Contribution à la parole augmentée : production et surdité

Denis Beautemps

UNIVERSITE DE GRENOBLE

THESE

Pour obtenir le grade de

HABILITATION A DIRIGER DES RECHERCHES DE L'UNIVERSITE DE GRENOBLE

Spécialité : Signal, Image, Parole, Télécom

Arrêté ministériel : 7 août 2006

Soutenue le 1er Avril 2015 par

Denis Beautemps

Contribution à la parole augmentée : production et surdité

Jury :

Madame Jacqueline Leybaert, Président du Jury, Professeur, Université Libre de Bruxelles
Madame Martine Adda-Decker, Rapporteur, Directeur de Recherche au CNRS
Monsieur Christophe d'Alessandro, Rapporteur, Directeur de Recherche au CNRS
Monsieur Gang Feng, Rapporteur, Professeur des Universités, Grenoble-INP
Monsieur Sébastien Schmerber, Examinateur, Professeur des Universités-Praticien Hospitalier, CHU de Grenoble
Madame Régine André-Obrecht, Examinateur, Professeur des Universités, Université de Toulouse III

Table des matières

Liste des abréviations……..5
INTRODUCTION 7
Chapitre I - Etude de la production du Cued Speech 11
Introduction 11
I.1 Résultats sur la coordination oro-bracho-faciale 13
I.2 Perception de l'anticipation de la main 18
I.3 Bilan 19
Chapitre II - Transcodage entre espaces visuel et audio pour le Cued Speech 21
II.1 Contexte du projet TELMA de Téléphonie à l'usage des malentendants 22
II. 2 Reconnaissance du Cued Speech 24
II.2.1 Méthodes de fusion 24
II.2.2. Résultats par application de la méthode de fusion ID 25
II.2.3. Résultats par application de la méthode de fusion IS 26
II.3. Conversion audio vers le Cued Speech : mapping des espaces 27
Chapitre III – Parole, surdité et audition réhabilitée 37
Introduction 37
III.1 Contexte du projet PLASMODY 37
III.2 Etude des aspects physique en production de parole 38
III.2.1- Etude du comportement des cordes vocales 39
III.2.2- Etude de l'effet du mouvement des parois du conduit vocal 42
III.3 Close shadowing et collaboration audiovisuelle 44
III.4 Relations perceptuo-motrices et implants cochléaires 45
Chapitre IV : Projet de recherche - Adaptabilité en parole, variabilité et plasticité 47
A. Articulation labiale complétée : du mapping à la reconnaissance 49
B. Implant cochléaire et fonctionnalités audio-visuelles 52
C. Rhébotique : de la syllabe au discours 53
C.1. Méthodologie 55
C.2. Axe 1 - Empathie, adhésion, interaction en discours 55
C.3. Axe 2- Intelligibilité, multimodalité et complémentarité en parole 56
CONCLUSION 59
Références bibliographiques 63
Annexe : 7 publications significatives 67

Liste des abréviations

ACP	Analyse en Composantes Principales
ALLSHCS	Pôle de recherche « Art, Lettres, Langues, Sciences Humaines Cognitives et Sociales »
ATER	Attaché Temporaire d'Enseignement et de Recherche
ATR	Advanced Telecommunications Research Institute International
ATT	Axe transversal transformant
BQR	Bonus Qualité Recherche de l'INPG
BEDEI	Banc Expérimental Dédié à l'Enregistrement In vivo
CAC	Conseil Académique
CerCo	Centre de Recherche Cerveau & Cognition
CHU	Centre Hospitalier Universitaire
CODIR	Conseil des Directeurs d'unité
COMUE	Communauté d'Universités et Etablissements, site Grenoblois
CR	Chargé de Recherche
CRISSP	Cognitive Robotics, Interaction & Speech Processing
CS	Cued Speech
DCT	Discrete Cosine Transform
DEA	Diplôme d'Etude Approfondie
DPC	Département Parole-Cognition
DR	Directeur de Recherche
EDISCE	École doctorale Ingénierie pour la santé la Cognition et l'Environnement
EGG	Electo GlottoGraphe
EEATS	Ecole doctorale « Electronique, Electrotechnique, Automatique, Traitement du Signal »
EM	Algorithme Expectation-Maximization
ENSE^3	Ecole d'ingénieurs en énergie eau et environnement
ENSERG	Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble
EVA	Evaluation Vocale Assistée
GIPSA-lab	Laboratoire Grenoble, Image, Parole, Signal, Automatique
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICP	Institut de la Communication Parlée
ID	Modèle de fusion à Identification Directe
IE	Ingénieur d'Etude
IR	Ingénieur de Recherche
IS	Modèle de fusion à Identification Séparée
INA	Institut National de l'Audiovisuel
INS2I	CNRS - Institut des Sciences de l'Information et de leurs interactions
INSHS	CNRS - Institut des Sciences Humaines et Sociales
INSIS	CNRS - Institut des Sciences de l'Ingénierie et des Systèmes
IUT	Institut Universitaire de Technologie
K-MEANS	Algorithme k-means ou k-moyennes
LaBRI	Laboratoire Bordelais de Recherche en Informatique
LIG	Laboratoire d'Informatique de Grenoble
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

LLSH	Ecole Doctorale « Langues Littératures et Sciences Humaines »
LPC	Langue Française Parlée Complétée
LPNC	Laboratoire de Psychologie et NeuroCognition
LSP	Line Spectral Pairs
LTCI	Laboratoire Traitement et Communication de l'Information
MAGIC	Machines parlantes, Gestes oro-faciaux, Interaction Face-à-face, Communication augmentée
MCF	Maître de Conférences
MFCC	Mel frequency Cepstral Coefficients
MPACIF	Equipe Machines Parlantes, Communication, Interaction Face-à-Face
MSTIC	Pôle de recherche « Mathématiques, Sciences et technologies de l'information et de la communication »
ORL	Oto-Rhino-Laryngologie
PFE	Projet de Fin d'Etude
PHELMA	Ecole nationale supérieure de PHysique, ELectronique, Matériaux
PUPH	Professeur des Universités Praticien Hospitalier
ROI	Region of Interest
RARE	Rhétorique de l'Antiquité à la Révolution
MD	Modèle de fusion par recodage dans la Modalité Dominante
RM	Modèle de fusion par Recodage Moteur
RD	Recodage dans la Modalité Dominante
SDV	Sciences de la Vie
SPI	Sciences pour l'Ingénieur
STIC	Les Sciences et Technologies de l'Information et de la Communication
UJF	Université Joseph Fourier
UGA	Université Grenoble-Alpes

INTRODUCTION

Mon cadre de recherche est le domaine de la communication parlée augmentée. La communication augmentée consiste à enrichir les signaux de la communication afin d'améliorer leur robustesse en condition de communication adverse, de s'adapter aux capacités de communication, ou au style de communication des interlocuteurs par exemple (confidentialité, situation face à face,…). Ce programme s'est appuyé sur des travaux théoriques et applicatifs innovants dans le domaine de l'analyse/modélisation des productions multimodales de l'activité langagière et de leur perception à partir d'enregistrement de corpus articulatoires, audio et visuels.

Citons tout d'abord l'analyse/modélisation statistique qui appliquée à la géométrie 2D du conduit vocal permet de capturer les traces non visibles ou partiellement visibles de l'activité motrice portée par les articulateurs de la parole (lèvres, langue, vélum par exemple) et réalise le lien (l'interface) avec l'activité labiale et les caractéristiques acoustiques associées. Les résultats ont fait l'objet d'une publication à la revue internationale avec comité de lecture *Journal of The Acoustical Society of America* en 2001 (Beautemps et al., 2001, voir aussi Figure-1).

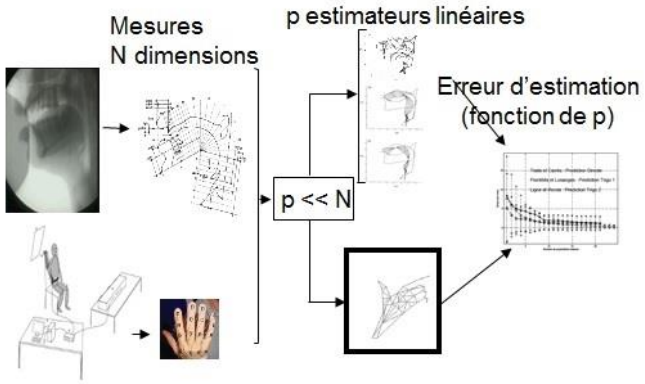


Figure-1 : Système d'acquisition de données articulatoires à gauche et analyse de l'explication de la variance à droite. Illustrations issues en partie de Beautemps et al., 2001 et de Sacher et al., 2008.

Cette approche a permis d'alimenter un volet de travaux en modélisation oro-faciale 3D dans le cadre de l'activité sur les « têtes parlantes virtuelles et la synthèse audiovisuelle de la parole » et à la modélisation de la collaboration bracho – faciale (Figure-1) telle qu'elle apparaît naturellement dans la « Langue Française Parlée Complétée » (version en langue Française du Cued Speech, système de mouvements codés de la main en appoint du mouvement des lèvres) dans la communication avec les personnes sourdes (projet RNRT ARTUS, cadre télévisuel -

chaîne ARTE-, projet TELMA de téléphonie à l'usage des malentendants - ANR, dont France Telecom division R&D a été partenaire ainsi que le CHU de Grenoble pour le lien avec les personnes sourdes).

Les résultats en modélisation ont permis des avancées novatrices dans le domaine de la synthèse audio-visuelle de la parole avec l'intégration d'une nouvelle modalité, le geste codeur, en reconnaissance labiale ainsi qu'en mapping entre espaces audio et visuels. Ils ont fait l'objet de six contributions dans les revues prestigieuses du domaine (*Journal of the Acoustical Society of America, Speech Communication, IEEE Signal Processing Letters*) et d'une contribution qui a été soumise récemment auprès de la revue internationale *Computer Speech and Language*.

Les travaux en multimodalité ont permis l'ouverture d'un nouveau volet de recherche sur la parole en situation de surdité profonde remédiée par l'implant cochléaire (projet ANR / PLASMODY). Les premiers résultats en perception viennent de faire l'objet d'une publication à la revue en ligne internationale avec comité de lecture *Frontiers in Psychology*. Les travaux en production font quant à eux l'objet d'une autre publication soumise à la revue internationale avec comité de lecture *Journal of the Acoustical Society of America*.

Les travaux dans ces différents volets ont pu être fructueux grâce au concours de trois thèses soutenues que j'ai encadrées principalement et de deux thèses actuellement en cours, de plusieurs projets contractualisés localement (BQR), auprès de la région Rhône-Alpes, au CNRS ou à l'ANR et d'une alliance pluri-disciplinaire alliant domaines d'expertise en traitement du Signal, Sciences Cognitives, Sciences du Langage et en Physique pour les Sciences de l'Ingénieur. Enfin, ces travaux n'auraient pu être menés sans le recours aux plateformes expérimentales de l'Institut de la Communication Parlée puis celles de GIPSA-lab pour l'enregistrement des corpus et les diverses expérimentations.

La suite du document est divisé en quatre chapitres présentant chacun une synthèse de mes travaux de recherche avec renvoi à des publications dans des revues ou conférences internationales avec comité de lecture du domaine (en annexe). Le premier chapitre est dédié à l'étude de la production du Cued Speech dans sa version en langue Française (code LPC), au centre de la thèse de Virginie Attina. Le second chapitre a trait au transcodage audio-visuel pour le Cued Speech qui regroupe les travaux en reconnaissance visuelle au cœur de la thèse de Noureddine Aboutabit et du projet ANR TELMA ainsi que les travaux sur le mapping audio-visuel au cœur de la thèse de Zuheng Ming. Enfin un troisième chapitre est dédié à la production et perception de la parole dans le cadre de la surdité et de l'audition réhabilitée qui sont l'objet de deux thèses en cours de Lucie Scarbel et de Louis Delebecque dans le contexte du projet

ANR PLASMODY et du projet Région Rhône-Alpes Cibles 2011. Enfin le dernier chapitre présente des enjeux de prospective dans le cadre de mon projet de recherche.

Chapitre I - Etude de la production du Cued Speech

Introduction

Les bénéfices de l'information visuelle pour la perception de la parole (lecture labiale) sont bien connus. Depuis les travaux de Sumby et Pollack (1954), à ceux de Benoit et collègues (1992) en passant par Summerfield et collègues (Summerfield, 1979 ; Summerfield et al., 1989), il est bien établi que l'information fournie par le mouvement du visage (principalement celui des lèvres), est utilisée pour améliorer la perception de la parole dans des situations de bruit ambiant. Les expériences en shadowing (répétition de la parole de l'autre) ont montré le bénéfice de l'apport de la collaboration audiovisuelle en situation de parole « audio claire » (Reisberg et al., 1987). L'effet McGurk manifeste dans le cas où les deux modalités fournissent des informations incohérentes la capacité d'intégrer les informations issues des deux modalités par l'identification d'un percept différent de celui porté par chacune des deux modalités (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978). Ces résultats montrent que les personnes entendantes ont des compétences en lecture labiale sans entrainement spécifique. Cependant, les performances varient grandement d'une personne à l'autre, les meilleurs lecteurs labiaux se trouvant dans la population des personnes sourdes profondes et oralistes (Bernstein et al., 2000). Les scores en identification de mots isolés atteignent 43,6 % (Auer and Bernstein, 2007; Bernstein et al., 2010), les femmes étant plus performantes selon une étude de 2009 de Strelnikov et collègues (2009).

La lecture labiale seule est donc loin d'atteindre la perfection. La raison principale étant liée à l'ambiguïté du pattern labial, des sons de parole différents pouvant avoir des formes aux lèvres similaires (voir Figure I-1 pour une illustration à partir de données de production).

Cependant, pour la population des personnes sourdes oralistes (5 à 6 millions de personnes en France touchées par la surdité), l'utilisation de la lecture labiale reste un mode important pour la perception de la parole. C'est ce qui a motivé Cornett (Cornett, 1967) à développer le système du Cued Speech comme geste codé en appoint des lèvres (Figures I-2 et I-3). Cette méthode développée depuis pour plus de 60 langues s'appelle Langue Française Parlée Complétée ou encore code LPC pour le Français.

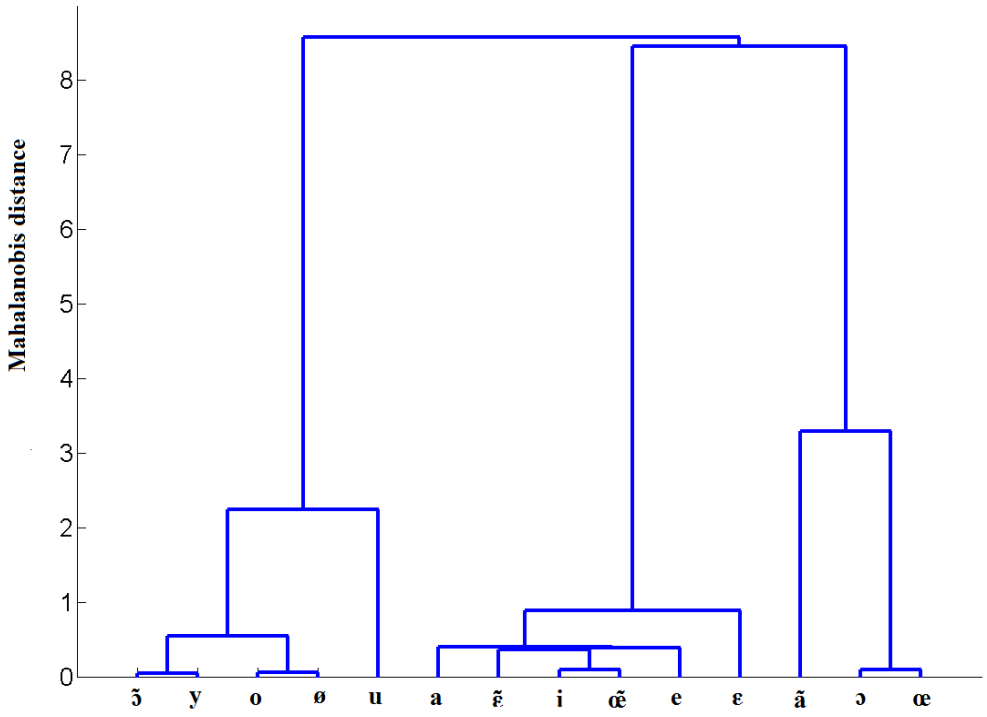


Figure I-1 : Dendrogramme des distances entre 1167 réalisations de voyelles du Français produites par une locutrice calculées à partir des paramètres extraits du contour interne des lèvres (étirement, ouverture, aire interolabiale). En se fixant un seuil à 4, on peut distinguer sur cette figure trois groupes de voyelles (visèmes) pour le Français à l'intérieur desquels les formes aux lèvres sont similaires: le groupe v1 [ɔ̃, y, o, ø, u] des voyelles fermées et arrondies aux lèvres, le groupe v2 [ã , ɔ, œ] des voyelles arrondies aux lèvres et mi-ouvertes, le groupe v3 [a, ɛ̃, i, œ̃, e, ɛ] des voyelles non arrondies et ouvertes aux lèvres. A noter que l'appartenance de la voyelle [œ̃] au groupe v3 est vraisemblablement due à une opposition réduite avec [ɛ̃] (voir Carton F., 1974 ; Durand et al., 2009). Figure issue de la thèse de Noureddine Aboutabit (2007).

Dans le système du Cued Speech la main présentant une configuration particulière vue de dos (parmi 8 pour le Français) pour indiquer la consonne vient pointer du doigt une position précise (parmi 5 pour le Français) sur le visage, à côté ou à la base du cou pour spécifier la voyelle. Chaque configuration, respectivement position de la main est utilisée pour un groupe de consonnes respectivement voyelles, correspondant à des phonèmes facilement discriminables aux lèvres. Inversement, des phonèmes ayant des formes aux lèvres similaires (les sosies labiaux) correspondent en Cued Speech à des configurations de main différentes s'il s'agit de consonnes ou des positions différentes s'il s'agit de voyelles. D'une certaine manière, c'est la vue conjointe de la main et des lèvres qui permet la perception d'un phonème, chacune de ces deux composantes portant une partie complémentaire de l'information. Ce système particulièrement efficace permet la perception complète de la parole et pour les enfants sourds l'utilisant depuis le plus jeune âge, des représentations complètes du système phonologique,

l'acquisition du langage avec des compétences en lecture et écriture comparables à des enfants entendants (Leybaert et al., 2000).

Un grand nombre de travaux ont eu trait à l'efficacité du Cued Speech en perception mais peu ont traité de la production, de son organisation, en particulier la coarticulation des différentes composantes du Cued Speech.

Comment un système artificiel datant de 1967 pourrait-il encore avoir de l'intérêt dans les années 2000 ? A part l'efficacité du Cued Speech pour l'acquisition d'un autre système artificiel comme la lecture, il a été l'opportunité d'étudier la coordination main-lèvres au niveau syllabique. Enfin, avec le développement grandissant des implants cochléaires, cette méthode facilite d'une certaine manière l'accès à la modalité auditive (Leybaert & Lassasso, 2010) ce qui renouvelle son intérêt.

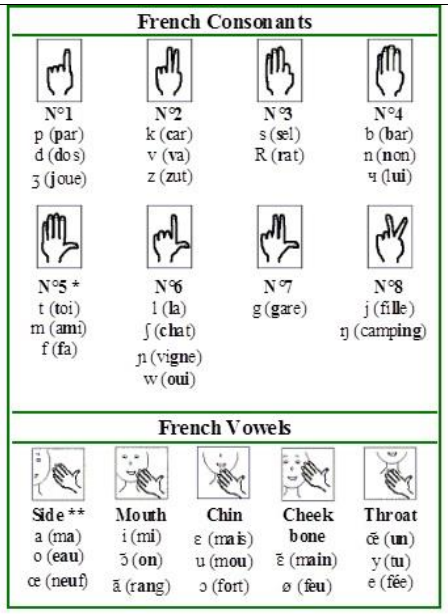


Figure I-2 : Système LPC. Figure issue de Beautemps et al., 2012

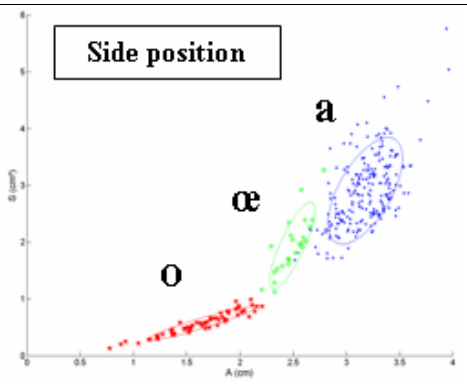


Figure I-3 :
Distribution des paramètres labiaux pour les voyelles dans le cas où l'information manuelle LPC est donnée (ici « side position »). La classification des voyelles est alors possible. Données de la thèse de Noureddine Aboutabit (2007).

I.1 Résultats sur la coordination oro-bracho-faciale

Nous présentons les travaux sur la coordination temporelle main-lèvres en relation avec des indices de la production audio afin d'établir la nature de la structure syllabique en Cued Speech en référence à la co-articulation de la parole. La production du code LPC n'avait jamais été étudiée et nous l'avons fait par une technique de suivi des mouvements labiaux et de main (Figure I-4) de quatre codeuses professionnelles à partir de l'enregistrement vidéo de séquences de parole avec LPC. Les figures I-5 et I-6 sont des exemples de signaux extraits des images vidéo par traitement d'image permettant de caractériser les différents déplacements.

Figure I-4 : Image d'une séquence extraite de l'enregistrement vidéo d'une codeuse portant un gant de données pour capturer la flexion des segments de doigts sur lequel des pastilles de couleur sont utilisées afin de faciliter l'extraction de leurs coordonnées par traitement d'images (utilisation d'artifice de maquillage des lèvres pour les mêmes raisons). Les axes en rouge définissent le repère des coordonnées. Figure issue de Attina et al., 2004.

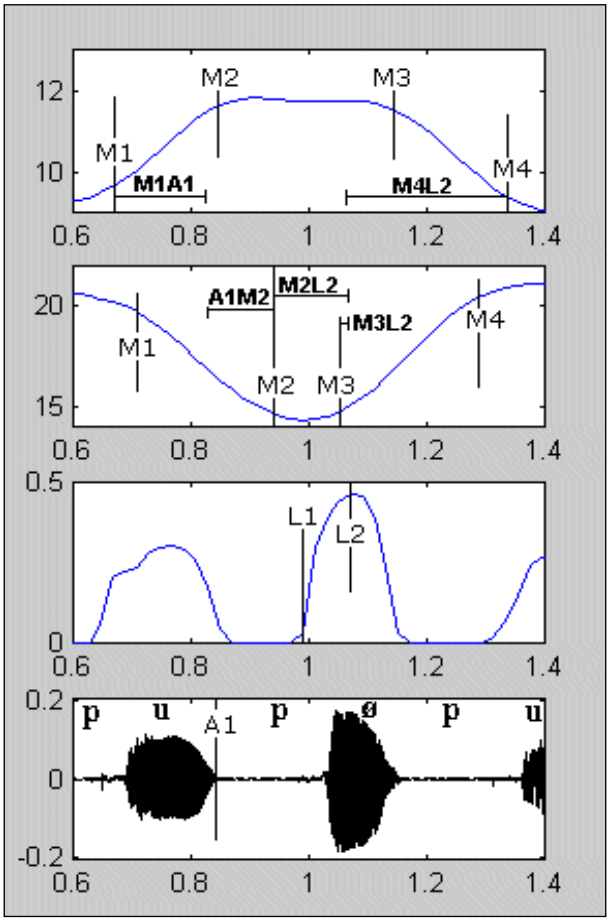


Figure I-5. De haut en bas: Déplacements horizontal x (cm) et vertical y (cm) de la main définis à partir de la pastille placée sur le dos (une augmentation de x indique un mouvement vers le côté droit du visage du sujet, une augmentation de y indique un mouvement vers le bas du visage) ; l'aire intero-labiale (cm^2) et le signal audio correspondant pour une séquence [pupøpu]. Pour la main, M1 et M3, instants de début de déplacement, M3 et M4 instants d'atteinte de la position cible. Pour les lèvres, L1, instant de début d'ouverture et L2 instant d'atteinte du climax dans la voyelle. Pour le signal audio, A1 repère le début du silence acoustique. Figure issue de Beautemps et al., 2012.

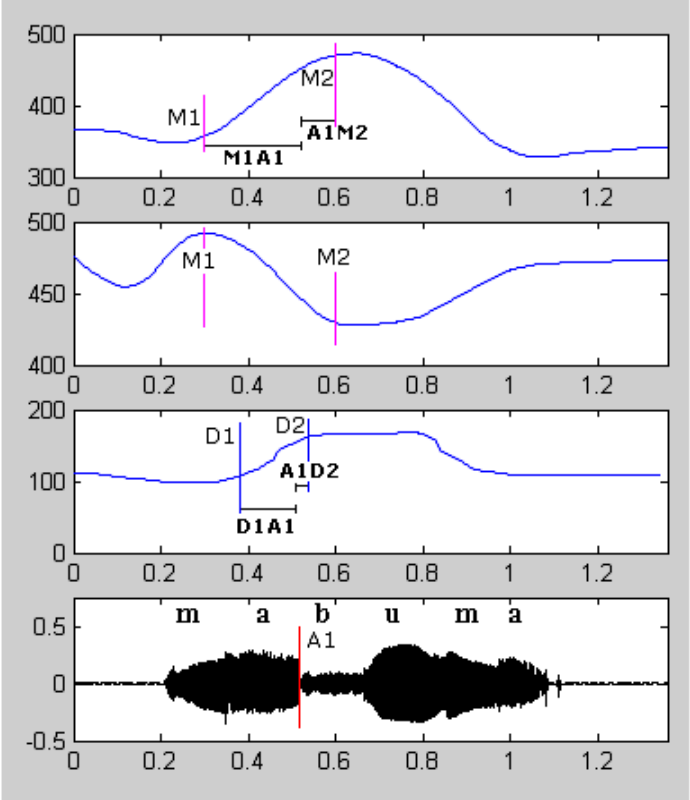


Figure I-6. De haut en bas: Déplacements horizontal x (cm) et vertical y (cm) de la main définis à partir de la pastille placée sur le dos (une augmentation de x indique un mouvement vers le côté droit du visage du sujet, une augmentation de y indique un mouvement vers le bas du visage) ; donnée brute montrant la flexion du pouce issue du capteur du gant de données situé sur l'articulation du pouce et le signal audio correspondant pour une séquence [mabuma]. Pour le pouce, D1 est l'instant de début du mouvement de flexion et D2 la fin. Figure issue de Beautemps et al., 2012.

La figure I-5 montre un mouvement de main qui se caractérise par une phase de transition entre M1 et M2 suivie d'une phase de tenue en position cible entre M2 et M3 et un déploiement de la clé digitale de configuration de la main entre D1 et D2 (Figure I-6). Le décours temporel de l'aire intero-labiale indique par sa valeur nulle une fermeture aux lèvres durant la production de la consonne [p] puis une détente en L1 suivie d'une phase d'ouverture atteignant son maximum en L2 dans le cas de la voyelle [a].

La Figure I-7 résume l'ensemble des résultats de co-production des différentes composantes que son main, doigts, lèvres et son de parole impliquées dans la production de parole avec LPC. Cette figure indique ainsi que dans une succession de syllabes de type Voyelle-Consonne-Voyelle, la mise en forme de la clé digitale de la consonne se superpose au geste de transition de main en direction de sa cible pour la voyelle finale. Le geste de formation de la clé digitale dure en moyenne 171 ms et termine son déploiement en tout début de consonne. La main commence son déplacement vers la cible correspondant à la voyelle finale alors que la cible aux lèvres de la première voyelle est tout juste atteinte, et 205 à 239 ms avant

le début de la réalisation sonore de la consonne pour arriver à sa position 33 à 37 ms après le début de cette consonne, et bien avant que la voyelle soit visible aux lèvres (de 172 à 256 ms).

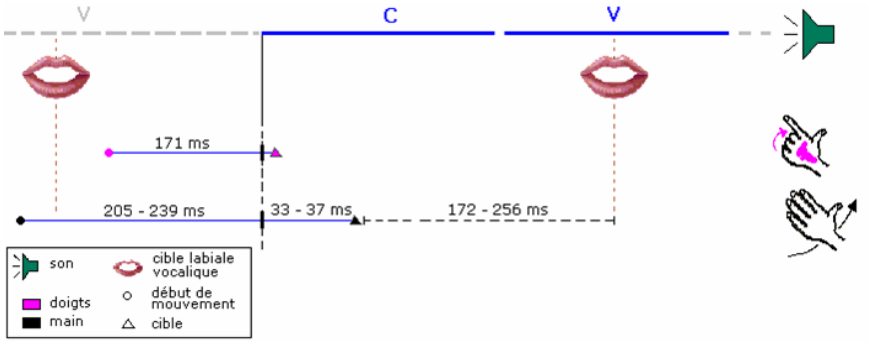
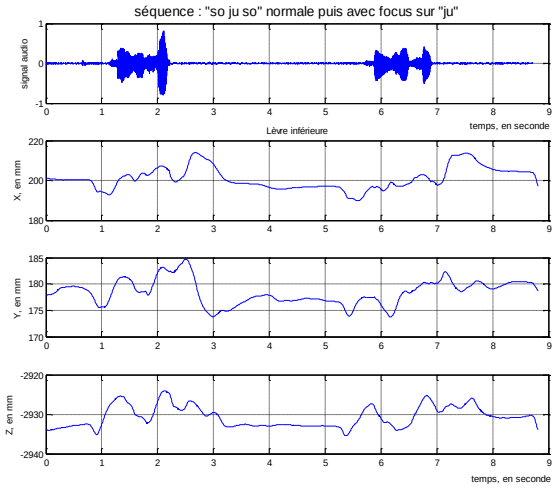


Figure I-7 : Schéma montrant la coordination temporelle main-lèvre-son du code LPC ou Cued Speech en Langue Française. Figure issue de Attina et al., 2004.

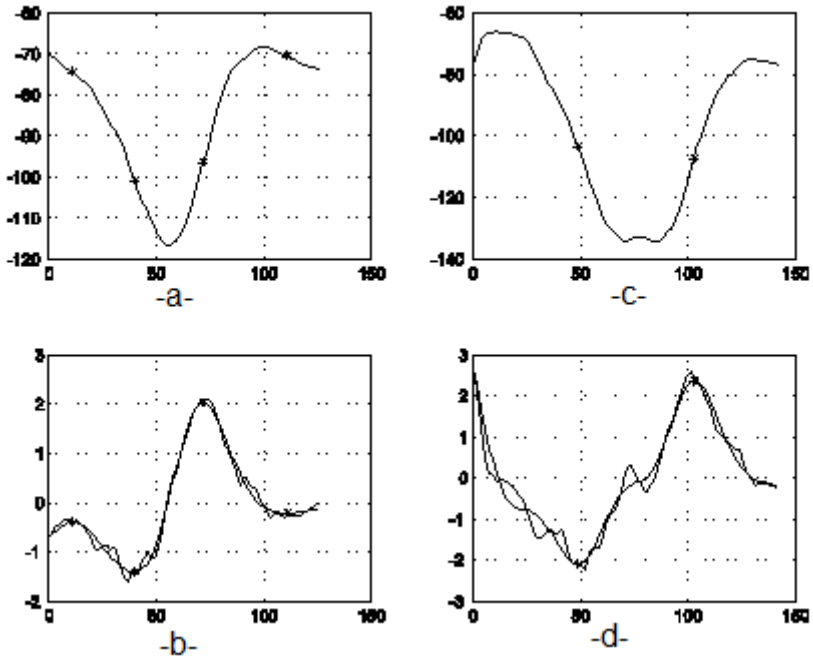
Le résultat majeur est que le geste de la main – contre toute attente – précède le geste des lèvres d'environ 200 ms. Cette anticipation donne un rôle inattendu à la parole visible : celui de venir désambiguïser le geste manuel, conçu au départ pour désambiguïser la parole. Des règles de phasages entre main, lèvre et son de parole ont été déduites de ces résultats et utilisées pour réaliser le premier prototype de synthèse du LPC à partir du texte dont l'utilisation ne nécessite aucune phase d'entrainement par les utilisateurs (Attina et al., 2004).

Nous poursuivons actuellement ces travaux par l'étude de la robustesse de ce phasage à partir de l'analyse de l'effet de focus sur la syllabe en utilisant les signaux issus d'un enregistrement en optotrak 3D d'un sujet codeur en LPC (Figures I-8 et I-9). Les résultats non encore publiés confirment le schéma de phasage et montrent l'allongement de la phase de tenue de la main en cas de focus sur la syllabe (Figure I-10, voir tracés (b) et (c)) en lien avec l'augmentation de la durée correspondante de la réalisation audio, renforçant au passage l'hypothèse de l'ancrage du geste codeur sur la parole.

Figue I-8 : Dans la chambre anéchoïque de la plateforme BEDEI de GIPSA-lab, sujet équipé des diodes à émission Infra rouges du système optotrak.



Figue I-9 : Séquence [sojuso] en condition normale de production puis avec focus sur la syllabe cible [ju]. De haut en bas le signal audio, les coordonnées X, Y et Z de la diode située sur la lèvre inférieure.



Figue I-10 : Séquence [sojuso] avec geste codeur en Cued Speech (i) en haut tracé de la première composante principale sur les coordonnées X,Y, Z d'une des diodes du dos de la main, (ii) en bas tracé de la vitesse et de la vitesse filtrée pour la condition normale (à gauche (a) et (b)) et la condition de focus (à droite (c) et (d)). On peut ainsi observer une tenue plus longue de la syllabe intermédiaire [ju] dans le cas de la condition « focus » qui s'étend entre la coordonnée 60 et 90 en abscisse, unité d'échantillon à la fréquence de 120 Hz (tracé (c)).

I.2 Perception de l'anticipation de la main

Nous venons de discuter comment le geste de main peut anticiper celui des lèvres. Nous avons mené une expérience de perception pour étudier si cette avance est exploitée par les personnes sourdes utilisant le LPC. L'organisation du « phasage » des composantes du LPC observé en production nous a conduit à penser que pour la perception, l'information de main disponible tout d'abord permet de prédire un ensemble réduit de deux ou trois voyelles puis la sélection d'une seule parmi ces possibilités une fois l'information de voyelle visible aux lèvres. Cette hypothèse a été testée à partir d'une expérience perceptive de dévoilement progressif de la réalisation d'une syllabe CV. A partir d'enregistrements vidéo de séquences « mutumaCVma » dans lesquelles la consonne pouvait être parmi [p, k, d, v] et la voyelle parmi [e, ɛ̃, ø, o], le test consistait à identifier la consonne et la voyelle à partir du déroulement de la vidéo jusqu'à un des six points de troncature choisi aléatoirement (voir Figure I-11 pour la définition des points de troncature).

Figure I-11 : Exemple de points de troncature classés de 1 à 6 en allant de haut en bas et de gauche à droite. Le premier point correspond au début du mouvement de la main, le 2nd au début de la mise en forme de la clé digitale pour la consonne, au point 3 la main commence son déplacement, au point 4 la clé est déployée et indique la position cible sur le visage qui est atteinte au point 5 alors que la consonne est visible aux lèvres, enfin le point 6 où la voyelle est visible aux lèvres. Figure issue de la thèse de Virginie Attina (2005).

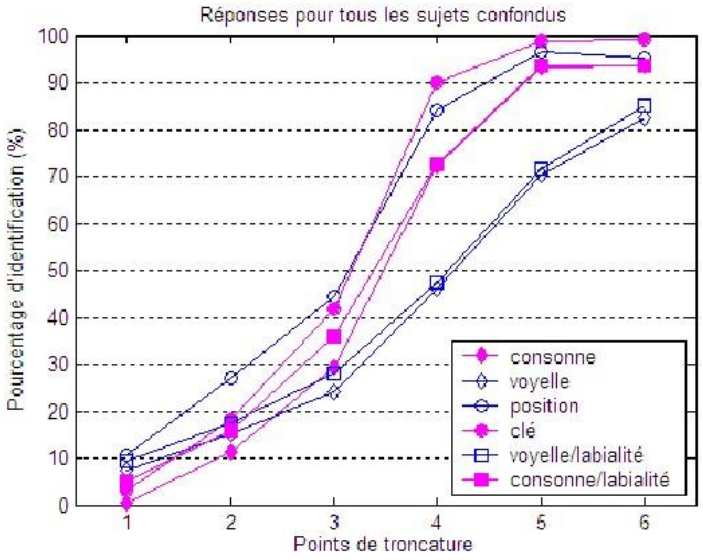


Figure I-12 : Scores moyen pour les 16 sujets d'identification de la consonne, la voyelle, la consonne classée selon son niveau de labialité, la voyelle classée selon son arrondissement, la clé et la position manuelle obtenus aux différents points de troncature de séquences « mutumaCVma ». Figure issue de la thèse de Virginie Attina (2005).

