.*, 2006). Dans les deux cas, il faut un algorithme permettant d'extraire les contours interne et externe des lèvres ou en général la forme du visage (pour la synthèse par exemple).

Capture des contours des lèvres : Après la localisation de la région de la bouche sur une image, un algorithme permettant d'estimer les contours des lèvres est alors appliqué. Nous retrouvons dans la littérature plusieurs méthodes qui rentrent dans ce cadre. Parmi ces méthodes trois d'entre elles émergent et semblent être les plus connues : la méthode dite "*snakes*" (Kass *et al.*, 1988), la méthode "*templates*" (Yuille *et al.*, 1992; Silsbee, 1994), et la méthode s'appuyant sur les modèles actifs d'apparence ou de forme (Cootes *et al.*, 1995, 1998; Daubias et Deleglise, 2002).

Le *snake* [3] est défini comme une courbe élastique (ou surface) paramétrée (représentée par un ensemble de points de contrôle mais qui peut aussi être implicite) qui se déforme en réponse à deux types de forces ; forces internes et forces externes correspondant respectivement à une énergie interne et à une énergie externe. En général, la courbe change constamment sa topologie en modifiant de façon itérative les coordonnées des points de contrôle de la courbe. La courbe finale représentant un contour des lèvres (externe ou interne) est obtenue lorsqu'un critère local (souvent défini par l'utilisateur) est optimisé. Le modèle *snake*, appelé aussi modèle de contour actif, a été utilisé par Chiou et Hwang (1997) dans leur système de lecture labiale automatique.

La seconde méthode pour la capture des contours des lèvres utilise des modèles déformables dits "*templates*". Le principe de cette méthode consiste en une description paramétrée des contours des lèvres. Des points caractéristiques sont repérés sur l'image. Puis, la variation de

[3] *mot en anglais qui signifie "serpent"*

la position de ces points est observée sur un lot d'images d'apprentissage. Le modèle résultant décrit les déformations habituelles des points caractéristiques. Sur une image de test, les points "règlent" la forme du modèle pour atteindre la forme désirée du contour et ceci en minimisant la valeur d'un certain nombre d'intégrales le long des contours pertinents. Cette valeur est définie comme une fonction d'énergie (ou de pénalité). Elle contient des termes qui font appel aux caractéristiques remarquables du *template* (l'intensité, les pics et les vallées, les bords). De ce fait, en comparant la position des points caractéristiques obtenus sur une image de test, il est possible de vérifier si la variation de la position de ces points est dans les limites du modèle déformable. Si c'est le cas, cela veut dire que l'image est similaire au lot d'apprentissage. Sinon, l'image est considérée comme différente. Hennecke *et al.* (1994) ont utilisé un simple modèle de la bouche et des lèvres sous forme de *templates* composés de 2 paraboles symétriques et 3 bi-quadratiques ("*quartics*") pour capturer les contours des lèvres. Leur modèle est contrôlé par 12 paramètres permettant de reconstruire le mouvement, l'angle d'orientation et le centre des coordonnées (voir figure 2.2). La région de la bouche est d'abord détectée par un filtrage gradient. Ensuite, le modèle de lèvres est superposé sur les contours réels d'une image test en minimisant une fonction de coût (ou d'énergie). Un tel modèle sur les contours des lèvres est très sensible à la variation d'éclairage et à l'initialisation des paramètres de contrôle. Pour éviter cette sensibilité, Chandramohan et Silsbee (1996) ont proposé un système de détection des paramètres labiaux fondé sur des modèles déformables multiples (*multiple déformable template*). La méthode proposée consiste en deux phases. Dans un premier temps, l'image test est classifiée selon plusieurs catégories d'images issues d'une phase d'apprentissage. Cette classification fournit une série de paramètres sur le choix du modèle ainsi qu'un ensemble de paramètres d'initialisation. Dans un second temps, le modèle désigné est appliqué aux contours et ajusté en optimisant la fonction de pénalité. La structure topologique est identique à chacun des *templates* (modèles). Chaque modèle est représenté par un ensemble de points reliés par des segments de droites comme le montre la figure 2.3 (chaque point est relié à son voisin par un segment de droite). Les modèles ont tous une géométrie identique, c'est-à-dire le même ensemble de segments. Par conséquent, pour les différencier, les auteurs déterminent des "ressorts" définissant l'interaction entre les différents points de chaque modèle. Ces "ressorts" peuvent être activés ou désactivés suivant l'image de test. Le nombre de "ressorts" actifs est alors différent d'un modèle à l'autre, ce qui permet donc de choisir avec un tel critère un seul modèle.

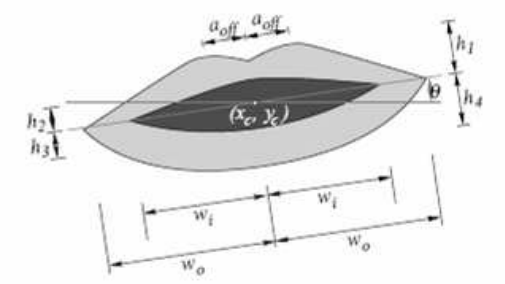

FIG. 2.2 – Modèle de lèvres avec ses 12 paramètres de contrôle Hennecke *et al.* (1994).

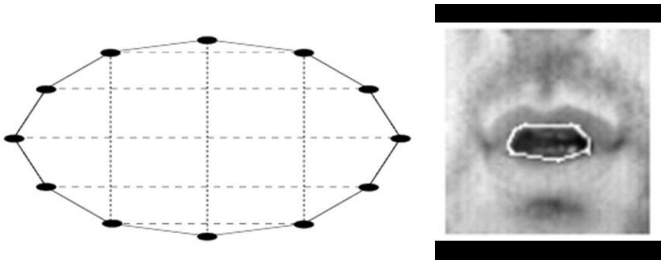

FIG. 2.3 – Modèle de lèvres (à gauche) appliqué sur une image (à droite). Les "ressorts" sont représentés par des traits en pointillés (Chandramohan et Silsbee, 1996).

Nous arrivons à la troisième méthode d'estimation des contours des lèvres. Ce type de méthode s'appuie sur un modèle actif de l'apparence ou de la forme. Ici, on construit des modèles statistiques de la forme ou de l'apparence de la région des lèvres. On retrouve deux algorithmes proposés par Cootes et al. (un pour le modèle de la forme (Cootes *et al.*, 1995) et l'autre pour le modèle de l'apparence (Cootes *et al.*, 1998)). Luettin *et al.* (1996) ont utilisé ce type d'algorithme. Précisément, ils s'appuient sur les modèles actifs de forme. Dans leur méthode, les auteurs représentent les frontières des lèvres à l'aide d'un ensemble de points étiquetés. Ces points se déforment suivant des modèles appris à partir d'un ensemble d'apprentissage en utilisant une analyse en composantes principales (ACP ou PCA[4]). De ce fait, les modes principaux de la variation de la forme issue de l'ensemble d'apprentissage peuvent être décrits par un nombre réduit de paramètres. Pour une image courante, les contours sont ré-estimés en minimisant la distance de niveau de gris entre cette image et le modèle. La figure 2.4 présente un exemple de résultats montrant des contours capturés par Luettin *et al.* (1996). Notons que les modèles actifs de forme semblent à première vue analogues aux modèles actifs de contour (*snake*) définis précédemment. La différence majeure entre les deux est que les premiers peuvent seulement déformer les données filtrées conformément à l'ensemble d'apprentissage ; alors que les seconds déforment un contour pour enfermer l'objet à extraire sur l'image. Daubias et Deleglise (2002) ont construit deux modèles statistiques de la forme et de l'apparence pour extraire les informations labiales des images d'un locuteur. Les auteurs définissent leurs modèles comme étant des modèles *à posteriori* qui sont appris à partir d'un corpus. Les deux modèles nécessitent pour leur apprentissage des méthodes de localisation des lèvres sur les images. Pour ceci, les auteurs présentent des méthodes automatiques s'appuyant sur le maquillage des lèvres en bleu. Dans un premier temps, Daubias et Deleglise (2002) définissent un modèle de la forme par deux polygones décrivant les contours interne et externe des lèvres. la phase d'apprentissage de ce modèle consiste à apprendre statistiquement, à partir des contours des lèvres extraits facilement grâce au maquillage en bleu, la forme moyenne et les déformations. Dans un second temps, pour apprendre le modèle de l'apparence, les lèvres sont localisées grâce à un étiquetage automatique s'appuyant sur l'utilisation du maquillage en bleu et de la bimodalité de la parole. En effet, les auteurs enregistrent deux répétitions d'une même phrase par le même locuteur, avec et sans maquillage en bleu sur les lèvres. Les images de la première répétition permettent d'extraire de

[4] *Principal Component Analysis*

façon automatique les contours des lèvres. Ensuite, un alignement des informations acoustiques associées aux deux répétitions permet d'estimer la forme des lèvres sur les images de lèvres sans maquillage ("naturelles"), à partir des formes obtenues sur les images avec maquillage. Enfin, le modèle d'apparence est entraîné en utilisant des réseaux de neurones.

FIG. 2.4 – Exemples d'images avec les contours capturés par la méthode de Luettin *et al.* (1996).

Finalement, quelque soit la méthode utilisée, les contours interne et externe ont été examinés avec un degré de réussite différent d'une étude à l'autre. C'est surtout le contour interne qui pose le plus de difficulté dans l'extraction. Par exemple, Luettin *et al.* (1996) ont trouvé que l'extraction du contour interne est beaucoup plus difficile que celle du contour externe en raison du contenu non uniforme de l'intérieur de la bouche. A l'intérieur de la bouche, il existe des zones qui peuvent avoir le même aspect que les lèvres (gencives et langue), des zones brillantes (dents), ainsi que des zones très sombres (cavité orale). Chaque zone pouvant apparaître et disparaître continuellement pendant la production de parole. Les méthodes utilisant les premiers modèles actifs de contour (exemple : Kass *et al.* (1988) et les *templates*) déformables (Yuille *et al.*, 1992) ont extrait seulement le contour externe. D'autres études simplifient le problème en maquillant les lèvres en bleu pour augmenter ainsi le contraste des lèvres et permettre ainsi l'extraction du contour interne avec une certaine facilité et une précision meilleure (Lallouache, 1991).

Plus récemment, des études menées au Département Image Signal (DIS) du laboratoire GIPSA (Grenoble Image Parole Signal Automatique) tentent d'extraire le contour interne et externe des lèvres. Pour extraire le contour externe, Eveno *et al.* (2003) ont développé un algorithme s'appuyant sur des contours actifs et des modèles paramétriques pour extraire le contour externe. Des modèles représentent la forme a priori de la bouche et un "*jumping snake*" ajuste leurs positions. Le "*jumping snake*" est un nouveau type de modèle de contour actif. La différence entre ce modèle et les *snakes* classique vient du fait que celui-ci peut être initialisé de loin depuis l'extrémité finale. De plus, l'ajustement de ses paramètres est facile et intuitif. Avec ce "*jumping snake*", les frontières supérieures de la bouche sont détectées ainsi que plusieurs points caractéristiques. Pour la segmentation du contour extérieur, les auteurs utilisent un algorithme s'appuyant sur un modèle paramétrique. La forme des lèvres est approchée par un ensemble de courbes qui sont décrites uniquement par quelques paramètres. Plus précisément, le modèle utilisé est composé de plusieurs courbes cubiques. Ce modèle est donc suffisamment flexible pour reproduire les spécifités des différentes formes de lèvres. Pour extraire le contour interne, les travaux sont encore inachevés mais les premiers résultats sont encourageants. Dans une première version, à partir du contour extérieur obtenu, des points clefs sont détectés, puis il faut faire converger 2 "*jumping snakes*" et définir 2 modèles paramétriques différents (selon que la bouche

est fermée ou ouverte) pour extraire le contour intérieur. Des résultats de segmentation sont proposés sur la figure 2.5.

FIG. 2.5 – Exemple de segmentation des contours des lèvres effectué au DIS-GIPSA (d'après Eveno *et al.* (2003).

Paramètres géométriques des lèvres : Dans cette catégorie, un ensemble de paramètres géométriques est extrait à partir des contours des lèvres obtenus par une des méthodes d'estimation que nous avons vu précédemment. Nous trouvons des paramètres tels que la hauteur (ou l'aperture) et la largeur (ou l'étirement) des contours des lèvres ainsi que les aires contenues à l'intérieur des contours. Plusieurs systèmes audio-visuels ont utilisé ce type de paramètres ou du moins une partie d'entre eux pour caractériser la modalité visuelle. Nous trouvons parmi ces études : Petajan (1984); Lallouache (1990); Adjoudani et Benoît (1996); Alissali *et al.* (1996); Jourlin (1997); Rogozan *et al.* (1997); Teissier *et al.* (1999); Heckmann *et al.* (2001) ...etc. Pour la plupart de ces études, les lèvres étaient colorées en bleu pour faciliter l'extraction des contours des lèvres. Les paramètres géométriques sont ensuite calculés en utilisant des équations établies par Lallouache (1991). Cette méthode sera détaillée dans la deuxième partie de ce manuscrit (partie expérimentale, chapitre 5).

Nous pouvons ajouter à cette catégorie certaines études utilisant d'autres paramètres visuels dérivés des contours des lèvres et qui peuvent améliorer la lecture labiale automatique. Potamianos *et al.* (2006) rapportent que les moments centraux ou normalisés d'une image binaire du contour interne (définis par Dougherty et Giardina (1987)) peuvent être considérés comme des paramètres visuels. Des coefficients normalisés des séries de Fourier calculés sur les paramètres d'un contour (Dougherty et Giardina, 1987) peuvent aussi être utilisés pour augmenter les paramètres géométriques dans certains systèmes de lecture labiale.

Paramètres issus du modèle des lèvres : Les méthodes telles que les *templates* déformables ou les *snakes* servent souvent dans la littérature pour capturer les contours des lèvres. Ces méthodes estiment paramétriquement des modèles déformables de contours. Les paramètres de ces modèles peuvent être considérés comme des paramètres visuels représentant l'information labiale. Par exemple, Chiou et Hwang (1997) utilisent un nombre de vecteurs radiaux de snakes comme paramètres visuels. Quant à Chandramohan et Silsbee (1996), ils utilisent plutôt les paramètres du *template* des lèvres. Hennecke *et al.* (1994) commandent leur *templates* par des paramètres caractérisant la localisation et l'orientation de la forme du modèle prototype ainsi que les modes de déformation. C'est précisément ces paramètres qui sont utilisés comme entrée

visuelle dans leur système de reconnaissance audio-visuelle.

D'autres méthodes utilisent des modèles actifs de forme. Ces derniers sont des modèles flexibles obtenus de façon statistique et représentent un objet par un ensemble de points étiquetés. Cet objet peut être soit le contour interne/externe des lèvres Luettin *et al.* (1996) ou soit une union de contours représentant la forme de différents visages[5] (Matthews *et al.*, 2001). Les paramètres à injecter ensuite dans un système de reconnaissance par exemple varient selon la technique statistique utilisée. Dans le cas de Luettin *et al.* (1996) par exemple, les paramètres du modèle de forme et du modèle de luminance servent d'entrée pour le système de reconnaissance automatique.

Sur ce dernier exemple, nous pouvons même remarquer que traiter la luminance de l'image de la bouche est en effet un aspect caractéristique de l'approche "image". En même temps, les auteurs proposent un modèle de déformation. Ceci est en fait une sorte de combinaison entre les deux approches.

2.1.3.3 Comparaison image-modèle

Les deux approches "modèle" et "image" ont toutes les deux des avantages et des inconvénients. En dépit des différences évidentes entre ces deux approches, une caractéristique qu'elles partagent toutes les deux est le besoin éventuel d'une intervention manuelle. En effet, on peut intervenir manuellement pour étiqueter des données ou définir une région d'intérêt (d'habitude c'est la région de lèvres). Cependant, l'utilisation de l'une ou l'autre dépend globalement de la difficulté de la méthode, de sa robustesse et de la pertinence de la paramétrisation visuelle résultante. Par ailleurs, il existe dans la littérature peu d'études comparant les deux approches. Nous présentons ci-dessous trois études les comparant :

- Brunelli et Poggio (1993) comparent les performances obtenues par deux techniques automatiques pour la reconnaissance du visage, à partir d'images prises en vue frontale. La première technique (qu'on peut qualifier d'approche "image") s'appuie sur le calcul d'un ensemble de paramètres géométriques à partir de l'image du visage. La seconde technique est fondée sur une adaptation d'un modèle du visage sur l'image réelle (*Template Matching*). La comparaison entre ces deux techniques nous semble intéressante même si l'objet à traiter dans l'étude était le visage et non pas seulement la bouche. Elle peut nous livrer certains aspects utiles pour fonder des arguments sur l'utilisation de ces techniques. Les auteurs ont obtenu, en terme de reconnaissance, des performances supérieures en utilisant la seconde technique ("*template matching*").

- Matthews *et al.* (1998) comparent deux techniques différentes pour caractériser les formes de la bouche pour la reconnaissance visuelle de la parole (lecture labiale automatique). La première technique extrait les paramètres requis pour adapter un modèle actif de forme (*Active Shape Model*, ASM) aux contours des lèvres. La seconde utilise des paramètres dérivés d'une analyse spatiale multi-échelle (*Multiscale Spatiale Analysis*, MSA) de la région de la bouche. Les résultats semblent avantager l'analyse spatiale multi-échelle. Ils montrent que cette technique est plus robuste, rapide et plus précise. En effet, dans les tests de reconnaissance avec des locuteurs multiples et utilisant seulement les données visuelles, la précision de reconnaissance des lettres

[5] *ou bouches si nous ne nous intéressons qu'aux lèvres*

est de 45% pour la méthode MSA et de 19% pour ASM. Pour reconnaitre des digits, la précision est la même pour les deux méthodes (77%). Cette performance relativement faible de l'ASM peut être expliquée par l'incorporation de connaissanes a priori dans la méthode qui peuvent être inexactes. Le fait de représenter le contour des lèvres par un modèle simple semble être aussi trop limite pour diffuser des informations plus précises. En général, l'ASM est confronté comme toutes les techniques de l'approche "modèle" à des erreurs de modélisation et de capture.

- Matthews *et al.* (2001) comparent, dans une tâche de reconnaissance audio-visuelle continue à large vocabulaire, quatre techniques différentes de paramétrisation visuelle. Trois de ces techniques appartiennent à l'approche "image". Il s'agit de la transformée en cosinus discrète (DCT), la transformée en ondelettes discrète (DWT) et l'analyse en composante principale (PCA). Ces trois méthodes nécessitent de localiser la région de la bouche. La quatrième technique, utilisant l'approche modèle active d'apparence (AAM), tente de modéliser le visage entier par un modèle déformable de l'apparence du visage et inclut un algorithme de capture. Il est évident a priori qu'utiliser le visage entier devrait être bénéfique. Le visage entier peut inclure des caractéristiques visuelles supplémentaires qui pourraient être utiles et bénéfiques à la reconnaissance. Toutefois, les résultats obtenus dans un test de reconnaissance visuelle de mots semble contredire cette évidence. Les résultats expérimentaux montrent que les performances des méthodes de l'approche "image" sont meilleures (en taux d'erreurs : autour de 59% pour les trois méthodes "image" *vs.* 64% pour l'AAM). La méthode AAM est probablement désavantagée par les problèmes que rencontrent toute méthode de l'approche "modèle", à savoir les erreurs d'apprentissage du modèle.

En résumé, ces quelques comparaisons donnent un petit avantage à l'approche "image". Ceci dit, comme nous l'avons évoqué précédemment, l'approche "modèle" dépend beaucoup des algorithmes employés pour l'apprentissage du modèle. Une amélioration de ces algorithmes et l'incorporation de connaissances a priori qui rendent mieux compte de la structure de déformation de l'objet considéré, augmentera probablement la robustesse de cette approche.

2.1.3.4 Combinaison image-modèle

Des combinaisons des deux approches ont été employées dans plusieurs systèmes de reconnaissance labiale automatique. Dans la plupart de ces systèmes, les paramètres issus de chaque catégorie sont juste concaténés. Comme nous l'avons vu ci-dessus, l'étude de Luettin *et al.* (1996) peut rentrer dans ce cas. Tout comme Dupont et Luettin (2000), qui ont combiné les paramètres issus d'une analyse en composantes principales avec ceux d'un modèle actif de forme (ASM). Chiou et Hwang (1997) utilisent deux types de paramètres visuels combinés dans un système de reconnaissance labiale à partir d'une vidéo couleur. D'un coté, des paramètres extraits de l'espace géométrique en utilisant des "*snakes*" et de l'autre coté, des composantes principales extraites par l'application d'une transformée de Karhunen-Loève (KLT[6]) dans l'espace propre des couleurs. Chan (2001) utilise une méthode combinant une technique "image" et une autre de type "modèle". La première consiste en des projections par analyse en composantes principales (PCA) d'un sous-ensemble de pixels contenus à l'intérieur de la bouche. La seconde consiste en

[6] *Karhunen-Loève Transform*

une estimation de paramètres géométriques à partir des images. Ces paramètres peuvent être par exemple la largeur et la hauteur des lèvres ainsi que leurs dérivées temporelles.

En terme de performance, toutes ces études soulignent l'avantage de combiner les deux approches pour extraire les paramètres visuels les plus pertinents. Par exemple, Chiou et Hwang (1997) avancent le pourcentage 94% de mots isolés correctement reconnus. Cependant, dans ces études, à l'exception de celle de Chan (2001), les scores de reconnaissance fournis ne sont pas comparés à ceux obtenus dans le cas où chacune des deux méthodes est employée seule. Chan (2001) obtient un score de reconnaissance correcte de digits de 98% en utilisant des paramètres issus des deux approches sont combinés et normalisés. Ce score est à comparer au 94% obtenu à partir de paramètres issus de l'approche géométrique (ou "modèle") et au 96% pour le cas de l'approche pixel (ou "image").

2.1.4 Résumé

Suivant l'utilisation finale des paramètres visuels, leur extraction peut se faire avec une approche "image" ou "modèle". Le peu d'études comparatives des deux ne permet pas clairement de favoriser l'une ou l'autre. Il faut dire que chacune a ses avantages ($\oplus$) et ses inconvénients ($\ominus$) que nous synthétisons ci-dessous :

Approche "image"

- $\oplus$ région des lèvres globalement traitée,
- $\oplus$ absence d'un traitement d'image requis pour la paramétrisation du signal visuel,
- $\oplus$ absence d'une sélection a priori des informations visuelles,
- $\ominus$ informations visuelles réparties sur un nombre important de paramètres,
- $\ominus$ sensibilité aux conditions d'éclairage et à la position de la tête.

Approche "modèle"

- $\oplus$ extraction facile des lèvres grâce au modèle,
- $\oplus$ représentation du signal visuel compacte,
- $\oplus$ possibilité de prendre en compte la variabilité de la position et de l'orientation des lèvres,
- $\ominus$ sensible aux erreurs de la modélisation et de la capture,
- $\ominus$ perte de l'information contenue dans certaines parties de la bouche comme les dents et la langue.

Enfin, il est possible de combiner les deux approches. Ceci augmentera certainement les performances d'extraction et rehaussera les informations visuelles. Cependant, il n'existe à notre connaissance aucune évaluation réelle et suffisament significative de l'apport d'une telle combinaison en rapport avec une complexité éventuelle du système. L'utilisateur reste donc maître du choix !

2.2 Reconnaissance des gestes de la main

Dans cette section, nous nous intéressons aux gestes manuels du code LPC (position et configuration de la main). Plus précisément, nous présentons un état de l'art des techniques permettant de détecter et de reconnaître ces gestes. Il n'existe malheureusement dans la littérature

que de rares études traitant ce sujet. C'est pourquoi nous élargissons notre champ de recherche aux nombreuses techniques de reconnaissance des gestes qu'englobe le domaine de la vision artificielle[7].

2.2.1 Reconnaissance des gestes : cas de la main

Dans le domaine de la reconnaissance des gestes, l'objectif est de reconnaître des événements issus de capteurs physiques en employant des techniques informatiques. Ces techniques procèdent généralement par deux phases : représentation et décision. Il s'agit premièrement de projeter les données brutes numériques en un espace de représentation. Les événements sont alors modélisés dans cet espace par un ensemble de paramètres de représentation. Ensuite, ces paramètres sont injectés dans un système de décision qui produit en sortie le geste reconnu. En général, dans les systèmes de reconnaissance des gestes, le système de décision compare le vecteur de paramètres à un lot de vecteurs de référence représentant chacun une classe d'un dictionnaire. Ce dernier, appelé aussi *vocabulaire*, est appris lors d'une étape d'apprentissage antérieure. La figure 2.6 illustre le schéma fonctionnel d'un système de reconnaissance des gestes.

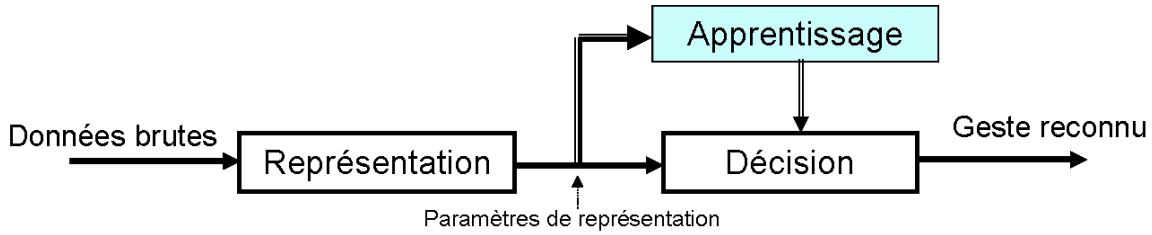

FIG. 2.6 – Scéma d'un système de reconnaissance des gestes.

Nous avons employé depuis le début le terme "geste". Mais à quoi correspond-il ? dans le domaine général de la reconnaissance des gestes, un *geste* désigne une séquence de vecteurs de données. Chaque vecteur de données, désigné par *posture*, peut contenir une configuration, une position ou/et une orientation de l'objet concerné. De cette définition, nous pouvons distinguer deux aspects caractérisant la reconnaissance des gestes : un aspect statique décrivant la posture ou la forme spécifique de l'objet à identifier, et un aspect dynamique caractérisant la succession des postures de l'objet et donc sa trajectoire.

Dans le cas de la reconnaissance des gestes de la main, la posture de la main est définie comme un mouvement statique. Par exemple, former un poing et le garder dans une certaine position est considéré comme une posture. Si les doigts de la main sont dans un état étendu ou plié, la posture est dite simple. Dans le cas contraire, donc si les doigts sont inclinés (formant un angle différent de 0° ou 90°), la posture est considérée comme complexe. Pour bien faire la différence, prenons l'exemple d'une posture où la main indique quelque chose. Cette posture est simple. En revanche, indiquer un "OK" par la main est une forme de posture complexe.

Le geste de la main est défini comme un mouvement dynamique, qui, à l'instar de la posture, peut être simple ou complexe. Un geste simple peut se faire de deux façons. Soit à partir d'une

[7] *aussi appelée vision par ordinateur, vision numérique ou vision cognitive.*

posture simple ou complexe tout en changeant l'emplacement et l'orientation de la main ; soit en bougeant les doigts d'une quelconque manière tout en laissant la position et l'orientation de la main inchangées. Un geste complexe implique des mouvements des doigts et un changement de la position et de l'orientation de la main. Les signes de la Langue des Signes sont des exemples d'un tel geste.

De ces différentes définitions, nous pouvons dire que les gestes manuels sont en général des événements spatio-temporels qui impliquent la main et son mouvement. Ce mouvement peut être décomposé en deux composantes : mouvement local et global. Le mouvement global capture le mouvement de la main entière. Le mouvement local peut être le mouvement des doigts ou le changement de la forme de la main. Il est ainsi clair que la reconnaissance des gestes de la main exige des techniques d'analyse spatiale et temporelle. La plupart des techniques de reconnaissance des gestes (ou des postures) de la main se basent sur des outils de reconnaissance de gestes (ou de postures) isolés. Par ce procédé, une séquence de gestes (ou de postures) est décomposée en un ensemble de gestes (ou postures) segmentés et isolés. De ce fait, il est nécessaire de segmenter la main et de la détecter parmi d'autres objets immobiles ou en mouvement. Cette tâche, même s'elle est parfois simple, ne doit pas être négligée.

Reconnaître des gestes de la main nécessite donc deux grandes étapes : extraire et capturer la main dans une image, et reconnaître le geste de la main. Les outils employés dans ces deux étapes doivent répondre à deux questions majeures. La première est : quelle technique permettrait de collecter un vecteur de données contenant les informations nécessaires pour la reconnaissance du geste de la main ? La seconde est : quelle technique de reconnaissance des gestes permettrait de maximiser la robustesse et la précision ?

2.2.2 Techniques pour la collecte des données

En général, les techniques de détection de la main peuvent être partagées en deux approches principales. Il y a tout d'abord les techniques qui utilisent des appareils portés par l'utilisateur. Cette catégorie consiste généralement en une utilisation d'un ou deux *gants instrumentés*[8] qui permettent de mesurer les divers angles de la main ainsi que certains degrés de liberté qui renseignent sur l'orientation et la position de la main. En d'autres termes, ces gants permettent de capturer et numériser les mouvements de la main. Ensuite, il y'a aussi les techniques basées sur l'approche *vision*[9]. Dans cette approche, une ou plusieurs caméras enregistrent un nombre d'images et les envoient vers un système de traitement des images qui permettent de réaliser la reconnaissance des gestes. Nous pouvons aussi ajouter une approche "*hybride*" qui combine les deux approches précédentes dans le but d'améliorer la précision de la reconnaissance.

2.2.2.1 Approche "gant instrumenté"

Dans cette approche, les données sont collectées en utilisant des gants instrumentés et des traqueurs. Les gants instrumentés sont équipés de capteurs mécaniques ou optiques qui trans-

[8] *Désigné aussi par : gants de données, gant numérique, gant électronique ou gant sensitif.*
[9] *En anglais : computer-vision-based approach.*

mettent, par des signaux électriques, les flexions et les abductions[10] des doigts de la main vers des instruments pour déterminer la posture de la main. Les traqueurs sont en général des capteurs supplémentaires (de type magnétique ou acoustique) attachés au dos de la main ou au dessus du poignet[11] et qui rendent des données sur la position et l'orientation de la main. De nombreux gants ont été développés pour des objectifs bien différents. Les gants développés diffèrent dans leur conception suivant plusieurs éléments : le type de gant, le nombre et le type de capteurs, ainsi que le type de traqueurs et leur position par rapport au gant.

Parmi les gants instrumentés les plus utilisés il y'a ceux développé par *VPL research*. Il s'agit des appareils *DataGlove* et *Z-Glove*. Ces deux gants ont plusieurs caractéristiques identiques mais ont aussi quelques différences. Les deux gants sont équipés par 5 à 50 capteurs à fibre optique détectant la flexion des doigts pour un total de 10 degrés de liberté. Dans certains cas, des capteurs supplémentaires d'abduction sont ajoutés pour mesurer les angles entre les doigts adjacents. La différence majeure entre les deux gants concerne l'emplacement et l'orientation des mécanismes utilisés avec chacun. Le *DataGlove* utilise un système magnétique traditionnel de détection, tandis que le *Z-Glove* emploie un système encastré ultrasonique qui place deux transmetteurs sur les cotés opposés des métacarpes. La figure 2.7 montre une image d'un gant de type *DataGlove*.

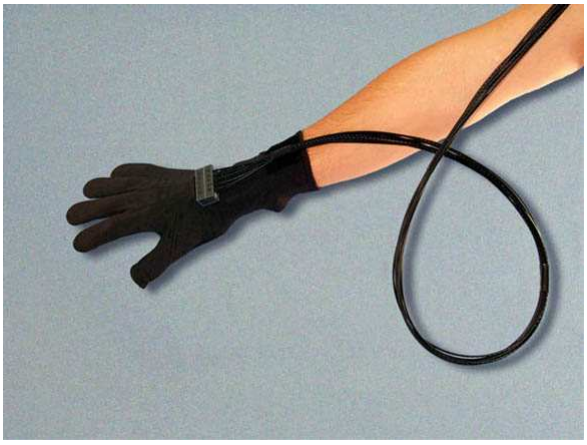

FIG. 2.7 – Image du gant "*5DT Data Glove 16 MRI*". Ce gant ne contient aucune partie métallique ou magnétique et il est connecté au boîtier d'interface par un ruban long de 5-7m via une fibre optique (image copiée du site internet de *Fifth Dimension Technologies (5DT)* : http ://www.5dt.com/products/.

Nous trouvons aussi d'autres gants instrumentés tels que *Digital Data Entry Glove* développé en 1981 dans les laboratoires *Bell Telephone*, *Power Glove* crée en 1989 par *Mattel*, *Pinch Glove* développé par Mapes à l'université Centrale de Floride (Mapes et J., 1995), *CyberGlove* développé par Kramer (Kramer et Liefer, 1989) ...etc. Pour une revue détaillée sur les appareils qui existent nous renvoyons vers Sturman et Zeltzer (1994) et LaViola (1999).

Tous ces gants permettent d'interpréter les mouvements de la main en analysant les signaux

[10] *Pour la main, l'abduction consiste à éloigner les doigts de l'axe de la main, qui passe par le medius.*
[11] *la position des traqueurs dépend du type de gant à utiliser.*

envoyés par leurs capteurs. Plusieurs systèmes conçus dans le domaine de la réalité virtuelle ont intégré ces appareils bénéficiant ainsi des avantages de leur utilisation. Ces avantages incluent :

- Facilité d'utilisation,
- Des mesures directes des paramètres de la main et des doigts tels que les angles entre les doigts, la rotation du poignet et les informations spatiales sur la position et l'orientation de la main,
- Des données suffisamment échantillonnées et qui sont indépendantes à toute translation de la main.

Cependant, les inconvénients de ces appareils ne sont pas de moindre importance. Ils sont les suivants :

- les systèmes précis sont onéreux, et les versions moins chères sont beaucoup bruitées,
- L'utilisateur est forcé de porter un appareil encombrant, réduisant le confort et la liberté du mouvement,
- La calibration est souvent délicate.

2.2.2.2 Approche "vision"

Les problèmes posés par les gants instrumentés ont mené à des recherches sur l'utilisation de la vision par ordinateur pour capturer les mouvements de la main et extraire les données pour une reconnaissance des gestes. Les systèmes fondés sur cette approche sont en général utilisés dans les interfaces d'interaction gestuelle homme-machine. Ils exploitent des algorithmes de traitement des images pour reconnaître visuellement les gestes de la main. Cette approche apparaît plus naturelle, mais en l'utilisant il est plus difficile d'implémenter des systèmes efficaces. Cinq éléments importants sont à considérer dans le développement et la réussite de tels systèmes comme une solution de collecte des données pour la reconnaissance des gestes et des postures de la main (Qutaishat *et al.*, 2007; Ong et Ranganath, 2005; LaViola, 1999; Starner, 1995) :

- Le placement et le nombre de caméras utilisées,
- La visibilité de la main sur la caméra,
- L'extraction des caractéristiques à partir du flux de données brutes extrait de l'image,
- La capacité des algorithmes de reconnaissance à extraire les paramètres requis,
- L'efficacité des algorithmes de reconnaissance appliqués pour fournir le maximum de précision et de robustesse.

Tout d'abord, la visibilité de la main est importante à cause des nombreux problèmes d'occlusion[12] rencontrés lors de la capture de la main. Le nombre de caméras utilisées est donc important pour balayer un champ de vision maximal. Leur position est aussi importante pour une vision complète de la main. Toutefois, une seule caméra suffit en général pour collecter les données pour la reconnaissance. Starner et Pentland (1996) ont montré qu'une seule caméra peut suffire en reconnaissance des gestes et des postures de la main. L'utilisation de plus d'une caméra est nécessaire pour capturer le mouvement de la main quand une information 3D ou

[12] *le terme est utilisé dans le domaine de la vision par ordinateur pour décrire la manière dans laquelle un objet tout près du champ de vision cache un autre objet plus loin. A ceci nous pouvons ajouter le fait que l'objet sort complètement du champ de vision de la caméra.*

profonde est requise. En effet, il est évident que lorsque le nombre de caméras augmente, la visibilité est améliorée. Cependant, dans ce cas la complexité algorithmique est accrue.

D'un autre coté, l'analyse du mouvement de la main exige de la repérer dans un environnement complexe, de la segmenter et d'extraire les paramètres caractéristiques à la reconnaissance de gestes et de postures. Cette tâche n'est pas triviale. La main doit en effet être beaucoup plus visible sur la caméra pour une extraction simple de ses données. Deux méthodes se distinguent pour cette tâche. La première approche utilise des gants spécialement conçus avec des marqueurs. Cette méthode est nommée "reconnaissance visuelle des gestes avec des gant-marques" ou "VBGwGM[13]" (Qutaishat *et al.*, 2007). Nous nous contentons dans la suite de l'expression "avec artifices" pour désigner cette méthode. L'utilisation de ces artifices (gants, gants marqués ou seulement des marqueurs) aide à la détermination des postures et des gestes de la main et réduit ainsi la complexité des traitements. La seconde méthode tente d'éviter l'utilisation de ces artifices et d'aboutir à un système plus naturel. Cette méthode est désignée par "reconnaissance visuelle pure des gestes" ou "PVBG[14]" (Qutaishat *et al.*, 2007). Nous la nommons dans notre cas pour simplifier par l'expression "sans artifices".

Avec artifices : La forme géométrique de la main est hautement non convexe (courbé). Par conséquent, il est très difficile de détecter la configuration de la main à partir des images produites par une caméra. Pour contourner ce problème, des artifices ont été utilisés par certaines techniques de reconnaissance des gestes de la main. Ces artifices sont soit actifs ou passifs. Les premiers sont généralement des LEDs[15]. D'habitude, les artifices passifs sont soit des marqueurs placés aux bouts (extrémités) des doigts soit des gants colorés qui peuvent porter eux aussi dans certains cas des marqueurs. Plusieurs travaux précédents ont utilisé ce type de technique.

Davis et Shah (1993) ont développé un système qui emploie un gant noir marqué par des rubans blancs sur le bout des doigts[16]. Le système calcule des trajectoires en repérant dans un premier temps les extrémités des doigts dans plusieurs trames avec un arrière-plan uniforme et en utilisant dans un second temps une correspondance de mouvement[17]. Ces trajectoires servent ensuite pour déterminer le début et la fin de la position du geste. Ainsi, chaque geste est modélisé par des vecteurs début-fin. Comme résultat obtenu par ce système, avec une vitesse de 4 fps[18], une segmentation de 7 gestes prédéfinis de la main peut être effectuée.

Starner (1995) ont décrit un système d'interprétation de la langue américaine des signes. L'utilisateur porte un gant coloré dans chacune des mains pour faciliter la détection temps réel de la main. Cette détection s'effectue par un algorithme appelé croissance de région (en anglais : Region Growing). En effet, dans ce système l'utilisateur se place assis devant une caméra en portant un gant jaune sur sa main droite et un autre orange sur sa main gauche.

[13] *visual-based gesture with glove-markers.*

[14] *pure visual-based gesture*

[15] *diode électroluminescente (abrégée en DEL), également appelée LED de l'anglais pour light-emitting diode.*

[16] *ce type de gant est en général nommé gant marqué binaire (Binary marked glove)*

[17] *Une correspondance de mouvement mappe les points dans une image sur les points de l'image suivante de façon à ce que deux points ne seront pas mappés sur le même point*

[18] *feet per second (pied par seconde) - une unité anglaise pour mesurer des vitesses. 1 fps = 1 ft/s = 0.3048 m/s.*

Pour repérer chacune des mains, l'algorithme parcoure l'image jusqu'à ce qu'il repère un pixel de la couleur appropriée. Ensuite, tous les pixels avoisinants à ce pixel et qui ont la même couleur sont supposés appartenir à la même région dans l'image. Cette région est amenée à croître par incorporation de pixels (ayant la même couleur) jusqu'à ce que toute l'image soit parcourue. Finalement, un vecteur de 8 paramètres contenant des informations sur les trajectoires des mains et leurs orientation, est extrait pour chaque image.

Sans artifices Malgré le fait que les gants colorés se différencient des gants instrumentés par leur simplicité qui ne réduit pas la liberté de mouvement des utilisateurs, la situation idéale est de reconnaître des gestes sans utiliser d'artifice. Sans artifices, hélas, la détection de la main présente d'importantes difficultés. Par exemple, le fait que la couleur de la peau de la main soit semblable à la couleur des autres parties du corps, rend sa distinction compliquée. Pour faciliter cette distinction, certaines restrictions sont imposées (porter des vêtements longues manches, une seule main visible ...etc). Ong et Ranganath (2005) dressent une liste des restrictions et des contraintes relatives à l'imagerie trouvées dans différentes techniques de l'approche vision pour reconnaître les gestes de la main sans utiliser aucun artifice. Cette liste est rapportée dans la table 2.2.

des vêtements à manches longues,
un arrière-plan uniforme,
un visage immobile ou qui a moins de mouvement que les mains,
un mouvement constant des mains est exigé,
un emplacement et une pose fixes du corps ou un emplacement initial spécifique de la main,
une des deux mains et/ou le visage exclu du champ de vue,
une prise de vue restreinte à la main qui conserve une orientation et une distance fixes par rapport à la caméra,
un ensemble restreint de signes (ou gestes) à reconnaître,
positionner la main au dessus du visage.

TAB. 2.2 – Restrictions et contraintes dans l'imagerie utilisées dans l'approche vision (Ong et Ranganath, 2005).

En imposant une ou plusieurs de ces restrictions, plusieurs recherches ont été menées afin de détecter la main dans une image. Si dans un grand nombre de ces recherches la détection de la couleur de la peau est utilisée, de nombreux problèmes d'ambiguité sont rencontrés notamment ceux dûs aux événements d'occlusion d'une main par l'autre ou par le visage. Dans le cas d'une capture d'image tri-dimensionnelle, certains de ces problèmes sont tout simplement évités (par exemple l'occlusion entre les deux mains). Mais dans le cas contraire, des solutions relativement laborieuses sont en général proposées.

Dans ce registre, Starner (1995) adoptent une représentation en tâches pour détecter les mains sans aucun artifice. L'idée sur laquelle se fondent les auteurs est que toutes les mains ont approximativement la même teinte et la même saturation et c'est principalement la luminance qui les différencie. Utilisant cette information, un modèle a priori de la couleur de la peau peut être construit. Avec ce modèle, chaque main est détectée comme une tâche qui s'élargit en fusionnant les pixels ayant la teinte de la peau. Le traitement se déroule normalement à l'exception de certains cas où la main est obstruée par l'autre main ou par le visage. Dans ce cas, la détection de la couleur ne peut résoudre ce problème d'ambiguïté. Pour le visage, les auteurs supposent qu'il reste dans la même zone de l'image (il bouge moins) et donc peut être déterminé et écarté. En revanche, les deux mains bougent toutes les deux et lorsque elles s'obstruent, une seule tâche apparaît pour les deux. Les auteurs remarquent que cette tâche est plus large qu'une tâche normale pour une seule main. De plus, les moments de cette tâche sont significativement différents de ceux de chacune des deux mains dans la trame précédente. Utilisant ces informations, les auteurs suggèrent d'assigner aux deux mains la même information sur les moments et la position de la grande tâche dans le cas d'occlusion. Par ailleurs, les auteurs assignent la tâche la plus à gauche (respectivement droite) à la main gauche (respectivement droite). Finalement, la méthode présentée par les auteurs retient comme information à injecter dans un système de classification la combinaison des informations de la position et des moments.

Imagawa *et al.* (1998) et Yang *et al.* (2002) utilisent eux aussi un principe relativement similaire s'appuyant sur une détection de la couleur de la peau. Les seules différences concernent les solutions proposées pour distinguer la main du visage et pour pallier le problème de l'occlusion. En effet, Imagawa *et al.* (1998) considèrent que la tête est relativement statique et donc le visage peut être facilement écarté. Yang *et al.* (2002) considèrent que la région du visage est la plus grande. Quant à l'apparition de l'occlusion, dans Imagawa *et al.* (1998) des filtres de Kalman sont utilisés pour chaque main, tandis que Yang *et al.* (2002) évitent ce problème en jouant sur l'angle de la caméra.

Les techniques de *tracking* des gestes de la main sont nombreuses dans la littérature. Nous nous contentons de ce que nous avons décrit ci-dessus. Si nous voulons résumer ces méthodes, la plupart se fondent sur :

- des détecteurs de mouvement,
- des détecteurs de la couleur de la peau,
- des détecteurs de contours,
- une combinaison de plusieurs de ces détecteurs.

2.2.2.3 Comparaison vision *vs.* gants instrumentés

Nous avons vu les différentes techniques pour la collecte de données de la main en les séparant en deux grandes approches : techniques orientées gant instrumenté et techniques orientées vision. Nous avons trouvé utile d'établir une comparaison entre les deux approches en discutant leurs avantages et leurs inconvénients. La table 2.3 présente cette comparaison.

2.2.3 Les techniques de classification pour la main

Après le suivi de la main et l'extraction de données concernant la forme et le mouvement de la main, l'étape suivante dans un système de reconnaissance consiste à classifier ces données pour reconnaître les gestes et/ou les postures de la main. Le domaine de l'intelligence artificielle regorge de nombreuses techniques de classification de ce type de données. LaViola (1999) classifie ces techniques en trois classes (classification reprise aussi par Qutaishat *et al.* (2007)) :

- Les techniques orientées paramètres, statistiques et modèles (*template matching*, extraction et analyse des caractéristiques, modèles actifs de forme, modèles géométriques de la main, analyse causale ...) ;
- Les techniques orientées algorithmes d'apprentissage (réseaux de neurones, modèles de Markov cachés (modèles HMM[19]), apprentissage par l'exemple) ;
- Les autres techniques telles que : l'approche linguistique, l'analyse du mouvement s'appuyant sur l'apparence et l'analyse de vecteurs spatio-temporels ;

Cependant, cette manière de classifier ne semble pas vraiment partager les techniques. Par exemple, la technique fondée sur les modèles actifs de forme (voir section I de ce chapitre), considérée dans la première classe, peut aussi être placée dans la seconde classe, puisqu'elle nécessite aussi une phase d'apprentissage. De même, la technique utilisant les modèles HMM s'appuie sur des outils statistiques (voir définition en chapitre 4) et donc peut aussi être considérée dans la première classe.

Selon le principe de chacune de ces techniques, certaines d'entre elles s'appuient sur la construction de modèles, d'autres s'appliquent directement sur les données. Deux classes sont donc distinguées : les techniques orientées "modèle" et les techniques orientées "données". A ces deux classes, une troisième classe peut être ajoutée. En effet, les techniques de la troisième classe proposée par LaViola (1999) ne peuvent être classées comme des techniques orientée "modèle" ni "données". Dans la suite, nous décrivons quelques techniques pour chacune de deux premières classes. Pour une description des techniques de la troisième classe nous renvoyons le lecteur vers la revue de LaViola (1999).

2.2.3.1 Techniques orientée "modèle"

Parmi les techniques évoquées ci-dessus, celles dont le principe s'appuie sur des modèles sont :

- modèles HMM,
- modèles actifs de forme,

[19] *Abréviation de son expression en anglais : Hidden Markov Models*

	Vision	Gants instrumentés
Confort d'utilisation	⊕ : Parfois avec des gants colorés mais qui ne gênent pas le mouvement de la main et ne sont pas encombrants	⊖ : Les gants utilisés sont encombrants et connectés à des ordinateurs donc libérté du mouvement restreint
Taille de la main (les mains varient en taille et en forme)	⊕ : Pas de problème	⊖ : Les gants ne peuvent être utilisés que par des personnes ayant des mains avec des tailles correspondantes
Puissance de calcul	⊖ : Puissance de calcul importante due aux algorithmes de traitement d'image	⊕ : Les données envoyées à l'ordinateur sont facilement enregistrables et transformables
Coût	⊕ : Le coût est moins élevé surtout que les différents postes de travail sont équipés par des caméras	⊖ : Le coût pour une reconnaissance robuste et complexe des gestes et des postures est élevé (plusieurs milliers d'Euros)
Portabilité	⊙ :L'indépendance de l'utilisateur du poste de travail est un peu difficile	⊙ : Portabilité possible dans le cas où le *tracking* n'est pas exigé
Précision	⊙ : Dépend en grand partie de la complexité des gestes à reconnaître et des algorithmes employés	⊙ : Idem
Robustesse au bruit	⊕ : Le bruit issu des capteurs est minimal	⊖ : Nécessité de filtrage pour réduire les bruits même si certains de ces bruits sont considérés comme des données à exploiter
Calibration	⊕ Importante notamment dans le cas d'un système multi-utilisateur	⊖ Importante et critique à cause de la différence de l'anatomie des mains entre les personnes

TAB. 2.3 – Synthèse des avantages et des inconvénients des deux approches vision et gants instrumentés pour la collecte des données de la main. Nous mettons des signes pour marquer l'avantage d'une approche sur l'autre : ⊕ approche avantagée, ⊖ approche désavantagée et ⊙ aucune des deux approches ne semble avantagée.

– modèles géométriques de la main *Linear fingertips models*
– réseaux de neurones,

Les modèles HMM, que nous définirons dans le chapitre 4 puisque ce sont des outils théoriques que nous allons utiliser dans nos travaux, s'appuient sur des modèles statistiques. Grobel et Assan (1996) ont utilisé ces modèles pour reconnaître des signes isolés réalisés par la main. Les auteurs extraient des paramètres caractéristiques à partir de l'enregistrement vidéo des codeurs qui portent des gants colorés. Pour un vocabulaire de 262 signes, ils obtiennent un taux de reconnaissance de 91,3%.

Les modèles actifs de forme fonctionnent ici tout comme nous l'avons décrit précédemment pour le cas des lèvres (voir section I de ce chapitre). La différence concerne l'objet à reconnaître (la main). Cette technique, appelée par certains "*smart snakes*", place un contour dans l'image qui est approximativement la forme du trait à extraire. Le contour évolue ensuite en se déplaçant de manière itérative vers les bords (frontières[20]) avoisinants qui déforment le contour pour convenir au trait. Tout comme Heap et Samaria (1995), Liu et Lovell (2005) ont récemment développé un système pour reconnaître les postures et les gestes de la main, s'appuyant sur cette technique. Une première étape consiste à construire un ensemble d'apprentissage à partir duquel les propriétés statistiques des classes de la main sont apprises. Pour chaque classe, les auteurs ont posé des pastilles sur les contours de la main pour extraire un ensemble de points définissant ainsi la forme de la main. L'étiquetage des pastilles a été fait à la fois de façon manuelle et automatique. Dans une seconde étape, une analyse en composantes principales est appliquée sur les points de chaque classe pour extraire les directions indépendantes de leur variation. En dernière étape vient l'ajustement du modèle pour l'adapter à la forme de la main dans l'image. L'avantage majeur de cette méthode est qu'elle permet une reconnaissance des gestes et des postures en temps réel. En revanche, elle n'est capable actuellement d'extraire que des gestes et des postures assez limités. De plus, elle ne peut extraire et reconnaître que des mains ouvertes.

La technique *Linear fingertips models* s'appuie sur une hypothèse simplificatrice : la majorité des mouvements des doigts est linéaire et ne comprend que de très peu de mouvements de rotation. Ceci simplifie donc le modèle de la main qui permet de n'utiliser que les bouts des doigts comme entrée. Ainsi, un modèle représentant les trajectoires de chaque extrémité de doigt, peut être représenté juste par un simple vecteur. Davis et Shah (1993) ont utilisé cette approche dans un système de reconnaissance des gestes. Dans ce système, la détection et l'extraction des extrémités des doigts repose sur des marques colorées posées sur ces extrémités et sur une segmentation d'histogramme. Ensuite, les trajectoires des bouts de doigts sont calculées en utilisant une correspondance des mouvements. A partir d'un petit ensemble d'apprentissage, les postures sont modélisées en enregistrant le code du mouvement, le nom du geste et les vecteurs de direction et de magnitude pour chaque bout de doigt. Finalement, si tous les vecteurs de direction et de magnitude coïncident avec un geste de la base enregistrée, la posture est reconnue. Cette technique semble simple et donne de bons scores de reconnaissance. Cependant, le nombre de gestes reconnus dans cette seule étude (Davis et Shah, 1993) est limité (7 au total), nous

[20] *Ces frontières sont localisées dans les zones où l'intensité change.*

ne pouvons par conséquent mesurer vraiment la robustesse de cette technique. Il faut en effet élargir le nombre de gestes et de postures pour mesurer la robustesse de cette méthode. Par ailleurs, ce type de système semble coûteux en temps de calcul et ne fonctionne pas en temps réel.

Les réseaux de neurones sont réellement utilisés dans plusieurs domaines. Principalement, ils étaient utilisés en intelligence artificielle pour construire certains types d'agents autonomes et pour reconnaître des motifs. Un réseau de neurones est un système de traitement de l'information qui vient à l'origine d'une tentative de modélisation du cerveau humain. Tout d'abord, on se donne une unité simple, appelé neurone ou nœud, capable d'effectuer quelques calculs élémentaires. On relie ensuite entre elles un nombre important de ces unités. La connections entre un neurone et un autre (appelé synapse) peut posséder un poids associé pouvant ainsi se comporter comme un mécanisme de mémoire. Chaque neurone peut être représenté par une fonction possédant plusieurs entrées et une sortie. Cette fonction possède deux composantes (cf. figure 2.8). La première est la fonction d'entrée qui consiste à calculer la somme pondérée des valeurs des entrées. La seconde est appelée la fonction d'activation[21] qui transforme la somme obtenue en une valeur finale de sortie. Cette fonction introduit une non-linéarité dans le fonctionnement du neurone. Il existe plusieurs modèles de neurones suivant la fonction d'activation utilisée. Parmi les fonctions d'activation classique, il y'a par exemple : la fonction sigmoïde[22], la fonction tangente hyperbolique et la fonction de Heaviside[23]. Il est important de noter qu'une

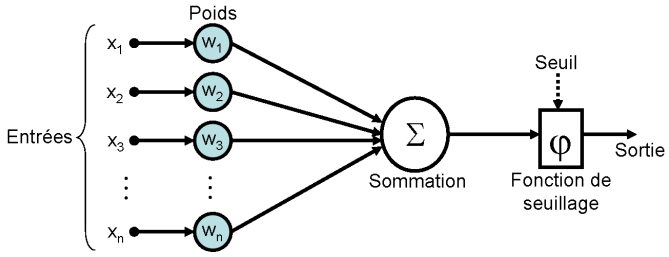

FIG. 2.8 – Structure d'un neurone. Le neurone calcule la somme pondérée de ses entrées puis cette valeur passe à travers la fonction de seuillage pour produire sa sortie.

cellule élémentaire dans un réseau de neurones peut manipuler des valeurs binaires ou réelles. Les valeurs binaires sont représentées par 0 et 1 ou -1 et 1. Concernant le calcul de sortie, plusieurs fonctions peuvent être utilisées et leur calcul peut être déterministe ou probabiliste.

En général, les réseaux de neurones ont deux structures élémentaires. La première est une structure dite en feed-forward (sans rétroaction) dans laquelle les connections entre les unités ne forment pas un cycle dirigé. Elle est la première et apparemment la plus facile à concevoir. Dans cette structure, l'information se propage en une direction unique, vers l'avant, à partir des nœuds d'entrées, passant par des éventuels nœuds cachés jusqu'aux nœuds de sorties. Il n'y a aucun cycle et aucune boucle dans un réseau ayant cette structure. La seconde structure est

[21] *appelée aussi fonction de transfert ou de seuillage*

[22] *Fonction définie par :* $f(x) = \frac{1}{1+\exp(-\lambda x)}$

[23] *Est une fonction discontinue prenant la valeur 0 en les réels strictement négatifs et la valeur 1 partout ailleurs.*

dite avec rétroaction. Dans cette structure, les connections forment un cycle dirigé. Un réseau avec rétroaction a un avantage sur un réseau feed-forward par le fait qu'il peut modéliser des systèmes avec des états de transition. Cependant, il nécessite plus de descriptions mathématiques complexes et peut devenir chaotique.

Dans les deux topologies, il est à noter qu'il n'y a aucune restriction sur le nombre de couches à mettre dans un réseau. Augmenter le nombre de couches dans un réseau améliorera certes la puissance de calcul et de représentation mais ceci est au prix d'une complexité de l'apprentissage. Ce dernier est très important dans les réseaux de neurones. Globalement, l'apprentissage peut être effectué par deux mécanismes : apprentissage supervisé ou non-supervisé. D'un coté, un apprentissage est dit supervisé lorsque l'on force le réseau à converger vers un état final précis, en même temps qu'on lui présente un motif. Le réseau va se modifier jusqu'à ce qu'il trouve la bonne sortie, c'est-à-dire celle attendue, correspondant à une entrée donnée. A l'inverse, lors d'un apprentissage non-supervisé, le réseau est laissé libre de converger vers n'importe quel état final lorsqu'on lui présente un motif. Les deux stratégies d'apprentissage supervisé et non-supervisé ne sont pas mutuellement exclusives. Il est aussi possible de les combiner en un apprentissage hybride . Pour une discussion exhaustive sur les algorithmes d'apprentissage employés dans les deux stratégies nous renvoyons vers Mehrotra *et al.* (1997) et vers Neocleous et Schizas (2002) pour une revue comparative.

Les réseaux de neurones sont des méthodes très utilisées pour reconnaître des gestes et des postures de la main. Elles peuvent être utilisées avec l'approche vision ou avec l'approche gant instrumenté. L'un des premiers systèmes à les utiliser a été développé par Murakami et Taguchi (1991). D'un coté, pour reconnaître des postures de la main, le système utilise un réseaux de neurones à trois couches contenant 13 nœuds d'entrée, 100 nœuds cachés et 42 nœuds de sortie. Le réseau utilise une topologie avec rétroaction et un mécanisme d'apprentissage qui minimise l'erreur entre une sortie cible et la sortie produite par le réseau. Avec un ensemble initial d'apprentissage de 42 postures, le réseau atteint 77% de précision. Quand l'ensemble d'apprentissage passe de 42 à 206, ce score s'élève à 98%. De l'autre coté, les gestes de la main sont reconnus aussi avec un réseau de trois couches, mais qui s'appuie sur une structure recurrente (avec rétroaction). Le nombre des nœuds diffère aussi : 16 d'entrée, 150 cachés et 10 de sortie. Un réseau est dédié pour les 10 possibles gestes à reconnaître. Le taux de reconnaissance est initiallement de 80%, mais augmente pour atteindre 96% dans le cas où les données brutes sont filtrées.

Récemment, en s'appuyant sur une approche "vision", Qutaishat *et al.* (2007) ont développé un système pour traduire automatiquement les gestes statiques des alphabets et signes en langage américain des signes. Après avoir extrait des vecteurs caractéristiques des gestes de la main à partir d'images, les auteurs utilisent un réseau de neurones pour classifier ces vecteurs. Le réseau repose sur une structure *feed-forward* avec une rétroaction et est composé de trois couches de neurones. La première contient (214×3) neurones, la deuxième (214×2) et la couche de sortie 214. Le système proposé par Qutaishat *et al.* (2007) atteint un taux de reconnaissance de 98,5% pour les données d'apprentissage et 80% pour les données de test.

Les réseaux de neurones permettent de reconnaître un large nombre de postures et de gestes. En utilisant un apprentissage adéquat, de hauts scores de reconnaissance peuvent être obtenus

par ces méthodes. En revanche, l'apprentissage peut être assez lourd. En plus, si une posture ou un geste est introduit ou enlevé du corpus, le réseau doit être entièrement ré-appris.

2.2.3.2 Techniques orientées "données"

Cette catégorie de techniques s'applique directement sur les données caractéristiques de la forme et du mouvement de la main. Parmi ces techniques, on trouve :

- *Template matching*,
- Extraction et analyse des caractéristiques (*Feature extraction and analysis*),
- Analyse causale (*Causal analysis*)
- Apprentissage par l'exemple (*Instance-Based Learning*, IBL).

Les deux premières techniques reposent, toutes les deux, sur une corrélation entre la forme à tester et un ensemble de formes "modèles". La seule différence entre ces deux techniques réside dans la nature des données utilisées. En effet, la technique *template matching* exploite directement les données brutes extraites en s'appuyant sur l'approche "gant instrumenté" ou sur l'approche "vision" ; alors que la technique *Feature extraction and analysis*, analyse ce type de données (information de bas-niveau) pour produire une information sémantique de haut-niveau. Cette dernière information est utilisée ensuite pour reconnaître les postures et les gestes de la main. Le principe des deux techniques est en revanche identique et consiste en deux étapes. La première consiste à construire, en général avec une intervention manuelle, une base de référence de formes "modèles" caractérisant les postures ou les gestes de la main à reconnaître. La seconde étape consiste à comparer la forme à tester avec l'ensemble des formes "modèles". Cette comparaison peut se faire de plusieurs façons par exemple par minimisation d'un critère de distance.

La technique *template matching* a été appliquée notamment sur des données obtenues avec un gant instrumenté (Sturman, 1992; Watson, 1993) et seulement pour reconnaître des postures de la main. Elle peut être appliquée aussi sur des données extraites par une technique de l'approche "vision", mais elle semble relativement peu robuste, notamment quand il s'agit de trouver un *template* approprié pour toute posture de main. De plus, elle peut rencontrer quelques problèmes de robustesse liés aux variations de lumière et d'échelle. Par ailleurs, Rubine (1991) est le premier à utiliser un système s'appuyant sur une analyse des données brutes (le système est appelé en anglais : *2D single-storke gesture recognizer*) pour reconnaître des gestes en 2D. Il calcule, à partir de ces données, un ensemble de 13 caractéristiques telles que le cosinus et le sinus de l'angle initial du geste, la distance entre le premier et le dernier point et le maximum de la vitesse du geste. Sturman (1992) et ensuite Wexelblat (1995), ont étendu ce système en 3D. La technique fondée sur l'extraction et l'analyse des données est robuste pour reconnaître les postures et les gestes de la main qu'ils soient simples ou complexes. Son plus grand défaut est qu'elle pourrait être très coûteuse en calcul dans le cas où la taille des données extraites devient importante.

La technique d'analyse causale (*Causal analysis*) ou par règles est souvent utilisée en analyse de scènes. Son principe consiste en une extraction de représentations (ou informations) d'une scène à partir d'un flux vidéo continu en utilisant des connaissances sur les actions dans cette

scène ainsi que sur la façon dont elle est reliée aux autres scènes et à l'environnement physique. Les représentations employées sont causales dans la mesure où elles décrivent la physique sous-jacente de la scène. Dans le cas de la compréhension des gestes, trois types de connaissances peuvent être utilisées : (i) connaître le mécanisme du système qui produit le mouvement (donc le corps) ; (ii) connaître la façon selon laquelle ce comportement se transforme en des symboles significatifs (par exemple quels gestes sont privilégiés) ; (iii) et connaître pourquoi certains gestes prennent des sens dans une séquence. Brand et Irfan (1995) ont appliqué l'analyse causale dans un système de reconnaissance des gestes en se fondant sur l'approche vision. Ils ont utilisé des connaissances sur les aspects dynamiques du corps pour retrouver à partir du flux vidéo les caractéristiques qui peuvent être utilisées pour identifier les gestes. Leur système extrait d'abord des informations sur les positions jointes de l'épaule, du coude et du poignet dans le plan de l'image. Ensuite, à partir de ces positions, un lot de caractéristiques, incluant l'accélération et la déccélération du poignet, la taille du geste, l'aire entre les bras, l'angle entre les avant-bras, est extrait. En normalisant et en combinant ces caractéristiques ainsi qu'en utilisant des connaissances causales sur la manière avec laquelle les humains interagissent avec les objets dans l'environnement physique, des filtres gestuels sont conçus pour reconnaître des gestes tels que l'ouverture, la poussée, l'arrêt ...etc. Ce système proposé par Brand et Irfan (1995) semble être intéressant dans la mesure où il utilise des informations sur l'interaction des humains avec le monde physique ambiant pour identifier les gestes. Cependant, cette technique présente certaines limitations. En effet, elle n'est utilisée que pour un nombre limité de gestes et n'utilise ni les données sur la position et l'orientation de la main ni les données des doigts.

Cette technique est définie comme une généralisation d'une nouvelle cible à classifier à partir d'exemples d'apprentissage emmagasinés. Les exemples d'apprentissage sont traités quand une nouvelle entité cible arrive.

Les techniques s'appuyant sur l'apprentissage par l'exemple (IBL)[24] consistent à classifier une cible à partir d'un ensemble d'exemples appris. Ces exemples sont traités chaque fois une nouvelle entité cible arrive. Une table (*instance*) est en général un vecteur de caractéristiques de l'entité à classifier. Chaque fois qu'une interrogation sur une nouvelle cible est rencontrée, sa relation avec les exemples emmagasinés précédemment est examinée pour assigner une valeur de fonction cible pour cette *instance*. En reconnaissance des gestes et des postures, le vecteur des caractéristiques d'une *instance* peut être la position et l'orientation de la main ainsi que les valeurs de pliage des doigts.