La Figure I-12 montre un saut à 85 % - 90 % au point de troncature 4 (où la clé digitale est visible, la main indiquant la position cible de la voyelle, la consonne commençant à être formée aux lèvres) pour l'identification de la clé respectivement la position de main, niveaux qui ne seront atteints qu'au point de troncature 5 pour la consonne et point de troncature 6 pour la voyelle une fois que l'information des lèvres aura été visible. Ces résultats montrent donc que quand l'information de main est disponible en avance des lèvres, celle-ci est bien perçue et exploitée par les sujets sourds utilisant le Cued Speech (voir le DEA de Florence Bouaouni et la thèse de Virginie Attina, 2005).

I.3 Bilan

La coordination observée entre main, lèvres et son confirme selon nous, le principe d'une avance de la main sur la réalisation sonore, programmée empiriquement par Duchnowski et collègues (2000) dans leur système automatique d'affichage du Cued Speech. Cependant, l'ampleur de ce comportement anticipatoire peut varier en fonction du locuteur, du débit de parole, et d'autres paramètres de niveau segmental ou supra-segmental.

De l'ensemble de nos résultats, il ressort une vision chamboulée de la fonction du Cued Speech (Beautemps et al., 2012). Dans le langage commun, le Cued Speech est considéré comme un augment venant en appoint des lèvres. De l'étude de nos données, un schéma général autre semble cependant se dessiner dans l'organisation temporelle lors de la production de

syllabes CV. La main atteint sa cible désignant la voyelle au début de la syllabe CV et la quitte pour se déplacer vers la position correspondant à la syllabe suivante avant même que la cible de la voyelle ait été atteinte aux lèvres. Cette anticipation de la main sur les lèvres est de plus exploitée en perception.

Il semble donc que dans le Cued Speech, le contrôle de la production contraigne l'organisation temporelle de la main et des lèvres. Ainsi le contrôle des contacts (ou cibles) vocaliques manuels va se trouver en phase avec celui des contacts consonantiques visibles (occlusions ou constrictions labiales). Ce phasage est assez précis pour que, quelles que soient les variations de la durée de production de la syllabe CV, l'aboutissement de la détente du système main-bras se produise dans la phase de tenue de l'attaque consonantique. Notre hypothèse est que le système de Cornett a été recodé en termes neuralement compatibles pour le contrôle des gestes des voyelles et des consonnes dans le LPC et la parole.

L'anticipation de la main sur les lèvres est importante à prendre en compte dans les systèmes de synthèse de la parole intégrant le code LPC comme nous l'avons fait dans notre prototype, et de façon plus large dans les traitements automatiques et les questions d'intégration perceptive.

Comme je le précisais dans la partie introductive du mémoire, les résultats sur la production ont pu être observés grâce à un banc expérimental précieux alliant vidéo (son et image) et gant de données (système de capture du mouvement des doigts à l'aide de capteurs sensibles aux flexions et placés dans un gant au niveau des articulations des main-segments de doigts et entre segments de doigts). L'ajout du système de gant de données dans le banc expérimental audio-visuel initial a nécessité une solution de synchronisation de l'ensemble.

Le gant de données a été acquis dans le contexte d'une collaboration en enseignement avec l'atelier de réalité virtuelle de la filière ENSERG devenue PHELMA de Grenoble-INP, pour lequel j'ai conçu un TP (allant de l'idée du sujet à sa conception en passant par le choix du matériel et la rédaction d'un document de type « poly » destiné aux étudiants, expliquant des éléments théoriques en analyse de données, ACP et classification gaussienne).

Contexte contractuel : projets « Jeune Equipe » du CNRS, programme « Cognitique » du ministère ;

Liste des intervenants : Denis Beautemps (CR), Marie-Agnès Cathiard (MCF), Virginie Attina (DEA puis Doctorat), Matthias Odisio (Doctorant), Florence Bouaouni (DEA), Pablo Sacher (DEA), Coriandre Vilain (IR), Christophe Savariaux (IR), Sara Hamdouchi (PFE), Edwin Corolleur (Stage d'étude d'ingénieur 2A, ENSE^3), Simon Rousseau (Stage d'étude d'ingénieur 2A, PHELMA).

Chapitre II - Transcodage entre espaces visuel et audio pour le Cued Speech

Introduction

Le défi adressé par cet axe de recherche est d'estimer un ensemble de signaux multimodaux à partir d'autres signaux collectés sur la production de parole produite par un sujet pour un interlocuteur. Cette opération de transcodage vise à permettre/améliorer la communication parlée entre ces deux interlocuteurs allant jusqu'à un transcodage inter-modalité, appelé aussi substitution sensorielle. Elle permet de compenser divers déficits moteurs du locuteur (laryngectomie, etc.), perceptifs (surdité, etc.) de l'interlocuteur, de s'adapter aux dispositifs de capture ou de restitution des signaux disponibles ainsi que les conditions environnementales (rapport signal sur bruit, etc.), les conditions de production (parole chuchotée voire articulation silencieuse, etc.) ou de perception (position/distance au haut-parleur, etc.) voire de transmission des signaux (exploitation de la redondance multimodale pour compression, etc.).

Deux sources de connaissance peuvent contribuer à ce transcodage : (a) la connaissance a priori des corrélations entre les diverses signatures – activités neuronales et neuro-musculaires, mouvements articulatoires, paramètres aérodynamiques, géométrie du conduit vocal, déformations faciales et structure acoustique du son - de l'activité vocale ; (b) la connaissance a priori des contraintes phonétiques, phonologiques et linguistiques de la langue parlée.

Des outils puissants permettent de modéliser ces connaissances et de les associer de manière optimale avec les signaux disponibles à l'entrée et désirés à la sortie. Sur un axe ordonnant les méthodes suivant leur indépendance à la langue utilisée, on trouve aux deux extrémités :

- Les méthodes utilisant un pivot phonétique, combinant reconnaissance et synthèse de parole pour assurer le respect d'organisation phonologique du langage parlé. Notons que les opérations de reconnaissance et de synthèse peuvent ici faire appel à des techniques de modélisation très différentes. Ainsi si les modèles phonétiques à base de Chaînes de Markov Cachés (HMM) constituent la base de la plupart des systèmes de reconnaissance, la synthèse par concaténation d'unités multi-représentées et de longueur variable reste très populaire. Notons cependant la montée en puissance de la synthèse par modèles de trajectoires basés HMM permettant plus de symétrie voire un apprentissage conjoint des modèles de reconnaissance et de synthèse.

- Les méthodes capturant les corrélations des signaux de manière directe sans pivot phonétique par techniques de mapping. Ces techniques capturent les corrélations entre trames de signaux en entrée et en sortie par quantification vectorielle ou par mélange de Gaussiennes GMM.

La dichotomie entre ces méthodes est de plus en plus floue car il semble intéressant de combiner les deux types d'approche permettant de gagner en robustesse et précision.

Deux projets sur le Cued Speech se sont inscrits dans cet axe : (1) Le transcodage de la Langue Française Parlée complétée (code LPC) en parole audio: l'enjeu est ici le décodage phonétique des mouvements faciaux articulés silencieusement et accompagnés par des gestes de main par des méthodes de reconnaissance automatique. Le projet ANR/TELMA est articulé autour d'une telle technologie dont l'objectif est de permettre à un sourd et un bien-entendant de converser par visiophonie assistée par ordinateur ; (2) Le transcodage de parole audio en parole visuelle complétée par le code LPC : l'enjeu est ici d'inférer les mouvements des lèvres et de la main permettant d'expliquer au mieux la séquence audio observée. Le problème générique est l'inversion visuo-acoustique par des approches stochastiques opérant directement à partir de données de type articulatoires collectées sur des sujets.

II.1 Contexte du projet TELMA de Téléphonie à l'usage des malentendants

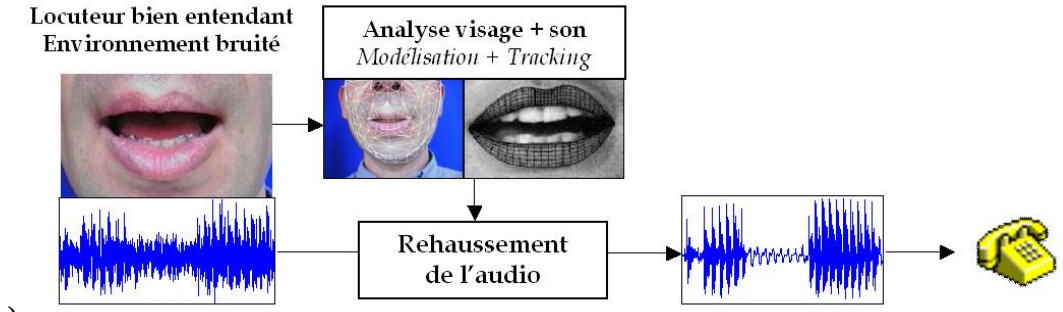
Le téléphone pour les sourds, une idée qui peut surprendre, est l'objectif du projet qui se propose de poser les bases d'une interface à l'usage des malentendants, utilisant la *modalité visuelle* pour permettre de communiquer avec les utilisateurs entendants utilisant la *modalité audio*. Pour contribuer au besoin de multimodalité, le projet s'est fixé de développer des briques autour des techniques du débruitage audiovisuel, et du transcodage entre parole acoustique et lecture labiale augmentée de la Langue Française Parlée Complétée. Les études sont axées sur les techniques de débruitage audiovisuelles et sur l'analyse et l'animation faciale centrées sur la lecture labiale et la LPC. Il s'agit pour la dimension LPC de réaliser un système automatique de traduction lecture labiale+LPC vers parole acoustique et inversement sachant que le vecteur qui transporte l'information dans le réseau téléphonique reste le son.

Le projet Telma s'est fixé l'objectif de développer les briques technologiques nécessaires à la mise en œuvre des fonctionnalités visées et de les intégrer dans un démonstrateur. Dans ce cadre, objectifs sont multiples (voir Figure II-1) :

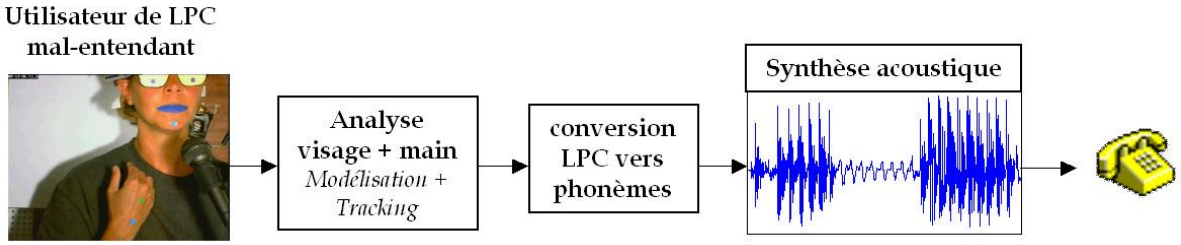
- L'exploitation de la modalité visuelle dans le débruitage dans des conditions de communication réalistes ;

- La reconstruction d'une chaîne phonétique par reconnaissance du LPC ;
- Le contrôle de la synthèse visuelle avec LPC à partir du son de parole.

a)



b)



c)

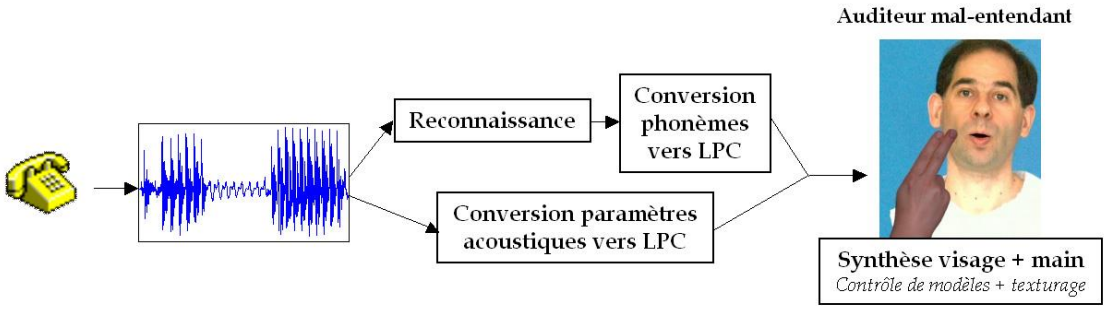


Figure II-1 : Les 3 fonctionnalités de TELMA

Partenaires GIPSA-lab et ICP (16): Denis Beautemps, Laurent Girin, Noureddine Aboutabit, Gérard Bailly, Marie-Agnès Cathiard, Frédéric Elisei, Panikos Heracleous, Bertrand Rivet, Pablo Sacher, Christophe Savariaux, Coriandre Vilain, Alice Caplier, Vincent Girondel, Christian Jutten, Stéphane Mancini, Sébastien Stillitano.

Partenaire LIG (3): Laurent Besacier, Jean-François Sérignat, Viet-Bac Le.

Partenaire Orange/France Telecom (8): Gaspard Breton, Thomas Burger, Denis Chêne, Danièle Pelé, Pascal Perret, Mélody Tribout, Sylvie Vidal, Oxana Govokhina.

Partenaire LTCI (2): Yves Mathieu, Zahir Larabi.

Partenaire CHU-ORL (12): Sébastien Schmerber, Martine Marthouret, *Clémentine Huriez, Myriam Douibi, Aurélie Chevallier, Nicolas Deffois, Laurie Fabbri, Juliette Huriez, Louis Magnin, Richard Nomballais, Godefroy Vannier, Nadège Clauss*

II. 2 Reconnaissance du Cued Speech

La reconnaissance du Cued Speech est composée d'un module de caractérisation de l'information visuelle de la main et des lèvres vue de face et d'un module de fusion de ces deux composantes (Figure II-2). Dans le contexte TELMA, cette partie a trait à la fonctionnalité illustrée par la Figure II-1 (b). Il s'agissait d'un axe de recherche totalement innovant, aucun travail n'existant antérieurement. Nous avons d'emblée choisi de marquer l'information visuelle pertinente pour le code LPC dans les enregistrements vidéo afin de faciliter l'extraction des composantes manuelles et labiales et ainsi consacrer nos efforts sur la modélisation de la fusion, véritable enjeu de ce travail.

Les modèles de fusion classiquement utilisés en reconnaissance audiovisuelle de la parole (voir par exemple la taxinomie de leur classification sur une échelle allant de fusion précoce à fusion tardive selon Schwartz et al., 1998 ou leur classement selon une typologie en fusion de données *vs.* fusion de décisions chez Potamianos, 2012) ont été appliqués à la fusion des deux composantes visuelles main et lèvres du code LPC.

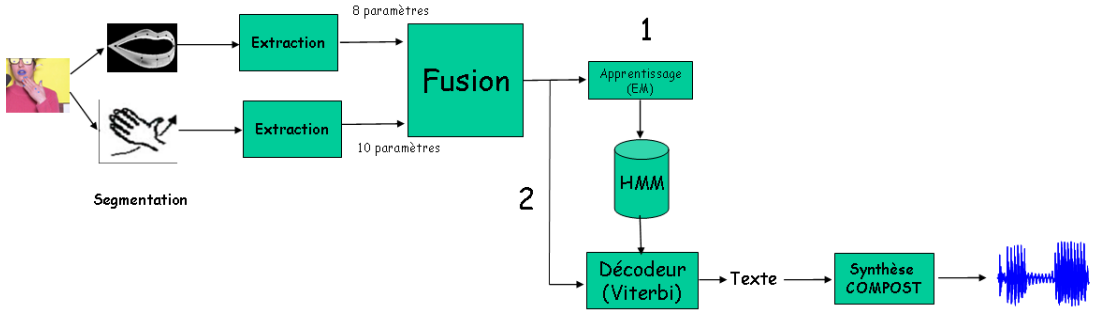


Figure II-2 : Schéma de principe de la chaîne de reconnaissance du code LPC et de la conversion en son de parole dans le cas de la modélisation par HMM. (1) Phase d'apprentissage ; (2) Phase de test.

II.2.1 Méthodes de fusion

Le modèle à recodage moteur (RM) renvoie à un encodage dans un espace articulatoire. Dans cet espace, les voyelles sont caractérisées par la forme globale du conduit vocal, définie par la position de la mandibule, l'ouverture aux lèvres et la position de la langue à l'intérieur du tractus. Dans la fusion audio-visuelle, la position de la langue, qui est le seul articulateur qui ne soit pas toujours visible, peut-être inférée à partir du contenu spectral du son de parole dans le cas des voyelles. En Cued Speech, la position de main qui est prévue pour coder plusieurs voyelles prédit donc plusieurs formes de conduit vocal dont un seul sera cohérent avec la composante labiale. Un module de comparaison des formes de lèvres peut permettre de sélectionner la forme du conduit vocal qui devra ensuite être soumise à un classifieur afin de

terminer la phase de reconnaissance. Cette approche nécessite d'avoir des données articulatoires en nombre suffisant sous forme de dictionnaires de conduits vocaux par exemple pour le locuteur considéré ou un modèle articulatoire adapté au locuteur.

Dans le modèle à recodage dans la modalité dominante (RD), une des modalités est considérée comme la modalité principale et la seconde réalise une prédiction dans l'espace principal. En audiovisuel, la composante audio est la modalité principale car elle porte toute l'information. La forme des lèvres peut prédire des formes de spectres dans le domaine audio. En Cued Speech, aucune des deux composantes (main et lèvre) ne portent l'information phonétique complète sans ambiguïté. Il n'y a donc pas de modalité dominante évidente. De plus ces deux composantes sont complémentaires et peu de corrélation sont donc attendues entre leurs paramètres. Ces deux raisons ont fait que nous avons écarté le modèle RD.

II.2.2. Résultats par application de la méthode de fusion ID

Pour la méthode par Identification Directe (ID), les paramètres des deux modalités (ici les coordonnées des pastilles de main et de doigt ainsi que les paramètres labiaux d'aperture, d'étirement et d'aire du contour interne des lèvres) sont concaténés dans un même vecteur ou une même matrice avant application d'un classifieur de type HMM par exemple (Figure II-2). Cette modélisation appliquée à des sujets entendants et sourds a permis d'atteindre (voir Table II-1) des taux de reconnaissance de 87,6 % pour les voyelles, de 89 % à 94,9 % pour le cas de mots isolés, à comparer au score maximum de 76,1 % dans le cas des lèvres seules (« lip-reading »). La modélisation HMM étendue à la reconnaissance multi-locuteur LPC, a permis d'obtenir des résultats entre 87 et 92%, montrant ainsi la capacité à modéliser la variabilité inter sujets en LPC et ouvrant des perspectives prometteuses pour la reconnaissance indépendante du locuteur (Voir en annexe, Heracleous, Beautemps, Aboutabit, 2010).

* Codeur LPC normo-entendant

Données	Mixture de gaussiennes par état		
	1	2	4
Lèvres	56	59,4	72,0 %
Fusion main-lèvres	92,8	93,5	**94.9 %**

* Codeur LPC sourd

Données	Mixture de gaussiennes par état		
	1	2	4
Lèvres	67,1	75,5	76,1 %
Fusion main-lèvres	87.8	87,8	**89.0 %**

* Reconnaissance multi-codeur LPC

Données	Modélisation HMM		
	Normo-entendant	Sourd	Normo-entendant + sourd
Codeur LPC normo-entendant	94,9	0,6	**92,0 %**
Codeur LPC sourd	2,0	89,0	**87.2 %**

Table II-1 : Résultats de reconnaissance sur un vocabulaire de 50 mots isolés

II.2.3. Résultats par application de la méthode de fusion IS

Dans le modèle à Identification Séparée IS, chacun des canaux main et lèvres est tout d'abord considéré dans un processus de décisions qui sont ensuite fusionnées. Nous avons implémenté ce modèle pour le LPC dans une démarche de comparaison mais avec l'intérêt supplémentaire qu'il permettait une implémentation des données de perception du LPC indiquant un dévoilement progressif de l'information de la main puis des lèvres (voir chapitre précédent). Les différentes modélisations dans ce cadre ont été l'objet des travaux de la thèse de Noureddine Aboutabit.

Nous avons implémenté le modèle IS de la façon suivante (voir aussi en annexe Aboutabit, Beautemps, Besacier, 2007, actes de AVSP 2007) : un module de décodage de l'information de la main LPC composé d'un classifieur gaussien est appliqué aux coordonnées de la main à l'instant M2 d'atteinte de sa position cible permettant de sélectionner le classifieur gaussien des lèvres pour ses paramètres extraits à l'instant L2 d'atteinte de la cible aux lèvres. Dans notre implémentation M2 < L2 pour prendre en compte l'avance de la main. Les résultats obtenus sur les voyelles atteignent 89 % sur le corpus de test lorsque l'information de main est connue sans erreur. Ils se stabilisent à près de 78 % dans le cas où le module de décodage automatique de la main est utilisé. Ces résultats sont à comparer au score de 85,6 % obtenu précédemment avec le modèle ID (mais en utilisant un grand nombre de gaussiennes). Une amélioration du module de décodage automatique de la main d'une part et de l'appariement

main-lèvre (instants M2 et L2) devraient permettre d'améliorer ces résultats qui restent malgré tout encourageants.

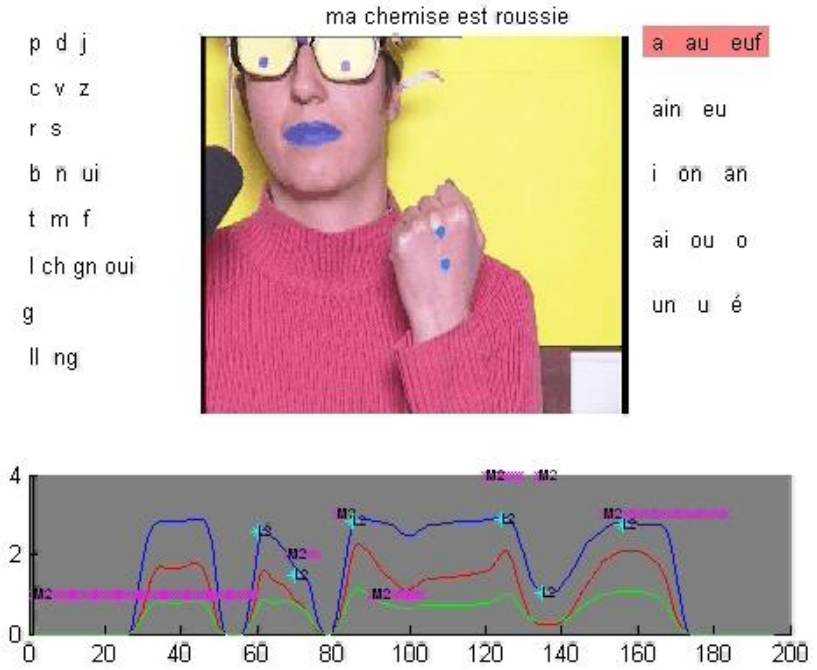


Figure II-3 : Reconnaissance progressive au cours du dévoilement progressif de l'information. En bas, décours temporel des paramètres de lèvre (aperture, étirement, aire du contour interne des lèvres). En superposition décodage de la main (couleur magenta), et instants L2 (croix bleues claires) de fusion des décisions main-lèvre.

II.3. Conversion audio vers le Cued Speech : mapping des espaces

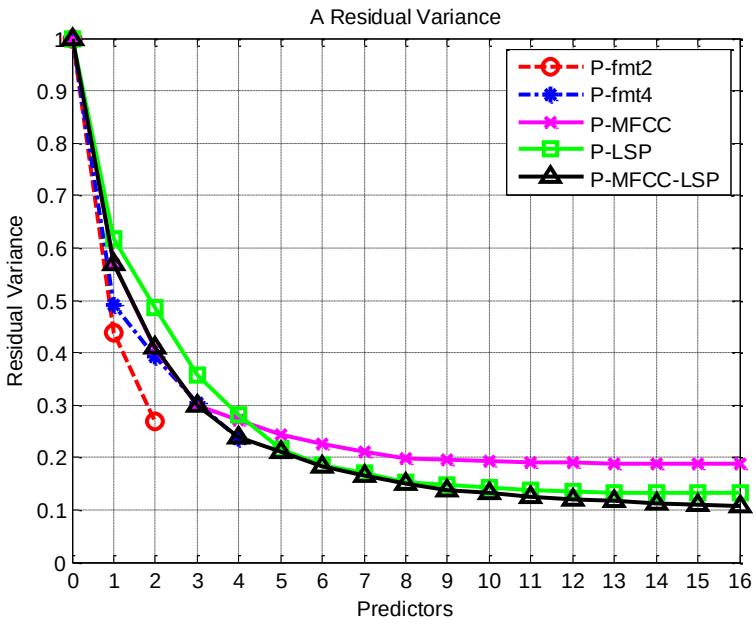
L'objectif visé dans cette partie a été d'étudier les méthodes de mapping des paramètres acoustiques du son de parole vers les paramètres visuels (labiaux et Cued Speech) en utilisant un bas niveau d'interfaçage de type signal et donc sans le recours à la reconnaissance automatique de la parole. Dans le contexte du projet TELMA, cette partie renvoie à la fonctionnalité de la figure II-1 (c). L'introduction de la composante manuelle du Cued Speech dans ce programme constitue une véritable originalité de ce travail avec des retombées claires pour les systèmes de communication utilisant le geste associé à la parole ou non, tels que le Cued Speech mais aussi des gestes de pointage ou la Langue des Signes.

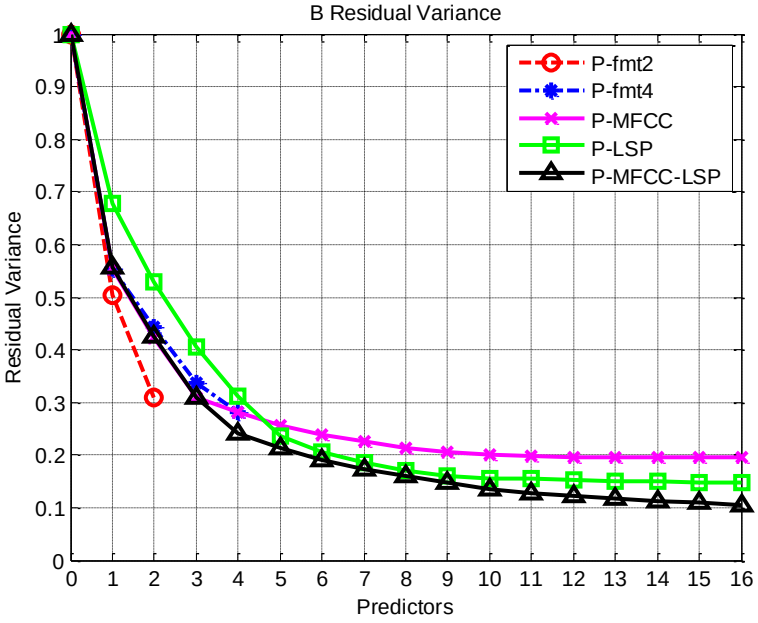
Nous avons abordé ce programme en traitant le cas des voyelles orales du Français avec le mapping des paramètres spectraux du son de parole vers les paramètres caractéristiques de la géométrie des lèvres vues de face et des paramètres de position de la main codant le Cued Speech sur le visage. Le mapping a consisté ici à déterminer les coefficients d'une combinaison linéaire reliant les paramètres de l'espace acoustique (les prédicteurs) aux paramètres de l'espace visuel (lèvres et Cued Speech) en minimisant l'erreur au sens des moindres carrés entre le résultat de la prédiction et les valeurs des paramètres visuels.

Nous avons comparé les méthodes en régression linéaire multiple et à mixture de gaussiennes GMM. L'introduction de l'outil GMM a pu se faire grâce à une discussion très fructueuse avec mon collègue Thomas Hueber, chercheur dans la même équipe que moi. Il avait en effet dans sa thèse utilisé cette méthode pour l'inversion de la relation articulatori-acoustique.

La difficulté supplémentaire dans le cas du Cued Speech était que d'une part les paramètres de position de main sont discontinus et d'autre part qu'il n'y a pas de causalité entre la réalisation acoustique et la position cible de la main sur le visage. En restant dans le contexte du « mapping », nous pouvons préciser qu'il s'agit d'une opération d'appariement (ici entre paramètres spectraux et LPC). Ainsi, les bons résultats obtenus dans l'application de ces outils sont d'autant plus intéressants, tant sur les aspects de l'objectif visé que sur l'augmentation de la connaissance des outils de modélisation, notamment leur comportement dans ces situations.

Les travaux dans ce cadre ont été au cœur de la thèse de Zuheng Ming. Les méthodes linéaires et en GMM ont donné toutes deux de bons résultats pour la prédiction des paramètres labiaux (Figures II-4, II-5 et II-6). Pour les paramètres de main, étant donné la forte non linéarité avec les paramètres spectraux, seule l'approche GMM a permis d'obtenir des résultats satisfaisants en terme de réplication des données. La modélisation la plus optimale a été obtenue lorsque les trois gaussiennes pour les lèvres ont été distribuées dans l'espace spectral en fonction de leur caractéristique labiale (les visèmes, voir figure I-1). De même pour les paramètres de position du Cued Speech, les cinq gaussiennes ont été distribuées dans l'espace spectral en fonction de la répartition des voyelles dans les cinq groupes de voyelle du LPC (Figure II-4). Ces résultats n'ont pas été améliorés par les traitements automatiques à base d'algorithmes K-means, EM et d'augmentation du nombre de gaussiennes.


A Residual Variance
Residual Variance
Predictors
P-fmt2
P-fmt4
P-MFCC
P-LSP
P-MFCC-LSP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16


B Residual Variance
Residual Variance
Predictors
P-fmt2
P-fmt4
P-MFCC
P-LSP
P-MFCC-LSP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

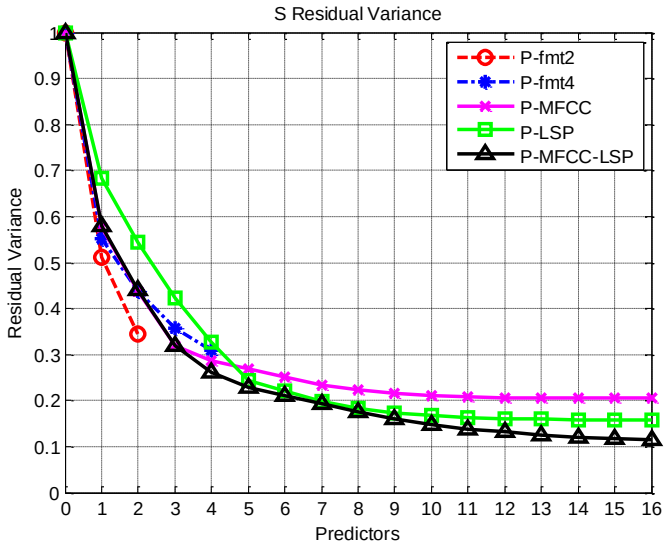
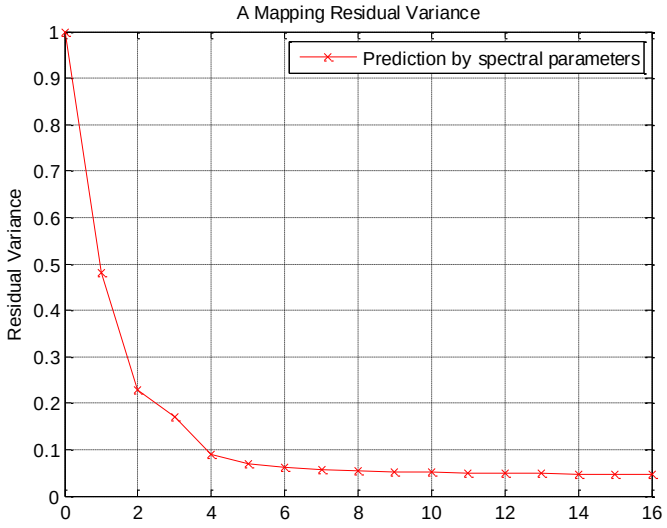


Figure II-4 – Modélisation linéaire. Variance résiduelle de la prédiction des paramètres labiaux, de haut en bas respectivement A (étirement aux lèvres), B (aperture aux lèvres) et S (aire du contour interne des lèvres), exprimés chacun relativement à leur variance totale, en fonction du nombre de prédicteurs (leurs composantes principales) et pour chaque ensemble de prédicteurs : 2, 4 formants (fmt), MFCC, LSP et l'ensemble de MFCC-LSP. Figure issue de Ming, Beautemps, Feng, 2013, contexte de la thèse de Zuheng Ming.



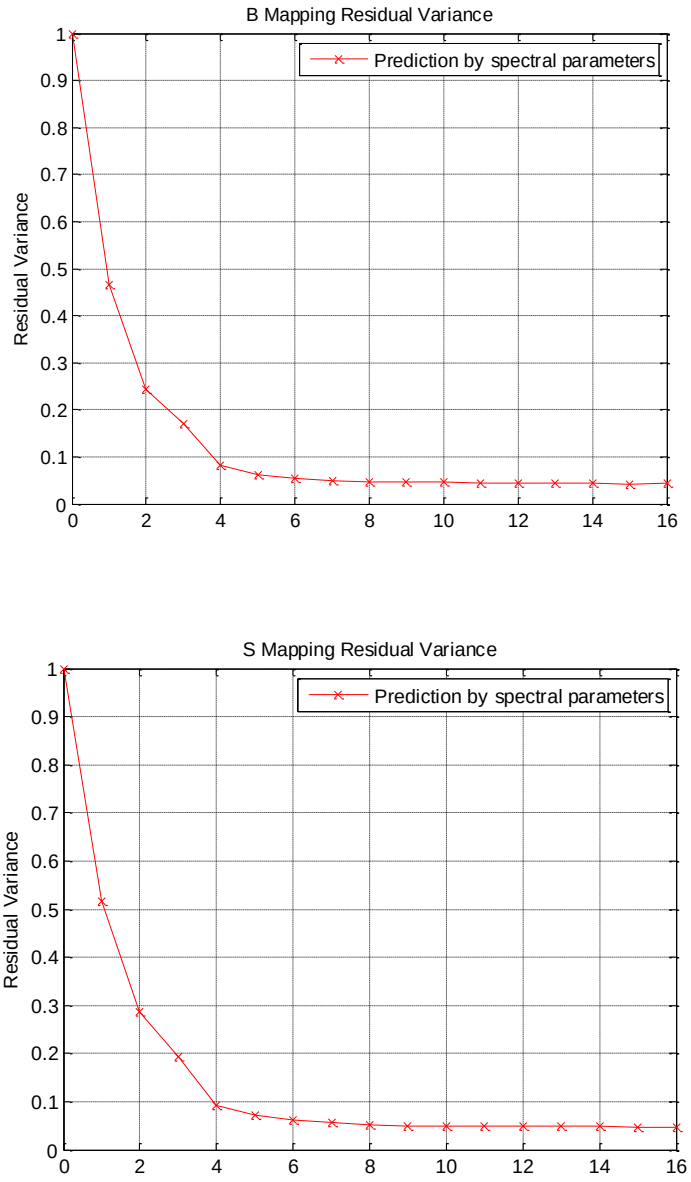


Figure II-5 – Modélisation GMM. De haut en bas, variance résiduelle des paramètres labiaux exprimés chacun relativement à leur variance totale, en fonction de la dimension de l'espace spectral. Figure issue de Ming, Beautemps, Feng, 2013, contexte de la thèse de Zuheng Ming.

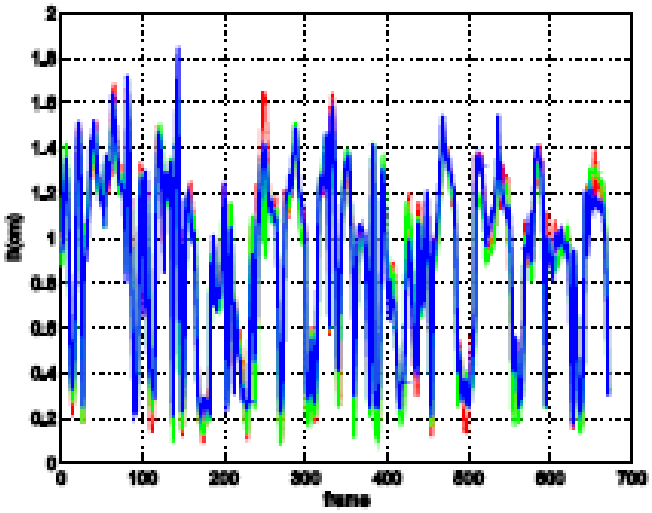
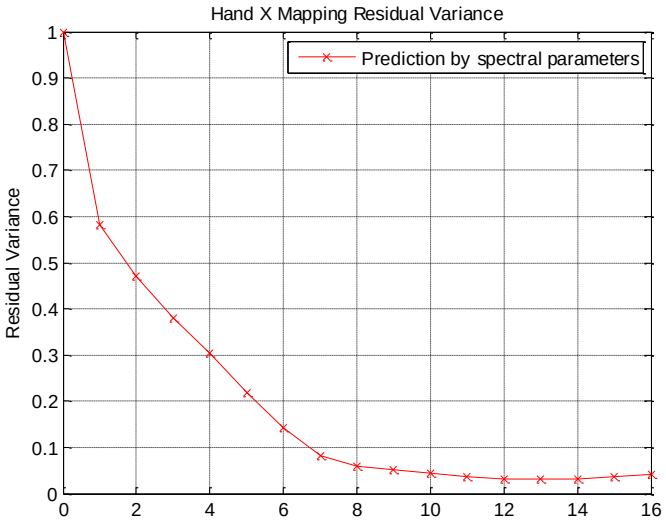


Figure II-6 – Estimation du paramètre d'aperture B sur un corpus de test de 671 voyelles. En rouge, valeurs réelles ; En vert, estimation utilisant la modélisation par GMM ; En bleu, estimation utilisant la modélisation linéaire. Figure issue de la thèse de Zuheng Ming.



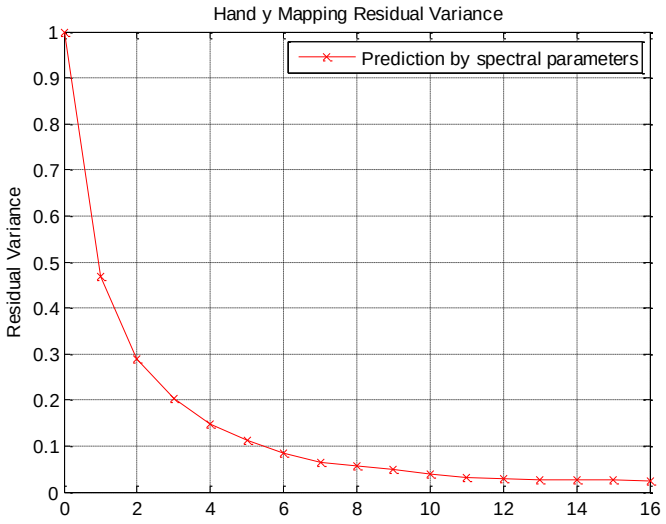


Figure II-7 – Modélisation GMM. De haut en bas, variance résiduelle des paramètres de main exprimés chacun relativement à leur variance totale, en fonction de la dimension de l'espace spectral. Figure issue de Ming, Beautemps, Feng, 2013, contexte de la thèse de Zuheng Ming.

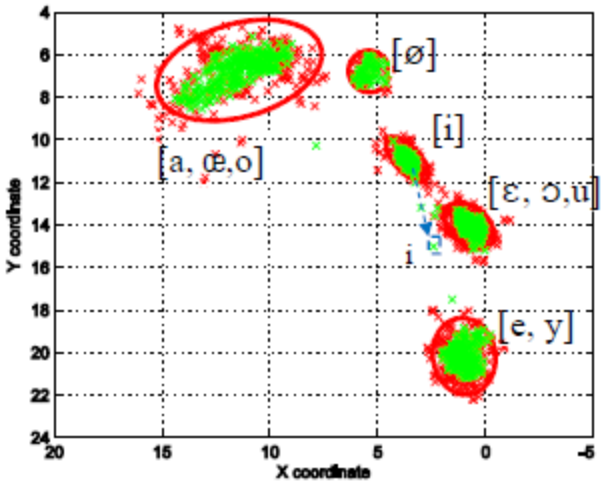


Figure II-8 – Estimation des coordonnées X et Y utilisant la modélisation par GMM. En rouge, données d'origine, en vert valeurs estimées. Figure issue de la thèse de Zuheng Ming.

L'ensemble de ces résultats en modélisation GMM a été étendu après la thèse de Zuheng Ming : les trois gaussiennes placées dans l'espace spectral en fonction du regroupement des voyelles dans les trois visèmes (voir Figure I-1 pour la définition des visèmes) ont été utilisées dans la prédiction de la transformation à base de DCT de toute la région d'intérêt des lèvres avec là aussi de bons scores de réplication (voir Figures II-9, II-10, contexte du projet d'étude PHELMA de Louis Gallet). Ces derniers résultats ne sont pas encore publiés.

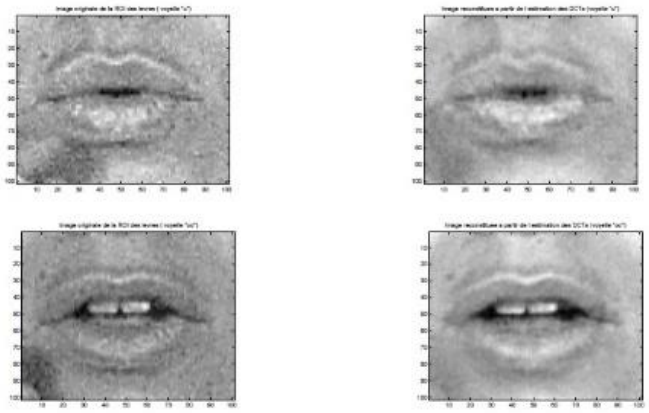


Figure II-9 : Composante Y du codage YUV de la ROI des lèvres. Colonne de gauche, ROI (101x101 pixels) originales de la voyelle [u] en haut et de la voyelle [ɔ] en bas ; colonne de droite, résultats de reconstruction des ROI correspondantes par un mapping GMM (3 gaussiennes à 16 dimensions placées dans l'espace spectral) des paramètres spectraux sur une transformation à base de DCT puis application de la DCT inverse pour revenir à la composante Y. La racine carrée de l'erreur quadratique moyenne sur un corpus de 700 images entre prédictions et données est de 11,3 en niveau de gris codé sur 8 bits, ce qui correspond à une erreur moyenne de 9,3 %.

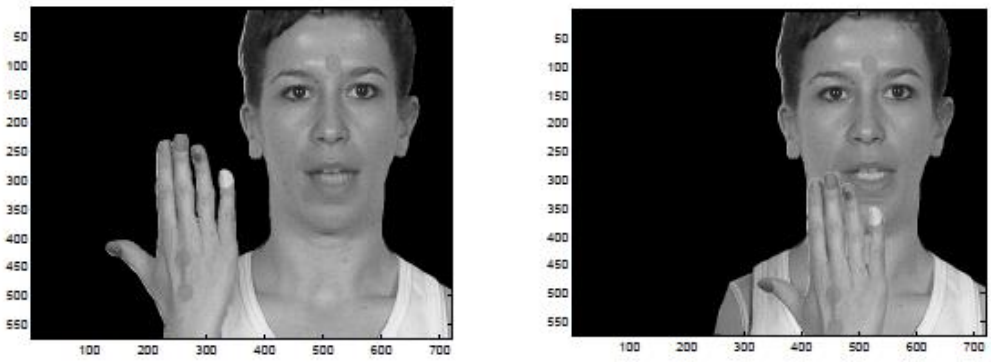


Figure II-10 : Résultat du mapping GMM (3 gaussiennes) pour la ROI des lèvres et pour la position cible de la main du LPC (qui définit ici les coordonnées x-y de l'extrémité du majeur) avec 5 gaussiennes. A gauche cas d'un [a], à droite, cas d'un [i]. Le reste du visage est identique dans les deux cas.

Enfin, un autre résultat important a été obtenu en traitement d'image de la région d'intérêt des lèvres où nous avons pu mettre en évidence la forte relation linéaire entre paramètres d'apparence et paramètres de forme des lèvres (Figures II-11 et II-12 ; voir aussi Ming, Beautemps, Feng, 2010). Nos résultats soulignent un lien entre ces deux approches. Des modélisations linéaires et en GMM de ces relations ont ainsi pu être développées et évaluées. Ces résultats pourront permettre de s'affranchir de l'utilisation d'artifices (bleu sur les lèvres) pour les sujets enregistrés, dans le cas d'apprentissages possibles pour les modèles de prédiction. Enfin les approches en « apparence » et en « forme » son classiquement présentées en miroir l'une de l'autre dans la littérature (Potamianos et al., 2003 ; Potamianos et al., 2012), et nos résultats montrent le pont qui peut exister au travers de cette dichotomie.

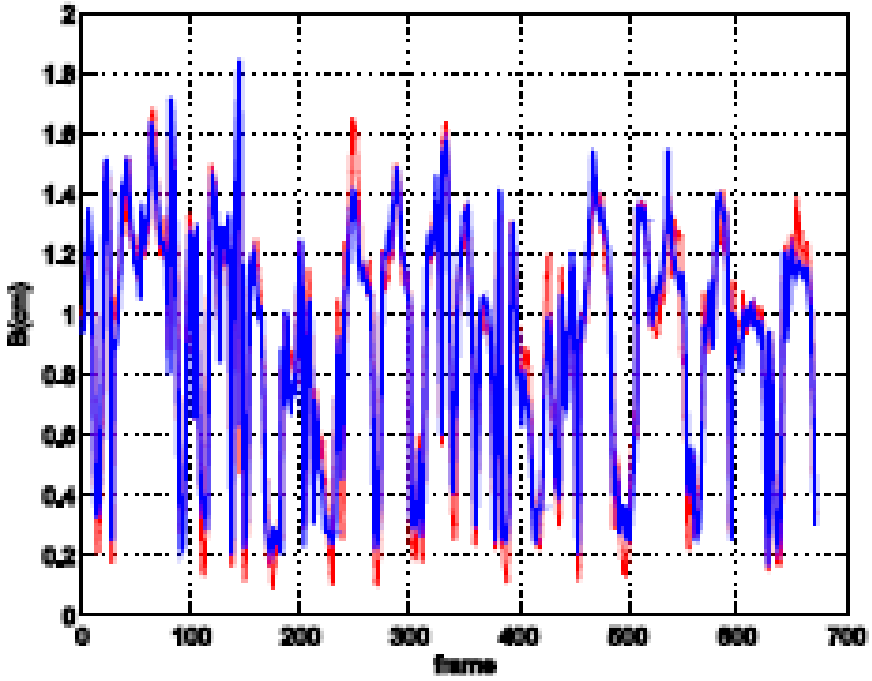


Figure II-11 – Estimation du paramètre d'aperture B sur un corpus de test de 671 voyelles. En rouge, valeurs réelles ; En bleu, estimation linéaire à partir des 20 composantes principales de l'ACP sur les 100 coefficients DCT sélectionnés (voir Figure II-13). Figure issue de la thèse de Zuheng Ming.

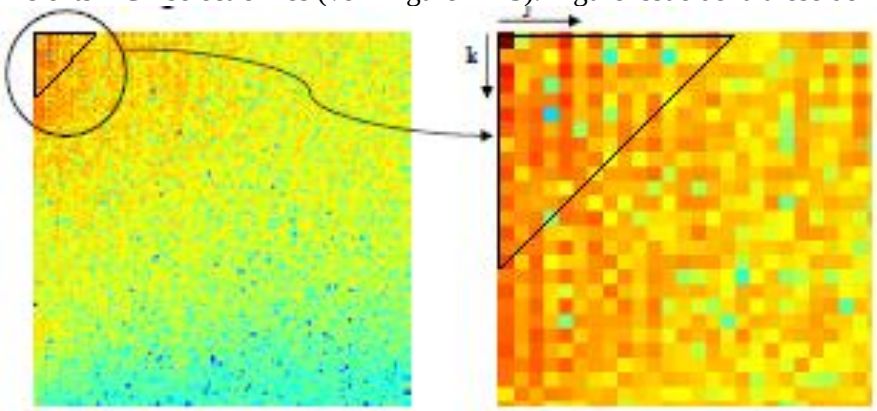


Figure II-12 – A gauche image (101x101) de la transformation en DCT à deux dimensions d'une ROI des lèvres. A droite, zoom sur les basses fréquences avec le masque de sélection triangulaire de sélection des 100 coefficients DCTs. Code couleur : Du plus sombre au plus clair les valeurs DCTs des plus élevées aux plus faibles. Figure issue de la thèse de Zuheng Ming.

Listes des intervenants : Denis Beautemps (CR), Gang Feng (PR), Noureddine Aboutabit (doctorant), Zuheng Ming (doctorant), Panikos Heracleous (Post-doc), Oxana Govokhina (post-doc), Jeanne Clarke (DEA), Louis Gallet (Stage d'ingénieur 2A, PHELMA), Thomas Burger (DEA) ;

Chapitre III – Parole, surdité et audition réhabilitée

Introduction

Ce thème vise à étudier et modéliser la production de parole et les processus de co-perception et production dans le contexte de la surdité.

Nous avons traité cette thématique à partir de trois actions principales sur un axe avec à une de ses extrémités la production sans retour auditif et à l'autre extrémité une situation de collaboration de la perception auditive voire audiovisuelle dans la production dans le cas de l'audition réhabilitée par l'implant cochléaire.

La première action traite de la production de parole liée à la physique. Nous avons voulu séparer les aspects liés au contrôle et le domaine de la surdité nous a paru intéressant à ce point de vue avec la limitation des effets de feedback auditif. Cette action n'est clairement pas ma dominante. J'y apparais par le contexte de la surdité et l'intérêt d'accompagner ce travail dans le cadre d'une co-direction de thèse avec Xavier Pelorson, dans le contexte de l'ANR PLASMODY et le projet Région « Cible » dans lesquels ces travaux se situent et dont je suis un des porteurs ou le porteur principal. La seconde action a consisté à introduire un nouveau champ expérimental en « shadowing » dans l'étude de la dimension visuelle de la parole, action réalisée dans le contexte du DEA d'Yvon Leborgne en 2002 que j'ai co-encadré avec Marie-Agnès Cathiard. La dernière action a trait à l'étude des relations sensori-motrices en surdité remédiée par l'implant cochléaire. Cette dernière action ne correspond pas non plus à la dominante de mon travail. Mais j'y apparais par l'apport d'une expérience sur le paradigme en close-shadowing et l'idée de l'application de ce paradigme à la surdité réhabilitée par l'implant cochléaire avec la proposition d'une expérience de co-perception et production intégrant ce paradigme ainsi que la dimension multimodale audio-visuelle dans le contexte d'une co-direction de thèse et du projet ANR PLASMODY.

III.1 Contexte du projet PLASMODY

L'objectif principal du projet est de mieux comprendre la perception et la production de la parole avant et après l'implantation cochléaire. Un accent particulier est mis sur le rôle des interactions multisensorielles et les substitutions intermodale pour l'accès à l'oralité chez les patients sourds implantés cochléaires. Sur le plan scientifique le projet PlasMody sous l'angle de la réhabilitation auditive est une opportunité d'étudier et questionner les mécanismes d'interactions multisensorielles qui jouent un rôle fondamental dans la compréhension de la

parole. Du point de vue clinique, nos études sur la récupération auditive fourniront des informations importantes pour suggérer des pistes en rééducation orthophonique et en technologies assistives.

Dans ce cadre de compensation multimodale, un premier objectif de PlasMody est de comprendre comment les personnes implantées cochléaires exploitent la vision et la combinent efficacement avec l'audition au travers du « filtrage » opéré par l'implant de données aussi complexes que la parole normale, comprenant le niveau segmental et l'information prosodique du contenu linguistique, ainsi que l'identité du locuteur et son état émotionnel.

L'hypothèse est que la synergie entre vision et audition lors de la récupération post-implantation agit comme une boucle de rétroaction positive conduisant à une augmentation de la performance de la modalité auditive. De même, en lien avec la théorie motrice de la perception de la parole, PlasMody permet d'évaluer si les interactions sensori-motrices améliorent les processus de co-perception et production de la parole après implantation cochléaire. PlasMody est multidisciplinaire dans plusieurs aspects de son architecture, elle englobe la physique, la cognition et les technologies assistives de communication avec l'objectif commun de comprendre les mécanismes de substitutions et de compensations induits par la perte de l'audition et permettre aux patients sourds implantés cochléaires d'accéder à la communication orale.

Partenaire GIPSA-lab: Denis Beautemps (CR), Marc Sato (CR), Jean-Luc Schwartz (DR), Xavier Pelorson (DR), Anne Vilain (MCF), Louis Delebecque (Doctorant), Lucie Scarbel (Doctorant).

Partenaire LPNC: Olivier Pascalis (DR).

Partenaire CerCO: Pascal Barone (DR) et Olivier Deguine (PUPH).

III.2 Etude des aspects physique en production de parole

Cette partie vise à étudier et modéliser les phénomènes liés à la physique dans la production de parole. En effet, beaucoup de travaux ont été menés sur la production sous l'angle articulatoire et acoustique, mais peu se sont concentrés sur les aspects physiques. Le paradigme de surdité profonde sans retour auditif s'est trouvé être bien adapté pour observer ces aspects en évitant toute perturbation d'un contrôle involontaire induit par le feedback auditif.

L'étude physique implique une description précise du comportement des cordes vocales et des parois à l'avant du conduit vocal, du débit à la constriction et des interactions de ces différentes composantes. L'objectif visé est de modéliser les phénomènes physiques plutôt que les effets qui en résultent, de façon à limiter les paramètres de contrôle en simulation numérique. L'ensemble de ces travaux est au cœur de la thèse en cours de Louis Delebecque. J'aborderai donc cette partie par des illustrations de la démarche de ce travail dont certains des résultats

sont en voie de publication (soumission actuelle à la revue internationale avec comité de lecture *Journal of the Acoustical Society of America*).

La démarche consiste à l'enregistrement de données aérodynamiques et acoustiques sur des sujets permettant d'émettre des hypothèses sur les comportements observés au regard des modèles théoriques. Ces hypothèses ainsi que la pertinence des modèles théoriques sont ensuite évaluée avec des mesures obtenues sur des maquettes qui sont des répliques simplifiées du système phonatoire humain (Figure III-1). Enfin des simulations numériques sont réalisées et les résultats comparés aux données mesurées sur les participants.

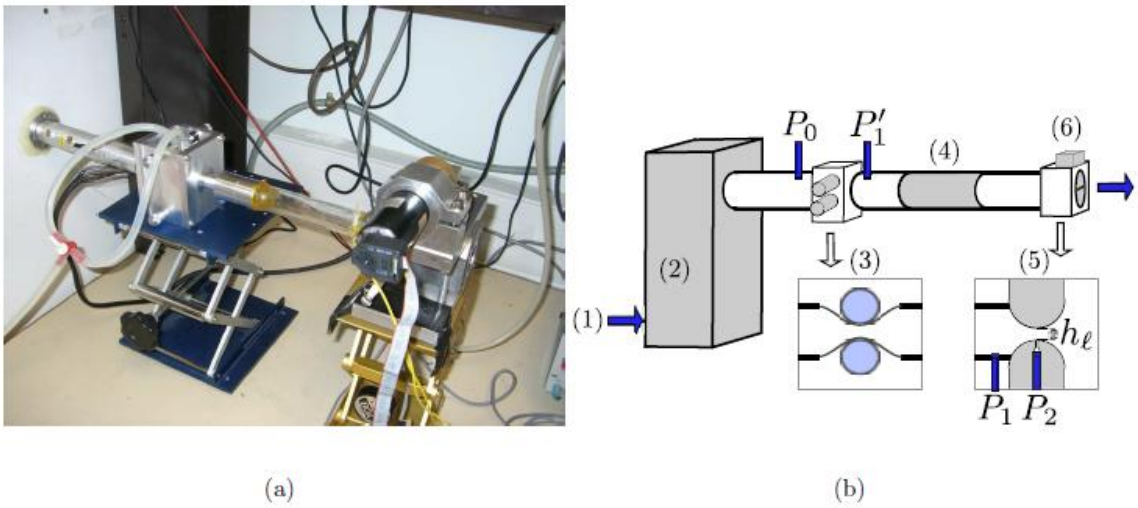


Figure III-1 : (a) : image photo de la maquette du système phonatoire, (b) : diagramme du système phonatoire composé (1) : d'un compresseur, (2) d'un réservoir de pression, (3) : d'une maquette des cordes vocales, (4) : d'un tube rigide en plexiglass (qui peut être remplacé par un tube en latex), (5) : d'une maquette en métal des lèvres et (6) : moteur et capteurs de déplacement des lèvres. Figure extraite de l'article Delebecque et al. soumis à la revue *Journal of the Acoustical Society of America*.

III.2.1- Etude du comportement des cordes vocales

Dans ce cadre expérimental un travail sur les cordes vocales est en cours avec l'analyse de la production d'un sujet sourd post-lingual privé de retour auditif. Les données aérodynamiques, acoustiques et EGG ont été enregistrées par le système EVA (Giovanni et al., 2006) auquel était adjoint un EGG (Figure III-2).

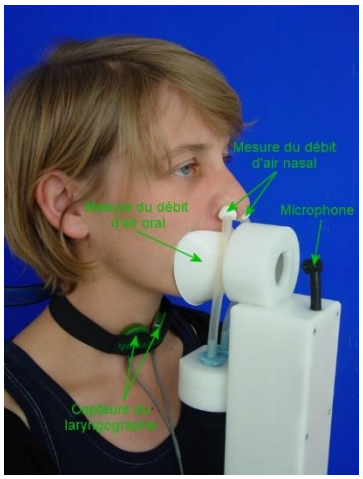


Figure III-2 : EGG couplé au système de mesure aérodynamique et acoustique EVA. Image issue de la page Web de la plateforme BEDEI de GIPSA-lab.

Ce système a permis de faire ressortir pour le sujet sourd dans le cas de la production d'une voyelle [u] tenue avec une forte intensité, une valeur très élevée du F0 (entre 500 et 550 Hz) suivie d'une brusque diminution (à 400 Hz) (Figure III-3) liée à une baisse de la pression acoustique et associée à un changement de mécanisme laryngé comme l'indique les signaux de l'EGG (Figure III-4). Ces observations suggèrent que la pression sous glottique permettant une forte intensité génère un mécanisme de vibration engendrant un F0 élevé. En fin de tenue, à partir de 0,7s, la pression acoustique baisse suggérant une baisse de la pression sous glottique avec pour conséquence un changement de mécanisme de régime vibratoire engendrant une chute du F0. Ces hypothèses restent à être validées sur maquette. Enfin il est à noter que le F0 n'est corrigée ni à la hausse ni à la baisse ce qui aurait pu se réaliser par un contrôle particulier des cordes vocales par feedback auditif, ce qui n'a pas été le cas. Cette observation confirme l'intérêt de l'utilisation du paradigme de surdité pour mettre en évidence les phénomènes de liés à la physique dans le comportement des cordes vocales.

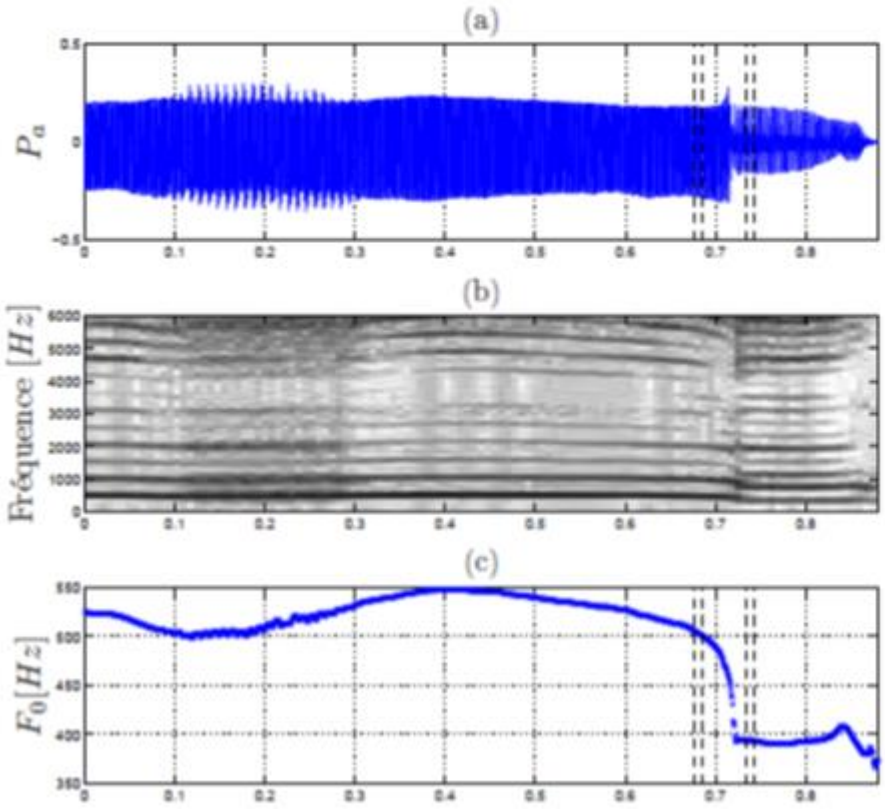


Figure III-3 : Mesures pour la voyelle [u] produite avec une forte intensité par le sujet sourd. (a) : Signal de pression acoustique Pa. (b) : Spectrogramme calculé sur les fenêtres de 10 ms. (c) : Fréquence fondamentale F0 extrait du signal Pa. Les traits pointillés verticaux désignent les agrandissements des signaux EGG présentés en figure III-4.

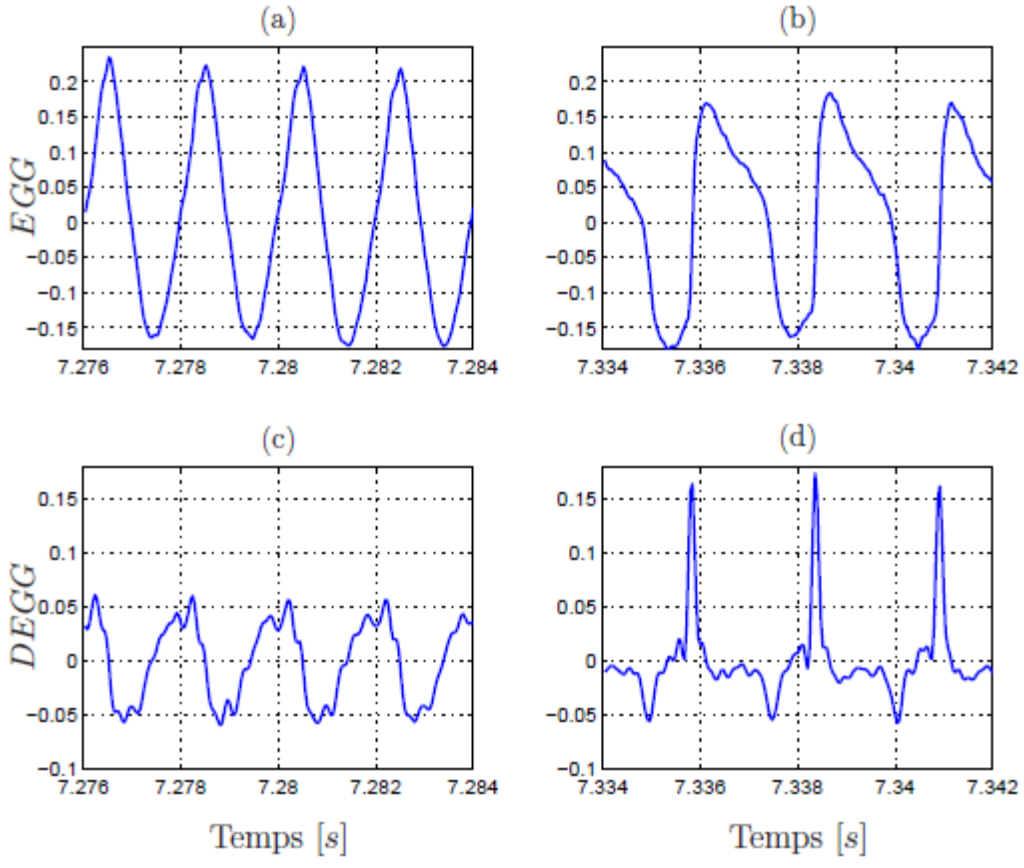


Figure III-4 : Signaux EGG (a, b) et leur première dérivée (c, d) mesurés pour la voyelle [u] produite avec une forte intensité par le sujet sourd, avant (a, c) et après (b, d) le saut de fréquence fondamentale observée.

III.2.2- Etude de l'effet du mouvement des parois du conduit vocal

Cet ensemble expérimental a permis une autre étude sur l'effet du mouvement des parois du conduit vocal sur la production de séquences Voyelle-plosive bilabiale-Voyelle. La pression intra-orale, mesurée par le système EVA, présente une dissymétrie du profil (Figure III-5c) entre la phase d'occlusion (intervalle [0,15s ; 0,2s]) et la phase de détente aux lèvres (intervalle [0,4s ; 0,45s]). L'augmentation plus lente de la pression intra-orale se superpose à la phase d'extinction acoustique de la voyelle (intervalle [0,15s ; 0,2s], Figure III-5b). L'hypothèse proposée dans ce cadre est que la hausse plus lente de la pression intra-orale observée à l'occlusion comparée à la chute à l'ouverture des lèvres pouvait s'expliquer par une simple expansion des joues lors de l'occlusion responsable dans le même temps du maintien de l'oscillation des cordes vocales jusqu'à son arrêt avec l'atteinte de l'expansion limite. L'hypothèse sur l'effet de l'expansion des joues a été validée in vitro sur maquette en comparant les mesures avec tube rigide et tube en latex (Figure III-6) et in vivo sur des sujets contraignants avec les mains le mouvement des joues (Figure III-7). On observe ainsi après la fermeture aux lèvres une augmentation rapide de la pression intra-orale aussi bien dans le cas du tube rigide que dans le cas des joues dont le mouvement est empêché. Inversement, on observe une pente plus lente de l'augmentation de la pression intra-orale aussi bien dans le cas du tube en latex que dans le cas d'un mouvement laissé libre des joues.

L'effet de l'expansion des joues a été simulé numériquement par un système masse-ressort. Rappelons-le, l'objectif visé est de modéliser les phénomènes physiques de façon à limiter les paramètres de contrôle en simulation numérique. La finalité est d'améliorer le réalisme de la simulation du point de vue de son comportement physique et de pouvoir prévoir l'évolution des paramètres physiques qui régissent la production de parole. À notre connaissance, avant que ne débute ce travail dans le contexte de la thèse de Louis Delebecque, l'expansion de la cavité buccale, n'était pas pris en compte dans les modèles physiques de production de parole.

Des simulations ont donc été réalisées pour des séquences [apa]. La figure III-8b fait apparaître la persistance d'un voisement 50 ms après la fermeture complète des lèvres dans le [p] ce qui n'est pas possible dans le cas de la simulation d'un conduit vocal rigide. L'amplitude de l'oscillation mesurée par la dérivée du débit glottique décroit progressivement. Ces résultats de simulation rendent compte des données enregistrées sur l'humain (voir Figure III-5).

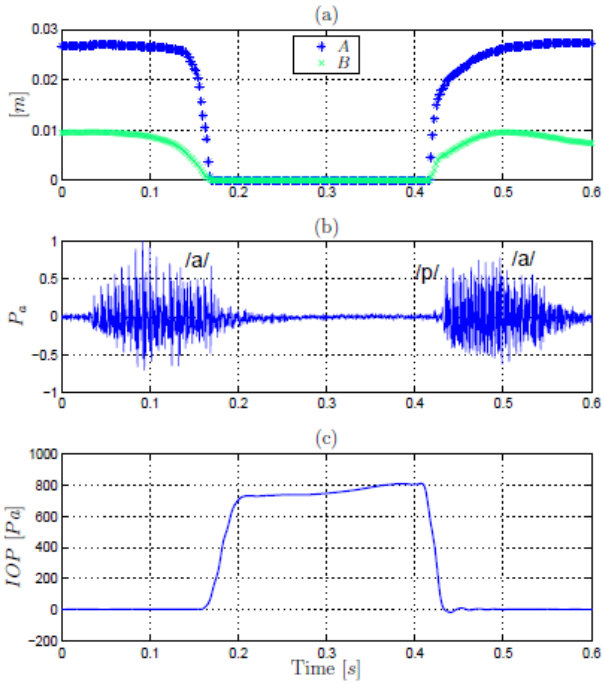


Figure III-5 : Mesures à partir de la production d'une séquence [apa], (a) : paramètres de lèvres (étirement A et aperture B) extraits d'un enregistrement vidéo, (b) : pression acoustique Pa mesurée à 50 cm des lèvres, (c) : pression intra-orale IOP. Figure extraite de l'article Delebecque et al. soumis à la revue *Journal of the Acoustical Society of America*.

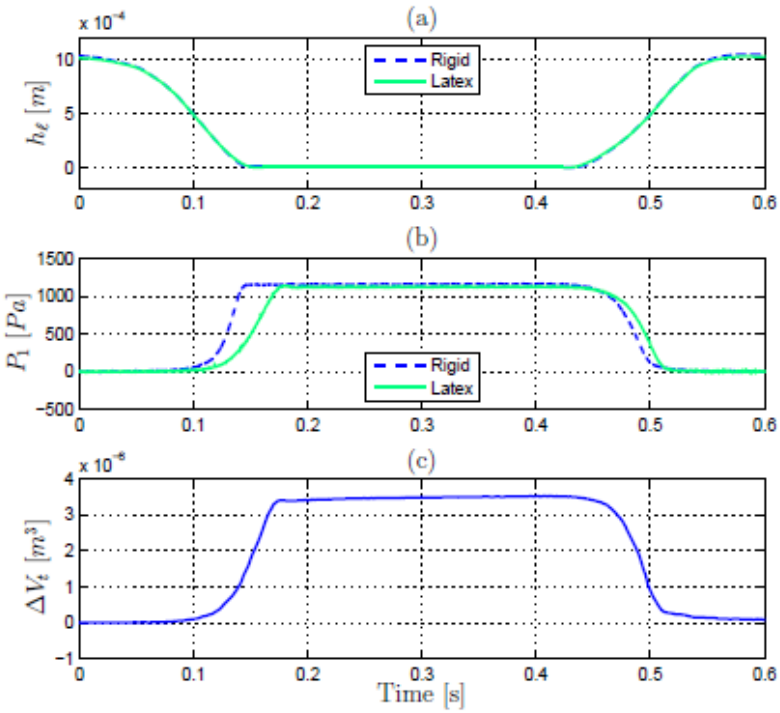


Figure III-6 : Résultats expérimentaux pour l'expansion du tube en latex. (a) : aperture hl aux lèvres dans le cas des configurations rigides et latex du tube. (b) : pression expérimentale P1 mesurée pour les deux configurations. (c) : variation du volume du tube en latex. Figure extraite de l'article Delebecque et al. soumis à la revue *Journal of the Acoustical Society of America.*

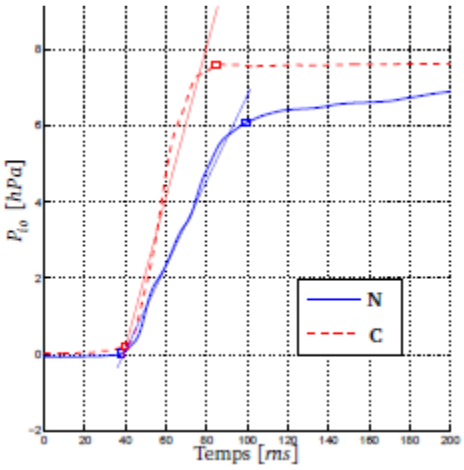


Figure III-7 : Pressions intra-orale mesurées sur une fenêtre de 200 ms autour de l'instant de fermeture aux lèvres pour une séquence [apa] avec deux conditions de production: normale (N) et contrainte (C) pour laquelle le sujet applique ses paumes de mains au niveau des joues pour limiter son mouvement. Figure extraite de Delebecque et et al. (2012).