Les algorithmes IBL consistent simplement d'abord à emmagasiner des exemples d'apprentissage (données). Ensuite, quand une nouvelle *instance* est rencontrée, un ensemble d'*instances* relativement similaires est récupéré de la mémoire et utilisé après pour classifier l'instance en question. Les techniques IBL peuvent construire différentes approximations de la fonction cible pour chaque *instance* distincte en question. Certaines techniques construisent seulement des approximations locales de la fonction cible qui s'applique dans le voisinage de la nouvelle *instance* candidate. Ces techniques ne construisent pas d'approximation destinée s'appliquer sur l'espace entier des *instances*. Ceci a un avantage significatif quand la fonction cible est très complexe.

[24] *Les techniques fondées sur l'apprentissage par l'exemple sont aussi appelées parfois apprentissage local (Lazy learning) du fait qu'elles retardent le traitement jusqu'à ce qu'arrive une nouvelle cible à classifier.*

Les algorithmes suivants peuvent être mis dans la catégorie des méthodes IBL :

* La méthode des k plus proches voisins (*k-Nearest Neighbor*),
* La régression localement pondérée (*Locally weighted regression*),
* Les fonctions de base radiales (*Radial basis functions*),
* Raisonnement à base de cas[25] (*Case-based reasoning*).

Ces algorithmes ont l'avantage d'être relativement simples à implémenter, à l'exception du raisonnement à base de cas. Avec ces algorithmes, un ensemble large de postures et de gestes de la main peut être reconnu avec une précision modérément élevée. Cependant, les inconvénients ne manquent pas. D'abord, la quantité de mémoire de travail augmente sensiblement avec la taille de l'ensemble d'apprentissage. Ensuite, tous les calculs doivent être fait à chaque classification d'une nouvelle *instance*. Ceci implique des problèmes concernant le temps de réponse qui augmente avec le traitement d'un ensemble large d'exemples d'apprentissage.

Très peu d'études ont été menées en utilisant l'approche IBL pour reconnaître des gestes et des postures. Notons seulement que Aha *et al.* (1991) ont décrit un système et une méthodologie s'appuyant sur technique IBL, qui génère des prédictions de classification en utilisant seulement des *instances* spécifiques. Ils ont ainsi développé trois algorithmes d'apprentissage par table tout en essayant d'atténuer certains problèmes que connaît l'approche apprentissage par l'exemple. Utilisant ces algorithmes, Kadous (1996) a reconnu, avec une précision de 80%, 95 postures discrètes de la main provenant du language des signes et décrites par des caractéristiques extraites par un gant instrumenté *Power Glove*.

2.2.4 Résumé

Ce que nous pouvons retenir de toutes ces expériences est que quelque soit la technique de détection de la main et de ses gestes, les ambiguïtés dues aux occlusions posent de sérieux problèmes. Les solutions données par certaines études à ces problèmes restent laborieuses et dépendent beaucoup des conditions de chaque expérience. Il est certes facile d'éviter ces problèmes en imposant certaines restrictions concernant les couleurs et les objets qui doivent apparaître, mais ces dernières enlèvent le caractère naturel au système. D'un autre coté, les méthodes de classification, que ce soit avec modèle ou non, ont des performances très variables. Souvent, un compromis est nécessaire entre la complexité de la méthode et sa robustesse. Certaines n'étaient à ce jour jamais utilisées pour reconnaître des gestes et des postures de la main. Leurs performances dans ce cas restent encore à évaluer.

2.2.5 Systèmes pour la reconnaissance des gestes manuels du code LPC

A partir des définitions que nous avons donné aux gestes et postures en général, il est maintenant possible de faire le lien avec les gestes du code manuel LPC. Ainsi, dans ce cas, le mouvement local peut être la formation de la configuration de la main, tandis que le mouvement global est matérialisé par le déplacement de la main d'une position à l'autre. Il est important de noter que les gestes manuels du code LPC sont des gestes particuliers qui diffèrent de ceux d'autres systèmes manuels tels que la langue des signes. D'une part, la main dans le code LPC

[25] *Ou système de raisonnement par cas.*

peut être en contact direct avec le visage qui a la même apparence que la main. D'autre part, le code LPC nécessite la détection précise de la position de la main par rapport au visage.

De très rares travaux ont concerné la reconnaissance des gestes LPC de la main. Gibert *et al.* (2005) ont utilisé 12 paramètres pour modéliser statistiquement les différentes configurations de la main dans un objectif visant à réaliser un avatar codant en LPC. Pour arriver à ces paramètres, les auteurs ont enregistré les positions 3D de 50 marqueurs, sensibles aux infra-rouges, placés sur la main d'un sujet en utilisant un système de capture de mouvement appelé Vicon©avec 12 caméras. Les auteurs notent que construire un modèle statistique des déformations de la main est très complexe. En effet, ils considèrent que les 50 marqueurs subissent le mouvement rigide de l'avant-bras qui est considéré comme le porteur de la main. De ce fait, les mouvements du poignet, de la paume et des phalanges des doigts ont une influence non linéaire tout à fait complexe sur les positions 3D des marqueurs. Ces positions ne reflètent pas suffisamment les rotations implicites des articulations. Pour résoudre ces contraintes anatomiques, les auteurs utilisent un modèle non linéaire calculant tous les angles possibles entre le segment de la main et l'avant-bras ainsi qu'entre les phalanges successives. Avec une analyse statistique, les auteurs retiennent seulement 12 paramètres pour contrôler le modèle de la main. Les auteurs ont ensuite testé la reconnaissance de la forme et le placement de la main. Pour choisir les formes cibles à classifier, Gibert *et al.* (2005) se sont fondés sur le fait que l'extension/rétraction maximale des doigts était approximativement synchrone avec le début acoustique de la consonne (résultat obtenu par Attina *et al.* (2002)). Ainsi, les images cibles contenant des formes de la main ont été sélectionnées aux environs du début acoustique de la consonne. Elles étaient ensuite labelisées avec une valeur appropriée (les auteurs ont choisi de les numéroter de 0 à 8 en affectant les numéros 1 à 8 aux 8 configurations LPC de la main et 0 pour le reste). Ces trames ont été prudemment choisies en traçant les valeurs de sept paramètres au cours du temps. Ces paramètres concernent différentes distances absolues : pour chaque doigt entre le point de fléchissement de la première phalange la plus près de la paume, et celle près du bout du doigt, entre les bouts de l'index et du majeur, et entre le bout du pouce et la paume. Au final, 4114 formes de la main étaient identifiées et labelisées. Les sept paramètres caractéristiques associés aux formes cibles de la main sont collectés et de simples modèles Gaussiens sont estimés pour chaque forme de la main. Ainsi, la probabilité *a posteriori* pour chaque trame de s'accorder avec chacune des 8 modèles des formes de la main est estimée. En testant les 4114 trames, les auteurs obtiennent un taux de reconnaissance des formes de la main de 98,78%. Sur ces mêmes images, les auteurs ajoutent un autre label concernant la position de la main. En effet, ils attribuent en plus du label de la forme, une valeur appropriée pour désigner la position de la main (un nombre de 0 à 5, 0 est attribué à toute position différente des cinq positions LPC). Les placements de la main pour ces configurations cibles sont caractérisés par la position 3D du doigt le plus long dans un repère réferentiel lié à la tête. Ce doigt est le majeur dans le cas où il est considéré dans une configuration sinon c'est l'index. Une fois que les coordonnées 3D de ce doigt sont collectées, des modèles Gaussiens simples sont estimés pour chaque placement de la main. En testant les 4114 placements de la main, le système identifie correctement 96,76% de ces placements. Il est important de noter que ces hauts scores ont été obtenus sur des images utilisées par la suite pour construire des modèles qui ont servi pour la synthèse de la main.

Dans un registre plus proche de notre travail qui porte sur la reconnaissance des gestes LPC de la main, les travaux effectués par Burger à FT R&D[26] et au DIS-GIPSA[27] ont pour objectif de développer un outil pour la reconnaissance des gestes manuels du code LPC. En utilisant un gant coloré, la main est détectée pour ensuite reconnaître la configuration et la position de la main. L'architecture globale du système de reconnaissance des gestes du code LPC ainsi développé est constituée des modules suivants : - segmentation colorimétrique de la main suite à un apprentissage de la couleur du gant porté par le codeur (figure 2.9) ; - détection des images cibles par utilisation d'un filtre rétinien qui permet de détecter les images pour lesquelles on enregistre un fort ralentissement du mouvement global de la main et du mouvement des doigts (Burger *et al.*, 2006b) ; - localisation du visage et des yeux et de la bouche par un algorithme proposé par Garcia et Delakis (2004) et qui se base sur une architecture neurale ; - détection de la zone pointée par la main par analyse de la position du doigt pointeur par rapport aux traits permanents du visage que sont les yeux et la bouche (figure 2.9) ; - classification de la configuration par utilisation de la théorie de l'évidence, des machines à vecteurs de support (SVM)[28] et de la généralisation de la transformation pignistique. Ces choix ont permi, selon les auteurs, de s'affranchir de certaines limitations liées aux méthodes probabilistes classiques (Burger *et al.*, 2006a; Aran *et al.*, 2007).

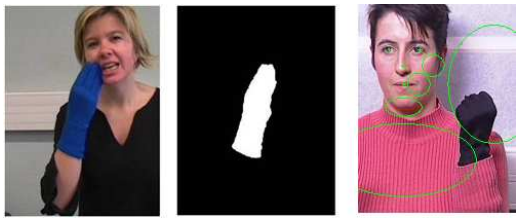

FIG. 2.9 – A gauche, segmentation de la main ; A droite, différentes zones de pointage.

Enfin, les deux techniques présentées ci-dessus se classe dans l'approche "'vision"' avec artifices d'après notre classification précédente.

2.3 Conclusion

Le traitement automatique de la lecture labiale nécessite l'extraction de caractéristiques visuelles contenant l'information phonétique. Deux approches sont généralement considérées : approche "modèle" et approche "image". Dans la première approche, les contours interne et externe des lèvres sont extraits à partir des images du locuteur. Un modèle de contour peut être

[26] *France Télécom Recherche et Développement.*

[27] *nous rappelons que c'est le Département Image Signal du laboratoire Grenoble Images Parole Signal Automatique.*

[28] *Une machine à vecteurs de support ou séparateur à vaste marge ou encore machine à support vectoriel (en anglais Support Vector Machine ou SVM) est une technique de discrimination. Elle consiste à séparer deux (ou plus) ensembles de points par un hyperplan. Selon les cas et la configuration des points, la performance de la machine à vecteurs de support peut être supérieure à celle d'un réseau de neurones ou d'un modèle de mixture gaussienne.*

obtenue d'une manière statistique ou paramétrique. Ainsi, l'ensemble des paramètres du modèle contient l'information visuelle de parole. Dans d'autres études, des paramètres géométriques des contours de lèvres sont utilisés. Dans la seconde approche, des transformations appropriées, telles que la transformation en cosinus discret ou l'analyse en composantes principales, sont appliquées aux pixels de l'image correspondant à la région de la bouche du locuteur.

Ces deux approches ont été peu comparées et ces comparaisons ne semblent pas avantager une par rapport à l'autre. En revanche, certaines études les ont combinées dans le but d'augmenter les performances et de rehausser les informations visuelles extraites. Cependant, il y a peu d'évaluation significative permettant de confirmer cette amélioration.

Pour reconnaître les gestes (ou les postures) de la main, deux étapes sont nécessaires. La première étape consiste à collecter des données contenant des informations suffisamment pertinentes des gestes à reconnaître. Deux approches sont utilisées. Dans la première, des gants instrumentés portés par la main et équipés de capteurs sont utilisés pour enregistrer et transmettre des paramètres contenant les informations spatiale et et temporelle sur les doigts et la main. Dans la seconde, les informations sont extraites à partir des images d'un enregistrement vidéo. Cette extraction peut se faire de deux façons selon qu'on utilise des artifices ou non. La seconde étape consiste à classifier ces données extraites pour reconnaître enfin les gestes de la main. Les méthodes de classification utilisées dans la littérature peuvent s'appuyer sur des modèles ou s'appliquer directement sur les données. Les performances de ces méthodes sont très variables.

Dans le cas du code LPC, très peu d'études ont tenté de reconnaître les gestes de la main. Le fait que, en code LPC, la main est souvent en contact direct avec le visage a exigé, dans toutes ces études, d'employer des artifices pour faciliter l'extraction des données.

Enfin, il est à noter que nous n'avons trouvé aucun travail, à notre connaissance, consacré à l'étude d'un système fusionnant les deux informations du code LPC (manuelle et labiale) dans le cadre de la reconnaissance automatique. Nous verrons dans le chapitre suivant quels sont les modèles de fusion qui peuvent être adaptés pour la reconnaissance des gestes du code LPC.

Chapitre 3

Intégration des flux

Le code LPC s'appuie sur deux composantes de la même modalité visuelle pour permettre la perception complète de la parole. Dans une tâche de reconnaissance automatique des gestes du code LPC, deux flux caractéristiques, chacun pour une composante (manuelle ou labiale), sont disponibles pour la reconnaissance. Un système combinant ces deux flux manuel et labial du code LPC doit permettre d'obtenir des résultats supérieurs à ceux obtenus avec une seule composante, et suffisamment performants pour permettre la transcription vers la parole. Il n'existe aucun travail visant à étudier un tel système. En terme d'intégration, il s'agit d'un problème de fusion de deux flux (l'un manuel, l'autre labial) qui pose des problèmes similaires à l'intégration audio-visuelle. Malgré les différences qui peuvent exister entre les flux manuel et labial du code LPC d'un coté et les traditionnels flux audio et visuel, les systèmes de reconnaissance automatique de la parole audio-visuelle (en anglais : Audio-Visual Automatic Speech Recognition (A-V ASR)) nous proposent quelques solutions pour fusionner des données issues des deux composantes du code LPC. L'aspect "parole" comme résultat des deux systèmes de reconnaissance (audio-visuel et main-lèvres) est une de nos motivations. D'autre part, il est à remarquer l'existence d'un flux commun : le flux labial qui caractérise la lecture labiale. D'ailleurs, les méthodes d'extraction des caractéristiques de ce flux peuvent être utilisées dans les deux cas (voir chapitre précédent). De ce fait, nous commençons avant tout par décrire les modèles d'intégration caractérisant la combinaison des informations audio et visuelle dans la perception humaine de la parole qui est bimodale par nature. Nous nous consacrerons ensuite aux éléments permettant de choisir un modèle tout en décrivant quelques études comparant ces modèles.

3.1 Modèles d'intégration audio-visuelle de la parole

Nous avons vu précédemment comment la parole peut être considérée comme bimodale. Les résultats présentés dans le chapitre 1 témoignent de la nécessité d'un processus d'intégration audiovisuelle de la parole. Les deux voies auditive et visuelle se concentrent de façon à ce que la prise de décision sous forme phonétique ou lexicale reflète les informations issues des deux modalités. De nombreuses études ont été menées pour rendre compte de la manière avec laquelle interagissent les deux modalités audition et vision pour la compréhension de la parole. Ces études menées tant par des psychologues, linguistes que par des ingénieurs, s'étendent sur plusieurs

domaines allant de la cognition, aux sciences de l'ingénieur en passant par la neurophysiologie. Ainsi, plusieurs modèles ont été proposés. Mentionnons par exemple, le célèbre modèle *Fuzzy-Logical Model of Perception* (FLMP) proposé par (Massaro, 1987, 1998). Les premières travaux se concentraient spécialement sur les architectures de fusion en considérant arbitrairement des représentations internes mono-modales (représentation visuelle seule et auditive seule). Sur ces représentations, les différents travaux consistaient à appliquer un certain nombre de calculs afin de prédire la performance bimodale. Dans ces études, le traitement de la représentation des informations des modalités est souvent négligé. Schwartz *et al.* (1998); Schwartz (2002), en croisant des modèles issus de la psycho-physique et de la fusion des capteurs, ont classé les modèles d'intégration audiovisuelle en quatre grandes architectures : (i) modèle à "Identification Directe" noté ID ; (ii) modèle à "Identification Séparée" noté IS ; (iii) modèle à "Recodage dans la modalité Dominante" noté RD ; et (iv) modèle à "Recodage commun des deux modalités sensorielles vers la modalité Motrice" noté RM.

Pour simplifier la compréhension du système d'intégration audio-visuelle dans la perception de la parole, nous pouvons le considérer comme une boîte qui a en entrée deux flux de nature différente (vision et audio) et en sortie une décision ou un code qui peuvent être de nature phonétique ou lexicale. Le schéma de la figure 3.1 illustre un tel système.

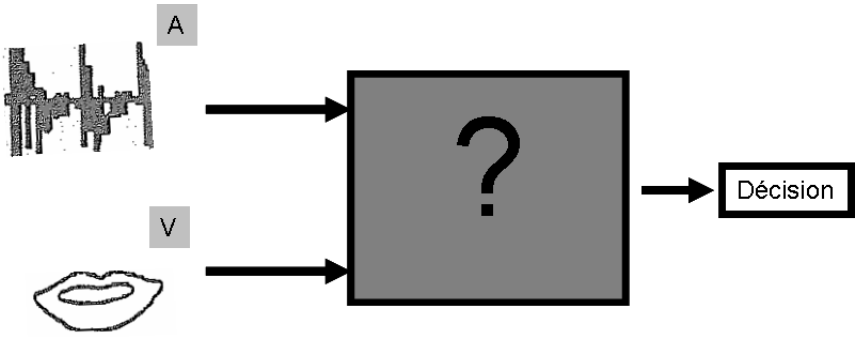

FIG. 3.1 – Le noyau d'un processus d'intégration audio-visuelle dans la perception de la parole (d'après Schwartz *et al.* (1998)).

Dans la suite, nous survolerons rapidement les 4 architectures classiques de l'intégration audio-visuelle. En plus de les définir, nous donnerons des exemples réalisés pour chacune de ces architectures.

3.1.1 Modèle ID

Dans ce modèle, appelé aussi modèle données-vers-décision, les deux sources d'information sont injectées directement dans un classifieur bimodal qui effectue le traitement de l'information des deux modalités (figure 3.2). La classification se fait donc directement sans aucun niveau intermédiaire de mise en forme commune des données. Le classifieur prend une décision dans l'espace des caractéristiques bimodales, dans lequel des prototypes bimodaux ou des règles de décision bimodales ont été appris. Ce modèle est une extension du modèle "*Lexical Access From Spectra*" (LAFS) de Klatt (1979) vers "*Lexical Access From Spectra and Face Parameters*".

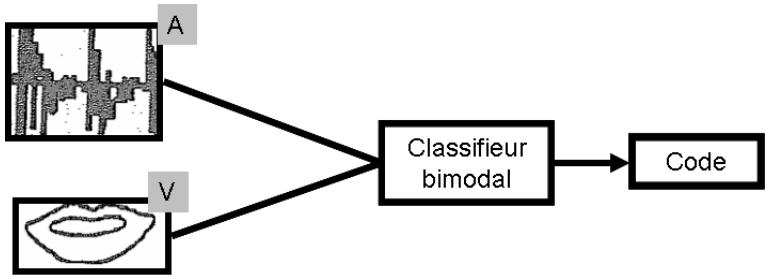

FIG. 3.2 – Modèle à identification directe.

Il existe diverses implémentations de l'architecture ID en reconnaissance de la parole (voir par exemple : Braida *et al.* (1986); Duchnowski *et al.* (1994); Adjoudani et Benoît (1996); Dalton *et al.* (1996); Krone *et al.* (1997); Nakamura *et al.* (1997); Teissier *et al.* (1999); Potamianos *et al.* (2001)) et en modélisation psychophysique (voir : Campell (1988); Braida (1991)).

Adjoudani et Benoît (1996) ont implémenté le modèle d'identification directe pour la reconnaissance audio-visuelle et ont évalué les performances pour une grande plage de rapport signal sur bruit. Ils injectent un vecteur d'observation audiovisuel dans un processus de reconnaissance s'appuyant sur les chaînes de Markov Cachées (HMM). Le vecteur audiovisuel est obtenu en concaténant des paramètres acoustiques issus d'une analyse acoustique à six paramètres géométriques des lèvres et leur dérivée. Dans une structure semblable, l'implémentation de Teissier *et al.* (1999) du modèle ID implique un classifieur Gaussien dans un espace de six dimensions. Le vecteur d'entrée bimodal de ce classifieur est composé de six paramètres : trois paramètres acoustiques issus d'une analyse ACP[1] et trois paramètres géométriques du contour interne des lèvres. Dans cette implémentation, un paramètre supplémentaire est ajouté dans le processus de fusion. Les deux flux d'entrée audio et vidéo sont pondérés. Ceci permet ainsi de contrôler les poids respectifs de chaque entrée conformément à leur efficacité pour la décision. Potamianos *et al.* (2001) ont proposé une technique de fusion des flux visuel et auditif en appliquant deux transformées l'une après l'autre. Ils utilisent tout d'abord une Analyse Discriminante Linéaire (ADL, en anglais LDA pour *Linear Discriminant Analysis*) pour réduire de façon discriminante les dimensions du vecteur concaténé des caractéristiques audio-visuelles. Puis, une Transformée Linéaire de Maximum de Vraisemblance (TLMV, en anglais MLLT pour *Maximum Likelihood Linear Transform*) est appliquée pour améliorer la modélisation des données. Ces deux transformées sont aussi utilisées pour prendre en compte l'information dynamique dans les flux des données audio-visuelles avant la fusion. Les auteurs réalisent ainsi un schéma hiérarchique d'intégration audio-visuelle.

3.1.2 Modèle IS

Le modèle d'identification séparée (IS) est fondé sur ce que les psychologues cognitifs appellent "*intégration tardive*" du fait que l'intégration vient après la classification phonétique dans chaque voie sensorielle séparée par opposition au modèle ID qui est une intégration "*précoce*"

[1] *Analyse en Composantes Principales*

car s'appliquant directement aux données. Dans le modèle IS, les informations visuelles et auditives sont traitées séparément chacune par un classifieur. Puis, la fusion des résultats des deux classifieurs dans un module d'intégration permet la reconnaissance du code (voir figure 3.3). Le modèle IS est aussi appelé décision-vers-décision en référence à la caractéristique de base de la fusion qui est une fusion de décisions. Dans ce type de modèle, la fusion peut être réalisée soit sur des valeurs logiques, à l'instar du modèle VPAM (*Vision-Place*, *Audition-Manner*) dans lequel chaque modalité est en charge d'un groupe spécifique de caractéristiques phonétiques (distinctives), soit par un processus probabiliste, comme dans le cas du modèle FLMP de Massaro (Massaro, 1987, 1998).

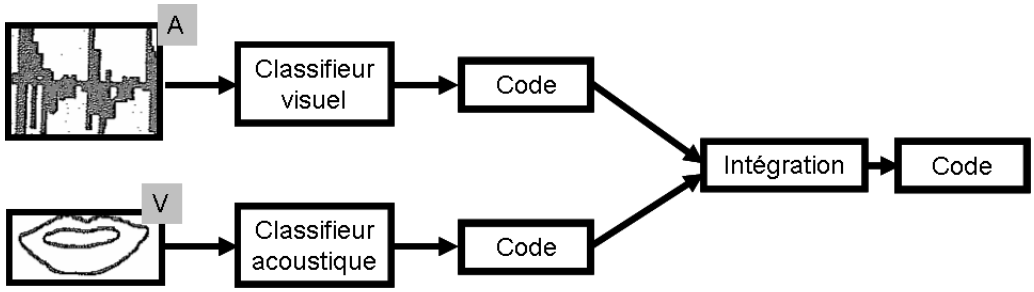

FIG. 3.3 – Modèle à identification séparée.

L'implémentation de ce modèle dans des systèmes de reconnaissance de la parole a été faite dans de nombreuses études (Petajan, 1984; Bregler *et al.*, 1993; Duchnowski *et al.*, 1994; Adjoudani et Benoît, 1996; Krone *et al.*, 1997; Nakamura *et al.*, 1997; Cox *et al.*, 1997; Huang *et al.*, 1997; Jourlin, 1997; Luettin et Dupont, 1998; Potamianos *et al.*, 1998; Teissier *et al.*, 1999; Neti *et al.*, 2000).

Adjoudani et Benoît (1996) ont aussi implémenté le modèle IS dans leur système de reconnaissance audio-visuelle. Ils ont utilisé deux réseaux HMM acoustique et visuel séparés. Dans cette implémentation, chaque modèle HMM est entraîné avec des données visuelles ou acoustiques. Les deux classifieurs fonctionnent ainsi indépendamment l'un de l'autre. En test, les vecteurs d'observations visuels ou acoustiques sont présentés séparément à l'entrée de chaque modalité. Les auteurs présentent ensuite trois méthodes pour le module d'intégration. La première, utilisée également dans d'autres études de reconnaissance de la parole audio-visuelle (Movellan et Chadderdon, 1996; Stork *et al.*, 1992), consiste à calculer le maximum des produits des probabilités conjointes des deux modalités. En d'autres termes, l'intégration s'appuie sur une sélection, pour chaque entité à reconnaître (phonème, syllabe, mot ...), d'un candidat qui maximise la vraisemblance dans les deux canaux. Le schéma synoptique de la figure 3.4 résume le processus d'intégration suivant ce principe. La seconde méthode repose sur une sélection du meilleur candidat d'une des deux modalités acoustique ou visuelle selon son degré de certitude (ou confiance). Ce dernier est évalué à partir des probabilités de sortie de chaque modèle HMM et sert à commander un "interrupteur" qui sélectionne la voie ayant une plus grande certitude dans sa sélection. Le principe de cette méthode ne permet pas de fusionner les données provenant des deux canaux. De ce fait, cette méthode ne peut être considérée comme une architecture

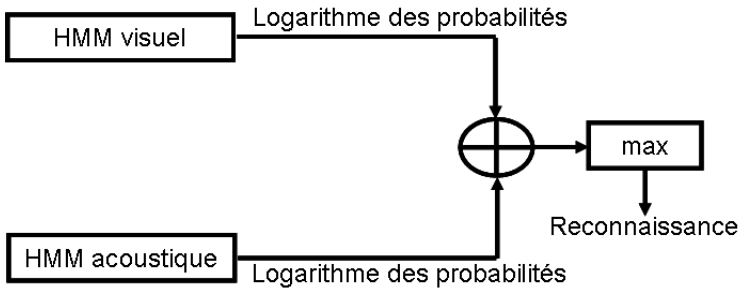

FIG. 3.4 – Modèle d'intégration basé sur la maximisation des produits des probabilités conjointes (D'après Adjoudani (1998)).

d'intégration. La figure 3.5 illustre le principe de cette dernière.

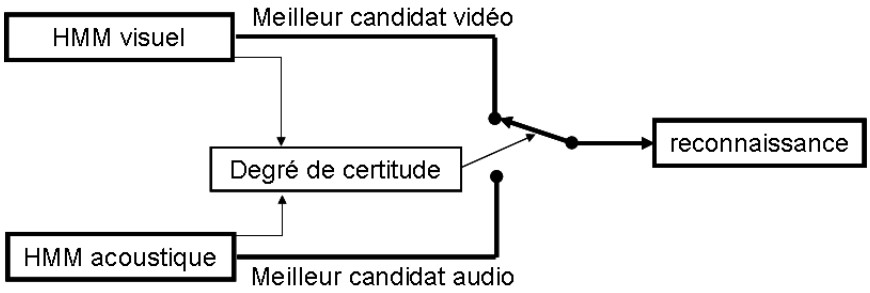

FIG. 3.5 – Méthode de sélection du meilleur candidat acoustique ou visuel (D'après Adjoudani (1998)).

La troisième méthode consiste à intégrer les informations auditives et visuelles suivant une pondération de chaque modalité en fonction de l'indice de confiance (voir figure 3.6). Le principe de cette méthode est identique au principe de la première sauf qu'ici les probabilités sont pondérées. D'abord, un indice est estimé de la même façon que dans la seconde méthode, c'est-à-dire à partir des probabilités de sortie de chaque voie. Le résultat de cette estimation définit ensuite le coefficient normalisé de pondération. Puis, en maximisant le produit des probabilités pondérées, un candidat est sélectionné.

(Teissier *et al.*, 1999) ont proposé une implémentation du modèle IS pour reconnaître des voyelles et utilisent deux classifieurs Gaussiens, chacun pour une modalité, et un processus de fusion attribuant à chaque canal un poids. Chaque classifieur délivre une probabilité d'appartenir à chacune des 10 catégories de voyelles à reconnaître. Ainsi, deux ensembles de valeurs de probabilités sont obtenus : (1) $p_A = p(i/x_A)$: Probabilité à posteriori que l'entrée acoustique corresponde à la catégorie i ; et (2) $p_V = p(i/x_V)$: Probabilité a posteriori que l'entrée visuelle corresponde à la catégorie i. Ces probabilités sont ensuite pondérées par des poids reflétant la contribution de chaque modalité dans le processus de fusion. Enfin, la probabilité audiovisuelle a posteriori est calculée en utilisant la formule suivante :

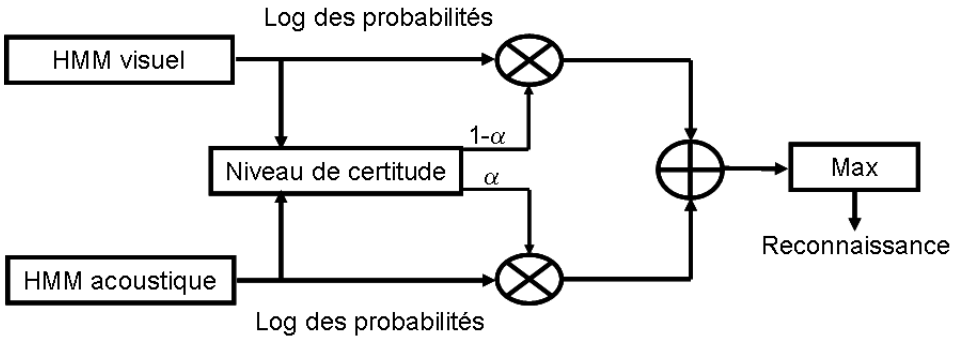

FIG. 3.6 – Architecture d'intégration audiovisuelle par pondération (D'après Adjoudani (1998)).

$$p(i/x_A, x_V) = \frac{p_A(i)p_V(i)}{\sum_{j=1}^{10} p_A(j)p_V(j)}$$

3.1.3 Modèle RD

Dans ce type de modèle, les informations visuelles sont codées dans un format compatible avec les représentations de la modalité auditive qui est considérée comme la modalité dominante. Un tel format peut être la fonction de transfert du conduit vocal. Cette fonction de transfert est estimée séparément par un module de traitement du signal et par les indices visuels à partir des deux entrées auditive et visuelle. L'estimation de la fonction de transfert peut être effectuée par exemple par association à partir de l'entrée visuelle et par un traitement cepstral à partir de l'entrée auditive. Les deux estimations sont ensuite fusionnées et l'ensemble ainsi obtenu est présenté à un classifieur phonétique (voir figure 3.7). Il s'agit là d'une fusion précoce.

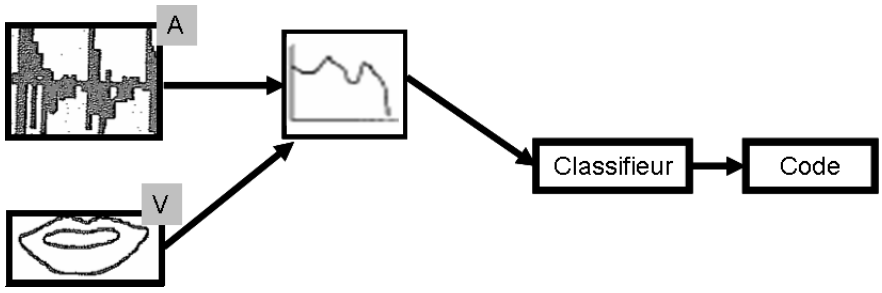

FIG. 3.7 – Modèle à recodage dans la modalité dominante.

Il existe relativement peu d'implémentations de ce modèle (Yuhas *et al.*, 1989, 1990; Watanabe et Kohda, 1990; Robert-Ribes *et al.*, 1996; Teissier *et al.*, 1999). Dans une tâche de reconnaissance audiovisuelle de neuf voyelles en anglais, Yuhas *et al.* (1989, 1990) ont implémenté le modèle RD. Le recodage des informations visuelles dans l'espace de la modalité acoustique (en un spectre acoustique) est fait grâce à un réseau de neurones. Le spectre estimé à partir des caractéristiques visuelles est combiné avec le spectre provenant de l'analyse acoustique pour finalement obtenir le spectre audiovisuel. La combinaison des deux spectres est réalisée en

pondérant chaque entrée par un poids variant suivant le niveau de bruit de l'audio. Le spectre audiovisuel résultant alimente ensuite un deuxième réseau de neurones pour enfin identifier la voyelle produite. Cette implémentation a été adapté par Robert-Ribes *et al.* (1996) aux voyelles du Français avec quelques différences. En effet, le classifieur audiovisuel employé par Robert-Ribes *et al.* (1996) est un classifieur gaussien tandis que le recodage de la modalité visuelle en une représentation auditive est réalisé par association utilisant des distances euclidiennes. Teissier *et al.* (1999) emploient aussi un schéma identique avec juste une différence dans la manière de convertir l'entrée visuelle en un spectre auditif équivalent. En effet, les auteurs s'appuient sur une association linéaire qui apprend la régression entre des entrées visuelles 3D et les composantes acoustiques 3D du son non bruité correspondant à l'entrée.

3.1.4 Modèle RM

Ce modèle est inspiré en partie de la théorie motrice de la perception de la parole proposée par Liberman et Mattingly (1985). Selon cette théorie, l'information phonétique est perçue par un module spécialisé dans la détection des gestes planifiés par le locuteur qui sont le fondement des catégories phonétiques. Dans ce type d'architecture, les deux entrées sont codées dans une nouvelle représentation commune dans l'espace moteur avant d'être classifiées. Dans ce modèle, le choix de l'espace moteur est crucial pour l'intégration. En général, les paramètres du conduit vocal sont les plus choisis comme représentation commune. Dans ce cas, à partir de chaque entrée, visuelle ou acoustique, les principales caractéristiques articulatoires sont estimées. Ensuite, la représentation finale est définie en additionnant les deux projections avec une certaine pondération et elle est fournie au classifieur pour la reconnaissance du code (voir figure 3.8).

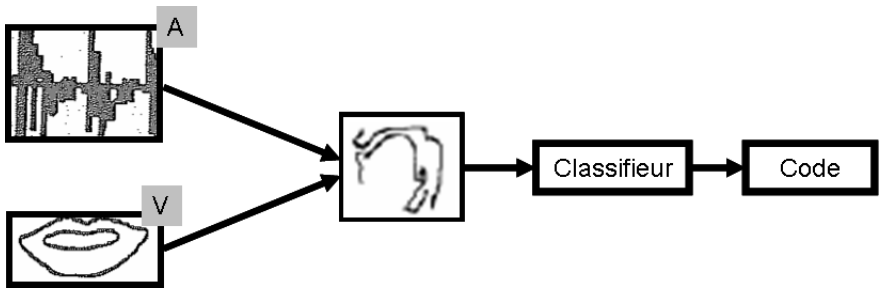

FIG. 3.8 – Modèle à recodage dans la modalité motrice.

A notre connaissance, seuls Teissier *et al.* (1999) et Robert-Ribes *et al.* (1996) ont proposé une implémentation de ce type de modèle. Dans l'implémentation de Teissier *et al.* (1999), qui a pour objectif la reconnaissance de voyelles du Français, la transformation des deux entrées en représentation motrice est réalisée par des associations linéaires. Les auteurs ont choisi comme espace moteur des caractéristiques articulatoires représentées par trois paramètres qui fournissent les corrélas articulatoires des dimensions d'arrondissement, d'ouverture-fermeture et d'avant-arrière : les coordonnées horizontale et verticale, respectivement X et Y, du point le plus haut de la langue et l'étirement, noté A, du contour interne des lèvres. Le réglage des associateurs est obtenu en définissant ces trois paramètres pour chaque voyelle d'un corpus d'apprentissage.

Le paramètre A est mesuré directement sur l'entrée visuelle. Par contre, les auteurs ont utilisé comme coordonnées X et Y des valeurs prototypiques provenant d'un expert phonétique. La classification est ensuite réalisée de la même façon que pour le modèle RD, c'est-à-dire avec un classifieur Gaussien.

3.2 Eléments du choix d'une architecture : théoriques et expérimentaux

Dans une tâche de fusion de deux modalités, un des principaux problèmes réside dans le choix du modèle d'intégration le plus approprié. Suivant la perspective envisagée, modélisation des processus cognitifs ou reconnaissance de la parole, le modèle retenu doit rendre compte au mieux des données au niveau reconnaissance automatique. Dans ce sens, Robert-Ribès (1995) propose une taxinomie mettant en correspondance les 4 modèles d'intégration décrits précédemment avec les modèles généraux de la psychologie cognitive (figure 3.9). Cette taxinomie s'organise autour de 3 questions :

1. Peut-on considérer, en fonction de l'interaction entre les modalités, une représentation intermédiaire commune ? Sinon, c'est un modèle ID à préconiser.
2. Dans le cas de l'existence d'une représentation intermédiaire, l'intégration est-elle tardive ou précoce pour accéder au code ? Une intégration est tardive quand elle suit l'intervention d'un processus de décodage ; c'est-à-dire qu'il y'a d'abord extraction des informations auditives et visuelles, puis fusion (c'est le cas du modèle IS). Dans le cas où la fusion intervient au cœur du processus d'extraction de l'information, l'intégration est dite précoce.
3. Si l'intégration est précoce, quelle forme prend le flux commun des données après fusion ? Plus précisément, existe-t-il une modalité dominante susceptible de fournir la représentation intermédiaire commune dans une architecture à intégration précoce (cas du modèle RD) ? ou cette représentation est elle amodale (cas du modèle RM) ?

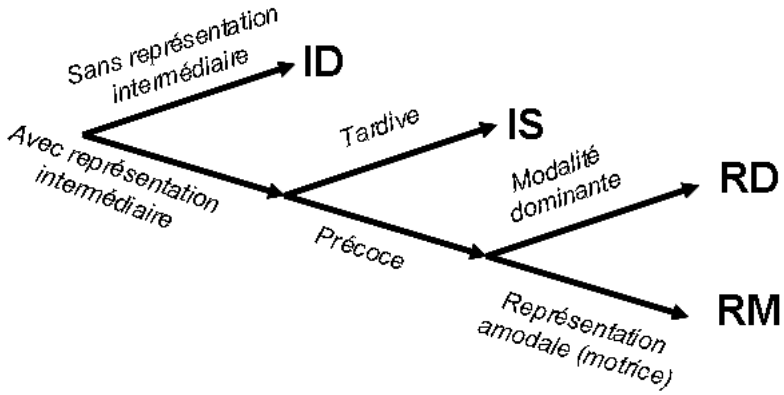

FIG. 3.9 – Taxinomie des modèles d'intégration (d'après Robert-Ribès (1995)).

Parmi les 4 architectures, les modèles ID et IS sont ceux qui sont les plus fréquemment utilisés en reconnaissance de parole (Schwartz, 2004). Les deux autres modèles sont très rarement implémentés et ceci malgré le fait qu'ils semblent être les plus pertinents au regard des données

issues de la psychologie expérimentale. C'est précisément ces données qui ont conduit Schwartz *et al.* (1998) à privilégier le modèle RM.

3.2.1 Etudes comparatives

Dans cette sous-section nous passons en revue quelques études comparant les quatre architectures d'intégration.

3.2.1.1 ID *vs.* IS

Adjoudani (1998) rapporte plusieurs études menées dans le domaine de la reconnaissance audiovisuelle de la parole, parmi lesquelles Robert-Ribès (1995); Movellan et Chadderdon (1996); Alissali *et al.* (1996); Jourlin (1996); Su et Silsbee (1996); Silsbee et Su (1996), comparant les deux modèles IS et ID. Il conclut que la grande partie de ces études semblent avantager le modèle IS (Duchnowski *et al.*, 1994; Robert-Ribes *et al.*, 1996; Alissali *et al.*, 1996; Su et Silsbee, 1996) tout en notant le statut quo entre ces deux modèles relevé dans d'autres études (Silsbee, 1994; Silsbee et Su, 1996; Jourlin, 1996). L'auteur a aussi procédé, en tenant compte des résultats de ces études comparatives, à un regroupement des avantages ($\oplus$) et des inconvénients ($\ominus$) de chacun de ces deux modèles.

Modèle ID

- $\oplus$ Modèle facile à implémenter : l'observation bimodale peut se former à partir d'une concaténation des indices des deux modalités.
- $\oplus$ Possibilité de pondérer chaque canal à condition de disposer d'un corpus d'apprentissage de taille importante ((Silsbee et Su, 1996)).
- $\ominus$ Modèle nécessitant un corpus de taille relativement grande par rapport au modèle IS (Jacob et Sénac, 1996) car la taille des modèles à apprendre est plus importante.
- $\ominus$ Nécessité d'une topologie identique des deux sources.
- $\ominus$ Conservation de la coordination temporelle entre les deux modalités durant la fusion.
- $\ominus$ Le problème de déphasage n'est pas géré.
- $\ominus$ Apprentissage adapté à chaque niveau du Rapport Signal sur Bruit (RSB) de l'entrée acoustique (Silsbee et Su, 1996).

Modèle IS

- $\oplus$ Nécessité d'un corpus moins important pour l'apprentissage que pour le modèle ID : grâce au traitement séparé de chaque modalité.
- $\oplus$ Les deux modalités ne demandent pas forcément d'avoir la même architecture de reconnaissance.
- $\oplus$ Le modèle s'approche plus des hypothèses faites sur la perception audiovisuelle (Robert-Ribès, 1995; Massaro, 1996).
- $\oplus$ Capable de traiter l'asynchronie : par exemple dans le cadre d'un mot entre son état initial et final.
- $\ominus$ Le module d'intégration peut être complexe et dépendant du corpus.

Après avoir comparé les modèles IS et ID, Adjoudani (1998) a implémenté, comme nous l'avons vu précédemment dans la section précédente, ces deux modèles et en a comparé les

performances dans une tâche de reconnaissance audiovisuelle de la parole avec un niveau de bruit variant sur l'entrée auditive. Les résultats obtenus montrent que malgré que le modèle ID améliore significativement les scores de reconnaissance quand l'entrée acoustique est bruitée (on passe de 3% en reconnaissance acoustique à 33% en audiovisuelle pour la condition d'un RSB acoustique de -6 dB), l'intégration reste encore non optimale. Par contre, avec une pondération de chaque canal par son degré de confiance, le modèle IS peut donner des résultats meilleurs. Enfin, l'auteur conclut que la complémentarité audio/ vision est mieux exploitée en IS et ceci grâce au traitement séparé des deux modalités, même si dans ce cas la coordination audiovisuelle semble perdue mais peut être retrouvée à certains points d'ancrage. Inversement, le modèle ID exploite bien les covariations des entrées visuelle et auditive mais dans le cas où l'entrée auditive est bruitée la complémentarité entre l'entrée propre et l'entrée atténuée n'est pas aussi prise en compte à cause du traitement conjoint des deux sources.

3.2.1.2 RD *vs.* RM

Comme ces deux modèles sont peu utilisés dans la reconnaissance audiovisuelle de la parole, les comparaisons sont rares pour déterminer le plus performant des deux. Il est important de rappeler que la différence entre ces deux modèles est la nature de leur représentation commune au niveau de la fusion. Le modèle RD appliqué à la fusion en parole considère la modalité auditive comme dominante alors qu'elle peut ne pas l'être. De ce fait, la complémentarité naturelle entre le son et l'image est difficilement exploitable dans ce modèle. Robert-Ribès (1995), l'un des rares à implémenter les modèles RD et RM, démontre que le modèle RM est mieux adapté que le modèle RD à la structure de l'information audiovisuelle et à la complémentarité audio-visuelle.

3.2.1.3 Un vainqueur ?

Quel modèle choisir ? Il est difficile de répondre à cette question d'autant plus que très peu d'études ont comparé les quatre modèles de façon systématique. En implémentant ces quatre architectures, Teissier *et al.* (1999) ont tenté une comparaison. Selon les auteurs, une hiérarchie globale émerge confirmant en grande partie les conclusions décrites précédemment. Cette hiérarchie peut se résumer par :

$$ID = IS > RM > RD$$

De cette hiérarchie, il semble donc que les modèles de fusion *data-to-data* (modèles RD et RM) sont moins efficaces que les modèles *data-to-decision* (ID) ou *decision-to-decision* (IS). Ceci pourrait être expliqué par le fait que les modèles RD et RM réduisent trop tôt le nombre de dimensions nécessaires pour atteindre un niveau élevé de performance. Concernant les deux meilleurs modèles, à savoir ID et IS, les auteurs n'arrivent pas à obtenir des différences significatives permettant ainsi de favoriser un des deux. Ils montrent que le modèle ID peut être adapté à des flux d'entrée largement dégradés, et en même temps notent que le modèle IS est plus facile à contrôler (le réglage s'opère après la reconnaissance de chaque entrée).

A partir de cette étude et des conclusions faites par Adjoudani (1998), nous pouvons déjà faire une première sélection et se focaliser sur les deux modèles IS et ID. Les autres modèles, à

défaut d'optimiser les représentations au niveau des données de fusion, sont moins performants. Enfin, il est important de noter que Teissier *et al.* (1999) ont comparé les 4 modèles tout en choisissant de pondérer les entrées de chaque fusion de manière similaire. Ceci nous ramène au second problème à résoudre dans un système d'intégration de modalités : comment fusionner les données ?

3.2.2 Nature de la fusion

Après avoir choisi le modèle de fusion adéquat, un second problème reste à résoudre. Si le choix du modèle de fusion est capital, la nature du mécanisme de fusion est de première importance pour l'efficacité du système de reconnaissance de parole. Plus particulièrement, il faut déterminer comment contrôler le processus de fusion en fonction d'un certain nombre de paramètres. A cet égard, il existe, en traitement de l'information, trois types de processus de fusion des données se différenciant selon les actions qu'ils déclenchent en fonction du contexte (Bloch, 1996) :

- Le processus Indépendant du Contexte et à Comportement Constant (noté ICCC) : dans ce type, les opérateurs de fusion gardent le même comportement quelque soit les valeurs des informations à combiner qui sont calculées sans aucune information contextuelle ou externe. En d'autres termes, la loi de la fusion dans ce cas est fixe (exemple multiplication ou addition).
- Le processus Indépendant du Contexte et à Comportement Variable (noté ICCV) : les opérateurs dans ce cas restent indépendants du contexte comme dans le processus ICCC. Cependant, leur comportement dépend cette fois des valeurs des informations à fusionner. La loi de fusion est donc ici variable selon les valeurs des entrées. Un tel processus pourrait être par exemple un système qui additionne les informations d'entrée avec des coefficients de pondération différents suivant le niveau de ces entrées. L'entrée la plus importante pourrait donc, dans ce cas, être amplifiée par un coefficient plus important.
- Le processus Dépendant du Contexte (noté DC) : ce processus prend en compte des connaissances sur l'environnement extérieur qui peuvent lui permettre de privilégier plus ou moins une entrée suivant la nature de ces informations contextuelles. Par exemple, dans une application de reconnaissance de la parole où une des voies (auditive ou visuelle) est bruitée, une variable estimant le niveau de bruit dans ce canal vient ajuster l'importance relative des entrées auditive et visuelle.

Si nous nous concentrons sur la fusion en perception de la parole, il apparaît que les données de psychologie expérimentale semblent plaider en faveur d'une fusion contrôlée (Schwartz, 2004). Plus particulièrement, ces données semblent favoriser le processus DC. Cependant, ce fait n'est pas toujours vérifié expérimentalement. Il suffit par exemple de rappeler le célèbre modèle FLMP (Massaro, 1987, 1998). Ce modèle, sans trop rentrer dans les détails, consiste à identifier séparément les classes phonétiques dans chaque modalité et à les fusionner ensuite suivant un processus ICCC. C'est en effet un exemple du modèle d'intégration IS avec une fusion constante.

3.3 L'asynchronie audio-visuelle dans la fusion

Bien que les données audio et visuelles soient corrélées, elles ne sont pas synchrones et l'activité visuelle précède souvent le signal audio. En effet, lorsqu'un locuteur prononce une suite de sons, les organes phonatoires qui produisent ne réagissent pas de manière synchrone. Ce phénomène est appelé *rétention et anticipation* labiale (Abry et Lallouache, 1991; Yehia *et al.*, 1997). Il ne s'agit pas uniquement d'un décalage entre les deux flux auditif et visuel, mais la durée de l'événement acoustique et celle de l'événement visuel peuvent être différentes.

Dans un système de reconnaissance automatique de la parole audio-visuelle, il faut prendre en compte cette asynchronie. La question qui se pose donc est la suivante : en utilisant un des quatre modèles décrits précédemment comment gérer cette asynchronie ?

Mais avant tout, revenons sur la classification de ces modèles. La taxinomie proposée par Robert-Ribès (1995) peut être vue d'une autre façon. Potamianos *et al.* (2003), dans une revue de questions sur le domaine, proposent de classifier les 4 modèles de base en trois familles : fusion de représentations, fusion des décisions (scores) et fusion hybride. Dans la première famille, les modèles s'appuient sur une application d'un seul classifieur sur un vecteur concaténé de données audio et visuelles, ou sur n'importe quelle transformation appropriée (modèles de type ID, RM et RD). En revanche, les méthodes de la deuxième famille s'appuient sur la combinaison de décisions, éventuellement partielles, prises indépendamment sur les informations audio d'une part et visuelles d'autre part (modèles de type IS). Dans la troisième famille, les modèles combinent des modèles des deux premières familles dans une structure de type fusion de décisions.

Concernant l'asynchronie, sa gestion est implicite dans les modèles de type fusion de décisions et fusion hybride, où la classification et la segmentation sont effectuées séparément pour chaque flux. De nombreuses études proposent des modèles d'intégration audio-visuelle pouvant être classés dans ces deux familles (Bregler *et al.*, 1993; Alissali *et al.*, 1996; Jacob et Sénac, 1996; Jourlin, 1998; André-Obrecht *et al.*, 1997; Rogozan et Deléglise, 1998). En revanche, les modèles de la famille fusion de représentations ne permettent pas de gérer l'asynchronie (Adjoudani et Benoît, 1996; Teissier *et al.*, 1999; Neti *et al.*, 2000; Chen, 2001).

Alissali *et al.* (1996) proposent un modèle hybride ID + IS d'intégration audio-visuelle. Le principe de ce modèle consiste tout d'abord en un sous-système acoustico-visuel s'appuyant sur une intégration directe (ID), qui propose N meilleures suites de phonèmes possibles. Ces suites sont évaluées par un second sous-système purement visuel, sans remettre en cause les frontières inter-unités. Ensuite, les sorties de ces deux sous-systèmes sont combinées. Une fonction de décision permet enfin d'obtenir la suite de phonèmes reconnue. Les auteurs enrichissent ensuite ce modèle en utilisant des visèmes au niveau du sous-système visuel. Ainsi, les paramètres du sous-système visuel sont mieux estimés grâce au nombre de visèmes inférieur au nombre des phonèmes. Les résultats expérimentaux montrent une dégradation des performances de reconnaissance par rapport à un modèle IS classique. Cette dégradation est attribuée, selon les auteurs, à un choix inadéquat de visèmes. Par ailleurs, ce modèle a la particularité de gérer dynamiquement l'asynchronie entre les flux acoustique et labial.

Ce modèle a été repris par Rogozan et Deléglise (1998) avec quelques modifications concer-

nant les classifieurs utilisés et surtout l'introduction d'une pondération adaptative des flux acoustique et labial. En effet, les auteurs utilisent des modèles de Markov cachés continus (CHMM) comme classifieurs dans les deux sous-systèmes. De plus, les modalités visuelle et acoustique sont pondérées chacune par un poids adapté dynamiquement pendant le processus en fonction du contexte et du rapport signal sur bruit. Par ailleurs, la gestion de l'asynchronie ne semble pas être influencée par ces modifications.

André-Obrecht *et al.* (1997) proposent un modèle maître-esclave segmental s'appuyant sur deux modèles HMM corrélés parallèles. Le premier est un modèle HMM maître qui consiste en un modèle HMM classique avec trois états et trois fonctions de densité de probabilités (PDFs), correspondant à trois traits labiaux caractérisant les visèmes : ouverture, semi-ouverture et fermeture des lèvres. Le second est un modèle HMM esclave construit de façon hiérarchique en introduisant soit les parties correspondant au phonème soit les transitions entre deux phonèmes (*pseudo-diphones*). La particularité d'un tel modèle maître esclave est que le modèle maître (labial) à N états (ici 3) conditionne les distributions et les transitions du modèle esclave (acoustique) à chacune des unités de reconnaissance. Pour chaque trame, N mises en correspondances d'observations acoustiques et labiales sont possibles. Ainsi, l'asynchronie est gérée sur un intervalle temporel de N trames. André-Obrecht *et al.* (1997) ont comparé les performances en reconnaissance de leur modèle maître esclave à un modèle ID classique en conditions de signal acoustique propre et bruité. Les résultats expérimentaux ne montrent pas une différence significative. Enfin, il est à noter que le modèle maître esclave nécessite un corpus de grande taille, vu le nombre important de paramètres à estimer. Souvent, en pratique, avec des corpus de taille faible, le modèle maître (labial) est simplifié comme ce fut le cas dans l'étude de André-Obrecht *et al.* (1997).

Jourlin (1998) propose un produit de modèles HMM qui vise à gérer l'asynchronie qui existe entre les flux audio et visuel. Pour une suite d'observations données, ce modèle présente l'avantage de conserver le produit des probabilités données par le modèle acoustique et le modèle labial. De plus, cette méthode prend en compte les différences de topologie entre les deux modalités puisque le produit n'est réalisé qu'après apprentissage séparé des deux modèles.

3.4 Conclusion

Dans ce chapitre, nous avons décrit un ensemble de modèles d'intégration audio-visuelle. Cette intégration peut être réalisée avec quatre modèles basiques : ID, IS, RD et RM. Ces derniers peuvent être classifiés en deux grandes familles. La première famille, fusion de représentations, regroupe les modèles s'appuyant sur l'entraînement d'un seul classifieur appliqué sur un vecteur des représentations audio et visuelles concaténées, ou sur toute transformation sur ce vecteur (modèles ID, RM, RD). La seconde famille, fusion de décisions, regroupe des modèles reposant sur une fusion des sorties de deux classifieurs mono-modal. A ces deux familles, une troisième famille, fusion hybride, peut être considérée, qui consiste à combiner deux modèles des deux familles précédentes.

La comparaison entre les quatre modèles classiques semblent plutôt favoriser les modèles ID et IS. Cependant, ces derniers ne peuvent être départagés. Par ailleurs, il semble que les modèles

orientés fusion de décisions et modèles orientés fusion hybride permettent de gérer l'asynchronie entre les flux. Dans le cas du code LPC, l'asynchronie entre le flux manuel et labial a été démontrée (Attina *et al.*, 2004; Gibert *et al.*, 2005). Il est donc impératif de prendre en compte cette asynchronie dans un modèle de fusion de ces deux flux.

Chapitre 4

Outils statistiques et d'analyse de données

Dans ce chapitre nous définissons les différentes méthodes que nous utiliserons par la suite dans la partie expérimentale.

4.1 Les Modèles de Markov Cachés

Un modèle de Markov caché (MMC) ou chaîne de Markov cachée (CMC)[1] - connu sous le sigle HMM[2] pour *Hidden Markov Model* - est un modèle statistique dans lequel le système modélisé est supposé être un processus Markovien avec des paramètres inconnus. Le défi est donc de déterminer ces paramètres à partir d'autres paramètres observables. Puis, les paramètres extraits peuvent ainsi être employés pour réaliser une analyse telle que la reconnaissance de formes. Historiquement, les modèles de Markov cachés ont été introduits dans les années 60-70 par Baum et ses collaborateurs. Ils ont été ensuite utilisés dans les systèmes de reconnaissance de la parole à partir des années 80 et appliqués après dans d'autres domaines tels que la bioinformatique, l'intelligence artificielle, la reconnaissance des gestes ...

4.1.1 Définition (Rabiner, 1986)

Un modèle HMM est défini comme un ensemble d'*états*, chacun d'entre eux associé à une distribution de probabilité (en général multidimensionnelle). Les transitions entre les états sont régies par un ensemble de probabilités appelées *probabilités de transition* Dans un état particulier, un résultat ou *observation* peut être généré conformément à la distribution de probabilité associée. Par opposition à un modèle de Markov classique où l'état est directement observable par un observateur externe, dans un modèle HMM, l'état n'est pas directement observable et seulement des variables influencées par l'état le sont. Les états sont donc cachés, d'où le nom de modèle de Markov caché.

Un modèle HMM est défini par les éléments suivants :

[1] *La dénomination correcte est automate de Markov à états cachés mais elle est moins employée*

[2] *C'est ce que nous allons utilisé dans la suite pour faire référence à un modèle de Markov caché.*

- N : le nombre d'états du modèle. Les états seront notés x_i pour $1 \leq i \leq N$
- M : le nombre de symboles d'observation. Dans le cas où les observations sont continues, M est infini. Dans notre notation, les symboles d'observation de l'alphabet sont notés $Y = \{y_j\}$ pour $1 \leq j \leq M$.
- π : le vecteur de probabilités initiales des états. Concernant cet élément, un autre type de HMM utilise des états *start* et *end* et non une distribution d'états initiaux. Ce type d'HMM est notamment employé en bioinformatique.
- A : la matrice de transition où sont définies les probabilités de transition entre les états. Ces probabilités $A = \{a_{ij}\}$ sont définies comme :
$$a_{ij} = p(x_t = i | x_{t-1} = j), 1 \leq i, j \leq N$$
avec x_t désigne l'état courant à l'instant t. Les probabilités de transition a_{ij} doivent satisfaire les contraintes stochastiques :
$$a_{ij} \geq 0 \text{ et } \sum_{j=1}^{N} a_{ij} = 1, 1 \leq i, j \leq N$$
- **B** : la matrice de confusion (ou matrice d'observation) contenant les probabilités d'observation (ou probabilités d'émission) $B = \{b_j(k)\}$ associées aux états. Ces probabilités sont définies comme :
$$b_j(k) = p(y_t = v_k | x_t = j), 1 \leq, j \leq N, 1 \leq k \leq M$$
avec v_k dénote le k^{eme} symbole d'observation dans l'alphabet, et y_t le vecteur de paramètres actuel (où simplement observation actuelle) à l'instant t. Les probabilités d'observation satisfont aussi les contraintes stochastiques. Dans le cas d'observations continues, des densités de probabilités continues sont à utiliser.

Pour dénoter un modèle HMM le triplet $\lambda = (\pi\ , A, B)$ est généralement utilisé. Il est important de noter que chaque probabilité dans la matrice de transition (de confusion) est indépendante du temps. En d'autres termes, les matrices ne changent pas dans le temps quand le système évolue. En pratique, ceci est l'une des suppositions les plus discutables des modèles de Markov à propos des processus réels.

Dans la théorie des HMMs, des hypothèses sont faites pour une docibilité mathématique et informatique :

- Hypothèse markovienne : concernant la définition des éléments de la matrice de transition A, la probabilité de transition vers un état ne dépend que de l'état actuel et non des états rencontrés précédemment. Ainsi, la séquence des états constitue une chaîne de Markov[3] simple.
- Hypothèse de stationnarité : comme nous l'avons déjà évoqué, la matrice des probabilités de transition est indépendante de l'actuel temps. dans lequel les transitions prennent place. Mathématiquement :
$$p(x_{t_1+1} = j | x_{t_1} = i) = p(x_{t_2+1} = j | x_{t_2} = i) \text{ pour tout } t_1 \text{ et } t_2,$$
- Hypothèse d'indépendance des sorties (observations) : l'observation courante est stati-

[3] *Elle est caractérisée par la distribution conditionnelle de probabilité des états futurs, étant donné l'instant présent, ne dépend que de ce même état présent et pas des états passés. Mathématiquement, si $X(t)$, $t > 0$, est un processus stochastique, la propriété de Markov est définie ainsi : $p[X(t+h) = y | X(s) = x(s), s \leq t] = p[X(t+h) = y | X(t) = x(t)], \forall h > 0$*

quement indépendante des observations précédentes. Mathématiquement, cette hypothèse peut être formulée pour un HMM λ par :

$$p(Y|x_1, x_2, ..., x_T, \lambda) = \prod_{t=1}^{T} p(y_t|x_t, \lambda).$$

4.1.2 Utilisation et algorithmes

Une fois qu'un système est décrit comme un HMM , trois problèmes doivent être résolus. Les deux premiers sont des problèmes qu'on peut associer à la reconnaissance : détermination de la probabilité d'une séquence observée étant donné un HMM (c'est le problème de l'*évaluation*) ; et, étant donné un modèle HMM et une séquence d'observations, déterminer quelle séquence d'états cachés dans le modèle est la plus probable (c'est le problème de décodage). Le troisième problème est la génération d'un HMM étant donné une séquence d'observations (c'est le problème d'apprentissage).

4.1.2.1 Evaluation et l'algorithme de Forward

Ce problème se pose notamment quand nous avons, par exemple, plusieurs HMMs décrivant différents systèmes, et une séquence d'observations. Nous voulons ainsi connaître quel est le HMM ayant la plus forte probabilité d'avoir généré cette séquence. En d'autres termes, pour un modèle $\lambda = (\pi\ ,\ A,\ B)$ et une séquence d'observations $Y = y_1, y_2, ..., y_T$, nous avons à calculer la probabilité $p(Y|\lambda)$. Un calcul de cette probabilité implique un nombre d'opérations de l'ordre de N^T. Heureusement, une autre méthode, ayant une complexité inférieure, existe. Cette méthode utilise une variable intermédiaire appelée variable "avant" ou *forward* ; d'où le nom de l'algorithme Forward (ou "avant").

Algorithme Forward : Cet algorithme est utilisé pour calculer la probabilité d'une séquence d'observation de longueur T :

$$Y = y_{1,2}\ , ..., y_T$$

avec chaque y est un élément de l'ensemble observable. La variable intermédiaire $\alpha_t(i)$ est définie comme la probabilité de la séquence d'observation partielle $Y^t = y_{1,2}\ , ..., y_t, t \leq T$, qui se termine à l'état i. Les probabilités intermédiaires (ou partielles) sont calculées de manière récursive en calculant premièrement ces probabilités pour tous les états à $t = 1$.

$$\alpha_1(j) = \pi(j).b_j(1)\ \text{, pour } 1 \leq j \leq N$$

Ensuite, pour chaque instant, $t = 2, ..., T$, les probabilités partielles sont calculées pour chaque état par la relation récursive suivante :

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} (\alpha_t(i)a_{ij})b_j(t)\ \text{, pour } 1 \leq j \leq N, 1 \leq t \leq T-1$$

Avec cette relation, nous pouvons alors calculer la probabilité intermédiaire à l'instant T pour chaque état j, $\alpha_T(j)$. Et finalement, la somme de toutes les probabilités partielles à l'instant T fournit la probabilité requise :

$$p(Y|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Pour récapituler, chaque probabilité partielle (à l'instant $t > 2$) est calculée à partir de tous les états précédents. De façon similaire, nous pouvons définir une variable "arrière" ou *backward* $\beta_t(i)$ comme la probabilité de la séquence d'observation partielle $y_{t+1}, y_{t+2}, ..., y_T$, étant donné que l'état courant est i. Pour calculer les $\beta_t(i)$, il existe aussi, comme pour les $\alpha_t(i)$, une relation récursive :

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(t+1) \text{ , pour } 1 \leq i \leq N, 1 \leq t \leq T-1$$

avec

$$\beta_T(i) = 1 \text{ , pour } 1 \leq i \leq N.$$

Si nous cherchions un lien entre les deux variables intermédiaires $\beta_t(i)$ et $\alpha_t(i)$, nous pouvons remarquer que :

$$\alpha_t(i)\beta_t(i) = p(Y, y_t = i|\lambda), \text{ pour } 1 \leq i \leq N, 1 \leq t \leq T$$

Ainsi, la somme de ce produit donne une autre façon pour calculer la probabilté $p(Y|\lambda)$, tout en utilisant les probabilités forward et backward :

$$p(Y|\lambda) = \sum_{i=1}^{N} p(Y, y_t = i|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i), \text{ pour } 1 \leq t \leq T$$

4.1.2.2 Décodage et l'algorithme de Viterbi

Le problème du décodage se pose quand, étant donné une série d'observations, nous avons à trouver la séquence la plus probable des états cachés d'un modèle HMM. Ce problème est d'autant plus intéressant que dans plusieurs cas, les états cachés du HMM représentent quelque chose de non observable directement. Pour déterminer la séquence des états cachés la plus probable, étant donné une séquence d'observations, $Y = y_1, y_2, ..., y_T$, et un HMM $\lambda = (\pi$, A, $B)$, l'algorithme de Viterbi est le plus utilisé. Dans cette méthode, la séquence complète des états avec le maximum de vraisemblance est trouvée.