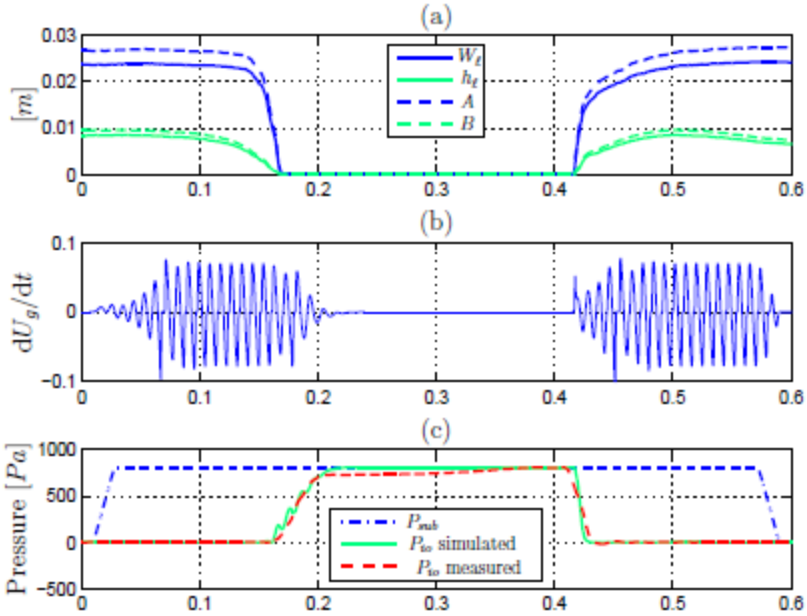


Figure III-8 : (a) Paramètres d'étirement et d'aperture aux lèvres ; (b) Dérivée du débit glottique simulé ; (c) Pression intra-orale simulée (trait vert) et mesurée ~~sur la maquette~~ (trait pointillé rouge). Figure extraite de l'article Delebecque et al. soumis à la revue *Journal of the Acoustical Society of America*.

III.3 Close shadowing et collaboration audiovisuelle

Le shadowing consiste en la répétition de la parole produite par le locuteur au fur et à mesure qu'elle est perçue. Lorsque cette répétition consiste à suivre en ligne la parole de l'autre en la répétant aussi vite que possible au fur et à mesure qu'elle est perçue, on parle alors de « close shadowing » ou « shadowing rapproché ». Ce paradigme nous semble particulièrement

adapté pour comprendre les processus cognitifs en jeu dans le monitoring de sa propre production linguistique par rapport au monitoring de la parole de l'autre. Ces processus de self-monitoring et de monitoring de l'autre semblent particulièrement impliqués en situation d'apprentissage des langues et de remédiation auditive au cours duquel l'apprenant doit gérer, probablement de manière très imbriquée, à la fois le monitoring du tuteur et son propre monitoring. Enfin ce paradigme a l'avantage de tester les traitements cognitifs en perception et production simultanées en restant dans le domaine du comportemental, beaucoup plus accessibles que l'IRMf ou d'autres méthodes d'imagerie cérébrale.

Dans le contexte de la multimodalité, la question de la collaboration audiovisuelle en situation de répétition s'est posée naturellement. Le paradigme en « close shadowing » utilisé pour tester la fonction audiovisuelle a de plus l'intérêt de travailler avec des signaux audio « propres » contrairement aux plans expérimentaux classiques utilisant la dégradation de l'information de la composante audio par du bruit. Une étude de Reisberg et al. (1987), testant l'intelligibilité de parole en langue étrangère, a montré un avantage dans la précision de la réponse d'une présentation audiovisuelle du message par rapport à une présentation seulement auditive. De notre côté, nous avons pu observer un effet positif de la présentation de stimuli audiovisuels (caractérisés par des gestes consonantiques) sur les temps de réaction des sujets avec un gain de 13,4 % à 14,6 % (DEA d'Yvon Le Borgne en 2002, voir aussi Beautemps et al., 2003).

III.4 Relations perceptuo-motrices et implants cochléaires

Si les études sur la perception post-implantation sont légion, celles sur la production de parole sont plus rares, et plus rares encore celles sur les liens entre perception et production, dont on considère de plus en plus qu'ils structurent toute la communication. C'est l'enjeu du travail de thèse en cours de Lucie Scarbel qui vise à étudier la relation perceptuo-motrice chez les sujets sourds porteurs d'un implant cochléaire en réhabilitation de l'audition. Dans ce cadre, trois expériences ont été élaborées avec pour objectif d'observer dans quelles mesures ces interactions entre systèmes sensoriels et moteur s'établissent chez des patients sourds post-linguaux après implantation cochléaire et, notamment, de déterminer la possible implication du système moteur lors de tâches de perception et de production de la parole à différents stades postopératoires.

La première expérience est une tâche de catégorisation rapide de consones à partir d'une expérience en « close shadowing » visant à répondre le plus rapidement à des stimuli syllabiques (Figure III-9).

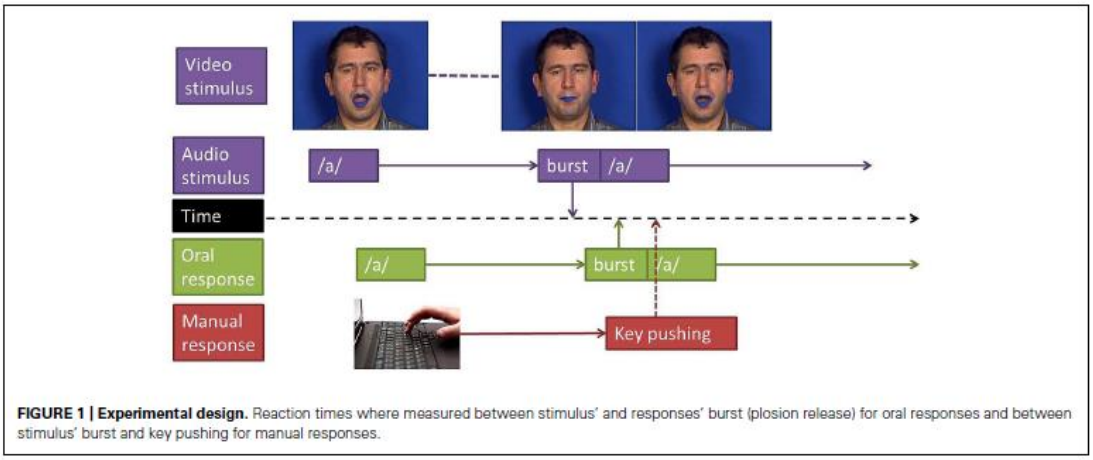


Figure III-9 : Schéma illustrant l'expérience en close-shadowing à partir de la présentation de stimuli audio-visuels. Réponse orale et réponse clavier pour la comparaison des processus de traitement. Figure issue de Scarbel et al., 2014.

La seconde expérience comporte une tâche de production et une tâche de perception des voyelles orales du français et vise à mettre en évidence des idiosyncrasies couplées entre perception et production. Enfin, la troisième expérience a pour but de déterminer les capacités d'imitation volontaire et de convergence phonétique chez ces patients. Les résultats sur une population de sujets entendants en « close-shadowing » de consonnes montrent la capacité des sujets à répondre rapidement et à exploiter l'information visuelle lorsqu'elle est disponible pour améliorer les délais de réponse (voir publication de Scarbel et al., 2014 à la revue internationale avec comité de lecture « Frontiers in Psychology »), plus rapidement en réponse orale qu'en réponse manuelle, suggérant une prédiction motrice dans le cas de la réponse orale pour les deux conditions de présentation des stimuli, c'est-à-dire en audio et audio-visuelle.

Les tous premiers résultats chez des sujets sourds montrent des performances similaires en close shadowing avec signaux audio «propres » à celles des sujets entendants en close-shadowing mais avec signaux audio bruités avec dans la condition audiovisuelle à la fois des temps de répétition plus courts et une meilleure précision en catégorisation. Ces résultats semblent compatibles avec l'hypothèse motrice d'un traitement de la réponse orale au niveau de la commande motrice.

Contexte contractuel : projet Région Rhône-Alpes CIBLE, projet ANR PLASMODY ;

Listes des intervenants : Denis Beautemps (CR), Xavier Pelorson (DR), Jean-Luc Schwartz (DR), Marc Sato (CR), Lucie Scarbel (Doctorant), Louis Delebecque (Doctorant), Xavier Laval (IE), Christophe Savariaux (IR), Yvon Leborgne (DEA), Balbine Maillou (DEA);

Chapitre IV : Projet de recherche - Adaptabilité en parole, variabilité et plasticité

Introduction

Mon programme de recherche est un projet en production de parole par ses aspects variabilité et plasticité. Je souhaite le mener à la lumière de questionnements en surdité et oralité, d'analyses des relations entre modalités et de modélisation fonctionnelle dans le contexte du traitement automatique de la parole.

En effet, les handicaps auditifs et la surdité sont un problème d'une importance croissante dans nos sociétés touchées par le vieillissement, au point d'apparaître comme un problème aigu de santé publique, et un enjeu socio-économique majeur. De nos jours, on estime en France de 5 à 6 millions les personnes malentendantes ou sourdes profondes ou devenues sourdes. Pour la population des personnes devenues sourdes progressivement, la lecture labiale appelée communément lecture sur les lèvres est devenue au fur et à mesure de la perte en capacités auditives la modalité principale pour percevoir la parole, tout comme pour la population des enfants nés sourds profonds. La lecture labiale est en effet une compétence naturelle que possède tout à chacun, avec cependant une grande variabilité d'une personne à l'autre, les lecteurs labiaux les plus performants se trouvant parmi la population des personnes sourdes. Cependant les meilleurs scores en perception de mots atteignent seulement 43,6 % comme cela est rapporté par la littérature (Auer & Bernstein, 2007 ; Bernstein et al., 2010) du fait de l'existence de sosies labiaux (des phonèmes différents pouvant avoir des formes labiales très similaires). Sans connaissance du contexte sémantique ou d'informations complémentaires auditives, codées gestuellement, ou de texte, il peut être difficile de percevoir la parole confortablement.

Plus largement, dans le domaine du traitement automatique de la parole dans sa dimension visuelle, des problématiques se posent en termes de segmentation de séquences d'unités de parole, de repérage des régions d'intérêt, de définition de leurs paramètres descripteurs, de fusion et/ou de classification de paramètres. Ces problématiques font l'objet d'un premier axe de recherche intitulé «Articulation labiale complétée : du mapping à la reconnaissance ».

Durant les trois dernières décades, des progrès très significatifs ont été faits sur les techniques d'implantation cochléaire, permettant aux sourds profonds de récupérer des capacités auditives très significatives. Néanmoins ces progrès restent tributaires des stratégies de rééducation et limités par l'inévitable réduction de l'information acoustique transmise par

l'implant, et par la plus ou moins longue déprivation d'information auditive vécue par les individus. Parallèlement, nos anciens souffrent de plus en plus de mauvaise audition, et les appareillages les assistent très imparfaitement. Dans ce contexte de réhabilitation, quelle place pour la modalité visuelle ? En particulier en situation de bruit ambiant, quel apport dans les traitements réalisés par l'implant ? Ces questionnements sont l'objet d'un second axe de recherche intitulé «Implant cochléaire et fonctionnalités audio-visuelles».

Dans un cercle plus large de recherches sur la variété des moyens de la communication, de nouvelles collaborations sont en cours de développement autour du thème de la parole hyper-hypo articulée, visant à mieux comprendre comment les locuteurs adaptent leur clarté de production en fonction du niveau d'information nécessaire à leur auditeur pour percevoir le message et de la recherche d'adhésion dans le discours codé. C'est ici une approche de type analyse de corpus qui sera utilisée, avec des corpus constitués de matériaux audio-visuels et/ou issus de systèmes à capture du mouvement existants ou à enregistrer pour différents types de production et d'interaction, de tests perceptifs d'évaluation, d'étiquetage de données, d'extractions de paramètres quantitatifs ou qualitatifs et de leurs analyses. Cette partie fait l'objet d'un troisième axe de recherche intitulé « Rhébotique : de la syllabe au discours» qui est extrait d'un projet soumis au dernier appel AGIR-PEPS de l'université, dont je suis l'instigateur, le rédacteur principal et le porteur.

L'ensemble de ce programme de recherche s'inscrit dans la politique de recherche du laboratoire GIPSA-lab. Il est une contribution aux enjeux de prospectives du DPC (voir Figure-IV-1) que j'ai eu le plaisir d'animer en Octobre/Novembre 2014 en tant que responsable du département. En particulier, le projet contribue aux actions «Parole, langage et relations sensori-motrices » pour sa dimension multimodale et les problématiques d'apprentissage automatique, « Troubles de la parole, voix et langage » pour les intérêts sur les déficiences sensori-motrices et les questions de plasticité liés à la surdité et à l'hyper/hypo et à des « nouveaux paradigmes expérimentaux » par ses plans expérimentaux originaux.

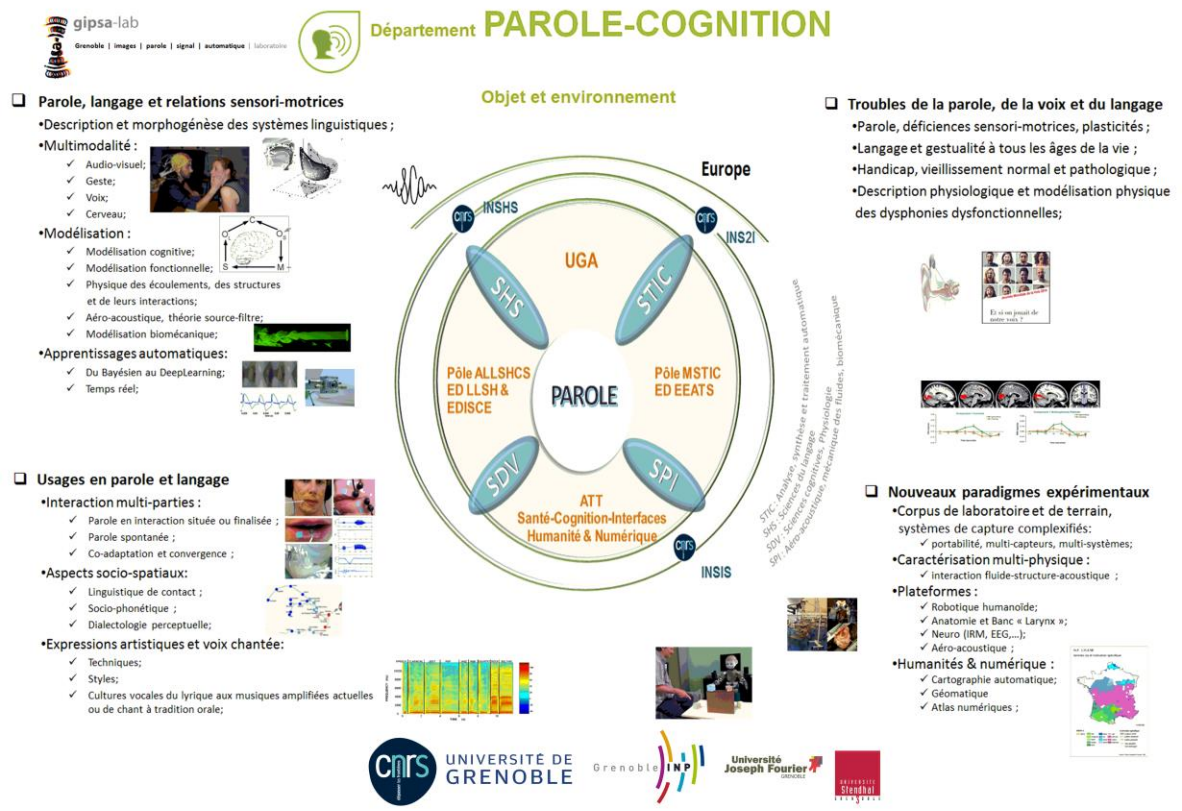


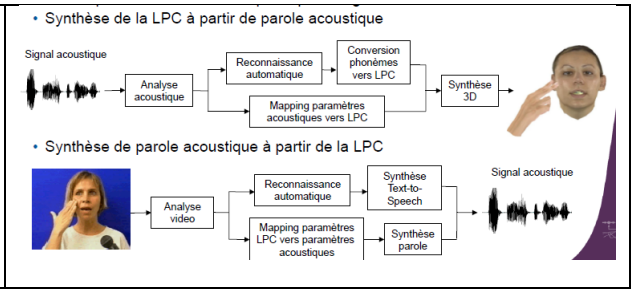
Figure-IV-1 : Enjeux de prospective au sein du DPC de GIPSA-lab

Les enjeux de prospective en parole s'inscrivent dans le contexte local Grenoblois de l'Université-Grenoble-Alpes en construction au moment de la rédaction de ce manuscrit (UGA : regroupement des trois universités fusionnées de Grenoble avec Grenoble-INP et le CEA,…). Il vise à se développer principalement au sein de deux des six pôles du site, les pôles MSTIC et ALLSHCS ainsi que sur deux axes transversaux transformant « Santé, Cognition et leurs interfaces » et « Humanités & Numérique ». Il se nourrit de quatre grands domaines disciplinaires relevant des STIC par l'analyse/synthèse et le traitement automatique du signal de parole, des SHS par une compétence importante en « Sciences du Langage », des SPI pour son expertise en aéro-acoustique, mécanique des fluides et biomécanique, et des SDV par son ouverture sur les sciences-cognitives et la physiologie. Cette alliance entre disciplines, données de terrain, de laboratoire, et modèles est rare et unique en France dans un contexte de laboratoire très majoritairement INS2I. Elle est très bien identifiée à l'échelon national, non seulement par l'Institut INS2I du CNRS mais aussi par l'INSHS et l'INSIS. Elle a déjà porté de nombreux fruits en permettant de contractualiser non seulement auprès de l'ANR mais aussi récemment au niveau de l'Europe ce qui n'était pas le cas dans le contrat quinquennal précédent.

A. Articulation labiale complétée : du mapping à la reconnaissance

Le contexte de cet axe est l'analyse/synthèse audio-visuelle en parole dans un cadre vidéo (Figure-IV-2). L'objectif est d'explorer et modéliser les relations entre les espaces audio et visuels de la parole. Nous nous placerons sur une échelle allant de bas niveaux d'interfaçage « signal » pour le mapping entre paramètres audio et visuels à des interfaces nécessitant des traitements de haut niveau en reconnaissance automatique pour le décodage phonétique en passant par le niveau combinant mapping et classification.

Figure-IV-2 : deux systèmes d'analyse / synthèse en parole avec LPC (version Française du Cued Speech) pilotés en haut par l'audio et en bas par l'image. Dans les deux situations, deux méthodes sont possibles : l'une utilisant la reconnaissance automatique, l'autre le mapping.



Ce programme, s'appuiera sur de grands corpus audio-visuels complets (phonèmes, syllabes, mots, phrases) intégrant composantes gestuelles porteuse d'informations linguistique relevant de différents niveaux (phonétique pour le Cued Speech, lexical, sémantique voire pragmatique pour les gestes co-verbaux). Ils seront produits oralement par plusieurs sujets enregistrés en vidéo.

Les paramètres considérés sont ceux porteurs des traces langagières. Typiquement, l'espace audio peut être défini par des paramètres issus des composantes spectrales du signal acoustique. L'espace visuel peut être décrit par des paramètres extraits de la région d'intérêt des lèvres et caractéristiques de l'articulation labiale (paramètres de formes ou d'apparence : nous avons pu montrer que certains des paramètres caractéristiques de ces deux approches sont en relation linéaire) complétés par des paramètres de la région d'intérêt de gestes manuels codant ou co-verbaux porteurs d'information linguistique et éventuellement de paramètres supplémentaires issus d'autres régions d'intérêt intervenant dans l'interaction.

A partir de ces données, un premier volet de travaux d'ampleur en traitement d'image sera mené. Il s'agira de traiter des problématiques de segmentation de séquences d'images et d'objets (visage, les régions d'intérêt ROI labiales, de main, …) à l'intérieur des images ainsi que leur mise en correspondance entre images pour prendre en compte des situations de désynchronisation. Des méthodes par apprentissage, des outils de reconnaissance d'objets utilisés en Traitement d'Image et en vision par ordinateur de façon plus générale pourront être éprouvés pour le repérage des ROI et le suivi d'objets. Ils pourront être couplés à des méthodes de traitement des collisions d'objets comme celles rencontrées en Cued Speech avec la main et le visage. Les paramètres de description au sens de l'information contenue dans les ROI seront étudiés en fonction de leur dimensionnalité et de leur robustesse aux conditions d'enregistrement (éclairage, situation ou non de face à face,…).

Les résultats de ce premier volet pourront alimenter deux autres volets de recherche. Le premier se concentrera sur la modélisation du mapping image vers audio et inversement audio vers vidéo. Les résultats obtenus précédemment dans le cas des voyelles (Figue-3) pourront servir de fondement à une extension aux autres unités de parole. Les méthodes de mapping

seront étudiées. Elles pourront s'appuyer sur des méthodes d'apprentissage automatique avec les outils de Machine Learning et de logique floue.

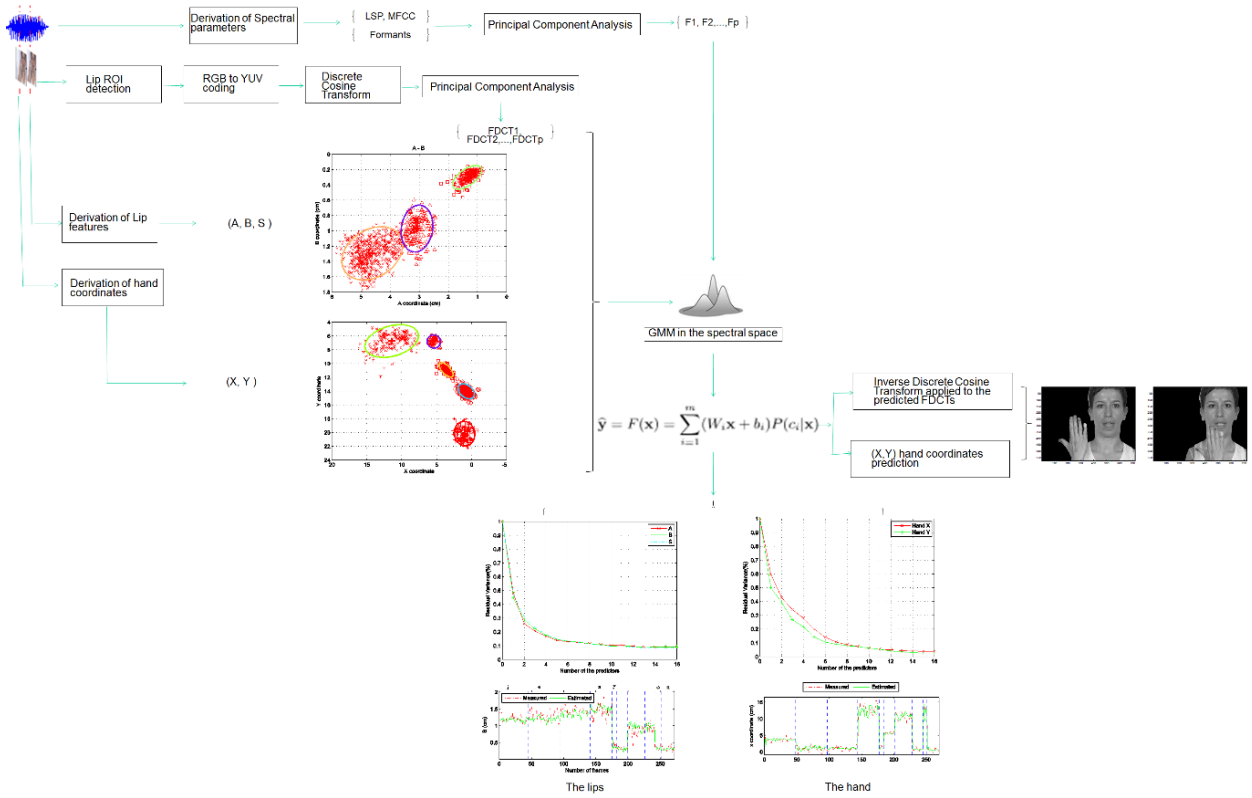


Figure-IV-3 : Mapping Audio-Visuel.
Analyse/synthèse appliquée à la main et aux lèvres pour le cas des voyelles.

Le dernier volet aura pour objectif le décodage phonétique à partir de la reconnaissance de l'articulation labiale y compris complétée dans le contexte de la conversion du visuel vers l'audio. Ce programme répond à des questions en reconnaissance labiale, domaine qui connait un regain d'intérêt, non seulement sur le plan scientifique mais aussi dans le contexte socio-économique en réhabilitation mais aussi en surveillance. Ainsi les derniers systèmes en reconnaissance labiale permettent d'atteindre 76 % (travaux récents du professeur Ahmad Hassanat de l'Université de Mu'tah en Jordanie), performances comparables que nous avions de notre côté déjà atteints dès 2010. Vous noterez que ce résultat dépasse déjà largement les performances rencontrées chez les meilleurs lecteurs labiaux. Les performances supérieures à 76% peuvent être obtenues en rajoutant des traitements supplémentaires pouvant faire appel à des modèles de langage (qui modélisent le contexte linguistique mais qui limitent la généralisation) ou par le moyen de traitements supplémentaires qui intègrent des informations linguistiques complémentaires comme celles issues des gestes du Cued Speech par exemple.

Nous avons ainsi pu atteindre des performances de 94,9% avec le Cued Speech sur un vocabulaire de mots isolés (Heracleous, P., Beautemps, D. & Aboutabit, N., 2010), performances comparables aux résultats en reconnaissance automatique à partir du seul son de parole. Il s'agira donc dans ce contexte d'étendre nos travaux précédents aux autres unités de parole. En particulier, les méthodes de fusion de paramètres seront étudiés. Les approches en fusion directe de paramètres, ou en fusion de décisions ou encore hybrides, selon le classement proposé par Potamianos et al., (2012) pourront être évalués.

Enfin, dans cet axe, l'ensemble des paramètres audio manipulés sont compatibles avec ceux utilisés en synthèse audio incrémentale. Les résultats de cet axe pourront donc naturellement s'ancrer sur les problématiques incrémentales.

B. Implant cochléaire et fonctionnalités audio-visuelles

L'implant cochléaire est un appareillage composé d'une partie externe contenant un microphone qui capte le son, un processeur qui code l'information spectrale en signaux numériques, une antenne magnétique en charge de transmettre les signaux numériques à la partie interne, le récepteur qui est la partie interne de l'appareillage, composé d'un ensemble d'électrodes posées sur la cochlée et dont la fonction est de stimuler les terminaisons nerveuses.

Le processus d'audition avec un implant cochléaire diffère du processus d'audition naturel. Dans le cas de l'audition naturelle, les vibrations sonores sont transmises aux cellules ciliées à l'intérieur de la cochlée. Un implant cochléaire utilise une impulsion électrique via les électrodes pour stimuler plus précisément les cellules du ganglion spiral. Les cellules du ganglion spiral sont situées dans une zone différente de la cochlée et reliées aux cellules ciliées. La plupart des cellules du ganglion spiral se trouve dans une zone précise, la « zone auditive ». C'est dans cette zone que la réponse à la stimulation électrique générée par l'implant cochléaire est la plus forte. La zone auditive n'est pas très étendue dans la cochlée. Le concept consiste donc à fournir une stimulation électrique aux cellules du ganglion spiral situées dans la zone auditive pour obtenir des performances auditives optimales et réduire la profondeur d'insertion afin de minimiser les risques liés à la stimulation apicale ou les traumatismes liés à l'insertion.

L'implant cochléaire est indiqué aux personnes ayant une surdité sévère ou totale et aux personnes devenues sourdes. Enfin, c'est une chirurgie envisagée lorsque les appareils auditifs classiques ne permettent pas d'amélioration, notamment quand la perte auditive est principalement due à la transmission.

Dans le contexte de situation de parole avec bruit ou plus largement avec des sons complexes (sources multiples, sons apériodiques,…), les performances de l'implant sont très dégradées. Une étude récente avec le test *FrMatrix* (Jansen et al., 2012) en bruit adaptatif menée au CHU sur des implants équipées des plus récentes technologies indique un écart de 6dB en rapport signal sur bruit dans les performances (pour les meilleures) comparées à celles de sujets entendants (communication personnelle du Professeur Sébastien Schmerber, PUPH, chirurgien au service ORL du CHU de Grenoble). Cet écart de 6 dB en défaveur de l'implant est considérable pour la perception de la parole. Il reste donc une marge de progression qui renvoie à des enjeux de recherche en traitement automatique. Ils se déclinent en termes (i) de localisation de la source de parole, (ii) de stratégie de codage –la dynamique disponible n'étant que de 30 dB-, (iii) de réglages du gain des électrodes (iv) de mesure de la charge cognitive.

L'apport de la dimension visuelle est à étudier. En effet, à notre connaissance aucun système n'intègre l'information visuelle qui permettrait pourtant des améliorations dans les solutions de localisation et de débruitage, en particulier dans les cas d'implant mono-oral. Pour le très jeune enfant, où tout est à apprendre, un implant augmenté de fonctionnalités visuelles permettrait à la fois de rehausser l'audio lorsque celui-ci fixerait son interlocuteur en face à face et renforcerait son intérêt pour les mouvements des lèvres de son interlocuteur.

C. Rhébotique : de la syllabe au discours

Le projet Rhébotique porte sur la parole multimodale à travers une approche très originale, aussi bien par l'objet de la recherche que par la collaboration entre deux équipes apparemment très éloignées mais unies par l'intérêt pour la production orale. Le mot-valise Rhébotique marie en effet une dimension Rhétorique (équipe RARE) et un horizon en robotique humanoïde (GIPSA-lab).

Le point de départ de la réflexion et du rapprochement a été l'intérêt commun des deux unités de recherche pour la production de la voix et de la parole en situation de discours institutionnelle et de parole travaillée. Une thèse sur cette problématique vient d'être soutenue au sein de GIPSA-lab, sur la perception du charisme en politique (thèse de R. Signorello, 2014, qui a donné lieu à un article dans *Le Monde*).

Pour l'équipe RARE, voici comment Francis Goyet Professeur à l'Université Stendhal décrit le contexte de l'implication dans le projet : « Pour RARE (Rhétorique de l'Antiquité à la Révolution), l'objectif est de tester un certain nombre d'hypothèses : corrélation entre effet de saillance vocale et moments d'amplification, au sens rhétorique de ce mot ; importance du moment d'un discours pour définir le type de ton utilisé ; variation continuelle du ton. L'intérêt

pour la voix et les gestes est ainsi l'aboutissement d'une réflexion sur la composition d'ensemble des discours, grâce à un corpus ancien qui montre comment les professeurs de rhétorique analysaient précisément et pas à pas un discours donné (réel ou fictif). RARE a ainsi développé une expertise très opératoire en termes de codes rhétoriques, anciens mais aussi modernes».

Pour GIPSA-lab, l'objectif est d'explorer les relations voix-gestes-parole dans le contexte de la variabilité de la parole générée par l'hyper-hypo articulation. La théorie hyper-hypo de Lindblom (1990) selon laquelle les locuteurs adaptent leur clarté de production en fonction du niveau d'information nécessaire pour que l'auditeur perçoive le message confortablement, a donné lieu à nombre de travaux, mais qui n'ont eu trait qu'au monitoring audio de la parole.

Peu de travaux se sont intéressés aux autres modalités de la parole : la modalité visuelle, le geste codeur et/ou co-verbal, les aspects liés à la voix. Le GIPSA-lab a déjà des travaux sur les corrélats visuels (M. Dohen), les gestes co-verbaux (thèse de B. Roustan, 2012), les travaux sur l'articulation labiale (Beautemps et al., 1999) ou la parole Lombard (Garnier et al., 2010). Enfin, pas ou peu de travaux se sont intéressés aux inter-relations entre l'ensemble de ces modalités dans le contexte hyper/hypo.

Le projet vise donc à étendre le cadre théorique de Lindblom aux composantes multimodales de la parole en traitant l'hyper/hypo à partir de deux objectifs de communication qui constituent une véritable originalité du projet: l'intelligibilité optimale du signal produit et la recherche d'adhésion dans le discours codé (de la rhétorique en situation institutionnelle au code switching en contexte bilingue). « Codé », parce qu'il est sujet à des conventions sociales, des règles d'acceptabilité ou de connivence, et donc à des phénomènes de convergence ou au contraire de divergence dans l'interaction.

Il s'agit en résumé d'un projet pluridisciplinaire sur un objet d'étude en parole adaptative qui se situe à l'interface STIC/SHS. Le projet est composé d'actions conjointes d'anthropologie linguistique et de linguistique de terrain, de rhétorique dans les productions de parole instituées (discours officiels, professionnels, artistiques) dans le cadre SHS en lien avec les disciplines expérimentales du domaine de la parole et du traitement du signal dans le cadre STIC. L'alliance des deux unités dans ce projet est donc une opportunité d'une extension à la multimodalité avec un éclairage nouveau issu du croisement des disciplines SHS et STIC. Enfin, je tiens à préciser que la rédaction de ce projet est le fruit de plusieurs discussions avec Francis Goyet et Christine Noille de l'équipe RARE, Professeurs en Rhétorique à l'Université Stendhal, et de collègues du laboratoire GIPSA-lab, que je citerai en particulier, et dans l'ordre

alphabétique, Elisabetta Carpitelli, Maëva Garnier, Giovanni Depau, Marion Dohen, Nathalie Henrich, Jean-Luc Schwartz, Nathalie Vallée et Coriandre Vilain entre autres collègues.

C.1. Méthodologie

La méthode utilisée dans ce projet s'appuie sur l'analyse de corpus constitués de matériaux audio-visuels et/ou issus de systèmes à capture du mouvement existants ou à enregistrer pour différents types de production et d'interaction, de tests perceptifs d'évaluation, d'étiquetage de données, d'extractions de paramètres quantitatifs ou qualitatifs et de leurs analyses. Les paramètres qui caractérisent ces productions relèvent du niveau segmental à l'échelle de la syllabe ou du syntagme et de niveau supra-segmental à l'échelle du discours, auxquels s'ajoutent des paramètres caractéristiques de l'effort vocal, de la réalisation acoustique, de l'articulation labiale, des paramètres décrivant le geste codé ou le geste co-verbal. Ce projet est un projet de données de production de parole et d'analyse de traits. Le programme scientifique s'organise autour d'un premier axe sur « l'empathie, l'adhésion, l'interaction en discours » et d'un second axe intitulé « intelligibilité, multimodalité et complémentarité en parole » qui partageront des méthodes, outils et des questionnements. Dans la suite, à toute fin de compréhension des enjeux de ce projet, je présente une version synthétique de l'axe 1 sous forme de deux actions principales que j'ai proposées à la lumière des discussions menées mais qui ne seront pas au centre de mon travail même si j'apporterai mon expertise sur la constitution de corpus et les aspects d'analyse du signal (extraction automatique et analyse de paramètres segmentaux et supra-segmentaux). Je détaillerai par contre davantage le second axe, plus au cœur de ma prospective personnelle.

C.2. Axe 1 - Empathie, adhésion, interaction en discours

Dans cette partie, nous traiterons des corpus où l'adhésion et/ou l'empathie est/sont recherchée(s) dans le discours en contexte d'interaction sociale.

Nous concrétiserons le travail dans cet axe par deux types de travaux s'appuyant sur l'analyse vidéo : i) de discours préparés tels que des discours écrits pour l'oral et pertinents du point de vue de l'archéologie des codes rhétoriques et dont on dispose de l'enregistrement vidéo dans les ressources de l'INA et ii) de discours spontanés en bilinguisme utilisant le code switching en contexte d'interaction sociale (voir par exemple Depau, 2012). Les analyses en rhétoriques seront croisées par l'analyse de la production orale constituée de l'étiquetage des différentes séquences, l'extraction de paramètres typiques de la production de la parole et de la voix (pitch, formants, durées des pauses, unités de parole –leurs durées, leur frontières, …),

l'étude de leurs alignements en relation avec les choix lexicaux, les dimensions sémantique et syntaxique, et les tours de parole le cas échéant. De plus l'adhésion pourra être évaluée par des tests perceptifs sur une population cible avec l'aide de questionnaires et l'apport de la multimodalité pourra être évalué en comparant les résultats obtenus en mode audio seul et en mode audio-visuel.

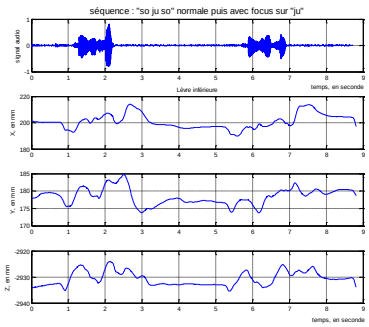
Les résultats de ces analyses permettront de préciser les hypothèses et d'en déduire un corpus à enregistrer en multimodalité dans le cadre de l'axe 2.

C.3. Axe 2- Intelligibilité, multimodalité et complémentarité en parole

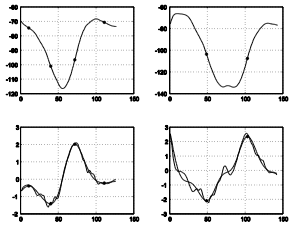
Dans cette partie, on se situera sur une dimension allant du style de production « tout audio et pas de visuel » (cas de la confidentialité et/ou du ventriloquisme) au style de « purement visuel sans audio ». Des corpus audio-visuels seront enregistrés pour les différents styles de parole considérés, dans le cadre d'un continuum sur la dimension audio-visuelle. Ils pourront être induits au travers d'une interaction contrôlée avec l'aide d'un paradigme en « Magicien d'Oz », de focus contrastif ou de code switching se contraposant à une situation de non focalisation ou de neutralisation de l'effet d'alternance. En particulier on induira différents degrés d'effort vocal et différentes intensités de parole. D'autre part, les résultats de l'axe 1 permettront de travailler sur un corpus de discours restitué oralement et enregistré en multimodalité, ce qui constituera une seconde action de l'axe 2.

Les sujets seront enregistrés dans une chambre anéchoïque en vidéo (son + image) couplé à un système d'électroglottographie (EGG), et d'un système optotrak de capture du mouvement (Figue-IV-4). Les lèvres des sujets seront maquillées en bleu afin de faciliter l'extraction automatique de leur contour à partir de l'enregistrement vidéo. Un système EGG pourra être posé sur la peau du coup pour enregistrer l'activité des cordes vocale. Des diodes optotrak seront placées sur les articulations des membres supérieurs, le dos de la main et l'extrémité des doigts et permettront de suivre en 3 dimensions les gestes brachiaux-manuels. Une diode sera placée sur le front afin d'enregistrer par ce même système les mouvements de tête. Enfin ces systèmes de capture et de vidéo seront synchronisés afin d'assurer la cohérence des mesures. Ce système complet est une originalité et constituera un des résultats en expérimentation du projet.

Figue-IV-4
Dans la chambre anéchoïque de la plateforme BEDEI de GIPSA-lab, sujet équipé des diodes à émission Infra rouges du système optotrak



Figue-IV-5
Séquence [sojuso] en condition normale de production puis avec focus sur la syllabe cible [ju]. De haut en bas le signal audio, les coordonnées X, Y et Z de la diode située sur la lèvre inférieure.



Figue-IV-6
Séquence [sojuso] avec geste codeur en Cued Speech (i) en haut tracé de la première composante principale d'une des diodes du dos de la main, (ii) en bas tracé de la vitesse et de la vitesse filtrée pour la condition normale (à gauche) et la condition de focus (à droite). On peut ainsi observer une tenue plus longue de la syllabe intermédiaire [ju].

A partir des paramètres extraits des composantes multimodales (paramètres audio – intensité, fréquence fondamentale, …–, paramètres extraits de la région d'intérêt des lèvres, coordonnées 3D, paramètres laryngés d'effort vocal,…), les décours temporels (les signaux) et/ou leurs éventuelles combinaisons pourront être analysés (voir sur la Figue-IV-5 un exemple en production de parole Cued Speech de signaux issus du système optotrak et en Figue-IV-6 un exemple de combinaisons en composantes principales). Il pourra être identifié des patrons en situation de parole normale (issus de l'analyse). Nous explorerons comment ceux-ci résistent ou se renforcent en situations de parole hyper-hypo. Les collaborations mises en jeu au travers des relations temporelles entre paramètres seront analysées ainsi que les stratégies de compensation développées en situation d'effacement d'une des composantes multimodales. En plus de ces questionnements, une attention particulière sera apportée à la question de la syllabe et du phénomène de re-syllabification qui peut être mis en œuvre dans ce contexte. Enfin ces analyses seront croisées avec celles issues de l'archéologie des codes rhétoriques et plus précisément en rhétorique de l'amplification et de l'atténuation (Macé, 2014).

La population étudiée sera composée de participants n'ayant pas de pathologie de la parole et de la voix. Cette population sera complétée par l'étude de cas particuliers : les ventriloques, pour lesquels la production de parole doit être non visible ; la parole produite par des personnes aveugles de naissance, qui utilisent moins l'articulation labiale dans leur production que les personnes voyantes et les personnes sourdes de naissance et oralistes qui utilisent la modalité labiale comme modalité principale en parole. Ces populations qui produisent naturellement la parole perturbée seront comparées avec la population de participants dont la production est contrainte.

Les résultats du projet pourront s'évaluer en termes de nouvelles données multimodales numériques mises à disposition de la communauté scientifique, nouveaux systèmes expérimentaux et nouveaux paradigmes. Des publications dans les conférences et dans les journaux phares des domaines concernés ainsi que l'organisation d'une journée d'études sont aussi des attendus du projet. Ce projet pluri-disciplinaire sur la thématique interdisciplinaire en parole devrait permettre en plus de capitaliser de nouvelles données en adaptabilité, d'obtenir des avancées et des éclairages nouveaux en retour sur les différents domaines disciplinaires du consortium. Enfin ce projet possède des enjeux en technologie des systèmes de parole, plus particulièrement en synthèse multimodale de la parole et robotique humanoïde, avec l'alimentation de ces systèmes en connaissances supplémentaires en adaptabilité.

CONCLUSION

Cet exercice m'a permis de rassembler et d'ordonner les travaux que j'ai pu diriger (thèses, stages de recherches, projets contractualisés) dans le contexte de ma recherche et présenter des éléments de prospective. Le bilan est décliné en trois chapitres qui présentent des problématiques en parole augmentée dans l'ordre de la chronologie de leur réflexion. Tout d'abord c'est l'étude de la production du code LPC, un augment manuel des lèvres, qui a constitué une contribution inédite dans le domaine du Cued Speech. Elle a ainsi permis de comprendre comment le geste accompagne la parole en mettant en évidence une avance moyenne de l'ordre de 200 ms du geste LPC sur celui des lèvres, avance exploitée par les personnes sourdes dans leur perception de la parole. Les résultats sur la coordination main-lèvres ont permis d'alimenter le premier système de synthèse audiovisuelle de parole augmenté de code LPC intégrant des données de production. De nouveaux champs de recherche en perception se sont nourris de ce travail pionnier en production.

La fusion de l'information de main et celle disponible aux lèvres, tenant compte de leur décalage temporel, a été un des enjeux principaux dans les traitements automatiques de la reconnaissance labiale augmentée par le code LPC. Les meilleurs scores obtenus à partir de ces informations purement visuelles rivalisent ceux de la reconnaissance audio de la parole. L'étude du mapping entre paramètres audio et paramètres visuels caractéristiques du mouvement des lèvres et de main a fait ressortir la possibilité avec la modélisation par GMM de prédire la région d'intérêt des lèvres et la position à atteindre par la main sur le côté du visage avec de bons scores de réplication. Cet ensemble de travaux a permis d'alimenter le projet TELMA labellisé par l'ANR et visant une interface de conversion automatique entre son de parole et parole avec LPC. Ils sont aussi source d'enjeux de prospective en traitements automatiques visant à généraliser les approches à des unités de parole plus complexes et à des conditions de production moins contraintes.

L'ensemble des résultats sur la production du Cued Speech a permis d'éclairer comment ce système est ancré sur la parole. Ce système purement visuel avait été conçu pour compléter l'information issue de la lecture labiale et permettre aux personnes sourdes oralistes une perception complète de la parole. Ce système d'une certaine manière renvoie à l'audition, ce qui est d'autant plus intéressant avec les développements récents de l'implant cochléaire et la réhabilitation de l'audition qu'il permet, même si elle n'est pas parfaite. L'implant cochléaire et la lecture labiale peuvent en quelque sorte être aussi considérés comme complémentaires, et nous avons pu en effet mettre en évidence qu'une collaboration audiovisuelle pouvait être

exploitée efficacement en co-perception et production de la parole. Là aussi des enjeux de perspective tant en traitements automatique qu'en perception s'ouvrent.

Enfin, la problématique de la surdité a permis d'utiliser la surdité profonde comme paradigme intéressant afin d'étudier la production sonore de parole sous les aspects physique. En effet, elle permet d'éviter tout contrôle lié au feedback auditif.

Ce projet en parole augmentée et adaptative apporte des contributions significatives aux domaines du traitement automatique de la parole, de la production et de la perception dans les dimensions audio et visuelles. Il se réalise grâce à une alliance entre disciplines allant de la physique pour les sciences de l'ingénieur à la socio-linguistique voire la rhétorique en passant par le traitement du signal et la phonétique grâce à des compétences relevant des SHS, STIC et SPI que j'ai pu regrouper dans des projets communs de thèse, de contrats passés ou à venir.

Enfin, la rédaction de cette habilitation a nécessité que je prenne un temps de retrait et de solitude pour mener à bien l'approfondissement de ma réflexion. Cela a été clairement une solitude choisie pour cet exercice. Mais puisqu'il s'agit d'un état dans lequel le chercheur que je suis a pu être amené à se trouver et risque encore d'y être confronté, que ce soit le « simple » ressenti dans le cas de la sensation d'être dans une impasse ou de solitude subie comme celle du responsable face à une situation difficile, source de souffrance, ou encore lorsqu'il s'agit de la solitude acceptée voire recherchée comme un moyen de la réflexion, source de créativité et de dépassement personnel, je souhaite vous partager quelques citations philosophiques et les conclusions que j'en ai tirées.

Ainsi, "La solitude offre à l'Homme intellectuellement haut placé un double avantage : le premier, d'être avec soi-même, et le second de ne pas être avec les autres". Cette citation du philosophe allemand Schopenhauer exprime l'idée selon laquelle la solitude permettrait une réflexion sur soi-même. Cependant "l'homme seul est quelque chose d'imparfait; il faut qu'il trouve un second pour être heureux". A travers cette citation, Pascal soutient que l'on ne peut vivre heureux dans la solitude et que l'Homme a besoin d'autrui pour atteindre le bonheur en renvoyant à la société.

Solitude et société devraient donc composer et se succéder selon la conception de Sénèque. Mais c'est justement cette solitude qui s'installe lorsqu'on réalise le fossé qui peut exister entre certains enjeux scientifiques et les réalités des situations de terrain que nous pouvons croiser chaque jour. Car il n'est pas toujours facile de trouver la matérialité du trait d'union entre le monde de la pensée et celui de l'action. Et à ce paradoxe s'ajoute celui de l'homme moderne qui par son exigence de liberté et d'autonomie appartient à une société individualiste alors qu'il éprouve pourtant dans le même temps le besoin de créer et de maintenir des liens sociaux

comme en témoigne Jean-François Mattei dans son discours de clôture de la journée dédiée à la solitude en 2013 organisée par la *Fondation Croix-Rouge Française*. Mon engagement depuis longtemps dans ce mouvement humanitaire concrétise pour moi ce trait d'union. Il me permet de toucher des enjeux de société, avec au centre la personne dans son corps, son esprit et dans toute sa dimension sociale, d'être à la fois dans le concret des actions qui peuvent apporter des solutions rapidement aux situations rencontrées ce qui en retour me permet de relativiser les soucis, qui dans un autre contexte pourraient être caractérisés comme exagérés, tout en alimentant mon travail de recherche en intérêts de société.

Références bibliographiques

Aboutabit, N., 2007. Reconnaissance de la Langue Française Parlée Complétée. Manuscrit de thèse, Université de Grenoble.

Attina, V., 2005. La Langue française Parlée Complétée : production et perception. Phd dissertation. Institut National Polytechnique de Grenoble, Grenoble, France.

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

Auer, E.T. Bernstein, L.E., 2007. Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. Journal of Speech, Language, and Hearing Research, Vol.50, pp. 1157-1165.

Beautemps D., Borel P., Manolios S., 1999. Hyper-articulated Speech: Auditory and Visual intelligibility. In Proceedings of the 6th European international conference EUROSPEECH, Budapest, Hungary, 5-9 September, 1999.

Beautemps, D., Badin, P. & Bailly, G., 2001. "Linear degrees of freedom in speech production: Analysis of cineradio and labio-films data for a reference subject, and articulatory-acoustic modelling". Journal of the Acoustical Society of America, 109, 5, 2165 – 2180.

Beautemps, D., Cathiard, M. A. & Leborgne Y., 2003. Benefit of audiovisual presentation in close shadowing task. In Proceedings of ICPHS conference, Barcellona, Spain.

Beautemps, D., Cathiard, M. A., Attina, V., Savariaux, C., 2012. Temporal organisation of Cued Speech Production. In Audiovisual Speech Processing, G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds), pp. 104-120.

Benoit, C., Lallouache, T., Mohamadi, T., Abry, C., 1992. A set of French visemes for visual speech synthesis. In: Bailly, G., Benoit, C. (Eds.), Talking Machines: Theories, Models and Designs. Elsevier Science Publishers, Amsterdam, pp. 485-504.

Bernstein, L.E., Demorest, M.E., Tucker, P.E., 2000. Speech perception without hearing. Perception & Psychophysics 62(2), 233-252.

Bernstein, L.E., Auer, E.T., Jr, & Jiang, J., 2010. "Lipreading, the lexicon, and Cued Speech", in C. la Sasso, J. Leybaert, K. Crain (Eds.), Cued Speech for the Natural Acquisition of English, Reading, and Academic Achievement. Oxford University Press.

Carton, F., 1974. Introduction à la phonétique du Français. Collection "Etudes", Série de Langue française dirigée par Jean Batany. Bordas, Paris/ Bruxelles/ Montréal.

Cornett, R., 1967. Cued Speech. American Annals of the Deaf 112, 3-13.

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

Depau G. (2012), Italien et sarde dans l'espace urbain de Cagliari : plurilinguisme et contexte

d'interaction. CoReLa – Cognition Représentation Langage, HS11 « Cotexte, contexte, situation » [on line : ttp://corela.edel.univ-poitiers.fr].

Durand, J., B. Laks & C. Lyche (eds.)(2009). Phonologie, variation et accents du français. Paris: Hermès.

Garnier, M., Henrich, Dubois, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect, Journal of Speech, Language, and Hearing Research, American Speech-Language-Hearing Association, 2010, 53 (3), p. 588-608.

Giovanni A., Yu P., Révis J., Guarella MD., Teston B., Ouaknine M. (2006). "Analyse objective des dysphonies avec l'appareillage EVA. Etat des lieux.", Revue Oto-Rhino-Laryngologie Française, 90, p3 183-192.

Heracleous, P., Beautemps, D., Aboutabit, N., 2010. Cued Speech automatic recognition in normal-hearing and deaf subjects. Speech Communication 52(6): 504-512.

Jansen S., Luts H., Wagener K.C., Kollmeier B., Del Rio M., Dauman R., James C., Fraysse B., Vormes E., Frachet B., Wouters J., & Van Wieringen A. (2012). Comparison of three types of French speech-in-noise tests: A multi-center study. Int J Audiol, 51 (3), 164-73.

Leybaert, J., 2000. Phonology acquired through the eyes and spelling in deaf children. Journal of Experimental Child Psychology 75, 291-318.

Leybaert J, LaSasso CJ., 2010. Cued speech for enhancing speech perception and first language development of children with cochlear implants.Trends Amplif. 2010;14(2):96-112.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H & H theory, In W. J. Hardcastle & A. Marchal: "Speech Production and, 403-439, Kluwer Academic Publishers, Dordrecht.

McGurk and John MacDonald, 1976. "Hearing lips and seeing voices", Nature 264, 746-748.

MacDonald, J., McGurk, H., 1978. Visual influences on speech perception processes. Perception and Psychophysics 24, 253-257.

Macé, S. (dir.) (2014). Dossier « Sur l'amplification ». Exercices de rhétorique, n. 4, décembre 2014.

Potamianos, G., Neti, C, Gravier, G., Garg, A. and Senior, A.W, 2003. Recent advances in the automatic recognition of audiovisual speech," in Proceedings of the IEEE, vol. 91, Issue 9, pp. 1306–1326.

Potamianos, G., Neti, C., Luettin, J., and Matthews I. (2012). Audiovisual automatic speech recognition. In G. Bailly, P. Perrier, E. Vatikiotis-Bateson (Eds), Audiovisual Speech Processing, pp. 193-247.

Reisberg, D., Mclean, J., Goldfield, A., 1987. Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In: Dodd, R., Campbell, R. (Eds.), Hearing by Eye : The Psychology of Lipreading. Lawrence Erlbaum Associates Ltd, Hillside, NJ, pp. 97-113.

Roustan, B. (2012). Etude de la coordination gestes manuels/parole dans le cadre de la désignation. Manuscrit de thèse, Université de Grenoble.

Scarbel, L., Beautemps, D., Schwartz J.-L., and Marc Sato, 2014. The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close-shadowing.

Schwartz, J., Robert-Ribès, J., Escudier, P., 1998. Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In: Campbell, R., Dodd, B. (Eds.), Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech. Psychology Press, Hove, UK, pp. 85-108.

Signorello, R. (2014). La voix charismatique : aspects psychologiques et caractéristiques acoustiques. Manuscrit de thèse, Université de Grenoble.

Strelnikov, K., Rouger, J., Lagleyrec, S., Frayssec, B., Deguine, O., Barone, P., 2009; Improvement in speech-reading ability by auditory training: Evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. Neuropsychologia 47, 972–979.

Sumby, W.H., Pollack, I. (1954). Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26 (2), 212-215.

Summerfield, Q., 1979. Use of visual information of phonetic perception. Phonetica 36, 314-331.

Summerfield, A., 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, R., Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lipreading. Lawrence Erlbaum Associates Ltd., Hove, UK, pp. 3-51.

Annexe : 7 publications significatives

Beautemps, D., Badin, P. & Bailly, G., 2001. “Linear degrees of freedom in speech production: Analysis of cineradio and labio-films data for a reference subject, and articulatory-acoustic modelling”. Journal of the Acoustical Society of America, 109, 5, 2165 – 2180.

Beautemps, D., Cathiard, M. A., & Leborgne, Y., 2003. Benefit of audiovisual presentation in close shadowing task. In Proceedings of ICPHS conference, Barcellona, Spain.

Aboutabit, N., Beautemps, D. & Besacier, L., 2007. Automatic identification of vowels in the Cued Speech context. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP), 2007.

Heracleous, P., Beautemps, D., Aboutabit, N., 2010. Cued Speech automatic recognition in normal-hearing and deaf subjects. Speech Communication 52(6): 504-512.

Beautemps, D., Cathiard, M. A., Attina, V., Savariaux, C., 2012. Temporal organisation of Cued Speech Production. In Audiovisual Speech Processing, G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds), pp. 104-120.

Ming, Z., Beautemps, D., Feng, G., 2013. GMM Mapping Of Visual Features of Cued Speech From Speech Spectral Features. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP), 2013.

Lucie Scarbel, Denis Beautemps, Jean-Luc Schwartz, S. Schmerber, Marc Sato. (2014). L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles. Journées d'Etudes sur la Parole (JEP 2014), June 2014, Le Mans, France.

Beautemps, D., Badin, P. & Bailly, G., 2001. “Linear degrees of freedom in speech production: Analysis of cineradio and labio-films data for a reference subject, and articulatory-acoustic modelling”. Journal of the Acoustical Society of America, 109, 5, 2165 – 2180.

Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling

Denis Beautemps,[a] Pierre Badin,[b] and Gérard Bailly[c]
Institut de la Communication Parlée, UMR CNRS 5009, INPG–Université Stendhal, 46, Av. Félix Viallet, F-38031 Grenoble Cedex 1, France

(Received 15 August 1999; accepted for publication 8 February 2001)

The following contribution addresses several issues concerning speech degrees of freedom in French oral vowels, stop, and fricative consonants based on an analysis of tongue and lip shapes extracted from cineradio- and labio-films. The midsagittal tongue shapes have been submitted to a linear decomposition where some of the loading factors were selected such as jaw and larynx position while four other components were derived from principal component analysis (PCA). For the lips, in addition to the more traditional protrusion and opening components, a supplementary component was extracted to explain the upward movement of both the upper and lower lips in [v] production. A linear articulatory model was developed; the six tongue degrees of freedom were used as the articulatory control parameters of the midsagittal tongue contours and explained 96% of the tongue data variance. These control parameters were also used to specify the frontal lip width dimension derived from the labio-film front views. Finally, this model was complemented by a conversion model going from the midsagittal to the area function, based on a fitting of the midsagittal distances and the formant frequencies for both vowels and consonants. © *2001 Acoustical Society of America.* [DOI: 10.1121/1.1361090]

PACS numbers: 43.70.Bk [AL]

I. INTRODUCTION

''Speech is rather a set of movements made audible than a set of sounds produced by movement,'' posited Stetson in 1928 (p. 29). This statement could be more properly rephrased as ''Speech can be regarded as the *audible* and *visible* signals resulting from articulatory movement,'' as stated, for instance, in the *speech robotics* approach fostered in the collaborative European project *Speech Maps* (Abry *et al.*, 1994). In this framework, the speech apparatus is viewed as a *plant* driven by a *controller* so as to recruit articulators and coordinate their movements, which have simultaneous acoustic and visual consequences.

The concept of plant and controller implies the notion of a relatively small number of *independent degrees of freedom* for the articulatory plant, i.e., the specification, for each articulator, of a limited set of movements that can be executed independently of each other by the articulator. As emphasized by Kelso *et al.* (1986), however, the speech production apparatus is made of a large number of neuromuscular components that offer a potentially huge dimensionality and which must be functionally coupled in order to produce relatively simple gestures [this view forms the basis of the concept of coordinative structures in speech, cf. Fowler and Saltzman (1993)]. Maeda (1991) refers to a similar concept in terms of ''elementary articulators.''

One *independent degree of freedom* may be more precisely defined for a given speech articulator as one variable that can completely control a specific variation of shape and position of this articulator, and that is statistically independent of the other degrees of freedom over a set of tasks. These degrees of freedom can be determined by observing the correlations between the various parameters that constitute the accurate geometrical description of the articulators shapes and positions, and retaining only independent parameters. These correlations stem from mainly three levels of implicit or explicit constraints: (1) physical continuity of the articulators (the tongue cannot have a jigsaw shape for instance); (2) biomechanical constraints (the range of possible articulators shapes and positions is limited by the physiological properties of the bony structures and of the muscles); and (3) the nature of the task in relation with control (chewing involves lateral translations of the jaw, but speech does not, and thus jaw has different degrees of freedom depending on the task observed). The correlations observed on articulatory measurements thus cannot always be ascribed with certainty to either biomechanical constraints or to strategies related to the task. For instance, Hoole and Kroos (1998) observed that larynx height and lip protrusion are inversely correlated: this correlation obviously cannot be explained by biomechanical links between lips and larynx, but should be ascribed to control strategies related to the speech task. It thus appears understandable that determining which properties of speech can be attributed to the plant and which to the controller is a recurrent issue in speech motor control (cf., e.g., Perkell, 1991; Scully, 1991; Abry *et al.*, 1994).

It has long been known that midsagittal profiles constitute a privileged representation of speech articulation (cf., e.g., Boë *et al.*, 1995, for a review on vowel representations). Indeed, for most phonemes, the complete vocal tract shape can be fairly well inferred from the midsagittal plane, the most notable exception being lateral sounds. Moreover, midsagittal profiles allow linking of vocal tract articulation and

[a]Electronic mail: beautemps@icp.inpg.fr
[b]Electronic mail: badin@icp.inpg.fr
[c]Electronic mail: bailly@icp.inpg.fr

the resulting acoustics. Articulatory models can therefore be viewed as one of the most efficient means of manipulating vocal tract shapes, and midsagittal profiles as a privileged interface between *motor control* on the one hand and *acoustic and visual* modules of the speech production system on the other hand. Developing and evaluating such articulatory models finally constitutes a good means for identifying the degrees of freedom of speech articulators.

Among the large number of studies devoted to articulatory modeling since the seventies, two main approaches can be identified: *functional articulatory modeling*, where the position and shape of articulators are algebraic functions of a small number of articulatory parameters, and explicit *biomechanical modeling*, where the position and shape of articulators are computed from physical simulations of the forces generated by muscles and of their consequences on the articulators.

In linear articulatory models, the relations between articulator positions and shapes and the control parameters can either be defined in geometrical terms, in which case the degrees of freedom of the articulatory plant are decided *a priori* and fitted to the data *a posteriori* (cf., e.g., Coker and Fujimura, 1966; Liljencrants, 1971; Mermelstein, 1973), or based on articulatory data measured on one or several subjects, in which case the degrees of freedom of the plant emerge from the data (cf., e.g., Lindblom and Sundberg, 1971; Maeda, 1990; Stark *et al.*, 1996).

The general approach of biomechanical articulatory models consists in modeling muscular forces and articulator structure by means of methods inspired from mechanical analysis and numerical simulation (cf., e.g., Perkell, 1974; Wilhelms-Tricarico, 1995; Laboissière *et al.*, 1996; Payan and Perrier, 1997). These models present the advantage of being physical models with intrinsic dynamics, although necessarily extremely simplified, but their control remains very complex, in particular due to the high number of degrees of freedom represented by each individual muscle command. Sanguineti *et al.*'s (1998) work constitutes a good illustration of this. They fitted, with their model, the articulator shapes and positions measured from the x-ray database already used by Maeda (1990) and determined, by optimization, the commands of the 17 muscles involved in their model. They identified then, by linear component analysis applied in the so-called λ-space corresponding to the biomechanical tongue control parameter space, the synergies between these commands, and showed that six independent components could account for most of the data variance of the midsagittal tongue shape. These first six components are closely related to the degrees of freedom that could be extracted directly from the original x-ray contours (Maeda, 1990). It thus appears that, from the point of view of the degrees of freedom, such complex biomechanical models are not a prerequisite to the accurate description of static speech articulation.

Rather than developing *a priori* complex biomechanical models with degrees of freedom in large excess and then reducing this high dimensionality based on articulatory data, we have adopted a dual approach in the present work. More precisely, our objectives were to determine the linear degrees of freedom of one subject's articulators in the midsagittal plane, and to build an *articulatory-acoustic plant* that could be considered a faithful and coherent representation of this subject's articulatory and acoustic capabilities.

The present paper describes our approach to this problem: (1) design of the corpus and collection of articulatory-acoustic data for one subject, (2) analysis of the data and extraction of the independent linear degrees of freedom of the articulators, and (3) the development of a linear articulatory-acoustic model based on these degrees of freedom.

II. ARTICULATORY AND ACOUSTIC DATA ACQUISITION METHODOLOGY

A. The experimental setup: Synchronized cineradio- and labio-film

Cineradiography was chosen as the best compromise between good spatial and temporal resolutions for the articulatory data. This technique, which has been used successfully for speech studies at the Strasbourg Phonetic Institute (cf., e.g., Bothorel *et al.*, 1986), was used in synchrony with the video labiometric method developed at ICP by Lallouache (1990) (cf. also Badin *et al.* 1994a). The recordings were performed at the Strasbourg Schiltigheim Hospital, France. The subject's head was positioned at a distance of 50 cm from the x-ray emitter and 20 cm from the radiance amplifier. An aluminum filter was placed in the lip region to avoid overexposure of the lips, thus improving the contrasts in this region (cf. Bothorel *et al.*, 1986). The vocal tract images produced by the radiance amplifier were captured and recorded by a 35-mm film camera. The subject's lips, painted in blue to allow the lip contours to be extracted by an image processing procedure, were recorded using a video camera. Both cameras were operating at a rate of 50 frames per second. The speech signal, captured by a directional microphone placed at a distance of 10 cm from the subject's mouth, was synchronously recorded.

B. The subject

The choice of the set of subjects always poses a dilemma: a single subject study surely reduces the generality of the work but allows us to gather rich and detailed data, whereas a study with a larger panel of subjects may permit us to draw some general conclusions but limits the extent of the data that can be practically acquired and processed.

Under the auspices of the European collaborative project *Speech Maps*, Abry *et al.* (1994) aimed to gather a variety of converging and complementary articulatory/acoustic data for one subject uttering the same speech material in a controlled manner in different experimental setups. This policy resulted in a large set of data of potential use in speech production modeling, such as vocal tract acoustic transfer functions (Djéradi *et al.*, 1991), acoustic and aerodynamic pressure and flow in the tract (Stromberg *et al.*, 1994; Badin *et al.*, 1995; Shadle and Scully, 1995), electropalatography data (Badin *et al.*, 1994b), and more recently 3D MRI vocal tract images (Badin *et al.*, 1998, 2000) and video face data (Badin

TABLE I. Corpus used for the cineradiographic recordings.

[aɛeiyuoø]
[pavapavipavupivipivupivy]
[pazapazipazupizipizupizy]
[paʒapaʒipaʒupiʒipiʒupiʒy]
[abaabiabuibiibuiby]
[adaadiaduidiiduidy]
[agaagiaguigiiguigy]

et al., 2000). As the present study was conducted in the framework of this project, the same subject was therefore chosen as a *reference* subject.

C. The corpus

As mentioned in the previous section, one of the major aims of the present study was to determine the degrees of freedom of a subject's articulators for speech, excluding any other nonspeech movements. Attaining this goal would ideally require recording a large corpus of speech material containing all possible combinations of phonemes. This is obviously not practical in general, and particularly inappropriate in the case of cineradiography, due to health hazards related to this method. The corpus was thus designed to include as many combinations of Vowel Consonant Vowel (VCV) sequences as possible in a very limited amount of time.

The voiced French plosive and fricative consonants C=[vzʒbdg] were chosen in six vocalic contexts involving the four French extreme vowels V=[aiuy]:aCa, aCi, aCu, iCi, iCu, iCy. It was assumed that the voiceless cognates of these consonants correspond approximately to the same articulation. The presence of voicing was expected to simplify the tracking of formants during fricative consonants. The French [l] was excluded because of the impossibility of getting information on the lateral channels from midsagittal x-ray pictures. Nasals were also excluded because velar movements do not directly influence other articulators movement (although nasal vowels in French seems to imply some additional tongue backing compared to other vowels, cf., e.g., Zerling, 1984), and will thus be studied in the near future. Finally the French [ʁ] was also not included because it was not hypothesized to require extra degrees of freedom for the midsagittal profile. The extreme vowels were expected to represent the most extreme vocalic articulations in French. Moreover, the fricative items were interspersed with [p]'s in order to allow the estimation of subglottal pressure during the fricatives as the intraoral pressure during the closure of [p] (Demolin *et al.*, 1997). In addition, a series of connected vowels [aɛeiyuoø] was recorded in order to test formant/cavity affiliation hypotheses (Bailly, 1993). Finally, the corpus duration could be reduced to about 24.5 s of signal (actually leading to 1222 pictures), with the following distribution of phonemes: 30 [ai], 12 [u], 6 [y], 6 [vzʒbdg], 18 [p]. Table I presents the complete corpus.

D. Processing of the x-ray and video images

For each picture, the sagittal contours were first drawn by hand from a projection of the picture onto a piece of paper, digitized by a scanner, and finally 11 subcontours were hand-edited by means of an interactive software. These subcontours correspond to different articulators or vocal tract regions: the upper and lower lips, the hard palate, the velum, the different components of the pharynx, the larynx, the tongue, the jaw, and the hyoid bone. Figure 1 presents an example of the resulting midsagittal contour. Note that rigid structures (hard palate, jaw, hyoid bone) were not drawn again for each image, but were given reference contours that best fitted most of the shapes observed on the whole set of images; for these structures, the operator's task was then only to optimally position the reference shapes for each image by roto-translation. This procedure presented four advantages: (1) the operator's task was made easier and faster; (2) it reduced the noise due to manual drawing; (3) it avoided the difficulty of precisely determining reference landmarks such as incisor edges from images where the contrast is not always very high; (4) it offered the possibility of determining in a straightforward manner the positions of these rigid bodies in the midsagittal plane (see the discussion on jaw analysis below).

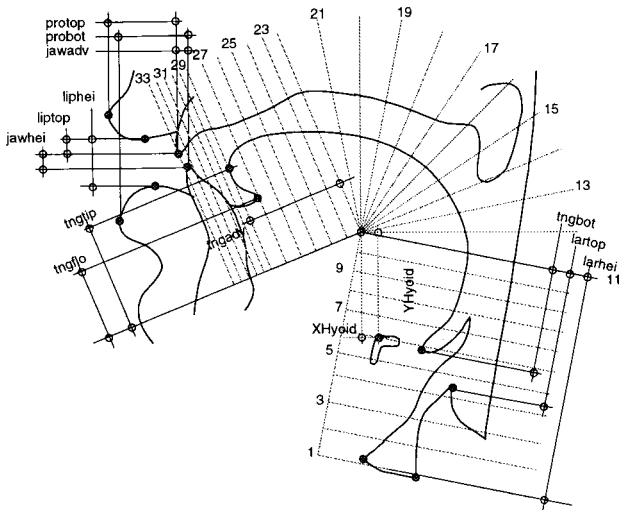


FIG. 1. Example of manually drawn *VT* contours and associated articulatory measures: upper lip protrusion *ProTop*, lower lip protrusion *ProBot*, upper lip elevation *LipTop*, jaw height *JawHei*, jaw advancement *JawAdv*, tongue tip advancement *TngAdv*, tongue tip height *TngTip*, tongue floor height *TngFlo*, tongue bottom *TngBot*, height of the larynx top *LarTop*, height of the larynx bottom *LarHei*, *x/y* coordinates of the hyoid bone *XHyoid,YHyoid*.

Concerning the lips, the blue of the video front lip images was converted into absolute black by means of an analogue Kroma-key, and the inner contour was then automatically determined by simple adapted thresholding (cf. Lallouache, 1990, or Badin *et al.*, 1994a, for more details).

E. Articulatory measurements

Before going into some details, it is useful to specify the midsagittal coordinate system attached to the skull structure and used in this study. The lower edge of the upper incisors is given arbitrary *x/y* coordinates (5,10) in cm. The *x*-axis is positive in the *posterior* direction of the head (toward the back), and negative in the *anterior* direction (toward the nose); the *y*-axis is positive in the *superior* direction (toward the brain), and negative in the *inferior* direction (toward the

feet). Finally, it happens that the direction of the *maxillary occlusal plane* (defined as the plane ''given by the tips of the central incisors and at least two other maxillary teeth on opposite sides of the mouth,'' Westbury, 1994), is oriented at an angle of 4.6° from the $y=0$ axis.

The values of a number of geometrical parameters (see Fig. 1) have been determined from the midsagittal contours: upper *ProTop* and lower *ProBot* lip protrusions, upper lip elevation *LipTop*, jaw height *JawHei*, jaw advancement *JawAdv*, tongue tip advancement *TngAdv* and height *TngTip*, tongue floor height *TngFlo*, tongue bottom *TngBot*, height of the larynx top *LarTop* and bottom *LarHei*, x/y coordinates of the hyoid bone *XHyoid*, *YHyoid*. In addition, three parameters were extracted from the video front views of the lips: lip height *B*, lip width *A*, and the intra-labial lip area *S*.

Note that the distance between the upper and lower lips can be determined either from the midsagittal profile *LipHei* or from the front view *B*: as expected, the two measures are very close to each other, *B* being less accurate when the lip opening is close to zero, particularly for the rounded vowels [uy], due to the fact that the subject's upper lip tends to mask the intra-labial orifice in such cases. The correlation coefficient between the measures is 0.99, while the rms error is 0.1 cm.