Algorithme de Viterbi : L'algorithme peut se résumer formellement de la façon suivante :

- Pour chacun des états, calcul par récurrence de la variable intermédiaire :

  $$\delta_t(i) = \max \quad p(x_1, x_2, ..., x_{t-1}, x_t = i, y_1, y_2, ..., y_{t-1}|\lambda)$$

  Le maximum étant calculé sur toutes les séquences d'états possibles, $x_1, x_2, ..., x_{t-1}$. Ce calcul se fait de manière récursive en deux étapes : - Initialisation :

  $$\delta_1(j) = \pi(j).b_j(1) \text{ , pour } 1 \leq j \leq N$$

  - Relation récursive :

  $$\delta_{t+1}(j) = b_j(t+1)\{\max_{1 \leq i \leq N} \quad \delta_t(i)a_{ij}\}, \text{ pour } 1 \leq j \leq N, 1 \leq t \leq T-1$$

- Calcul de $\delta_T(j)$, $1 \leq j \leq N$, en utilisant cette dernière récursion et en retenant toujours un pointeur sur l'état "élu" dans une opération de maximisation.

– Détermination de l'état final du système ($t = T$) le plus probable :

$$i_t = \arg\max_{1 \leq j \leq N} \quad (\delta_T(i))$$

– Suivi du chemin le plus probable en revenant en arrière, soit : Si on note :

$$\phi_t(i) = \arg\max_{1 \leq j \leq N} \quad (\delta_{t-1}(j))$$

la séquence d'état la plus probable peut être trouvée par :

$$i_t = \phi_{t+1}(i_{t+1})$$

Et en fin, la séquence $i_1, i_2, ..., i_T$ est la séquence la plus probable des états cachés pour la séquence d'observation considérée.

4.1.2.3 Apprentissage

Le troisième, et le plus difficile, problème associé aux HMMs est de prendre une séquence connue d'observations pour représenter un ensemble d'états cachés, et d'obtenir le HMM $\lambda = (\pi\ ,\ A,\ B)$ qui est le modèle le plus probable décrivant ce qui est observé. En d'autres termes, dans plusieurs cas d'applications, le problème de l'apprentissage concerne la façon avec laquelle les paramètres du HMM sont ajustés, étant donné un ensemble d'observations (appelé *ensemble d'apprentissage*). Les paramètres du HMM à optimiser peuvent être différents d'une application à l'autre. De ce fait, il peut y avoir divers critères d'optimisation pour l'apprentissage, chacun d'entre eux étant choisi selon l'application considérée. Parmi ces critères, nous trouvons le critère du maximum de vraisemblance et de l'Information Maximum Mutuelle (MMI pour Maximum Mutual Information). Nous nous contentons ici de décrire un seul algorithme permettant de générer les paramètres d'un HMM à partir d'une séquence d'observations. Il s'agit de l'algorithme de Baum-Welch avec un critère de maximum de vraisemblance. Cet algorithme est aussi connu sous le nom de *Forward-Backward*.

Algorithme de Forward-backward Cet algorithme est utilisé quand les matrices A et B d'un HMM ne sont pas directement mesurables, comme c'est souvent le cas dans plusieurs applications réelles. Plus formellement, on considère une unique séquence d'observation $Y = y_1, y_2, ..., y_T$. Notre but est de trouver les paramètres $\lambda = (\ A,\ B)$ qui maximisent la probabilité de générer Y avec le modèle. Formellement, les calculs doivent maximiser la quantité :

$$Q(\lambda, \overline{\lambda}) = \sum_x p(x|Y, \lambda) log\{p(Y, x, \overline{\lambda})\}$$

où x désigne un état donné et $\overline{\lambda}$ le modèle estimé. Pour décrire l'algorithme nous avons à définir deux variables intermédiaires : - $\xi_t(i, j) = p(x_t = i, x_{t+1} = j|Y, \lambda)$: la probabilité d'être dans l'état i à l'instant t et dans l'état j à l'instant $t+1$. - $\gamma_t(i) = p(x_t = i|Y, \lambda)$: la probabilité d'être dans l'état i à l'instant t étant donné la séquence d'observation et le modèle HMM. Ces deux variables peuvent être exprimées en fonction des variables forward, $\alpha_t(i)$, et backward, $\beta_t(i)$, définies précédemment. Pour résumer, l'algorithme peut être décrit de la façon suivante :

Initialisation : Des paramètres arbitraires pour le modèle sont choisis ; entre autre, les valeurs de π sont choisies aléatoirement tandis que les variables A et B sont initialisées. Par exemple, les valeurs de A sont fixées à priori et celles de B sont initialisées par une quantification vectorielle.

Itération : – Les variables A et B sont placées à leurs valeurs de pseudo-comptes.

– Calcul des variables $\alpha_t(i)$ et $\beta_t(i)$ pour chaque état i, en utilisant respectivement les algorithmes forward et backward.

– En déduire les variables $\xi_t(i,j)$ et $\gamma_t(i)$ en utilisant les expressions suivantes qui les lient aux variables forward et backward :

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}\beta_{t+1}(j)b_j(t+1)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}\beta_{t+1}(j)b_j(t+1)}$$

et

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$$

De ces deux expressions, il facile de remarquer que :

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j)$$

– L'étape suivante consiste à actualiser les paramètres du HMM en utilisant ce qu'on appelle les *formules de re-estimation* :

$$\overline{\pi} = \gamma_1(i), \text{ pour } 1 \leq i \leq N$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T}\xi_t(i,j)}{\sum_{t=1}^{T}\gamma_t(i)}, \text{ pour } 1 \leq i \leq N,\ 1 \leq j \leq N$$

$$\overline{b}_j(k) = \frac{\sum_{t=1}^{T}\xi_t(i,j)}{\sum_{t=1}^{T}\gamma_t(i)}, \text{ pour } 1 \leq j \leq N,\ 1 \leq k \leq M$$

L'algorithme est arrêté si le changement de la log-vraisemblance est inférieur à un seuil prédéfini ou si le nombre maximum d'itération est atteint.

4.1.3 Différents types de modèles HMM

Depuis le début de cette section, nous avons traité en général le modèle HMM en supposant qu'il est caractérisé par une matrice de transition des états pleine ; c'est-à-dire que les transitions peuvent s'effectuer à partir de n'importe quel état vers n'importe quel autre état. On parle ici de modèle ergodique. Un tel modèle est défini comme un HMM tel que tous les états sont accessibles à partir de n'importe quel autre état. Pour certaines applications, il est demandé d'imposer certaines contraintes sur la matrice de transition ; ce qui rend le modèle non ergodique. Dans ce sens, la littérature nous donne deux exemples types de modèles non-ergodique largement employés (Rabiner et Juang, 1986). Ces deux modèles sont appelés gauche-droite du fait que la séquence des états produisant la séquence d'observations doit toujours avancer de l'état le plus à gauche à l'état le plus à droite. Ils diffèrent par le fait qu'un est un simple gauche-droite dans lequel il y a qu'un seul chemin à travers les états, et l'autre est un parallèle gauche-droite dans lequel il y a plusieurs chemins. Un modèle gauche-droite (parallèle ou simple) impose une structure temporelle ordonnée pour le HMM dans laquelle l'état numéroté avec un numéro inférieur précède toujours l'état avec un numéro supérieur. La figure 4.1 illustre les trois structures HMM.

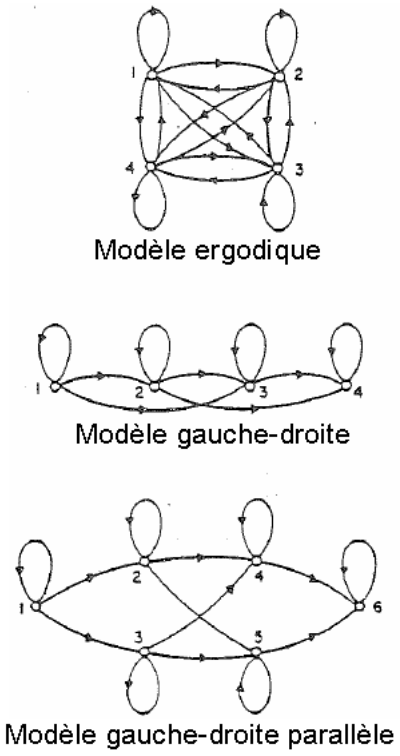

FIG. 4.1 – Trois types distincts de modèles HMM. Illustration avec un exemple de HMM à 4 état. (d'après Rabiner et Juang (1986).

4.1.4 Résumé

Le modèle de Markov caché est un outil statistique qui peut être défini quand les états d'un processus ne sont pas directement observables, mais sont indirectement et probabilistiquement observables comme un autre ensemble d'états. De tels modèles, appliqués dans des processus réels, imposent de résoudre trois problèmes :

- Evaluation : avec quelle probabilité un modèle donné génère-t-il une séquence d'observations donnée. L'algorithme forward résout efficacement ce problème.
- Décodage : quelle est la séquence d'états cachés la plus probable qui génère une séquence d'observations. L'algorithme de Viterbi résout ce problème.
- Apprentissage : comment optimiser (apprendre) les paramètres d'un modèle HMM à partir d'un échantillon donné de séquences d'observations. Ce problème peut être résolu en utilisant l'algorithme forward-backward.

Enfin, il est à noter un défaut habituel des modèles HMM qui concerne la sur-simplification associée à l'hypothèse markovienne ; c'est dire qu'un état dépend seulement de ses prédécesseurs directs et que cette dépendance est indépendante du temps. Cependant, les HMMs ont prouvé leur grande valeur dans des systèmes réels d'analyse et restent l'un des outils les plus utilisé en reconnaissance automatique de la parole.

4.2 L'analyse de variance (ANOVA)

4.2.1 Définition

L'analyse de la variance (ou test ANOVA du terme anglais : ANalysis Of VAriance) est un test statistique utilisé pour révéler les effets principaux et d'interaction d'un nombre de variables catégoriquement indépendantes (souvent appelées "facteurs") sur une variable dépendante. D'un coté, un effet "principal" est un effet direct d'un facteur sur la variable dépendante. De l'autre coté, un effet "d'interaction" est l'effet conjoint de deux ou plusieurs facteurs selon la variable dépendante.

Dans une ANOVA, la statistique clé est le test de Fisher (F-test)[4] qui teste si les moyennes de groupes formés par des valeurs du facteur (ou combinaisons de valeurs pour des facteurs multiples) sont suffisamment différentes. Si les moyennes des groupes ne diffèrent pas significativement alors le facteur n'a pas d'effet sur la variable dépendante. Si le test Fisher montre que globalement le (ou les) facteur(s) est (sont) liée(s) à la variable dépendante, alors de multiples tests de comparaison de signification sont utilisés pour explorer les valeurs du (des) facteur(s) qui sont le plus en relation avec la variable dépendante.

4.2.2 Hypothèses

L'ANOVA teste l'hypothèse nulle que les moyennes des groupes sont égales. Elle n'est pas un test pour étudier les différences de variances entre les groupes, mais plutôt assume une homogénéité relative des variances. Par conséquent, l'une des hypothèses principales de l'ANOVA est "l'homogénéité des variances" ; c'est-à-dire que les groupes formés par les facteurs sont relativement égaux en taille et ont des variances similaires sur la variable dépendante. l'ANOVA est sensible à cette hypothèse. Il est donc nécessaire de la tester avant toute utilisation ; même si dans le cas où les échantillons sont grands et les variances ne sont pas trop différentes, le résultat est tout de même significatif. Par ailleurs, l'ANOVA fait une autre hypothèse aussi importante. Elle suppose que toute variable dépendante a une distribution normale pour chaque catégorie de valeurs des variables indépendantes (hypothèse de la "normalité" des variables). La méthode est cependant assez robuste à la non normalité, ce qui permet de l'utiliser dans une grande variété de conditions.

4.2.3 Principe

L'ANOVA procède en décomposant d'abord la variance totale de l'échantillon en deux variances partielles, la variance inter-groupes et la variance résiduelle (souvent appelée aussi variance intra-groupe), et en comparant ensuite ces deux variances. Formellement, le déroulement de l'ANOVA peut être décrit de la façon suivante :

Données : Considérons notre échantillon composé de p groupes d'observations. Chaque groupe k est caractérisé par n_k observations $(x_{k,1}, ..., x_{k,n_k})$ d'une variable X_k.

[4] *Pour tester l'égalité de deux variances, en faisant le rapport des deux variances et en vérifiant que ce rapport ne dépasse pas une certaine valeur théorique que l'on cherche dans la table de Fisher (pour plus de détails voir http ://en.wikipedia.org/wiki/F-test).*

Notations : Nous notons : $N = n_1 + ... + n_p$ le nombre total des valeurs observées, et μ_k la moyenne du groupe k.

Hypothèse nulles : Les moyennes $\mu_1, ..., \mu_p$ sont égales.

Phases du test : – Calcul de la moyenne empirique (nous la différencions de la moyenne théorique μ_k en la notant m_k) de chaque classe :

$$m_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{k,i}$$

Et en déduire la moyenne empirique M totale de l'échantillon :

$$M = \frac{1}{N} \sum_{k=1}^{p} n_k m_k$$

– Calcul des variances : Variance de chaque classe :

$$V(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_{k,i} - m_k)^2$$

Variance intra-groupes = moyenne des variances :

$$V_{intra} = \sum_{k=1}^{p} \frac{n_k}{N} V(k)$$

Variance inter-groupes = variance des moyennes :

$$V_{inter} = \sum_{k=1}^{p} \frac{n_k}{N} (m_k - M)^2$$

– Comparaison des deux variances V_{intra} et V_{inter} : variable de test :

$$F_{p-1,N-p} = \frac{N-p}{p-1} \frac{V_{inter}}{V_{intra}}$$

– Test de signification : Si la valeur calculée de $F_{p-1,N-p}$ est supérieure à 1, alors il y a plus de variation entre les groupes qu'à l'intérieur de ces groupes. Ainsi, la variable liée au regroupement fait une différence. Si la valeur de $F_{p-1,N-p}$ est suffisamment au-dessus de 1, elle est comparée à la valeur critique de la loi de Fisher-Snedecor[5] en se donnant un risque de degrés de liberté $p-1$ et $N-p$. La différence entre les moyennes est significative si la valeur de $F_{p-1,N-p}$ est supérieure à la valeur critique. Dans le cas où la valeur de $F_{p-1,N-p}$ est environ 1, les différences dans les moyennes des groupes sont seulement des variations aléatoires. Notons pour bien comprendre ce test que si la valeur de $F_{p-1,N-p}$ s'éloigne de plus en plus au-dessus de 1, l'hypothèse nulle est de plus en plus rejetée.

4.2.4 MANOVA

La MANOVA (ou "Multivariate ANalysis Of VAriance") est une analyse de variance traditionnelle (ANOVA) adaptée à plusieurs variables dépendantes, qui ne peuvent être simplement combinées. L'analyse examine s'il y a des effets principaux et d'interaction des facteurs sur l'ensemble de ces variables considérées. Elle crée une combinaison linéaire des variables dépendantes et passe ensuite des tests pour examiner les différences dans la nouvelle variable en utilisant des méthodes similaires à l'ANOVA. Le facteur utilisé pour grouper les classes est catégorique. La MANOVA teste si cette variable catégorique explique une quantité significative de variabilité dans la nouvelle variable dépendante.

La MANOVA fait certaines hypothèses, comme l'ANOVA :

[5] *Les valeurs critiques de Fisher-Snedecor sont relevées dans une table dressée en fonction des degrés de liberté pour un intervalle de confiance donné*

- Distribution normale : les distributions des variables dépendantes doivent suivre une loi normale.
- Homogénéité des variances : cette homogénéité assume que les variables dépendantes exposent des niveaux de variances égaux.
- Homogénéité des covariances : dans le cas multivariable, avec de multiple variables dépendantes, il est exigé en plus de l'homogénéité des variances, que les inter-corrélations (covariances) soient aussi homogènes.
- Linéarité : la MANOVA suppose qu'il y a des relations linéaires entre toutes les paires de variables dépendantes, toutes les paires de covariances, et toutes les paires variable-covariance. Cependant, quand la relation dévie de la linéarité, la puissance de l'analyse peut être compromise.

Deuxième partie

Partie expérimentale

La transcription phonétique du code LPC nécessite de fusionner les informations de main et de lèvres correspondantes. Dans cette perspective, nous décrirons, dans le chapitre 5, nos données permettant de modéliser les informations issues des deux flux main et lèvres en vue de la fusion. Nous consacrerons ensuite le chapitre 6 au flux manuel où nous présentons nos méthodes pour reconnaître les positions et les configurations de la main. Dans le chapitre 7, nous analyserons le flux labial du point de vue de la modélisation et de la classification à partir des signaux des paramètres labiaux. Enfin, nous discuterons dans le chapitre 8 les modèles de fusion de ces deux flux en vue de la reconnaissance.

Chapitre 5

Description des données

Dans chaque expérience qui sera détaillée par la suite, les données d'étude proviennent d'un corpus global. Ce corpus a été enregistré tout en se fixant certains choix techniques concernant le type de matériel et les conditions d'enregistrement qui nous permettent d'extraire les informations sur les gestes main-lèvres. Ces choix sont établis en tenant compte de nos objectifs scientifiques et des limites techniques rencontrées dans certaines applications. Dans ce chapitre, nous discutons dans un premier temps les choix faits en les comparant avec l'état de l'art. Dans un second temps, nous décrivons le corpus global et le système d'enregistrement nous permettant d'acquérir les données souhaitées.

5.1 Les choix pour l'acquisition des données labiales et manuelles

Nous avons vu dans la partie *état de l'art* différentes approches d'extraction que ce soit pour les données labiales ou pour les gestes de la main. Pour les lèvres, deux approches sont généralement considérées : l'approche "image" et l'approche "modèle". Dans la première approche, des transformations appropriées (ACP, DCT ...) sont appliquées sur les pixels de l'image correspondant à la région de la bouche du locuteur. Dans la seconde approche, les contours des lèvres sont extraits à partir de l'image montrant le visage du locuteur. Un modèle des contours des lèvres peut alors être extrait. Ainsi, deux types de paramètres peuvent être utilisés comme porteurs de l'information labiale : des paramètres de contrôle du modèle ou des paramètres géométriques.

Pour les gestes de la main, la revue de la littérature sur la reconnaissance des gestes de la main, révèle deux approches : l'approche "gant instrumenté" et l'approche "vision". Dans la première, les mouvements de la main sont directement obtenus en utilisant des capteurs placés dans un gant instrumenté que porte le participant étudié. Dans la seconde, les mouvements de la main sont dérivés à partir d'un traitement de l'image vidéo du participant. Dans cette dernière approche, la main est soit considérée avec artifices, soit sans.

La question qui nous posons dans cette section est la suivante : quelle approche choisissons-nous dans notre étude pour chacune des composantes main-lèvres du code LPC ? Avant de répondre à cette question, il est important de rappeler l'objectif du projet TELMA qui est le

cadre global de notre étude. L'objectif est de réaliser un système de communication téléphonique (par l'image et le son) pour les personnes mal-entendantes ou sourdes. De ce fait, il est clair que nos données brutes sont des séquences vidéo et donc l'approche "gant instrumenté" est évidement exclue. De plus, il est aussi important de fixer un facteur important pour réaliser un enregistrement vidéo. Il s'agit de l'angle de vue. Ce facteur, comme nous l'avons vu précédemment, a une importance capitale pour l'extraction de l'information pertinente pour les lèvres et semble être le cas pour la main. Comme nous l'avons vu, la vue de face semble être la plus adaptée du point de vue de plusieurs chercheurs. Cette préférence est renforcée dans notre cas par le fait que, dans le système du code LPC, la communication se fait face-à-face ; c'est-à-dire que le codeur LPC fait face à la personne sourde. Ainsi, **nous avons enregistré nos données vidéo en vue de face**.

Pour répondre à la question posée, nous rappelons que notre objectif est de modéliser les flux labial et manuel du code LPC et de les intégrer. Ainsi, il est clair que pour une telle étude, la base de données doit être fiable et sans bruit de mesure. Par conséquent, nous imposons certaines contraintes afin d'avoir la base de données la plus fiable possible, avec des informations labiale et manuelle extraites de façon précise.

5.1.1 Choix pour les lèvres

Les deux approches "image" et "modèle" sont difficilement comparables et aucune ne prend le dessus sur l'autre. Dans les deux cas, certaines méthodes n'imposent aucune contrainte. D'autres nécessitent certaines conditions de pré-traitements, de vue, d'éclairage ou/et de matériel sophistiqué. Il clair qu'en imposant plus de contraintes, la complexité des méthodes diminue tout en augmentant la précision. Dans le cadre de notre travail, les lèvres du locuteur sont maquillées en bleu afin de séparer le vermillon des lèvres du reste du visage. En effet, le contraste chromatique naturel entre les lèvres et le reste du visage n'est pas toujours suffisant pour localiser les lèvres avec précision. D'autre part, il est facile de voir, en représentant la couleur du visage dans un espace de couleurs RVB[1], que la composante bleu est faible comparée aux deux autres composantes. Ainsi, le maquillage en bleu facilite la localisation et l'extraction de la zone des lèvres seulement par un simple seuillage par chrominance. En plus, en utilisant cet artifice, la fiabilité de mesure a été testée et prouvée dans plusieurs expériences de production, de perception ou même de synthèse (Lallouache, 1991; Benoit *et al.*, 1992; Cathiard, 1994; Robert-Ribès, 1995; Adjoudani, 1998).

Ce choix de maquiller les lèvres en bleu **nous place directement dans une approche "modèle"**. En effet, il n'est pas judicieux d'appliquer des transformations sur l'image de la zone en bleu alors que les contours des lèvres sont faciles à obtenir. En plus, en appliquant ces transformations sur ce type d'image, il est fort probable que le bleu perturbe la pertinence des informations en vue de la classification. A partir de ces contours, nous calculons **des paramètres géométriques** des lèvres en s'appuyant sur le système développé par Lallouache (1991). Ces paramètres labiaux ont démontré dans plusieurs études leur efficacité pour décrire les phonèmes (voir par exemple Benoit *et al.* (1992) pour consonnes et voyelles ; Robert-Ribès (1995) pour des

[1] *RVB = Rouge Vert Bleu*

voyelles).

5.1.2 Choix pour la main

Le fait que la main est en contact direct avec le visage (les deux ont la même couleur) pose un grand problème pour localiser d'abord la main et ensuite extraire ces gestes (position et configuration) à partir des images vidéo. D'ailleurs, à ce jour aucune étude n'est parvenue à réaliser cette tâche sans artifices (marqueurs ou gant coloré). Dans une démarche similaire (**donc avec artifices**), nous choisissons de placer deux pastilles en bleu (le bleu est la composante d'amplitude faible contenue dans la couleur de la peau de la main) sur le dos de la main et cinq autres sur les extrémités des doigts. Les deux pastilles sur le dos de main sont placées dans le but essentiel de suivre la trajectoire complète de la main et ainsi nous permettre de détecter la position de la main. Les pastilles sur les bouts des doigts servent principalement à déterminer la configuration de la main.

5.2 Corpus et acquisition des données

Les données proviennent de l'enregistrement vidéo d'un sujet prononçant et codant en code LPC un corpus de phrases. Pour obtenir des données précises (dans la mesure du possible) et prêtes à être analysées, nous avons enchaîné des phases de pré-traitements. Ainsi il a fallu tout d'abord traiter notre enregistrement vidéo pour extraire un lot de signaux afin de caractériser chaque modalité (les lèvres, la main et l'audio).

5.2.1 Sujet

Le sujet enregistré dans cette étude est une locutrice française titulaire du diplôme professionnel de codeur LPC que l'on appellera codeuse dans la suite (nous l'identifions aussi par la suite par SC). Le codeur est une femme agée de 30 ans à l'époque de l'enregistrement. Elle a obtenu son diplôme en 2001. Au moment de l'enregistrement, elle pratiquait de manière professionnelle le code LPC au lycée avec une moyenne de 24 heures par semaine.

5.2.2 Corpus

5.2.2.1 Présentation

Dans un système de reconnaissance de parole, le choix du corpus est essentiel. Dans notre tâche d'étudier la reconnaissance automatique des gestes du code LPC, le corpus est sensé être constitué pour couvrir l'ensemble des unités élémentaires nécessaires à la production du code LPC. Le corpus doit être en effet riche et réduit. D'un coté, il doit être phonétiquement équilibré, et donc contenir un minimum d'unités phonétiques élémentaires (diphones par exemple), pour pouvoir disposer d'une base de données représentative de toutes les combinaisons possibles de ces unités. De l'autre coté, le corpus doit être de taille réduite pour alléger les traitements et pré-traitements nécessaires notamment ceux qui demandent une intervention manuelle. C'est donc un compromis entre la taille du corpus et le nombre d'unités élémentaires présentes.

Notre corpus est composé des mêmes phrases que celles utilisées par Gibert *et al.* (2006) dans son travail portant sur la synthèse du code LPC à partir du texte. Les phrases sont initialement au nombre de 238. Pendant l'enregistrement, il a été demandé à la codeuse de répéter chaque phrase, qu'elle entend, au moins deux fois. De plus, quand la codeuse se trompe dans une phrase (erreur sur une syllabe ou un mot), celle-ci n'est pas rejetée mais considérée comme une phrase supplémentaire. Certaines phrases sont trop longues pour que la codeuse les répète correctement après les avoir entendues. Pour remédier à ce problème, les phrases longues ont été présentées en deux parties mais seulement pour la première répétition. Le nombre total de phrases ainsi obtenues est de 267 (référencées en annexe E) et en tenant compte des répétitions de 638 phrases. Le corpus initial (sans compter les répétitions ni les phrases ajoutées) se compose de 1814 diphones distincts avec un total de 7279 diphones.

En plus de l'équilibre phonétique, ce corpus possède certaines caractéristiques relatives aux transitions entre les clés (Gibert *et al.*, 2006). En effet, dans ce corpus, toutes les transitions de main sont présentes que ce soit pour de la configuration ou la position. Par contre, les 1680 diclés[2] (configuration+position) possibles ne sont pas toutes présentes. Un détail sur le nombre de chaque transition de main au niveau la configuration et de la position est présenté en annexe E.

5.2.2.2 Evaluation du corpus en réception

Afin de s'assurer du bon code de la codeuse, nous avons évalué notre corpus en réception. Nous avons présenté seulement les vidéos (sans le son) de notre corpus séquence par séquence à une participante sourde profonde âgée de 21 ans et qui utilise régulièrement le code LPC depuis son jeune age (nous l'appellerons dans la suite *décodeuse*). La tâche de la participante est de transcrire en Français chaque séquence vidéo (vers une phrase décodée). Notre évaluation du corpus consiste à comparer les phrases décodées avec les phrases prononcées par la codeuse (phrases codées). Pour éviter les problèmes avec la transcription orthographique, nous avons comparé les chaînes phonétiques des phrases. Nous avons aussi développé une interface graphique (sous Matlab) qui permet de comparer automatiquement chaque phrase codée avec sa correspondante décodée. Cette interface permet aussi de comparer en boucle plusieurs phrases. Pour faciliter la tâche, nous nous somme appuyés sur le transcripteur automatique texte vers chaîne phonétique de notre laboratoire.

Nous avons choisi un sous-corpus composé de **124 phrases** (la seconde répétition des 124 premières phrases du corpus) représentant le corpus global. Dans un premier temps, chaque phrase codée est comparée automatiquement à sa version décodée. Sur ce sous-corpus, la comparaison automatique délivre **55 phrases** ne contenant aucune erreur (les deux chaînes phonétiques codée et décodée sont identiques au caractère près). Dans un second temps, nous avons analysé les erreurs individuellement, afin de rechercher parmi les phrases contenant des erreurs les voyelles bien identifiées. Nous avons commencé par localiser et vérifié les erreurs et vérifié tout d'abord si l'erreur n'étant pas due au transcripteur automatique (comme tout système automa-

[2] *Une diclé est le nom donné par Gibert* et al. *(2006), en analogie avec le diphone, pour désigner une transition entre un doublet (configuration + position) et un autre.*

tique, des erreurs sont possibles). Dans le cas contraire, nous avons vérifié si la transcription de la phrase décodée était cohérente avec le code produit. Finalement, pour les voyelles, nous obtenons **94,8% de voyelles correctement identifiées** par la décodeuse. Ce score de décodage des voyelles en LPC, sera utilisé ultérieurement comme une référence, dans l'évaluation des différentes étapes du processus automatique de reconnaissance que nous présentons par la suite.

5.2.3 Procédé expérimental

L'enregistrement a été réalisé dans la chambre sourde du département Parole et Cognition[3] du laboratoire GIPSA-lab, à l'aide de deux caméras analogiques enregistrant à 50 Hz. Le locuteur était assis et sa tête maintenue fixe par un casque solidaire du mur afin que la tête reste dans le champ des caméras. Le locuteur portait des lunettes aveugles afin de protéger ses yeux du fort éclairage. Sur les lunettes, des pastilles bleues sont utilisées comme repère de référence pour les mesures de position de main et de doigts. Les lèvres ont été maquillées en bleu et des pastilles de couleur bleue ont été posées sur le dos de la main et à l'extrémité des doigts afin d'extraire les paramètres labiaux et de suivre le mouvement de la main (voir figure 5.1 et figure 5.2). Une première caméra en champ large dédiée à la main et au visage a été synchronisée à une seconde caméra en zoom sur les lèvres, chacune d'entre elle étant reliée à un magnétoscope Bétacam.

FIG. 5.1 – Image des lèvres de la codeuse (enregistrement en mode zoom).

En début d'enregistrement, un pavé de LEDs placé dans le champ des deux caméras est allumé durant 20 ms afin d'avoir un phénomène physique commun permettant de caler par la suite les étiquettes temporelles des deux enregistrements Bétacam. De plus un tableau plan quadrillé en centimètres a été enregistré par les 2 caméras en début de session afin de permettre par la suite une conversion des pixels en centimètres pour les différents paramètres extraits des images vidéo. Utilisant le poste Image-Parole du département Parole & Cognition de GIPSA-lab, la partie image de la bande vidéo a été numérisée en des images tramées toutes les 40 ms, en synchronie avec la bande son numérisée à 22050 Hz. Cette numérisation a nécessité un traitement préliminaire qui consistait à repérer chaque phrase sur la bande magnétique en la caractérisant

[3] *Ex ICP : Institut de la Communication Parlée.*

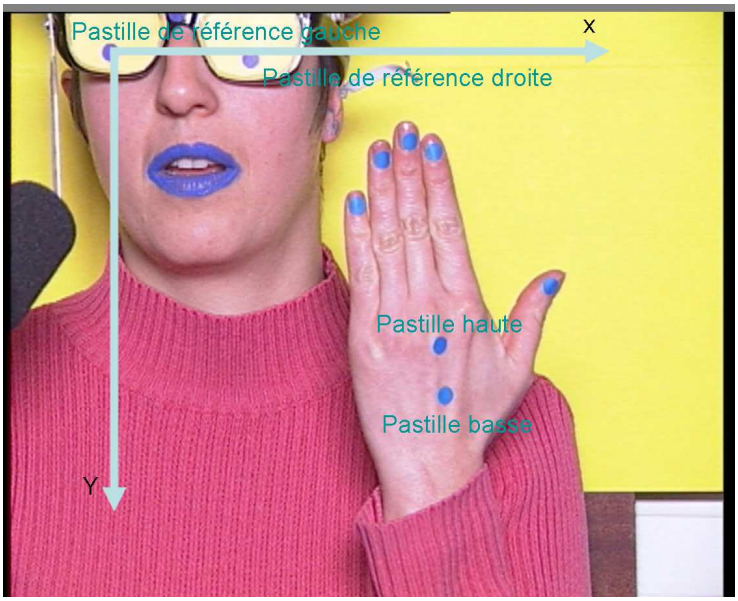

FIG. 5.2 – Image de la codeuse avec les pastilles de couleur sur la main et les axes servant de repère (enregistrement en plan large). Nous noterons le choix de la pastille sur la lunette gauche (**pastille de référence G**) comme origine de ce repère. Les axes sont (GX) pour les coordonnées horizontales (**x**) et (GY) pour les coordonnées verticales (**y**). Nous utilisons les appellations **pastille basse** (pastille vers le poignet de la main) et **pastille haute** (pastille vers les doigts) pour nommer les deux pastilles sur le dos de la main.

par ses *time-codes* de début et de fin. Un"*time-code*" est une chaîne de caractère contenant des informations temporelles qui permettent de repérer de manière unique une phrase dans une bande magnétique. La figure 5.3 montre un extrait du fichier qui contient les informations nécessaires pour la numérisation.

Après numérisation, les images des deux enregistrements (caméra champ large et caméra en mode zoom) sont stockées au format "bitmap 24" bits chacune avec un nom spécifique. Les noms des images d'une même phrase ont une racine de nom commune. Avant de commencer les traitements, ces images sont détramées en séparant les trames paires des trames impaires et ainsi récupérer une information toutes les 20 ms. En effet, si nous considérons, pour simplifier, une image comme étant une matrice (tri-dimensionnelle en espace RVB), le détramage se fait tout d'abord en séparant les lignes paires et impaires de cette matrice dans deux autres matrices de même dimension (et donc images). Les lignes vides des deux nouvelles matrices sont reconstituées par interpolation linéaire entre les lignes précédents et suivants. Ainsi, pour chaque image initiale nous obtenons deux trames avec les mêmes dimensions qui sont classées dans l'ordre impaire-paire.

Enfin, il est important de noter que nous enregistrons le signal audio pour obtenir une modalité de référence à chaque fois que nous traitons une des deux modalités principales du code LPC.

#tc1	tc2	sampling repertoire	prefixe	fichier audio	trames audio	comp0	yc1	incrust	X	Y			
00,01,41,12	00:01:45:09	D:\temp\sc_lev_plu001_1_0	sc_lev_plu001_1_0	sc_lev_plu001_1_0.wav	0	0	1	1	0	0			
00,01,47,19	00:01:50:10	D:\temp\sc_lev_plu002_1_0	sc_lev_plu002_1_0	sc_lev_plu002_1_0.wav	0	0	1	1	0	0			
00,01,50,24	00:01:53:02	D:\temp\sc_lev_plu002_2_0	sc_lev_plu002_2_0	sc_lev_plu002_2_0.wav	0	0	1	1	0	0			
00,01,54,09	00:01:56:18	D:\temp\sc_lev_plu003_1_0	sc_lev_plu003_1_0	sc_lev_plu003_1_0.wav	0	0	1	1	0	0			
00,01,57,11	00:02:00:09	D:\temp\sc_lev_plu003_2_0	sc_lev_plu003_2_0	sc_lev_plu003_2_0.wav	0	0	1	1	0	0			
00,02,01,17	00:02:04:00	D:\temp\sc_lev_plu004_1_0	sc_lev_plu004_1_0	sc_lev_plu004_1_0.wav	0	0	1	1	0	0			
00,02,05,10	00:02:07:14	D:\temp\sc_lev_plu004_2_0	sc_lev_plu004_2_0	sc_lev_plu004_2_0.wav	0	0	1	1	0	0			
00,02,09,07	00:02:11:18	D:\temp\sc_lev_plu005_1_0	sc_lev_plu005_1_0	sc_lev_plu005_1_0.wav	0	0	1	1	0	0			
00,02,12,19	00:02:15:06	D:\temp\sc_lev_plu005_2_0	sc_lev_plu005_2_0	sc_lev_plu005_2_0.wav	0	0	1	1	0	0			
00,02,16,13	00:02:19:05	D:\temp\sc_lev_plu006_1_0	sc_lev_plu006_1_0	sc_lev_plu006_1_0.wav	0	0	1	1	0	0			
00,02,19,14	00:02:22:10	D:\temp\sc_lev_plu006_2_0	sc_lev_plu006_2_0	sc_lev_plu006_2_0.wav	0	0	1	1	0	0			

FIG. 5.3 – Extrait d'un fichier contenant les informations nécessaires pour la numérisation. *tc1* et *tc2* sont les time-codes respectivement du début et de la fin de la séquence sous les formes suivantes : tc1 = heure,minute,seconde,N° image ; tc2 = heure :minute :seconde :N° image.

5.2.4 Traitements des données

Le but de ces traitements est d'obtenir des signaux cohérents contenant l'information précise permettant d'analyser les mouvements de la main et des lèvres. Pour les lèvres, nous recherchons à obtenir les contours des lèvres et par la suite calculer les paramètres géométriques. Pour la main, le but est de détecter les positions et les configurations de la main. Pour ceci, il faut au préalable extraire les coordonnées de toutes les pastilles placées sur la main. Enfin, le signal audio est traité directement pour obtenir un étiquetage des phonèmes prononcés.

Il est à noter que dans un premier temps, nous avons effectué les traitements de la main à partir des images de l'enregistrement en plan large et les traitements des lèvres à partir des images de l'enregistrement en mode zoom. Dans un second temps, nous nous sommes contenté seulement de l'enregistrement en plan large pour les deux types de traitements et ceci pour des raisons que nous développons par la suite.

5.2.4.1 Traitement des données manuelles

Seuillage numérique pour détecter le bleu : Avant de commencer l'extraction des coordonnées des pastilles, il faut trouver une méthode de seuillage qui permet de détecter le bleu des pastilles dans une image. Nous avons dans un premier temps utilisé une méthode simple fondée sur un seuillage appliqué sur les pixels de chaque composante couleur de l'espace RVB. Cette méthode a été utilisée auparavant par Attina (2005). Etant donné que les couleurs des pastilles sont choisies de façon à être différentes de la couleur de la peau de la main, cette méthode consiste plus précisément à fixer un seuil pour chacune des composantes rouge, vert et bleu des pixels de l'image candidate. En fait, ce seuil correspond aux valeurs maximales et minimales, apprises sur un échantillon d'images, que les pixels d'une des pastilles peuvent avoir selon la luminosité et l'orientation de la main.

Problèmes : Cette méthode ainsi définie est lourde à appliquer sur un corpus assez grand. En effet, il faut fixer un seuil pour chaque pastille et pour chaque composante couleur ; ce qui fait 21 seuils à fixer au total. De plus, nous avons remarqué que si l'éclairage change légèrement (l'intensité de l'éclairage peut varier légèrement ou la main change d'orientation par rapport à la source d'éclairage), ces seuils ne sont plus adaptés.

Notre solution : Pour pallier ces problèmes, nous avons développé une autre méthode qui consiste en trois étapes. Tout d'abord, nous calculons, pour une image donnée, la luminance[4]. Ensuite, nous soustrayons cette luminance de la composante bleu. Nous obtenons ainsi une image contenant le niveau de bleu pour chaque pixel de l'image initiale. En dernière étape, nous appliquons un seuil à cette nouvelle image pour ainsi obtenir une image bi-chromatique (chaque pixel ayant une valeur supérieure à ce seuil se fait attribuer la valeur 255 sinon 0). Le seuil appliqué est fixé en tenant compte de l'intensité du bleu des pastilles. La figure 5.4 illustre les trois étapes de cette méthode appliqué à une image.

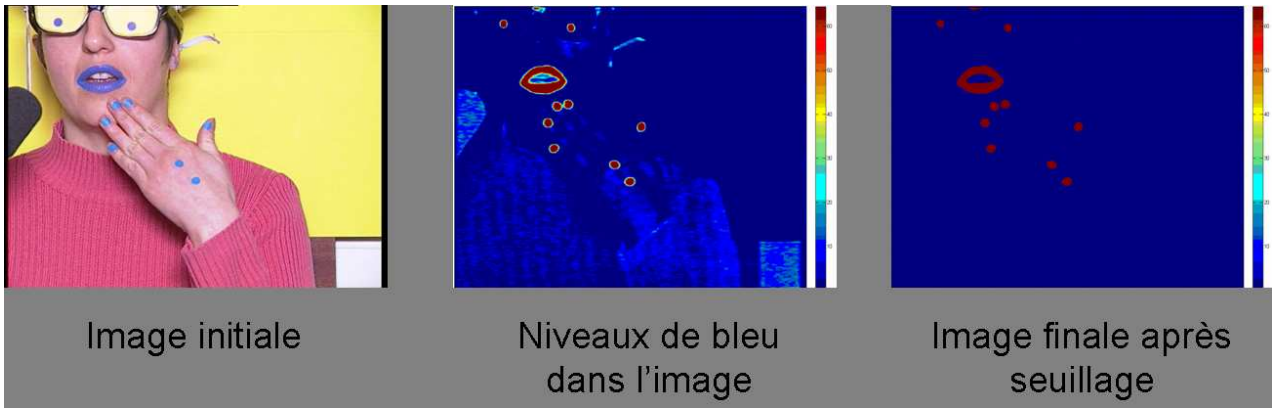

FIG. 5.4 – Exemple de détection des objets en bleu dans une image.

Suivi des pastilles : Après avoir détecté les objets en bleu dans une image, il s'agit de calculer les coordonnées de chaque pastille. Pour ceci, nous avons utilisé une méthode, héritée du travail d'Attina (2005), et qui consiste en un suivi des pastilles tout au long de la séquence d'images. Pour une séquence d'images (bi-chromatiques) donnée, le suivi d'une pastille peut se décrire de la façon suivante :

- A partir d'un point de départ sur la pastille, une fenêtre de traitement contenant la pastille est délimitée.
- La fenêtre est parcourue et tous les pixels rencontrés qui ont la valeur 255 sont stockés et considérés comme les pixels de la pastille.

[4] *En traitement des images, la luminance est l'image en noir et blanc, et souvent appelée "en niveaux de gris". Elle est définie en général comme une somme pondérée des trois composantes rouge, vert et bleu. Elle peut s'écrire formellement à partir des trois composantes rouge, vert et bleu de la façon suivante : luminance* $= 0,299rouge + 0,587vert + 0,114bleu$.

- A partir de l'ensemble de ces pixels, un contour est déterminé et ensuite les coordonnées x et y du centre de gravité de la pastille sont calculées. Ce centre est stocké dans un vecteur résultat.
- Le suivi continue sur l'image suivante en considérant automatiquement comme point de départ le centre de gravité de la pastille de l'image précédente.

De cette manière, les coordonnées des centres de gravité des pastilles de référence sur les lunettes et des deux pastilles sur le dos de la main (pastilles de position) sont déterminées. Le suivi peut ainsi s'effectuer automatiquement pour toutes les images d'une séquence. Le résultat final du suivi donne pour toutes les séquences les coordonnées x et y en pixels dans l'image des quatre pastilles toujours visibles sur l'image ; c'est-à-dire les pastilles sur le dos de la main et sur les lunettes. Aux coordonnées des deux pastilles du dos de la main sont retranchées les coordonnées de la pastille de référence sur les lunettes et les coordonnées relatives ainsi obtenues sont alors transformées en centimètre (cm) en utilisant la conversion pixel vers centimètre.

Inconvénients : Cette méthode a toutefois deux inconvénients. En effet, il est indispensable, pour chaque séquence, de déterminer manuellement le point de départ du suivi sur la première image de la séquence. De plus, cette méthode ne nous permet de suivre que les quatre pastilles toujours visibles sur l'image. Les pastilles du bout des doigts ne peuvent être suivie par cette méthode puisque selon la configuration de la main, elles peuvent être cachées et donc invisibles sur l'image.

Notre solution : Une alternative à cette méthode est de s'appuyer sur un algorithme de marquage de composantes connexes[5]. La technique de marquage de composantes connexes est fondée sur une détection des contours des composantes. Suite à cette détection des contours, les pixels d'une composante donnée sont marqués par un même label suivant leur relation de connexité (4 ou 8 connexe). Pour marquer nos pastilles, nous avons utilisé la fonction de marquage de composantes connexes ***bwlabel()*** disponible dans la bibliothèque image de *MATLAB*. Avec cette fonction, les pixels correspondant à une pastille dans une image sont labelisés avec une même valeur (un entier). Donc chaque pastille correspond à un label (une valeur). Etant donné les pixels de chaque pastille, il reste maintenant à calculer les coordonnées du centre de gravité et à poursuivre le calcul de la même façon que ce qui est décrit ci-dessus. Nous obtenons finalement pour chaque séquence, des signaux traçant les trajectoires en x et en y de toutes les pastilles.

Le seul problème qui se pose avec cette méthode est de déterminer quelle pastille correspond à un centre de gravité calculé. Pour les pastilles de référence le problème peut se résoudre facilement puisque les deux pastilles ont pratiquement les mêmes coordonnées dans toutes les images. En revanche, pour les autres pastilles la tâche est un peu plus compliquée. Tout d'abord, la pastille basse est considérée comme la pastille qui a la coordonnée y la plus grande (la pastille basse est toujours la pastille la plus en bas dans l'image). A partir de cette pastille, une mesure de distance permet de localiser la pastille haute. Les pastilles restantes sont bien évidemment les pastilles sur les doigts. Dans la suite, nous aurons besoin seulement d'identifier précisément

[5] *La composante connexe en traitement d'image représente une agrégation de pixels connectés sur une image.*

certains doigts et non pas tous. En effet, nous aurons besoin de déterminer la pastille du doigt directeur[6]. En code LPC, ce doigt est en général soit le majeur soit l'index selon la configuration de la main. De ce fait, pour identifier le doigt, directeur nous déterminons d'abord l'axe de la main (c'est la droite qui passe par les deux pastilles sur le dos de la main) et nous y projetons toutes les pastilles des doigts. La pastille qui a la coordonnée la plus élevée sur cet axe est directement considérée comme la pastille du doigt directeur.

Pour illustrer les résultats que nous obtenons de ce traitement, nous traçons sur la figure 5.5 les coordonnées x et y des pastilles haute et basse pour une séquence. De même la figure 5.6 montre la trajectoire du doigt directeur.

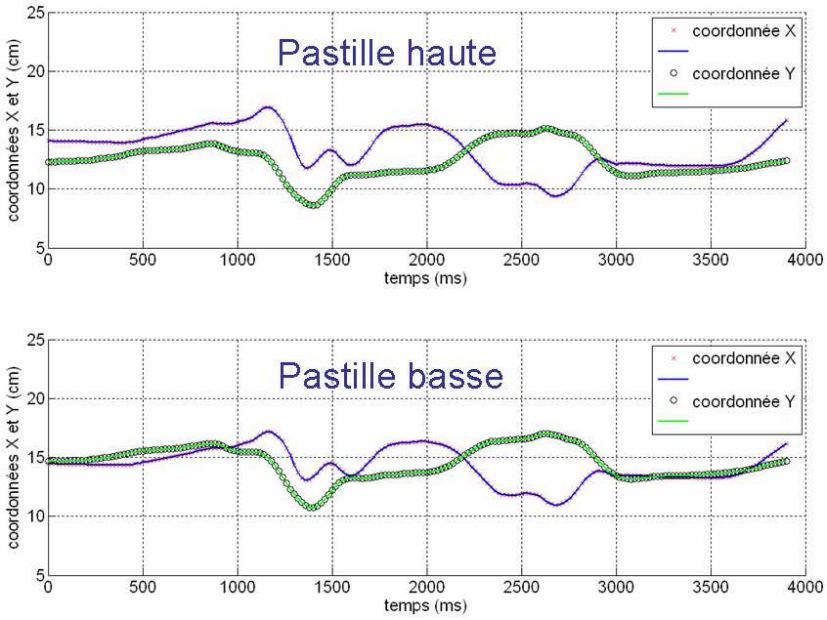

FIG. 5.5 – Exemple de signaux représentant les coordonnées x et y des pastilles haute et basse pour la séquence :"ma chemise est roussie".

5.2.4.2 Traitement des données labiales

Méthode de traitement initiale : Dans un premier temps, nous avons commencé à traiter nos images des lèvres provenant de l'enregistrement en mode zoom à l'aide du logiciel TACLE (Traitement Automatique du Contour des LEvres, un logiciel développé dans notre laboratoire) pour suivre automatiquement les contours des lèvres. Ce logiciel permet d'extraire, à partir d'une séquence d'images numérisées, des paramètres géométriques descripteurs des lèvres vues de face. Ces paramètres, calculés sur les contours externe et interne des lèvres, sont : l'étirement des lèvres[7] (on le note A pour le contour interne et A' pour le contour externe), l'aperture[8] (on le note B pour le contour interne et B' pour le contour externe) et l'aire interolabiale (on le note S pour le contour interne et S' pour le contour externe). Ce logiciel contient un module

[6] *Le doigt directeur est le doigt qui pointe la position*
[7] *La séparation horizontale entre les lèvres.*
[8] *La séparation verticale entre les lèvres.*

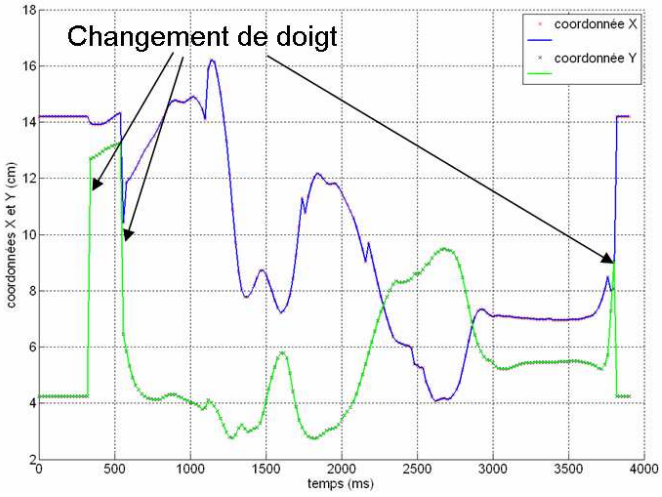

FIG. 5.6 – Exemple de signaux représentant les coordonnées du doigt directeur pour la séquence :"ma chemise est roussie".

qui permet de passer des images numérisées (deux images successives sont séparées de 40ms) aux trames paires et impaires pour reconstituer deux images séparées de 20 ms. Le principe de cette unité de "*détramage*" est identique à ce que nous avons décrit auparavant pour les images en plan large.

Le logiciel TACLE nécessite en entrée un fichier "report" où sont listés les time-codes et les images associées, et génère en sortie un fichier dont chaque ligne correspond à un time-code (et donc une trame) et qui contient les différents paramètres calculés[9] (voir figure 5.7). Il fonctionne en plusieurs étapes que nous synthétisons de la façon suivante :

1. Initialisation : une fenêtre de traitement est déterminée manuellement pour détecter le contour des lèvres. Cette fenêtre est associée à des valeurs de teinte et de luminance (codage en TSL[10]) pour le seuillage numérique (ou chroma-key). Ces valeurs peuvent entre autre être modifiées. Dans la fenêtre déterminée il faut indiquer un point de départ par un clic de la sourie. Ce point sert comme point de départ pour la recherche automatique des contours.
2. Seuillage numérique des images :les valeurs de teinte et de luminance fixées au préalable sont utilisées comme des seuils qui sont appliqués afin de localiser les zones maquillées en bleu et ensuite de les noircir.
3. Filtrage : Les zones noires obtenues sont filtrées par un filtre médian pour éliminer des éventuels pixels noirs isolés, ce qui permettrait une bonne détection des contours par la suite.
4. Détection des contours : les contours (interne et externe) des masses noires des lèvres

[9] *Le logiciel TACLE peut aussi extraire des paramètres qui décrivent des lèvres vues de profil. Nous nous contentons de décrire seulement les paramètres de vue de face.*

[10] *Les images sont codées dans l'espace TSL (Teinte, Saturation et Luminance) et non pas dans l'espace RVB essentiellement pour des raisons d'ergonomie au niveau de l'interface utilisateur. Le codage TSL est en effet plus proche de la perception "naturelle" des couleurs par l'humain.*

sont ensuite déterminés automatiquement par un algorithme utilisant une exploration à 8 voisins dans le sens trigonométrique. Les contours détectés sont sauvegardés. Pour chaque contour, une liste des coordonnées x et y de ces points ainsi qu'un mot clé déterminant le type du contour sont obtenus.

5. Extraction des paramètres : cette étape consiste à calculer les paramètres labiaux à partir des points constituant les contours interne et externe. Le calcul repose sur les équations établies par Lallouache (1991). Les paramètres S et S' sont calculés directement à partir des points du contour. En revanche, pour calculer les autres paramètres (A, B, A', B') des contours, on se sert des formules donnant, pour chaque contour caractérisé par ses points (X_k, Y_k), l'aire S (ou S' pour le contour externe), le barycentre (X_G, Y_G), et l'axe principal d'inertie. Ces formules sont décrites par :

$$S = \tfrac{1}{2}\sum_k (X_k Y_{k+1} - X_{k+1} Y_k)$$

$$X_G = \tfrac{1}{2}\sum_k \frac{X_k^2 (Y_{k+1} - Y_k)}{S}$$

$$Y_G = \tfrac{1}{2}\sum_k \frac{Y_k^2 (X_{k+1} - X_k)}{S}$$

L'axe principal d'inertie qui passe par le barycentre G peut être déterminé par son angle α. Pour calculer cet angle, nous calculons les moments d'inertie J_x, J_y et J_{xy} avec les formules :

$$J_x = \tfrac{1}{3}\sum_k Y_k^3 (X_{k+1} - X_k) - Y_G^2 S$$

$$J_x = \tfrac{1}{3}\sum_k Y_k^3 (X_{k+1} - X_k) - Y_G^2 S$$

$$J_{xy} = \tfrac{1}{3}\sum_k X_k^2 Y_k (Y_{k+1} - Y_k) - X_G^2 Y_G^2 S$$

Pour calculer A, nous considérons les points $A1$ et $A2$ définis comme étant les extrema en X du contour interne des lèvres. Le paramètre A est simplement la distance entre $A1$ et $A2$. Même chose pour $A'1$, $A'2$ et A'.

Pour calculer B, nous nous servons de deux points $B1$ et $B2$ qui sont déterminés en cherchant d'abord l'axe principal d'inertie du contour interne des lèvres. Ensuite, les intersections de cet axe avec le contour sont considérées comme les points $B1$ et $B2$ (nous trouvons forcément deux points, un entre la partie inférieure du contour et l'autre pour la partie supérieure du contour). En pratique, les intersections sont déterminées d'abord en prenant seulement les points du contour les plus proches de l'axe et en appliquant des approximations paraboliques au voisinage de ces points. Les points $B1$ et $B2$ sont les intersections des paraboles avec l'axe d'inertie. Enfin, B est la distance entre $B1$ et $B2$. Le calcul de B' est identique à celui du contour externe des lèvres.

Nous obtenons enfin pour chaque séquence après le traitement avec le logiciel TACLE, les valeurs des paramètres A, A', B et B' en cm ainsi que celles de S et S' en cm^2 toutes les 20 ms. La figure 5.8 montre le décours temporel de ces 6 paramètres calculés à partir d'un exemple de séquence.

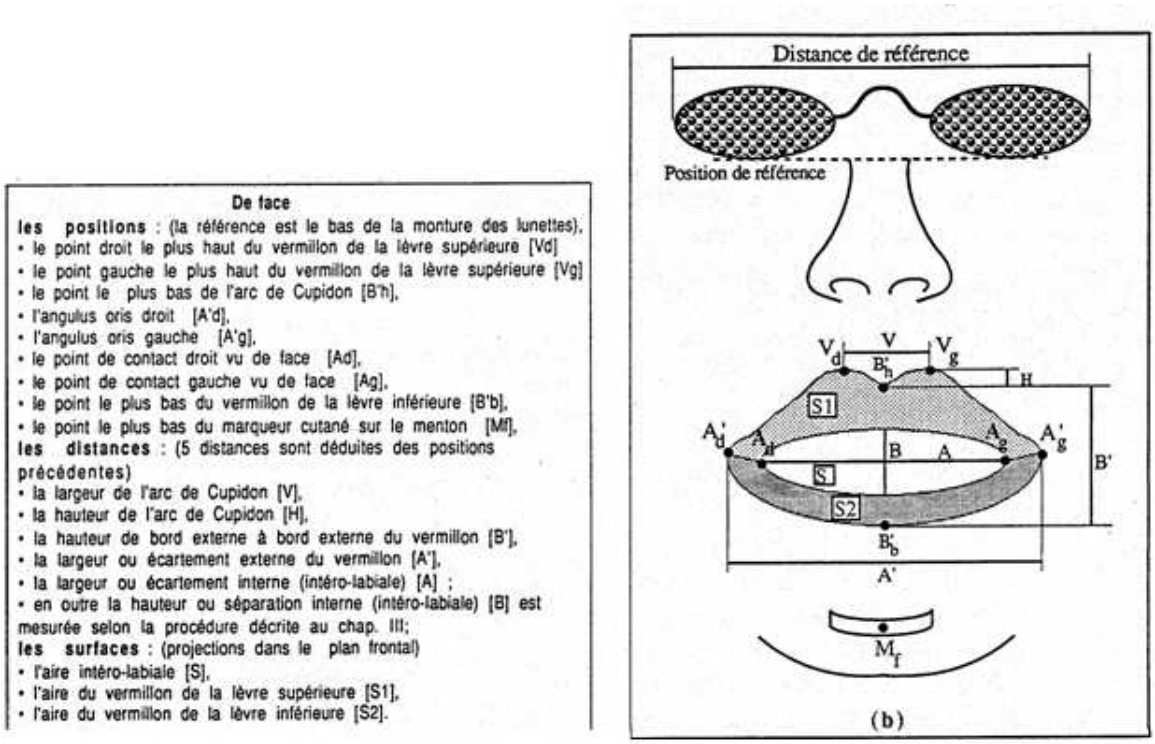

FIG. 5.7 – Paramètres descripteurs des lèvres (d'après Lallouache (1990)).

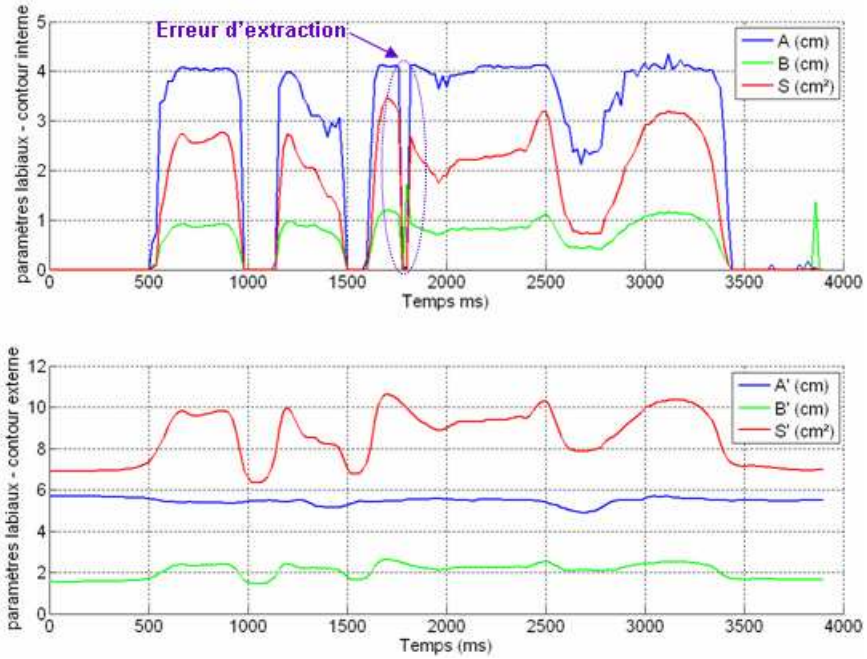

FIG. 5.8 – Exemple de décours temporel des 6 paramètres labiaux A, A', B et B' en cm ainsi que S et S' en cm^2 calculés à l'aide du logiciel TACLE à partir de la séquence :"ma chemise est roussie". Les erreurs d'extraction des paramètres sont par ailleurs marquées.

Problèmes : Les principes implémentés dans le logiciel TACLE ont été l'outil d'extraction des paramètres labiaux de plusieurs travaux (Adjoudani, 1998; Attina, 2005) où ils ont montré leur efficacité spécialement sur des corpus très réduits. Dans le cas d'un corpus d'une taille moyenne ou grande, il se peut que les paramètres d'initialisation (taille de la fenêtre, les valeurs de la teinte et de la luminance) ne soient pas bonnes pour toutes les séquences du corpus. En effet, nous avons remarqué, en appliquant le traitement de TACLE sur notre corpus, de nombreuses erreurs dans les signaux des paramètres labiaux (souvent plusieurs dans une même séquence) (voir figure 5.8 par exemple). Ces erreurs sont provoquées essentiellement par de mauvaises détection des contours des lèvres pour deux raisons. Premièrement, la bouche peut sortir, souvent partiellement, de la fenêtre d'analyse initiale au cours du traitement de tout le corpus[11]. Deuxièmement, les valeurs de la teinte et de la luminance fixées sont sensibles aux variations même légères de l'éclairage. En utilisant toujours ce logiciel, la solution est de corriger ces erreurs en initiant pour chaque séquence ses propres paramètres d'initialisation. Vu la taille de notre corpus (638 phrases), cette solution paraît trop lourde à effectuer et enlève de plus le caractère automatique du traitement.

Notre solution : La solution que nous proposons est de remplacer les 4 premières étapes du traitement TACLE par un traitement identique à celui que nous avons décrit ci-dessus pour détecter les pastilles en bleu. A partir des contours labiaux obtenus par ce traitement, l'étape de calcul des paramètres décrite ci-dessus est alors effectuée de la même manière qu'avec le traitement TACLE. De plus, ce traitement peut être appliqué soit sur des images en mode zoom soit en plan large. Dans ce dernier cas, le seule souci est de séparer la composante connexe des lèvres de celles des pastilles. Pour ceci, nous avons considéré comme lèvres toute composante connexe dont le nombre des pixels est le plus grand. En d'autres termes, la tâche bleue la plus grande dans l'image bi-chromatique est celle des lèvres. Notre solution tolère en plus un déplacement de la tête puisque le traitement ne nécessite pas de délimiter une fenêtre initiale fixe. Pour illustrer les performances de la méthode, nous présentons dans la figure 5.9 les signaux temporels des paramètres labiaux de la même séquence que celle de la figure 5.8.

Evaluation de la solution : La solution que nous proposons pour extraire les paramètres labiaux nécessite d'évaluer sa précision pour pouvoir la considérer dans la suite. Pour ceci, nous avons sélectionné un échantillon de 70 images représentant différentes formes aux lèvres. A partir de ces images, nous avons extrait le contour interne (le contour externe ne semble pas pouvoir nous donner une bonne précision) des lèvres d'abord automatiquement en utilisant notre méthode sur les images provenant de l'enregistrement en plan large et ensuite manuellement à partir des images correspondantes de l'enregistrement en mode zoom (pour avoir une résolution augmentée des lèvres et pour permettre ensuite de valider le choix de travailler en plan large). Pour pouvoir comparer les contours extraits avec les deux méthodes, nous avons calculé les paramètres labiaux A, B et S. Ensuite, nous avons calculé les écarts absolu et arithmétique entre ces paramètres pour chaque image de la façon suivante :

[11] *le sujet peut bouger sa tête légèrement et si la fenêtre n'est pas de taille assez grande, des parties de la bouche peuvent se retrouver à l'extérieur de la fenêtre*

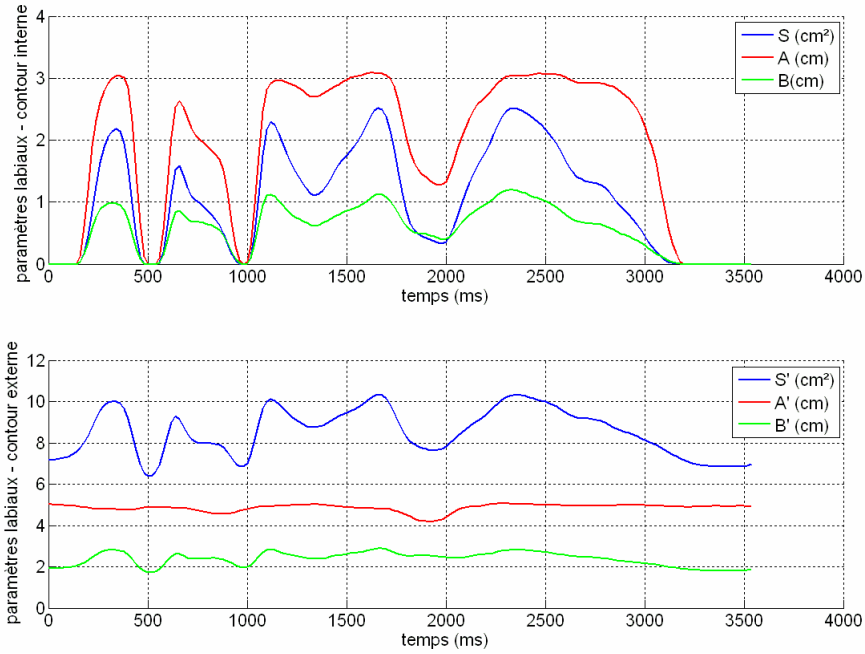

FIG. 5.9 – Exemple de décours temporel des 6 paramètres labiaux A, A', B et B' en cm ainsi que S et S' en cm^2 calculés à l'aide de la solution que nous proposons appliquée à la séquence :"ma chemise est roussie".

$$Ecart\_arithmetique =$$
$$\left[A_{automatique} - A_{manuelle} \quad B_{automatique} - B_{manuelle} \quad S_{automatique} - S_{manuelle}\right]$$

$$Ecart\_absolu =$$
$$\left[|A_{automatique} - A_{manuelle}| \quad |B_{automatique} - B_{manuelle}| \quad |S_{automatique} - S_{manuelle}|\right]$$

Sur l'ensemble des images de l'échantillon, nous avons calculé les moyennes et les écarts types de ces écarts. La table 5.1 présente les valeurs obtenues.