The articulators are expected to follow relatively smooth trajectories due to their long time responses (cf., e.g., the jaw characteristic resonance frequency of 5–6 Hz mentioned by Sorokin *et al.*, 1980). Deviations of geometric measures from their smooth trajectories revealed that the noise added by the whole chain of acquisition was in the range of 1-mm peak-to-peak.

F. Midsagittal contours

A semi-polar grid has been used to describe the midsagittal contours, as has traditionally been done since Heinz and Stevens (1965) or Maeda (1979). However, as proposed by Gabioud (1994), two parts of this grid have been made adjustable in order to follow the movements of the larynx (gridlines 1 to 6, line 1 being the lowest one near the glottis) and the movement of the tongue tip (in fact, the *tongue blade*, defined as the linguistic class *coronal articulation*; grid lines 24 to 28). This grid presents a double advantage: (1) the number of intersection points between the grid and the tongue contour is constant, whatever the extension of the tongue tip or of the larynx, which is a crucial feature for further statistical analysis: (2) the fact that the measurement grid follows tongue tip movements implies that all the points in the vicinity of the tongue tip present a behavior close to that of *flesh-points*, i.e., points mechanically attached to the tongue surface, which is more than just a description of tongue shape (this is useful when recovering tongue contours from flesh-points coordinates measured by electromagnetic articulometry; see Badin *et al.*, 1997). Finally, a third part of the grid has been introduced to describe the alveolar dental cavity with adjustable grid lines (grid lines 29 to 33) equally spaced between the tongue tip and the lower edge of the upper incisor. The inner and outer vocal tract midsagittal contours intersect thus the grid lines at 2×33 points; each contour can thus be represented by the 33-element vector (referred to as *Int* and *Ext*) of the abscissa of these intersection points along the grid lines. Note that, as mentioned by Westbury (1994), the dimensionality of the articulators may change depending on the coordinate system used. For instance, a point running on a fixed circle appears to have two *linearly* independent degrees of freedom in a Cartesian system, but only one single degree of freedom in a polar coordinate system. The choice of the dynamically adjustable semi-polar grid system seems a good solution to avoid artificial overdimensionality.

The velum is terminated in the midsagittal plane by the uvula. It can be in contact with the upper surface of the tongue, however, without creating a real constriction in the vocal tract, since air remains free to flow on each side. This fact has been approximately taken into account by making the velum artificially thinner by a factor linearly increasing from 0 at its extremity to about 40% at its base, and by shifting the result so as to align its posterior wall with the pharyngeal wall.

G. Formants

The speech signal was digitized at 16 kHz, and the first four formant values were estimated using LPC analysis with a 20-ms window centered on the times where the midsagittal views were acquired, leading to formant trajectories sampled at 50 Hz. Because of the background noise due to the x-ray emitter, the signal-to-noise ratio was rather poor (about 25 dB), and thus some of the formants had to be hand-edited. This was done in reference to another version of the same corpus recorded by the subject in good recording conditions. The $F1/F2$ and $F1/F3$ spaces for the pooled vowels and consonants are shown in Fig. 2.

III. ANALYSIS OF THE INDEPENDENT LINEAR DEGREES OF FREEDOM OF THE MIDSAGITTAL CONTOURS

A. Principles

1. Identifying degrees of freedom

As mentioned in Sec. I, the approach taken in the present study to determine the degrees of freedom of the various speech articulators is based on articulatory data obtained from one *subject* producing a given *corpus* in a given *language*.

In general, speech articulators possess excess degrees of freedom, i.e., a given articulation can be achieved by means of different combinations of the available degrees of freedom of the articulators (cf. bite-block experiments performed by Lindblom *et al.*, 1979). Control strategies finally aim at recruiting these degrees of freedom when they are needed to attain given articulatory/acoustic/visual goals, and leaving them free to anticipate other goals whenever possible (this is one basic principle of *coarticulation*). In the present data driven approach, the problem is to decide the repartition of the variance of the measured articulatory variables between the different variables associated with the degrees of freedom. The present work rests on a common consideration in speech motor control modeling: what is explained by the biomechanics of the speech plant does not need to be worked

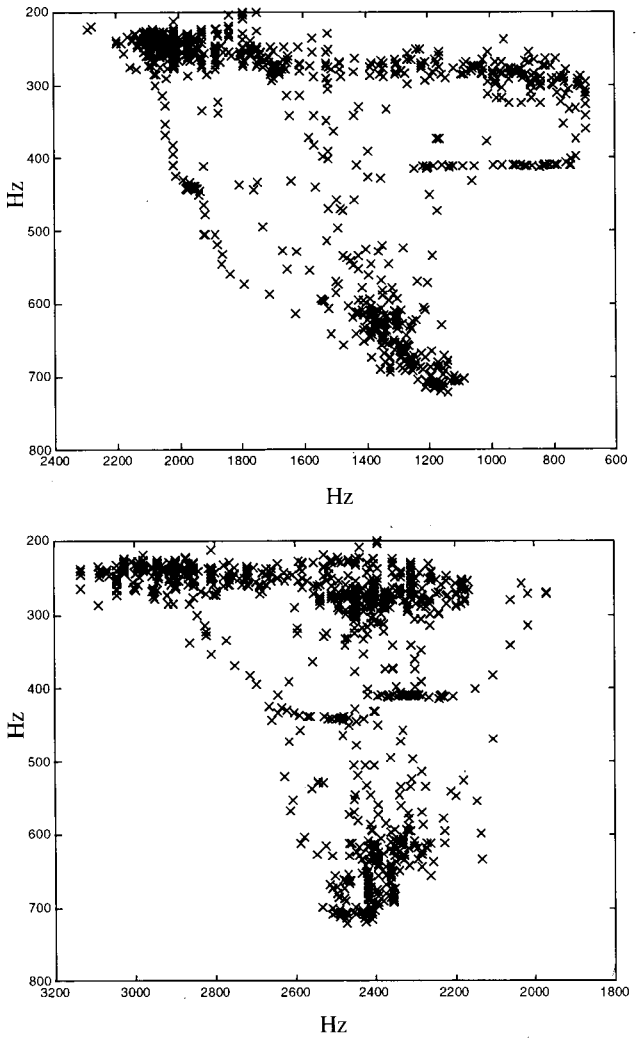


FIG. 2. Measured $F2/F1$ (top) and $F3/F1$ (bottom) formant spaces (in Hz) for the vowels and consonants pooled.

out by the controller (Abry *et al.*, 1994; Perrier *et al.*, 1996, 2000). In other words, any correlation observed between the articulatory variables should be used to reduce the number of degrees of freedom of the articulators. However, this approach must be carefully balanced by another criterion, the *biomechanical likelihood*. For instance, if larynx height and lip protrusion are inversely correlated due to the subject's articulatory control strategy (cf., e.g., Hoole and Kroos, 1998), two separate degrees of freedom should nevertheless be considered, even at the price of some residual correlation between the corresponding parameters.

2. Linear component analysis

Another important assumption in the present work is the *linearity* of the analysis and of the associated model: the shape data vectors *DT* are decomposed into linear combinations of a set of basic shape vectors *BV* weighted by loading factors *LF*, in addition to their average *neutral* shape $\overline{DT}$:

$$DT=\overline{DT}+LF\cdot BV.$$

Each loading factor *LFi* corresponds to an independent linear component, if its cross correlation with the other loadings is zero over the corpus of data. The dimensionality of the articulators' shapes and positions can thus be explored by classical linear analysis techniques such as principal component analysis (PCA) and linear regression analysis, as carried out by Maeda (1990, 1991), whose approach largely inspired the present work.

Maeda's approach to this decomposition was to iteratively determine each linear component in the following way: (1) the loading factor *LFi* is determined from the data as described below; (2) the associated basis shape vector *BVi* is determined by the linear regression of the current residual data for the whole corpus over *LFi*; (3) the corresponding contribution of the component is computed as the product of the loadings by the basis shape vector, and is finally subtracted from the current residue in order to provide the next residue for determining the next component.

For some of the linear components, the loading factors were arbitrarily chosen as the centered and normalized values of specific geometric measurements extracted from the contours, such as jaw or larynx height. For the other linear components, loading factors were derived by standard PCA applied to specific regions of the tongue contour.

Note that the solution of this type of linear decomposition is not unique in general: PCA delivers optimal components explaining the maximum data variance with a minimum number of components, but Maeda's linear component analysis allows a certain room of maneuver to control the nature and repartition of the variance explained by the components (for instance to make them more interpretable in terms of control), at the cost of a suboptimal variance explanation.

In this rest of this section, the various geometric measures are studied using statistical linear analysis in order to determine the correlations between these articulatory variables and to determine the degrees of freedom of the articulatory plant.

B. Jaw

The tongue is naturally identified as an important articulator in speech production, and its midsagittal contours, obtained from x-ray profile views of the vocal tract, have been the focus of most modeling efforts. The jaw has long been recognized as one of the main speech articulators, because it carries both the tongue and the lips. Its specific contribution to tongue shape has been clearly identified and related to the phonetic features of vowels (Lindblom and Sundberg, 1971). The dimensionality of jaw motion has been studied by many researchers (cf., e.g., Westbury, 1988; Edwards and Harris, 1990; Ostry *et al.*, 1997). The jaw, a rigid body, possesses six geometrical degrees of freedom (three rotations and three translations); however, it appears that for speech, movements are mostly restricted to the midsagittal plane if the rotation around the jaw axis is neglected (cf., e.g., Ostry *et al.*, 1997), which reduces the degrees of freedom of the jaw to three. From simple geometrical considerations, it is clear that the position of the jaw as a rigid body in a plane is uniquely defined by one rotation (defined here as *JawRot*) and by the two *x/y* translations of a reference point attached to the body, chosen as the upper edge of the lower incisors (defined here

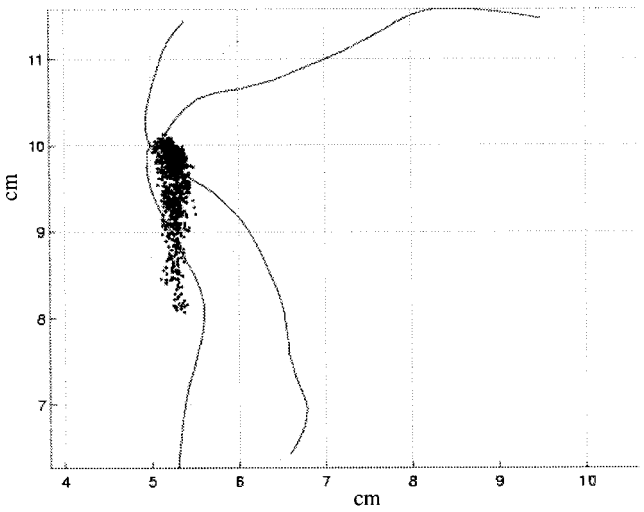


FIG. 3. Dispersion of the lower incisor upper edge superimposed on the contours of the hard palate and of the jaw.

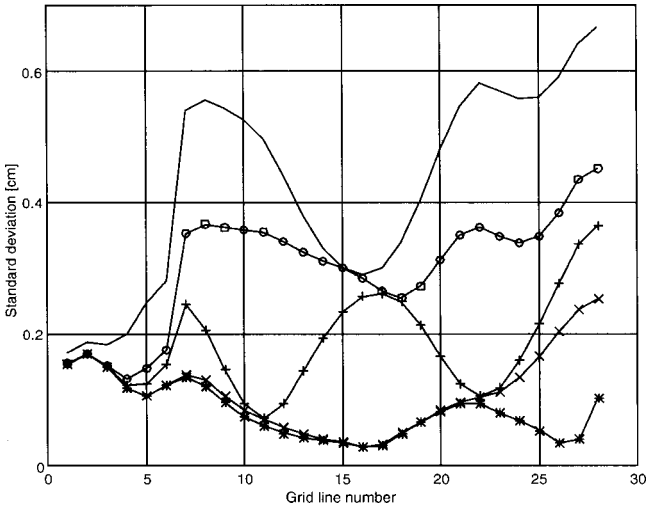


FIG. 4. Standard deviation (in cm) against gridline number for the inner contour of the vocal tract for the successive residues when the effect of the parameters are removed one by one (Raw data: solid line; data after suppression of the contribution of *JH* ''○,'' then of *TB* ''+,'' then of *TD* ''×,'' and finally of *TT* ''*'').

as *JawHei* and *JawAdv*; see Fig. 1). Figure 3 displays the resulting distribution of lower incisor position for the whole corpus.

PCA was applied to the centered—but nonnormalized—jaw position data *JawAdv* and *JawHei*. The first component explains 97.0% of the *JawAdv* and *JawHei* data variance. The overwhelming importance of this component could be predicted intuitively from the fact that the standard deviations of *JawAdv* and *JawHei* are, respectively, about 0.075 and 0.419 cm.

The jaw height component *JH* corresponding to the first degree of freedom of the jaw data was thus defined as the *JawHei* variable centered on its mean and normalized by its standard deviation. A second component, corresponding to jaw advance, *JA* was defined as the residue of *JawAdv* centered and normalized once the linear contribution of *JH* was removed. It can be concluded that for the present subject and corpus, the jaw possesses two independent degrees of freedom in the midsagittal plane, although the second component would have a rather limited influence on the tongue and the lips, as will be discussed further. The jaw was observed to be most retracted for labio-dentals: indeed this retraction allows the lower incisors and the upper lip to get in contact. Maximum jaw protrusion was observed for the coronal fricative [z]: this facilitates the creation of a constriction between the anterior region of the tongue blade and the front region of the alveolar ridge.

C. Tongue

The tongue shape is defined by the vector *Int* of the abscissa of its intersections with the grid lines. Since the jaw carries the tongue, the contribution of its movements should first be subtracted from tongue movements to maintain some biomechanical likelihood. However, due to the complexity of the muscular links between tongue and jaw (cf., e.g., Sanguineti *et al.*, 1998), it is very difficult to separate tongue movements induced by jaw movements from those due to active actions of tongue muscles themselves. In a study involving three subjects (including the subject of the present study), Bailly *et al.* (1998) showed that the slope of the regression line that links jaw height and tongue abscissa can be substantially greater than unity (by almost 100%). This means that tongue movements are apparently larger than the associated jaw movements, indicating that the subjects tend to actively move both jaw and tongue in synergy. In such a case, the passive tongue movement due to jaw movement needs to be determined. However, for the present subject, this synergy was rather weak compared to that of the other subjects (regression slopes lower than 1.15), and did not need to be taken into account: the *JH* parameter was directly considered as the first linear loading factor for each element of *Int*, and the corresponding prediction coefficients were obtained as the coefficients of the linear regression between *Int* and *JH* computed over all the items. Finally, the residual vector *Int_JH*, computed as the difference between predicted and measured values, for all the items, represents the tongue shape from which the contribution of the jaw has been removed.

The variance of the original *Int* data and the variance of the *Int_JH* residual data can be examined in Fig. 4 in terms of standard deviation (i.e., as the square root of the variance), as a function of grid line number. Table II gives, in addition, the global percentage of the total *Int* data variance explained by *JH* (numerical column 1).

The influence of the second jaw parameter *JA* upon tongue contours will be addressed later in this section.

The next step of the analysis consisted of extracting the degrees of freedom of the residual vector *Int_JH*. Gabioud (1994) showed that PCA applied to the whole tongue contour led to poor modeling of the tongue tip, even using three components. It was thus decided to apply PCA separately to the tongue body (gridline 7 to 24) and to the tongue tip (lines 24 to 28).

A first PCA procedure was thus applied to the residues of the 18 points considered for the tongue body, *Int_JH*(7:24). The first two components were retained. The

TABLE II. Summary of data variance explanation for the tongue contours. Column *Design* indicates how the factor was extracted. First column *Var* shows the ratio of data variance explained by the factor for the case the influence of jaw movements is taken into account by one parameter only. The second and third columns *Var* show the ratio of data variance explained when jaw is taken into account by two factors, the second factor being imposed at two different stages of the analysis.

Param.	*Design*	*Var.*	*Var.*	*Var.*
JH	Jaw height	52.2%	52.2%	52.2%
JA	Jaw advance		1.4%	
TB	PCA/tongue body	28.8%	28.6%	28.8%
TD	PCA/tongue body	11.4%	10.4%	11.4%
TT	PCA/tongue tip	3.6%	3.6%	3.6%
JA	Jaw advance			0.2%
	Total	**96.0%**	**96.1%**	**96.1%**

corresponding principal axes are characterized by the eigenvectors associated with the highest two eigenvalues of the cross-correlation matrix computed from these residues. The projections of the centered and normalized residues on these two principal axes give the values of the two associated components: *tongue body* component *TB*, and *tongue dorsum* component *TD*, which describe, respectively, the *front–back* and *flattening–arching* movements of the tongue (see also the nomograms in Fig. 8). These components were then used as predictors for the whole tongue contour. Table II presents a summary of the proportion of the total tongue data variance explained by each component, while Fig. 4 shows the details of the variance of the residues.

The tongue tip was found to possess two independent degrees of freedom [its coordinates, measured as *TngAdv* and *TngTip*[1] (see Fig. 1) are plotted in Fig. 5]: indeed, the residues of *TngAdv* and *TngTip*, after subtraction of the contributions of *JH*, *TB*, and *TD* (determined by the linear regression of *TngAdv* and *TngTip* for the whole corpus over *JH*, *TB*, and *TD*) are clearly not correlated. A first component, more generally dedicated to the representation of the apical region of the tongue, was then extracted: the *tongue tip* component *TT* is defined as the first component determined by the PCA of the residues of the tongue tip region (lines 24 to 28), from which the contributions of *JH*, *TB*, and *TD* have been removed. Its effects can be observed in Fig. 4 and Table II.

The *tonge advance* parameter *TA* was defined as the centered and normalized residue of the measured tongue advance *TngAdv* from which the contributions of *JH*, *TB*, *TD*, and *TT* were subtracted. Since it was, as expected, found to have a negligible predictive power on the tongue abscissa, it was not used as a loading factor for *Int*, but just to control the longitudinal extension of the grid in the front mouth region.

In order to test the influence of the *JA* parameter upon tongue contours, two experiments were carried out: in a procedure similar to that applied to *JH*, *JA* was used as the second imposed loading factor for the tongue analysis in one experiment, and as the loading factor imposed after *JH*, *TB*, *TD*, and *TT* in the other experiment. It was found that *JA* explained only 1.3% of the tongue data variance in the first

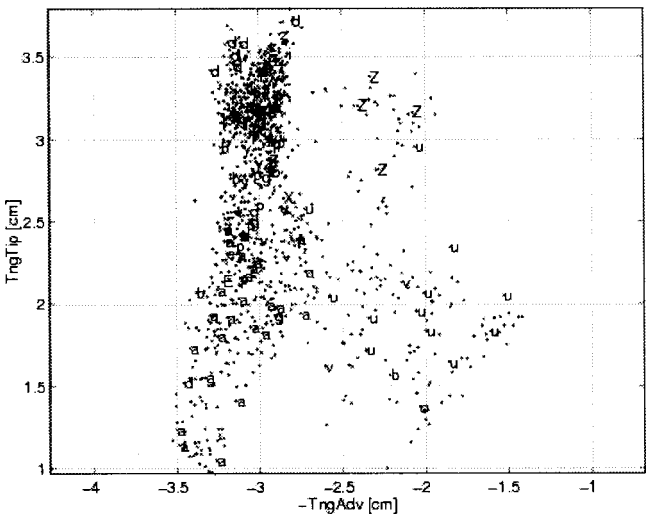


FIG. 5. Plot of the *TngAdv/TngTip* coordinates of the tongue tip (note that these are expressed in the rotated coordinate system attached to the gridline for the front part of the vocal tract; see Fig. 1). Phonemes /ø/, /ʒ/, /ɛ/ are referred to by symbols X, Z, and E, respectively.

case (see Table II, numerical *Var* column 2), and 0.2% in the second case (numerical *Var* column 3). A comparison of the associated nomograms in Fig. 6 suggests that the data variance explained by *JA* in the first case is actually explained by the other components *TB*, *TD*, and *TT*, in the second case. This hypothesis is also supported by results in Table II (and by a more detailed analysis of tongue shape data). *JA* was therefore not used as a control parameter of tongue shape.

In summary, the tongue contours in the grid line system possess four degrees of freedom, controlled by components *JH*, *TB*, *TD*, and *TT*. These four components account for 96% of the tongue variance data, which is only 1.5% less than the variance explained by the first four independent components (but with no direct articulatory interpretation) of a principal component analysis. The standard deviation of the residual error (normally distributed around zero on each gridline) reaches a maximum of 0.15 cm in the vicinity of the pharynx and of 0.1 cm at the tongue tip. The rms reconstruction error for the tongue, i.e., the root mean square error between the measured tongue data and the data calculated with the linear decomposition, amounts to a global value of 0.09 cm, while reaching maxima of 0.15 cm in the vicinity of the pharynx and of 0.1 cm at the tongue tip. The relatively poor modeling of the tongue tip extremity (which results in

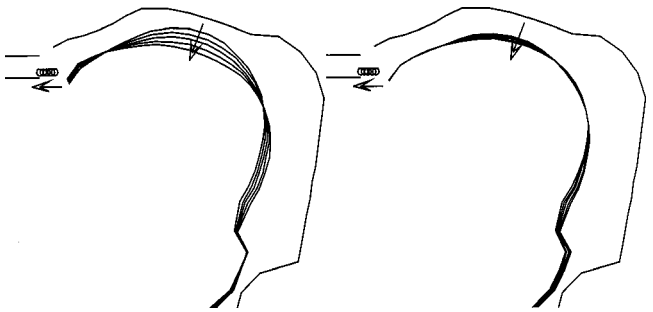


FIG. 6. Articulatory nomograms for *JA*. Left: case where *JA* is the second loading factor in the analysis; right: case where *JA* is the fifth loading factor after *JH*, *TB*, *TD*, and *TT*.

TABLE III. Correlation coefficients of some articulatory measurements. Absolute values higher than 0.6 are in bold face.

	JawHei	*JawAdv*	*LipHei*	*LipTop*	*ProTop*	*ProBot*	*TngTip*	*TngAdv*	*TngFlo*	*TngBot*	*LarHei*	*XHyoid*	*YHyoid*
JawHei	1.000												
JawAdv	−0.226	1.000											
LipHei	**0.641**	−0.039	1.000										
LipTop	−0.448	0.127	0.176	1.000									
ProTop	−0.463	0.148	−0.443	0.060	1.000								
ProBot	**−0.608**	0.062	−0.673	0.024	**0.912**	1.000							
TngTip	**−0.734**	0.207	−0.325	0.375	0.027	0.150	1.000						
TngAdv	0.412	−0.108	0.366	−0.242	−0.568	−0.571	0.142	1.000					
TngFlo	**−0.593**	0.053	−0.290	0.219	−0.083	0.056	**0.877**	0.454	1.000				
TngBot	−0.038	−0.109	−0.347	−0.325	0.352	0.411	−0.219	−0.205	−0.160	1.000			
LarHei	−0.358	−0.068	**−0.550**	−0.136	**0.574**	**0.664**	0.005	−0.380	−0.008	0.797	1.000		
XHyoid	**0.733**	−0.128	0.544	−0.259	−0.416	−0.533	−0.556	0.288	−0.469	−0.169	−0.421	1.000	
YHyoid	0.355	0.100	0.522	0.099	−0.460	−0.572	−0.047	0.351	−0.033	−0.815	**−0.859**	0.574	1.000

only minor acoustical effect) is mainly due to measurement inaccuracies related to the difficulty of precisely defining this tongue tip extremity. A supplementary articulatory control parameter could be extracted to more precisely control the pharyngeal region as implied in the ±Advanced Tongue Root languages (cf., e.g., Tiede, 1996).

Recall finally that the grid system is controlled, in addition, by two parameters, i.e., *TA*, and a parameter related to *LarHei* that will be defined in Sec. III E.

D. Lips

A PCA analysis revealed that 98.4% of the variance of the lip measures *LipHei*, *LipTop*, *ProTop*, and *ProBot* can be explained by three independent components, in addition to the natural contribution of jaw height to lip shape. This is expected, as lip protrusions *ProBot* and *ProTop* are strongly correlated (cf. Table III). Note also that in another study on the same subject, where the lip shape was more accurately described as a three-dimensional mesh of points controlled by the 3-D coordinates of 30 control points (Revéret and Benoît, 1998), Badin *et al.* (2000) also found that three degrees of freedom were sufficient to describe the position of the lips on a corpus of 34 sustained articulations (French vowels and consonants), in addition to the *JH* contribution (the *JA* contribution explained only 1% of the lip data variance). These degrees of freedom are related to three gestures: lip protrusion/rounding, lip closure, and a sort of simultaneous vertical movement of both lips as needed for the subject to realize labio-dentals. In order to simplify the model and its relations to simple articulatory measurements and acoustic interpretations of the lip horn, we decided to use an equivalent set of components: (1) a component related to *LipHei*, taken into account by *LH*, the centered and normalized residue of *LipHei* after removing the *JH* contribution; (2) a component related to *ProTop*, *LP*, the centered and normalized value of the residue of *ProTop* after removing the *JH* contribution; and (3) a component related to a mere vertical, roughly synchronous, movement of both upper and lower lips relative to upper incisors lower edge, taken into account by the *lip vertical position* parameter *LV*, the centered and normalized residue of *LipTop* after removing *JH*, *LH*, and *LP* contributions. Note that this approach results in a slight correlation between *LP* and *LV*. Note also that the horizontal jaw retraction aiming at producing labio-dental constrictions is not taken into account as such, but that its acoustical consequences are dealt with in an indirect way (cf. Sec. IV B 3).

E. Other articulatory measurements

Finally, a number of other articulatory measurements were analyzed. Table III provides the linear correlation coefficients between these measurements.

Table III shows that *LarHei* is partially correlated with *ProBot*, *ProTop*, and *LipHei*. These correlations cannot be explained *a priori* by obvious biomechanical effects, and will thus be ascribed to the speaker control strategies. Indeed, it is clearly established that lip rounding and larynx lowering constitutes, for some subjects, a synergetic strategy for high rounded vowels [uy] (Hoole and Kroos, 1998). Larynx height was thus represented by its centered and normalized value *LY*, and further used to control the grid system (cf. Sec. IV A 2).

The horizontal position of the hyoid bone, *XHyoid*, is very highly correlated to jaw height, while its vertical position, *YHyoid*, is even more strongly correlated to larynx height (see Table III and Fig. 7). These two components are

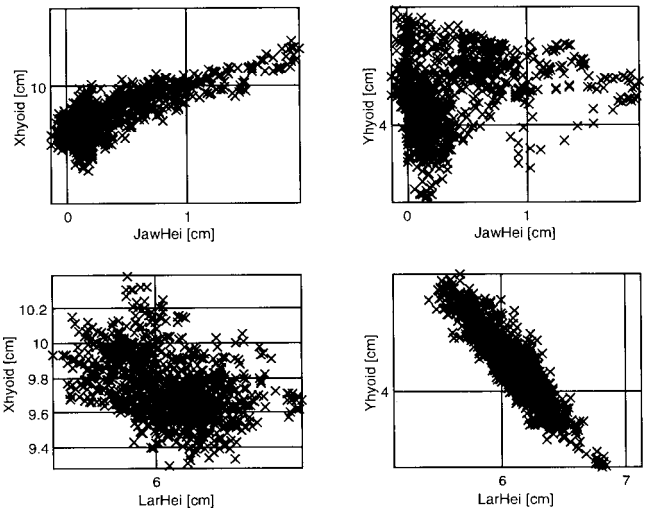


FIG. 7. Plots showing the relations between hyoid bone coordinates and jaw and larynx heights.

TABLE IV. Correlation coefficients of the articulatory control parameters of the model.

	JH	*TB*	*TD*	*TT*	*TA*	*LY*	*LP*	*LH*	*LV*
JH	1.000								
TB		1.000							
TD			1.000						
TT				1.000					
TA					1.000				
LY	−0.358	−0.041	0.272	−0.499	−0.103	1.000			
LP		0.215	0.472	−0.125	−0.118	0.461	1.000		
LH		−0.256	−0.039	0.236	−0.075	−0.417	−0.215	1.000	
LV		0.215	0.065	−0.006	−0.164	−0.045			1.000

clearly less correlated with each other (correlation coefficient $R=0.574$), than found by Westbury (1988) using a more restricted corpus for a single subject ($R=0.871$).

Note that the position of the highest connection point between tongue and epiglottis, referred to as *TngBot* (see Fig. 1), is highly correlated with *LarHei*, as expected. The elevation of tongue floor *TngFlo* is correlated with *TngTip* and *JawHei*.

IV. BERGAME: AN ARTICULATORY-ACOUSTIC MODEL

As stated above, the main function of an articulatory model is to offer a compact representation of articulation, i.e., a representation that needs as few control parameters as possible and is nevertheless accurate enough to be meaningful for speech. The analysis presented in the previous section prepared the ground for establishing such a model, which is necessarily the result of a compromise between a minimum number of control parameters and a maximal explanation of the data variance (or minimal data reconstruction error). The present section describes *Bergame*, an articulatory-acoustic model developed at ICP with the aim of mimicking as closely as possible experimental data gathered on the reference subject.

Bergame consists of: (1) a physiologically oriented linear articulatory model, based on the articulatory data measured from the cineradiofilm and the video labiofilm made on the reference subject; (2) a model of midsagittal-to-area function conversion based on the same subject; (3) an acoustic model.

A. The linear articulatory model

The principle of a linear articulatory model is to *calculate the position and shape of the various articulators as linear combinations of the articulatory control parameters.* The development of the model thus amounts to defining the control parameters and to determining the coefficients of these linear combinations. The nine parameters chosen for controlling the articulatory model stem directly from the previous component analysis: *JH*, *TB*, *TD*, *TT*, *TA*, *LY*, *LH*, *LP*, and *LV*, which are dimensionless, centered, and normalized. These parameters are, in most cases, orthogonal to each other, the exceptions (see Table IV) being due to the subject and language specific control strategies. The model equations are described in some detail in the following. The model behavior is illustrated in Fig. 8 by *articulatory nomograms*, i.e., the variations of the midsagittal contours resulting from variations of the articulatory control parameters from −3 to +3 with +1 steps.

1. Jaw

The jaw has been shown above to possess essentially one degree of freedom for this subject and the corpus analyzed. Jaw position is therefore controlled by the single parameter *JH* that defines *JawHei_mod* by the simple linear relation:

$$JawHei\_mod = JawHei\_mean + JawHei\_std \cdot JH,$$

where *JawHei_std* is the standard deviation of *JawHei* and *JawHei_mean* its mean over the corpus.

2. Tongue, midsagittal distances, and vocal tract outer contours

Since the tongue contours are attached to the grid lines, the next necessary step is to determine the position of the mobile parts of the grid system, namely *TngAdv* and *LarHei*. The modeled tongue advance, *TngAdv_mod*, was found to be almost linearly related to *TA*, *JH*, *TB*, and *TD*, and was therefore controlled by:

$$TngAdv\_mod = TngAdv\_mean + pred\_TngAdv\_JH\_TB\_TD\_TA \cdot [JH,TB,TD,TA],$$

where [*JH,TB,TD,TA*] is the matrix of control parameters, and *pred_TngAdv_JH_TB_TD_TA* are the associated coefficients determined by multiple linear regression. As seen in Sec. II E, *LarHei* is controlled only by *LY*, and not *LP* and *LH*, despite a slight correlation between lips and larynx, in order to ensure an independent control of lips and larynx in the model. *LY* is therefore partially correlated with a number of other control parameters, as seen in Table IV. Note that *TngAdv* and *LarHei* are reconstructed without error.

Finally, the abscissa of the whole tongue contour *Int_mod* (lines 1 to 28) is determined as linear combinations of the parameters *JH*, *TB*, *TD*, and *TT*:

$$Int\_mod = Int\_mean + pred\_Int\_JH\_TB\_TD\_TT \cdot [JH,TB,TD,TT].$$

Figure 8 displays articulatory nomograms for *JH*, *TB*, *TD*, *TT*, and *TA* as well.

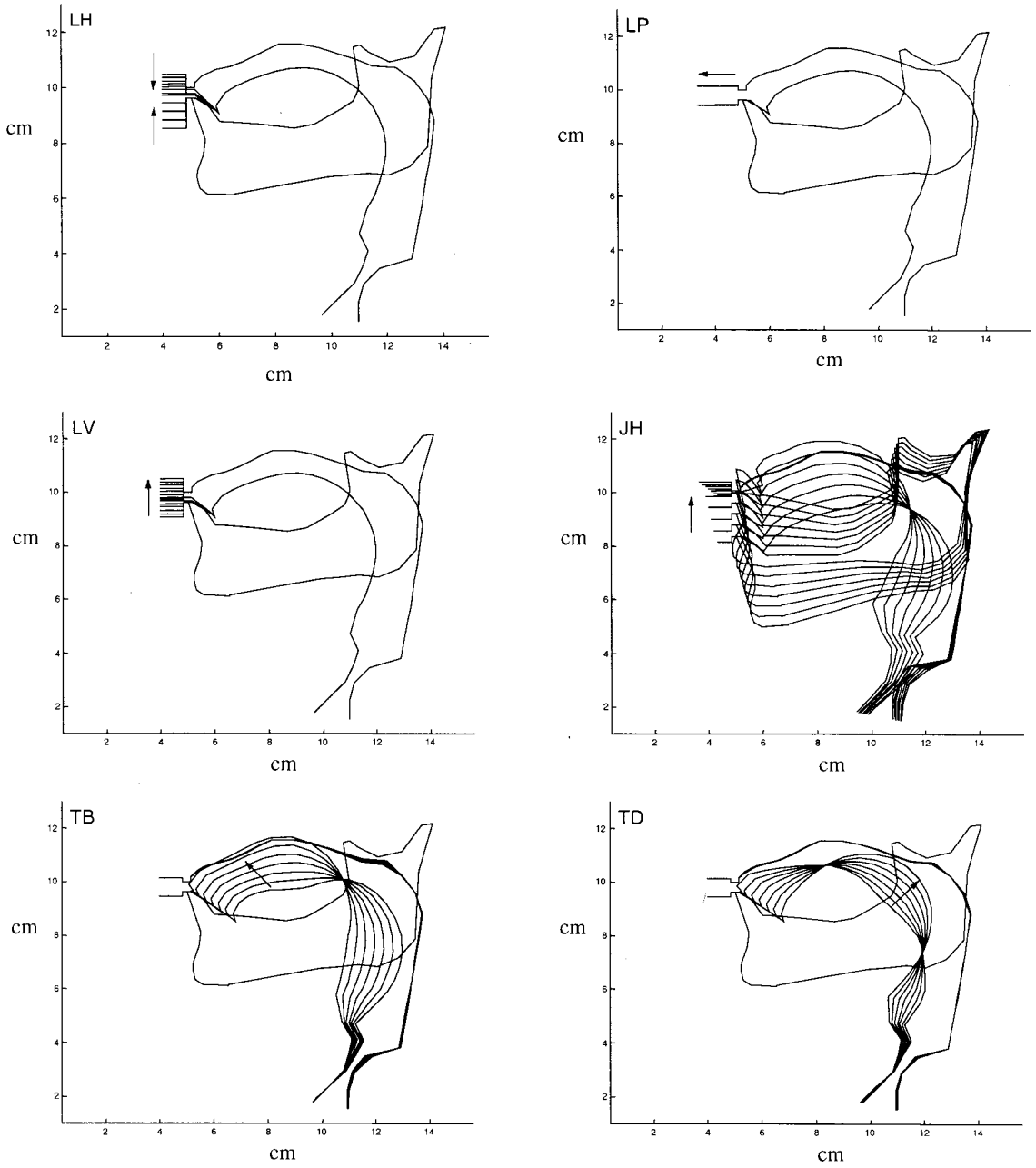


FIG. 8. Articulatory nomograms: variations of the midsagittal contours resulting from variations of the articulatory control parameters from −3 to +3 with +1 steps. Note that the movements of upper and lower lips are *opposite* for the *LH* nomogram, but *parallel* for the *LV* nomogram.