	A en cm	B en cm	S en cm^2
Ecart arithmétique : moyenne	0,1593	0,0822	0,1501
Ecart arithmétique : écart type	0,2326	0,0728	0,1919
Ecart absolu : moyenne	0,2115	0,0831	0,1686
Ecart absolu : écart type	0,1857	0,0717	0,1756

TAB. 5.1 – Moyennes et écarts types des écarts arithmétique et absolu des valeurs des paramètres labiaux A, B et S obtenus automatiquement par notre méthode et manuellement.

Avec seulement ces valeurs d'écarts entre les paramètres mesurés manuellement et automatiquement, et même s'elles sont en valeur numérique faibles, nous ne pouvons conclure sur la précision de notre solution. Il faut en effet comparer ces écarts avec ceux obtenus avec une autre méthode largement utilisée dans la littérature. Bien évidement, nous disposons de TACLE

comme méthode qui a bien sûr de bonnes performances quand les paramètres d'initialisation sont bien choisis pour chaque image. Ainsi, pour chaque image de notre échantillon de test nous avons bien réglé les paramètres d'initialisation pour un traitement avec TACLE. Les écarts entre les paramètres labiaux extraits du contour interne des lèvres par TACLE et ceux manuellement sont calculés de la même façon que précédemment et sont présentés dans la table 5.2. Il est à noter que TACLE a été appliqué sur les images zoomées.

	A en cm	B en cm	S en cm^2
Ecart arithmétique : moyenne	0,2809	0,0416	0,1480
Ecart arithmétique : écart type	0,2463	0,0480	0,1997
Ecart absolu : moyenne	0,2820	0,0438	0,1492
Ecart absolu : écart type	0,2449	0,0460	0,1988

TAB. 5.2 – Moyennes et écarts types des écarts arithmétique et absolu des valeurs des paramètres labiaux A, B et S obtenus par TACLE et manuellement.

Si nous comparons les deux tables, nous pouvons remarquer que notre méthode est plus précise dans la mesure du paramètre A. Certes, TACLE semble bien mesurer le paramètre B, mais les valeurs de ces écarts sont faibles. Le paramètre S semble être mesuré avec presque la même précision avec les deux méthodes.

En résumé, notre méthode permet d'obtenir les paramètres labiaux avec une précision aussi bonne qu'avec TACLE appliqué avec une bonne initialisation. Elle peut être appliquée sur des images en plan large (résolution réduite) et donc diminuer le nombre de caméras. De plus, le traitement avec cette méthode est automatique. Il suffit en effet de fixer un seul seuil pour binariser les images de niveaux de bleu. Ainsi, avec une précision presque identique que TACLE, notre méthode est moins contraignante.

5.2.4.3 Etiquetage de l'audio

Le signal acoustique a été automatiquement étiqueté au niveau phonétique en utilisant les outils d'alignement du Laboratoire LIG[12](une description un peu plus détaillée du système de reconnaissance automatique de la parole peut notamment être trouvée dans Lamy *et al.* (2004)). En effet, la transcription de chaque phrase prononcée par le codeur étant connue, un dictionnaire de prononciation a été utilisé pour produire la séquence de phonèmes correspondant à chaque signal. Cette séquence est ensuite alignée avec le signal en utilisant des modèles acoustiques HMM du Français appris sur la base BRAF100 (Vaufreydaz *et al.*, 2000). A l'issue de cette étape, un étiquetage phonétique temporel du signal acoustique est disponible, pouvant comporter un certain nombre d'erreurs dues au dictionnaire de prononciation ou à l'alignement automatique. Nous pouvons voir un exemple d'étiquetage d'une séquence donnée sur la figure 5.10.

[12] *Laboratoire d'informatique de Grenoble*

5.2.4.4 Bilan des données extraites

L'ensemble des traitements a conduit à un ensemble cohérent de signaux : les coordonnées x et y des centres des pastilles haute et basse placées sur le dos de la main ainsi celles du doigt directeur, toutes les 20 ms, les valeurs des paramètres labiaux extraits des contours internes et externes, également toutes les 20 ms, ainsi que la réalisation acoustique du signal correspondant accompagnée de sa segmentation et de son étiquetage phonétique, corrigé manuellement (voir figure 5.10 pour des exemples de ces signaux). Il est à noter que tous les signaux (à l'exception de l'information extraite du signal audio) ont été lissés par des filtres passe-bas.

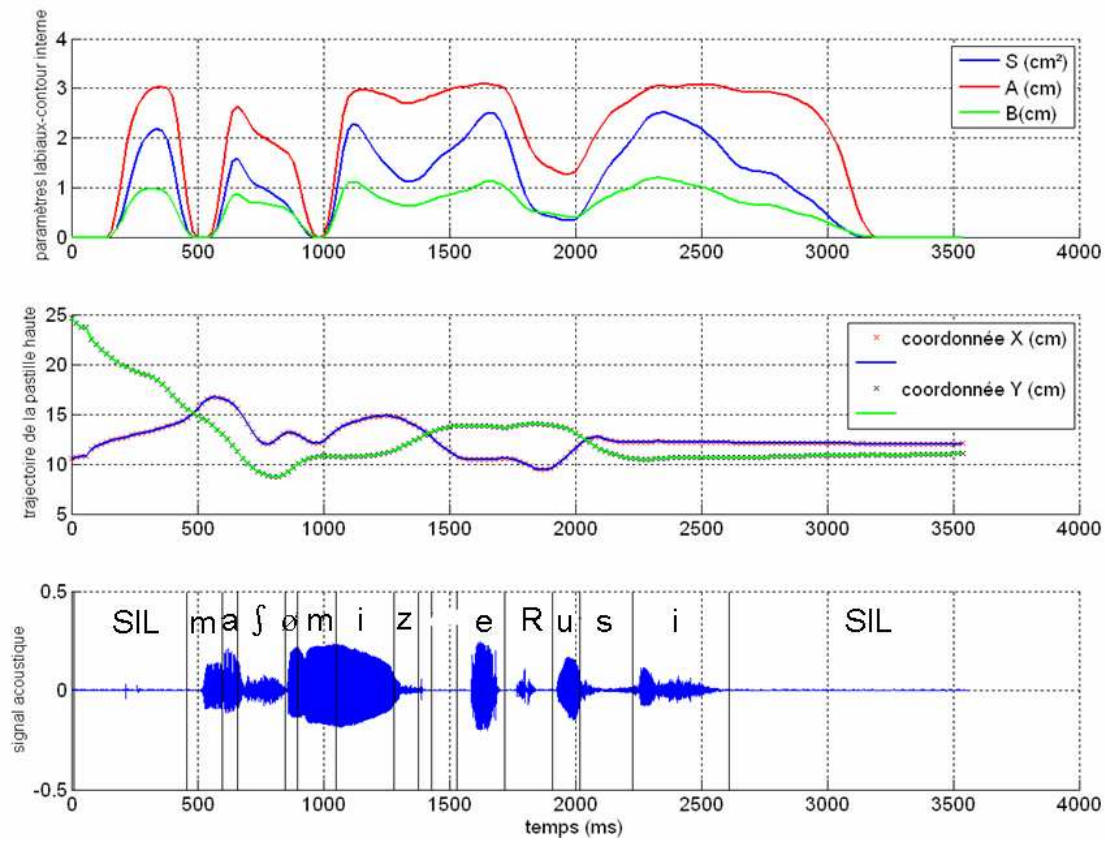

FIG. 5.10 – De haut en bas : les paramètres du contour interne des lèvres (A, B et S), les coordonnées x et y de la pastille haute, la réalisation acoustique.

Chapitre 6

Codage de la main : détection et classification

L'identification de la position LPC et de la configuration digitale de la main constitue une étape dans le processus de fusion avec les formes labiales. D'autre part, la segmentation temporelle du flux de la position de la main LPC est aussi un point important, compte tenu de la désynchronisation naturelle entre les flux manuel et labial (comme nous l'avons évoqué dans la partie *état de l'art*, (Attina *et al.*, 2004)). Cette partie présente tout d'abord un algorithme de segmentation temporelle automatique de la position LPC. Cet algorithme permet dans un premier temps de détecter les positions LPC à chaque instant et dans un second temps de déterminer les limites entre les positions cibles et les transitions. Ensuite, une évaluation de la méthode d'identification de la position cible de la main en LPC est discutée. Enfin, la forme de la main est reconnue grâce à un second algorithme dont les performances sont évaluées aussi.

6.1 Segmentation temporelle automatique de la position LPC

6.1.1 Description de la méthode

Si nous observons les trajectoires x et y des pastilles sur le dos de la main (pastille haute ou basse), nous pouvons remarquer qu'elles présentent des transitions plus ou moins rapides entre valeurs extrêmes, de durée plus ou moins longues, caractéristiques des cibles du code LPC. La segmentation temporelle consiste à déterminer automatiquement les limites entre positions cibles et transitions, c'est-à-dire déterminer les instants de début de transition, d'atteinte de la position LPC et de fin de tenue de cette position.

6.1.1.1 Affectation des numéros de position LPC

Cette première étape consiste à affecter à chaque couple de coordonnées x et y, c'est-à-dire toutes les 20 ms, un numéro de position de main parmi les cinq du code LPC. La méthode s'appuie sur le maximum de vraisemblance selon une modélisation gaussienne des coordonnées x et y des pastilles de la main et des doigts. Cette classification a été choisie pour sa simplicité et notamment du fait de l'homogénéité des dispersions des positions. En effet, nous supposons

que les distributions des cinq positions dans le repère (GXY)(voir chapitre précédent) suivent une loi normale à deux dimensions définie par leur densité de probabilité :

$$P_i(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \exp^{-\frac{1}{2}\left[\left(\frac{x-m_x}{\sigma_x}\right)^2+\left(\frac{y-m_y}{\sigma_y}\right)^2-2r\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y}\right]}$$

où m_x, m_y et r représentent respectivement les moyennes et le coefficient de corrélation des coordonnées x et y (coordonnées de la pastille haute du dos de la main ou pastille du doigt directeur) associées à la position i et où σ_x et σ_y sont leurs écarts types. $P_i(x,y)$ représente donc la densité de probabilité que le couple (x,y) appartienne à la position i.

Chacune des cinq positions a finalement été modélisée par deux gaussiennes bidimensionnelles construites à partir d'un dictionnaire de 30 images cibles sélectionnées par un expert. La première modélise la position 2D de la pastille haute du dos de la main (voir figure 6.1), la seconde modélise celle de la pastille placée à l'extrémité du doigt directeur (voir figure 6.2). Ce second modèle est utilisé pour pondérer le premier afin d'améliorer la robustesse de la méthode de classification. En effet, les coordonnées de la pastille haute ne sont pas suffisantes étant donné que la main est susceptible d'effectuer des rotations ; ainsi, une même position peut se réaliser de plusieurs manières différentes. Les coordonnées de la pastille haute pour la position "bouche", par exemple, sont susceptibles d'être les mêmes que celles de la position "pommette" si la direction formée par la pastille haute et celle du doigt directeur est horizontale (voir figure 6.3).

Pour la classification d'une image, les coordonnées x-y de ces deux pastilles sont donc considérées. Ainsi, à chacun des 2 couples de coordonnées x-y (pastilles haute et de doigt directeur) est associé un vecteur de cinq valeurs de densité de probabilité. Le produit scalaire de ces deux vecteurs fournit un vecteur de cinq composantes contenant chacune le résultat de la pondération du premier par le second modèle. La plus grande composante (celle dont la valeur maximise la vraisemblance) définit ainsi le numéro de la position de la main (entre 1 et 5). Le résultat de cette première étape de classification conduit à affecter à chaque image un numéro de position cible. Pour une séquence phonétique, le résultat donne une suite de positions cibles numérotées toutes les 20 ms pour former des plateaux cibles (voir figure 6.4 pour un exemple de résultat).

Nous pouvons résumer cette première étape par le schéma de la figure 6.5.

A ce niveau de classification, il n'est pas possible de définir des transitions entre les positions cibles de la main du fait que la méthode du maximum de vraisemblance fournit toujours une solution, même si celle-ci peut être peu probable.

6.1.1.2 Les limites des plateaux cibles

Une seconde étape a donc consisté à filtrer les cibles potentielles par application d'un critère non linéaire, afin d'affiner la taille des plateaux cibles. Le critère est un seuil ajouté au minimum de vitesse de déplacement de la pastille de référence du dos de la main, afin de définir l'intervalle de ralentissement caractéristique de l'atteinte d'une position cible. La vitesse $v(t)$ au point de coordonnée $(x(t),\ y(t))$ du centre de gravité de la pastille haute a été définie comme étant la distance euclidienne entre les deux points $(x(t),\ y(t))$ et $(x(t+\Delta),\ y(t+\Delta))$ successifs ramenée à l'espacement temporel Δ de 20 ms :

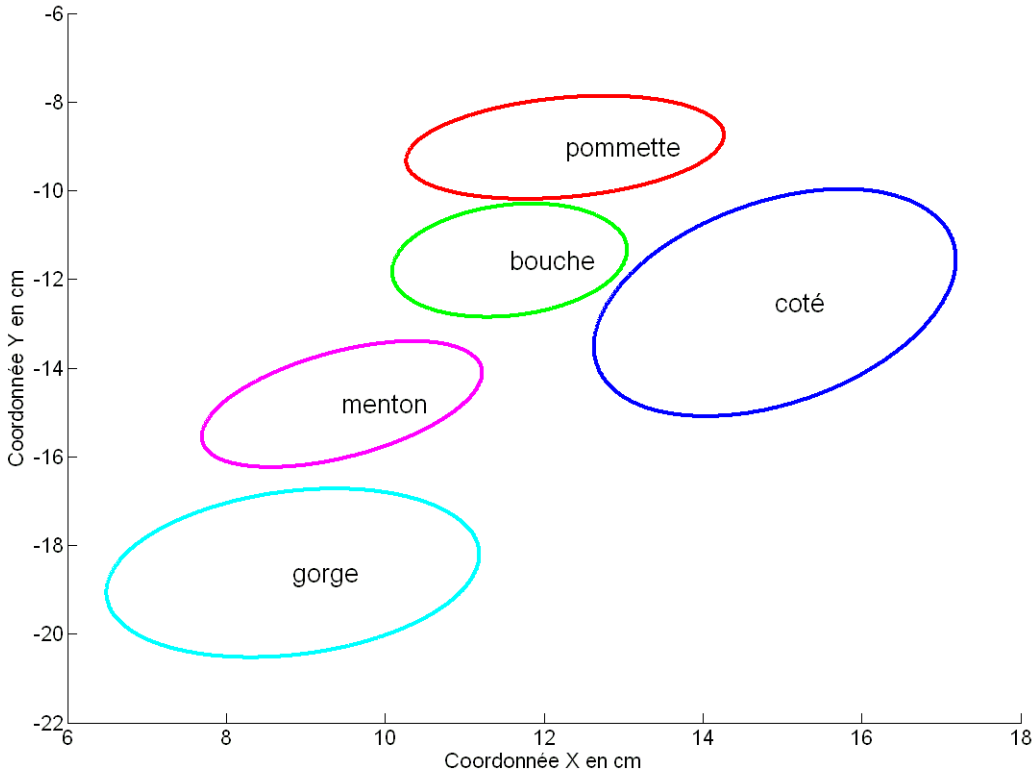

FIG. 6.1 – Ellipses de dispersion d'ordre 2 (écart type égal à 2) des différentes positions tracées à partir des coordonnées x et y de la pastille haute pour les données issues de l'échantillon d'apprentissage.

$$v(t) = 50 * \sqrt{(x(t+\Delta) - x(t))^2 + (y(t+\Delta) - y(t))^2}$$

Avant d'affiner les plateaux cibles, nous avons remarqué que certains plateaux ne contiennent pas de minimum local de vitesse (voir figure 6.6). Et donc la main ne ralentit pas dans ces plateaux, ce qui nous permet de considérer que ces plateaux ne représentent pas des positions cibles. Par conséquent, ces plateaux sont simplement éliminés. Pour le traitement d'un plateau donné de positions cibles identiques, l'instant de vitesse minimum est repéré. Afin de prendre en compte la rapidité variable de déplacement de la main dans les transitions qui influe la valeur du minimum de vitesse, le contraste de vitesse entre le pic de vitesse précédent (recherché dans l'intervalle défini par le milieu du plateau précédent et celui du plateau considéré) et le minimum de vitesse considéré est utilisé (voir figure 6.7). Un pourcentage de 40 % (fixé de manière empirique[1]) de ce contraste est retenu et ajouté au minimum de vitesse, ce qui définit une valeur seuil de la

[1] *Si le seuil a été choisi empiriquement, celà tient au fait que toutes les personnes ont des manières de coder sensiblement différentes (au niveau de la vitesse de déplacement de la main). Ainsi, le seuil appliqué sur le contraste est fonction de chaque codeur ; si la personne code vite, le contraste sera faible, le seuil, donc, le sera encore plus, de telle sorte que certaines cibles manuelles auront des durées qui seront plus faibles que leur durée réelle ; dans ce cas, il s'agirait d'augmenter le seuil.*

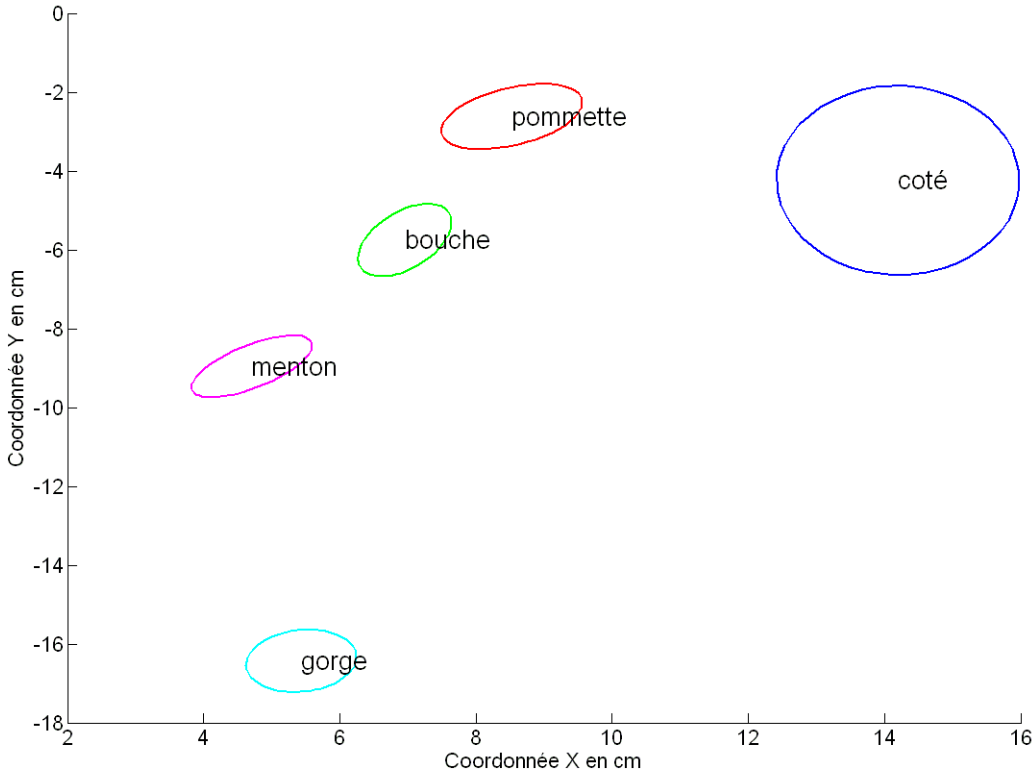

FIG. 6.2 – Ellipses de dispersion d'ordre 2 (écart type égal à 2) des différentes positions tracées à partir des coordonnées x et y de la pastille du doigt directeur pour les données issues de l'échantillon d'apprentissage..

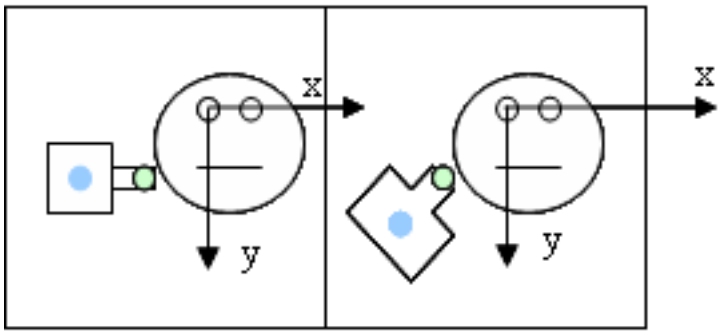

FIG. 6.3 – Pour la même position, les coordonnées de la pastille haute (en bleu), suite à sa rotation, peuvent beaucoup varier, d'où l'intérêt d'utiliser, en plus, les coordonnées du doigt directeur (en vert).

vitesse notée Vs au dessus de laquelle les points $(x(t), y(t))$ du plateau sont exclus de la position cible et considérés dans la transition. Inversement, les points du plateau dont la vitesse $v(t)$ est

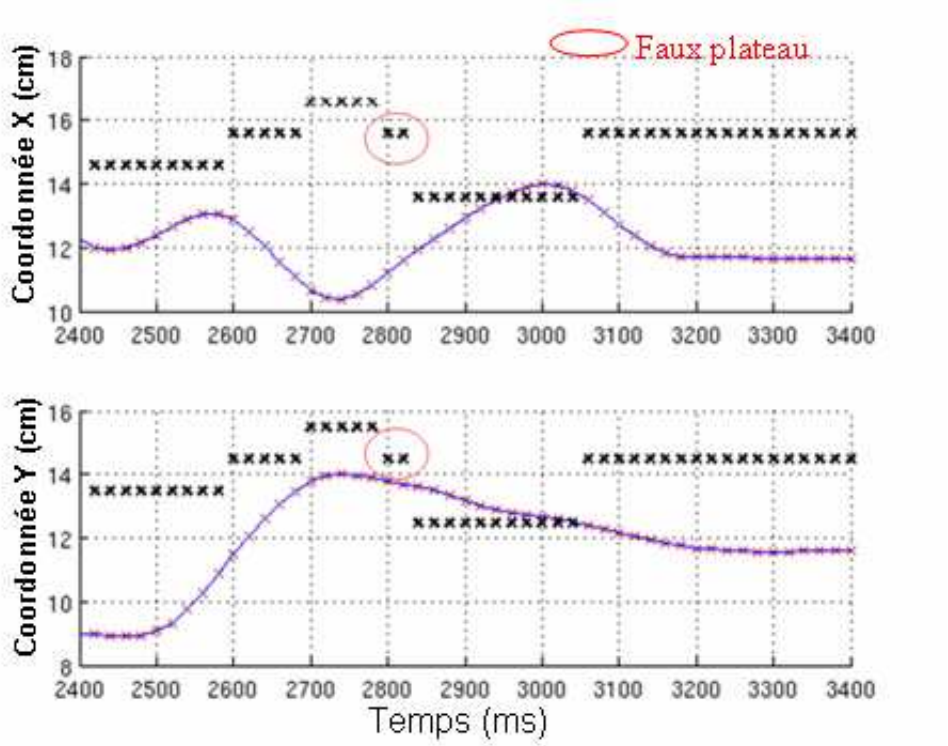

FIG. 6.4 – Exemple de plateaux cibles détectés par le système.

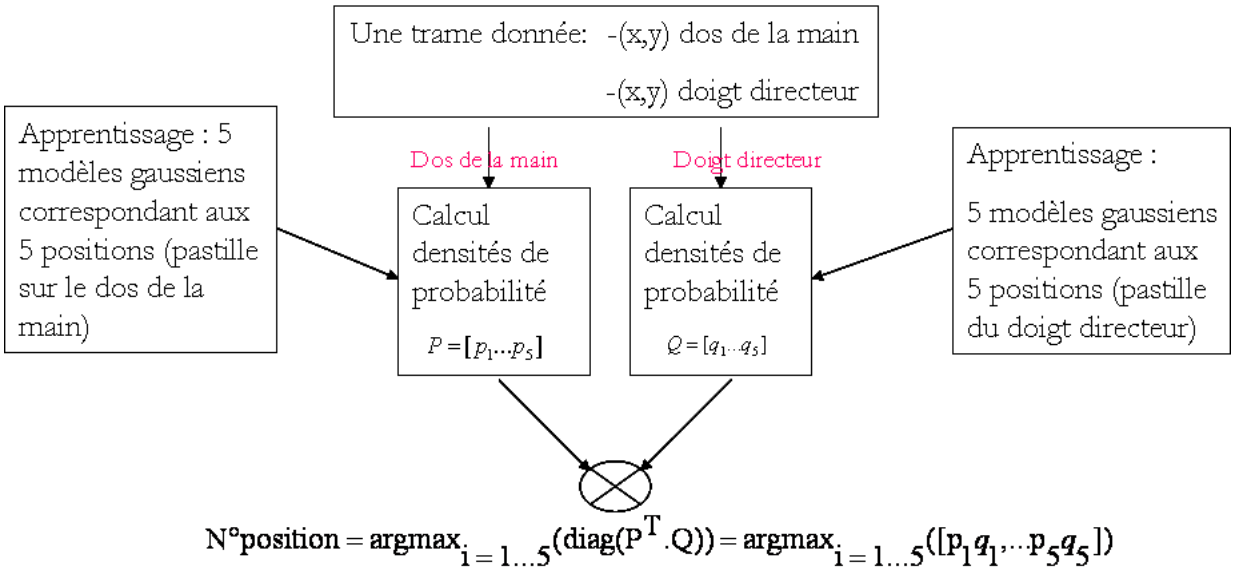

FIG. 6.5 – Schéma résumant la première partie de l'algorithme de détection des positions LPC.

en dessous du seuil Vs sont définis comme étant dans la cible. Cette étape de filtrage permet de supprimer de faux plateaux cibles. Le résultat final (voir figure 6.7) définit des bornes de plateau qui correspondent à l'instant d'atteinte de position cible LPC (noté $M2$) et l'instant de fin de tenue ($M3$), correspondant aussi à l'instant de début de transition vers la position cible

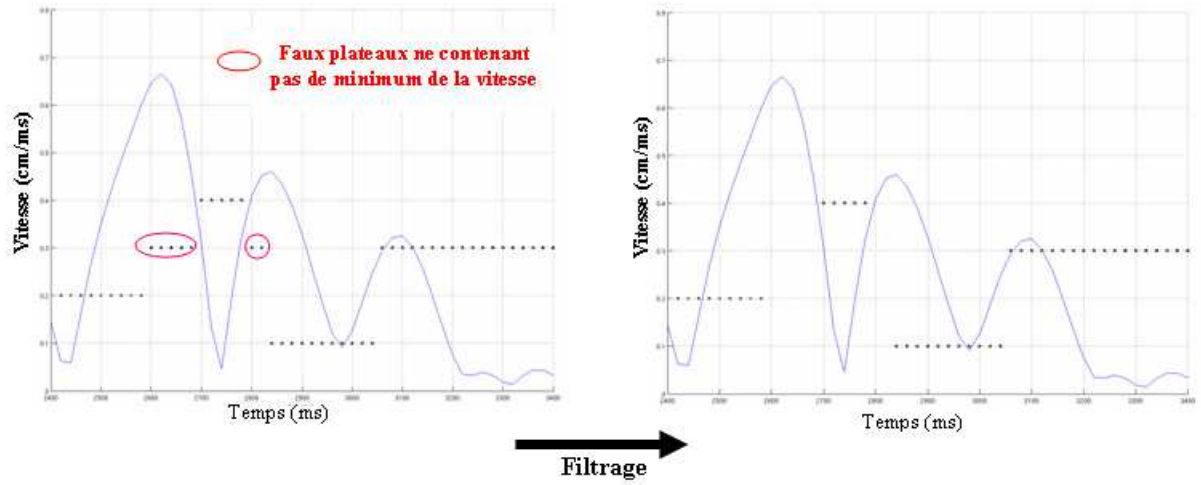

FIG. 6.6 – Les faux plateaux qui ne contiennent pas des minima de vitesse sont à supprimer (filtrage).

suivante (noté $M1$), selon la nomenclature définie par Attina et al. (2004).

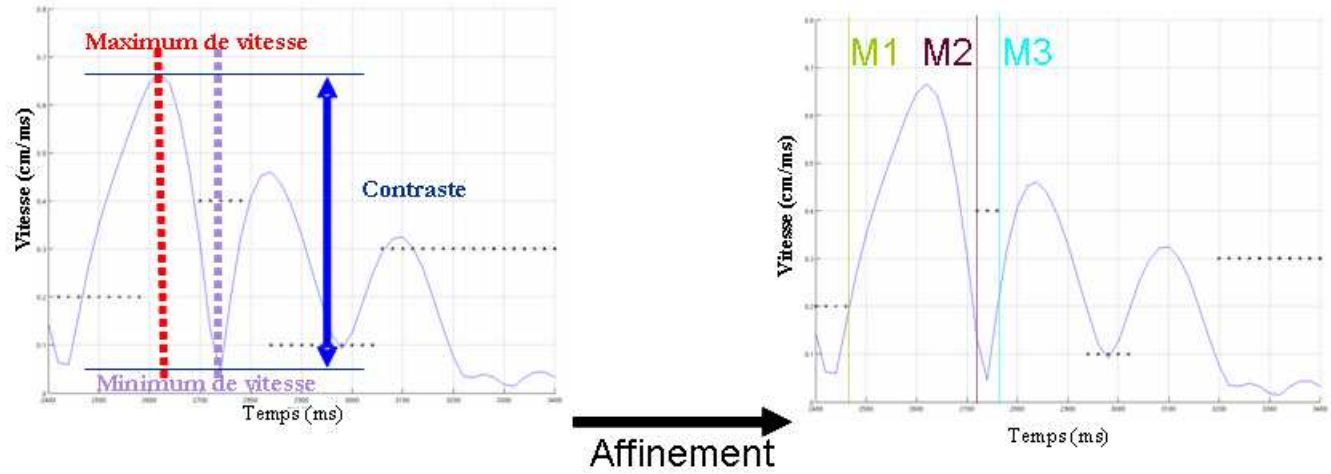

FIG. 6.7 – Calcul du contraste pour affiner les plateaux de positions cibles.

6.1.2 Evaluation

La détection automatique des positions LPC de la main et des instants $M1$, $M2$ et $M3$ est appliquée à toutes les séquences du corpus. Afin de mesurer les performances de cette méthode au niveau de l'identification de la position de la main et de la précision sur ces instants nous avons effectué une évaluation sur une partie du corpus. En plus, une étude sur la synchronisation du flux de la main par rapport au flux labial et acoustique permet aussi d'évaluer indirectement la technique en comparant les résultats obtenus à ceux trouvés dans la littérature.

6.1.2.1 Taux d'erreur

Pour cette première évaluation, nous avons utilisé un sous-ensemble composé des 60 premières phrases répétées au moins deux fois, ce qui a conduit à 130 phrases. Le premier niveau consiste à évaluer la première étape, c'est-à-dire la modélisation et la classification gaussiennes de 1001 positions cibles en comparant le résultat à celui obtenu par un expert identifiant à l'œil ces cibles. Nous savons qu'à l'instant de minimum local de vitesse, la main est par définition en position cible. La table 6.1 présente la matrice de confusion pour l'identification des cinq positions cibles du code LPC, en se plaçant à l'instant du minimum local de vitesse contenu dans chaque plateau à l'issue du traitement final.

	Coté	Pommette	Bouche	Menton	Gorge
Coté	373	1	4	0	0
Pommette	1	75	3	0	0
Bouche	2	1	215	3	0
Menton	0	0	3	163	0
Gorge	1	0	0	0	156
% d'identification	98,9	97,4	95,6	98,2	100

TAB. 6.1 – Matrice de confusion de la classification des positions cibles LPC par le système (colonne) et par l'expertise (ligne).

Le taux global d'identification est de 98,1 %. Celui-ci correspond au pourcentage de classification commune entre l'expert et le système. Ce pourcentage élevé indique que le choix de la modélisation et la classification gaussienne est pertinent. Cependant on remarque que le taux varie entre 95,6 % et 100 % d'une position à l'autre. En fait, il reste de faux plateaux de position cibles qui reflètent des ralentissements dans les transitions. Ceci a pour conséquence de présenter à l'expert une image d'une main en transition ne lui permettant pas de l'affecter de manière certaine à une position cible. Il peut s'en suivre une différence entre l'affectation par le système et l'expert.

Le deuxième niveau d'évaluation consiste à évaluer la pertinence du choix de 40 % du contraste entre pic de vitesse et vitesse minimum utilisé dans l'affinement des plateaux (seconde étape). Pour ce, ont été présentés à l'œil de l'expert les images des instants M2 obtenus avec l'application d'un seuil de 40 % sur le contraste de vitesse (table 6.2) et d'un seuil de 50 % (table 6.2).

Le taux global d'identification est de 96,5 % dans la table 6.2, très proche du niveau de référence (98,1 %) de la table 6.1. Le plus grand écart s'obtient pour la position « bouche » (91,1 % vs 95,6 %). **Le taux global d'identification est de 94,7 %** dans la table 6.3, inférieur à celui de la table 6.2. On remarque de plus une forte différence pour la position « pommette » par rapport à la référence de la table 6.1 (77,9 % vs 97,4 %) ce qui traduit que pour cette position, l'image présentée à l'expertise se trouvait dans la transition dans près de 20 % des cas et donc pas en position cible. Cette différence justifie a posteriori le choix empirique du seuil de 40 % sur le contraste de vitesse pour définir l'instant d'atteinte de la position LPC,

	Coté	Pommette	Bouche	Menton	Gorge
Coté	373	4	11	0	0
Pommette	1	72	3	0	0
Bouche	3	1	205	3	0
Menton	0	0	6	161	1
Gorge	0	0	0	2	155
% d'identification	98,9	93,5	91,1	97	99,4

TAB. 6.2 – Matrice de confusion de la classification des positions cibles LPC à l'instant $M2$ par le système (colonne) avec un seuil de 40% sur le contraste de vistesse et par l'expertise (ligne).

	Coté	Pommette	Bouche	Menton	Gorge
Coté	370	16	15	1	0
Pommette	2	60	2	0	0
Bouche	5	1	202	3	0
Menton	0	0	6	162	2
Gorge	0	0	0	0	154
% d'identification	98,1	77,9	89,8	97,6	98,7

TAB. 6.3 – Matrice de confusion de la classification des positions cibles LPC à l'instant $M2$ par le système (colonne) avec un seuil de 50% sur le contraste de vistesse et par l'expertise (ligne).

c'est-à-dire $M2$.

Le système d'identification de la position LPC décrit ici atteint un taux global de 98,1% et de 96,2% à l'instant de segmentation $M2$. Ces bons scores cachent cependant quelques phénomènes qui expliquent les quelques erreurs de détection et de segmentation et qui sont attribuées à un ralentissement de la main entre deux positions cibles pour différentes raisons :

- La vitesse avec laquelle la main se déplace entre deux positions est trop élevée pour laisser le temps à la forme de la main de se déployer complètement. La main marque alors « un temps d'arrêt » avant de reprendre de la vitesse pour atteindre sa position cible. Le système prend donc le ralentissement en compte et définit un plateau clairement dans la transition.
- Une hésitation de la codeuse peut entraîner un ralentissement de la main, et donc un plateau de positions indésirable.
- Le mauvais placement de la main entre deux positions cibles, qui constitue une erreur de codage, peut causer une mauvaise identification de la position.

La prise en compte de la dynamique du déploiement de la forme de la main pour coder la consonne pourra améliorer l'algorithme pour palier ces effets.

6.1.2.2 Phasage main-lèvres

Le but de cette partie est d'analyser la désynchronisation entre la main et les lèvres en relation avec la réalisation acoustique correspondante dans le cas de syllabes Consonne-Voyelle (CV). En ce qui concerne cette partie, 57 syllabes sont extraites des séquences. Les syllabes sont de type CV avec les consonnes [p, t, k, b, m] pour C et les voyelles [a, i, u, y, e, ɛ, œ, ø, o, ɔ] pour V. Le choix de ces consonnes est motivé par la facilité de pouvoir les identifier sur les signaux acoustiques et articulatoires. D'un autre coté, les syllabes CV qui se situent au début et à la fin des séquences ainsi que celles précédées ou suivies d'une pause prosodique ne sont pas considérées. Ceci dans le but d'éviter les cas spécifiques dus à l'initialisation du geste de la main et la relaxation de la main. Afin d'éviter des éventuelles erreurs de précision que peut générer le système d'étiquetage automatique du signal acoustique, les débuts des réalisations acoustiques de la consonne et de la voyelle ainsi que la fin de la voyelle nommés respectivement $A1$, $A2$, $A3$, sont labelisés manuellement. De plus, l'instant de la réalisation de la voyelle sur les lèvres nommé $L2$ (nomenclature d'Attina, 2004) est aussi labelisé à la main. A partir de tous ces labels, les durées des intervalles $M1A1$, $A1M2$, $M1M2$, $M2M3$, $A1A3$, $A1A2$, $M2L2$ et $M3L2$ sont calculées. Pour un intervalle donné, la durée est calculée comme la différence arithmétique, c'est-à-dire le second label moins le premier label (par exemple : la durée de $M1A1$ est $A1-M1$). Toutes ces durées sont présentées dans la table 6.4.

Intervalles	Moyennes (ms)	Ecarts types (ms)
Syllabe CV (A1A3)	284,10	74,58
Consonne (A1A2)	142,51	34,95
M1M2	215,09	64,17
M2M3	94,74	69,23
M1A1	151,79	86,97
M1M3	309,83	100,57
A1M2	63,30	71,57
M2L2	144,17	80,68
M3L2	49,43	96,65

TAB. 6.4 – Moyennes et écarts types des durées des différents intervalles.

Sur la figure 6.8, nous présentons la position temporelle relative de ces intervalles par rapport à la durée acoustique d'une syllabe CV (c'est-à-dire $A1A$). Sur cette figure, nous observons une large anticipation du mouvement de la main. En d'autres termes, la transition vers une position commence largement avant le début de la réalisation acoustique de la consonne ($M1A1$). La position de la main est alors atteinte dans la première partie de la consonne ($A1M2$), et donc largement en avance par rapport à la réalisation aux lèvres de la voyelle ($M2L2 > 0$). Ces résultats sont globalement compatibles avec les résultats obtenus par Attina (2004, 2005). Cependant, une différence est à noter pour l'intervalle $A1M2$ (nous obtenons 22,98% alors que Attina obtient 10% pour le même sujet, mais une variation large allant de 6 à 18% pour d'autres sujets). Cette différence peut être expliquée par la définition de l'instant $M2$. En effet, notre M2

est défini à partir d'un seuillage automatique de vitesse de la main alors que le $M2$ est défini par Attina à partir d'un marquage manuel en se basant sur les pics de décélération. En outre, contrairement aux syllabes utilisées par Attina où elles sont dans un contexte isolé, nos syllabes CV sont dans un contexte de phrases dans lesquelles des syllabes plus complexes (CCV, VCV...) se succèdent l'une à l'autre. Ceci peut aussi expliquer la différence observée sur l'intervalle $M2M3$ pour lequel nous obtenons 36,34% tandis qu'Attina obtient 64% : la main est maintenue en position pendant une courte durée, ce qui explique la différence sur M3L2 (15% contre -6%).

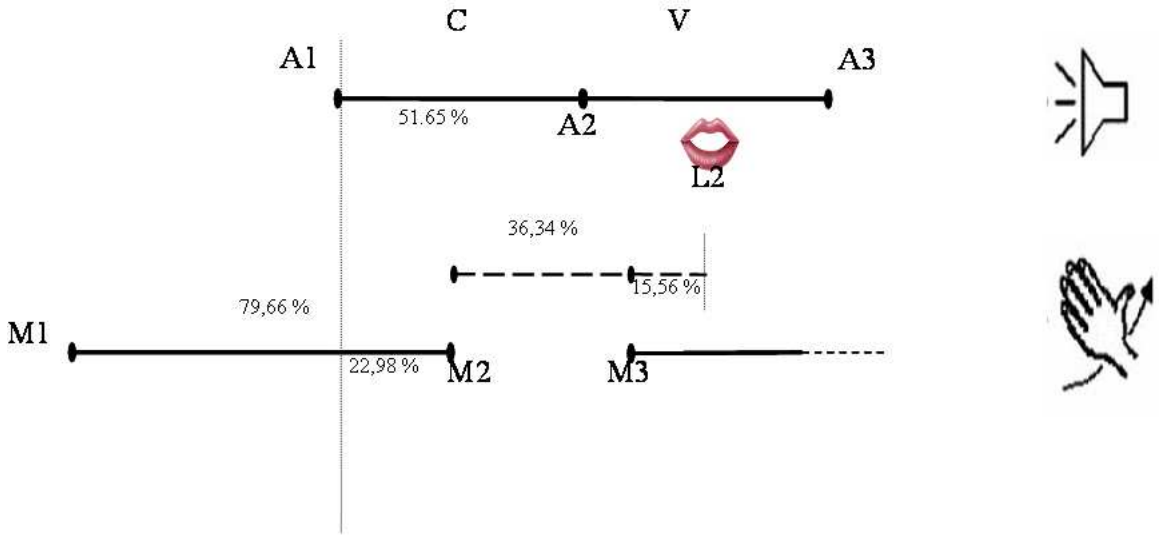

FIG. 6.8 – Patron temporel de la coordination entre le son, les lèvres et la position de la main dans le cas de la production du code LPC. La durée de chaque intervalle est exprimée en pourcentage de la durée $A1A3$.

Le point principal de cette étude est l'importance de l'instant $M2$. Dans cette étude, l'instant $M2$ apparaît comme le principal point de contact pour la coordination des gestes du code LPC ($M1M2$ et $M2M3$). Effectivement, si nous comparons les intervalles $A1M2$, $M2L2$, $M3L2$ et $M1A1$, nous constatons que la variance de l'intervalle $A1M2$ est la plus faible (voir table 6.4). Ceci révèle, en termes de contrôle, une plus grande précision du contact du début de la position de la main avec le commencement de la syllabe CV. D'un autre coté, si nous comparons les intervalles $M3L2$ et $M2L2$, nous constatons que les écarts types de ces deux intervalles sont relativement larges (96,65 ms contre 80,68 ms, voir table 6.4). Ceci montre une large variabilité du positionnement de l'instant $M3$ par rapport à l'instant $L2$, ce qui semble prouver que $M3L2$ ne peut être la principale relation de contrôle. Il est à noter aussi que l'intervalle $M1A1$ a une variation large par rapport à $A1M2$, ce qui confirme notre observation sur $M3$, puisque ce dernier n'est pas seulement la fin de la tenue de main en position cible mais aussi définit le début de la transition vers la position cible suivante, et donc comme est lié à $M1$. La figure 6.9montre le positionnement temporel des différents événements. **En résumé, cette étude confirme bien, et dans un contexte plus complexe, l'importance de l'instant** $M2$ **relevé par Attina (2005)**. L'instant $M2$ est défini comme l'instant où la main atteint sa

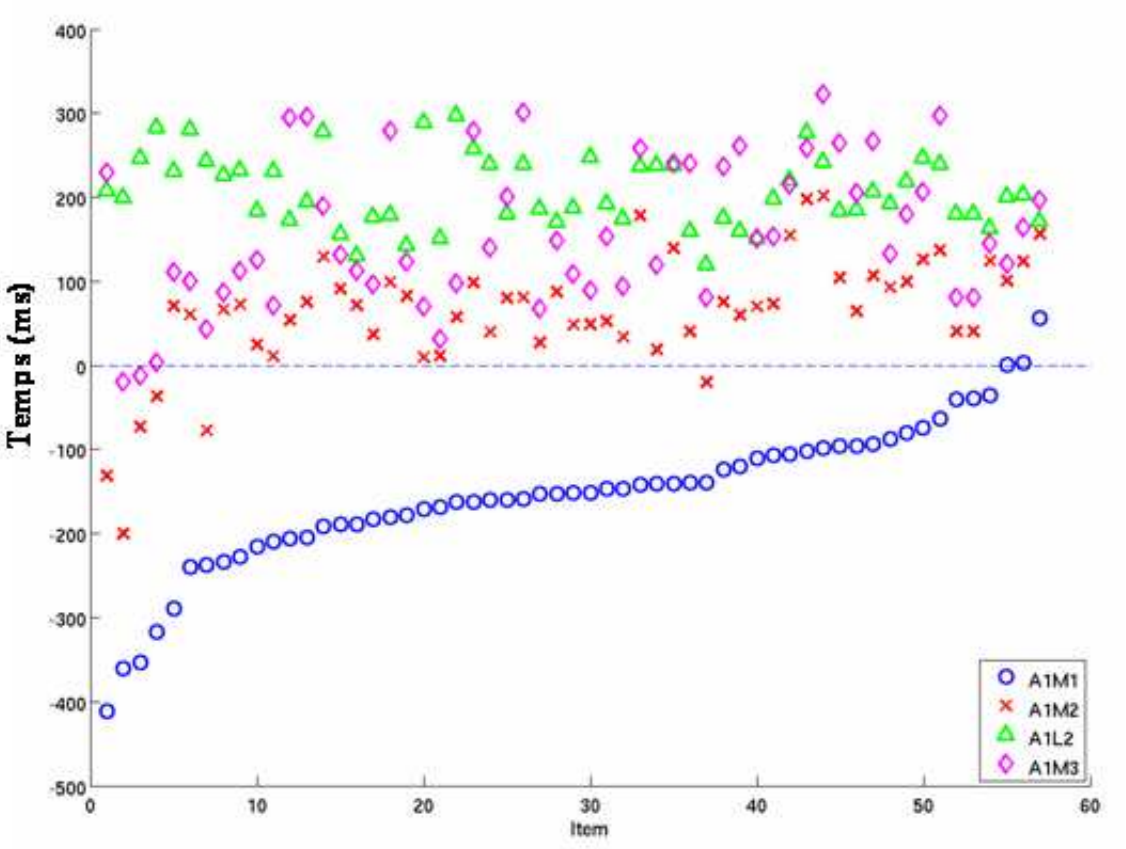

FIG. 6.9 – Distribution temporelle relative des labels $M2$, $M3$ et $L2$ par rapport au label $A1$ classé dans l'ordre croissant des valeurs de $M1A1$.

position cible et en même temps où la configuration digitale est complètement formée. A cet instant, l'information de la consonne contenue dans la configuration LPC de la main est donc connue. Ainsi, dans la perspective de la fusion des flux d'information de la main et des lèvres, la main définit l'instant dans lequel les lèvres peuvent être analysées afin de déterminer uniquement la consonne. Par contre, puisque l'instant $L2$ suit l'instant $M2$, **l'identification complète de la voyelle nécessite de prendre en compte le décalage entre l'information donnée en avance par la main (plateau $M2M3$) et celle délivrée par les lèvres à l'instant $L2$.**

6.1.3 Extension à un sujet malentendant

Cette méthode automatique de segmentation temporelle des positions LPC de la main a été appliquée telle qu'elle, pendant le travail de stage de Pablo Sacher, étudiant en Master recherche que j'ai -co-encadré, à des données provenant d'un enregistrement d'un autre sujet mal-entendant. Ce dernier prononce et code des phrases dans un contexte de dialogue défini (réservation téléphonique) ; c'est-à-dire, le sujet construit ses propres phrases. Les conditions d'enregistrement, notamment les artifices employés, étaient dans cette expérience similaires à ce que nous avons imposé pour notre enregistrement avec certaines exceptions. Dans cette expérience, la tête du codeur n'est pas tenue fixe par un casque mais laissée libre. Le codeur

ne porte pas de lunettes (le sujet codeur est mal-entendant donc il devait voir son interlocuteur codant aussi avec le code LPC, pour dialoguer), ce qui a pour conséquence que l'éclairage est moins puissant pour éviter l'éblouissement et l'emplacement des pastilles de référence diffère. En effet, une seule pastille de référence est utilisée et est placée sur le front du codeur ; ce qui permet d'avoir un repère relatif à la tête. La reconnaissance de la position LPC suit la même méthodologie décrite précédemment et les plateaux de positions cibles ainsi que les transitions sont obtenus (les instants $M1$, $M2$ et $M3$).

Dans cette expérience aussi, un patron caractérisant la coordination temporelle main-lèvres est établi. La figure 6.10 présente les moyennes des durées obtenues pour les différents intervalles.

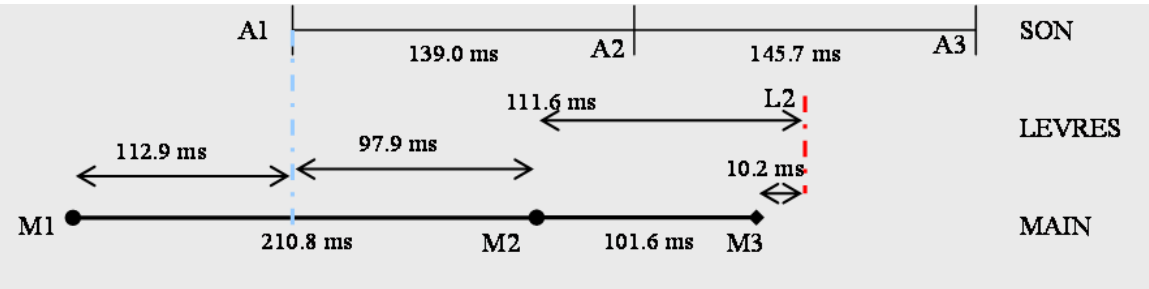

FIG. 6.10 – Patron temporel de coordination main-lèvres obtenu pour un sujet malentendant.

Il est à noter que, contrairement à notre étude, les instants $A1$, $A2$ et $A3$ ont été obtenus en utilisant le système d'étiquetage phonétique automatique, du signal acoustique, que nous avons décrit dans le chapitre précédent. De même, l'instant $L2$ (cible vocalique aux lèvres) est obtenu avec une méthode de détection automatique que nous décrivons dans le chapitre suivant.

En comparaison avec notre patron, la coordination temporelle chez ce sujet mal-entendant est globalement similaire du point de vue de l'organisation générale des relations main-lèvres avec une atteinte de la position manuelle dans la réalisation de la consonne, et un redémarrage vers une nouvelle position au moment de l'atteinte de la cible labiale de la voyelle. Les résultats obtenus dans notre cas (table 6.4 et celui de cette expérience semblent assez cohérents, si ce n'est quelques différences entre certains intervalles ($M1A1$, $A1M2$, $M2L2$ et $M3L2$). Ces différences ne sont dues qu'à un décalage (environ 30 ms) de la main sur les lèvres. Les durées de transition de la main d'une position à l'autre ($M1M2$) et les durées de tenue de la cible manuelle sont similaires (par rapport à nos résultats). En définitive, nous retenons donc que les patrons temporels de coordination main-lèvres entre cette étude et la notre sont comparables. **Ainsi, l'organisation temporelle des geste main-lèvres de la codeuse normo-entendante est confirmée chez un codeur mal-entendant**.

En résumé, la position LPC de la main est détectée en deux phases. Dans un premier temps, l'objectif est d'associer chaque point représentant les coordonnées de la main à une position LPC par une catégorisation automatique qui repose sur un modèle gaussien utilisant le critère du maximum de vraisemblance. Dans un second temps, un seuil sur le contraste de vitesse est appliqué afin d'associer l'intervalle au sein duquel la vitesse est inférieure à ce seuil avec les positions trouvées précédemment. L'évaluation directe de cette méthode, avec des taux d'identification, établit la précision de notre méthode et donc nos choix d'imposer des artifices. L'obtention avec cette méthode d'un patron temporel de coordination main-lèvres similaire à celui observé pour la première fois par Attina et al. (2004), pour le même sujet, valide à un autre niveau la méthode. Enfin, cette méthode appliquée cette fois-ci à l'analyse du code d'un sujet malentendant a permis d'obtenir rapidement le patron de coordination.

6.2 Reconnaissance de la configuration

6.2.1 Méthode

La reconnaissance des huit configurations LPC de la main est un cas particulier dans la reconnaissance des postures de la main. En effet, elles peuvent être distinguées en partie seulement avec un comptage du nombre de doigts visibles (non pliés). Dans le cas où certaines configurations font apparaître le même nombre de doigts, la dispersion de ces doigts et l'angle entre eux peuvent être utilisés comme des informations complémentaires. L'algorithme que nous proposons pour identifier les formes LPC de la main est le suivant :

- Si le nombre des pastilles des doigts détectées = 1 $\longrightarrow$ configuration n°1 ;
- Si le nombre des pastilles des doigts détectées = 4 $\longrightarrow$ configuration n°4 ;
- Si le nombre des pastilles des doigts détectées = 5 $\longrightarrow$ configuration n°5 ;
- Si le nombre des pastilles des doigts détectées = 3 $\longrightarrow$ configuration n°3 ou configuration n°7 : détection de l'apparition du pouce (en utilisant les modèles de dispersion précédents) ;
- Si le nombre des pastilles des doigts détectées = 2 $\longrightarrow$ configuration n°2 ou configuration n°6 ou configuration n°8 :

  * si pouce détecté $\longrightarrow$ configuration n°6 ;
  * sinon nous calculons l'angle entre les deux pastilles et nous fixons fixe un seuil empirique : si l'angle entre les deux pastilles est supérieure à ce seuil $\longrightarrow$ configuration n°8 sinon configuration n°2 ;

Appliqué à chaque séquence de notre corpus, cet algorithme nous permet d'obtenir une configuration de la main toutes les 20ms. La figure 6.11 illustre les configurations obtenues pour un exemple de séquence. Dans cette figure, nous pouvons observer que les configurations cibles de la main sont représentées, comme pour les positions, par des plateaux. La longueur de ces

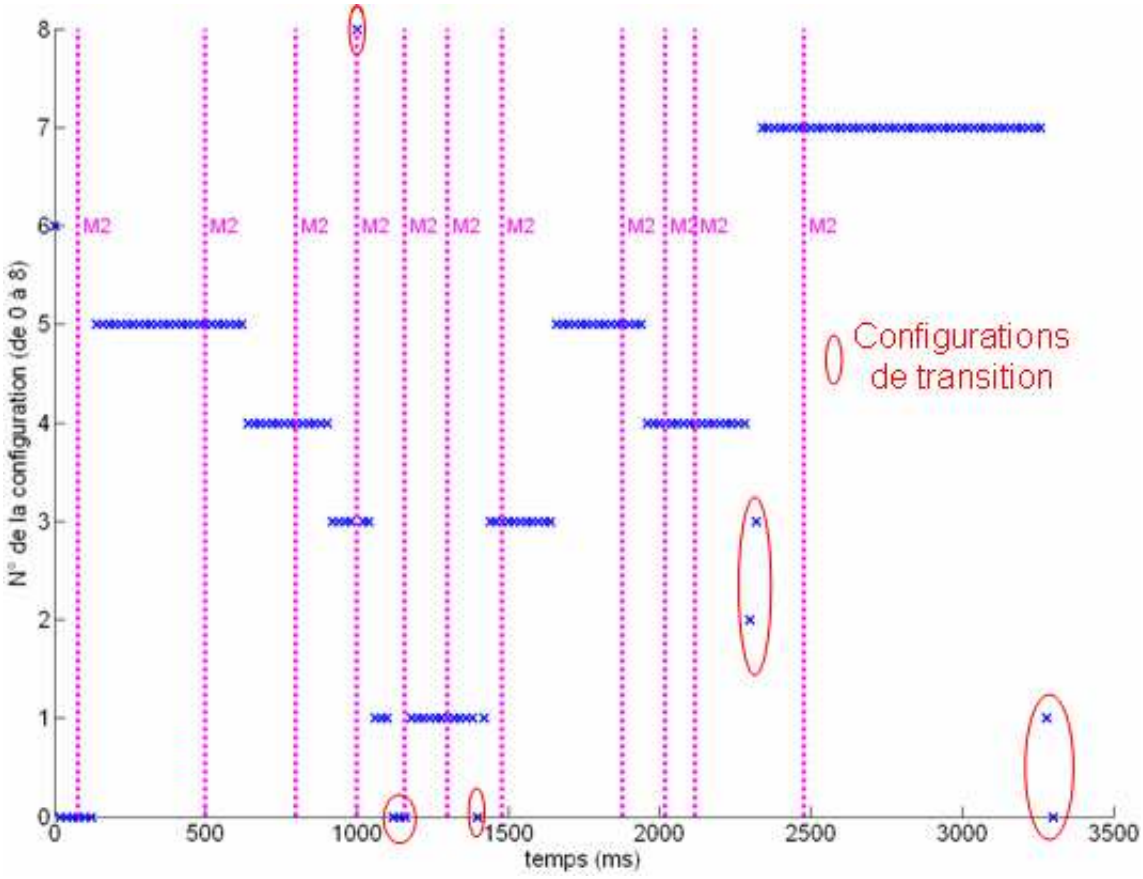

FIG. 6.11 – Configurations obtenues pour la séquence "une réponse ambiguë". Les configurations de transition d'une configuration cible à une autre sont marquées.

plateaux est variable selon la durée de maintien de la configuration par la codeuse. Cependant, il ne faut pas considérer les plateaux de courte durée (20 à 60 ms [2]) comme des cibles, car ils sont des plateaux de transitions ; c'est-à-dire, des plateaux de configurations intermédiaires pour passer d'une configuration LPC à une autre. De ce fait, un filtrage de ces plateaux est nécessaire. Ainsi, tout plateau dont la durée est inférieure à 60 ms, est d'abord supprimé et la configuration correspondant à chaque instant de la première moitié de ce plateau prend la valeur de la configuration cible précédente et celle de la seconde moitié la configuration cible suivante. Avec ce filtrage simple, nous pouvons réduire les transitions notamment celles qui ne correspondent pas à des configurations LPC. La figure 6.12 montre le résultat du filtrage sur l'exemple de la figure 6.11.

6.2.2 Evaluation

Nous avons montré précédemment l'importance de l'instant $M2$, qui correspond, nous le rappelons, à l'instant où la main atteint sa cible. A cet instant, la configuration LPC de la main doit être déjà formée en raison du contact de la main avec le visage ou le torse pour

[2] *Valeurs déterminées empiriquement.*

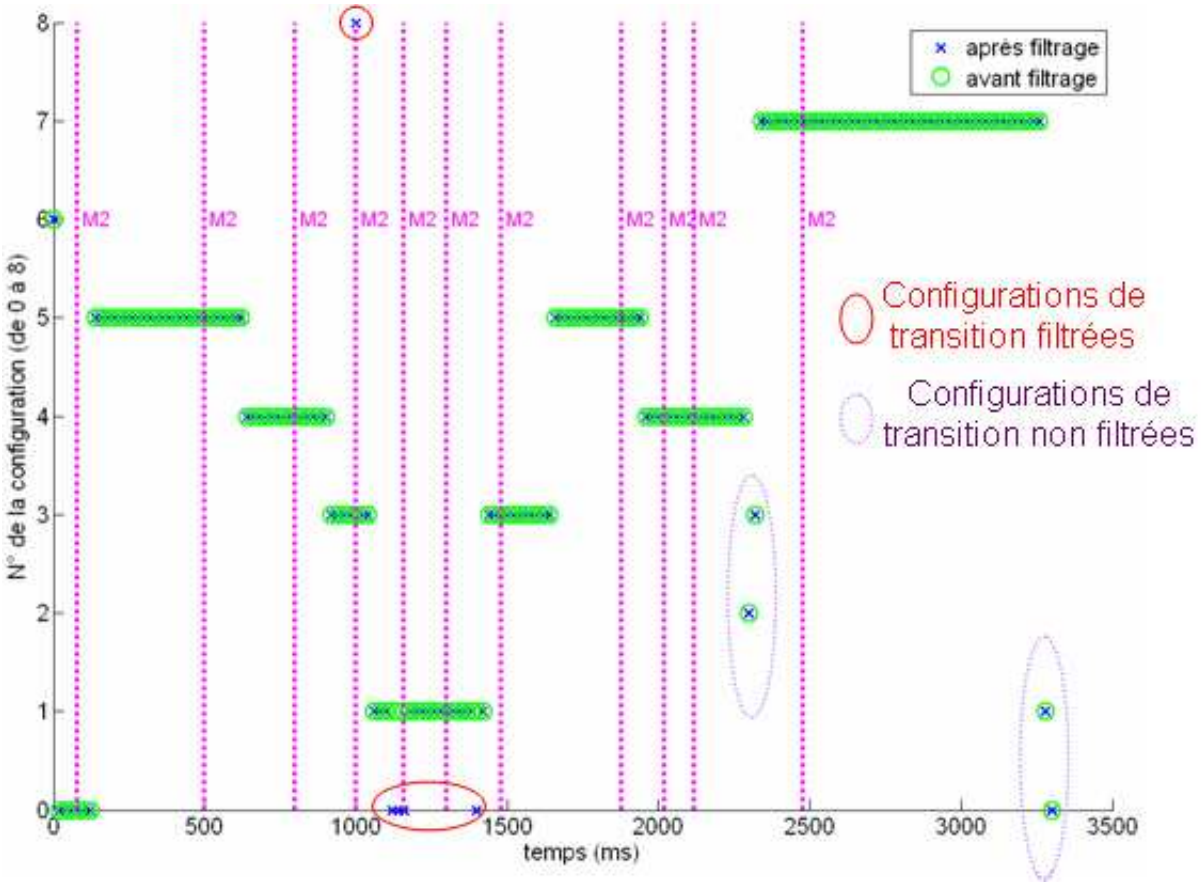

FIG. 6.12 – Configurations obtenues après filtrage en comparaison avec celles obtenues avant filtrage pour la séquence "une réponse ambiguë". Le filtrage supprime les points (configurations) isolés mais pas deux points isolés successifs.

4 des cinq positions LPC (voir figures 6.11 et 6.12). Ainsi, pour évaluer la performance de notre système à reconnaître correctement les configurations de la main, nous sélectionnons aux instants $M2$ les images correspondantes sur les 60 premières phrases répétées au moins deux fois. Nous construisons ainsi un sous-ensemble de 1009 images que nous présentons à un expert qui détermine quelle est la configuration de la main. Pour faciliter la confrontation des résultats de l'expert et du système automatique, nous numérotons les configurations LPC de la main de 1 à 8 (voir figure 6.13). Dans le cas où la configuration formée ne peut être catégorisée comme une configuration LPC, nous affectons à cette configuration le numéro 0. La table 6.5 présente la matrice de confusion de la classification des configurations LPC reconnues par le système automatique et par l'expert.