The midsagittal distances have been similarly handled. The abscissa of the vocal tract outer contours are computed as the sum of the abscissa of the tongue and the corresponding midsagittal distances, except for the hard palate region that is considered as a fixed contour.

3. Lips

It has been shown above that the lip geometry of the subject is best described with three degrees of freedom, represented by *LH*, *LP*, and *LV*, in addition to the contribution of *JH*. The lip horn, considered as the vocal tract region anterior to the upper incisor plane, is represented by a single tube section, with a length *ProLip_mod* proportional to the prediction of the *LipTop* dimension, with a proportionality factor of 0.6. This length reduction aims at approximately taking into account the fact that the lip corner position is not known, and that the effective acoustical end of the lip horn is located between the lip corner and the extremities of the lips measured by *ProTop* and *ProBot*. The parameter *ProLip_mod* is thus defined by:

$$ProLip\_mod = 0.6 \cdot (ProTop\_mean + pred\_ProTop\_JH\_LH \cdot [JH, LH]),$$

where the coefficients *pred_ProTop_JH_LH* are obtained by multiple linear regression.

Lip height *LipHei_mod* is similarly modeled as a linear combination of *JH*, *LH*, and *LP*, as well as lip width *A_mod*. Lip vertical position *LipTop_Mod* is also a linear combina-

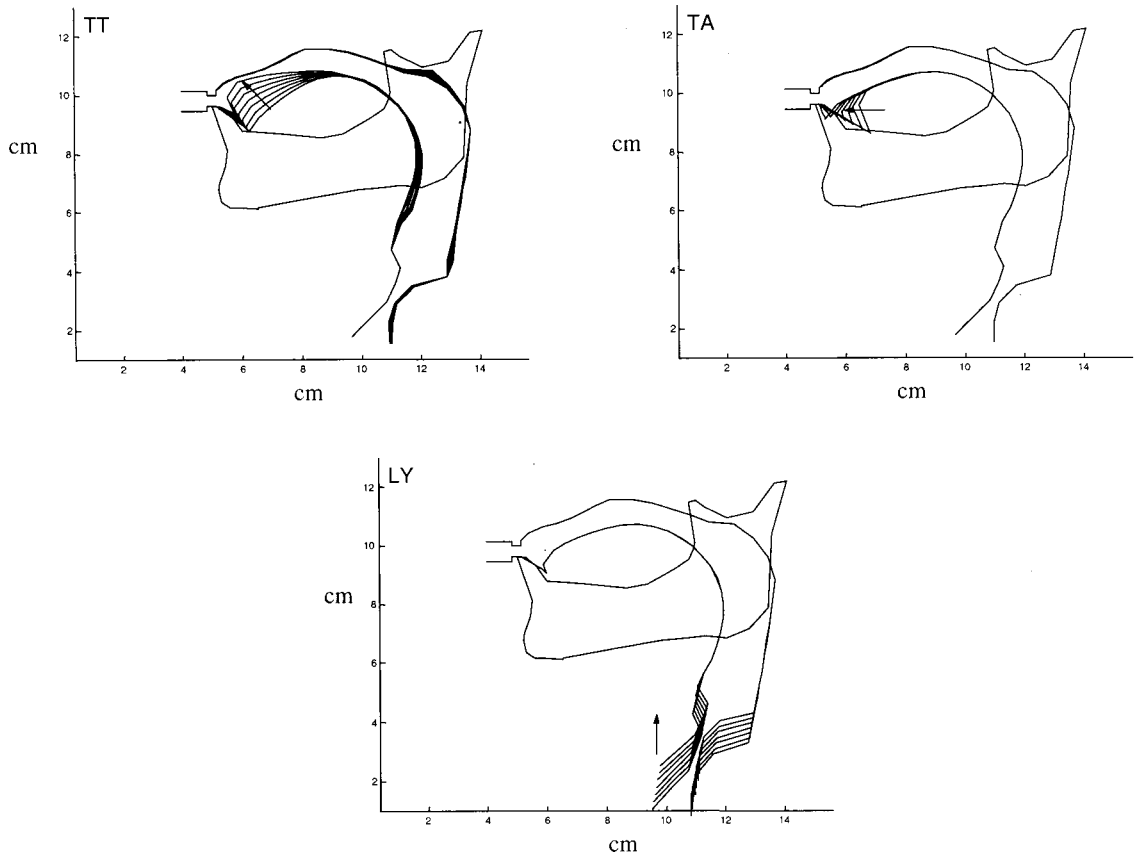


FIG. 8. (*Continued.*)

tion of *JH*, *LH*, *LP*, and *LV*. Figure 8 shows also the articulatory nomograms relating these lip parameters.

B. Midsagittal and area functions—Acoustic models

The midsagittal contour alone is not sufficient to derive the corresponding vocal tract acoustic features. Acoustic vocal tract models are indeed based on tube acoustics, and thus need a description of the vocal tract in terms of *area function*. Most studies devoted to the problem of converting midsagittal distances d to area functions S resulted in solutions based on the ''α,β model'' proposed by Heinz and Stevens (1965), where $S=\alpha\cdot d^{\beta}$: the principle consists in calculating the area of each vocal tract section as a power function of the corresponding midsagittal distance (cf., e.g., Beautemps *et al.*, 1995, for more details). Finally, vocal tract aeroacoustic simulations in the time or in the frequency domain allow the computation of the speech signal or speech acoustic characteristics from the area function (cf., e.g., Maeda, 1982; Badin and Fant, 1984; Mawass *et al.*, 2000).

1. Vocal tract

The midsagittal function represents the sagittal distances between the tongue contour and the outer vocal tract contour along the vocal tract midline, estimated for each section enclosed between two consecutive measurement grid lines. For each section, a quadrilateral can be defined in the midsagittal plane by the intersection points of the tongue contour and of the vocal tract outer contours with the corresponding two lines of the grid. The midsagittal distance for this section is calculated as the surface of the quadrilateral divided by the length of the section. Following Heinz and Stevens (1965), the vocal tract area function is estimated from the midsagittal function. It uses an extended version of a conversion model (Beautemps *et al.*, 1996) optimized for both vowels and consonants: the area function is then derived from the midsagittal function using a polynomial expression where the cross-sectional area S depends on both the midsagittal distance d and the x distance from the glottis measured along the vocal tract midline:

$$S(x,d)=\alpha_1(x)\cdot d+\alpha_2(x)\cdot d^{1.5}+\alpha_3(x)\cdot d^{2}+\alpha_4(x)\cdot d^{2.5}.$$

The $\alpha_i(x)$ functions are expressed as Fourier series, up to the third order, of $\pi.x/l_{\text{tot}}$, where the l_{tot} is the vocal tract length (including the lips):

$$\alpha_i(x)=a_{i0}(x)+\sum_{n=1}^{3} a_{in}\cdot\cos\left(n\frac{\pi}{l_{\text{tot}}}x\right)+\sum_{n=1}^{3} b_{in}\cdot\sin\left(n\frac{\pi}{l_{\text{tot}}}x\right).$$

The values of the Fourier coefficients (altogether, 28 parameters) were optimized so as to minimize, for the N selected configurations, the χ^2 distance between the four formants F_{ik} computed from the area function derived from the synthesized contours and the formants F_{ik}^{c} measured on the acoustic signal of the original data:

$$\chi^2=\sum_{i=1}^{N}\sum_{k=1}^{4}\frac{(F_{ik}^{c}-F_{ik})^2}{F_{ik}^{c}}.$$

At first, the optimization procedure was applied to a restricted set of eight vowels [aɛeiyuoø], in order to ensure an easy convergence. The results were then refined by applying

the same optimization procedure to the whole corpus, excluding only the data for which measurement of the four formants was not possible, i.e., excluding the 347 configurations mainly associated with occlusive consonants [pbdg].

Dang and Honda (1998) developed a similar polynomial decomposition where the coefficients, a function of the distance from the glottis, are determined by minimizing the difference between the estimated and MRI-based area functions for five Japanese vowels. The present procedure, developed before any 3-D vocal tract data were available for the subject, does not make use of 3-D data. However, Badin *et al.* (1998) subsequently acquired 3-D MRI data allowing the direct determination of both midsagittal contours and area functions for a set of vowel articulations for the same subject. These data have therefore been used to assess the quality of this algorithm: for the ten French vowels [aɛeiyuoɔøœ] the comparison between the areas directly estimated from the 3-D measurements (excluding the larynx region and the lips for which no MRI data were available) and those computed by the present procedure from the midsagittal contours estimated from the 3-D measurements has revealed a global root mean square (rms) error value lower than 0.6 cm^2. The relatively important rms error (more than 1 cm^2) observed in the low pharyngeal region for [aɔøœ] is probably due to the whispered production mode used to maintain the articulation during the long 3-D MRI data recording duration whose main consequence is a more constricted tongue in the back region (cf. Matsuda and Kasuya, 1999). This implies a decrease of the cross-sectional areas between the glottis and the epiglottis and probably modifies the relation between the midsagittal distances and the related area functions. Finally, in the uvular region, the small midsagittal distance measured is not representative of the entire cross-section, due to the fact that the main part of the velum body is concentrated in the midsagittal plane with free air flow on both sides. The consequence is an underestimation of the area inherent to the conversion model.

The fit between measured and reconstructed data was also assessed at the level of midsagittal functions. The midsagittal functions of the synthesized vocal tract contours have thus been compared to the midsagittal functions of the original data calculated with the measurement gridline system implemented in the model. The rms errors of the length and of the midsagittal distances are almost zero except for the larynx region where the errors can respectively reach 0.1 and 0.3 cm, probably due to the poor modeling of the tongue in this region (see Fig. 4). A maximum of 0.2 cm for the rms error on the midsagittal distances is also noted in the front part of the vocal tract between the last tongue point and the teeth, due to the fact that the sublingual cavity is not taken into account in the model.

2. Lip area

A possibility for computing lip area *S_mod* is the relation established for the first time by Fromkin (1964):

$$S\_mod = pred\_S\_A\_B \cdot A \cdot B,$$

where *A* and *B* are, respectively, the intra-oral lip width and height measured from the video labio-film. For the present subject, a value of 0.80 was found for *pred_S_A_B* by linear regression applied to the whole corpus. However, the lack of accuracy of *A_mod* resulted in a poor modeling of small areas. This method was therefore abandoned, and lip area was in practice calculated as a second order multilinear regression of the *JH*, *LH*, and *LP* components, for which the coefficients were optimized as to obtain the best fit to the lip area measured on video front pictures. With this modeling, we obtained 0.2 cm^2 for the rms error.

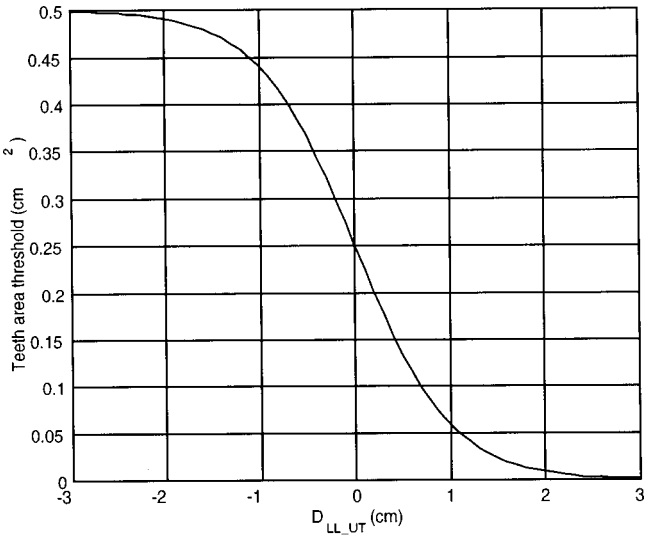


FIG. 9. Minimum threshold for the area at the incisors as a function of difference between lower lip position and upper incisor edge.

3. Acoustic effect of LV

In the absence of the horizontal jaw control parameter *JA*, only vertical movements are taken into account for the lower incisors, i.e., through *JH*. Therefore, there is no straightforward provision in the model for producing labio-dental constrictions by a combination of jaw retraction and lower lip elevation movements, which is the standard articulatory strategy for producing labio-dental fricatives.

This problem is overcome by a mechanism that uses the *LV* parameter for the production of the labio-dental constriction at the incisor section. The incisor section area is made an indirect function of lower lip vertical position, and thus of *LV*, by limiting it to a minimum threshold value function of the difference between lower lip position and upper incisor edge (see Fig. 9). This allows *LV* to be audible, i.e., to have acoustic consequences, at least in circumstances typical of labio-dentals where the lower lip has to be higher than the upper incisor edge in order to produce the proper constriction. This feature was particularly useful for the inversion of the articulatory-to-acoustic relation for fricatives (Mawass *et al.*, 2000).

4. Acoustic model

Acoustic transfer functions as well as formants and bandwidths were determined from these area functions by means of a frequency domain vocal tract acoustic model (Badin and Fant, 1984). A time domain reflection-type line analogue (Bailly *et al.*, 1994), extended to include improved

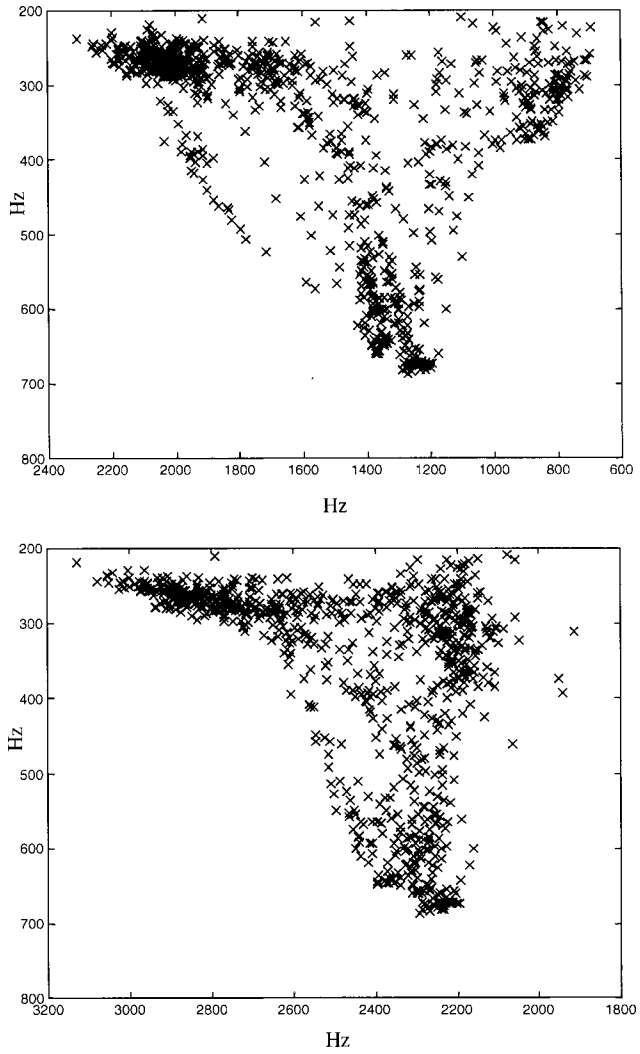


FIG. 10. Predicted $F2/F1$ (top) and $F3/F1$ (bottom) formant spaces (in Hz) for the vowels and consonants pooled.

voice (Pelorson *et al.*, 1996) and noise source models (Badin *et al.*, 1995), can also be driven by these area functions, in association with lung pressure and vocal cords parameters, to produce high quality articulatory synthesis (cf Mawass *et al.*, 2000).

The area functions and the derived formants have been computed for the assessment of the midsagittal to area function conversion, starting from the synthesized midsagittal contours. The rms error and the rms relative error on formants have both been calculated for the whole corpus (excluding 347 configurations for which the measurement of the four formants was not possible): 45 Hz (12.86%), 100 Hz (7.32%), 162 Hz, (6.47%), and 173 Hz (5.21%), respectively, for $F1$, $F2$, $F3$, and $F4$. The mean differences between the formants obtained by the model and those measured are 14 Hz, −65 Hz, and 26 Hz, respectively, for $F2$, $F3$, $F4$ (no significant difference was found for $F1$, except 27 Hz for the vowels).

The predicted maximal formant spaces are comparable to the measured ones (cf. Fig. 2 and Fig. 10). However, the computed formant $F1$ of [a] is about 34 Hz too low.

When modeling the area function of the four point vowels of their two American subjects, Baer *et al.* (1991) reported a deviation (in terms of the rms of the relative error) of 13%, 31%, and 13% on the measures for, respectively, $F1$, $F2$, and $F3$; these deviations are noticeably higher than those in the present study. From a set of 16 disyllabic utterances [hə'CV] and one sentence, Mermelstein (1973) obtained 10.3%, 4.9%, and 5.5% for the average absolute error on $F1$, $F2$, and $F3$; in terms of rms, these errors are still lower. Mermelstein's fits are thus clearly better than ours; however, it should be recalled that they were obtained on a much more restricted set of data.

V. DISCUSSION AND PERSPECTIVES

A. Summary

A linear component analysis of tongue contours and articulatory measures extracted from cineradio- and labio-films made on a reference subject revealed a relatively small number of degrees of freedom. The jaw appears, for the subject studied, to have mainly two degrees of freedom, related to the lower incisors vertical and horizontal movements. However, only the vertical component exerts a significant effect on tongue shape. The residue of tongue shape, once the contribution of the jaw has been removed, possesses four degrees of freedom: tongue body, tongue dorsum, tongue tip, and tongue advance. An extra parameter takes into account the larynx height variance. Similarly, the lip shape possesses, in addition to the jaw contribution, three degrees of freedom: lip protrusion, lip height, and lip vertical elevation. These nine parameters are mostly independent of each other (cf. Table IV), except for *LY* that is correlated with lips and tongue parameters and the correlation between *LP* and *TD*. These degrees of freedom are specific to the vocal tract and articulators of one subject uttering one specific corpus in one language. The corpus was designed to include as many French vowels and consonants as possible. A linear articulatory model was developed based on these data; it explains 96% of the tongue data variance, with an rms reconstruction error of about 0.09 cm. It was complemented by a model converting the midsagittal contours to an area function based on a fitting of midsagittal functions and formant frequencies. Finally this model allows the calculation of formants with rms errors of 45 Hz for $F1$, 100 Hz for $F2$, and 162 Hz for $F3$ over the corpus. To the knowledge of the authors, no such comprehensive model has been developed so far; most of the available models deal with vowels only, while others do not include acoustics.

B. Choice of subject and corpus

The development of such an articulatory-acoustic model based on a specific reference subject was motivated by the need for a model that could fit a real subject's midsagittal profiles of French fricative consonants, plosives, and vowels, as well as formants, with a fairly high degree of accuracy for a large number of configurations. The possibility now exists to investigate in detail the articulatory strategies employed by the subject, and in particular coarticulatory strategies (cf. Mawass *et al.*, 2000, or Vilain *et al.*, 1998). One may argue that no general conclusions may be drawn from such studies, as they are supported by one single subject's data. However,

we were very much aware of the risk of blurring out clear individual articulatory strategies employed by individual subjects when merging together several subjects' data, and therefore made the choice of a single subject for the present study. Similar analyses are under way for other subjects, in order to determine which features may be considered as general and which ones as more subject-specific (Bailly *et al.*, 1998; Vilain *et al.*, 1998; Engwall and Badin, 1999). These studies will also allow us to investigate the number of degrees of freedom of the jaw involved in speech.

The influence of the number of items used for the linear analysis was studied for the present subject by Badin *et al.* (1998); they found that, by choosing the contour samples, i.e., by selecting only vowel and consonant targets in the initial corpus, an articulatory model was produced that represented the whole corpus data with an accuracy close to that obtained when the full model based on the whole corpus was used. More specifically, they showed that the data reconstruction error, computed as the rms error of the abscissa of the tongue contour along each grid line for the 1222 images of the available corpus of midsagittal contours, was 0.09 cm, 0.11 cm, and 0.17 cm when the model was elaborated using, respectively, 1222, 20, and 8 configurations. This justifies the elaboration of models from a much lower number of articulations, and thus, in particular, the use of MRI images instead of x-ray images.

C. Comparison of the degrees of freedom found in other studies

Degrees of freedom are clearly subject and corpus-dependent but their number and their definition are closely related to the method used to explain the whole data set variance. The linear component analysis used by Maeda (1979, 1990) is sometimes referred to as a two-way factor analysis of the variance, where one mode corresponds to the predictors and the other one to the matrix of coefficients of the linear combinations. Using this principle, Maeda (1979) extracted one loading factor for the jaw, and three for the residual midsagittal tongue data to explain 98% of the variance for a corpus made of 400 frames of $[pV_1CV_2]$ ([aiu] and [dg]) sequences uttered by one subject. In an extended corpus of 519 frames corresponding to 10 French sentences, three supplementary components were obtained for the lips including the frontal lip-opening shapes, and four tongue degrees of freedom explaining 88% of the variance (Maeda, 1990). Finally, for these data, Sanguineti *et al.* (1998) imposed two degrees of freedom for the jaw (protrusion and rotation), and one for the larynx, and obtained three other degrees of freedom for the tongue residue from a similar analysis in the so-called λ-space.

The PARAFAC method (Harshman *et al.*, 1977; Nix *et al.*, 1996; Hoole, 1999) is a three-way factor generalization where the third mode corresponds to linear coefficients that account for differences between subjects. Harshman *et al.* (1977) derived two components for the description of representative midsagittal tongue and lip shapes of ten English vowels uttered by five subjects. Hoole (1999) proposed a two-factor PARAFAC model of the German vowels in a symmetrical stop consonant context, plus an additional PCA component to capture the subject-specific nonvocalic behavior of the tongue. More recently, Hoole *et al.* (2000) extracted a two-factor PARAFAC solution that explained 90% of the variance from a set of MRI midsagittal tongue contours measured during the production of seven German vowels by nine speakers. The first component captured the dimension low-back to high-front, and the second was associated with the mid-front to high-back motion. The complex effect of the first component can be decomposed in a co-variation of *JH* and *TB*, the second component being related to *TD*. Ultimately, and to the knowledge of the authors, no analysis based on the PARAFAC method has been realized with an imposed jaw component. To conclude, no analysis based on PARAFAC principles has shown success in explaining the large and phonetically varied data obtained from multiple speakers.

Bailly *et al.* (1998) studied the synergy between tongue and jaw for three subjects, including the present reference subject. They found that the two other subjects used a fairly strong synergy: the amplitude of the tongue movements measured at the tip and at the root that were correlated with jaw movements were about twice as large as might be expected from the simple mechanical carrying effect of the jaw. In other words, the jaw and the tongue shared the execution of the tongue movements. The present subject does not use this synergy: the tongue is not so active, and appears to be passively carried by the jaw. However, all three subjects' articulators had qualitatively the same degrees of freedom. This synergy is still a crucial issue for understanding coarticulation strategies.

D. Perspectives

The principles of this work have been duplicated for the modeling of Swedish midsagittal tongue shapes (Engwall and Badin, 1999). Over 90% of the variance is explained by the four tongue degrees of freedom *JH*, *TB*, *TD*, and *TT*.

One of the main issues in the analysis of speech degrees of freedom is the possibility to build a linear articulatory model that takes into account the redundant feature of the articulators shapes. For instance, it can help to reconstruct complete tongue shapes from a reduced number of articulatory measurement points, such as those provided by electromagnetic articulometry. Badin *et al.* (1997) used the present model to retrieve, from one coil on the lower incisor and three coils on the tongue of the reference subject, the tongue shape as well as the articulatory control parameters *JH*, *TB*, *TD*, *TT*, and *TA* with a fairly good accuracy. This may be useful for investigating speech coarticulation and synergetic strategies (cf., e.g., Vilain *et al.*, 1998), for testing hypotheses of the Frame/Content concept in the child's language development (Vilain *et al.*, 1999), or evaluating the adaptability of speech articulation to various linguistic tasks and environmental conditions such as changes illustrated by the Lombard reflect (Beautemps *et al.*, 1999).

The present articulatory-acoustic model can also be used to derive, by inversion, articulatory control parameters from formants measured in other utterances produced by the same subject (Mawass *et al.*, 2000). These data, in conjunction

with aerodynamic data obtained for the same subject, have been used for the articulatory synthesis of French fricatives (Mawass *et al.*, 2000).

Another extension of the present study is the third dimension. 3-D MRI images have been recorded for the same subject, and a 3-D linear articulatory model is being developed according to the same approach (Badin *et al.*, 1998, 2000); the new model has been elaborated in such a way that part of its control parameters are identical with those of the present midsagittal model, which opens the possibility of inheriting knowledge already acquired for the midsagittal plane, while acquiring new features such as the capability of producing lateral consonants. Finally, the modeling of the velum from MRI midsagittal data should complement the present model.

ACKNOWLEDGMENTS

This work has been partially funded by the European Community (ESPRIT/BR project *Speech Maps* No. 6975), and by the Rhône-Alpes Agency for Social and Human Sciences (ARASSH) (project ''A Virtual Talking Head: Data and models in speech production''). It has benefited from the valuable help of many people to whom the authors are very much indebted: Bernard Gabioud (who initiated a part of this work in the framework of the *Speech Maps* project), Tahar Lallouache (for the lip measurements), Shinji Maeda (for the initial version of the contour edition program, and more importantly for having largely inspired this work), Gilbert Brock, Péla Simon, and Jean-Pierre Zerling (for their expertise on cineradiography), Agnes Hennel (for access to the cineradiography equipment at the Strasbourg Schiltigheim Hospital), Thierry Guiard-Marigny and the late Christian Benoît (for their help on data gathering and processing), Christian Abry (for many stimulating discussions), Marija Tabain (for polishing our French English), as well as many other colleagues at ICP, Grenoble. We have also greatly appreciated the pertinent comments and careful editorial advice of Anders Löfqvist and two anonymous reviewers.

[1]Note that *TngTip* is identical to the last point of the tongue contour abscissa vector.

Abry, C., Badin, P., and Scully, C. (**1994**). ''Sound-to-gesture inversion in speech: The *Speech Maps* approach,'' in *Advanced Speech Applications*, edited by K. Varghese, S. Pfleger, and J. P. Lefèvre (Springer, Berlin), pp. 182–196.

Badin, P., Baricchi, E., and Vilain, A. (**1997**). ''Determining tongue articulation: from discrete fleshpoints to continuous shadow,'' in *Proceedings of the 5th EuroSpeech Conference* (University of Patras, Wire Communication Laboratory, Patras, Greece), Vol. 1, pp. 47–50.

Badin, P., Mawass, K., and Castelli, E. (**1995**). ''A model of frication noise source based on data from fricative consonants in vowel context,'' in *Proceedings of the 13th International Congress of Phonetic Sciences*, edited by K. Elenius and P. Branderud (Arne Strömbergs Grafiska Press, Stockholm, Sweden), Vol. 2, pp. 202–205.

Badin, P., Bailly, G., Raybaudi, M., and Segebarth, C. (**1998**). ''A three-dimensional linear articulatory model based on MRI data,'' in *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis* (Jenolan Caves, Australia), pp. 249–254.

Badin P., Motoki, K., Miki, N., Ritterhaus, D., and Lallouache, T. M. (**1994a**). ''Some geometric and acoustic properties of the lip horn,'' J. Acoust. Soc. Jpn. (E) **15**, 243–253.

Badin, P., Borel, P., Bailly, G., Revéret, L., Baciu, M., and Segebarth, C. (**2000**). ''Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images,'' in *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling* (Kloster Seeon, Germany), pp. 261–264.

Badin, P., and Fant, G. (**1984**). ''Notes on vocal tract computation,'' Speech Transmission Laboratory—Quarterly Progress Status Report Vol. 2-3/1984, pp. 53–108.

Badin, P., Shadle, C. H., Pham Thi Ngoc, Y., Carter, J. N., Chiu, W., Scully, C., and Stromberg, K. (**1994b**). ''Frication and aspiration noise sources: contribution of experimental data to articulatory synthesis,'' in *Proceedings of the 3rd International Conference on Spoken Language Processing* edited by Mike Edington (Yokohama, Japan), Vol. 1, pp. 163–166.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). ''Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels,'' J. Acoust. Soc. Am. **90**, 799–828.

Bailly, G. (**1993**). ''Resonances as possible representations of speech in the auditory-to-articulatory transform,'' in *Proceedings of the 3rd Eurospeech Conference on Speech Communication and Technology* (Berlin), Vol. 3, pp. 1511–1514.

Bailly, G., Badin, P., and Vilain, A. (**1998**). ''Synergy between jaw and lips/tongue movements: Consequences in articulatory modelling,'' in *Proceedings of the 5th International Conference on Spoken Language Processing*, edited by R. H. Mannell, J. Robert-Ribes, and E. Vatikiotis-Bateson (Australian Speech Science and Technology Association, Inc., Sydney, Australia), Vol. 5, pp. 1859–1862.

Bailly, G., Castelli, E., and Gabioud, B. (**1994**). ''Building prototypes for articulatory speech synthesis,'' in *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis* (New York), pp. 9–12.

Beautemps, D., Badin, P., and Laboissière, R. (**1995**). ''Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data,'' Speech Commun. **16**, 27–47.

Beautemps, D., Borel, P., and Manolios, S. (**1999**). ''Hyper-articulated speech: Auditory and visual intelligibility,'' in *Proceedings of the 6th European Conference on Speech Communication and Technology*, Vol. 1 (Budapest, Hungary), pp. 109–112, September 1999.

Beautemps, D., Badin, P., Bailly, G., Galván, A., and Laboissière, R. (**1996**). ''Evaluation of an articulatory-acoustic model based on a reference subject,'' in *Proceedings of the 4th Speech Production Seminar* (Autrans, France), pp. 45–48.

Boë, L. J., Gabioud, B., Schwartz, J. L., and Vallée, N. (**1995**). ''Towards the unification of vowel spaces,'' in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, edited by K. Elenius and P. Branderud (Arne Strombergs Grafiska Press, Stockholm, Sweden), Vol. 4, pp. 582–585.

Bothorel, A., Simon, P., Wioland, F., and Zerling, J. P. (**1986**). ''Cinéradiographie des voyelles et consonnes du francais [Cineradiography of vowels and consonants in French],'' Trav. de l'Inst. de Phonétique de Strasbourg, 296 pp.

Coker, C., and Fujimura, O. (**1966**). ''Model for specification of the vocal-tract area function,'' J. Acoust. Soc. Am. **40**, 1271.

Dang, J., and Honda, K. (**1998**). ''Speech production of vowel sequences using a physiological articulatory model,'' in *Proceedings of the 5th International Conference on Spoken Language Processing*, edited by Robert H. Mannell and Jordi Robert-Ribes, Vol. 5 (Sydney, Australia R.H.), pp. 1767–1770.

Demolin, D., Giovanni, A., Hassid, S., Heim, C., Lecuit, V., and Soquet, A. (**1997**). ''Direct and indirect measurements of subglottic pressure,'' *Proceedings of Larynx 97* (Marseille, France), pp. 69–72.

Djéradi, A., Guérin, B., Badin, P., and Perrier, P. (**1991**). ''Measurement of the acoustic transfer function of the vocal tract: a fast and accurate method,'' J. Phonetics **19**, 387–395.

Edwards, J., and Harris, K. S. (**1990**). ''Rotation and translation of the jaw during speech,'' J. Speech Hear. Res. **33**, 550–562.

Engwall, O., and Badin, P. (**1999**). ''Collecting and analyzing two- and three-dimensional MRI data for Swedish,'' Tal Musik Hörsel, Quarterly Progress Status Report, Stockholm Vol. 3-4, pp. 11–38.

Fowler, C. A., and Saltzman, E. (**1993**). ''Coordination and coarticulation in speech production,'' Language and Speech **36**, 171–195.

Fromkin, V. A. (**1964**). ''Lip positions in American English vowels,'' Language and Speech **7**, 215–225 .

Gabioud, B. (**1994**). ''Articulatory models in speech synthesis,'' in *Funda-*

mentals of Speech Synthesis and Speech Recognition, edited by E. Keller (Wiley, Chichester), pp. 215–230.
Harshman, R., Ladefoged, P., and Goldstein, L. (**1977**). ''Factor analysis of tongue shape,'' J. Acoust. Soc. Am. **62**, 693–707.
Heinz, J. M., and Stevens, K. N. (**1965**). ''On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech,'' *Proceedings of the Fifth International Congress of Acoustics* (Liége, Belgium), Paper A44.
Hoole, P. (**1999**). ''On the lingual organization of the German vowel system,'' J. Acoust. Soc. Am. **106**, 1020–1032.
Hoole, P., and Kroos, C. (**1998**). ''Control of larynx height in vowel production,'' in *Proceedings of the 5th International Conference on Spoken Language Processing,* edited by R. H. Mannell, J. Robert-Ribes, and E. Vatikiotis-Bateson (Australian Speech Science and Technology Association, Inc., Sydney, Australia), Vol. 2, pp. 531–534.
Hoole, P., Wismüller, A., Leisinger, G., Kroos, C., Geumann, A., and Inoue, M. (**2000**). ''Analysis of tongue configuration in multi-speaker, multi-volume MRI data,'' in *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling* (Kloster Seeon, Germany), pp. 157–160.
Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (**1986**). ''The dynamical theory of speech production: Data and theory,'' J. Phonetics **14**, 29–60.
Laboissière, R., Ostry, D. J., and Feldman, A. G. (**1996**). ''Control of multi-muscle systems: Human jaw and hyoid movements,'' Biol. Cybern. **74**, 373–384.
Lallouache, M. T. (**1990**). ''Un poste Visage-Parole. Acquisition et traitement de contours labiaux [A ''Face-Speech'' workstation. Acquisition and processing of labial contours],'' *Proceedings of the 18th Journées d'Etude sur la Parole* (Montréal, Canada), pp. 282–286.
Liljencrants, J. (**1971**). ''A Fourier series description of the tongue profile,'' Speech Transmission Laboratory—Quarterly Progress Status Report Vol. 4/1971, pp. 9–18.
Lindblom, B. E. F. and Sundberg, J. E. F. (**1971**). ''Acoustical consequences of lip, tongue and jaw movements,'' J. Acoust. Soc. Am. **50**, 1166–1179.
Lindblom, B. E. F., Lubker, J., and Gay, T. (**1979**). ''Formant frequencies of some fixed-mandible vowels and a model of speech-motor programming by predictive simulation,'' J. Phonetics **7**, 141–161.
Maeda, S. (**1979**). ''Un modèle articulatoire de la langue avec des composantes linéaires,'' *Proceedings of the 10th Journées d'Etude sur la Parole* (Grenoble, France), pp. 152–163.
Maeda, S. (**1982**). ''A digital simulation method of the vocal tract system,'' Speech Commun. **1**, 199–299.
Maeda, S. (**1990**). ''Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model,'' in *Speech Production and Speech Modeling,* edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, The Netherlands), pp. 131–149.
Maeda, S. (**1991**). ''On articulatory and acoustic variabilities,'' J. Phonetics **19**, 321–331.
Matsuda, M., and Kasuya, H. (**1999**). ''Acoustic nature of the whisper,'' in *Proceedings of the 6th European Conference on Speech Communication and Technology,* Vol. 1 (Budapest, Hungary), pp. 133–136, September 1999.
Mawass, K., Badin, P., and Bailly, G. (**2000**). ''Synthesis of French fricatives by audio-video to articulatory inversion,'' Acta Acoustica **86**, 136–146.
Mermelstein, P. (**1973**). ''Articulatory model for the study of speech production'' J. Acoust. Soc. Am. **53**, 1070–1082.
Nix, D. A., Papcun, G., Hogden, J., and Zlokarnik, I. (**1996**). ''Two cross-linguistic factors underlying tongue shapes for vowels,'' J. Acoust. Soc. Am. **99**, 3707–3717.
Ostry, D., Vatikiotis-Bateson, E., and Gribble, P. (**1997**). ''An examination of the degrees of freedom of human jaw motion in speech and mastication,'' J. Acoust. Soc. Am. **40**, 1341–1351.
Payan, Y., and Perrier, P. (**1997**). ''Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis,'' Speech Commun. **22**, 185–205.
Pelorson, X., Hirschberg, A., Wijnands, A. P. J., Bailliet, H., Vescovi, C., and Castelli, E. (**1996**). ''Description of the flow through the vocal cords during phonation. Application to voiced sounds synthesis,'' Acta Acoustica **82**, 358–361.
Perkell, J. S. (**1974**). ''A physiological-oriented model of the tongue activity during speech production,'' Ph.D. dissertation, MIT, Cambridge.
Perkell, J. S. (**1991**). ''Models, theory and data in speech production,'' *Proceedings of the XIIth International Congress of Phonetic Sciences* (Université de Provence, Aix-en-Provence, France), Vol. 1, pp. 182–191.
Perrier, P., Ostry, D. J., and Laboissière, R. (**1996**). ''The equilibrium point hypothesis and its application to speech motor control,'' J. Speech, Language, and Hear. Res. **39**, 365–578.
Perrier, P., Payan, P., Perkell, J. S., Zandipour, M., Pelorson, X., Coisy, V., and Matthies, M. (**2000**). ''An attempt to simulate fluid-walls interactions during velar stops,'' in *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling* (Kloster Seeon, Germany), pp. 149–152.
Revéret, L., and Benoît, C. (**1998**). ''A new 3D lip model for analysis and synthesis of lip motion in speech production,'' in *Proceedings of the International Conference on Auditory-Visual Speech Processing/Second ESCA ETRW on Auditory-Visual Speech,* edited by D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson (Terrigal-Sydney, Australia), pp. 207–212.
Sanguineti, V., Laboissière, R., and Ostry, D. J. (**1998**). ''A dynamic biomechanical model for neural control of speech production,'' J. Acoust. Soc. Am. **103**, 1615–1627.
Scully, C. (**1991**). ''The representation in models of what speakers know,'' in *Proceedings of the XIIth International Congress of Phonetic Sciences* (Université de Provence, Aix-en-Provence, France), Vol. 1, pp. 192–197.
Shadle, C. H., and Scully, C. (**1995**). ''An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences,'' J. Phonetics **23**, 53–66.
Sorokin, V. N., Gay, T., and Ewan, W. G. (**1980**). ''Some biomechanical correlates of jaw movements,'' J. Acoust. Soc. Am. Suppl. 1 **68**, S32.
Stark, J., Lindblom, B., and Sundberg, J. (**1996**). ''APEX: An articulatory synthesis model for experimental and computational studies of speech production,'' TMH-QPSR 2/1996, pp. 45–48.
Stetson, R. H. (**1928**). ''Motor phonetics. A study of speech movements in action,'' Archives Néerlandaises de Phonétique Expérimentale **3**, 216.
Stromberg, K, Scully, C., Badin, P., and Shadle, C. H. (**1994**). ''Aerodynamic patterns as indicators of articulation and acoustic sources for fricatives produced by different speakers,'' Institute of Acoustics **16**, 325–333.
Tiede, M. K. (**1996**). ''An MRI-based study of pharyngeal volume contrasts in Akan and English,'' J. Phonetics **24**, 399–421.
Vilain, A., Abry, C., and Badin, P. (**1998**). ''Coarticulation and degrees of freedom in the elaboration of a new articulatory plant: Gentiane,'' in *Proceedings of the 5th International Conference on Spoken Language Processing,* edited by R. H. Mannell, J. Robert-Ribes, and, E. Vatikiotis-Bateson (Australian Speech Science and Technology Association, Inc., Sydney, Australia), Vol. 7, pp. 3147–3150.
Vilain, A., Abry, C., Badin P., and Brosda, S. (**1999**). ''From idiosyncratic pure frames to variegated babbling: Evidence from articulatory modeling,'' in *Proceedings of the 14th International Congress of Phonetic Sciences,* edited by J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey (Congress organizers at the Linguistics Department, University of California at Berkeley, San Francisco, CA), Vol. 3, pp. 2497–2500.
Westbury, J. R. (**1988**). ''Mandible and hyoid bone movements during speech,'' J. Speech Hear. Res. **31**, 405–416.
Westbury, J. R. (**1994**). ''On coordinate systems and the representation of articulatory movements,'' J. Acoust. Soc. Am. **95**, 2271–2273.
Wilhelms-Tricarico, R. (**1995**). ''Physiological modeling of speech production: Methods for modeling soft-tissue articulators,'' J. Acoust. Soc. Am. **97**, 3085–3098.
Zerling, J. P. (**1984**). ''Phénomènes de nasalité et de nasalisation vocaliques: Étude cinéradiographique pour deux locuteurs [in French],'' Trav. de l'Inst. de Phonétique de Strasbourg **16**, 241–266.