A partir de ces résultats, nous observons que **notre système automatique identifie correctement 92% des configurations**. Ce taux justifie nos choix et montre que l'utilisation de seulement 5 pastilles, placées sur les bouts des doigts, ne diminue pas sensiblement la précision de reconnaissance des formes LPC de la main en comparaison avec la précision obtenue par Gibert et al. (2005), qui obtiennent avec 50 marqueurs placés sur la main un taux de reconnaissance des formes LPC de la main de 98,78%. En analysant nos résultats par configuration,

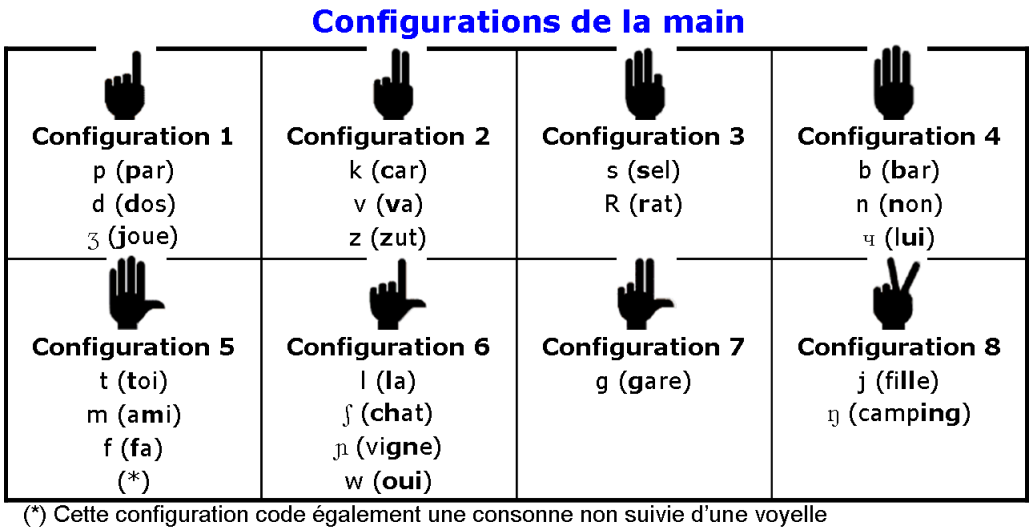

FIG. 6.13 – Les configurations LPC de la main numérotées de 1 à 8.

	configuration N°								
	0	1	2	3	4	5	6	7	8
config 0	33	16	1	0	3	2	0	0	1
config 1	2	151	2	2	0	0	0	0	0
config 2	0	0	93	0	0	0	0	0	0
config 3	0	0	0	163	0	4	0	0	2
config 4	0	0	0	2	100	4	0	0	0
config 5	0	0	0	0	0	193	0	0	0
config 6	0	1	0	0	0	0	124	0	0
config 7	0	0	0	3	3	0	5	17	0
config 8	0	5	6	9	0	1	0	0	58
% d'identification	94	87	91	91	94	95	96	100	95

TAB. 6.5 – Matrice de confusion de la classification des configurations LPC reconnues à l'instant $M2$ par le système (colonne) et par l'expert (ligne).

nous constatons que c'est la configuration 1 qui présente le plus faible taux d'identification. L'expertise identifie 16 fois, sur les 22 cas où l'expertise reconnait une autre configuration, que la configuration reconnue par le système automatique comme la configuration 1, est en fait une configuration 0. Ceci implique que 73% des erreurs de reconnaissance sur la configuration 1 sont dues à une confusion avec la configuration 0. Ceci s'explique par le fait que le système détecte une pastille alors que la codeuse ne forme aucune configuration LPC. Par exemple, nous avons remarqué qu'une pastille peut être visible quand la codeuse plie ces doigts (configuration 0). La figure 6.14 illustre ce cas de figure.

Par ailleurs, dans les autres cas, les erreurs sont dues en général au système de détection des pastilles. En effet, d'un coté il se peut que, dans certains cas, une ou plusieurs pastilles ne

FIG. 6.14 – Cas où une pastille est visible (ici la pastille du pouce) alors qu'aucune configuration LPC n'est codée. L'image est prise à un instant $M2$.

soient pas détectées à cause d'une rotation de la main ou des doigts. Par exemple, en position "gorge", la codeuse a tendance à appuyer fortement tout en pliant l'index quand elle code la configuration 1, ce qui fait que la pastille sur ce doigt devient invisible (voir figure 6.15. Par ailleurs, il se peut qu'une ou plusieurs pastilles sur des doigts pliés (ne doivent pas être comptés) restent visibles, ce qui se traduit, après comptage des pastilles, par une configuration différente de la réalité.

FIG. 6.15 – Cas où une pastille n'est pas détectée par le système, ce qui fausse le comptage des doigts et donc identifie une mauvaise configuration. L'image est prise à un instant $M2$.

En résumé, la configuration LPC de la main est déterminée par un algorithme s'appuyant sur un comptage des pastilles placées sur les bouts des doigts. Le choix d'utiliser des pastilles en nombre réduit (sept au total) est justifié d'abord par la simplicité de cet algorithme et par la bonne précision d'identification correcte des configurations obtenue (92%).

6.3 Conclusion

Certes, nos deux algorithmes utilisés pour la détection et la classification des positions et des configurations LPC de la main s'appuient sur des concepts simples (classification gaussienne, comptage des doigts, seuillage ...). Cependant, ceci ne veut dire en aucun cas que les performances sont diminuées. Les résultats obtenus lors des différents tests d'évaluation montrent de très bonnes précisions dans les deux cas (position et configuration), justifiant ainsi l'utilisation ce type d'artifices (pastilles en bleu).

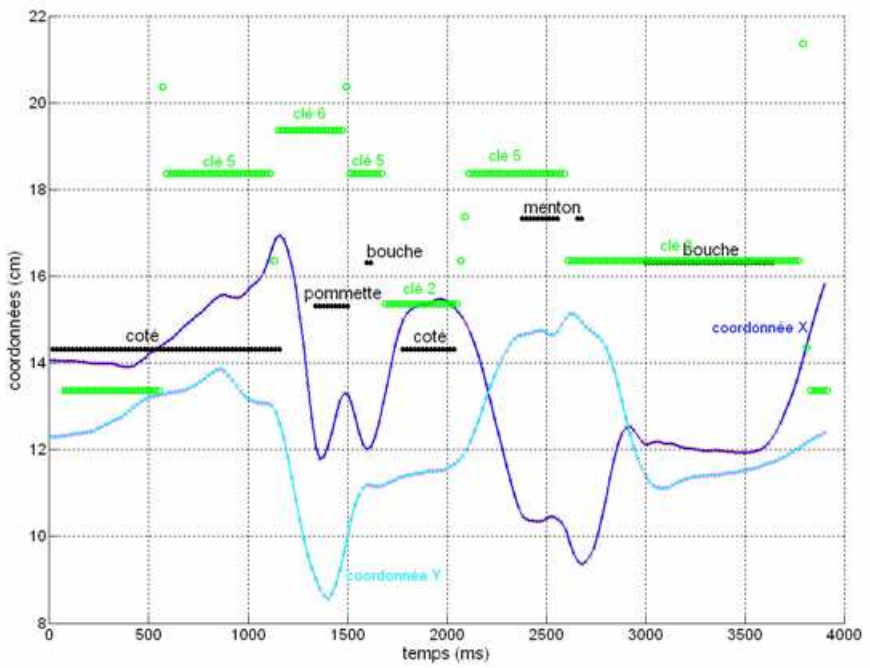

FIG. 6.16 – Résultats finaux de nos deux algorithmes d'identification de la position et la configuration LPC de la main appliqué sur la séquence "ma chemise est roussie". Sur cette figure le terme "clé" signifie "configuration".

L'une des pastilles les plus importantes est la pastille haute (sur le dos de la main). En effet, cette pastille, en calculant sa vitesse, permet de délimiter automatiquement les instants temporels $M1$, $M2$ et $M3$. Ces instants fournissent une segmentation temporelle des gestes de la main permettant de préciser l'instant où la main débute son mouvement vers la position cible ($M1$ ou $M3$) et l'instant d'atteinte de la position cible ($M2$).

Dans ce chapitre, nous avons confirmé dans un contexte plus complexe, en s'appuyant sur

la segmentation temporelle obtenue par notre algorithme d'identification de la position, l'importance de l'instant $M2$ pointé par Attina et al. (2004). Cet instant est défini comme l'instant auquel la main atteint sa position cible et où sa configuration est complètement formée. A cet instant, l'information de la consonne contenue dans le geste LPC de la main est donc connue. Ainsi, dans la perspective de la fusion des flux manuel et labial, la complète identification de la voyelle nécessite de prendre en compte la désynchronisation entre l'information donnée en avance par la main (dans l'intervalle $M2M3$) et celle délivrée par les lèvres à l'instant $L2$. Ce dernier point est traité dans le chapitre suivant. Enfin, le principe de l'algorithme de la segmentation temporelle des gestes de la main, s'appuyant sur les contrastes de la vitesse, sera appliqué également pour la segmentation temporelle automatique du flux labial.

Chapitre 7

Etude du flux labial

Ce chapitre est centré sur le flux labial et discute des classification des phonèmes du Français. Le flux labial est composé des variations temporelles de paramètres caractéristiques des contours des lèvres. Pour les voyelles, nous partons de l'hypothèse qu'un seul instant de mesure est suffisant pour caractériser une voyelle, puisque toutes les voyelles sont articulées aux lèvres. De plus, dans le cas des voyelles, les paramètres dérivés du contour interne des lèvres ont démontré leur efficacité dans plusieurs études (Benoit *et al.*, 1992; Robert-Ribès, 1995), à caractériser les voyelles. En revanche, cette hypothèse ne semble pas être réalisable pour les consonnes. En effet, toutes les consonnes ne sont pas articulées aux lèvres : il est donc difficile de déterminer un seul instant de mesure caractérisant chaque consonne. Ainsi, afin de pouvoir identifier les consonnes en plus des voyelles, nous prendrons en compte toute la transition de la syllabe consonne-voyelle (CV). Dans une première expérience, nous étudierons les voyelles que nous caractérisons à partir des paramètres du contour interne des lèvres mesurés à l'instant d'atteinte de la cible. Pour cette mesure, nous présentons une méthode de segmentation des cibles labiales des voyelles. Dans une second expérience, nous nous consacrerons à l'étude des syllabes CV à partir des paramètres extraits des contours des lèvres (externe et interne). Enfin, nous montrerons comment des modèles appris sur des syllabes CV peuvent être utilisés pour reconnaître des mots.

7.1 Expérience 1 : modélisation et classification des voyelles

7.1.1 Modélisation

7.1.1.1 Détection de la cible vocalique aux lèvres

L'objectif de cette partie est de repérer la cible labiale des voyelles (notées $L2$) contenues dans des phrases. Une solution est de s'appuyer sur l'étiquetage phonétique du signal acoustique qui fournit la segmentation des phonèmes (début et fin). L'hypothèse initiale étant que l'instant d'atteinte de cible labiale se trouve dans l'intervalle [début, fin]. Or, en plus de la désynchronisation possible et bien connue entre les flux auditif et visuel (Abry et Lallouache, 1991), se pose le problème de l'imprécision des instants de début et de fin[1] ce qui nous a conduit à rechercher la cible labiale autour de l'instant du milieu de cet intervalle sans être contraint par les bornes.

[1] *Ces instants étant obtenus automatiquement avec un alignement forcé, voir chapitre 5.*

Le critère est de définir la cible labiale à l'instant de minimum local de vitesse du paramètre labial considéré, le plus proche de l'instant milieu. La vitesse des lèvres est estimée en calculant la distance euclidienne entre les deux points $S(t)$ et $S(t+\Delta)$ successifs ramenée à l'espacement temporel Δ de 20 ms (S étant l'aire intérolabiale du contour interne des lèvres). Le choix du paramètre S est justifié par le fait que S est fortement corrélé au produit $A \times B$ ($r = 0.9591$ sur nos données). En effet, la vitesse des lèvres peut être calculée sur deux composantes verticale (sur B) et horizontale (sur A).

Le système de recherche des cibles pour les voyelles fonctionne donc en 4 étapes :

1. Calcul de la vitesse labiale sur le paramètre S,
2. Recherche de tous les minima locaux de cette vitesse,
3. Localisation de la voyelle et de l'étiquette milieu issu de l'audio,
4. Estimation de l'instant de cible labiale de la voyelle par l'instant de minimum local le plus proche de l'instant milieu.

La figure 7.1 illustre comment l'instant $L2$ est détecté.

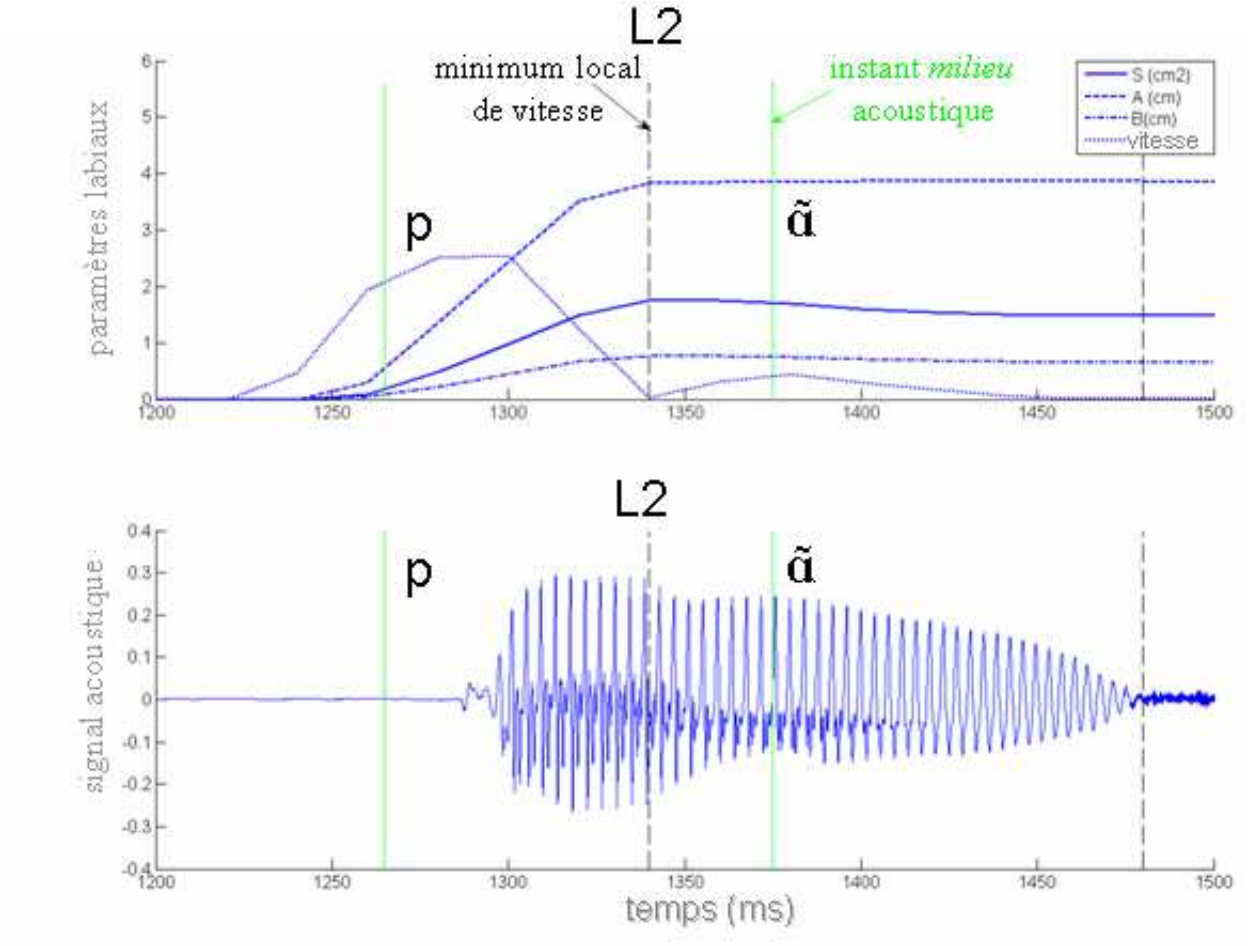

FIG. 7.1 – Détection de l'instant d'atteinte de la cible vocalique aux lèvres (L2), exemple d'une voyelle [ɑ̃].

Il est impératif de préciser que l'alignement forcé utilisé pour l'étiquetage phonétique du signal acoustique nécessite la connaissance a priori de la transcription phonétique des phonèmes. Cependant, l'objectif final de notre travail est précisément de reconnaître ces phonèmes à partir des paramètres visuels des lèvres fusionnés avec l'information de la main. En fait, nous avons utilisé cet alignement dans le but de contrôler le processus expérimental. Si nécessaire, il existe d'autres approches pour segmenter le signal acoustique de parole sans recours à la transcription phonétique (Golipour et O'Shaughnessy, 2007).

7.1.1.2 Arbre hiérarchique pour la classification des voyelles : le *dendrogramme*

Pour analyser la discrimination des voyelles à partir des paramètres labiaux, un arbre hiérarchique de *clusters* (appelé aussi *dendrogramme*) peut être calculé à partir de la distribution (A, B, S) des voyelles à l'instant d'atteinte de la cible labiale ($L2$). Le *dendrogramme* consiste en un ensemble de plusieurs axes verticaux connectant des objets (voyelles ou groupes de voyelles) dans un arbre hiérarchique. La hauteur de chaque axe vertical représente la distance entre les deux objets connectés, en utilisant une distance de Mahalanobis (voir figure 7.2 pour une illustration dans le cas des voyelles).

Distance de Mahalanobis Etant donné deux objets X et Y, représentés respectivement par deux matrices $N_x lignes * M colonnes$ et $N_y lignes * M colonnes$, la distance de Mahalanobis entre X et Y est une matrice D de $N_x lignes * N_y colonnes$. Si nous considérons :

$$D = \{D(x_i, y_j)\},$$

avec x_i : le $i^{éme}$ élément de la matrice X, y_j : le $j^{éme}$ élément de la matrice Y, et C : la matrice $M * M$ de covariance calculée à partir des données X et Y, alors :

$$D(x_i, y_j) = \sqrt{(x_i - y_j).C^{-1}(x_i - y_j)^T}., \text{ distance de Mahalanobis entre } x_i \text{ et } y_j.$$

Si C est diagonale, la distance de Mahalanobis correspond à la distance Euclidienne standardisée (c'est-à-dire normalisée par l'écart type). Si $C = I$, la matrice identité, la distance de Mahalanobis correspond à la distance Euclidienne.

Dans notre application de cette distance, nous retenons comme la distance entre deux objets (groupes de voyelles), la distance de Mahalanobis la plus petite (en anglais *shortest Mahalanobis distance*) définie comme :

$$min(D(x_i, y_j)), i \in [1, N_x], j \in [1, N_y]$$

Enfin, pour construire l'arbre hiérarchique, cette distance est calculée pour chaque paire d'objets et les distances sont ordonnées par ordre croissant. Les premiers niveaux du regroupement sont donc utilisés pour définir les groupes d'objets (les visèmes des voyelles dans notre cas).

a	o	œ	ɛ̃	ø	i	ɑ̃	ɔ̃	ɛ	u	ɔ	œ̃	y	e
216	63	24	26	110	176	67	41	96	69	32	26	97	124

TAB. 7.1 – Sous-ensemble 1 : 1167 voyelles extraites de la première répétition (ou autre que la seconde répétition) des 124 premières phrases du corpus.

7.1.2 Résultats

7.1.2.1 L'ambiguïté labiale et la complémentarité du code LPC

Ambiguïté labiale : Le fait que différent phonèmes sont produits avec des formes aux lèvres similaires cause des ambiguïtés pour la perception visuelle de la parole par le biais de la lecture labiale. Cette section analyse d'abord la proximité des voyelles en fonction de la distribution des paramètres labiaux produits et discute ensuite la complémentarité du code LPC pour la discrimination des voyelles.

Grâce à l'étiquetage phonétique du signal acoustique, nous avons sélectionné un sous-ensemble composé de 1167 voyelles (les 14 voyelles du français sont représentées) extraites des 124 premières phrases (cet ensemble sera appelé sous-ensemble 1) du corpus. La table 7.1 présente le nombre d'occurrence de chacune des voyelles.

pour ce sous-ensemble 1, nous avons tout d'abord extrait les paramètres labiaux du contour interne des lèvres à l'instant cible $L2$ et ensuite nous avons calculé un *dendrogramme*. Les résultats des différents niveaux de regroupement sont présentés dans la figure 7.2.

Rappelons qu'en ordonnée, la longueur des branches verticales est la distance entre deux objets X et Y. Ainsi, pour le calcul de la distance entre les deux voyelles [ɔ] et [œ] (un premier regroupement visible sur la figure 7.2, l'objet X est l'ensemble des triplets (A, B, S) des 32 réalisations labiales de la voyelle [ɔ] et Y est l'ensemble des triplets (A, B, S) des 24 réalisations labiales de la voyelle [œ] (voir table 7.1). L'ensemble D est composée des 32×24 distances deux-à-deux entre les réalisations de [ɔ] et celles de [œ]. La matrice de covariance est de dimension 3×3 calculée sur (A, B, S) à partir des 32 réalisations de [ɔ] et des 24 de [œ]. La distance de Mahalanobis est la plus petite des 32×24 distances et la hauteur de la branche regroupant [ɔ] et [œ] dans le *dendrogramme* de la figure 7.2.

Dans le regroupement de la voyelle [ɑ̃] et de la classe [ɔ, œ], cette classe est composée de $32 + 24$ éléments (les 32 réalisations labiales de [ɔ] et des 24 réalisations de [œ]) et est considérée comme l'objet X dans le calcul de la distance avec l'objet Y qui est la voyelle [ɑ̃].

Dans cet arbre, les 14 voyelles du sous-ensemble 1 sont présentés en abscisse. Au niveau de la distance 3,5 sur l'axe vertical, la figure 7.2 montre trois groupes contrastés, qu'on appellera désormais visèmes, à l'intérieur desquelles les voyelles sont proches : (i) visème 1 : les voyelles antérieures non arrondies [a,ɛ̃,i,œ̃,e,ɛ] ; (ii) visème 2 : les voyelles arrondies fermées ou moyennement fermées [y,ɔ̃,u,ø,o] ; (iii) et visème 3 : les voyelles arrondies ouvertes [ɔ,ɑ̃,œ]. Ces trois groupes sont globalement conformes aux quatre groupes obtenus par Robert-Ribès (1995); Robert-Ribès *et al.* (1998) dans leur étude visuelle des voyelles (*voir partie état de l'art*). A l'exception des voyelles nasales (non utilisées par les auteurs), la seule différence à noter concerne la

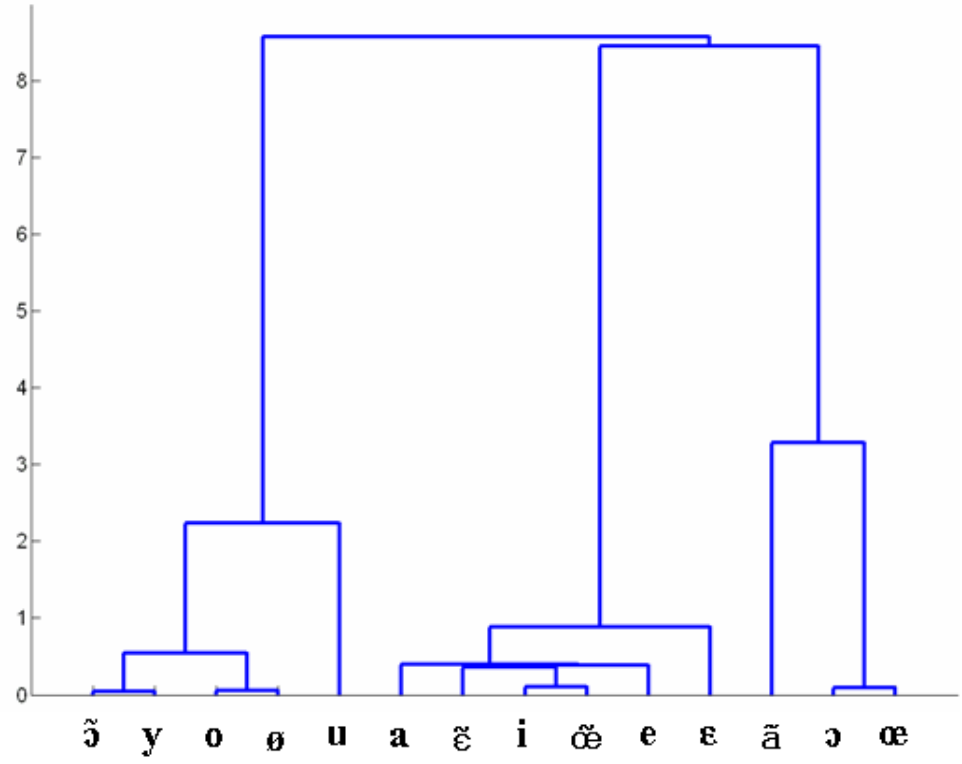

FIG. 7.2 – Arbre hiérarchique des voyelles obtenu à partir des paramètres du contour interne des lèvres pour un sujet normo-entendant.

voyelle [a]. Cette voyelle est catégorisée dans un groupe séparé par rapport au viseme 1. Cette différence peut être expliquée par le fait que, dans le cas de l'étude de Robert-Ribès (1995); Robert-Ribès *et al.* (1998), les voyelles ont été produites isolément (sans phrase porteuse) et tenues pendant une certaine durée (une seconde).

Utilisant les mêmes méthodes d'extraction des paramètres labiaux, appliquées sur les données d'un sujet sourd, et selon un processus identique d'analyse, l'étudiant de Master recherche que j'ai co-encadré, obtient un groupement similaire avec une seule exception (voir figure 7.3). Les visèmes obtenus dans cette étude sont : (i) visème 1 : [a,ɛ̃,i,œ̃,e,ɛ] ; (ii) visème 2 : [y,ɔ̃,u,ø,o, ɑ̃] ; (iii) visème 3 : [ɔ,œ]. L'exception concerne la voyelle [ɑ̃] qui se situe dans le visème 2 alors qu'elle devrait se trouver dans le visème 3. Cette exception est dûe probablement à l'articulation du sujet sourd, qui confond la voyelle [ɑ̃] avec la voyelle protrue du visème 2 : [ɔ̃]. La confusion entre ces deux voyelles se manifeste chez plusieurs personnes sans que pour autant la compréhension du message soit perturbée (c'était le cas exemple d'Antonin Artaud). **Cette étude**, malgré le cas de la voyelle [ɑ̃], **confirme pour un sujet mal-entendant notre regroupement de visèmes**.

Par ailleurs, le regroupement en trois visèmes est en conformité avec la description phonétique classique des voyelles. Les lèvres fournissent ici trois formes labiales contrastées (protrue, étirée et ouverte) groupant les 14 voyelles, qui ne peuvent être complètement discriminées à partir de l'information labiale seule. **Ce résultat illustre, du point de vue de la production, les**

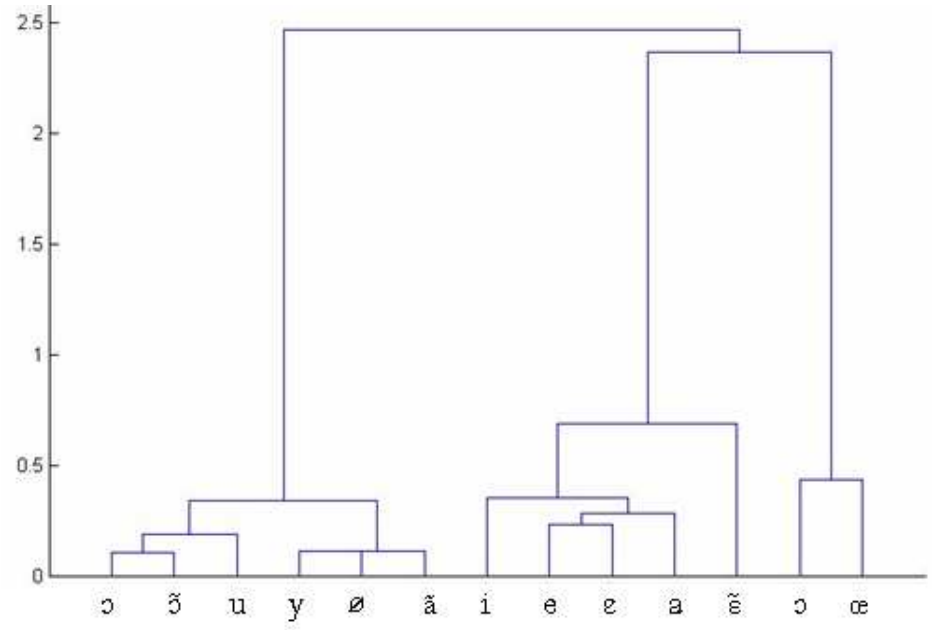

FIG. 7.3 – Dendrogramme représentant le regroupement hiérarchique des visèmes de voyelles obtenu pour un sujet sourd ; l'analyse a été effectuée sur 1022 voyelles.

causes de l'ambiguïté de la lecture labiale à laquelle sont confrontées les personnes sourdes.

Complémentarité du code LPC : Le regroupement en trois visèmes est maintenant comparé avec le système du code LPC. La table 7.2 présente les voyelles de chaque visème et la position LPC de la main correspondante, à partir du regroupement en visèmes des données de la figure 7.2.

Chaque voyelle du visème 2 ou du visème 3 est associée à une position LPC spécifique, et donc peut être discriminée en ajoutant la composante manuelle du code LPC. Pour le visème 1, nous remarquons une exception avec la position LPC "gorge". En effet, cette position est associée aux voyelles [e] et [œ] ; ce qui signifie que ces deux voyelles ne peuvent être distinguées. Ainsi, **à l'exception de ce dernier cas, le regroupement des voyelles en trois visèmes à partir des paramètres du contour interne des lèvres est compatible avec le système manuel du code LPC pour l'identification complète de la voyelle**.

La même comparaison avec les données du sujet sourd (figure 7.3) montre également une ambiguïté maintenue dans le visème 1 entre [e] et [œ], et aussi dans le visème 2 entre les voyelles [ɔ̃] et [ɑ̃].

Pour aller plus loin dans l'analyse et la modélisation de la discrimination des lèvres par le système manuel du code LPC, la distribution des voyelles à l'intérieur de chaque position LPC de la main est présentée sur la figure 7.4. Les données de chaque voyelle sont distribuées selon la loi normale autour d'une valeur moyenne. Sur cette figure, nous pouvons observer très peu de recouvrement entre les voyelles de chaque position à l'exception de la position "gorge". Pour

		Visème 1						Visème 2					Visème 3		
Voyelles		a	ɛ̃	i	ɛ	e	œ	o	ø	ɔ̃	u	y	œ	ɑ̃	ɔ
Position LPC de la main	Coté	X						X					X		
	Pommette		X						X						
	Bouche			X						X				X	
	Menton				X						X				X
	Gorge					X	X					X			

TAB. 7.2 – Voyelles des visèmes et les positions correspondantes de la main du système du code LPC.

cette dernière, les voyelles [œ̃] et [e] ont des distributions dans les plans (A, B) et (A, S) qui se chevauchent beaucoup (voir figure 7.4 pour le plan (A, S)). Ceci confirme le regroupement précédent de ces deux voyelles dans un même visème.

Comme vu, la discrimination entre les voyelles [œ̃] et [e] à partir des lèvres semble être difficile, malgré l'information de la position LPC de la main. Une des raisons de cette difficulté peut être le fait que les lèvres, pendant les réalisations de la voyelle [œ̃], ne sont pas suffisamment ouvertes. Dans ce cas, la voyelle [œ̃] devrait être classée dans le troisième visème, et donc la position de main "gorge" devrait être complètement efficace pour désambiguïser les lèvres (voir table 7.2). Cette observation (lèvres pas assez ouvertes) peut être expliquée par le fait que notre codeuse LPC ne semble pas différencier les voyelles [œ̃] et [ɛ̃] avec les lèvres, bien que ces deux phonèmes sont codés avec deux positions LPC de la main différentes. En effet, notre codeuse LPC produit des réalisations similaires de formes aux lèvres pour les deux voyelles [œ̃] et [ɛ̃], comme le montre la petite distance entre les deux distributions correspondantes (voir figure 7.4, les positions "pommette" et "gorge"). Pour confirmer cette observation, nous avons appliqué une analyse de variance (ANOVA, voir chapitre 4 de la partie *état de l'art* pour la définition de cette analyse) à chacun des trois paramètres labiaux (facteurs). L'hypothèse nulle d'égalité des distributions n'peut être rejetée (voir figure 7.5), ce qui démontre que les données s'appuyant sur les paramètres A, B et S du contour interne des lèvres ne permettent pas statistiquement de différencier les voyelles [œ̃] et [ɛ̃]. Par ailleurs, il est important de noter que l'opposition entre ces deux voyelles tend à disparaître dans la langue française (Carton, 1974). Finalement, nous pouvons dire que l'ambiguïté entre les voyelles [œ̃] et [e] est maintenue à cause du choix de

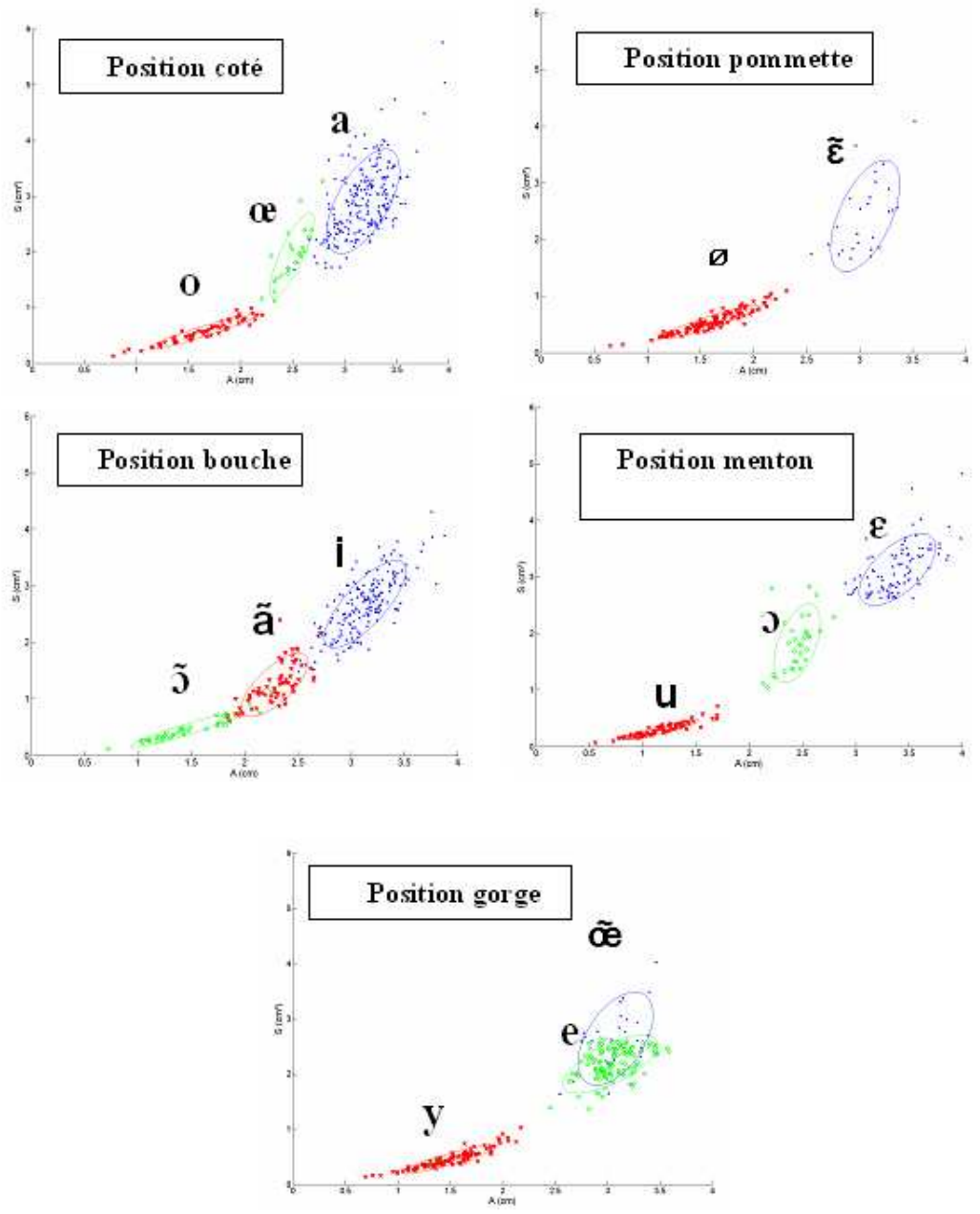

FIG. 7.4 – La distribution des voyelles à l'intérieur de chaque position LPC de la main dans le plan $[A(cm), S(cm^2)]$. Les ellipses représentent la dispersion des données à 2 écarts types par rapport à la moyenne.

codage de notre codeuse LPC. Dans ce cas, la discrimination complète nécessite des traitements de plus haut niveau.

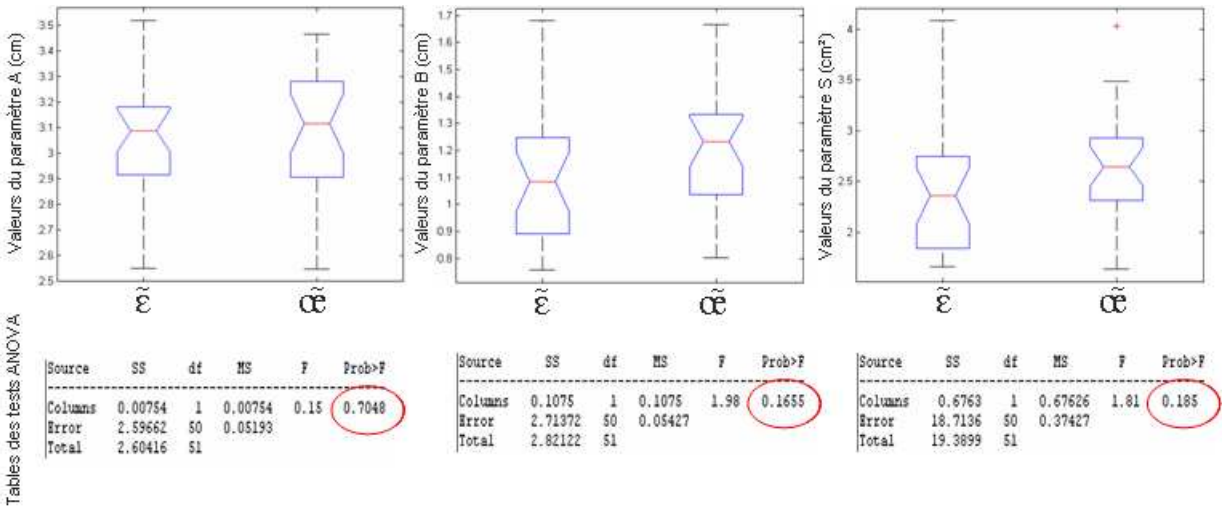

FIG. 7.5 – Trois tests ANOVA chacun sur un paramètre labial prouvant la non différenciation entre les deux voyelles [œ̃] et [ɛ̃] à partir des paramètres labiaux du contour interne.

7.1.2.2 Scores de classification à l'intérieur de chaque groupe LPC

D'après la section précédente, nous avons vu que connaître l'information labiale et la position LPC de la main permet de discriminer les voyelles à l'exception du cas des deux voyelles [œ̃] et [e]. De plus, comme nous l'avons vu aussi précédemment, la distribution des données suggère une modélisation Gaussienne pour la classification des voyelles. Aussi, nous avons construit un classifieur Gaussien tri-dimensionnel, s'appuyant sur les paramètres labiaux du contour interne des lèvres (A, B et S), pour chaque position LPC de la main. Pour cette classification, nous considérons deux sous-ensembles différents des mêmes voyelles : le premier est le sous-ensemble 1 décrit précédemment, le second (sous-ensemble 2) est composé des voyelles de la seconde répétition des 124 premières phrases. Pour ce deuxième sous-ensemble, nous présentons dans la table 7.3 le nombre d'occurrences de chacune des voyelles (au total 1105 voyelles). Comme pour le sous-ensemble 1, les paramètres labiaux des voyelles du sous-ensemble 2 sont extraits à l'instant $L2$ grâce à notre méthode de détection de la cible vocalique aux lèvres.

a	o	œ	ɛ̃	ø	i	ɑ̃	ɔ̃	ɛ	u	ɔ	œ̃	y	e
199	57	23	24	97	168	66	42	83	68	31	25	96	126

TAB. 7.3 – Sous-ensmble 2 : 1105 voyelles extraites de la seconde répétition des 124 premières phrases du corpus.

Les voyelles du sous-ensemble 1 sont utilisées pour estimer les différents paramètres des classifieurs (moyennes, écarts types et covariances), dans la phase d'apprentissage, tandis que les voyelles du sous-ensemble 2 sont réservées pour tester les performances de la classification. Nous présentons dans la table 7.4 les valeurs des moyennes et des écarts types pour chacune des 14 voyelles calculées à partir des paramètres labiaux des voyelles du sous-ensemble 1.

Dans un premier temps, la performance du classifieur est estimée sur les données du sous-

	$A(cm)$	$B(cm)$	$S(cm^2$
ø	1,65 (0,31)	0,48 (0,11)	0,55 (0,2)
œ	2,5 (0,14)	1,1 (0,22)	1,94 (0,51)
ɛ	3,4 (0,31)	1,32 (0,13)	3,13 (0,42)
ɔ	2,45 (0,15)	1,06 (0,23)	1,82 (0,47)
a	3,18 (0,23)	1,29 (0,23)	2,9 (0,63)
ɑ̃	2,28 (0,21)	0,79 (0,18)	1,25 (0,37)
e	3,03 (0,3)	1,02 (0,13)	2,18 (0,35)
i	3,13 (0,25)	1,22 (0,17)	2,67 (0,52)
ɛ̃	3,05 (0,22)	1,11 (0,26)	2,41 (0,66)
ø	1,64 (0,34)	0,49 (0,11)	0,57 (0,2)
ɔ̃	1,45 (0,35)	0,42 (0,11)	0,45 (0,21)
u	1,17 (0,24)	0,35 (0,1)	0,3 (0,12)
œ̃	3,08 (0,24)	1,2 (0,21)	2,64 (0,56)
y	1,49 (0,3)	0,42 (0,1)	0,46 (0,17)

TAB. 7.4 – Valeurs des moyennes et des écarts types (entre parenthèses) des paramètres labiaux A, B et S pour les données de la phase d'apprentissage.

ensemble 1 (voir figure 7.6). Le taux moyen de classification correcte est de 95,03%. Les erreurs sont partiellement causées par la non discrimination entre les voyelles [œ̃] et [e] de la position "gorge". Si nous considérons que ces deux voyelles constituent un seule voyelle le taux moyen s'élève à 97,74% ; ce qui veut dire que la faible discrimination entre ces deux voyelles cause une erreur de 2,71%. De plus, il faut rappeler que nous avons utilisé une méthode automatique pour détecter les instants d'atteinte de la cible vocalique aux lèvres. Cette méthode s'appuie aussi sur un système d'alignement automatique. Il est donc fortement probable que les résultats, obtenus à partir de ces deux méthodes, contiennent certaines erreurs. Ce qui peut probablement expliquer le reste des erreurs de la classification (environ 2,26%).

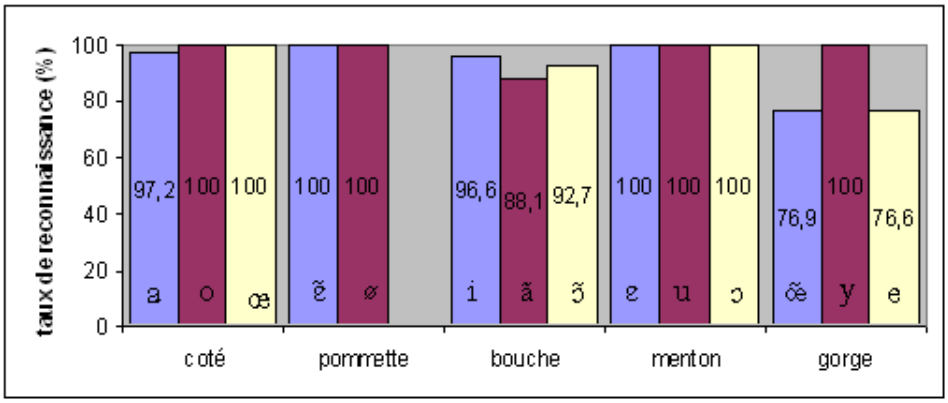

FIG. 7.6 – Taux d'identification correcte sur les données d'apprentissage.

Dans un second temps, la performance du classifieur est estimée sur les données de test (sous-ensemble 2). Le taux moyen d'identification correcte est de 89%. Comparé au taux de 95,03% obtenu sur les données d'apprentissage, ce taux, certes inférieur, semble montrer une performance satisfaisante pour cette classification. La différence entre les deux taux (sur les données d'apprentissage et de test) est de près de 6%. Ce dernier pourcentage quantifie des erreurs supplémentaires dues globalement à la fois à la précision de la transcription phonétique du système d'étiquetage du signal audio et aux erreurs du codage LPC. En effet, il est à rappeler que pour obtenir le taux d'identification de cette classification, les voyelles reconnues sont comparées aux voyelles fournies par l'étiquetage phonétique. Cet étiquetage, automatique, s'appuie sur un dictionnaire contenant la transcription phonétique de chaque mot du corpus. Or certains mots peuvent avoir plusieurs transcriptions (par exemple : le mot "autorisation" peut avoir comme transcription [otorizasjɔ̃] ou [otɔrizasjɔ̃], sur cette exemple le [o] peut être transcrit [ɔ]). Nous avons remarqué notamment que le système d'étiquetage confond dans certains cas les voyelles [o] et [ɔ], [e] et [ɛ], ainsi que [œ] et [ø]. Ces confusions sont ainsi une des causes des erreurs de classification. A ceci nous pouvons ajouter les erreurs dues au codage LPC de la codeuse. En effet, cette dernière peut articuler aux lèvres la voyelle [œ] comme [ø] alors qu'elle code avec la main la position "côté" (la voyelle [œ] appartient à la position LPC "côté" alors que [ø] est dans le groupe de la position "pommette"). Pour résumer, en plus des erreurs observées pour les données d'apprentissage, des erreurs supplémentaires (6%) sont observées notamment dans la distinction de [o] *vs.* [ɔ], [e] *vs.* [ɛ], ainsi que [œ] *vs.* [ø]. Finalement, il est important de noter que notre taux d'identification de 89% est légèrement inférieur par rapport à la performance de décodage par un humain de 94,8% obtenue dans l'évaluation du corpus (chapitre 5).

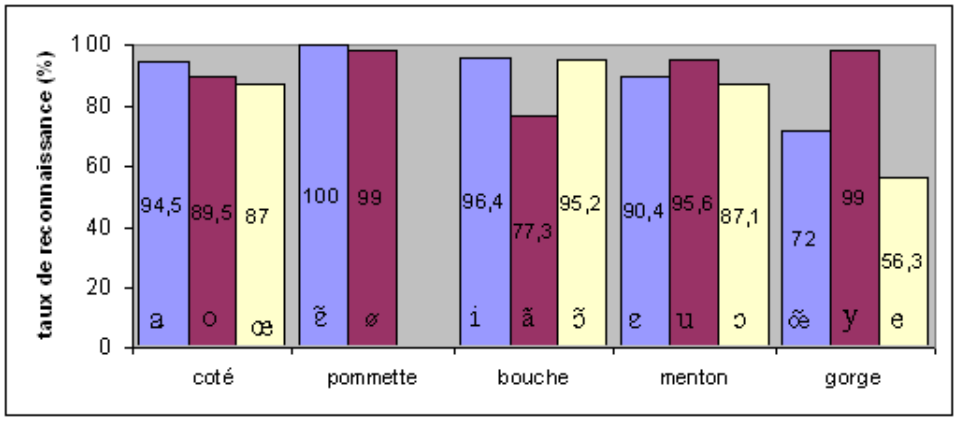

FIG. 7.7 – Taux d'identification correcte sur les données de test.

7.1.2.3 Scores de classification des visèmes

L'objectif de cette partie est d'évaluer les performances de reconnaissance des voyelles groupées en visèmes sans tenir compte de la position LPC de la main. En effet, nous avons vu précédemment que les voyelles peuvent être regroupées en trois visèmes. Ceci amène à modéliser les trois visèmes par des modèles gaussiens tri-dimensionnels s'appuyant sur les paramètres du contour interne

des lèvres (A, B et S) obtenus à partir des voyelles du sous-ensemble 1. La table 7.5 présente les valeurs des moyennes et des écarts types correspondant à ces modèles Gaussiens.

	$A(cm)$	$B(cm)$	$S(cm^2$
Visème 1	1,49 (0,35)	0,43 (0,12)	0,47 (0,2)
Visème 2	3,16 (0,28)	1,21 (0,21)	2,71 (0,61)
Visème 3	2,37 (0,2)	0,91 (0,25)	1,53 (0,53)

TAB. 7.5 – Valeurs des moyennes et des écarts types (entre parenthèses) des paramètres A, B et S pour les visèmes de voyelles obtenues sur les données d'apprentissage.

Dans cette expérience, les données d'apprentissage et de test s'appuient toujours sur les cibles labiales des voyelles obtenues à l'instant $L2$. Les voyelles du sous-ensemble 1 sont groupées en trois groupes et celles du sous-ensemble 2 en trois autres groupes différents. la table 7.6 détaille le nombre d'occurrences de chaque visème pour les deux phases apprentissage et test.

	Viseme 1	Viseme 2	Viseme 3
apprentissage	380	664	123
test	360	625	120

TAB. 7.6 – Nombre d'occurrences des visèmes dans les données d'apprentissage et de test.

Les résultats de la classification donnent un taux moyen d'identification correcte de 94,8% pour les données d'apprentissage. Ce taux baisse légèrement et atteint 92,6% pour les données de test. La figure 7.8 montre en détail les performances obtenues pour chaque visème.

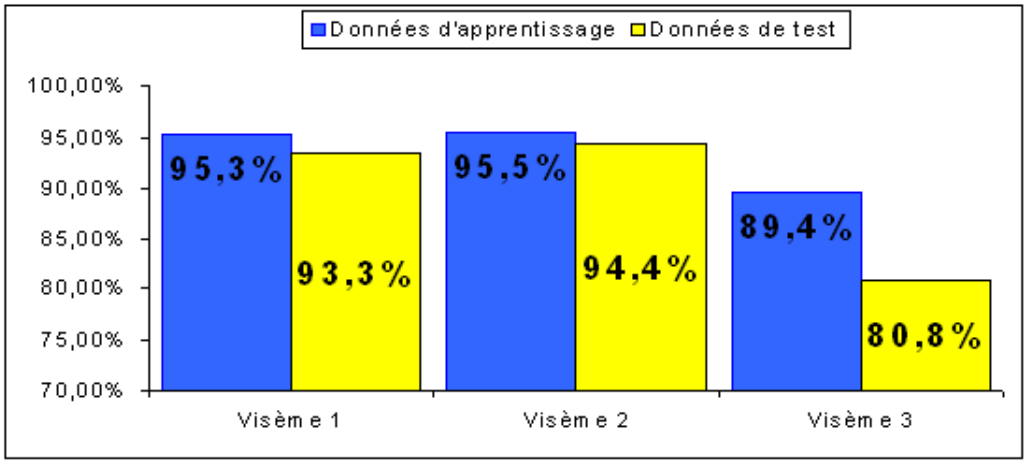

FIG. 7.8 – Taux d'identification correcte pour chaque visème.

Même si nous ne pouvons pas directement comparer des résultats à partir de classifications ayant des classes différentes à reconnaître, il est intéressant de noter que les taux obtenus dans le cas des visèmes montre une augmentation de 3,6% pour les données de test en comparaison

avec la première classification par groupe de position LPC de la main. Il est à noter que, dans ce cas, les voyelles [œ̃] et [e] sont considérées dans le même visème. Ainsi, la distinction entre ces deux voyelles, qui introduit 2,7% des erreurs dans notre classification précédente, n'est pas considérée. Ceci explique en partie l'augmentation du taux moyen.

7.1.3 Les points clés

En résumé, nos résultats sur les voyelles confirment les ambiguïtés de la lecture labiale à partir des formes labiales produites en parole continue. Dans ce contexte de production, nous avons, en plus, démontré l'efficacité de la composante manuelle du code LPC pour distinguer les différentes formes labiales. Nous avons aussi démontré qu'un simple classifieur Gaussien, s'appuyant sur les paramètres du contour interne des lèvres vue de face, peut être suffisant pour classifier les voyelles. En effet, quand la position LPC de la main est connue, de bonnes performances d'identification des voyelles sont obtenues (en moyenne 89%) avec seulement un seul instant de mesure, défini par notre méthode de segmentation des cibles vocaliques aux lèvres.
Par ailleurs, nous avons vu que les 14 voyelles du Français peuvent être catégorisées en trois visèmes. Un classifieur Gaussien permet aussi de classifier ces visèmes avec de bonnes performances. Dans ce cas, le nombre de modèles à construire passe de 14, pour la classification des voyelles par position LPC, à seulement 3, pour la classification de visèmes. Cependant, la modélisation en terme de visèmes ne conduit pas immédiatement à l'identification complète de la voyelle. Cette modélisation sera discutée dans le chapitre concernant la fusion des données pour la reconnaissance automatique de la voyelle.

7.2 Expérience 2 : modélisation et classification des syllabes consonne-voyelle (CV)

Cette section est consacrée à la reconnaissance des syllabes CV à partir de paramètres extraits des contours interne et externe des lèvres. Plus précisément, nous considérons les syllabes CV en terme de visèmes. Comme nous l'avons vu dans la section précédente, les voyelles peuvent être catégorisées en trois visèmes. Pour les consonnes, plusieurs ensembles de visèmes sont testés afin de déterminer le regroupement qui fournit le meilleur taux de reconnaissance avec la contrainte d'être compatible avec le regroupement de la composante manuelle du code LPC. Par ailleurs, nous considérons plusieurs intervalles pour caractériser au mieux les transitions des syllabes CV. De plus, nous introduisons et nous évaluons deux nouveaux paramètres relatifs aux pincements des lèvres supérieure et inférieure. Et enfin, la reconnaissance des syllabes CV est effectuée en utilisant les modèles HMMs (voir partie *état de l'art* pour leur définition).

7.2.1 Modélisation

7.2.1.1 Corpus de syllabes CV

Données : Pour cette expérience, nous avons construit notre corpus en extrayant toutes les syllabes CV des 267 phrases (répétées au moins deux fois) du corpus initial. Nous avons partagé ensuite les syllabes CV extraites en deux sous-ensembles. Le premier est constitué des syllabes CV des premières répétitions, et si une phrase est répétée plus de deux fois, ces répétitions supplémentaires sont également placées dans ce sous-ensemble. Ce sous-ensemble constitué est dédié à la phase d'apprentissage (ce sous-ensemble sera appelé dorénavant sous-ensemble apprentissage). Le second est composé des syllabes CV des secondes répétitions et il sera dédié à la phase de test (il sera appelé donc sous-ensemble de test). Numériquement, nous obtenons un sous-ensemble d'apprentissage composé de 2357 syllabes CV et un sous-ensemble de test de 1766 syllabes CV.

Nous rappelons que l'objectif de cette expérience est de reconnaître des syllabes CV en terme de visème à partir des paramètres labiaux. Ceci implique que les syllabes CV des deux sous-ensembles doivent être regroupées en visèmes. Cela veut dire qu'il faut définir pour chaque syllabe CV une classe visème en fonction des visèmes vocalique et consonantique. Pour les voyelles, nous disposons déjà, comme nous l'avons vu, de trois visèmes de voyelles. Par contre, la détermination des visèmes pour les consonnes ne semble pas aussi évidente. Ceci paraît en contradiction avec la littérature que nous avons développée dans la partie *état de l'art*. En effet, nous avons vu que de nombreuses études se sont concentrées plutôt sur la catégorisation des consonnes en visèmes que sur les voyelles. Pour ces études, les regroupements de consonnes résultant sont fructueux, mais en grande partie assez différents entre eux ; ce qui n'est pas le cas pour les voyelles. Les quelques regroupements des voyelles semblent être cohérents (par exemple, notre regroupement par rapport à celui de Robert-Ribès (1995). De plus, les regroupements des consonnes ont été effectués par le biais de la perception et dans des conditions expérimentales différentes. Notre regroupement des voyelles est obtenu comme un résultat de mesure direct des paramètres caractérisant les lèvres à l'instant où la cible vocalique est atteinte. Cette analyse ne peut être appliquée dans le cas des consonnes, car toutes les consonnes ne sont pas articulées aux lèvres, ce qui pose un problème de cible labiale pour les consonnes.

Pour avoir un regroupement des consonnes en visèmes, nous nous sommes inspirés du regroupement établi par Summerfield (1987) tout en l'adaptant au Français et en respectant la compatibilité avec les groupes de consonnes du système manuel du code LPC (nous notons ce regroupement : regroupement I). La table 7.7 présente les trois visèmes de voyelles et les dix visèmes de consonnes considérés. Pour chaque association différente d'un groupe de consonnes avec un groupe de voyelles, une classe CV est formée. Il y a autant de classes CV que le nombre de groupes de voyelles multiplié par le nombre de groupes de consonnes (dans ce cas, nous avons donc 30 classes CV).

Il est à noter que dans toute cette expérience les semi-voyelles [w,j,ɥ] ne sont pas considérées.

Intervalles considérés : Il reste maintenant à déterminer les transitions des syllabes CV ; c'est-à-dire, pour chaque syllabe CV, nous avons à délimiter l'intervalle caractérisant la transition

Visèmes de voyelles	Visèmes de consonnes
V1 = [y,ɔ̃,u,ø,o]	C1 = [p,b,m]
V2 = [a,ɛ̃,i,œ̃,e,ɛ]	C2 = [f,v]
V3 = [ɔ,ɑ̃,œ]	C3 = [d,n,t,k,g,ŋ]
	C4 = [ʃ,ʒ]
	C5 = [l]
	C6 = [s,z]
	C7 = [r]

TAB. 7.7 – Groupes de consonnes (regroupement I) et de voyelles pour la formation des classes des syllabes CV.

entre les cibles consonantique et vocalique aux lèvres. Nous nous sommes appuyés dans un premier temps sur l'étiquetage phonétique du signal audio pour déterminer ces transitions. Nous pouvons considérer tout l'intervalle correspondant à la réalisation acoustique de la syllabe CV (c'est-à-dire prendre l'intervalle $[A1, A3]$: [début acoustique de la consonne, fin acoustique de la voyelle]). Cependant, sur cet intervalle les paramètres labiaux de la syllabe CV subissent l'effet de la coarticulation (voir partie *état de l'art*). En effet, selon le phonème précédent la consonne de la syllabe CV, l'allure des paramètres labiaux de la consonne change. De même, selon le phonème suivant la voyelle de la syllabe CV, l'allure des paramètres labiaux de la voyelle change aussi. La figure 7.9 illustre ce phénomène pour le cas d'une syllabe avec une consonne occlusive (pas d'influence du contexte avant) et pour le cas d'une syllabe avec une consonne moins articulée aux lèvres.

Pour atténuer les effets de la coarticulation, il faut considérer l'intervalle entre la cible de la consonne aux lèvres (si cette cible existe) et la cible de la voyelle aux lèvres. Etant donné que toutes les consonnes ne sont pas articulées aux lèvres, nous ne pouvons pas déterminer une cible consonantique aux lèvres pour chaque consonne. Ainsi, nous considérons les deux instants *milieu* de la consonne et de la voyelle (AC et AV) comme des premières estimations des cibles consonantique et vocalique respectivement. Ces instants peuvent être obtenus à partir des instants $A1$, $A2$ et $A3$ (voir nomenclature Attina *et al.* (2004); Attina (2005) d'une syllabe CV de la façon suivante :

$$AC = \frac{(A1 + A2)}{2} \text{ et } AV = \frac{(A2 + A3)}{2}$$

Dans un second temps, nous ajustons cet intervalle en prenant l'instant cible aux lèvres de la voyelle ($L2$) à la place de l'instant AV. Avec cet instant, nous nous attendons que les résultats de reconnaissance s'améliorent. De plus, ceci nous permettra, en supposant que le $L2$ pourrait être obtenu sans utiliser les instants acoustiques, de faire un premier pas pour s'affranchir de l'étiquetage du signal acoustique, puisque ce signal ne serait pas disponible pour tout sujet.

Pour s'affranchir des étiquettes audio, nous remplaçons l'instant AC par l'instant $M2$. En effet, nous avons montré précédemment (chapitre précédent) que l'instant $M2$ se situe en moyenne dans la réalisation de la consonne, et précisément plus proche de l'instant milieu AC.

En résumé, nous effectuons les tests de reconnaissance des syllabes CV sur trois intervalles :

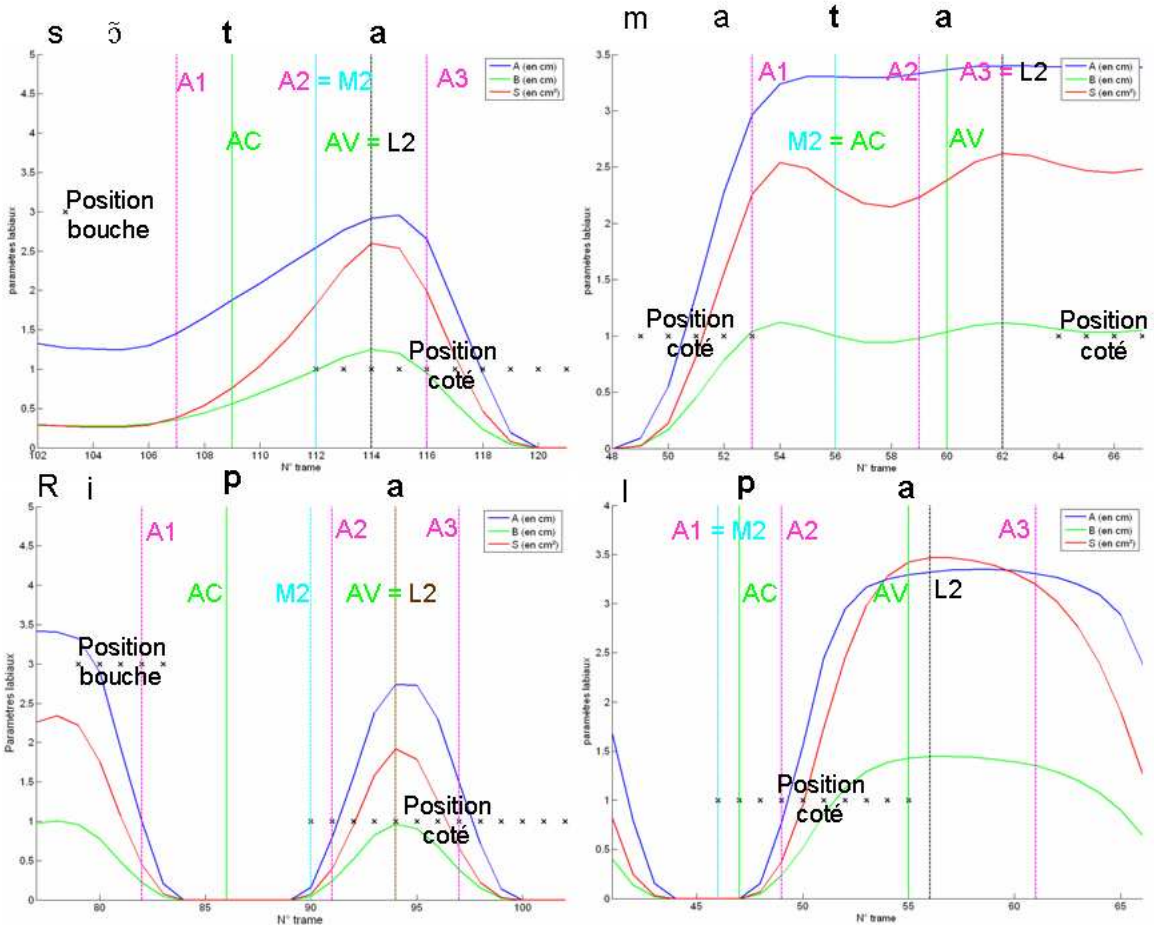

FIG. 7.9 – Effets de la coarticulation sur la forme aux lèvres de la consonne d'une syllabe CV caractérisée par des paramètres labiaux différents d'un contexte à l'autre. En haut : la consonne [t] de la syllabe [ta] est modifiée par le contexte avant (à gauche [õ] et à droite [a]), ce qui donne des allures différentes des paramètres labiaux (ici du contour interne) de la syllabe entre les instants $A1$ et $A3$. En bas : la syllabe [pa] est moins influencée par le contexte avant. Notons dans ces figures la variabilité des instants acoustiques par rapport au début et la fin de la transition labiale d'une syllabe.

- $[AC, AV]$
- $[AC, L2]$
- $[M2, L2]$

Pour chaque syllabe CV, l'évolution temporelle des six paramètres des contours interne et externe des lèvres (A, B, S,A', B', S'), extraits tous les 20 ms, est considérée.

7.2.1.2 Phase d'apprentissage

Cette phase consiste à estimer les paramètres de chacune des classes HMM modélisant les classes des visèmes CV. Si nous disposons d'un nombre N de visèmes CV, il nous faut donc apprendre N modèles HMMs. Dans cet objectif, nous disposons des séquences temporelles des

paramètres labiaux des syllabes (entre les instants de début et de fin définis précédemment) du sous-ensemble d'apprentissage. Pour chaque classe visème CV, les séquences correspondant aux syllabes de cette classe sont mises les unes après les autres sous forme de matrice. Ainsi, nous construisons pour chaque classe une matrice d'observation de la forme présentée dans la figure 7.10.

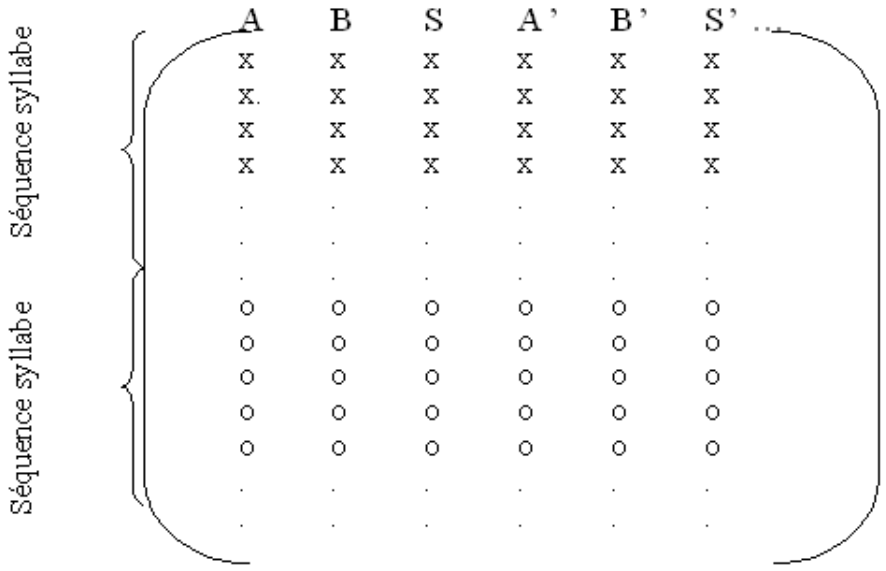

FIG. 7.10 – Disposition des séquences des syllabes CV pour une classe donnée.

Nous utilisons ensuite les outils Matlab crées par Cappé (2001) pour la phase d'apprentissage. Il faut avant tout initialiser certains paramètres (voir définition des modèles HMMs dans la partie *état de l'art*). Nous choisissons tout d'abord de fixer **le nombre d'états de nos modèles HMMs à trois**. Avec ce choix, nous supposons que la cible consonantique est modélisée par un état, la cible vocalique par un second état et la transition entre ces deux cibles par un troisième état (voir figure 7.11). Nous optons aussi pour des modèles HMMs gauche-droite avec une matrice des probabilités de transition initialisée de manière aléatoire et une matrice des probabilités d'observation initialisée en affectant la valeur 1 au premier état et 0 pour les deux autres (on considère qu'au début de chaque séquence nous sommes dans l'état 1). Les observations sont modélisées par des distributions mono-gaussiennes (une matrice des moyennes et une matrice des covariances).

Par ailleurs, l'algorithme d'apprentissage est itératif[2]. Ainsi, il faut fixer un nombre d'itération qui soit suffisant pour que l'algorithme converge. Après plusieurs tests, il nous semble que 15 itérations soient appropriées pour notre cas.

A l'issue de cette phase nous obtenons pour chaque classe visème CV un modèle HMM défini par :

- le nombre d'états fixé à 3,
- une matrice de transition contenant les probabilités des transitions entre les 3 états,

[2] *voir définition des modèles HMM.*

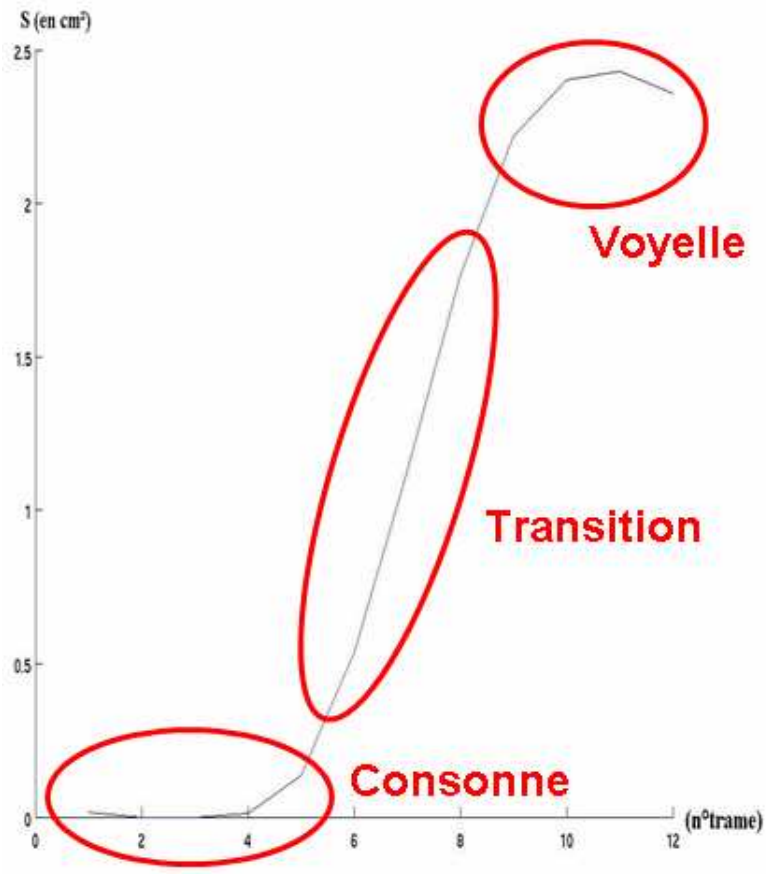

FIG. 7.11 – 3 états pour modéliser la transition pour une syllabe CV [pa.

– une matrice des moyennes et une matrice des covariances pour le calcul des probabilités d'observations,

La figure 7.12 illustre sur un exemple de classe HMM, pour le paramètre S, les 3 gaussiennes modélisant les 3 états.

7.2.1.3 Phase de test

Dans cette phase, nous utilisons aussi les outils de Cappé (2001) et plus précisément l'algorithme de Viterbi (voir définition des modèles HMMs). Nous présentons en entrée de cet algorithme une séquence d'observation correspondant à une syllabe CV et nous récupérons en sortie la classe HMM reconnue. Les séquences d'observations sont disposées de la même manière que celles de l'apprentissage.

Pour obtenir les performances de reconnaissance, nous comparons toute classe HMM reconnue pour chaque séquence d'observation avec la classe de référence correspondante. Cette dernière est obtenue en affectant manuellement les séquences d'observations à des classes de référence correspondantes. A partir de cette comparaison nous pouvons calculer un taux de reconnaissance global (ou moyen) sur les 1766 syllabes du sous-ensemble de test. Ce taux est considéré comme le nombre d'observations bien reconnues sur le nombre total des observations.

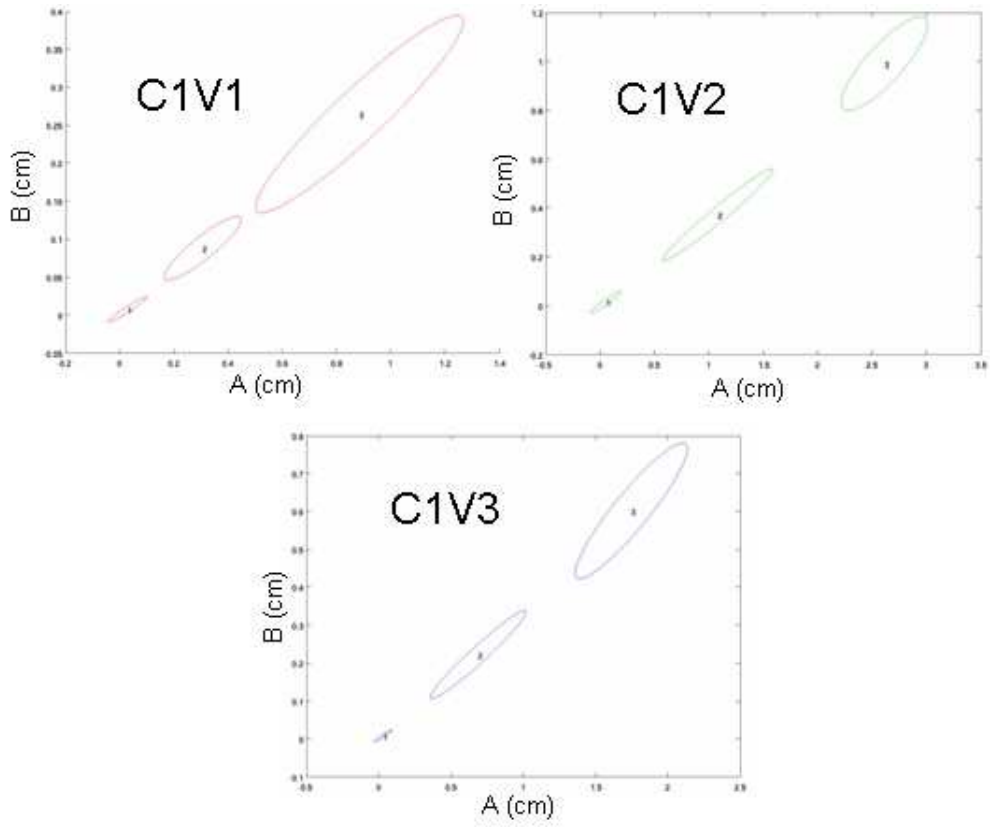

FIG. 7.12 – Exemple de 3 distributions Gaussiennes (ellipses de dispersion dans le plan (A, B), écart type de 1.5) des 3 états du modèle HMM obtenu pour la classe visème composée par le groupe C1 en contexte des trois visèmes de voyelles (V1, V2, V3). Les séquences d'observations sont prises sur l'intervalle $[ACAV]$.

Dans le but d'analyser les erreurs possibles, nous construisons pour chaque test une matrice de confusion. Cette matrice est carrée de dimension égale au nombre de classes. Les classes de référence sont placées en colonnes tandis que les classes reconnues en lignes. En diagonale, nous pouvons lire le nombre des classes correctement reconnues et hors diagonale les erreurs de reconnaissance.

7.2.1.4 Changement du regroupement des consonnes

Le regroupement des consonnes inspiré de Summerfield (1987) que nous avons choisi au départ pour construire nos visèmes de syllabes CV comporte dix groupes dont certains sont composés d'une seule consonne. Etant donné que certaines de ces consonnes ne peuvent se distinguer à partir des lèvres seulement, il est possible donc de les mettre ensemble dans un même groupe. Ceci implique évidement une réduction du nombre de groupes de consonnes et par la suite du nombre de classes de visèmes des syllabes CV. Cependant, il faut des critères supplémentaires pour un regroupement bien approprié. Pour les consonnes articulées aux lèvres ([p,m,b,f,v,ʃ,ʒ]), le seul critère est la discrimination à partir des paramètres labiaux. Pour les autres consonnes ([d,n,t,s,z,r,k,g,l,ŋ]), nous utilisons comme critère de discrimination le lieu d'articulation de la

langue, qui peut influencer la forme aux lèvres lors de la production de la consonne. Mais, il faut aussi garder en mémoire que tout regroupement fait doit impérativement être compatible avec les groupes de consonnes du système manuel du code LPC, qui par association, permettrait d'identifier une consonne unique. Un groupe de consonne ne doit en aucun cas contenir deux consonnes identiquement codées avec la main.

Si nous dressons une table regroupant les consonnes suivant le lieu d'articulation, nous obtenons déjà un premier regroupement. La table 7.8 présente ce premier regroupement.

Bilabial	Labiodental	Dental	Palatal	Vélaire
[p]	[f]	[d]	[ʃ]	[k]
[m]	[v]	[n]	[ʒ]	[g]
[b]		[t]	[ɲ]	[r]
		[s]		
		[z]		
		[l]		

TAB. 7.8 – Classification des consonnes selon le lieu d'articulation de la langue.

Sur cette table, nous observons qu'il y a un conflit de compatibilité avec le système manuel du code LPC dans le groupe des consonnes palatales. En effet, les deux consonnes [ʃ] et [ɲ] de ce groupe sont codées avec la même configuration de la main (configuration n°6). Il faut donc déplacer une de ces consonnes dans un autre groupe. Comme les deux consonnes [ʃ] et [ʒ] sont plus articulées aux lèvres que la consonne [ɲ] et qu'elles se ressemblent beaucoup, nous avons choisi de déplacer la consonne [ɲ]. Dans ce cas, deux possibilités se présentent : mettre cette consonne dans le groupe des consonnes vélaires ou dans le groupe des consonnes dentales. Les consonnes de ces deux groupes ne sont pas articulées aux lèvres tout comme la consonne [ɲ]. Dans le cas du deuxième groupe (dental), nous nous retrouvons encore avec un problème de compatibilité avec le système manuel du code LPC : les deux consonnes [l] et [ɲ] sont codées avec la même configuration de la main. Dans ce cas précis, nous déplaçons la consonne [l] dans le groupe des vélaires et nous gardons la consonne [ɲ] dans le groupe des dentales. Ceci est motivé par le fait que la consonne [l] n'était que très rarement regroupée avec les consonnes [d,n,t] dans les résultats des expériences sur les visèmes que nous avons présenté dans la partie *état de l'art*. Alternativement, la consonne [ɲ] est souvent mise avec ces consonnes (voir regroupements de Gentil (1981); Jutras *et al.* (1998). Nous construisons ainsi avec ce déplacement deux regroupements des consonnes que nous considérons lors de nos tests afin de déterminer celui qui donne le meilleur taux de reconnaissance. Ces deux regroupements des consonnes seront notés respectivement regroupement II et regroupement III et sont présentés dans la table 7.9.

7.2.1.5 Le pincement en renfort

Pour une meilleure caractérisation des visèmes des syllabes CV, il est possible d'ajouter aux paramètres utilisés (A, B, S,A', B', S') de nouveaux paramètres extraits à partir des contours des lèvres et qui peuvent caractériser des formes particulières de certaines consonnes. Dans

Regroupement II des consonnes	Regroupement III des consonnes
C1=[p,b,m]	C1=[p,b,m]
C2=[f,v]	C2=[f,v]
C3'=[d,n,t,s,z,l]	C3'=[d,n,t,s,z,ŋ]
C4'=[ʃ,ʒ]	C4'=[ʃ,ʒ]
C5'=[r,k,g,ŋ]	C5'=[r,k,g,l]

TAB. 7.9 – Changement du regroupement des consonnes : regroupements II et III.

cet objectif, nous introduisons deux paramètres : le pincement des lèvres supérieure ($Bsup$) et inférieure ($Binf$). Ces paramètres devraient permettre de mettre en relief le phénomène d'éversion des lèvres lors de la locution de consonnes telles que [ʃ] et [ʒ], ou encore le pincement de la lèvre supérieure pour les consonnes [f] et [v]. Ces deux paramètres ($Bsup$) et ($Binf$) se calculent respectivement comme les épaisseurs selon la verticale des lèvres supérieure et inférieure (voir figure 7.13).

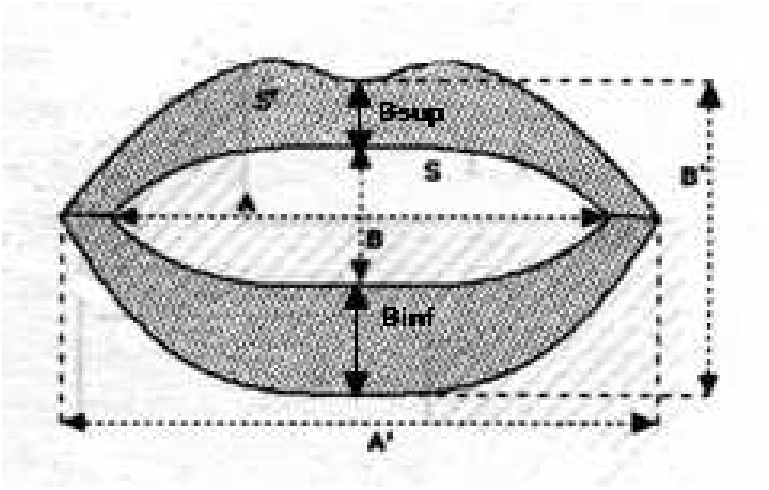

FIG. 7.13 – Illustration du pincement des lèvres supérieure et inférieure (respectivement Bsup et Binf).

7.2.2 Résultats

Nous avons effectué différents tests de reconnaissance suivant trois paramètres : le regroupement des consonnes, l'intervalle de transition caractérisant les syllabes CV et les paramètres labiaux utilisés (utilisation du pincement ou non). Dans un premier temps, nous avons considéré l'intervalle acoustique [AC,AV] pour les syllabes CV et nous avons testé la reconnaissance des syllabes CV en terme de visème pour les trois cas de regroupement des consonnes. Dans un second temps, nous avons ajouté les deux paramètres du pincement aux six paramètres des contours interne et externe des lèvres pour améliorer la discrimination de certaines consonnes. Enfin, en sélectionnant le regroupement qui donne le meilleur taux de reconnaissance, nous avons testé

l'influence du changement de l'intervalle d'observation des syllabes CV sur ces performances.

7.2.2.1 Avec l'intervalle [*AC*, *AV*]

Regroupement I des consonnes : Avec le regroupement I (celui inspiré de Summerfield (1987), **le taux moyen de reconnaissance est de 55,95%**. Ce faible taux implique une grande confusion entre les visèmes syllabiques. Entre les visèmes de voyelles, la confusion est estimée, à partir seulement des paramètres du contour interne des lèvres, à peu près à 7% (le taux obtenu précédemment pour la classification des visèmes de voyelles est de 92,6%). Ceci implique que la plus grande partie de la confusion entre les visèmes syllabiques vient en fait des confusions entre les consonnes. De ce fait, nous avons effectué des tests de reconnaissance avec des sous-ensembles de visèmes des consonnes. C'est-à-dire, nous avons testé différentes combinaisons des groupes de consonnes afin de déterminer le (s) groupe (s) de consonnes qui cause (nt) le plus de confusion. La table 7.10 présente les taux obtenus pour ces différents tests.

Combinaisons	Taux de reconnaissance des visèmes de syllabes CV (en %)
Cas 1 : C1, C3	91,18
Cas 2 : **cas 1** + C2	86,6
Cas 3 : **cas 1** + C4	80,8
Cas 4 : **cas 1** + C6	70,86
Cas 5 : **cas 2** + C4	77,29
Cas 6 : **cas 2** + C6	70,54
Cas 7 : **cas 3** + C6	65,22
Cas 8 : **cas 6** + C6	65,79
Cas 9 : **cas 8** + C5	62,1
Cas 10 : **cas 8** + C7	61,6
Cas 11 : **regroupement I**	55,95

TAB. 7.10 – Taux de reconnaissance des visèmes de syllabes CV pour différentes combinaisons des consonnes.

A partir de ces résultats, nous constatons que le taux de reconnaissance diminue de façon progressive à chaque fois qu'un groupe de consonnes est ajouté. Ceci s'explique par le fait qu'en ajoutant un groupe de consonne, le nombre de classes des syllabes CV augmente (une augmentation par trois classes pour chaque groupe ajouté) et dans le même sens, les confusions entre les classes syllabiques se multiplient. Ces confusions ne sont pas pour autant de la même ampleur selon le groupe de consonnes ajouté. En effet, si nous considérons le cas 1 comme notre point de départ, l'ajout du groupe C6 (cas 4) fait chuter le taux de reconnaissance de plus de 20% alors qu'avec le groupe C2 (cas 2), la chute n'est que de 5%. Ceci veut dire que le groupe C6, constitué des consonnes [s,z] introduit plus de confusion que le groupe C2 (consonnes [f,v]). C'est un fait tout à fait attendu puisque dans le cas 4, les consonnes [s,z] ne sont pas articulées aux lèvres, ce qui fait qu'elles sont confondues en grande partie avec les consonnes du groupes C3, qui elles mêmes ne sont pas articulées aux lèvres. Si nous passons à une reconnaissance à 12

classes de syllabes CV (cas 5, 6 et 7), la constatation précédente est aussi vérifiée. Pour le cas 5, nous remarquons que le taux de reconnaissance est pratiquement similaire à celui du cas 3 ; ce qui signifie que les groupes C2 et C4 sont bien discriminés, et que les erreurs sont dues à la confusion entre les groupes C3 et C4. Il en est de même pour le cas 6 où le résultat est proche du cas 4 (différence de moins de 1%). Pour le cas 7, le résultat obtenu est très inférieur à celui obtenu dans les cas 3 et 4. Ceci nous ramène à supposer qu'en plus des confusions qui existent entre les groupes C3 et C6, il y a confusion entre les groupes C4 et C6.

Avec 15 classes de CV (cas 8), les cinq groupes testés dans les cas précédents sont confrontés. Nous constatons cette fois que le taux global reste à peu près identique à celui du cas 7. Le groupe de consonnes C2 est bien discriminé par rapport aux autres groupes. En effet, les consonnes de ce groupe sont bien articulées aux lèvres et donc peuvent être bien discriminées à partir des paramètres des contours des lèvres.

Il reste donc à analyser les deux cas 9 et 10 où nous ajoutons séparément les groupes C5 et C7 constitués respectivement des consonnes [l] et [r]. Dans les deux cas, le taux de reconnaissance diminue par rapport au cas 8 avec presque la même ampleur (3,78% pour le cas 9 et 4,28%). Cependant, quand les groupes C5 et C7 sont ajoutés, la diminution du taux de reconnaissance est plus grande et approche les 10%. Ceci montre une confusion entre les deux consonnes [l] et [r] de l'ordre de 6%. Cette confusion peut s'expliquer par une forme labiale semblable pour ces deux consonnes. Il est donc préférable de les mettre dans un même groupe.

En conclusion de ces tests, il est clair que l'ordonnancement des consonnes engendre un fort taux d'erreur. Voyons ce qu'un nouveau regroupement des consonnes s'appuyant sur le lieu d'articulation de la langue donne comme résultat.

Regroupements II et III des consonnes : Si nous nous appuyons sur les regroupements des consonnes issus du classement articulatoire (selon le lieu de l'articulation de la langue), nous obtenons les résultats de classification des visèmes des syllabes CV suivants :

- **Avec le regroupement II des consonnes, le test de reconnaissance donne un taux de 65,29%**. Dans ce cas, nous avons une forte confusion entre les groupes C3' et C5' due essentiellement à la ressemblance des réalisations aux lèvres des consonnes [l] et [r] constatée précedemment.
- **Avec le regroupement III, le taux de reconnaissance est de 75,93%**. D'abord, nous constatons une nette augmentation (près de 20%) par rapport au taux obtenu avec le regroupement I des consonnes (avec le même nombre de consonnes) ; ce qui confirme nos constatations précédentes sur les confusions entre certaines consonnes. Ensuite, il est à remarquer aussi que par rapport au regroupement II, le taux augmente de près de 10%. Cette augmentation s'explique par le fait que dans le regroupement III, les deux consonnes [l] et [r] sont placées dans le même groupe ; ainsi, la confusion entre ces deux consonnes n'intervient pas.

Nous pouvons conclure de ces premiers tests que le regroupement des consonnes a une influence majeure sur les taux de reconnaissance des visèmes syllabiques CV. La plus grande partie des erreurs vient des confusions entre les consonnes non articulées aux lèvres. La réalisation aux lèvres de ces consonnes est généralement influencée par les voyelle qui précède ou/et celle qui suit.

Mettre ces consonnes dans un même groupe, certes, donne un taux meilleur de reconnaissance (77,8%), mais la contrainte que le regroupement soit compatible avec le système manuel du code LPC oblige de partager ces consonnes au moins dans deux groupes. De ce fait, le regroupement III est le regroupement à considérer dans la suite de cette expérience.

Le bénéfice du pincement : Ayant fixé un regroupement de consonnes, nous pouvons maintenant ajouter les deux paramètres du pincement dans l'objectif d'améliorer nos taux de reconnaissance. Avec ses deux paramètres en plus, nous disposons donc de huit paramètres labiaux.

Le taux de reconnaissance global obtenu pour ce test est de 80,3%. Ce résultat montre que l'information du pincement augmente les performances de reconnaissance de plus de 4% en moyenne. Cette augmentation varie selon les groupes de consonnes. La table 7.11 présente les taux de reconnaissance en fonction du groupe de consonnes avec et sans les paramètres du pincement.

	CV avec cons1	CV avec cons2	CV avec cons3'	CV avec cons4'	CV avec cons5'
Taux de reconnaissance sans avec le pincement	85.5%	70.5%	81.7%	64.5%	80.1%
Taux de reconnaissance avec le pincement	89.9%	75.8%	79.4%	74.8%	81.7%

TAB. 7.11 – Table de comparaison de l'effet du pincement pour la reconnaissance des syllabes CV avec les différents groupes de consonnes.

Sur cette table, il est intéressant de noter que les syllabes contenant une consonne du groupe C4' voient leur taux de reconnaissance augmenter de plus de 10%. Et bien que les consonnes du groupe C3' ne soient pas articulées aux lèvres, leur taux de reconnaissance augmente également (5%). Ces résultats montrent que le pincement des lèvres apporte une information supplémentaire à la discrimination des syllabes CV.

Le test avec le regroupement III et les paramètres du pincement donne globalement de bons résultats de reconnaissance de syllabes que la consonne soit articulée ou non aux lèvres. Cependant, les taux n'atteignent pas les 100% ; des erreurs subsistent toujours. Ainsi, pour analyser ces erreurs tant pour les consonnes que pour les voyelles, nous avons construit la matrice de confusion entre les classes de syllabes CV pour ce test (voir table 7.12).

D'une part, **les confusions entre les groupes de consonnes et de voyelles expliquent les erreurs avec des proportions différentes pour chaque syllabe CV**. En effet, nous constatons, pour les classes composées d'une consonne du groupe C1 (groupe des consonnes occlusives), une erreur de 10,1%. Cette dernière est due principalement à la confusion entre les groupes de voyelles (9,4%). Ceci confirme la pertinence des paramètres labiaux que nous utilisons pour décrire l'occlusion bilabiale. En revanche, pour les autres classes, les erreurs sont dues en grande partie aux confusions entre les consonnes. Par exemple, pour les classes de référence (classes attendues à la reconnaissance) avec une consonne du groupe C3, les erreurs causées

Syllabes CV reconnues \ Syllabes CV de référence	C1V1	C1V2	C1V3	C2V1	C2V2	C2V3	C3V1	C3V2	C3V3	C4V1	C4V2	C4V3	C5V1	C5V2	C5V3
C1V1	**85**	0	3	0	0	0	0	0	0	0	0	0	0	0	0
C1V2	0	**137**	1	0	0	0	0	1	0	0	0	0	0	1	0
C1V3	6	18	**45**	0	0	0	1	0	0	0	0	0	0	1	0
C2V1	0	0	0	**22**	1	1	11	0	0	0	0	0	6	0	0
C2V2	0	0	0	0	**77**	5	2	3	1	0	2	0	0	1	0
C2V3	0	0	0	2	0	**5**	0	1	0	0	0	0	0	0	0
C3V1	0	0	0	7	1	0	**171**	2	1	18	0	0	23	0	0
C3V2	0	0	0	0	3	0	1	**235**	0	0	2	1	0	13	0
C3V3	0	0	0	0	1	0	13	10	**64**	0	0	0	2	2	2
C4V1	0	0	0	1	0	0	36	0	3	**23**	0	0	26	0	0
C4V2	0	0	0	0	0	0	0	3	1	1	**67**	1	0	7	0
C4V3	0	0	0	0	0	0	0	5	6	0	2	**11**	0	1	0
C5V1	1	0	1	2	1	1	23	1	0	4	0	0	**122**	1	2
C5V2	0	0	0	0	1	0	0	22	0	0	1	0	0	**314**	0
C5V3	0	0	0	0	0	0	0	1	1	2	0	0	12	7	**40**
total	**92**	**155**	**50**	**34**	**85**	**12**	**258**	**284**	**77**	**48**	**74**	**13**	**191**	**348**	**44**
Taux de reconnaissance (%)	92%	88%	90%	65%	91%	42%	66%	83%	83%	48%	91%	85%	64%	90%	91%

TAB. 7.12 – Matrice de confusion pour le test de reconnaissance avec le regroupement III des consonnes et avec les paramètres du pincement en plus des six paramètres des contours interne et externe des lèvres. L'intervalle des observations est $[AC,AV]$. En colonne la référence, en ligne le résultat de classification.

par la confusion entre les groupes de consonnes atteignent 20% ; alors que celles causées par la confusion entre les groupes de voyelles n'est que de 4,2%.

D'autre part, **le geste d'ouverture est bien détecté**. En effet, d'après la matrice de confusion, les transitions d'une consonne vers une voyelle non arrondie (groupe V2) ou semi-arrondie (groupe V3) semble apporter des informations pertinentes à la discrimination des syllabes CV. Par contre, dans le contexte des voyelles du groupe V1, c'est-à-dire les voyelles arrondies, les syllabes CV sont largement moins reconnues à l'exception du cas de la classe C1V1. Notons que l'arrondissement des lèvres est bien identifié puisqu'il y a peu d'erreurs sur l'identification des voyelles seules du groupe V1. Nous pouvons en conclure que la transition d'une consonne vers une voyelle arrondie (V1)ne semble pas suffisante pour discriminer la consonne en utilisant les paramètres labiaux choisis. **Ce résultat est cohérent avec l'effet de coarticulation des voyelles arrondies** qui modifient la forme aux lèvres de la consonne qui précède d'une façon qui peut aller jusqu'à masquer cette consonne (Abry et Boë, 1986).

Enfin, le résultat global de 80,3% doit être comparé avec le taux de reconnaissance obtenu sur les données d'apprentissage. Ce taux atteint seulement 81,37%. Comme nous l'avons évoqué précédemment, il est facile d'attribuer la majeure partie des erreurs au seul fait que les consonnes des groupes C3' et C5' ne sont pas articulées aux lèvres. Ainsi, si nous fusionnons ces deux groupes en un seul groupe, le taux global de reconnaissance ne change pratiquement pas (80,58%). Mais, si ces deux groupes ne sont pas considérés dans le test, le taux de reconnaissance

augmente clairement (90,41%). Ce dernier obtenu dans un test de reconnaissance des syllabes CV est similaire au taux obtenu par la classification gaussienne des voyelles vue précédemment (89% pour la classification des voyelles par positon LPC et 92,6% pour la classification des visèmes de voyelles). Ainsi, l'erreur résiduelle peut être expliquée en grande partie par la confusion entre les voyelles.

Il est à noter qu'il ne faut pas tenir compte des résultats obtenus pour les deux classes C2V3 et C4V3. Ces résultats ne sont pas significatifs car les deux classes ont un effectif très faible.

7.2.2.2 Avec l'intervalle [AC, $L2$]

Comme le résultat obtenu en considérant le regroupement III des consonnes avec les huit paramètres labiaux (avec le pincement donc) est le meilleur, nous cherchons maintenant à s'affranchir des étiquettes audio, dans l'hypothèse où ces dernières n'interviennent pas dans la détection des instants labiaux, ce qui n'est pas notre cas. Dans un premier temps, nous remplaçons l'étiquette AV (milieu acoustique de la voyelle), qui définissait la fin de la transition de la syllabe, par l'instant $L2$ qui correspond à la cible labiale vocalique. Avec ce remplacement, nous espérons une amélioration du taux de reconnaissance puisque l'instant $L2$ est plus précis pour décrire la cible vocalique aux lèvres que l'instant AV. Ainsi, nous construisons de nouvelles séries d'observations (une pour l'apprentissage et une autre pour le test) et nous procédons à l'apprentissage des modèles et au test de reconnaissance.

Nous obtenons un taux de reconnaissance global de 79,28%. La différence avec le taux obtenu dans le test précédent (80,3%) n'est pas significative. Certes, ce résultat montre que l'instant $L2$ permet de s'affranchir d'une première étiquette audio (AV). Cependant, contrairement à nos attentes, nous n'obtenons pas d'amélioration sur le taux global.

En regardant le résultat des classes au cas par cas (table 7.13), nous pouvons dire que les classes avec les groupes C3' et C5', qui sont peu articulées aux lèvres, présentent des taux meilleurs avec l'instant $L2$ en contexte de voyelle arrondie (groupe V1), alors qu'avec les autres groupes de consonnes les taux de reconnaissance diminuent. En revanche, en contexte de voyelle non arrondie (groupe V2) les tendances s'inversent (les résultats pour les groupes C3' et C5' sont moins bons, et pour les autres groupes sont meilleurs). En contexte de voyelle semi-arrondie (groupe V3) les taux sont moins bons pour tous les groupes de consonnes à l'exception du groupe C1 (sûrement dû au fait de l'occlusion bilabiale bien différenciée).

7.2.2.3 Avec l'intervalle [$M2$, $L2$]

Dans un second temps, nous remplaçons l'étiquette acoustique AC (milieu acoustique de la consonne) par l'instant $M2$. Nous reconstruisons de nouveau nos séries d'observations (pour l'apprentissage et pour le test) avec la nouvelle durée [$M2$,$L2$] et **nous obtenons un taux global de reconnaissance de 71,29%**. Ce taux est inférieur à celui obtenu précédemment avec l'intervalle [AC,$L2$]. Cette baisse du taux global est due à une diminution de la reconnaissance de chaque classe. La table 7.14 présente la matrice de confusion de ce test.

Pour presque toutes les classes, nous remarquons que les erreurs de reconnaissance augmentent plus entre les groupes de consonnes qu'entre les groupes de voyelles. En effet, nous

Syllabes CV reconnues	Syllabes CV de référence														
	C1V1	C1V2	C1V3	C2V1	C2V2	C2V3	C3V1	C3V2	C3V3	C4V1	C4V2	C4V3	C5V1	C5V2	C5V3
C1V1	**81**	0	1	0	0	0	0	0	0	0	0	0	0	0	0
C1V2	0	**141**	0	0	0	0	0	1	0	0	0	0	0	2	0
C1V3	8	14	**48**	0	0	0	1	0	0	0	0	0	0	0	0
C2V1	0	0	0	**21**	0	1	16	0	0	0	0	0	8	0	0
C2V2	0	0	0	1	**78**	6	3	3	0	0	1	0	0	2	0
C2V3	0	0	0	0	1	**3**	3	2	0	0	0	0	0	0	0
C3V1	0	0	0	9	1	0	**184**	1	5	18	0	0	33	0	0
C3V2	0	0	0	0	3	0	1	**226**	0	0	1	1	0	21	0
C3V3	1	0	0	0	0	1	7	11	**54**	0	0	3	1	1	2
C4V1	0	0	0	1	0	0	24	1	3	**22**	0	0	11	0	1
C4V2	0	0	0	0	1	0	0	8	1	1	**68**	1	0	12	0
C4V3	0	0	0	0	0	0	0	4	11	0	2	**8**	0	0	0
C5V1	2	0	1	2	0	1	19	0	2	7	0	0	**128**	1	3
C5V2	0	0	0	0	1	0	0	28	0	0	2	0	0	**314**	0
C5V3	0	0	0	0	0	0	1	0	1	0	0	0	10	9	**38**
total	**92**	**155**	**50**	**34**	**85**	**12**	**259**	**285**	**77**	**48**	**74**	**13**	**191**	**362**	**44**
taux de reconnaissance (%)	88%	91%	96%	62%	92%	25%	71%	79%	70%	46%	92%	62%	67%	87%	86%

TAB. 7.13 – Matrice de confusion pour le test de reconnaissance avec le regroupement III des consonnes et avec les paramètres du pincement en plus des six paramètres des contours interne et externe des lèvres. L'intervalle des observations est $[AC,L2]$.

Syllabes CV reconnues	Syllabes CV de référence														
	C1V1	C1V2	C1V3	C2V1	C2V2	C2V3	C3V1	C3V2	C3V3	C4V1	C4V2	C4V3	C5V1	C5V2	C5V3
C1V1	**73**	0	5	1	0	0	0	0	0	0	0	0	3	0	0
C1V2	0	**118**	0	0	1	0	0	0	0	0	1	0	0	0	0
C1V3	3	5	**36**	0	0	0	0	0	0	0	0	0	0	0	1
C2V1	0	0	0	**19**	0	1	33	1	0	2	0	0	10	0	0
C2V2	0	0	0	0	**66**	3	2	6	0	0	2	0	0	4	0
C2V3	0	0	0	0	1	**2**	1	1	0	0	0	1	0	0	0
C3V1	3	0	0	8	1	1	**161**	1	3	12	0	0	21	0	1
C3V2	0	4	0	0	4	0	0	**201**	3	0	2	1	0	30	1
C3V3	0	6	0	2	0	3	8	11	**46**	1	0	1	1	1	2
C4V1	0	0	1	1	0	1	28	1	9	**23**	0	0	26	0	6
C4V2	0	0	0	0	7	1	0	14	2	2	**67**	1	0	14	0
C4V3	0	1	0	0	0	0	0	2	6	0	0	**8**	0	1	1
C5V1	13	0	3	3	0	0	25	1	1	7	0	0	**117**	1	2
C5V2	0	20	0	0	5	0	0	45	1	0	2	0	0	**292**	0
C5V3	0	1	5	0	0	0	1	0	6	1	0	1	13	5	**30**
total	**92**	**155**	**50**	**34**	**85**	**12**	**259**	**284**	**77**	**48**	**74**	**13**	**191**	**348**	**44**
taux de reconnaissance (%)	79%	76%	72%	56%	78%	17%	62%	71%	60%	48%	91%	62%	61%	84%	68%

TAB. 7.14 – Matrice de confusion pour le test de reconnaissance avec le regroupement III des consonnes et avec les paramètres du pincement en plus des six paramètres des contours interne et externe des lèvres. L'intervalle des observations est $[M2,L2]$.

constatons, par exemple pour la classe de référence C1V1, que la confusion avec la classe C5V1

passe de 2 à 13 tandis qu'avec la confusion avec la classe C1V3 passe de 8 à 3. En général, la tendance globale est que les confusions entres les classes syllabiques en contexte consonantique fixe diminuent ou restent presque inchangées, alors qu'en contexte vocalique fixe les confusions entre classes syllabiques augmentent. Ceci implique que l'instant $M2$ représente moins la cible consonantique que l'instant AC. Ce constat est conforme à nos attentes puisque nous avons déjà vu que la position de l'instant $M2$ est très variant autour de l'instant AC (la moyenne de l'intervalle [$A1$, $M2$] est de 63,30 ms et l'écart type est de 71,57 ms). Rappelons que le calcul de cette variance a été fait en détectant manuellement les instants acoustiques $A1$, $A2$ et $A3$, et donc AC et AV. Dans le cas présent, ces instants proviennent de l'étiquetage phonétique automatique décrit dans le chapitre 6 ; ce qui veut dire que cette variance s'amplifie davantage. Par conséquent, la différence de scores de reconnaissance entre le cas où nous utilisons l'intervalle [AC,$L2$] et le cas avec l'intervalle [$M2$,$L2$].

7.2.3 Résumé

Dans cette expérience, nous avons étudié la modélisation et la reconnaissance, en terme de visèmes, des syllabes CV en contexte du code LPC. Nous avons relevé, ainsi, trois problèmes à résoudre :

- Le premier problème concerne la catégorisation des syllabes en terme de visèmes. Etant donné que le cas des voyelles avait été déjà résolu, il restait à regrouper les consonnes en visèmes. Après plusieurs essais, un regroupement s'appuyant sur le lieu d'articulation de la langue et adapté pour qu'il soit compatible avec le système manuel du code LPC est choisi.
- Le second problème concerne le choix de la durée des observations en entrée des modèles HMM. nous avons testé trois intervalles fondés sur des instants acoustiques (AV et AC), labiaux ($L2$) et manuels ($M2$). Nous avons commencé par tester l'intervalle acoustique [AC,AV] et nous avons obtenu un taux de reconnaissance des visèmes de syllabes CV de 80,3%. Ensuite nous avons montré, en deux étapes avec les intervalles [AC,$L2$] et [$M2$,$L2$] que nous pouvions nous affranchir des étiquettes audio, si nous supposons que $L2$ peut être obtenu sans s'appuyer sur ces étiquettes. Pour la première étape, le taux de reconnaissance des visèmes de syllabes est resté stable, contrairement à nos attentes d'une amélioration grâce à la précision de l'instant $L2$. En revanche, comme la variance de l'instant $M2$ autour de l'instant AC est assez grande, le taux de reconnaissance a baissé pour atteindre 71,29% en utilisant l'intervalle [$M2$,$L2$].
- Le troisième problème porte sur le nombre de paramètres labiaux à utiliser. En plus des six paramètres classiques extraits des contours interne et externe des lèvres (A, B, S, A', B', S'), nous avons introduit deux paramètres caractérisant le pincement des lèvres supérieure et inférieure. L'ajout de ces deux paramètres fait augmenter les taux de reconnaissance de plus de 4% en moyenne.

Dans tous les cas de figure, les erreurs apparaissent principalement pour les classes de syllabes CV en contexte de voyelles arrondies (groupe V1) ; alors qu'en contexte de voyelles non arrondie (groupe V2) ou semi-arrondies (groupe V3) les visèmes de syllabes CV sont mieux discriminés.

Dans cette expérience, les syllabes CV modélisées sont dans un contexte de phrases. L'effet de ce contexte sur la reconnaissance doit être analysé afin d'optimiser les observations sélectionnées des syllabes CV. De même, d'autres paramètres labiaux peuvent être ajoutés tels que les aires des lèvres supérieure et inférieure.

7.3 Vers une reconnaissance de mots

Les systèmes de reconnaissance automatique de la parole dépendent beaucoup de l'unité minimale (indécomposable) de reconnaissance à utiliser. Les unités utilisées vont généralement du phonème jusqu'au mot (Ben Mosbah, 2005). La question qui se pose est comment connecter ces unités reconnues pour arriver à reconnaître le message d'entrée (mot ou phrase).

Dans le cas du code LPC, qui s'appuie sur une organisation reposant sur des unités syllabiques de type CV, l'unité la plus évidente à utiliser semble être la syllabe CV. Les résultats obtenus sur ces syllabes dans l'expérience 2 sont encourageants et ouvrent la porte vers une reconnaissance automatique de la parole continue en contexte LPC. Cependant, la connection de ces unités pour arriver au mot ou à la phrase n'est pas une tâche facile. En effet, la langue française ne se constitue pas uniquement de ce type d'unités ; des syllabes de type VC, VV ou CC peuvent aussi être présentes.

Dans cette section, nous tentons de montrer, à partir d'une étude exploratoire, qu'il est possible de modéliser des mots, composés de syllabes CV successives, par des modèles HMM construits en concaténant les modèles HMM des classes de ces syllabes. En disposant d'un dictionnaire de modèles HMM de toutes les classes de CV, l'avantage de cette méthode se situe dans l'absence de tout nouvel apprentissage chaque fois qu'un nouveau mot vient s'ajouter à la liste des mots à reconnaître. Cependant, le problème majeur de cette méthode concerne la manière de concaténer les modèles de syllabes. Dans la suite,en utilisant une concaténation simple (voir annexe F), nous testons la reconnaissance d'un corpus réduit de mots extraits de notre corpus global.

7.3.1 Résultat du test de reconnaissance

Vu la taille de notre corpus global, le nombre de mots composés de deux syllabes CV ne devrait pas être suffisamment grand pour pouvoir faire une vraie analyse du problème. Ainsi, cette expérience n'est qu'une exploration pour tester notre méthode de concaténation, nous permettant de relever certains aspects de l'application de cette méthode.

Tout d'abord, nous récupérons les occurrences de tout mot composé de deux syllabes CV. Comme les mots se trouvaient dans des phrases, il est nécessaire de définir les limites de leurs intervalles. Ainsi, nous choisissons de caractériser chaque mot par la transition entre l'instant AC de la première syllabe et l'instant $L2$ de la seconde syllabe. La liste des mots récupérés pour ce test est présentée dans la table 7.15.

Ensuite, nous concaténons les modèles HMM correspondant aux syllabes composant chaque mot. Nous choisissons d'utiliser les modèles obtenus sur l'intervalle [AC,$L2$]. Ainsi, dans la construction du modèle HMM de mot, la transition entre les deux syllabes, à savoir la transition entre l'instant $L2$ de la première syllabe et l'instant AC de la seconde, n'est pas prise en compte. La figure 7.3.1 illustre un exemple de transition entre les deux syllabes d'un mot.

Enfin, nous testons la reconnaissance des mots sélectionnés en terme de visèmes aussi, puisque les modèles le sont. Dans ce test, pour chaque mot, la séquence d'observations injectée en entrée du classifieur HMM peut contenir la transition entre les deux syllabes ou non. **En conservant la transition, nous obtenons un taux global de reconnaissance des visèmes de mots de**

mots	nombre de répétitions	CV1	CV2	classe de mot	mots	nombre de répétitions	CV1	CV2	classe de mot
jeté	5	10	8	1	maison	2	2	7	11
chuté	2				mangé	2	3	11	12
chacun	6	11	14	2	pendant	5	3	9	13
jamais	5	11	2	3	fichu	2	5	10	14
Corot	2	13	13	4	fendu	4	6	7	15
commun	7	13	2	5	secoue	2	7	13	16
carré	9	14	14	6	torrent	2	7	15	17
guéri	4				tombé	2	7	2	18
cadeau	2	14	7	7	demi	2			
rasé	2	14	8	8	neveu	2	7	4	19
campait	2	15	2	9	souvent	2	7	6	20
petit	4	1	8	10	défunt	2	8	5	21
posé	2				donné	4	9	8	22

TAB. 7.15 – Liste des mots composés de deux syllabes CV successives.

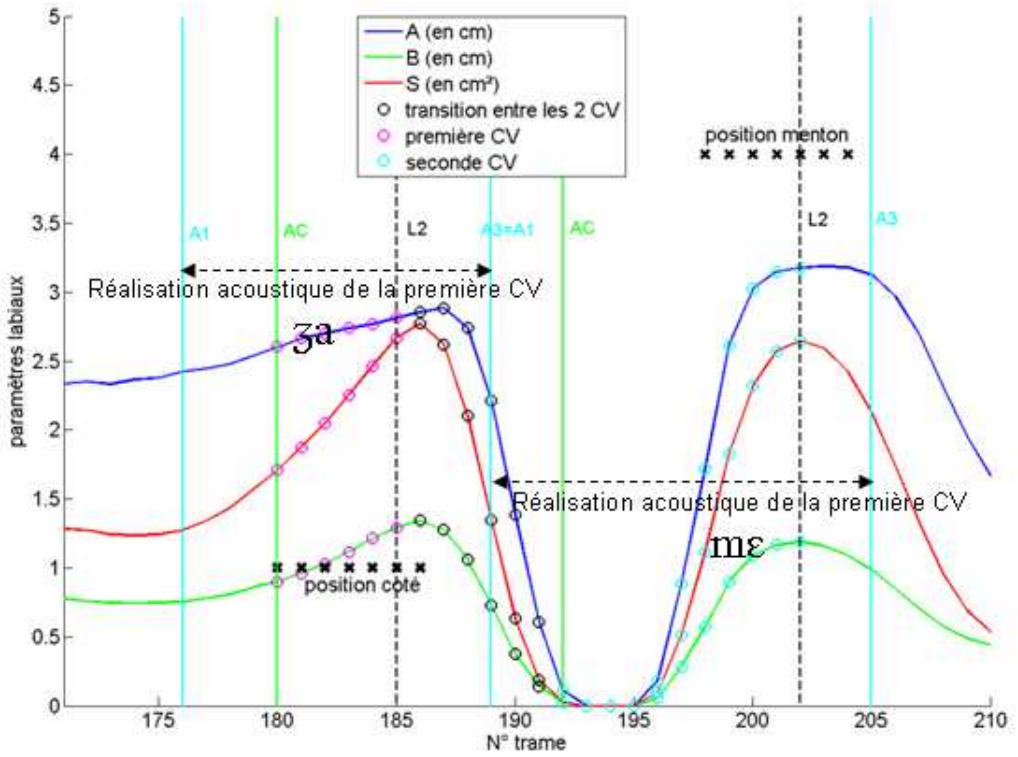

FIG. 7.14 – Représentation des signaux labiaux du mot [ʒamɛ]. La transition entre les 2 syllabes CV est marquée.

78,57%. Les erreurs se produisent globalement entre les modèles qui ont une classe de syllabes commune. **En enlevant la transition, ce taux augmente et atteint 88,1%**. Ceci implique que nous pouvons évaluer à 10% l'erreur due à la présence de la transition. En modélisant cette transition, nous devrions pouvoir éviter cette erreur.

7.4 Conclusion

Dans ce chapitre, l'étude du flux labial nous a permis de relever plusieurs points importants et essentiels en vue de la fusion avec le flux manuel. Tout d'abord, le fait que les voyelles soient toutes articulées aux lèvres implique qu'elles peuvent être caractérisées par des paramètres labiaux extraits à un seul instant (l'instant de l'atteinte de la cible vocalique aux lèvres $L2$). Notre méthode de détection de la cible vocalique aux lèvres, s'appuyant sur les minima de la vitesse labiale, permet de déterminer cet instant. De plus, les paramètres issus du contour interne des lèvres peuvent être suffisants pour modéliser toutes les voyelles. En s'appuyant sur ces paramètres, nous avons confirmé, d'un coté, les ambiguïtés de la lecture labiale en contexte de production de parole continue (phrases). D'un autre coté, une représentation hiérarchique des distributions des voyelles a montré que ces dernières peuvent être catégorisées en trois visèmes compatibles, à une exception près, avec les groupes du système manuel du code LPC ; ce qui démontre la complémentarité de ce code.

En terme de classification, un simple classifieur gaussien permet d'obtenir de bonnes performances de reconnaissance des voyelles par position LPC (taux global de reconnaissance de 89%) et des visèmes de voyelles (taux de 92,6%). L'analyse des erreurs d'identifcation dans ces deux tests montre que le problème principal réside dans les imprécisions de la méthode de détection de la cible vocalique, et plus précisement sur les imprécisions des instants déterminés par la segmentation du signal audio (étiquetage automatique phonétique). Nous notons aussi dans ce cas, les effets de la coarticulation qui influent beaucoup sur les cibles vocaliques aux lèvres. En effet, pour certaines voyelles, notamment arrondies, précédées par certaines consonnes non articulées aux lèvres (les consonnes fricatives par exemple), la cible vocalique aux lèvres se retrouve dans la réalisation acoustique de la consonne ; c'est-à-dire la réalisation labiale de la voyelle est anticipée dans la réalisation de la consonne.

Par ailleurs, l'étude effectuée sur les voyelles ne peut être appliquée directement sur les consonnes. La faute est attribuée à la non articulation aux lèvres de toutes les consonnes. Pour reconnaître les consonnes, la solution est de les associer avec les voyelles dans le cadre d'une syllabe CV. Ceci est d'un grand intérêt puisque le code LPC est un système qui s'appuie sur des unités syllabiques de type CV. Dans ce cas, c'est toute la transition entre la consonne et la voyelle qui est considérée. La modélisation des syllabes CV nécessite donc des systèmes qui prennent en compte ce type de données d'observation ; d'où l'emploi des modèles HMM. En absence de la fusion avec l'information de la main, les syllabes ne peuvent être considérées qu'en termes de visèmes.

La modélisation HMM des syllabes CV donne des performances encourageantes. En reconnaissance, le regroupement des consonnes en visèmes et les paramètres labiaux utilisés ont une importance primordiale dans l'amélioration des performances. Dans les meilleurs cas, c'est-à-dire avec le regroupement des consonnes s'appuyant sur le lieu d'articulation de la langue et avec les paramètres de pincement, le taux de reconnaissance des visèmes de syllabes CV avoisine les 80%. Cependant, cette étude a montré que les erreurs sont principalement observées sur des groupes de consonnes en contexte des voyelles arrondies. En revanche, les syllabes CV pour des voyelles non arrondies et semi-arrondies sont mieux reconnues (87%).

Dans cette étude, les syllabes CV modélisées se situent en contexte de phrases. L'effet du contexte a, par conséquent, une influence sur les modèles des classes syllabiques et par la suite sur les taux de reconnaissance. Ainsi, une optimisation des durées d'observations des syllabes CV est nécessaires pour réduire cet effet. Dans ce sens, les instants obtenus à partir des segmentations temporelles des flux labial et manuel ($M2$ pour les positions LPC de la main et $L2$ pour les voyelles) peuvent être considérés ; ce qui nous permet en plus de s'affranchir du besoin des étiquettes acoustiques. Les résultats obtenus montrent que nous pouvons compter sur l'instant $L2$. Par contre, les performances avec l'instant $M2$ sont diminuées, probablement à cause de sa variance par rapport aux instants acoustiques.

En dernière étude, nous avons montré que les modèles HMM des syllabes CV construites peuvent servir à reconnaître, en termes de visèmes, des mots composés de syllabes CV successives. Dans ce cas aussi, nous obtenons de bonnes performances encourageantes pour une éventuelle reconnaissance complète de la parole continue en contexte du code LPC.

Enfin, tous ces points peuvent être utiles pour établir des modèles de fusion des informations manuelle et labiale. Dans le chapitre suivant, nous présenterons les premiers modèles de fusion des gestes main-lèvres.

Chapitre 8

Reconnaissance phonétique des gestes main-lèvres

Le décodage phonétique des gestes main-lèvres du code LPC nécessite de fusionner les informations issues des flux LPC de la main et des lèvres (voir respectivement chapitres 6 et 7). La fusion consiste à combiner ces deux sources d'information comme nous l'avons présenté au chapitre 3, soit au niveau des données (fusion précoce) ou au niveau des décisions prises sur chacun des flux (fusion tardive). Pour prendre une décision, un classifieur est nécessaire (classifieur gaussien, HMM, quantification vectorielle... voir aussi chapitres 2 et 4). De plus cette fusion doit tenir compte de la désynchronisation entre les deux flux manuel et labial, relevée par Attina *et al.* (2004) en contexte de syllabes isolées et que nous avons confirmé précédemment dans le chapitre 6 en contexte plus complexe de phrases.

Dans ce chapitre, nous proposons des modèles de fusion issus du domaine de l'intégration audio-visuelle en parole. Ces modèles intègrent des informations issues des lèvres et des mouvements de la main, permettant la reconnaissance complète d'unités phonétiques. Ces premiers modèles s'appuient sur une approche déterministe. Nous consacrerons la première section à la discussion de ces modèles avec une focalisation sur la reconnaissance complète de la voyelle. Dans la seconde section nous montrerons comment ces modèles peuvent être adaptés pour reconnaître des syllabes CV.

8.1 Reconnaissance complète de la voyelle

8.1.1 Modèles de fusion

Dans le processus de fusion, les instants d'atteinte de la position cible de la main ($M2$) sont utilisés comme des instants de référence pour le flux manuel. Les paramètres labiaux sont extraits aux instants $L2$. Dans une étude récente, Alegria et Lechat (2005) concluent que l'intégration des gestes main-lèvres semble suivre des principes similaires à ceux observés dans la perception de la parole audio-visuelle. Ainsi, les modèles classiques de la fusion audio-visuelle[1], que nous avons décrit dans la partie *état de l'art*, semblent être potentiellement intéressants.

[1] modèles ID, IS, RM et RD.

Dans le modèle d'identification directe (ID), toutes les composantes sont réunies dans un même vecteur. Ce vecteur est alors considéré dans la phase de classification. Mais, ce modèle ne semble pas approprié à notre cas de fusion LPC. Une des raisons est que le système d'intégration à utiliser doit être capable d'intégrer des informations quantitatives provenant des lèvres (les valeurs des paramètres labiaux) et des informations qualitatives provenant des gestes LPC de la main (position et configuration de la main). Le modèle à recodage dans la modalité dominante (RD), dans lequel l'information d'une modalité est recodée dans l'autre, semble aussi non approprié, puisque aucune des deux composantes du code LPC ne porte le code phonétique complet. La transformation dans un troisième espace commun (amodal), le domaine des causes, reste une hypothèse théorique mais très intéressante. Ces trois modèles, ID, RM et RD sont des modèles de la famille fusion de représentations qui ne pas gèrent l'asynchronie (chapitre 3 de la partie *état de l'art*).

Par ailleurs, le modèle à identification séparée (IS) semble convenir à notre cas. Dans ce dernier, une décision est prise à partir du traitement de chaque flux et le résultat final est obtenu comme l'intersection des deux décisions. En considérant ce modèle, d'un coté, la position LPC de la main est connue à l'instant $M2$ et donc un premier groupe composé de deux ou trois voyelles candidates peut être obtenu. De l'autre coté, à l'instant $L2$ correspondant (tout dépend du système d'appariement main-lèvres), un second groupe de voyelles (visème) est dérivé d'une classification des paramètres labiaux. La correspondance entre les deux instants $L2$ et $M2$ est réalisée dans notre cas en considérant pour chaque instant $M2$ l'instant $L2$ qui suit directement, en tenant compte de l'avance de $M2$ sur $L2$ (voir chapitre 6). La reconnaissance de la voyelle résulte de l'intersection entre ces deux groupes de voyelles. La figure 8.1 illustre ce modèle de fusion. Cependant, il peut arriver que l'intersection soit vide, notamment avec le groupe de voyelles de la position "pommette" [ɛ̃,ø] et le visème des voyelles semi-arrondies [ɑ̃,ɔ,œ].

Pour résoudre ce problème, nous considérons pour chacune des cinq positions LPC de la main un classifieur ; donc, la classification labiale fournit une voyelle candidate pour chaque position LPC de la main (voir figure 8.2).

Finalement, afin de réduire le nombre de classifieurs de cinq à un, il est possible de contraindre la décision sur le flux labial par la décision sur le flux manuel, en considérant l'avance de l'instant $M2$ sur l'instant $L2$ dans le cas des voyelles. Ainsi, entre deux instants successifs $M2$, un instant $L2$ est localisé à l'exception du cas des syllabes consonne-consonne-voyelle (CCV)[2]. Au premier instant $M2$, la position LPC de la main permet d'identifier un premier groupe de voyelles. Un classifieur s'appuyant sur les paramètres labiaux extraits à l'instant $L2$ identifie un seul élément parmi les voyelles de ce groupe. La figure 8.3 illustre ce modèle que nous appelons "la main en premier, ensuite les lèvres".

Ce modèle ainsi construit "la main en premier, ensuite les lèvres" peut être considéré comme un modèle maître-esclave piloté par le flux manuel. Il est cohérent avec les résultats en perception du code LPC (Cathiard *et al.*, 2004) qui ont démontré que l'information de la main disponible en avance (par rapport à l'information labiale) est aussi perçue en avance et donc utilisée pour le décodage du code LPC (voir partie *état de l'art*, chapitre 1).

[2] *Dans ce cas, il n'y a pas de cible vocalique entre la position LPC correspondant à la première C et celle correspondant à la syllabe CV .*

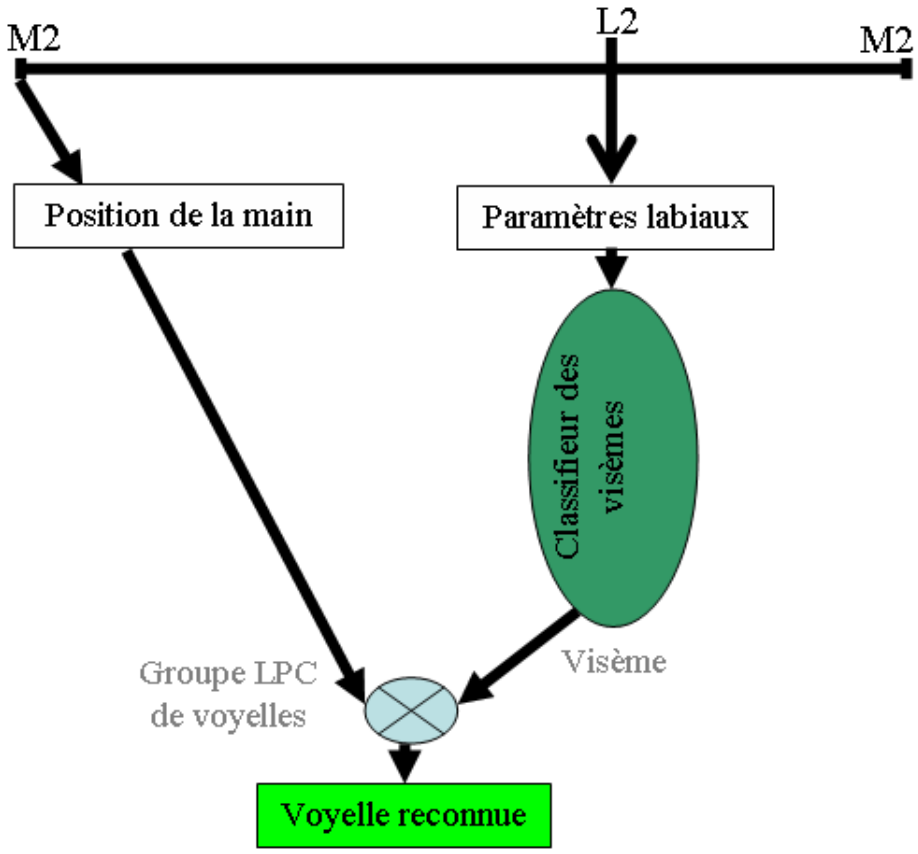

FIG. 8.1 – Première version du modèle IS appliqué à la fusion des gestes main-lèvres pour la reconnaissance de la voyelle.

8.1.2 Taux de reconnaissance

Le dernier schéma de fusion "la main en premier, ensuite les lèvres" est maintenant appliqué pour reconnaître automatiquement les voyelles contenues dans un sous-ensemble de phrases de notre corpus global. Dans ce cas, la position de la main identifiée à l'instant $M2$ fournit un groupe composé de deux ou trois voyelles (selon la position identifiée). Ensuite, nous considérons un simple classifieur Gaussien appliqué aux paramètres labiaux du contour interne pour identifier la voyelle codée parmi ces voyelles.

Nous nous appuyons, dans ce test, sur les mêmes sous-ensembles qui ont servi dans l'expérience 1 décrite dans le chapitre précédent. Dans cette expérience, rappelons-le, nous avons construit deux sous-ensembles : sous-ensemble 1 et sous-ensemble 2. Le premier a servi pour la phase d'apprentissage et le second pour le test. Dans le test présent, le sous-ensemble 1 (1167 voyelles) est aussi utilisé pour l'apprentissage des modèles gaussiens des voyelles. Pour le test, nous n'avons pas à extraire les voyelles du sous-ensemble 2 (1105 voyelles). Par contre, l'ensemble des phrases contenant ces voyelles est considéré et l'objectif est de détecter les voyelles dans ces phrases et de les identifier.

Le taux global de reconnaissance calculé sur les voyelles identifiées dans ces

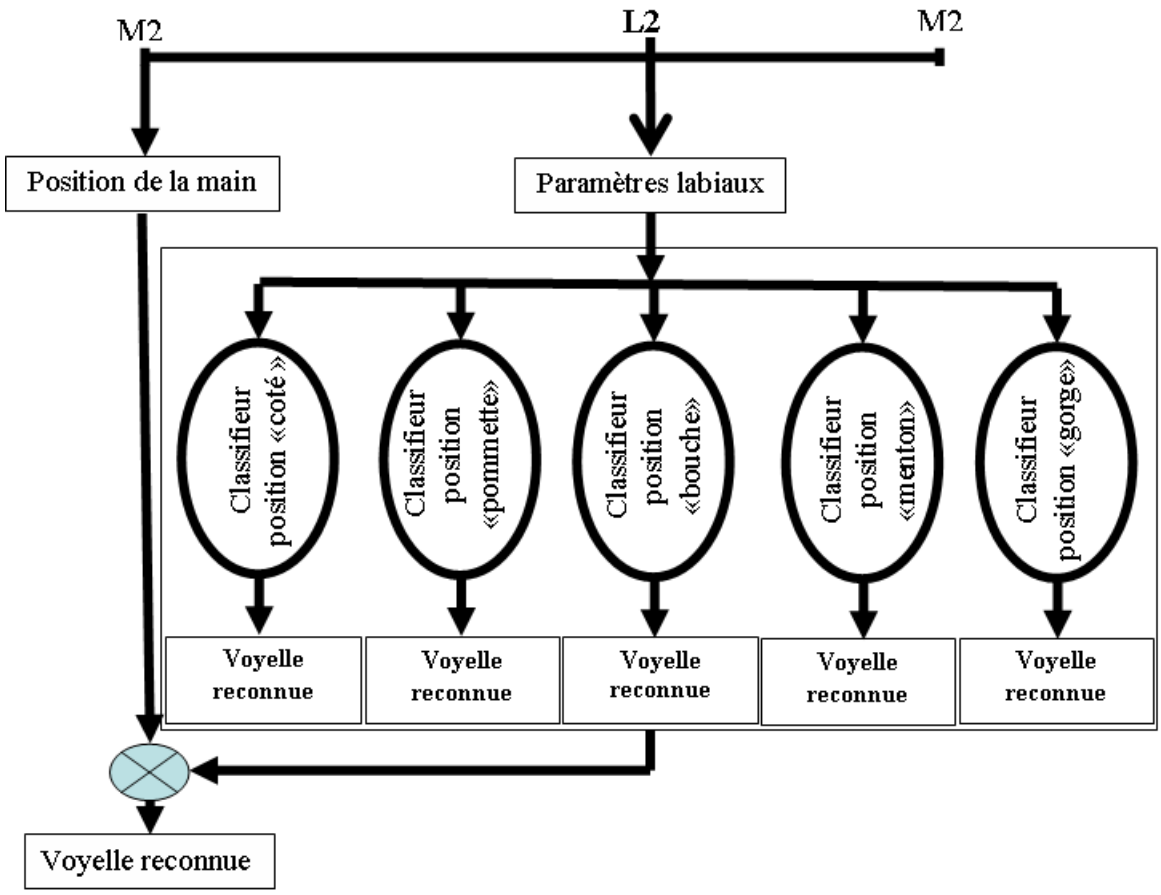

FIG. 8.2 – Identification de la voyelle : schéma possible de fusion des informations labiale et manuelle s'appuyant sur le modèle IS et en utilisant 5 classifieurs.

phrases est de 75%. Notons que, dans ce test, les deux voyelles [œ̃] et [e][3] sont considérées comme une seule voyelle. En se donnant la position LPC de la main, c'est-à-dire sans erreurs, la classification des voyelles par position LPC a donné un taux de reconnaissance de 89% en s'appuyant sur les mêmes données labiales (chapitre 7, expérience 1). La différence avec le taux obtenu dans le test présent peut être donc attribuée à la décision automatique sur la position LPC de la main et/ou à l'appariement des instants $M2$ et $L2$ (la correspondance entre les données du flux manuel et celles du flux labial).

Notre taux de 75% doit être comparé avec le taux de décodage perceptif des voyelles (94,8%) obtenu dans l'évaluation du corpus (voir chapitre 5). Dans ce cas de décodage par des auditeurs humains, l'identification de la voyelle était facilitée par le contexte sémantique ; ce qui majore le résultat. En revanche, notre résultat est tout à fait comparable avec le score de 83,5% obtenu par Nicholls et Ling (1982) qui mesure l'efficacité perceptuelle du *Cued Speech*[4] dans leur étude de la réception des syllabes de types CV et VC. Dans cette étude, les auteurs se sont appuyés sur des logatomes, donc sans possibilité d'une prédiction par le contexte sémantique. La comparaison

[3] *Nous avons vu que ces deux voyelles appartenant au même groupe LPC, sont aussi dans le même visème.*
[4] *Version originale du code LPC*

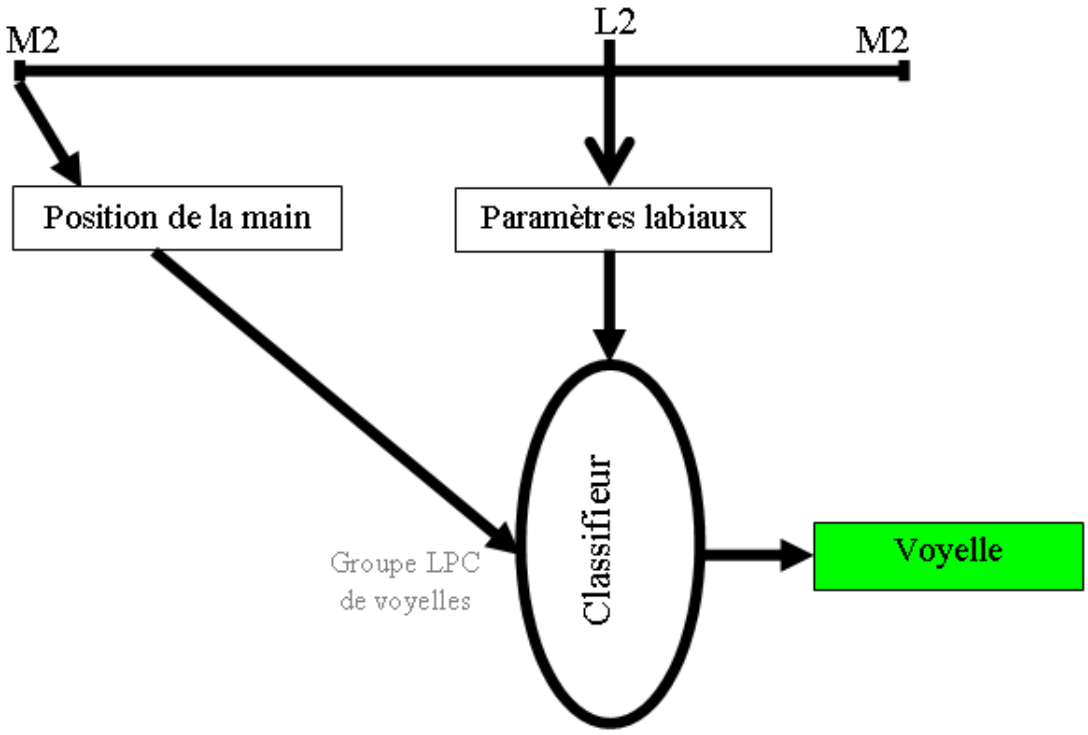

FIG. 8.3 – Identification de la voyelle : "la main en premier, ensuite les lèvres" pour le schéma de fusion des informations labiale et manuelle s'appuyant sur un modèle maître-esclave.

avec le taux de Nicholls et Ling semble donc plus judicieuse et notre résultat, légèrement en retrait, reste honorable même si des améliorations sont envisageables. La table 8.1 présente une comparaison des taux discutés dans cette section.

Taux de reconnaissance de la voyelle à partir du test avec le schéma de fusion «la main en premier, ensuite les lèvres »	*Taux de reconnaissance de la voyelle avec la position LPC de la main connue sans erreur*	*Taux de reconnaissance des CV et VC (Nicholls et Ling, 1982)*	**Taux d'identification des voyelles dans le test de décodage perceptif**
75 %	89 %	83.5 %	94.8 %

TAB. 8.1 – Table comparative des scores de reconnaissance des voyelles.

8.2 Perspectives : Modèle de fusion pour reconnaître les syllabes CV

Dans le chapitre 7, nous avons étudié la modélisation et la reconnaissance, en termes de visèmes, des syllabes CV. Nous avons ainsi montré qu'une modélisation HMM peut donner de bons résultats en reconnaissance des visèmes syllabiques. Pour reconnaître complètement la syllabe produite en LPC, il faut tenir compte de l'information de la main (complémentarité du code LPC). Ce qui nous ramène vers la problématique de la fusion des gestes main-lèvres du code LPC. Dans ce sens, notre modèle de fusion, défini précédemment pour la reconnaissance complète de la voyelle, peut être utilisé dans le cas des syllabes CV. En effet, dans l'intervalle délimité par les instants $M2$ et $L2$, nous possédons suffisamment d'informations manuelle et labiale pour pouvoir envisager une reconnaissance complète de la syllabe. D'une part, à l'instant $M2$, l'information manuelle, position et configuration LPC, est connue (voir chapitre 6 et 7). Ainsi, la position LPC de la main identifie un groupe de voyelles et la configuration de la main un groupe de consonnes. D'autre part, la transition des paramètres labiaux entre l'instant $M2$ et $L2$ porte des informations labiales suffisantes pour identifier le visème syllabique correspondant à la syllabe produite. Cette dernière est déterminée comme l'intersection entre les groupes de voyelles et de consonnes issus de la décision sur le flux manuel et le visème syllabique issu de la décision sur le flux labial.

Pour avoir la décision sur le flux labial, une classification HMM est utilisée. Si nous considérons que les consonnes sont regroupées en 5 visèmes (voir chapitre 7) et les voyelles en 3 visèmes, alors 15 modèles HMM sont à apprendre. Dans la phase de test, pour chaque séquence d'observations en entrée du classifieur HMM, les 15 modèles sont testés. Il est possible, comme nous l'avons fait pour le cas de la voyelle, de contraindre le classifieur HMM par la décision sur le flux manuel ; ce qui permet de sélectionner un nombre inférieur de modèles à tester (la décision sur la position donne au maximum 3 voyelles possibles tandis que la décision sur la configuration donne au maximum 4 consonnes, ce qui fait que dans ce cas 12 modèles peuvent être considérés). Contrairement au cas de la voyelle, reconnue en sortie du classifieur, nous obtenons dans le cas des syllabes CV en sortie du classifieur seulement un visème de syllabe. C'est après l'intersection de ce visème avec les deux groupes (de voyelles et de consonnes) obtenus par la décision sur le flux manuel, que la syllabe est enfin identifiée. Avec ce schéma ainsi conçu, nous pensons réduire l'erreur de reconnaissance. Ce schéma a un principe similaire au schéma de fusion "la main en premier, ensuite les lèvres" que nous avons testé pour les voyelles. La figure 8.4 illustre ce dernier schéma dans le cas des syllabes CV.

Dans la même optique de réduire le nombre de modèles HMM employés par le classifieur et d'améliorer la reconnaissance, ce schéma de fusion pour les syllabes CV est combiné avec un schéma de fusion permettant d'identifier complètement la voyelle. En effet, le classifieur HMM reçoit en entrée l'information sur la voyelle qui permet d'identifier le visème auquel elle appartient. Ainsi, le nombre de modèles HMM à utiliser est divisé par trois. Dans ce cas, le classifieur teste quatre modèles à la place de douze. En sortie de ce classifieur, un visème de syllabe CV est donc obtenu pour lequel la voyelle est déjà reconnue. Ensuite, l'intersection avec le groupe LPC des consonnes obtenu par le traitement de la main identifie ensuite la consonne

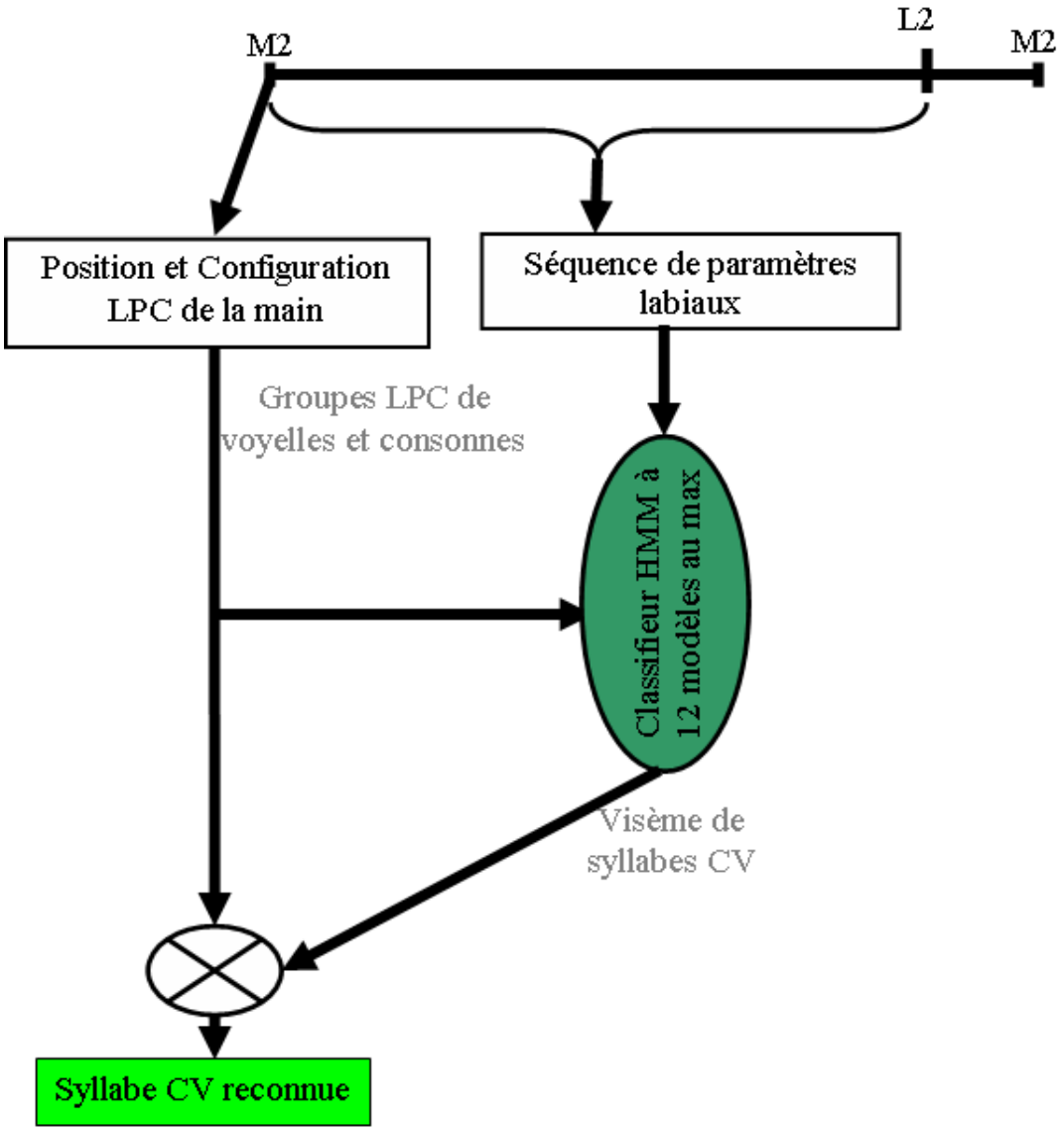

FIG. 8.4 – Identification des syllabes CV : schéma de fusion "la main en premier, ensuite les lèvres".

et en même temps la syllabe CV. La figure 8.5 montre ce processus.

Nous avons donc construit un modèle hybride "la main en premier, ensuite les lèvres" où la voyelle est reconnue séparément grâce à un modèle maître-esclave piloté par la main, et vient contraindre le système de reconnaissance de la syllabe CV.

8.3 Conclusion

Pour reconnaître complètement la voyelle, nous proposons un modèle de fusion maître-esclave "la main en premier, ensuite les lèvres" qui est un modèle à identification séparée pilotée par le flux manuel. Dans ce modèle, la décision sur la main obtenue à partir du codage automatique de la main sélectionne d'abord, à l'instant d'atteinte de la position LPC cible par la main, un

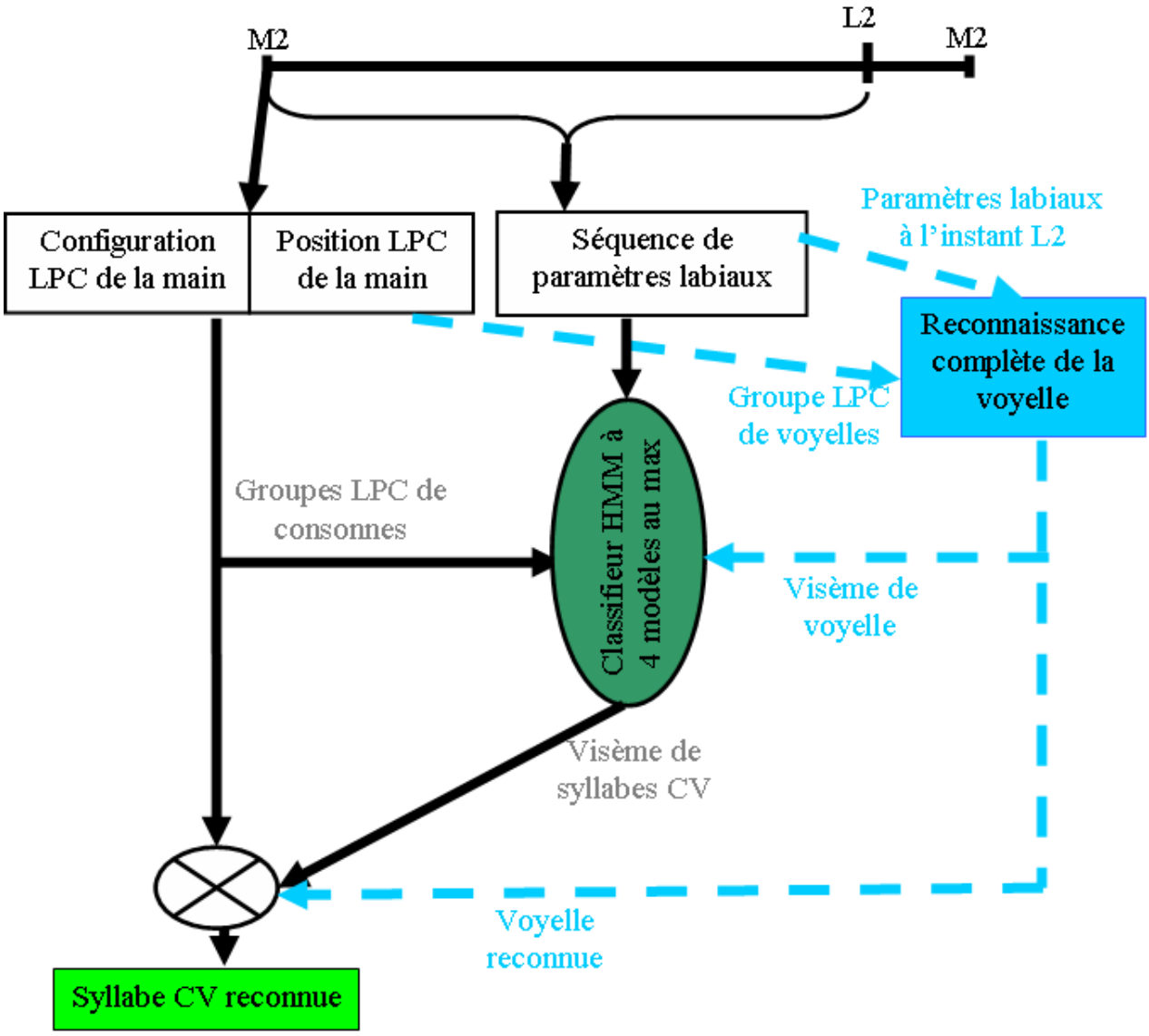

FIG. 8.5 – Identification des syllabes CV : schéma hybride de fusion "la main en premier, ensuite les lèvres" avec en plus l'information sur la voyelle .

groupe de voyelles candidates. Puis, cette information est injectée dans un classifieur (Gaussien dans notre cas) qui reconnaît la voyelle parmi les candidates à partir des paramètres labiaux extraits à l'instant d'atteinte de la cible vocalique aux lèvres. Le test expérimental de ce modèle a permis d'obtenir un taux honorable et encourageant de 75% de reconnaissance complète de la voyelle. En plus des erreurs causées par la classification labiale estimées à 11% (voir chapitre précédent), des erreurs estimées à 14% sont dues à la précision sur le codage de la position LPC de la main et à l'appariement main-lèvres.

Par ailleurs, notre modèle de fusion pourrait tout à fait être adapté au cas des syllabes CV. Dans ce cas, un classifieur permettant de prendre en compte toute la durée d'observation des syllabes CV (un HMM par exemple), remplace le classifieur Gaussien. De plus, l'information

manuelle injectée au classifieur est cette fois composée de l'information sur les voyelles et sur les consonnes. Afin de réduire le nombre des modèles HMM utilisés par la classification labiale, nous avons décrit un modèle hybride combinant un modèle maître esclave pour les syllabes CV en parallèle avec un autre modèle maître-esclave pour les voyelles. L'information de la voyelle en sortie de ce dernier contraint le premier modèle. Les expériences concernant ce modèle sont en cours.

Conclusion générale

En guise de résumé ...

Pouvons nous voir un jour un sourd avec un téléphone portable (vidéo y compris) en discussion avec ses amis entendants ? Arriver à réaliser cela passe par de multiples travaux de recherches. Si l'on s'en tient à l'hypothèse que le sourd utilise la Langue Française Parlée Complétée (code LPC) pour communiquer, notre thèse fait partie de ce cadre. Elle vise à étudier la reconnaissance phonétique des gestes main-lèvres du code LPC. Ceci passe par une analyse séparée de chacun des flux manuel et labial, et une analyse combinée des deux avant de les fusionner.

Flux manuel

Les deux méthodes que nous avons proposé permettent d'obtenir la position et la configuration LPC à chaque instant. Ces deux méthodes reposent sur des concepts simples (classification Gaussienne, seuillage, détection des ralentissements du mouvement et comptage des doigts) et les différents tests d'évaluation montrent qu'elles donnent de bonnes performances. En effet, le traitement complet de la position LPC de la main permet d'obtenir les positions avec une précision de 96,5% en utilisant seulement 7 pastilles placés sur la main en 2D. Ce résultat est identique à celui obtenu par Gibert *et al.* (2005) (96,76%) en utilisant 30 marqueurs sur la main en 3D. Le traitement de la configuration LPC en utilisant sept pastilles seulement permet d'obtenir une performance de 92% d'identification correcte, voisine de celle obtenue par Gibert *et al.* (2005) (98,78%).

De plus, le traitement fondé sur le ralentissement du mouvement de la main nous a permis d'obtenir automatiquement des durées caractérisant la tenue de la main dans une position cible et sa transition vers une autre. A partir des instants obtenus de cette segmentation temporelle, nous avons confirmé, en contexte complexe de phrases chez un participant normo-entendant et chez un participant sourd, la coordination entre la main et les lèvres obtenue par Attina *et al.* (2004) sur des logatomes. Dans cette coordination, la main atteint une position cible ($M2$) dans la réalisation de la consonne ; ce qui s'explique, selon l'hypothèse d' Attina (2005), par la compatibilité de deux contrôles locaux : celui du pointage de la cible LPC (à l'instant $M2$), et du geste articulatoire de la consonne dans le conduit vocal. Cette hypothèse reste valide avec des données dans ce cadre plus complexe de parole continue avec une robustesse du rendez-vous : "atteinte de la cible de la main avec la réalisation articulatoire de la consonne".

Flux labial

Nos résultats sur les voyelles analysent et confirment les ambiguïtés de la lecture labiale à partir des formes de lèvres produites en contexte de parole continue. L'efficacité de la composante manuelle du code LPC à distinguer entre les formes de lèvres a été aussi démontrée en contexte de production. De plus, nous avons démontré qu'un simple classifieur Gaussien s'appuyant sur des paramètres extraits d'une vue de face des lèvres peut être suffisant pour classifier les voyelles. En effet, quand la position LPC de la main est donnée, une bonne performance d'identification de la voyelle, en moyenne 89%, est obtenue avec seulement un seul instant de mesure défini par la méthode de segmentation de cibles vocaliques aux lèvres.

L'extension directe de cette modélisation aux consonnes semble être beaucoup plus complexe. En effet, certaines consonnes sont moins articulées aux lèvres, donc leur réalisation aux lèvres est influencée par le contexte de parole. Pour en tenir compte, nous nous sommes concentrés dans notre étude sur le contexte aval des consonnes dans le cadre d'une syllabe de type consonne-voyelle (CV).

A l'instant où la main atteint sa position cible ($M2$), la position et la configuration LPC sont connues. L'étude des syllabes CV s'est donc focalisée sur le flux labial. Pour modéliser les syllabes CV, trois questions sont posées : dans quel intervalle d'observation les syllabes CV doivent-elles être considérées ? Etant donné les ambiguïtés de la lecture labiale, et donc l'existence de visèmes, quel sont les regroupements des voyelles et des consonnes à considérer ? Et enfin, les paramètres labiaux considérés apportent-ils suffisamment d'information pour discriminer les syllabes CV ?

Dans notre étude, les syllabes CV ont été extraites dans un intervalle délimité par des instants acoustiques, labiaux et manuels. Nos résultats expérimentaux montrent qu'avec les instants acoustiques de bonnes performances sont obtenues. De même, avec un intervalle acoustico-labial ces mêmes performances sont confirmées. En revanche, l'intervalle délimité par l'instant d'atteinte de la position LPC cible de la main et l'instant d'atteinte de la cible vocalique aux lèvres, dégrade les performances en reconnaissance purement labiale. Cette dégradation est probablement la conséquence de la variance de l'instant $M2$ par rapport aux instants acoustiques.

Pour répondre à la deuxième question, l'analyse du flux labial dans le cas des voyelles a fourni un regroupement en trois visèmes compatibles avec la littérature (Robert-Ribès, 1995). En revanche, le regroupement des consonnes en visèmes a été l'objet de plusieurs tests. Le regroupement des consonnes finalement considéré repose sur le lieu d'articulation de la langue et les formes des lèvres (pour les consonnes articulées aux lèvres).

Par ailleurs, notre modélisation s'est appuyée dans un premier temps sur six paramètres extraits des contours interne et externe des lèvres (étirements, apertures et aires). Dans un second temps, l'ajout de deux paramètres caractérisant le pincement des lèvres améliore de 4% en moyenne le taux de reconnaissance. Ceci illustre l'importance d'utiliser des paramètres pertinents pour la discrimination des syllabes.

Dans le meilleur cas, un taux de reconnaissance des visèmes syllabiques de type CV autour des 80% est obtenu. Nous avons montré dans cette étude que les erreurs de reconnaissance sont principalement observées sur des groupes de consonnes en contexte de voyelles arrondies. Ceci est en accord avec l'effet de coarticulation des voyelles arrondies qui masquent complètement ou

partiellement les consonnes qui les précèdent (Abry et Boë, 1986). Par contre, les syllabes CV pour des voyelles non arrondies et semi-arrondies sont mieux discriminées (87%).

Enfin, les modèles de syllabes CV ainsi construits peuvent servir à la reconnaissance, en terme de visèmes, de mots composés d'une suite de syllabes CV successives. Dans un test exploratoire, nous avons montré la possibilité de concaténer les modèles HMM des syllabes CV pour obtenir des modèles HMM de mots. En terme de reconnaissance, cette méthode donne de bonnes performances qui varient de 78,57% à 88,1% selon que la transition entre les deux syllabes CV est considérée ou non.

Fusion main-lèvres

Si l'information de la main (position LPC) est obtenue à partir du processus de détection automatique et que l'appariement main-lèvres est traité dans notre modèle de fusion maître-esclave "la main en premier, ensuite les lèvres" piloté par le flux manuel, la performance de reconnaissance complète de la voyelle a certes baissé, mais atteint cependant un score honorable de 75%. Les erreurs peuvent être réduites par une amélioration de la qualité de l'extraction des contours de lèvres. De plus, une amélioration de l'appariement main-lèvres dans le modèle de fusion peut augmenter les performances de reconnaissance. En effet, sans remettre en cause le modèle 'la main en premier, ensuite les lèvres", la précision de la segmentation temporelle des cibles manuelle et labiale peut parfois entraîner des inversions entre ces instants cibles. Une amélioration de la précision permettrait ainsi de conserver l'ordre réel, et donc de réduire les erreurs d'appariement. De plus, les problèmes de sur-segmentation ou inversement de sous-segmentation engendrent forcément des erreurs supplémentaires. En l'absence de détecteur de ces deux problèmes, notre méthode touche ici ses limites.

Le modèle de fusion "la main en premier, ensuite les lèvres" dans le cas des voyelles est aussi la première implémentation de la réception du code LPC. Il tient en compte de l'avance temporelle de la main par rapport aux lèvres comme révélée par Attina *et al.* (2004) dans leur étude sur la production du code LPC. Il est aussi cohérent avec les résultats de perception du code LPC obtenus par les mêmes auteurs (voir aussi Cathiard *et al.* (2004) qui ont démontré une identification progressive de la parole avec le code LPC, s'illustrant d'abord par une sélection d'un sous-ensemble de consonnes et de voyelles candidates à partir de l'information manuelle, puis par le choix du phonème parmi les candidats lorsque l'information labiale est finalement disponible.

Des perspectives

Ce travail de thèse représente une première modélisation du code LPC en réception. L'objectif final du projet TELMA , certes, n'est pas achevé. Cependant, ce travail ouvre plusieurs chantiers qui peuvent d'abord achever ce projet et ensuite lui apporter des améliorations.

Tout d'abord, une approche probabiliste du modèle de fusion "la main en premier, ensuite les lèvres" peut aussi être expérimentée. Les probabilités qui sont disponibles à chacune des étapes du traitement (au niveau du codage LPC de la main et au niveau de la classification des

lèvres) peuvent être couplées pour obtenir un treillis de phonèmes. Ceci sera nécessaire dans la perspective d'appliquer des traitements de haut niveau tels que les modèles de langage, incluant des contraintes syntaxique et sémantique, qui peuvent sélectionner par exemple le chemin le plus probable dans le treillis. Cette approche peut améliorer notre processus de reconnaissance, étant donné qu'en perception du code LPC, les pseudo-mots (sans possibilité de prédiction par le contexte sémantique) sont moins bien identifiés que les mots (Alegria *et al.*, 1999).

Par ailleurs, les paramètres labiaux utilisés dans ce travail ne sont pas exhaustifs. Nous avons vu par exemple que l'ajout de nouveaux paramètres, comme ceux du pincement, augmente les performances de reconnaissance visuelle des syllabes CV. D'autres paramètres, qui peuvent être extraits des contours de lèvres peuvent apporter eux aussi des améliorations pour discriminer notamment les consonnes. Dans ce sens, l'aire de chaque lèvre, qui peut être obtenue à partir des contours labiaux, pourrait s'ajouter à nos paramètres. Etant donné que les consonnes ne sont pas toutes articulées aux lèvres, l'approche "modèle" pour extraire l'information labiale semble être limitée. Une combinaison avec l'approche "image" sera sûrement bénéfique à la discrimination des phonèmes aux lèvres. Par exemple, une transformation des pixels de l'image contenus dans le contour interne des lèvres peut donner en effet des indices sur le rôle de la langue et des dents dans l'articulation de certaines consonnes par exemple.

Ensuite, il est à noter que les résultats de ce travail sont obtenus sous certaines conditions. En effet, l'extraction des paramètres de la main et des lèvres est facilitée par l'utilisation de certains artifices. Un traitement d'image supplémentaire est nécessaire pour pallier cet aspect "non naturel". Dans le cas des lèvres, de nombreux travaux ont été effectués dans ce sens et d'autres sont en cours pour extraire notamment le contour interne des lèvres sans aucune contrainte. Pour le cas de la main dans le contexte du code LPC, se passer des artifices est un problème encore non élucidé du fait que la main est souvent en contact direct avec le visage. Cependant, on peut imaginer dans ce cas qu'un système décodeur des gestes LPC peut raisonnablement utiliser certains artifices sur la main (un gant coloré ou un gant avec des pastilles) sans avoir des conséquences sur les aspects d'usage.

Notre méthode de détection de la cible vocalique aux lèvres, rappelons-le, s'appuie sur l'étiquetage phonétique du signal acoustique obtenu par un un alignement forcé. Ce dernier utilise la connaissance a priori de la transcription phonétique des phonèmes. Une des améliorations possibles de notre chaîne de traitements est de développer une méthode de segmentation du signal acoustique de parole n'utilisant pas la transcription phonétique.

Enfin, cette première modélisation du code LPC en réception doit être étendue à d'autres participants pour évaluer sa robustesse. Dans ce sens, certaines parties de notre chaîne de traitements ont été déjà appliqué à un participant sourd (le décodage des gestes LPC de la main, analyse du flux labial pour le cas des voyelles). Il serait donc intéressant de voir si les modèles appris sur notre participant étudié ici seront exploitables pour ce participant sourd.

Annexes

Annexe A : visèmes des consonnes pour l'anglais

Dans cette annexe, nous présentons des notes descriptives des expériences pour catégoriser en visèmes les consonnes pour l'anglais et dont nous avons rapporté les résultats dans la table 1.1 dans le chapitre 1.

- Auteurs : Heider et Heider (1940)
- Stimuli : 20 syllabes [Cɔɪ] et 20 syllabes [Ci] présentées en ordre dispersé et a priori en direct
- Sujets et réponses : 39 enfants sourds dans une école pour les sourds
- Locuteur (s) : enseignants connaissant les enfants sujets
- Critère du regroupement : 75% des consonnes présentées sont identifiées dans leurs groupes
- Visèmes de consonnes : [p, b, m] ; [f,v] ; [r] ; [θ] ; [ʃ,tʃ,dʒ] ; [n,t,d] ; [l] ; [k,g].

- Auteurs : Woodward et Barber (1960)
- Stimuli : film avec le son en noir et blanc de paires de syllabes C_1V-C_2V (C_1V peut être la même que C_2, la voyelle=[ɑ]). Sujet filmé en vue de face avec la tête et les épaules.
- Sujets et réponses : les sujets sont tous des normo-entendants parlants anglais : 38 adultes s'intéressant à l'éducation des sourds et 147 étudiants universitaires. Les sujets indiquent comme réponse si les stimuli sont identiques ou différents.
- Locuteur (s) : étudiante diplômée en apprentissage linguistique.
- Critère du regroupement : hiérarchie du contraste visuel.
- Visèmes de consonnes : [p,b,m] ; [f,v] ; [w,r,hw] ; [t,d,n,l,θ,ð,s,z,tʃ,dʒ,ʃ,ʒ,j,k,g,h].

- Auteurs : Fisher (1968)
- Stimuli : consonnes localisées au début ou à la fin des mots. Film blanc & noir
- Sujets et réponses : 18 étudiants collégiens normo-entendants. Les sujets regardent en groupe le film sur un écran et désignent le mot parmi un ensemble de choix possibles en dessinant une ligne en dessous. Les distances entre les sujets et l'écran varient entre 1.8, 2.7 et 3.6 mètres
- Locuteur (s) : 6 adultes
- Critère du regroupement : les groupes sont sélectionnés sur la base des confusions de consonnes dont le taux est supérieur au taux du hasard.
- Visèmes de consonnes :
  - Initiales : [p,b,m,d] ; [f,v] ; [w,hw,r].
  - Finales : [p,b] ; [f,v] ; [ʃ,ʒ,dʒ,tʃ] ; [t,d,n,θ,ð,s,z,r,l].

- Auteurs : Binnie *et al.* (1974)
- Stimuli : une série de 16 syllabes CV formées avec 16 consonnes de l'Anglais combinées avec la voyelle [ɑ]. Conditions d'éclairage améliorées avec l'utilisation de deux réflecteurs (un à angle d'incidence de 45° et l'autre à incidence faciale) afin de faciliter la vision des mouvements articulatoires.
- Sujets et réponses : dix étudiants normo-entendants : 2 de sexe masculin et 8 féminin. Les sujets sont inscrits à des cours élémentaires concernant l'enseignement de la lecture labiale.
- Locuteur (s) : de sexe féminin et âgée de 24 ans, elle lit les stimuli (un toutes les 5 secondes) en Anglais américain.
- Critère du regroupement : matrice de confusion montrant des clusters importants.
- Visèmes de consonnes : [p,b,m] ; [f,v] ; [ʃ,ʒ] ; [θ,ð] ; [n,d,t,s,z,k,g].

- Auteurs : Walden. *et al.* (1977)
- Stimuli : 20 syllabes CV composées de 20 consonnes de l'Anglais américain dans le contexte de la voyelle [a]. Film en couleur et illumination intense pendant l'enregistrement.
- Sujets et réponses : 31 adultes mal-entendants de sexe masculin. Ils ont acquis la parole normale et développé des capacités linguistiques avant d'être confrontés à leurs pertes d'audition. Aucun n'a été formé à la lecture labiale auparavant. L'entraînement consistait à 38 exercices à la lecture labiale. Chaque exercice est organisé pour un ensemble de 4 à 9 des syllabes CV incluses dans le test.
- Locuteur (s) : adulte de sexe masculin avec une articulation normale. Il a été filmé vue de face en faisant apparaître la tête et les épaules.
- Critère du regroupement : seuil de 75% sur les présentations dans lesquelles les consonnes étaient bien identifiées dans leur classes de visèmes.
- Visèmes de consonnes :
  - Avant entraînement : [p,b,m] ; [f,v] ; [w] ; [θ,ð] ; [ʃ,ʒ,s,z] ; non classées : [t,d,n,k,g,r,l,j].
  - Après entraînement : [p,b,m] ; [f,v] ; [w] ; [ʃ,ʒ] ; [θ,ð] ; [r] ; [s,z] ; [t,d,n,k,g,j] ; [l].

– Auteurs : Walden *et al.* (1981)
– Stimuli : 22 consonnes en anglais sont prononcées dans le contexte [ɑ/-C-/ɑ] de façon claire et naturelle avec 5 secondes de silence entre les syllabes. La locutrice est filmée de face avec la tête et les épaules, l'éclairage est direct et relativement intense.
– Sujets et réponses : 35 adultes de sexe masculin âgés de 19 à 68 ans avec une perte d'audition à haute fréquence, inscrits dans un programme de réhabilitation auditive. Un pré-test est pratiqué pour se familiariser avec les consonnes et leurs symboles (pour noter les réponses). La liste des consonnes est toujours disponible.
– Locuteur (s) : une adulte expérimentée en communication auditive et visuelle avec les mal-entendants.
– Critère du regroupement : seuil de 75% sur les présentations dans lesquelles les consonnes étaient bien identifiées dans leur classes de visèmes.
– Visèmes de consonnes :
  - Avant entraînement : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ;[ʃ,ʒ,tʃ,dʒ] ; non classées : [t,d,n,s,k,g,l,j].
  - Après entraînement : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ;[ʃ,ʒ,tʃ,dʒ] ; [t,d,n,s,k,g,l,j].

– Auteurs : Kricos et Lesner (1982)
– Stimuli : 21 consonnes (C) dans un contexte [ɑCɑ] sont présentées en vidéo blanc et noir. Les locutrices étaient filmées de face avec la tête et les épaules. L'éclairage était focalisé sur les bouches des locutrices.
– Sujets et réponses : 12 étudiantes normo-entendantes âgées de 18 à 23 ans avec a priori aucune expérience en lecture labiale ni en phonétique. Les stimuli étaient présentés sur un écran TV éloigné de 1,5m aux sujets assis en paires. Les sujets écrivaient la consonne reconnue et choisie dans une liste de symboles orthographiques.
– Locuteur (s) : 6 étudiantes ayant une articulation et une intelligibilité auditive normales.
– Critère du regroupement : seuil de 75% sur les présentations dans lesquelles les consonnes étaient bien identifiées dans leur classes de visèmes.
– Visèmes de consonnes : pour les six locuteurs
  - Locuteur 1 : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,z,s] ; [l] ; [k,j,h,g,ŋ].
  - Locuteur 2 : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,z,s] ; [l] ; [k,j,h,g,n,ŋ].
  - Locuteur 3 : [p,b,m] ; [f,v,r] ; [w] ; [θ,ð] ; [ʃ,ʒ,tʃ,d] ; [k,g].
  - Locuteur 4 : [p,b,m] ; [f,v] ; [w,r] ; [θ,ð] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,z,s].
  - Locuteur 5 : [p,b,m] ; [f,v,z,s] ; [w,r] ; [ʃ,ʒ,tʃ,dʒ].
  - Locuteur 6 : [p,b,m] ; [ʃ,ʒ,tʃ,dʒ] ; [w,r,θ,ð] ; [t,d,z,s] ; [l,j,h,n,l].

- Auteurs : Owens et Blazek (1985)
- Stimuli : des séries de logatomes de type VCV formés par 23 consonnes placées dans le contexte de 4 voyelles. Enregistrement vidéo en couleur avec un son en haute qualité enregistré par un microphone séparé. Eclairage intense de face et direct par un parapluie lumière.
- Sujets et réponses : 5 sujets normo-entendants et 5 autres mal-entendants tous âgés entre 22 et 62 ans. Chaque sujet est testé seul dans une petite chambre sourde. Les réponses des sujets sont transcrites par un seul examinateur. La transcription de cet examinateur a été évalué satisfaisante sur la base d'une comparaison avec la transcription faite par un expert en transcription phonétique lors du test d'un sujet.
- Locuteur (s) : une audiologiste âgée de 26 ans expérimentée en communication avec des patients mal-entendants, sélectionnée parmi d'autres sur des critères qui la classent comme une personne relativement facile à lire sur ses lèvres.
- Critère du regroupement : seuil de 70-75%.
- Visèmes de consonnes :
  - Context [ɑCɑ] : [p,b,m] ; [f,v] ; [θ,ð] ; [w,r] ; [ʃ,ʒ,tʃ,dʒ] ; [n,k,g,l] ; [h].
  - Context [ʌCʌ] et [iCi] : [p,b,m] ; [f,v] ; [θ,ð] ; [w,r] ; [ʃ,ʒ,tʃ,dʒ] ; [t,d,s,z] .
  - context [uCu] : [p,b,m] et [f,v].

- Auteurs : Summerfield (1987)
- Stimuli :stimuli visuels de consonnes prononcées en contexte [ɑ ː] sont présentés à des sujets pour un test perceptif d'identification. Il utilise ensuite une classification hiérarchique pour regrouper les consonnes qui ont été confondues par les interlocuteurs.
- Sujets et réponses : sujets mal-entendants expérimentés (donc ayant eu un entraînement) en lecture labiale.
- Critère du regroupement : seuil de 75% sur les présentations dans lesquelles les consonnes étaient bien identifiées dans leur classes de visèmes.
- Visèmes de consonnes : [p,b,m] ; [f,v] ; [θ] ; [ʃ,ʒ] ; [d,t] ; n,g,k] ; [s,z] ; [l] ; [r] ; [w] ; [y].

- Auteurs : Massaro *et al.* (1993)
- Stimuli : 66 syllabes CV combinant 22 consonnes initiales avec 3 voyelles. Les syllabes sont répétées deux fois. Les 132 syllabes sont présentées de façon aléatoire en 3 conditions modales : audio seul, vidéo seul ou bimodale (audio+vidéo) et ceci selon que le débit est normal ou rapide.
- Sujets et réponses : étudiantes âgées de 18 à 32 normo-entendantes. Les sujets ont été testés et entraînés simultanément dans des pièces séparées atténuant le son.
- Locuteur (s) : de sexe masculin (stimuli enregistré par Bernstein et Eberhardt (1986) d'après Massaro et al. (1993)).
- Critère du regroupement : matrice de confusion.
- Visèmes de consonnes : [p,b,m] ; [s] ; [ʃ,tʃ,j] ; [d,t, n,g,k, h] ; [l] ; [r] ; [w].

Annexe B : complément sur les études sur l'efficacité du code LPC

Cette annexe traite plus en détail l'efficacité du code LPC. Dans la première partie, nous décrivons quelques études montrant l'efficacité du code LPC en perception. Dans la seconde partie, nous présentons d'autres études discutant de l'efficacité du code sur le développement du langage parlé.

Efficacité perceptive du code LPC

L'intérêt du code réside dans son efficacité à améliorer la perception de la parole. D'ailleurs c'est la raison principale pour laquelle ce système a été inventé. L'expansion du *Cued Speech* dans le monde entier par ses adaptations aux différentes langues et son utilisation croissante dans plusieurs milieux (en famille ou à l'école), témoignent de l'efficacité de ce système pour une bonne réception de la parole. Le site internet de l'ALPC présente des témoignages concrets de parents utilisant le code LPC pour communiquer avec leurs enfants sourds et qui atteste de l'apport perceptif du code LPC.

Sur le plan expérimental, plusieurs études ont été menées sur la réception des différentes versions du *Cued Speech* dans le monde. Pour ce qui nous concerne, nous nous focalisons dans cette partie que sur les études relative au *Cued Speech* dans sa version originale et à le code LPC. Pour le *Cued Speech*, les premières études avaient commencé par Ling et Clarke. Deux études réalisées par ces deux auteurs évaluaient un apport relativement modeste à la lecture labiale des clés du *Cued Speech* pour des enfants sourds.

La première étude (Ling et Clarke, 1975) portait sur la perception de la parole codée chez des enfants sourds. Ces derniers étaient au nombre de 12, âgés de 7 à 12 ans au moment de l'expérience et étaient exposés au code du *Cued Speech* durant une année auparavant. Leur tâche consistait à décoder les stimuli qu'ils avaient perçus en condition de lecture labiale seule et lecture labiale avec les clés du *Cued Speech*. Les stimuli testés étaient de deux types. Il s'agissait soit de mots placés dans des expressions de type "boy and girl" ; soit de phrases simples composées de quatre mots comme par exemple "she has five books". En condition "lecture labiale seule", les résultats étaient en moyenne de 9% pour les phrases entièrement reconnues et de 35% d'identification correcte pour les mots dans les expressions. Si, en plus, les mots dans les phrases sont eux aussi comptabilisés, le score moyen d'identification des mots s'élève à 50%. En condition lecture labiale + clés, les performances du décodage des enfants

augmentent certes, mais les écarts ne dépassent 18% dans les trois cas : 17% de bénéfice pour les mots dans les expressions, 18% pour tous les mots dans les expressions et dans les phrases et aucun apport suplémentaire dans le cas des phrases. Ces cas favorables conduisent les auteurs à conclure que les sujets ont une faible expérience avec le système et à suggérer une suite à cette étude.

Une année plus tard, les deux auteurs reviennent avec une nouvelle étude (Clarke et Ling, 1976) sur huit enfants sourds choisis parmi ceux de la première étude. Les enfants participant à cette deuxième étude avaient donc deux ans d'exposition au *Cued Speech*. Ceci est probablement la raison pour laquelle les scores d'identification des phrases augmentent. On note 23% d'identification correcte des phrases en condition lecture labiale seule et 68% avec en plus les clés du *Cued Speech*. Comparé au 9% obtenu lors de la première étude, on remarque que cette fois-ci, l'ajout des clés est beaucoup plus avantageux. Ainsi, on peut conclure de cette expérience que la durée d'exposition au code joue un rôle important dans l'efficacité perceptive du code LPC. Pour un détail sur ce rôle nous renvoyons vers une revue d'études menée dans Attina (2005).

Uchanski *et al.* (1994) vont plus loin et testent la réception de la parole conversationnelle suivant la complexité sémantique du contexte et des phrases complexes. Ils choisissent de tester quatre sujets adultes et sourds âgés de 18 à 27 sur des phrases codées en *Cued Speech*. Les sujets étaient suffisamment expérimentés au code puisqu'ils étaient exposés intensivement au *Cued Speech* pendant au moins huit ans à la maison et à l'école. Les phrases ont été extraites de deux listes contenant de la parole conversationnelle selon que le degré de prédictibilité du contexte (liste Clarke avec un fort contexte prédictible et liste CID) et d'une troisième liste contenant des phrases difficiles dont le contexte est très peu prédictible (liste Harvard). Les sujets avaient pour tâche d'identifier un certain nombre de mots clés dans les phrases et de noter les réponses par écrit. Les phrases ont été présentées soit dans la condition "lecture labiale seule" soit dans la condition "lecture labiale + clés du *Cued Speech*". Dans la condition "lecture labiale seule", les scores moyens d'identification correcte des mots sont de 62% pour les phrases de la liste CID, 45% pour la liste Clarke et 25% pour la liste Harvard. Avec les clés manuelles du *Cued Speech*, les scores moyens augmentent et atteinent 97% de mots clés correctement identifés pour la parole conversationnelle et 84% pour la liste Harvard. Ces résultats montrent que le *Cued Speech* apporte un complément important pour la perception des mots et même dans un contexte difficile et peu prédictible.

Deux autres études menés au MIT viennent confirmer ces résultats. Bratakos *et al.* (1998) dans une étude menée pour évaluer les différences entre un locuteur-codeur humain et un système de génération du code du *Cued Speech*, trouvent des résultats semblables à ceux d'Uchanski *et al.* (1994). Ils ont testé 6 sujets adultes âgés de 19 à 27 ans sur des phrases complexes provenant de la liste Harvard. Les sujets, dans cette expérience, sont expérimentés dans la réception du *Cued Speech* et ont utilisé le code durant une période de 12 à 23 ans avec un parent ou avec un professionnel de la translittération. Les phrases étaient présentées dans des conditions seulement visuelles (donc sans son) et les réponses étaient notées par écrit. Les scores d'identifications de mots clés dans les phrases s'améliorent nettement en ajoutant les clés du *Cued Speech* à la lecture labiale (ils passent de 30% en lecture labiale seule à 84% si on ajoute les clés manuelles). Duchnowski *et al.* (2000), dans une expérience semblable à celle de Bratakos *et al.* (1998) avec

un matériel similaire, trouvent aussi des scores comparables (35% de mots clés correctement identifés pour la lecture labiale seule et 91% en la complétant par les clés du *Cued Speech*).

En français, des études ont confirmé globalement pour la LPC les résultats vu précédemment dans le cas du *Cued Speech*. L'étude de Nicholls et Ling (1982) (voir chapitre 1) a été élargie par Charlier *et al.* (1990) à plusieurs sujets communicant avec le code LPC dans différents lieux. Les sujets étaient 55 enfants sourds âgés de 5 à 16 ans dont 14 d'entre eux utilisaient le LPC à la maison et l'école et les autres seulement à l'école. Leur tâche était d'identifier des phrases présentées dans deux conditions : lecture labiale seule ou avec le code LPC. La procédure consistait au choix d'un dessin pour une phrase présentée. Les phrases présentées étaient classées dans trois catégories (facile, moyennement difficile et difficile) selon le degré de l'ambiguïté de l'information labiale par rapport aux dessins. Les résultats obtenus montrent une amélioration nette des performances d'identification dans tous les groupes lorsque les clés du code LPC viennent assister la lecture labiale. De plus, avec des phrases isolées du contexte conversationnel, l'apport du code LPC était significatif. Ceci supposerait qu'en condition de parole conversationnelle les performances devraient augmenter.

En conclusion, la lecture labiale ne transmet qu'une partie de l'information. L'autre partie peut être transmise aussi visuellement par les clés du code LPC. L'ambiguïté de la lecture labiale peut donc être réduite par le code LPC. Ainsi, la perception d'un message de parole codé par les deux modalités visuelles ne diffère guère de la perception de la parole orale non codée. Ceci indique que les enfants sourds peuvent bénéficier de ce code pour développer leur langage oral avec des performances similaires aux personnes entendantes.

Efficacité sur le développement du langage parlé

En plus de son efficacité dans la perception d'un message oral, le code LPC permet l'accès à une représentation complète du système phonologique pour les malentendants exposés à cette méthode depuis leur plus jeune âge, avec un impact positif sur le développement du langage. C'est ce qui a été montré dans plusieurs études (notamment : (Kipila, 1985; Cornett, 1990; Hage *et al.*, 1990, 1991; Metzger, 1994; Leybaert et Charlier, 1996; Leybaert, 1998, 2000; Charlier et Leybaert, 2000; Leybaert et Lechat, 2001)) nous présenterons brièvement les conclusions dans cette section (pour une revue détaillée voir (Attina, 2005) p. 28-30 ; (Alegria et Leybaert, 2005)). Ces conclusions concernent spécialement le rôle positif du *Cued Speech* pour :

- Le développement du langage
- L'apprentissage de l'écriture avec le développement de la phonique et l'orthographe
- L'apprentissage de la lecture
- Le développement de la mémoire de travail

Concernant le premier point, Kipila (1985) dans une étude de cas a trouvé des structures morphologiques utilisées dans le codage d'enfants sourds âgés d'un peu plus de 5 ans (5 ans et 4 mois). Parmi ces structures, certaines étaient utilisées avec une précision de 100% et d'autres avec une précision inférieure. Les premières structures incluaient le temps (régulier et irrégulier), les pluriels, la troisième personne irrégulière et les possessives. Les autres structures incluaient par exemple les articles, le présent progressif et la troisième personne régulière. Beaucoup de ces

structures n'avaient été atteintes qu'au cinquième année du développement du langage chez les enfants normaux. Dans son étude, Metzger (1994) teste les mêmes enfants utilisés dans l'étude de Kipila (1985) à l'âge de 11 ans. Il se focalise sur six des structures relevées par Kipila : trois notées à une précision de 100% (le temps régulier et irrégulier ainsi que les pluriels) et trois notées à une précision inférieure (les articles, le présent progressif et le *contractible copula*). Les résultats montrent que ces six structures sont toutes démontrées à 100% de précision suggérant que la nature de développement du langage codé est, selon l'auteur, identique à celle du langage parlé.

D'un autre coté, les enfants sourds exposés intensivement et précocement au codage LPC développent des représentations phonologiques des mots dans leurs langage. Ils peuvent ainsi apprendre des généralisations phoniques de l'orthographe de la même façon que les enfants entendants. Ceci constitue un moyen fondamental pour ces enfants sourds, de développer des formes d'écriture du langage parlé de façon similaire aux enfants bien entendants. Leybaert et Charlier (1996) ont conduit une étude sur les compétences phoniques des étudiants pré-linguistiquement sourds profonds et les ont comparées à celles des étudiants entendants. Une partie des étudiants sourds était exposée au code LPC à la maison et à l'école (donc intensivement) et l'autre partie était exposée seulement à l'école. La tâche des participants consistait à épeler les mots qui correspondaient aux images qu'on leur montrait. Les résultats montraient une similarité entre erreurs commises par les étudiants sourds intensivement exposés au code LPC et celles des étudiants entendants. La majorité des erreurs étaient précises du point de vue phonologique (c'est-à-dire, les mots erronés pouvaient être prononcés identiquement aux mots épelés correctement). En revanche, les erreurs des étudiants sourds exposés au code LPC seulement à l'école, marquaient un important contraste avec celles des deux premiers groupes de participants. Ceci peut servir d'argument pour insister sur l'intensité de l'exposition au code LPC. L'apprentissage du code LPC est relativement facile mais l'exposition intensive et précoce à ce code n'est pas sans avantages.

Concernant le troisième point, les personnes ayant bénéficié d'une exposition précoce au codage LPC durant leur enfance montrent des facultés de compréhension de la lecture et de la phonologie au niveau de celles des personnes entendantes. Les enfants exposés au code LPC montrent des compétences de génération et jugement des rimes similaires à celles des personnes entendantes. Dans premier temps, Charlier et Leybaert (2000) ont étudié les compétences de jugement de rimes chez des enfants sourds répartis en deux groupes (des enfants ayant bénéficié précocement du code LPC en famille et des enfants n'ayant bénéficié du code LPC qu'à l'école). Ces compétences ont été comparées à celle d'enfants entendants effectuant la même tâche. Tous les participants avaient pour tâche de juger s'il y a rimes ou non sur des paires de dessins. Ces dernières étaient réparties en deux : des paires qui rimaient et des paires qui ne rimaient pas. Les dernières paires étaient elles aussi réparties en deux : des paires où les mots sont difficiles à discriminer en lecture labiale, et des paires de mots faciles à discriminer en lecture labiale. Les sujets bénéficiant du code à la maison présentaient des performances élevées (supérieures à 90%) et identiques à celles des sujets entendants, tandis que les performances des sujets ayant bénéficié du code à l'école étaient inférieures surtout pour deux types de paires (les paires qui rimaient avec une orthographe différente et les paires qui ne rimaient pas mais qui étaient difficiles à discriminer par la lecture labiale). Dans un second temps, les auteurs ont mené une autre expérience où des

sujets étaient sensés produire des rimes. Une série de mots cibles ont été présentés aux sujets pour qu'ils produisent par écrit deux mots rimant avec chacune des cibles. Les sujets utilisés étaient des enfants sourds et répartis en deux groupes : un groupe LPC-maison où les enfants étaient exposés au code LPC à la maison et à l'école et un groupe LPC-école avec des enfants éxposés au code seulement à l'école. Chaque groupe de sujets sourds était associé à un groupe de sujets entendants (groupe de contrôle) afin d'apparier chaque sujet sourd avec un autre entendant. Les résultats de cette expérience ne montrent pas de surprises. Les sujets du groupe LPC-maison présentaient des performances élevées (pourcentage de réponses correctes autour de 80%) significativement non différentes à celles de leur groupe de contrôle (90% de réponses correctes). Les performances du groupe LPC-école étaient quant à elles significativement inférieures à celles de leur groupe de contrôle.

En résumé, et pour ne pas aller plus loin dans notre description des expériences, le code LPC est avantageux pour les enfants sourds dans la mesure où il permet une acquisition de connaissances des mots (surtout ceux ayant une signification) et aussi de la morpho-phonologie de la langue parlée. Avec l'aide du code LPC, les enfants développent des représentations phonologiques précises de la parole qui leur permettent de juger des rimes correctes et de les produire. Enfin, les enfants exposés au code LPC peuvent développer des compétences en lecture et en orthographe similaires à celles des enfants entendants.

Annexe C : les effets mutuels du contexte entre les consonnes et les voyelles

Dans cette annexe nous décrivons des études ayant montré l'influence mutuelle du contexte entre les consonnes et les voyelles. La première section concerne les effets du contexte de la voyelle sur la lecture labiale de la consonne. La seconde section est centrée sur les effets du contexte de la consonne sur la lecture labiale de la voyelle.

Effets du contexte vocalique sur la lecture labiale de la consonne

Benguerel et Pichora-Fuller (1982) comparent l'intelligibilité de la lecture labiale pour neuf consonnes (C ∈ [p, t, k, f, s, w, ʃ, tʃ, θ]) dans le contexte V_1CV_2 de trois voyelles (V_1 et V_2 ∈ [i, u, æ]). Les séquences V_1CV_2 étaient prononcées par un locuteur choisi parmi six locuteurs de façon à faciliter la perception de ses gestes articulatoires par les lecteurs labiaux employés dans cette expérience (sujets normo-entendants et mal-entendants). Les résultats obtenus pour les sujets normo-entendants et mal-entendants confirment l'influence des effets de la coarticulation sur les performances de la lecture labiale. Cependant, ils montrent aussi que certains phonèmes sont beaucoup plus sensibles au contexte que d'autres. Les consonnes [p, f, w, θ] sont bien reconnues dans le contexte des trois voyelles. Par contre, le reste des consonnes ([t, k, s, ʃ, tʃ]) dépend considérablement du contexte vocalique adjacent. L'effet de ce contexte varie selon la voyelle avoisinant la consonne. En effet, l'étude montre que lorsque la voyelle arrondie [u] est contenue dans la syllabe, les scores d'identification des consonnes chutent, notamment pour la syllabe [uCu] qui est identifiée correctement à 58%. En revanche, les scores d'identification des syllabes qui ne contiennent pas la voyelle [u] varient de 70% à 78%. Les auteurs expliquent cette différence du comportement de la voyelle arrondie ([u]) par le fait qu'elle est bien visible sur les lèvres (l'arrondissment est le trait le mieux perçu visuellement) et qu'elle tend à dominer les consonnes qui la précèdent et principalement celles qui ne sont pas articulées au niveau des lèvres. Ceci montre que les consonnes sont influencées par le contexte phonétique de la voyelle et que cette influence est davantage importante si l'environnement phonétique contient une voyelle arrondie.

Le regroupement en visèmes des consonnes subit l'effet du contexte. C'est ce que Owens et Blazek (1985) ont tenté de montrer en étudiant la variation des groupes de visèmes des

consonnes par rapport au contexte vocalique. Ils ont utilisé des syllabes VCV formées par 23 consonnes de l'anglais dans le contexte des 4 voyelles [ɑ], [i], [u] et [ʌ]. Les résultats montrent clairement là aussi une grande influence du contexte vocalique. En effet, les groupes de visèmes consonantiques varient de façon considérable selon la voyelle avoisinante. Les visèmes passent ainsi de 7 groupes pour la voyelle [a] à 2 groupes seulement pour la voyelle [u] (les visèmes sont détaillés dans la table 1.1). Ainsi, nous trouvons là aussi que les voyelles arrondies dominent en terme de perception visuelle, les consonnes adjacentes.

Massaro *et al.* (1993) résument l'effet du contexte de la voyelle en classant les performances d'identification visuelle des syllabes CV selon le contexte vocalique. L'identification de la consonne est meilleure en contexte [a], intermédiaire en contexte [i] et démunie en contexte [u].

En ce qui concerne la langue française, les résultats obtenus précédemment pour l'anglais ont été confirmés pour leur majeure partie par Gentil (1981) et Benoit *et al.* (1994). Gentil (1981) montre en termes de pourcentages de reconnaissance visuelle en initiales et en finales de mots l'importance du contexte vocalique dans l'identification visuelle des consonnes et leur classement en visèmes. Plus précisément, l'auteur ordonne suivant les valeurs des pourcentages obtenus le degré de l'influence des voyelles [a], [u], [i] sur les consonnes. C'est dans le contexte vocalique [Ca] que les consonnes initiales de mots sont mieux reconnues. Elles sont ensuite moins reconnues dans le contexte vocalique [Ci] tandis que le contexte vocalique [Cu] donne les scores de reconnaissance les plus bas. Ces résultats rejoignent ainsi ceux obtenus par Massaro *et al.* (1993) pour l'anglais. D'autres part, il apparaît que l'effet du contexte de la voyelle [u] influe moins sur certaines consonnes en finale de mots ([z, ʒ, s, ʃ, f]) tandis que nous retrouvons une bonne perception des consonnes ([p, m, b, ʃ, ʒ]) placées en finale de mots dans le contexte de la voyelle [aC].

Le même classement des effets du contexte vocalique a été confirmé par Benoit *et al.* (1994) en considérant six consonnes ([b, v, r, l, z, ʒ]) placées dans des séquences de la forme [VCVCVz]. Ils ont choisi les mêmes voyelles que Gentil (excepté la voyelle [y] qui remplace la voyelle [u] mais les deux appartiennent au même visème) parce qu'elles représentaient les positions extrêmes du mouvement labial pour les voyelles du français.

Effets du contexte consonantique sur la lecture labiale de la voyelle

Benguerel et Pichora-Fuller (1982) montrent que les voyelles en initiale des séquences syllabiques V_1CV_2 sont mieux reconnues que celles en finale de syllabes. En particulier, ils observent pour le cas des voyelles en position finale que les erreurs de reconnaissance des voyelles [æ, i] sont relativement supérieures à celles de la voyelle [u]. Cette dernière est hautement visible sur les lèvres (grâce à l'arrondissement) et donc, quelque soit le contexte qui la précède et quelque soit sa position, sa reconnaissance visuelle est presque parfaite. En revanche, la voyelle [æ] en position finale est celle qui est la moins reconnue de toutes les voyelles. Les auteurs ont remarqué que les erreurs sont fréquentes (scores de reconnaissance inférieurs à 90%) quand cette voyelle est précédée par une des consonnes [t, k, s, ʃ, tʃ]. Etant elles-mêmes fréquemment mal peçues visuellement, ces cinq consonnes contribuent, par la variation de leurs articulations, aux

mauvaises reconnaissances de la voyelle [æ]. Pour la voyelle [i] en position finale, les scores de reconnaissance sont globalement meilleurs que celle de la voyelle [æ] et dependent du contexte VC qui la précède (on passe de 73% par exemple pour l'environnement [up-] à 100% pour l'environnement [æt-]). Les auteurs concluent tout de même que les voyelles en position finale sont moins reconnues que celles en position initiale. Ceci est dû, selon les auteurs, aux différences dans la production des deux voyelles et non pas à un effet sequentiel au niveau perceptif, rejoignant ainsi l'observation d'Öhman (1966) selon laquelle les voyelles en finale, dans une prononciation de VCV en monotone, tendent à se neutraliser plus que celles en initiale (la coarticulation trans-consonantale). Afin de confirmer cette explication, Benguerel et Pichora-Fuller (1982) ont effectué un test complémentaire où l'on a joué les séquences videos des syllabes V_1CV_2 en sens inverse.

En plaçant les voyelles [i, ɪ, ɑ, ʊ, u] dans des contextes consonantiques symétriques CVC et dans des contextes consonantiques assymétriques, Montgomery et Jackson (1983) explorent les effets du contexte consonantique sur la perception visuelle des voyelles par 30 sujets malentendants. Ils ont travaillé sur un corpus composé de 52 monosyllabes de type CVC (C est une consonne parmi [p, b, f, v, t, d, g, ʃ]) et des trois contextes [hVg], [wVg] et [rVg]. L'ensemble des consonnes a été choisi de façon à représenter en même temps les consonnes hautement visibles ayant des composantes articulatoires labiales ([p, b, f, v, ʃ]), et les consonnes moins visibles aux lèvres ([d, t, g, h]). Les résultats obtenus pour les locutrices enregistrées pour cette étude montrent d'abord que parmi les voyelles utilisées c'est la voyelle [ɑ] qui est la plus intelligible tandis que la voyelle [ʊ] est la moins intelligible. Ensuite, les auteurs distinguaient leurs cinq voyelles en deux groupes : les voyelles dites "tendues" (les voyelles [i, ɑ, u]) et les voyelles dites "molles" (les voyelles [ɪ, ʊ]). Ils concluaient que les voyelles du premier groupe sont plus facilement reconnaissables (en visuel) que celles du second groupe. Deux explications peuvent être données : la première concerne la durée des voyelles ; en effet, la durée des voyelles [i, ɑ, u], systématiquement plus longue que celle des voyelles [ɪ, ʊ], peut favoriser l'identification visuelle. La seconde consiste à dire que les voyelles "tendues" sont généralement plus résistantes aux influences coarticulatoires des consonnes environnantes et peuvent être produites avec des gestes articulatoires plus distinctifs. Un autre résultat important que nous pouvons tiré de cette étude est que la reconnaissance visuelle des voyelles est perturbée dans un contexte qui présente une composante labiale distinctive (par exemple dans le contexte des consonnes labialisées comme [p, b, f, v, ʃ]) et qu'elle est nettement meilleure dans un contexte neutre, du point de vue de la labialité (par exemple le contexte [hVg]). Cependant, même si dans le contexte des consonnes ayant des formes labiales fortes la reconnaissance des voyelles est moins bonne, il est à remarquer que dans le cas des occlusives, avec la rapidité des gestes d'ouverture et de fermeture, le labiolecteur a plus d'information sur la durée de la voyelle et donc une meilleure identification.

Les résultats vus précédemment pour l'anglais restent en grand partie observés pour le français. A ce sujet, Cathiard (1988) a mené une étude afin de déterminer l'importance de la protrusion des lèvres dans l'identification visuelle des deux voyelles [i] et [y] placées dans le contexte consonantique CV, C étant une des deux consonnes [s] et [ʃ]. En choisissant des phonèmes formant des oppositions minimales sur la protrusion des lèvres ([y] est plus protruse que [i/ ; idem [ʃ] par rapport à [s]), elle montre qu'il est plus difficile de reconnaître la voyelle

[i] dans le contexte [ʃ] que dans le contexte [s]. De plus, la protrusion de la consonne [ʃ] affecte plus l'identification de la voyelle [i] que la voyelle [y] puisque 69% des contextes [ʃ -] est identifié comme [ʃy].

Dans le même type de données, Tseva et Cathiard (1990) élargissent à 8 phonèmes du français le corpus de l'étude. Elles ont choisi de mettre en contexte CV les voyelles [i, e, y, ø] et les consonnes [s, z, ʒ, ʃ] ; donc, des phonèmes qui forment des oppositions minimales sur la protrusion labiale. Elles montrent que l'anticipation de la protrusion des consonnes [ʃ, ʒ] entame l'intelligibilité des voyelles [i, e] en passant d'un score d'identification de 72% en contexte [s] et [z] à 91% en contexte [ʃ] et [ʒ]. Cette anticipation se manifeste dans le fait que les voyelles [i, e], qui sont produites habituellement avec des formes de lèvres étirées, se retrouvent avec une frome protruse si elles sont mises en contexte de consonnes protruses. En revanche, la tendance s'inverse pour les voyelles protruses [y, ø/ : on passe de 70% en contexte de [s] et [z] à 84% en contexte [ʃ] et [ʒ]. C'est en effet les mêmes conclusions de Benoit *et al.* (1994) qui trouvent qu'une voyelle située en contexte de consonne protruse est moins bien reconnue.

Annexe D : l'influence de l'angle de vue

Cette annexe est consacrée à une discussion sur l'influence de l'angle de vue sur les résultats de certaines expériences en perception et en reconnaissance de la parole.

Dans les tests de perception visuelle de la parole, dans certains sont présentés dans le chapitre 1, les auteurs choisissent de présenter leurs stimuli visuels sous des angles de vue différentes. Ceci prouve en quelque sorte que l'information visuelle perçue dépend en partie de ce facteur de visibilité. Ce dernier a été l'objet de plusieurs études (parmi lesquels : Neely (1956); Larr (1959); Nakano (1961); Berger *et al.* (1971); Erber (1974); Cathiard (1988, 1994); Adjoudani (1998)).

A ce sujet, après avoir passé en revue les travaux effectués avant 1974 (parmi lesquels : Neely (1956); Larr (1959); Nakano (1961); Berger *et al.* (1971)), Erber (1974) a conduit une étude analysant les effets de trois facteurs sur la réception visuelle de la parole chez des enfants sourds profonds. L'étude concernait les effets de l'angle verticale de la lumière incidente, de l'angle horizontale d'observation et de la distance sur la lecture labiale de 3 sujets sourds profonds. Dans cette expérience, les 3 sujets lecteurs étaient placés à une distance de 6, 12 ou 24 pieds du locuteur. L'incidence de la lumière illuminant le locuteur variait suivant les angles suivantes : 0°, 45° ou 90°. De même, les sujets étaient positionnés par rapport au lecteur une angle de vue de : 0° (vue de face), 45° (vue en 3/4) ou 90° (vue de profil). La tâche des sujets était de reconnaître 240 mots prononcés par une locutrice âgée de 26 avec une articulation normale. Concernant l'angle de vue, les résultats montrent que les angles de vue de face et de 3/4 des pourcentages équivalents de reconnaissance des mots. Sous un angle de vue de profile, les scores sont inférieurs (près de 14,3 à 22,5% d'erreurs en plus). En montrant une équivalence sur les performances de la face et du 3/4, Erber (1974) contredit donc Neely (1956) qui classait les trois angles de vue suivant les pourcentages obtenus en : vue de face (66%), vue de 3/4 (62%) et vue de profil (58%). Il contredit aussi Larr (1959) et Nakano (1961) qui montraient un avantage du 3/4 sur la face. Pour expliquer ces différences, l'auteur met en cause les conditions d'éclairage et la composition des stimuli.

Erber (1974) constate donc un avantage nettement supérieur des vues de face et de 3/4 sur la vue de profil. Summerfield (1987) trouve cet avantage peu significatif. Cependant, il est clair qu'une perte d'à peu près 14% à 22%, sachant que le score maximal obtenu par Erber (1974) est de près de 85%, est à prendre en compte.

Cathiard (1988) a étudié la protrusion en français en testant l'identification visuelle des syllabes [si], [sy], [ʃi] et [ʃy]. Les sujets devaient identifier la syllabe prononcée et présentée en

image correspondant à la réalisation de la consonne ou de la voyelles sous deux vues : vue de face ou vue de profil. Les résultats montrent que les scores d'identification sont équivalents dans ces deux conditions. L'auteur a ensuite analysé les données avec une analyse factorielle des correspondances et a montré les paramètres géométriques prédominant dans l'identification de chaque classe :

- l'étirement pour le [i] et le [s] de la syllabe [si],
- la protrusion pour le [y] des syllabes [sy] et [ʃy],
- l'aperture pour le [i] de la syllabe [ʃi],
- l'étirement et la protrusion pour le [s] de la syllabe [sy],
- et l'aperture et la protrusion pour le [ʃ] de [ʃi] et [ʃy].

Notons pour bien comprendre ces résultats que la protrusion est plus visible en vue de profile qu'en face tandis que c'est le contraire pour l'aperture et l'étirement. Ceci peut expliquer l'équivalence des performances obtenues dans cette expérience dans les deux vues de profil et de face.

Dans une autre expérience, Cathiard (1994) a étudié la perception du mouvement de l'arrondissement entre les voyelles [i] et [y]. L'expérience consistait à identifier ces deux voyelles dans les transitions [i] [i] et [i] [y] sans geste consonantique. Les stimuli étaient présentés en images sous les 4 conditions visuelles suivantes : présentation en image statique de face, en image statique de profil, en séquence d'images de face, ou en séquence d'images de profil. Notons que l'image statique correspondait à une image choisie à un instant donnée alors que la séquence d'images était formée par suite d'images prises à des instants successives. Dans le cas d'une vue de profil, les résultats montraient des scores d'identification identiques en dynamique et en statique. Par contre, en vue de face, les scores en dynamique sont supérieurs à ceux en statique. Ce qu'il faut retenir de ces études de Cathiard (1988, 1994) c'est que dans certains cas, la vue de profil (protrusion) semble porteuse d'information plus importante que la vue de face.

Adjoudani (1998) montre que les paramètres extraits d'une vue de face transmettent plus d'information que ceux de profil. Dans son étude portée sur un test de reconnaissance visuelle, les paramètres de la vue de face obtenaient un score de 83,2% d'identification correcte des stimuli comparés au 61% obtenu avec ceux de profil. L'auteur a aussi testé les deux vues combinées ensemble et obtenait un score de 86,6%.

Ce qu'il faut retenir est que la vue de face apporte plus d'information que la vue de profil. A l'exception bien sur de cas spécifiques qui concernent la classification des traits labiaux de protrusion et d'étirement (Cathiard, 1988, 1994) où la vue de profil peut être plus souhaitable que la vue de face. La vue de 3/4 a été globalement équivalente à la vue de face. Cependant, dans notre cas d'étude du code LPC où la main et les lèvres doivent être simultanément visibles, la vue de 3/4 poserait quelques problèmes de visibilité notamment pour la forme de la main.

Annexe E : liste complète des phrases du corpus

Nous présentons dans cette annexe la liste des phrases de notre corpus ainsi que des statistiques concernant le nombre d'occurrences des transitions d'une clé LPC (position ou configuration) vers une autre.

Liste des phrases

001 ma chemise est roussie .
002 voilà des bougies .
003 donne un petit coup .
004 tiens toi assis .
005 il a du goût .
006 elle m'étripa .
007 une réponse ambiguë .
008 louis pense a ça .
009 un four touffu .
010 un tour de magie .
011 voilà du filet cru .
012 la force du coup .
013 prête lui seize écus .
014 vous êtes exclue .
015 il fait des achats .
016 chevalier du gué .
017 le jeune hibou .
018 il fume son tabac .
019 un piège a poux .
020 l'examen du cas .
021 je suis a bout .
022 elle a chu .
023 je vais chez l'abbé .
024 deux jolis boubous .

025 une belle rascasse .
026 il part pour vichy .
027 faire la nouba .
028 c'est louis qui joue .
029 c'est ma tribu .
030 gilles m'attaqua .
031 pas plus de quatre rubis .
032 une rocaille moussue .
033 un pied fourchu .
034 c'est lui qui me poussa .
035 la chaise du bout .
036 trop d'abus .
037 j'en ai assez .
038 jean est fâché .
039 le pied du gars .
040 vous avez réussi .
041 ils n'ont pas pu .
042 le vent mugit .
043 une autre roupie .
044 deux beaux bijoux .
045 tu ris beaucoup .
046 il se garantira du froid avec ce bon capuchon .
047 dès que le tambour bat les gens accourent .
048 les deux camions se sont heurtés de face .
049 annie s'ennuie loin de mes parents .
050 la vaisselle propre est mise sur l'évier .
051 vous poussez des cris de colère .
052 mon père m'a donné autorisation .
053 un loup est jeté immédiatement sur la petite chèvre .
054 j'ai un scorpion sec dans mon talon aiguille .
055 nos dalmatiens campaient au camping à la montagne .
056 les gangs infligent des bings et des bangs périlleux sur une île.
057 vend-on un cake intact a hong-kong .
058 noam chomsky balaie encore le club ce soir .
059 l'avoué a besoin d'un joint sous huitaine .
060 la sueur suinte du thon huileux .
061 le beau ouistiti suit le riche huissier à waterloo .
062 tout winipeg attend wendy sur le parking ouest .
063 huit jésuites très huileux se font un brushing yougoslave .
064 bud et buck font un bon whist à maubeuge .
065 youri fouette l'ail ionique de kohoutek .
066 beung j'ai heurté le puits dans la lueur .

067 j'avais honte car la fille huait les who .
068 sur le sentier du ring camille eut peur pour son maquillage .
069 l'africa song s'emballe en juillet sur un walkman muet .
070 vuitton fait cuire dix wapitis goitreux .
071 david bowie s'est rué sur le quai où j'ai organisé ce must .
072 young fait un petit huit avec un joueur noueux .
073 jean nohain a chargé watson de louer le huitième buisson .
074 li-peng met du nuoc-mam dans son amuse-gueule .
075 j'ai eugène au téléphone qui cueille joliment du gui .
076 les keums du wharf rament évidemment dans le paysage .
077 ivanhoe a fait un bug au huitième essai .
078 tu huiles l'étui du buzzer de deux watts .
079 c'est hervé qui fuit dans un yacht en leasing .
080 j'ai étudié le parking huit a plancoet .
081 walid a hué les pink floyd a rouen .
082 eh oui les forums de l'accueil sont chouettes .
083 des ewoks habitent la maison en paille du centre spatial .
084 la famille ouistiti a éternué sous les dolmens .
085 nous jouons aux billes dans les ruines muettes .
086 j'ai identifié un mohican dans un western pyrénéen .
087 le balai a fait un looping sur la toundra .
088 ce tuyau a voyagé très haut chez les martiens .
089 j'ai huilé un rayon du train huit à l'équinoxe .
090 les caïds jouent au ping-pong avec l'équipe de bosnie .
091 j'ai eu les symptômes de la presbytie en huit jours .
092 je souhaite que sa peau usée ne reçoive jamais cette greffe ridicule .
093 ce fou ordinaire fiche le turban indien dans le bain optionnel .
094 une agraphe géante a pu heurter son beau hors-bord .
095 de mauvaises gens privent victor de sa coiffe bretonne .
096 la grive perchée sur l'if noir couve toujours ce canif chinois .
097 pose calmement ta dague pointue sur cette étoffe carrée .
098 le vase zen a perdu aussi un anneau en roche grise .
099 la houle lave les hublots d'une case déserte .
100 il abrase chaque jour un pneu ancien avec ses griffes pointues .
101 le photographe garantit un gag tordu au goût incertain .
102 le bateau heurta les housses du hublot un peu humides .
103 la feuille fut sertie avec une dent usée de la biche docile .
104 le géologue trouve finalement la houille en vrac dans le gave de pau .
105 va dans une cave quelconque et caches-y ce drapeau honteux .
106 le loup oublie son plan astucieux dans une poche chinoise .
107 le prof mielleux triche souvent à ce jeu idiot .
108 ce jazz rythmé est un cadeau inespéré .

109 le veau heureux attend eudes dans le hameau indien .
110 l'âne bègue voit que la vache de joseph se vexe .
111 tu houspilles ton amant onctueux qui louche réellement .
112 lagaffe fabrique une ruche carrée si tu y coopères .
113 cette phrase particulière étouffe toute une strophe vertueuse .
114 rêves-y car l'extase vient de cette bague gracieuse .
115 ce chant hideux rase son héros venu en hâte .
116 il élague curieusement la houpe qui est récalcitrante .
117 le camp hostile coordonne le putsch dans la cohue .
118 cette pêche fameuse a vu onduler l'endive blanche .
119 il se lève chaque jour et attend hercule qui oublie .
120 quand je soulève ma hache le banc ondule .
121 il n'arrive nullement qu'une vague surgisse du hors-d'oeuvre .
122 un zébu heureux ne touche jamais au houblon .
123 son gant entoure la valise trouvée sur la digue droite .
124 la horde de hors-la-loi alpague bientôt l'épave galloise .
125 jean heurta une cuve large pleine de gouache verte .
126 le vent établi sèche bien le houx où crèche mon hibou .
127 tchang ôte sa toge cintrée d'une main innocente .
. 128 un très bon vin en bouteille exige un planning idoine .
129 dom juan drague finalement une jeune fille mal faite .
130 a eux la soif zoologique du bourgeon ouvert .
131 au yen la tache pénible de ce prêt embarrassant .
132 en haut la guêpe pense aux fleurs .
133 objectez à neuilly contre le gaz nocif des hommes .
134 l'anglaise lui offre ce qu'elle a au doigt ou a l'oreille .
135 elle joue uniquement avec la neige chantante .
136 la pin-up feind de tomber chez toi mais ne blague jamais .
137 on tua onze ou douze torchons archaïques .
138 oudini ignore le train où doit se produire le spectacle .
139 il est parti illico en avion ou en gondole .
140 il gobe douze fèves et bêche tout mon jardin .
141 la caisse seule a enflé sur le ring en bois .
142 votre crêpe chaude vise bien le haut du feu .
143 tailles-en un bien haut et travaille chaque nom .
144 fernand oublie de moudre son café .
145 l'abeille n'engrange pas de miel sur un chemin .
146 cherche ou est le thon obtu que je trouve sot .
147 éole aide sa robe fendue à se soulever .
148 bashung oublie aussi qu'il lègue quelque chose .
149 je passe chercher ce que j'ai lu avec vous .
150 un zoom ferait ce que neuf demis pensent faire .

151 le fou immerge son aiguille et brode finement .
152 ce buveur balte augmente sa masse veineuse à heure régulière .
153 chaque bout du rail carré est une tige tenue .
154 un argument élogieux échappe bien au rosbif .
155 le malade guéri attrape mon solide microbe .
156 zola demande notamment du bon lait à un mage zurichois .
157 cette dame veut galber un tube vertical .
158 nous traquions bien euler pendant son footing urbain .
159 j'ai vu un holding important sur un terre-plein escarpé .
160 pain et pudding gallois aident le petit hussard oublieux .
161 une bouteille de riesling heurta le balcon humide .
162 ce jeu invite un type joueur et une dame riche .
163 miss zazie effectue un travelling heureux sur un machin imposant .
164 une vache normande dirige rarement un jumping zèlé .
165 le viking honteux a mal chuté sur cette petite nappe .
166 le moteur du boeing ronronne dans la brouette .
167 le pape vient en yamaha dans une bourgade curieuse .
168 le rotring exige une page carrée dans une feuille verte .
169 le lapin utilise son yoyo et a besoin d'aide .
170 le dumping l'incite à jeter les prunes tombées .
171 les yetis mal rasés ont la bouille pâteuse .
172 ils oublièrent chuck dans un tube carré .
173 léon range le parking vendéen où on aime zoner .
174 le king charmeur porte une chemise rouge foncée .
175 yasmine aime ton standing japonais .
176 gaspard blague mollement sur le leasing omniprésent .
177 nous draguions le torrent pour trouver des crabes noirs .
178 eux aussi aiment la tripe glorieuse un peu euphorique .
179 oeuvrez pour l'ove du globe bleu des yeux !
180 mes juges vont manger ce fichu yaourt à la truelle .
181 ce soldat un peu honteux fait un job glorieux .
182 cet oeil globuleux porte une lentille luisante .
183 la sage baleine zoophile n'a aucune patte valide .
184 un pale zébu agnostique mange normalement une solide pizza .
185 le prieur brade tout centime gagné .
186 la caille revient sans eux dans l'herbage gourmand .
187 une guenon heureuse a vu un balcon ombragé .
188 chaque garçon aime que le soleil brille .
189 il y a un truc qui ondule dans la cage murale .
190 tapes-en au noir sur une petite zone .
191 la fausse reine en tailleur agace guy .
192 nous tuons chaque chiot qui a été heureux .

193 flambes-y une crêpe bretonne de gamme moyenne .
194 chaque zéro est un looping tordu .
195 la meilleure omelette du larzac peut rivaliser avec le yachting normand .
196 un nain heurta une bogue charnue un onze janvier .
197 une tombe ming ne passe jamais pour un karting belge .
198 un homme jeune ne tombe pas pendant cette java .
199 des rides charmantes aèrent cette robe choisie dans les pages jaunes .
200 la foule a afflué quand mon neveu heurta le RER .
201 le thon heurta un bleuet .
202 il a été heurté par un pêcheur .
203 intonnes un u ou un euh à intervalles réguliers .
204 ceux des gueux bigleux veulent libérer bob taylor .
205 ou était oxymel .
206 le jeu ôtait illico au parfum oublié un fin bouquet d'embruns .
207 le cousin chinois du tribun évalue au juge autrement le tissu invendu .
208 le c.e. isole les engins communs aux deux charlots .
209 une québécoise pleurnicheuse brandit euclide lors des réunions .
210 moreau étale immanquablement un déficit commun à la queue de l'UE .
211 aladin élève chacun en symbiose avec le vieux ouzbek .
212 un coup heureux et impétueux modifie un vulgaire pain onctueux en gnome .
213 chacun ignore son CE un peu un moment .
214 avec un aplomb imparable nous avons chacun un CE énergique .
215 cette énergie insensée grève un quinzième de ugines .
216 sur le zing chacun interprète l'atlas humblement posé sur l'ancien jabot .
217 sa tape un peu impolie heurta bernache un peu trop violemment .
218 sylvain ne suit pas le parfum imprévu .
219 ce cabot ombrageux fête son accession au pouvoir .
220 un noir de jais évoque le front eurasien .
221 ce suspect heurta le bibelot ancien un peu lourdement .
222 le bedeau euphorique secoue l'anneau un jour par an .
223 aux lilas violets européens corot eugène préfère vingt-et-un oeillets .
224 jojo heurta le défunt et le tua .
225 à jeun antoine le heurte et cet accident le hantera .
226 le LPE insiste et les PME ont signé .
227 regardes, il zigzague un peu vite !
228 un huit dans l'eau a huilé l'un des tiroirs .
229 railles un bourrin oisif .
230 prends - le euclide !
231 tailles huit brins ouatés .
232 je m'huile le corps dans ce lieu iodé .
233 jourdain rajoute un pneu huileux .
234 il se ouate le tein rebelle .

235 antoine avait ouint son numéro huit .
236 j'ai reçu ton dessin hier .
237 quantum suédois ou rituel wolof .
238 la secoueuse fait des percings linguaux .
240 les gangs infligent des bings et des bangs sur une île périlleuse .
241 les gangs infligent des bings et des bangs périlleux sur une île .
242 j'ai étudié le plan du parking huit à plancoet .
243 j'ai joué au ping-pong avec l'équipe de bosnie .
244 la grive perchée sur cet if noir couve toujours ce canif chinois .
245 le géologue trouve finalement la houille en vrac dans la gave de pau .
246 ce chant hideux rase ce héros venu en hâte .
247 il arrive nullement qu'une vague surgisse du hors-d'oeuvre .
248 la horde de hors-la-loi alpague bientôt l'épave de galloise .
249 tchang ôte sa toge sacrée .
250 un zoom ferait ce que pensent neuf demis .
251 ce buveur malte augmente sa masse veineuse .
252 zola demande du bon lait à un mage zurichois .
253 ce soldat un peu honteux fait un job honteux .
254 un pale zébu agnostique mange normalement une pizza .
255 ana heurta une bogue charnue un onze janvier .
256 des hommes jeunes ne tombent pas pendant cette java .
257 ces rides charmantes aèrent .
258 le jeu ôtait illico .
259 ce des gueux bigleux veux libérer bob taylor .
260 le cousin du tribun évalue au juge le tissu invendu .
261 le CE isole les engins communs aux deux CE .
262 moreau étale immanquablement un déficit commun aux deux UE .
263 sur le zing chaque interprète .
264 sa tape impolie .
265 sa tape un peu impolie heurta bernache un peu violemment .
266 le bedeau un peu euphorique secoue l'anneau un jour par an .
267 aux lilas violet européens eugène corot préfère vingt-et-un oeillets .

Statistiques sur la représentation des transitions entre les clés (d'après (Gibert, 2006))

Transition d'une configuration LPC à une autre (table 8.4)

Transition d'une position LPC à une autre (table 8.5)

de / vers		Configuration N°								
		0	1	2	3	4	5	6	7	8
Configuration N°	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. 8.4 – Nombre de représentants lors des transitions de configuration à configuration. La configuration 0 correspond à la forme de la main en début et fin de phrase (configuration "repos").

de / vers		Position N°					
		0	1	2	3	4	5
Position N°	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. 8.5 – 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").

Annexe F : Concaténation de deux modèles HMM de deux syllabes

Rappelons qu'un modèle HMM est défini par trois paramètres : la matrice des probabilités de transition (Λ), la matrice des probabilités d'observation (Θ, dans notre cas, les observations sont modélisées par des mono-gaussiennes, donc par des matrices de moyennes (M) et de covariances (cov)), et le vecteur des probabilités initiales (π). Supposons que nous voulons concaténer deux modèles HMM de syllabes définis par ($\Lambda 1$, $M1$, $cov1$, $\pi 1$) et ($\Lambda 2$, $M2$, $cov2$, $\pi 2$), le nouveau modèle HMM de mot résultant (Λ, M, cov, π) est alors défini de la façon suivante :

- La matrice de transition Λ : Si

$$\Lambda 1 = \begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ & a_{22}^1 & a_{23}^1 \\ & & a_{33}^1 \end{pmatrix}$$

et

$$\Lambda 2 = \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ & a_{22}^2 & a_{23}^2 \\ & & a_{33}^2 \end{pmatrix}$$

alors :

$$\Lambda = \begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 & & & \\ & a_{22}^1 & a_{23}^1 & & & \\ & & a_{33}^1 & a_{31}^{12} & & \\ & & & a_{11}^2 & a_{12}^2 & a_{13}^2 \\ & & & & a_{22}^2 & a_{23}^2 \\ & & & & & a_{33}^2 \end{pmatrix}$$

Avec a_{31}^{12} est la probabilité de transition de l'état 3 du premier modèle HMM vers l'état 1 du modèle suivant. Etant donné que la somme des probabilités en lignes d'une matrice de transition est égale à 1, cette probabilité vaut donc : $a_{31}^{12} = 1 - a_{33}^1$ (théoriquement, a_{33}^1 est égale à 1, mais en pratique ce n'est pas le cas). Il est à noter que le nombre d'états du nouveau modèle crée devient égal à 6 puisque nous concaténons deux modèles à 3 états ; d'où la dimension $6 * 6$ de la matrice de transition.

- Le vecteur moyenne du nouveau modèle est simplement la juxtaposition des deux vecteurs moyennes des deux modèles de syllabes :

$$M = [M1 \quad M2]$$

- La matrice de covariance est aussi construite en juxtaposant les deux matrices de covariances des deux modèles de syllabes :

$$cov = [cov1 \quad cov2]$$

– Le vecteur des probabilités initiales est :

$$\pi = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Références bibliographiques

C. Abry et L.-J. Boë . "laws" for lips. *Speech Communication*, 5:97–104, 1986.

C. Abry et M. T. Lallouache . Audibility and stability of artiulatory movements. deciphering two experiments on anticipatory rounding in french. *In XIIth International Congress of Phonetic Sciences*, volume 1, pages 220–225, Aix-en Provence, France, August 1991.

A. Adjoudani . *reconnaissance automatique de la parole audiovisuelle*. Thèse de doctorat, Institut National Polytechnique de Grenoble, 1998.

A. Adjoudani et C. Benoît . *Speechreading by Humans and Machines*, chapitre On the integration of auditory and visual parameters in an HMM-based ASR, pages 461–471. Springer, Berlin, Germany, 1996.

A. Adjoudani, T. Guiard-Marigny, B. Le Goff et C. Benoît . Un modèle 3d de lèvres parlantes. *In Actes des XX° Journées d'Étude sur la Parole (JEP)*, pages 143–146, 1994.

D.W. Aha, K. Dennis et A. Marc . Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

J. Alegria, B. Charlier et S. Mattys . The role of lip-reading and cued.speech in the processing of phonological information in french-educated deaf children. *European Journal of cognitive psychology*, 11(4):451–472, 1999.

J. Alegria et J. Lechat . Phonological processing in deaf children : When lipreading and cues are incongruent. *Journal of Deaf Studies and Deaf Education*, 10(2):122–133, 2005.

J. Alegria et J. Leybaert . *L'acquisition du langage par l'enfant sourd : les signes, l'oral et l'écrit*, chapitre Le langage par les yeux chez l'enfant sourd : lecture, lecture labiale et langage parlé complété, pages 213–251. Marseille : Editions SOLAL, collection Troubles du Développement psychologique et les Apprentissages, 2005.

J. Alegria, J. Leybaert, B. Charlier et C. Hage . *Analytic approaches to human cognition*, chapitre On the origin of phonological representations in the deaf : hearing lips and hands, pages 107–132. Elsevier Science Publishers, 1992.

M. Alissali, P. Deleglise et A. Rogozan . Asynchronous integration of visual information in an automatic speech recognition system. *In International Conference on Spoken Langage Processing*, volume 1, pages 34–37, Philadelphia, PA, 1996.

R. André-Obrecht, B. Jacob et N. Parlangeau . Audio-visual speech recognition and segmental master-slave hmm. *In International Conference on Auditory-Visual Speech Processing*, Rhodes (Grèce), Septembre 1997.

O. Aran, T. Burger, A. Caplier et L. Akarun . Sequential belief-based fusion of manual and non-manual signs. *In 7th International Workshop on Gesture in Human-Computer Interaction and Simulation*, Lisbon, Portugal, May 2007.

V. Attina . *La Langue française Parlée Complétée : production et perception.* Thèse de doctorat en sciences cognitives, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.

V. Attina, D. Beautemps et M.-A. Cathiard . Coordination of hand and orofacial movements for cv sequences in french cued speech. *In International Conference on Spoken Langage Processing*, pages 1945–1948, Denver, Colorado, 2002.

V. Attina, D. Beautemps, M. A. Cathiard et M. Odisio . A pilot study of temporal organization in cued speech production of french syllables : rules for cued speech synthesizer. *Speech Communication*, 44:197–214, 2004.

E. T. Jr. Auer et L. E. Bernstein . Enhanced visual speech perception in individuals with early onset hearing impairment. *Journal of Speech, Hearing, and Language Research.*, A paraître.

D. Beautemps, M.-A. Cathiard et Y. Le Borgne . Benefit of audiovisual presentation in close shadowing task. *In 15th International Congress of Phonetic Sciences*, volume vol. 1, pages 841–844, Barcelone, 2003.

B. Ben Mosbah . *Utilisation de la mémoire de parole pour la reconnaissance : Application pour des personnes handicapées.* Thèse de doctorat, ENST Paris, 2005.

Andre-Pierre Benguerel et Margaret Kathleen Pichora-Fuller . Coarticulation effects in lipreading. *J Speech Hear Res*, 25(4):600–607, 1982.

C. Benoit, T. Lallouache, T. Mohamadi et C. Abry . *Talking Machines : Theories, Models and Designs*, chapitre A set of French visemes for visual French speech synthesis, pages 485–504. Elsevier SC. Publishers, Amesterdam, 1992.

C. Benoit, T. Mohamadi et S. Kandel . Effects of phonetic context on audio-visual intelligibility of french. *Journal of Speech and Hearing Research*, 37:1195–1203, 1994.

C. Benoît, T. Guiard-Marigny, B. Le Goff et A. Adjoudani . *Speechreading by Humans and Machines*, chapitre Which components of the face do humans and machines best speechread ?, pages 315–328. NATO-ASI Series 150 Springer, Berlin, 1996.

K. W. Berger, M. Garner et J. Sudman . The effect of degree of facial exposure and the vertical angle of vision on speechreading performance. *Teacher of the Deaf*, 69:322–326, 1971.

L. E. Bernstein, E. T. Jr. Auer et P. E. Tucker . Enhanced speechreading in deaf adults : Can short-term training/practice close the gap for hearing adults ? *Journal of Speech, Langage and Hearing Research*, 44(1):5–18, 2001.

L.E. Bernstein, M.E. Demorest et P.E. Tucker . Speech perception without hearing. *Perception and Psychophysics*, 62(2):233–252, 2000.

C. Binnie, P. Jackson et A. Montgomery . Visual intelligibility of consonants : A lipreading screening test with implications for aural rehabilitation. *J. Speech Hearing Disorders*, 41:530–539, 1976.

C. A. Binnie, A. Montgomery et P. Jackson . Auditory and visual contributions to the perception of consonants. *Journal Speech and Hearing Research*, 17:619–630, 1974.

I. Bloch . Information combination operators for data fusion : A comparative review with classification. *IEEE Transactions on Systems , Man and Cybernetics*, vol. 26:52–67, Janvier 1996.

L. D. Braida . Crossmodal integration in the identification of consonant segments. *Quaterly Journal of Experimental Psychology*, 43(A-3):647–677, 1991.

L. D. Braida, M. A. Picheny, J. R. Cohen, W. M. Rabinowitz et J. S. Perkell . Use of articulatory signals in automatic speech recognition. *Journal of Acoustical Society of America*, vol. 80:S18, 1986.

M. Brand et E. Irfan . Causal analysis for visual gesture understanding. Perceptual Computing Section Technical Report No. 327, MIT Media Laboratory, 1995.

M. S. Bratakos, P. Duchnowski et L. D. Braida . Toward the automatic generation of cued speech. *Cued Speech Journal*, 18:299–320, 1998.

C. Bregler, H. Hild, S. Manke et A. Waibel . Improving connected letter recognition by lipreading. *In International Joint Conference of Speech and Signal Processing*, pages 557–560, Minneapolis, MN, 1993.

C. Bregler et Y. Konig . "eigenlips" for robust speech recognition. *In ICASSP '94*, pages 669–672, Adelaide, Austrailia, 1994. URL `citeseer.ist.psu.edu/article/bregler94eigenlips.html`.

R. Brunelli et T. Poggio . Face recognition : Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993. ISSN 0162-8828.

T. Burger, O. Aran et A. Caplier . Modeling hesitation and conflict : a belief-based approach for multi-class problems. *In International Conference on Machine Learning and Applications (ICMLA 2006)*, Orlando, December 2006a.

T. Burger, A. Benoit et A. Caplier . Intercepting static hand gestures in dynamic context. *In International Conference on Image Processing (ICIP)*, Atlanta, USA, October 2006b.

R. Campell . Tracing lip movements : Making speech visible. *Visible Language*, vol. 22:33–57, 1988.

O. Cappé . h2m : A set of matlab functions for the em estimation of hidden markov models with gaussian stateconditional distributions. ENST/Paris http ://www.tsi.enst.fr/ cappe/h2m/, 2001.

F. Carton . *Introduction à la phonétique du Français.* Bordas, Paris/ Bruxessels/ Montreal, 1974.

M.-A. Cathiard . Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-rétraction des lèvres en français. Mémoire de maîtrise, Univérsité Grenoble II, 1988.

M.-A. Cathiard . *La perception visuelle de l'anticipation des gestes vocaliques : cohérence des événements audibles et visibles dans le flux de la parole.* Thèse de doctorat de psychologie cognitive, UFR SHS, Université Pierre Mendès France, 1994.

M.A. Cathiard, V. Attina, C. Abry et D. Beautemps . La langue française parlée complétée (lpc) : sa coproduction avec la parole et l'organisation temporelle de sa perception. *Revue Parole. n° spécial sur " Handicap langagier et recherches cognitives : apports mutuels"*, 31/32 (N° 29/30/31/32):255–280, 2004.

M. T. Chan . Hmm based audio-visual speech recognition integrating geometric and apparence based visual features. *In Workshop on multimedia signal processing*, pages 9–14, Canne, France, 2001.

D. Chandramohan et P.L. Silsbee . A multiple deformable template approach for visual speech recognition. *In Proceedings International Conference on Spoken Language Processing (ICSLP'96)*, 1996.

B.L. Charlier, C. Hage, J. Alegria et O. Périer . Evaluation d'une pratique prolongée du la pc sur la comprehension de la parole par l'enfant atteint de déficience auditive. *Glossa*, 22:28–39, 1990.

B.L. Charlier et J. Leybaert . The rhyming skills of deaf children educated with phonologically augmented speechreading. *Quarterly Journal of Experimental Psychology*, 53A:349–375, 2000.

T. Chen . Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18(1):9–21, 2001.

G.I. Chiou et J.-N. Hwang . Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8):1192 – 1195, 1997.

B. R. Clarke et D. Ling . The effects of using cued speech : A follow-up study. *The Volta Review*, 78:23–34, 1976.

T. F. Cootes, G. J. Edwards et C. J. Taylor . Active appearance models. *Lecture Notes in Computer Science*, 1407:484–498, 1998.

T. F. Cootes, C. J. Taylor, D. H. Cooper et J. Graham . Active shape models - their training and application. *Computer Vision and Image Understanding*, 61 (1):38–59, 1995.

R.O. Cornett . Cued speech. *American Annals of the Deaf*, 112:3–13, 1967.

R.O. Cornett . Annotated bibliography of research in cued speech. *Cued Speech Journal*, 4:1–23, 1990.

S. Cox, I. Matthews et J. Bangham . Combining noise compensation with visual information in speech recognition. *In* C. Benoît et R. Campbell, éditeurs . *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, pages 53–56, Rhodes, 1997. URL `citeseer.ist.psu.edu/cox97combining.html`.

B. Dalton, R. Kaucic et A. Blake . *Speechreading by Man and Machine : Models, Systems and Applications*, chapitre Automatic speechreading using dynamic contours, pages 373–382. Springer-Verlag, NATO ASI Series,, Berlin, Germany, 1996.

P. Daubias et P. Deleglise . Statistical lip-appearance models trained automatically using audio information. *Journal on Advances in Signal Processing*, 2002 (11)(11):1202–1212, November 2002.

J. Davis et M. Shah . Gesture recognition. Rapport technique CSTR-93-11, Department of Computer Science, University of Central Florida, 1993.

P. B. Denes . On the statistics of spoken english. *The Journal of Acoustical Society of America*, 35(6):892–904, 1963.

F. Destombes . *Aides manuelles à la lecture labiale et perspectives d'aides automatiques*, chapitre Le projet VIDVOX, pages 35–36. Centre scientifique IBM-France, 1982.

B. Dodd . The phonological systems of deaf children. *Journal of Speech and Hearing Disorders*, 41:185–198, 1976.

B. Dodd . The role of vision in the perception of speech. *Perception*, 6:31–40, 1977.

B. Dodd . *Hearing by Eye : The Psychology of Lipreading*, chapitre The acquisition of lip-reading skills by normally-hearing children. Lawrence Erlbaum Associates Ltd, Hillsdale, NJ, 1987.

E. R. Dougherty et C. R. Giardina . *Image processing - Continous to Discrete, Vol. 1 Geometric, Transform, and Statistical Methods*. Englewood Cliffs, NJ : Pentice Hall, 1987.

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

P. Duchnowski, U. Meuer et A. Waibel . See me, hear me : integrating automatic speech recognition and lip-reading. *In International Conference on Spoken Language Processing*, pages 547–550, Yokohama, Japan, 1994.

S. Dupont et J. Luettin . Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, 2000.

N. P. Erber . Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, 12:423–425, 1969.

N.P. Erber . Effect of angle, distance, and illumination on visual reception of speech by profoundly deaf children. *Journal of Speech and Hearing Research*, 17:99–112, 1974.

N. Eveno, A. Caplier et P.Y. Coulon . Jumping snakes and parametric model for lip segmentation. *In International Conference on Image Processing*, volume volume 3, pages 867–870, 2003.

H. W. Ewertsen et H. Birk Nielsen . A comparative analysis of the audiovisual, auditive, and visual perception of speech. *Acta oto- laryngologica*, 72:201–205., 1971.

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

J.-P. Gagné . Visual and audiovisual speech-perception training. *Journal of the Academy of Rehabilitative Audiology (Monograph Supplement)*, 27:133–159, 1994.

J.-P. Gagné, V. Materson, K.G. Munhall, N. Bilida et Querengesser . Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology*, 1995.

C. Garcia et M. Delakis . Convolutional face finder : A neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1408–1423, 2004.

M. Gentil . Etude de la perception de la parole : Lecture labiale et sosies labiaux. Rapport technique, IBM France, 1981.

G. Gibert . *Conception et évaluation d'un système de synthèse 3D de Langue française Parlèe Complétée (LPC) à partir du texte.* Thèse de doctorat, Institut National Polytechnique de Grenoble, April 2006.

G. Gibert, G. Bailly, D. Beautemps, Elisei F. et 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.

G. Gibert, G. Bailly et Elisei F . Evaluating a virtual speech cuer. *In International Conference on Spoken Langage Processing*, Pittsburgh, PA, 2006.

L. Golipour et D. O'Shaughnessy . A new approach for phoneme segmentation of speech signals. *In Interspeech'07*, Antwerp , Belgium, 2007.

M.S. Gray, J.R. Movellan et T.J. Sejnowski . *Advances in Neural Information Processing Systems 9*, chapitre Dynamic features for visual speech-reading : A systematic comparison, pages 751–757. MIT Press, Cambridge, MA, 1997.

K. Grobel et M. Assan . Isolated sign language recognition using hidden markov models. *In International conference of system , man and cybernetics*, pages 162–167, 1996.

C. Hage, J. Alegria et O. Périer . Cued speech and language acquisition : with specifics related to grammatical gender. *Cued Speech Journal*, 4:36–46, 1990.

C. Hage, J. Alegria et O. Périer . *Advances in cognition, education and deafness*, chapitre Cued speech and language acquisition : The case of grammatical gender morpho-phonology, pages 395–399. Washington, DC : Gallaudet Univ. Press, 1991.

A. J. Heap et F. Samaria . Real-time hand tracking and gesture recognition using smart snakes. *In Proceedings of interface to real and virtual worlds*, Montpellier, 1995.

M. Heckmann, F. Bertommier et K. Kroschel . A hybrid ann/hmm audio-visual speech recognition system. *In International Conference on Auditory-Visual Speech Processing*, pages 190–195, Aalborg, Denmark, 2001.

F. Heider et G. Heider . An experimental investigation of lip-reading. *Psychological Monographs*, 52:124–153, 1940.

M. Hennecke, K. Prasad et D. Stork . Using deformable templates to infer visual speech dynamics, 1994.

S. E. G. Öhman . Coarticulation in vcvs utterances : Spectrographic measurements. *The Journal of Acoustical society of America*, 39:151–168, 1966.

T. S. Huang, C. P. Hess, H. Pan et Z. Liang . A neuronet approach to information fusion. *In 1st IEEE Workshop on Multimedia Signal Processing*, pages 45–50, Princeton, NJ, 1997.

K. Imagawa, S. Lu et S. Igi . Color-based hands tracking system for sign language recognition. *In Proceedings of the 3rd. International Conference on Face and Gesture Recognition*, 1998.

P. L. Jackson, A. A. Montgomery et C. A. Binnie . Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing Research*, 19(4):796–812, 1976.

P.L. Jackson . The theoretical minimal unit for visual speech perception : Visemes and coarticulation. *The Volta Review*, 11-5:99–115, 1988.

B. Jacob et C. Sénac . Un modèle maître-esclave pour la fusion de données acoustiques et articulatoires en reconnaissance. *In Actes des Journées d'Étude sur la Parole (JEP)*, pages 363–366, Avignon, Juin 1996.

P. Jourlin . Handling disynchronization phenomena with hmm in connected speech. *In Proceedings of European Signal Processing Conference*, pages 133–136, Trieste, 1996.

P. Jourlin . Word dependent acoustic-labial weights in hmm-based speech recognition. *In Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, pages 69–72, Rhodes, Greece, 1997.

P. Jourlin . *Approche Bimodale du Traitement Automatique de la Parole : application à la Reconnaissance du Message et du Locuteur.* Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, 1998.

B. Jutras, J.-P. Gagné, M. Picard et J. Roy . Identification visuelle et catégorisation de consonnes en français québécois. *Revues d'orthophonie et d'audiologie*, 22(2):8–87, 1998.

W. Kadous . Machine recognition of auslan signs using powerglove : towards large-lexicon recognition of sign language. *In The workshop on the Integration of Gesture in Language and Speech*, pages 165–174, Wilmington, 1996.

M. Kass, A. Witkins et Terzopoulos . Snakes : active contour models. *Journal Computer Vision*, 1(4):321–331, January 1988.

E. Kipila . Analysis of an oral language sample from a prelingually child's cued speech : a case study. *Cued Speech annual*, 1:46–59, 1985.

D. H. Klatt . Speech perception : A model of acoustic-phonetic analysis and lexical access. *Journal Phonetique*, vol. 7:279–312, 1979.

J. Kramer et L. Liefer . The talking glove : An expressive and receptive "verbal" communication aid for the deaf, deaf-blind, and non-vocal. Rapport technique, Department of Electrical Engineering, Stanford University, 1989.

P. B. Kricos et S. A. Lesner . Differences in visual intelligibility across talkers. *Volta Review*, 84:219–225, 1982.

P. B. Kricos et S.A. Lesner . Effect of talker differences on the speechreading of hearing-impaired teenagers. *The Volta Review*, 87:5–16, 1985.

G. Krone, B. Talle, A. Wichert et G. Palm . Neural architecture for sensor fusion in speech recognition. *In ESCA/ESCOP Workshop Audio-Visual and Speech Processing*, pages 57–60, Rhodes, Greece, 1997.

M. T. Lallouache . Un poste visage-parole. acquisition et traitement des contours labiaux. *In Journées d'Etude sur la Parole (JEP)*, Montréa, 1990.

M.-T. Lallouache . *Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, 1991.

R. Lamy, D. Moraru, B. Bigi et L. Besacier . Premiers pas du clips sur les données d'évaluation ester. *In Proceedings Journées d'Etude sur la Parole*, Fès, Maroc, 2004.

A. L. Larr . Speechreading through closed-circuit television. *Volta Review*, 61:19–21, 1959.

J. LaViola . A survey of hand posture and gesture recognition techniques and technology. Rapport technique, Department of Computer Science Brown University, Providence, Rhode Island, 1999.

B. Le Goff, T. Guiard-Marigny et C. Benoît . Read my lips ... and my jaw ! how intelligible are the components of a speaker's face ? *In Eurospeech'95*, Madrid, Spain,, 1995.

B. Le Goff, T. Guiard-Marigny et C. Benoît . *Progress in Speech Synthesis*, chapitre Analysis-synthesis and intelligibility of a talking face, pages 235–246. Springer, New York, 1996.

S. A. Lesner . The talker. *Volta Review*, 90-5:89–98, 1988.

J. Leybaert . Phonological representations in deaf children : The importance of early linguistic experience. *Scandinavian Journal of Psychology*, 39:169–173, 1998.

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

J. Leybaert et J. Alegria . Spelling development in deaf and hearing children : Evidence for use of morphophonological regularities in french. *Reading and Writing*, 7:89–109, 1995.

J. Leybaert et B. L. Charlier . Visual speech in the head : the effect of cued speech on rhyming, remembering and spelling. *Journal of Deaf Studies and Deaf Education*, 1:234–248, 1996.

J. Leybaert et J. Lechat . Phonological similarity effects in memory for serial order of cued speech. *Journal of Speech, Language and Hearing Research*, 44:949–963, 2001.

A. M. Liberman et I. G. Mattingly . The motor theory of speech production revised. *Cognition*, 21:1–36, 1985.

D. Ling et B. R. Clarke . Cued speech : An evaluative study. *American Annals of the deaf*, 120:480–488, 1975.

N. Liu et B. C. Lovell . Hand gesture extraction by active shape models. *In Digital Image Computing : Techniques and Applications*, volume 1(1), pages 6–8, Cairns, December 2005.

J. Luettin et S. Dupont . Continuous audio-visual speech recognition. *In 5th European Conf. on Computer Vision*, Freiburg, Germany, 1998.

J. Luettin, N. A. Thacker et S.W. Beet . Visual speech recognition using active shape models and hidden markov models, 1996.

J. MacDonald et H. McGurk . Visual influences on speech perception processes. *Perception and Psychophysics*, 24:253–257, 1978.

A. MacLeod et Q. Summerfield . Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, 12:131–141, 1987.

A. MacLeod et Q. Summerfield . A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise : rationale, evaluation, and recommendations for use. *British Journal of Audiology*, 24:29–43, 1990.

D. J. Mapes et Moshell M. J. . A two-handed interface for object manipulationin virtual environments. *PRESENSE : Teleperators and Virtual Environments*, 4(4):403–416, 1995.

K. Mase et A.P. Pentland . Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76, 1991.

D. W. Massaro . *Categorical Perception : The Groundwork of Cognition*, chapitre Categorical partition : a fuzzy logical model of categorization behavior. Cambridge, MA : University Press, 1987.

D. W. Massaro . *Speechreading by Humans and Machines : Models, Systems, and Applications*, chapitre Bimodal speech perception : a progress report, pages 79–101. Springer, Germany, 1996.

D. W. Massaro . *Perceiving talking faces : From speech perception to a behavioral principle.* Cambridge, Massachusetts : MIT Press, 1998.

D.W. Massaro, M.M. Cohen et A.T. Gesi . Long-term training, transfer, and retention in learning to lipread. *Perception and Psychophysics*, 53(5):549–562, 1993.

I. Matthews, J. A. Bangham, R. Harvey et S. Cox . A comparison of active shape model and scale decomposition based features for visual speech recognition. *In Workshop on Audio Visual Speech Processing*, volume 1407 de *Lecture Notes in Computer Science*, pages 514+, 1998.

I. Matthews, J.A. Bangham et S. Cox . Audio-visual speech recognition using multiscale nonlinear image decomposition. *In International Conference on Spoken Language Processing (ICSLP)'96*, 1996.

I. Matthews, G. . Potamianos, C. Neti et J. Luettin . A comparaison of model and transform-based visual features for audio-visual lvcsr. *In International Conference on Multimedia and Expo*, Tokyo, Japan, 2001.

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

Sharad Mehrotra, Henry F. Korth et Abraham Silberschatz . Concurrency control in hierarchical multidatabase systems. *Journal of Very Large Data Bases (VLDB)*, 6(2):152–172, 1997.

M. Metzger . *First language acquisition in deaf children of hearing parents : Cued English input.* Thèse de doctorat, Georgetown University, 1994.

T. Mohammed, R. Campbell, M. MacSweeney, E. Milne, P. Hansen et M. Coleman . Speechreading skill and visual movement sensitivity are related in deaf speechreaders. *Perception*, 34(2):205–216, 2005.

A. A. Montgomery et P. L. Jackson . Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America*, 73(6):2134–2144, 1983.

Allen A. Montgomery, Brian E. Walden et Robert A. Prosek . Effects of consonantal context on vowel lipreading. *Journal of Speech and Hearing Research*, 30(1):50–59, 1987.

L. Mourand-Dornier . *Le rôle de la lecture labiale dans la reconnaissance de la parole.* Thèse de médecine, Univérsité de Franche-Comté, 1980.

J. R. Movellan et G. Chadderdon . *Speechreading by Man and Machine : Models, Systems and Applications*, chapitre Channel separability in the audiovisual integration of speech : A Bayesian approach, pages 473–488. Springer-Verlag, NATO ASI Series, Berlin, Germany, 1996.

K. Murakami et H. Taguchi . Gesture recognition using recurrent neural networks. *In Proceedings of CHI'91 human factors in computing systems*, pages 237–242, 1991.

S. Nakamura, R. Nagai et K. Shikano . Adaptative determination of audio and visual weights for automatic speech recognition. *In Eurospeech*, pages 1623–1626, Rhodes, Greece, 1997.

Y. Nakano . A study on the factors which influence lipreading of deaf children. *Language research in countries other than the United States, Volta Review*, 68:68–83, 1961. Cited by Quigley (1966).

K. K. Neely . Effect of visual factors on the intelligibility of speech. *Journal of Acoustic Society of America*, 28:1275–1277, 1956.

C. C. Neocleous et C. N. Schizas . Neural network, review and critic : Methods and applications of artificial intelligence. *In 2nd Pan-Hellenic Conference on Artificial Intelligence*, pages 300–313, Thessaloniki (Aristotle University), 2002. Berlin, Springer.

C. Neti, G. Iyengar, G. Potamianos, A. Senior et B. Maison . Perceptual interfaces for information interaction : joint processing of audio and visual information for human-computer interaction. *In International Conference on Spoken Langage Processing*, volume vol.3, pages 11–14, 2000.

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

J. J. O'Neill . Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech Hearing Disorders*, 19:429–439, 1954.

S. Ong et S. Ranganath . Automatic sign language analysis : a survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6), 2005.

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

E.D. Petajan . Automatic lipreading to enhance speech recognition. *In Proceedings Global Telecommunications Conference (GLOBCOM)'84*, pages 265–272, Atlanta, 1984.

E.D. Petajan . Automatic lipreading to enhance speech recognition. *IEEE Computer society conference on computer vision and pattern recognition*, pages 19–23, 40–47, 1985.

E.D. Petajan et H.P. Graf . Robust face feature analysis for automatic speechreading and character animation. *In AFGR96*, pages 357–362, 1996.

G. Potamianos, H.P. Graf et E. Cosatto . An image transform approach for hmm based automatic lipreading. *In Proceedings International Conference on Image Processing (ICIP)'98*, volume III, pages 173–177, Chicago, 1998.

G. Potamianos, C. Neti, G. Gravier, A. Garg et A.W. Senior . Recent advances in the automatic recognition of audio-visual speech. *In Proceedings of the IEEE*, volume vol. 91, pages 1306–1326, 2003.

G. Potamianos, C. Neti, G. Iyengar, A.W. Senior et A. Verma . A cascade visual front end for speaker independent automatic speechreading. *Speech Technology*, 4:193–208, 2001.

G. Potamianos, C. Neti, J. Luettin et I. Matthews . *Audio-Visual Speech Processing*, chapitre Audio-Visual Automatic Speech Recognition : An Overview. MIT Press, ISBN : 0-26-222078-4, 2006.

K. Prasad, D. Stork et G. Wolff . Preprocessing video images for neural learning of lipreading, 1997.

O. Périer . L'enfant à audition déficiente. *Acta oto-rhino-laryng*, 41:125–420, 1987.

M. Qutaishat, H. Moussa et A. A.-M. Expert Syst. Appl. 32(1) : 24-37 (2007) Bayan, T.and Hiba . American sign language (asl) recognition based on hough transform and neural networks. *Expert systems with Applications*, 32(1):24–37, 2007.

B.H. Rabiner, R. et Juang . An introduction to hidden markov models. *IEEE Acoustics, Speech and Signal Processing Magazine*, 3 (1):4–16, January 1986.

D. Reisberg, J. Mclean et A. Goldfield . *Hearing by Eye : The Psychology of Lipreading*, chapitre Easy to hear but hard to understand : a lipreading advantage with intact auditory stimuli, pages 97–113. Lawrence Erlbaum Associates Ltd, Hillsdale, NJ, 1987.

G. Rizzolatti, G Luppino et M. Matelli . *Supplementary sensorimotor area*, chapitre The classic supplementary motor area is formed by two independent areas, pages 45–56. Philadelphia : Lippincott–Raven, 1996.

J. Robert-Ribes, M. Piquemal, J. L. Schwartz et P. Escudier . *Speechreading by Man and Machine : Models, Systems and Applications*, chapitre Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition, pages 193–210. Springer-Verlag, NATO ASI Series, Berlin, Germany, 1996.

J. Robert-Ribès . *Modèles d'intégration audiovisuelle de signaux linguistigues : de la perception humaine à la reconnaissance automatique des voyelles.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, 1995.

J. Robert-Ribès, J.-L. Schwartz, M.T. Lallouache et P. Escudier . Complementarity and synergy in bimodal speech : Auditory, visual and audio-visual identification oral vowels in noise. *The Journal of Acoustical Society of America*, 103(6):3677–3689, 1998.

A. Rogozan et P. Deléglise . Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication*, 26(1-2):149–161, 1998. ISSN 0167-6393.

A. Rogozan, P. Deléglise et M. Alissali . Adaptive determination of audio and visual weights for automatic speech recognition. *In Proceedings European Tutorial Research Workshop on Audio-Visual Speech Processing (AVSP)'97*, pages 61–64, 1997.

D. Rubine . Specifying gestures by example. *In Proceedings of SIGGRAPH'91*, pages 329–337. ACM Press, 1991.

J.-L. Schwartz . *Traitement automatique du langage parlé 2 : reconnaissance de la parole*, chapitre La parole multimodale :deux ou trois sens valent mieux qu'un, pages 141–178. Hermes,, Paris, 2002.

J.-L. Schwartz . La parole multisensorielle : Plaidoyer, problèmes et perscpectives. *In Actes des XXVèmes Journées d'Etude sur la Parole (JEP)*, pages 11–17, Fès, Maroc, 2004.

J.-L. Schwartz, J. Robert-Ribès et P. Escudier . *Hearing by Eye II : Advances in the Psychology of Speechreading and Auditory-Visual Speech*, chapitre Ten years after Summerfield : A taxonomy of models for audio-visual fusion in speech perception, pages 85–108. Psychology Press, Hove, UK, 1998.

P. L. Silsbee . Motion in deformable templates. *In In First IEEE Internationale Conference on Image Processing*, volume volume 1, pages 323–327, November 1994.

P. L. Silsbee et Q. Su . *NATO ASI : Speechreading by Humans and Machines*, chapitre Audiovisual sensory integration using hidden Markovmodels, pages 489–495. Springer-Verlag, 1996.

T. Starner . Visual recognition of american sign language using hidden markov models. Mémoire de master, Massachusetts Institue of Technology, 1995.

T. Starner et A. Pentland . Real- time american sign language recgnition from video using hidden markov models. Perceptual Computing Section N° 375, MIT Media Laboratory, 1996. Technical Report.

D. G. Stork, G. Wolff et E. Levine . Neural network lipreading system for improved speech recognition. *In IJCNN'92*, volume 2, pages 285–295, Baltimore, MD, 1992.

D. Sturman . *Whole-hand Input.* Thèse de doctorat, Massachusetts Institute of Technology, 1992.

D. Sturman et D. Zeltzer . A surveyof glove-based input. *IEEE Computer Graphics and Applications*, 14(1):30–39, 1994.

Q. Su et P. L. Silsbee . Robust audiovisual integration using semicontinuous hidden markov models. *In International Conference on Spoken Language Pocessing*, volume 1, pages 42–45, Philadelphia, PA, 1996.

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

Q. Summerfield . Use of visual information of phonetic perception. *Phonetica*, 36:314–331, 1979.

Q. Summerfield . Audio-visual speech perception, lipreading and artificial stimulation. *Hearing Science and Hearing Disorders*, pages 131–182, 1983.

Q. Summerfield . *Hearing by Eye : The Psychology of Lipreading*, chapitre Some preliminaries to a comprehensive account of audio-visual speech perception, pages 3–51. Lawrence Erlbaum Associates Ltd., Hove, UK, 1987.

Q. Summerfield, A. MacLeod, M. McGrath et M. Brooke . *Handbook of Research on Face Processing*, chapitre Lips, teeth, and the benefits of lipreading, pages 223–233. Elsevier Science Publishers, Amesterdam, The Netherlands, 1989.

P. Teissier, J. Robert-Ribès et J.-L. Schwartz . Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6):629–642, 1999.

A. Tseva . L'arrondissement dans l'identification visuelle des voyelles du français. premiers acquis. *Bulletin du Laboratoire de la Communication Parlée*, 3:149–186, 1989.

A. Tseva et M.-A. Cathiard . Paroles vues : la dimension d'arrondissement dans l'identification visuelle des voyelles du français. *In Actes du 1er Congré Français d'Acoustique, Colloque de physique, Colloque C2*, volume 51, pages 507–510, 1990.

R. M. Uchanski, L. A. Delhorme, A. K. Dix, L. D. Braida, C. M. Reed et N. I. Durlach . Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of cued speech. *Journal of Rehabilitation Research and Development*, 31(1):20–41, 1994.

D. Vaufreydaz, J. Bergamini, J. F. Serignat, L. Besacier et M. Akbar . A new methodology for speech corpora definition from internet documents. *In Proceedings 2nd International Conference on Language Resources and Evaluation (LREC2000)*, pages 423–426, Athens, Greece, 2000.

B. Walden., R. Prosek, A. Mongomery, C. Scherr et C. Jones . Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, pages 135–145, 1977.

B. E. Walden, S.A. Erdman, A. A. Montgomery, D. M. Schwartz et R. A. Prosek . Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech and Hearing Research*, 24(2):207–216, 1981.

T. Watanabe et M. Kohda . Lip-reading of japanese vowels using neural networks. *In International Conference Spoken Language Procesing*, pages 1373–1376, Kobe, Japan, 1990.

R. Watson . A survey of gesture recognition techniques. Rapport technique TCD-CS-93-11, Department of Computer Science, Trinity College Dublin, 1993.

A. Wexelblat . An approach to natural gesture in virtual environments. *ACM Transactions on Computer Human Interaction 2*, 3:179–200, 1995.

M. F. Woodward et C. G. Barber . Phoneme perception in lipreading. *Journal of Speech and Hearing Research*, 3:212–222, 1960.

W. Wouts . *Aides manuelles à la lecture labiale et perspectives d'aides automatiques*, chapitre L'AKA, pages 16–29. Centre scientifique IBM-France, 1982.

M. H. Yang, N. Ahuja et M. Tabb . Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1061–1074, 2002. ISSN 0162-8828.

H. Yehia, P. Rubin et E. Vatikiotis-Bateson . Quantitative association of orofacial and vocal-tract shapes. *In International Conference on Audio-Visual Speech Processing*, pages 41–44, 1997.

B. P. Yuhas, M. H. Goldstein et T. J Sejnowski . Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, pages 65–71, Novembre 1989.

B. P. Yuhas, M. H. Goldstein Jr., T. J. Sejnowski et R. E. Jenkins . Neural network models of sensory integration for improved vowel recognition. *Proc. IEEE*, 78 (10):1658–1667, 1990.

A. L. Yuille, P. W. Hallinan et D. S. Cohen . Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8 (2):99–111, 1992.

Liste des publications

- **Noureddine Aboutabit**, Denis Beautemps & Laurent Besacier. Hand and Lips desynchronization analysis in French Cued Speech : Automatic segmentation of Hand flow. *In Proceedings of ICASSP'06*, 2006.
- **Noureddine Aboutabit**, Denis Beautemps & Laurent Besacier. Characterization of Cued Speech vowels from the inner lip contour. *In Proceedings of ICSLP'06*, 2006.
- **Noureddine Aboutabit**, Denis Beautemps & Laurent Besacier. Vowel classification from lips : the Cued Speech production case. *In proceeding of International Seminar on Speech Production (ISSP)*, pp 127-134, 2006.
- **Noureddine Aboutabit**, Denis Beautemps & Laurent Besacier. Automatic identification of vowels in the Cued Speech context. *In International Conference on Auditory-Visual Speech Processing (AVSP)*, 2007.
- Beautemps, D., Girin, L., **Aboutabit, N.**, Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Le, V.B., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.F., Tribout, M., Vidal, S., 2007. TELMA : Téléphonie à l'Usage des Malentendants. Des modèles aux tests d'usage. *In : proc. ASSISTH'2007*, Toulouse, France.
- **Noureddine Aboutabit**, Denis Beautemps, Jeanne Clarke & Laurent Besacier. A HMM recognition of consonant-vowel syllabes from lip contours : the Cued Speech case. *In Interspeech 2007*. Antwerp, Belgium.
- **Noureddine Aboutabit**, Denis Beautemps & Laurent Besacier. Lips and Hand Modeling for Recognition of the Cued Speech Gestures : The French Vowel Case. *Speech Communication*, soumis.

Table des figures

Liste des tables

Table des matières

Résumé

La Langue Française Parlée Complétée (LPC) héritée du Cued Speech (CS) a été conçue pour compléter la lecture labiale par nature ambiguë et ainsi améliorer la perception de la parole par les sourds profonds. Dans ce système, le locuteur pointe des positions précises sur le côté de son visage ou à la base du cou en présentant de dos des formes de main bien définies. La main et les lèvres portent chacune une partie complémentaire de l'information phonétique. Cette thèse présente tout d'abord une modélisation du flux manuel pour le codage automatique des positions de la main et de la configuration. Puis les travaux sont centrés sur le flux labial en discutant la classification des voyelles et des consonnes du Français. Le flux labial est composé des variations temporelles de paramètres caractéristiques issus du contour interne et externe des lèvres. Dans le cas des voyelles la méthode de classification utilise la modélisation gaussienne et les résultats montrent une performance moyenne de 89 % en fonction de la position de la main LPC. Le contexte vocalique est pris en compte dans le cas des consonnes par une modélisation HMM de la transition labiale de la consonne vers la voyelle avec un taux d'identification de 80 % en termes de visèmes CV. Un modèle de fusion « Maître-Esclave » piloté par le flux manuel est présenté et discuté dans le cadre de la reconnaissance des voyelles et des consonnes produites en contexte LPC. Le modèle de fusion prend en compte les contraintes temporelles de la production et la perception du LPC, ce qui constitue aussi une première contribution à la modélisation du système LPC du point de vue perceptif.

Mots-Clés : Lecture labiale ; Modélisation des lèvres ; classification des voyelles et des consonnes ; visèmes, Langue Française Parlée Complétée ; Modèle de fusion ; Reconnaissance de gestes.

Abstract

Cued Speech (CS) is a visual communication system that uses handshapes placed in different positions near the face, in combination with the natural speech lip-reading, to enhance speech perception from visual input for deaf people. In this system, the speaker moves his hand in close relation with speech. Handshapes are designed to distinguish among consonants whereas hand positions are used to distinguish among vowels. Due to the CS system, both manual and lip flows produced by the CS speaker carry a part of the phonetic information. This work presents at first a method for the automatic coding of the manual flow in term of CS hand positions and CS handshapes. Then the lip-shape classification of the vowels and the consonants is discussed. The labial flow is composed of the temporal variations of lip parameters extracted from the inner and the outer contours of the lips. This work will show how the distribution of lip parameters inside each group of CS hand positions allows vowel discrimination. A classification method based on Gaussian modeling is presented and results demonstrate a good performance of this classification (89% as test score). The vocalic context is taken into account in the case of the consonants, with the use of HMM for the modeling of the lip transition from the consonant towards the vowel (80 % as test scores in term of CV visemes). Finally, the modeling of the lip information and the coding of the manual flow are included in a "Master-Slave" fusion model for recognition of the vowels and the consonants in the CS context. The fusion model integrates the temporal constraints of the CS production and perception. This work is thus also a first contribution to the modeling of the CS system from the perceptive point of view.

Keywords : Lipreading ; Lip Modeling ; Vowel and Consonant Classification ; Visemes ; Cued Speech ; fusion process ; gesture recognition.