D. Beautemps, M.-A. Cathiard & Y. Leborgne, 2003. Benefit of audiovisual presentation in close shadowing task. In Proceedings of ICPHS conference, Barcellona, Spain.

Benefit of audiovisual presentation in close shadowing task

Denis Beautemps, Marie-Agnès Cathiard and Yvon Le Borgne
Institut de la Communication Parlée, CNRS UMR5009 / INPG/ Université Stendhal
Grenoble, France
E-mail: beautemps@icp.inpg.fr, cathiard@icp.inpg.fr

ABSTRACT

Close shadowing paradigm consists in repetition of speech as soon as the speaker produces it. This paradigm was used for the evaluation of time co-perception and production process. We used this paradigm in order to evaluate the benefit of visual lip information in case of audiovisual presentation of stimuli with a perfectly audible signal. We presented French [VCVsa] sequences made of [a, i, u, y] vowels and [p, t, k] consonants pronounced by a male French speaker for both normally articulated and hyper articulated speech. Four subjects were asked to repeat *on line* syllables for stimuli presented in the two conditions : auditory alone and audiovisually. The average reaction times measured between stimulus onset acoustic closure of the consonant and the corresponding event on the response vary between 227 to 314 ms. We found no advantage for the hyper articulated production. But a significant advantage of audiovisual presentation with a visual gain of 35 ms and 49 ms (i.e. 13.4% and 14.6%) for two of the four subjects.

1. INTRODUCTION

Shadowing definition

"Speech shadowing is an experimental task in which the subject is required to repeat (shadow) speech as he hears it. When the shadower is presented with a sentence, he will start to repeat it before he has heard all of it. The response latency to each word of a sentence can therefore be measured" ([1] p. 252). This paradigm is used to study co-perception and production processing. Reaction time (*RT*) (i.e. response delay) generally observed for isolated words or non-sense syllables were as small as 200 to 250 ms [1]. In a close shadowing task of a 300-words passage of normal prose, Marslen-Wilson ([1]) obtained, for the closest subjects, *RT* ranged from 254 to 287 ms. The analysis of the errors showed that close shadowers process in a same way the syntactic and semantic structure of speech than the distant shadowers, thus revealing the natural processing of speech in shadowing task.

V-V syllable shadowing

Porter and Lubker ([2]) presented Vowel-Vowel synthesized stimuli as [ao], [ai] and [aæ] to four subjects. The first vowel was lengthened up to 1100 - 1500 ms. The authors compared reaction time in three conditions. (i) Subject under phonation of the first vowel was asked to pronounce the second vowel as soon as he perceived it (in average 147.6 ms for *RT*). (ii) Subject in phonation of the first vowel was asked to pronounce the [o] vowel as he perceived the change in vowel (162.4 ms in average for *RT*). (iii) A third one similar to the previous one where no preliminary phonation was used before the [o] response to the vowel change (in average 187.6 ms for *RT*).

In the first condition, subjects had to repeat the second vowel initially not known contrarily to the other conditions where a simple known response to the vowel change was asked. Taking into account that reaction time of the first condition is in average not much different from the others (147.6 ms *vs* 162.4 ms or 187.6 ms), these results showed that the auditory phase needed in the identification of the second vowel (first condition) did not delayed the response production phase. The authors concluded on the following noticeable result that motor control mechanisms involved in the response largely overhead the auditory process.

VCV syllable shadowing

Similar experiments on VCV sequences were performed by Porter and Castellanos [3]. Reaction time to three conditions were tested in the presentation of VCV auditory stimuli made of [a] vowel lengthened to 2 to 5 seconds and [p, b, m, k, g] consonants as for example in [apa] sequence. (i) A shadowing task where subjects were asked to repeat in line (*RT* = 223 ms in average), (ii) subjects under phonation of the first vowel were asked to answer [ba] as they perceived the consonant (*RT* = 171 ms in average), (iii) same as the precedent condition but with no initial phonation (*RT* = 296 ms in average). The second condition where no identification was needed (corresponding to a simple reaction time) gave the quickest reaction times. Once more the authors explained the slight elevation of reaction time (223 ms *vs* 171 ms) observed in the shadowing task by a consequent overhead of the perception process needed in the first condition with the response production mechanism.

Vision and shadowing

In a series of shadowing experiments, Reisberg et al.

ISBN 1-876346-49-3

([4]) evaluated the gain with vision in identification of intact auditory stimuli, i.e. not acoustically degraded. In the repetition of 10 phrases made of 110 words each in foreign languages (French stimuli and English subjects with fair level in French) the authors obtained a gain of 15 % on identification scores with the audiovisual presentation (AV) in comparison to the auditory alone (A). This gain was computed as following: $\left|\frac{AV-A}{A}\right|$. A supplementary shadowing task made of German phrases and English subjects with poor German level was performed to dismiss the hypothesis where the gain with vision is due to a better concentration of the subjects. In the audio condition, the view of the face with mouth and chin being masked was added. Results showed a gain for the audiovisual presentation higher than 21.5% in comparison with the audio alone condition where subjects could not see neither the lips, neither the chin. In a third experiment where English phrases pronounced by a Belgium speaker were repeated by English subjects, the authors obtained a gain of 4% with the vision. In a latter experiment, English students were asked to repeat in English a part of the famous philosophic "Critique de la raison pure" text of Kant, complex for comprehension. The authors also observed a gain with the vision (8%). These latter experiments showed that a shadowing experiment even with familiar vocabulary and syntax leads to better results in audiovisual condition. In this ensemble of experiments, the gain of vision for speech perception was evidenced even if the audio signal is not degraded by noise, nor in conflict with vision. However in these results, this benefit of vision was not based on reaction time measurements but simply on scores of correctly repeated words.

More recently, Davis and Kim [5] tested the visual gain obtained for correct repetition of shadow sentences and for words memorization in a foreign language. They asked to 10 English subjects to repeat Corean sentences after 3 successive presentations in 2 audiovisual conditions: one with only the high part of the face visible and another one with the low part of the face. Note that it is not a close shadowing task but a differed shadowing one since subjects repeat sentences after complete audition. The authors evidenced a benefit of the view of the low part of the face (lip, mandible, teeth and tongue) both for production with 100 ms of advance in *RT* and for memorization scores.

Hyper-articulated speech

In a series of work devoted to speech adaptability, Lindblom [6] explained that speech variability is controlled through negotiation between the listener and the speaker. In his hyper-hypo theory, the information carried by the acoustic signal depends inversely on context: if the context is sufficiently rich, the information in the signal can be poor. In the contrary, when the context gives no information, the speaker can hyper-articulate to increase information in the signal. Beautemps et al. [7] quantified the effect of hyper-articulation. In the identification of plosive consonants acoustically degraded, the authors compared the hyper condition with normal speech. They observed that identification scores decreased with the signal to noise ratio but with a better solidity for the hyper condition. In the hyper-articulation production, it was observed on the acoustic signal a clear control of the frequency distribution of the energy at the release instant in addition to a global volume effect [7, 8].

The present work aimed at evaluation of the benefit of vision and hyper-articulation on reaction times in shadowing task of French unvoiced plosive consonants [p, t, k], for sequences with perfect audible acoustic signal.

2. EXPERIMENTAL SETUP

Speech material:

A French speaker uttering a set of 28 /VCVsV/ sequences was audiovisually recorded in quiet environmental conditions and with a 80 dB SPL white noise presented at both ears for a corpus made of the French [a, i, u, y] vowels and [p, t, k] plosives. The noisy environmental condition was used to encourage the speaker to produce an auditory vocal effort (the so-called « Lombard reflex ») with a natural hyper-articulation. For each of the two environmental conditions the speaker was asked to pronounce the sequence and to repeat it with an emphasis on the first consonant. The perfect identification of the consonant from the quiet and hyper-articulated + emphasis conditions was reported in Beautemps et al. [7, 8].

The set of data involved in the shadowing experiment was made of with the 12 audiovisual stimuli of the quiet condition and normal production (so-called normal speech) and with the 12 audiovisual stimuli of the noisy environmental condition with emphasis on the consonant (so-called hyper-articulated speech). In order to dismiss the possibility to predict the consonant from the initial vowel duration (Reisberg et al. [4]), the initial vowel of each sequence was lengthened a multiple of 40 ms (i.e. image rate) up to successively 2, 3 and 4s. A few periods of the acoustic signal at the vowel center was duplicated using a TDPSOLA technique. The image preceding the cut off point was duplicated as to complete the video sequences. We thus obtained a set of 144 [VCVsV] stimuli (12 sequences x 2 speech conditions x 3 durations of the initial vowel x 2 speech modes (audio *vs* audiovisual)).

Finally, the energy of the acoustic signal of stimuli of the same vowel were normalized to the hyper-condition one.

Stimuli presentation

Four subjects with no known hearing damage and visual injuries were submitted to shadowing of [VCVsV] sequences with audio and audiovisual stimuli. They where asked to repeat as early as possible the [CVsV] part as they were listening and uttering the first vowel. For example, in

ISBN 1-876346-49-3 © 2003 UAB

case of listening [apasa] sequence with the first [a] lengthened up to 2s, 3 or 4s, the subject had to maintain [a] in phonation and to produce [pasa] sequence as soon as he identified the consonant. Stimuli were grouped by vowel and presented randomly in a same session, the information on the vowel being given to the subject. Thus duration of the initial vowel, consonant and speech articulation condition (normal *vs* hyper-articulated) were the unknown parameters.

A monitor placed one meter in front of the subject was used for presentation of the speaker face bottom part (nose, mouth, chin, larynx) in the audiovisual condition. In the audio alone condition, the monitor was switched off. In both conditions, the audio stimuli was presented at the subject through headphones and simultaneously recorded on the stereo first line of an audio DAT tape. Subject response to the stimulus was recorded on the synchronised second line of the same audio DAT tape, thus allowing a further reaction time calculation.

Subjects were preliminary trained with presentation of twelve assorted audiovisual stimuli.

Reaction time calculation

The recorded stimuli and the corresponding audio responses were digitalized in stereo. Reaction Time was derived from the duration between the instant of consonant acoustic closure of the stimulus and the instant of the corresponding acoustic event in the response. The acoustic closure was automatically marked at the point where the energy (integrated through a 10ms window) attained 20% of the energy in the preceding vowel center.

3. RESULTS

Error analysis

Responses for which the consonant was different from the stimulus one or with reaction time superior to 400 ms or containing interruption were considered as errors for the shadowing task and were not considered in reaction time analysis. The average rate error of 6.8 % observed for the responses was much less than the 23 % obtained by Porter & Catellanos [3]. Curiously the audiovisual condition gave much more errors (9 %) than with auditory alone (4 %). The error on the consonant was increased (9 %) in context of the rounded vowels [u] and [y] - context well known for masking visual consonant effect ([9])- in comparison to 3 % for [a] context and 6 % for [i]. For bilabial [p] and dental [t] the error is in average 4 %, in comparison to 12.5 % for the alveolar [k]. Surprisingly, the error attained 9 % in the hyper-articulated speech condition (5 % for the normal speech condition). Subjects did not take benefit of the focus on the consonant. Finally error rates in relation to the duration of the initial vowel attained 10 % in case of 4s against 5 % for 2 and 3s.

Average Reaction time

The average reaction times measured between stimulus onset acoustic closure of the consonant and the corresponding event on the response vary between 227.4 to 314.19 ms. These reaction times are somewhat larger than the 170 to 278 ms ones observed by Porter and Castellanos [3]: our subjects were less trained to the task (12 vs. 30 to 60 stimuli of their training phase). On the other hand we obtained much less identification errors (6.8 % vs. 23 %).

Analysis by subject

For each of the four subjects, we performed a three way ANOVA with 3 factors (consonant factor: [p, t, k]; speech articulation factor: normal and hyper-articulated and presentation: audio and audiovisual). Note that when a stimulus was not correctly repeated (error in the response), the responses were not considered both for the audio and audiovisual conditions.

No significant effect was observed for two subjects. Their average reaction time (RT) was 227.4 ms for CL subject and 296,91 ms for NL subject.

For the slower subject (314.19 ms for reaction time), only the presentation factor was significant ($F(1,57) = 10.19$, $p = 0.0023$), i.e. the audiovisual presentation (287.46 ms) allows faster *RT* than the audio presentation (336.75 ms). Following Reisberg et al. [4] formula, the visual gain was calculated: $\left|\frac{287.46-336.75}{336.75}\right| = 14.6\%$

For subject VA (average *RT* of 244.7 ms), the ANOVA showed the significant effect of presentation ($F(1,57) = 13.54$, $p = 0.0005$) with a significant consonant–presentation interaction $F(2,57) = 4.03$, $p = 0.023$). The *RT* in audiovisual presentation (225.8 ms) are shorter than audio *RT* (260.65 ms). The visual gain was: $\left|\frac{225.8-260.65}{260.65}\right| = 13.4\%$

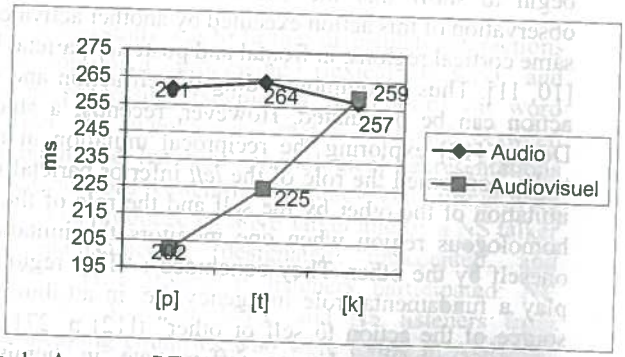


Figure 1: Average *RT* obtained in audiovisual and audio presentations of [p, t, k] consonants for VA subject.

Concerning interaction effect, post-hoc comparisons (Newman-Keuls test) showed an audiovisual benefit for [p] consonant ($p < 0.01$) (Figure 1). We can also observe that

ISBN 1-876346-49-3 © 2003 UAB

the audiovisual *RT* is shorter than audio *RT* for the consonant [t] but this tendency is not significant.

4. CONCLUSION

In this shadowing experiment, contribution of two main factors were evaluated: the hyper vs. normal articulation and the audiovisual vs. audio presentation effect.

No benefit of hyper articulation was obtained. It appeared that articulatory variations reflected in burst energy of the consonants were not exploited by our subjects during the shadowing task. It shall be recalled that the perceptual benefit observed in no shadowing task in beautemps et al. [7, 8] was related to a greater intelligibility robustness in noise. Since our shadowing task was performed with perfectly audible signals, it is possible that these variations could not processed fast enough to be taken into account.

It was observed an advantage of the audiovisual presentation for two of our four subjects with visual gains of 13.4 % and 14.6 %. These gains are relatively high and similar to those of Reisberg et al. [4] (0 to 21.5 %). The originality of these results is that our gain are obtained from reaction time in close shadowing task and not from intelligibility scores. The specific gain obtained for audiovisual [p], by one of our subjects, can be explained by the great visibility of the labial occlusion. Owens & Blazek [9] showed that labial plosives obtained the higher lipreading scores, whatever the vocalic context. We can explain in the same way the tendency to better identify the consonant [t] vs. [k]. However, the effectiveness of audiovisual presentation remains to be confirmed by further experiments with more trained subjects.

More generally, this paradigm of close shadowing seems particularly adapted to understand the cognitive processes implied in the linguistic self-monitoring *vs.* the monitoring of the speech of the other speaker. Many studies begin to show that the execution of an action or the observation of this action executed by another activated the same cortical regions, in frontal and posterior parietal lobes [10, 11]. Thus a common coding of self-action and other action can be postulated. However, recently, a study of Decety [12] exploring the reciprocal imitation in motor tasks, confirmed the role of the *left* inferior parietal in the imitation of the other by the self and the role of the *right* homologous region when one monitors the imitation of oneself by the other. They concluded : "This region may play a fundamental role in agency, i.e. in attributing the source of the action to self or other" ([12] p. 271). Will speech monitoring be as differentiate in hemisphere dominance when our rapid shadowers (hence left) were contrasted with the monitoring of another speaker as a follower of oneself ?

Acknowledgments : This research was supported by a "Jeune équipe" project of the CNRS (French National Research Center).

REFERENCES

[1] W. Marslen-Wilson, "Linguistic structure and speech shadowing at very short latencies," *Nature*, vol. 244, pp. **522-523,** 1973.

[2] R.J. Porter and J.F. Lubker, "Rapid reproduction of vowel-vowel sequences: evidence for a fast and direct acoustic-motoric linkage in speech", Journal of Speech and Hearing Research, 23, pp. 593-602, 1980.

[3] R.J. Porter and Jr. Castellanos, "Speech production measures of speech perception: rapid shadowing of VCV syllables" Journal of the Acoustical Society of America, 57(4), pp. 1349-1356, 1980.

[4] D. Reisberg, J. McLean and A. Goldfield, "Easy to hear but hard to understand: A lip reading advantage with intact auditory stimuli", in B. Dodd and R. Campbell (Eds), Hearing by Eye: The psychology of lip-reading, London, Lawrence Erlbaum Associates, pp. 97–113, 1987.

[5] C. Davis and J. Kim, "Repeating and remembering foreign language words: Does seing help?", In D. Burnham, J. Robert-Ribès and E. Vatikiotis-Bateson (Eds.), AVSP'98: International Conference on Audio-Visual Speech Processing, 121-125, Terrigal, Australia, 4-7 Dec. 1998.

[6] B. Lindblom, "Adaptive variability and absolute constancy in speech signals: Two themes in the quest for phonetic invariance", in PERILUS, 5, pp. 2-20, 1986.

[7] D. Beautemps, P. Borel, S. Manolios, "Hyper-articulated speech", in proceedings *of the 6th European Conference on Speech Communication and Technology* (Budapest), Vol.1, 109-112, 1999.

[8] D. Beautemps. « Parole hyper-articulée : données et analyses acoustiques pour des plosives en français » *Actes des XXIIIèmes Journées d'Etude sur la Parole*, 437-440, 2000.

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

[10] J. Decety, N. Grèzes, D. Costes, M. Perani, E. Jeannerod, F. Procyk, F. Grassi and F. Fazio, "Brain activity during observation of actions. Influence of action content and subject's strategy", Brain, 20, 1763-1777, 1997.

[11] G. Rizolatti, L. Fadiga, M. Matelli, V. Bettinardi, E. Paulesu, D. Perani and F. Fazio, "Localization of grasp representations in humans by PET. 1. Observation versus execution", Experimental brain Research, 111, 246-252, 1996.

[12] J. Decety, T. Chaminade, J. Grèzes and A.N. Meltzoff, "A PET exploration of the neural mechanisms involved in reciprocal imitation", NeuroImage, 15, 265-272, 2002.

ISBN 1-876346-49-3

N. Aboutabit, Denis Beautemps & Laurent Besacier, 2007. Automatic identification of vowels in the Cued Speech context. In International Conference on Auditory-Visual Speech Processing (AVSP), 2007.

Automatic identification of vowels in the Cued Speech context

Noureddine Aboutabit[1], *Denis Beautemps*[1], *Laurent Besacier*[2]

[1]Grenoble Images Parole Signal Automatique, département Parole & Cognition
46 Av. Félix Viallet, 38031 Grenoble, cedex 1, France

[2]Laboratoire d'Informatique de Grenoble, UMR 5217 - 681 rue de la passerelle - BP 72 - 38402 Saint Martin d'Hères, France

Noureddine.aboutabit@gipsa-lab.inpg.fr

Abstract

The phonetic translation of Cued Speech (CS) (Cornett [1]) gestures needs to mix the manual CS information together with the lips, taking into account the desynchronization delay (Attina et al. [2], Aboutabit et al. [3]) between these two flows of information. The automatic coding of CS hand positions and lip targets (Aboutabit et al. [3], Aboutabit et al. [4]) are thus a key factor in the mixing process. This contribution focuses on the identification of vowels by merging CS hand positions and vocalic lip information produced by a CS speaker. The hand flow is coded automatically as plateaus between transition phases. A plateau is defined as the interval during which the hand is maintained at a specific CS hand position. A transition is the interval during which the hand moves from a specific CS hand position to another one. The CS hand position is automatically obtained as the result of the hand 2d-coordinates Gaussian classification. The instants of reached hand targets are used as reference instants to define the interval inside which the lip target instant of the vowel is automatically detected. The lip parameters extracted at this instant are processed in a Gaussian classifier as to identify the vocalic lip feature of the vowel. The vowel is obtained as the result of the combination of the corresponding hand position and the lip feature. The global performance of the method attains 77.6% as correct identification score. This result does not take into account the CS coding errors. This result has to be compared with the global 83.5% score of speech perception by deaf people using CS (Nichols and Ling, 1982 [6].

Index Terms: Cued Speech production, lip target segmentation, vocalic lip classification, and CS gesture segmentation.

1. Introduction

The Cued Speech (CS) (Cornett, 1967 [1]) is a manual cues system used to disambiguate the lip-reading and enhance speech perception from visual input by deaf and impaired-hearing people. In this system, the speaker moves the hand in close relation with speech (see Attina et al., 2004 [2] for a detailed study on CS temporal organization). The hand (with the back facing the perceiver) is a cue that uniquely determines a phoneme when associated with the corresponding lip shape. A manual cue in this system is made up of two components: the shape of the hand and the hand position relative to the face. Handshapes are designed to distinguish among consonants and hand positions among vowels. A single manual cue corresponds to phonemes that can be discriminated with lip shapes, while phonemes with identical lip shapes are coded with different manual cues (see figure 1 which describes the complete system for French).

In the framework of communication between hearing and hearing impaired people, the automatic translation of CS components into a phonetic chain is a key issue. Due to the CS system, both hand and lip flows produced by the CS speaker carry a part of the phonetic information. Thus the recovering of the complete phonetic information needs to constrain the process of each flow by the other one (see Aboutabit et al., 2006 [3] for an example of a complete analysis of the hand flow).

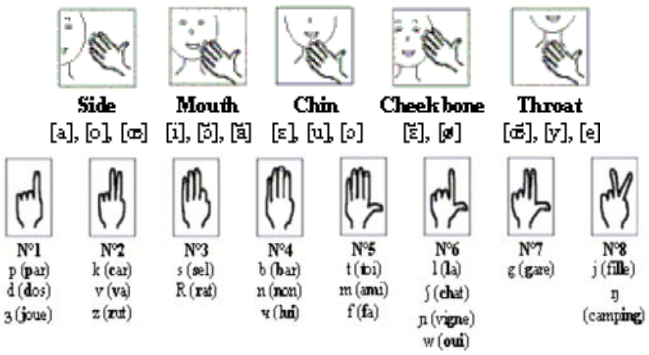


Figure 1: *CS Hand position (top) for vowels and CS handshapes (bottom) for consonants (adapted from [2]).*

This paper focuses on the automatic recognition of French vowels by combining the two flows of information (hand and lip). The first part presents a method to automatically segment the hand flow. The objective of this segmentation is to detect the CS hand position and also to identify wether the hand is maintained in a position or is in transition between two positions. The second part develops an automatic method of lip target segmentation applied to vowels. The last part discusses how both flows of information are merged to recognize vowels in a sequence. Finally, an experimental evaluation is presented to measure the performance of the merging system as well as the performance of the segmentation method for each separate modality.

2. Data

The data was obtained from a video recording of a speaker pronouncing and coding in French CS a set of 267 sentences, repeated at least twice.

Figure 2: *Image of the speaker.*

The French CS speaker is a native female speaker of French, certified in French CS. She regularly translates into French CS code in a school. The recording was made in a sound-proof booth at the Parole & Cognition department of Grenoble Images Parole Signal Automatique laboratory, (GIPSA-lab, department P&C), at 50 frames/second for the image video part. The speaker was seated and wore a helmet that served to keep her head in a fixed position and thus in the field of the camera. She wore opaque glasses to protect her eyes against a halogen floodlight. The camera in large focus was used for the hand and the face and was connected to a betacam recorder. The lips were painted in blue, and blue marks were placed on the speaker's glasses as reference points. Blue marks were placed on the left hand, on the back and at the extremity of the fingers to independently follow the displacement of the hand and the handshape formation. Blue marks were placed on the speaker's goggles as reference points (figure 2).

A square paper was recorded for further pixel-to-centimeter conversion. Using ICP's Face-Speech processing system, the audio part of the video recording was digitized at 22,050 Hz in synchrony with the image part, the latter being stored as Bitmap frames every 20 ms. A specific image processing was applied to the Bitmap frames in the lip region to extract the inner and outer contours and to derive the corresponding characteristic parameters (Lallouache, 1991 [5]): lip width (A), lip aperture (B) and lip area (S). These parameters were converted using a pixel-to-centimeter conversion formula. Finally the parameters were low-pass filtered. The x and y coordinates of the center of gravity of the hand landmarks were automatically extracted from the image as follows. A process based on image processing detected all marks on the image, and the knowledge of those on the back of hand and on the goggles allowed to extract the marks on the fingers. The coordinates initially in pixels were converted into centimeters using the pixel-to-centimeter conversion formula.

The acoustic signal was automatically labeled at the phonetic level using forced alignment (see Lamy, 2004 [7] for a description of the speech recognition tools used for this). Since the orthographic transcription of each sentence was known, a dictionary containing the phonetic transcriptions of all words was used to produce the sequence of phonemes associated with each acoustic signal. This sequence was then aligned with the acoustic signal using French ASR acoustic models trained on the BRAF100 database (Vaufreydaz, 2000 [8]).

The whole process resulted in a set of temporally coherent signals: the x and y hand position of the reference hand landmark placed on the back of the hand near the knuckles, every 20 ms, the lip parameter values every 20 ms and the corresponding acoustic signal (figure 4).

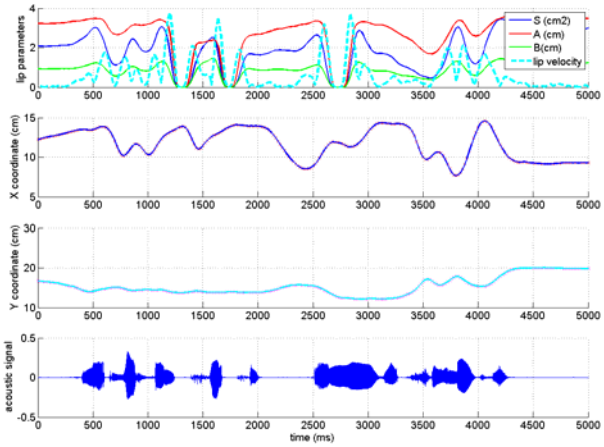


Figure 4: *Example of signals. From top to bottom: Inner lip parameters (A, B, S and lip velocity calculated from S), x and y coordinates of the reference hand landmark, the acoustic realization.*

3. Automatic hand segmentation

The x and y trajectories are characterized by smooth deviations between local extrema corresponding to spatial FCS Hand positions. The automatic segmentation process of the hand trajectories involved automatic temporal marking of the beginning and the end of each of these segments. The first step consisted of the automatic labeling of a hand position to each frame, i.e. every 20 ms.

The method uses the likelihood computed from a Gaussian model of the x-y coordinates that corresponds to the center of gravity of hand landmarks. This kind of classifier was chosen for its simplicity and especially for the homogeneous dispersion of the positions (see Figure 4, the results for the reference hand landmark). Each of the five hand positions was modelized by two 2-dimensional Gaussian models built from a dictionary of 30 images manually selected in the corpus. The first one is devoted to the reference hand landmark and the second one to the landmark placed at the extremity of the pointing finger. The use of the x-y coordinates of these two landmarks was needed to improve the robustness of the classifying method. For the classification phase, we consider a given frame with its x-y coordinates of both landmarks. For each landmark x-y coordinate, a vector made of five probability densities is delivered, thanks to the five Gaussian models. The two computed vectors are combined by a scalar product in order to obtain a final vector with five components. Thus recognized hand position is to the one that gives the maximum amongst the 5 components of the final vector.

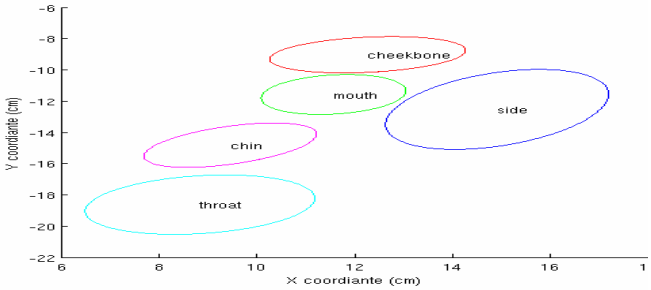


Figure 4: *2 standard deviation ellipsis around the corresponding average values of the reference hand landmark extracted from the learning data, for the 5 hand positions.*

This method applied to each frame of a sentence delivers a sequence of hand position numbers from 1 to 5, with a set of plateaus. A plateau is defined by a set of successive identical position numbers.

At this step of classification, it is not possible to distinguish the transitions between attained hand positions. Thus, a second step was needed to refine this result. Its principle was based on the use of the velocity minimum applied to the x-y coordinated of the reference hand landmark. The velocity was defined as the Euclidean distance between two successive (x-y) points temporally spaced by 20ms. Inside each plateau, the value of the velocity minimum is detected. In addition, the value of the velocity maximum is detected between the middle of the previous plateau and the middle of the considered one. The contrast is calculated as the difference between these two extreme values. A percentage (40%) of this contrast is added to the minimum value in order to define a threshold value. Thus for the considered plateau, the positions for which the velocity is lower than the threshold value are considered in the target hand position. In other hand, the positions for which the velocity is higher than the threshold value are considered in the transition. Finally, incorrect plateau detections (see comparison between Figure 5 and Figure 6) are considered as points in a transition.

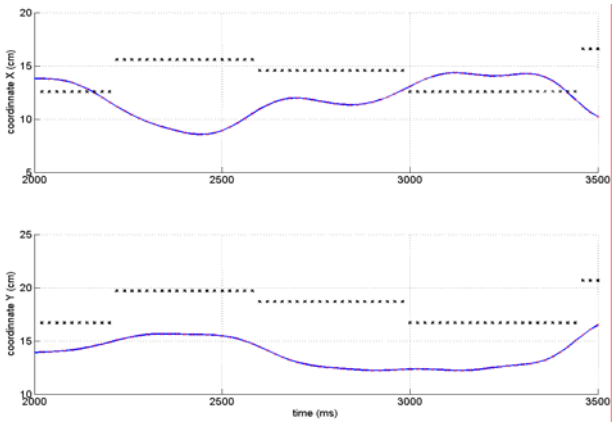


Figure 5: *Zoom of signals with hand positions plateaus delivered by the classifier.*

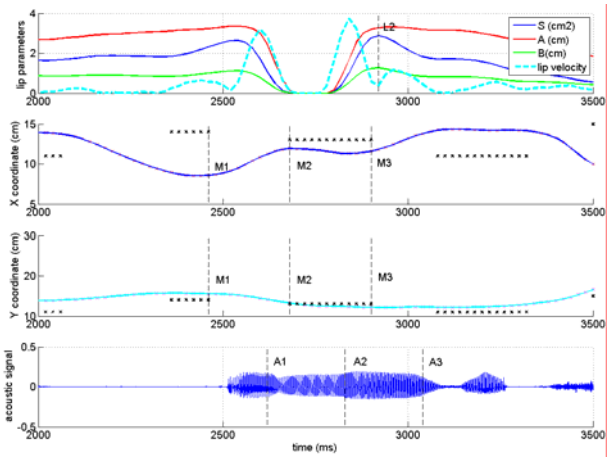


Figure 6: *Zoom of signals with M1, M2, M3, A1, A2, A3 and L2 labels.*

Following the nomenclature of Attina and colleagues, the extremity of the plateaus delimiting the attained hand position and the transitions were automatically labeled M1, M2 and M3 for the onset of the transition, the onset and the end of the reached hand position, respectively. For the acoustic signal, the beginning of the acoustic realization of the consonant and of the vowel and the end of the vowel were labeled A1, A2 and A3.

4. Automatic lip target segmentation

The lips were characterized at the instant the lip target was attained. The automatic definition of this instant, labeled L2 (see figure 6 for an example), is based on the temporally marked phonetic chain. Recall that the phonetic chain marks the acoustic realization. Note that the beginning and the end of each phoneme are obtained automatically with a forced alignment; this labeling may therefore include errors or fuzzy phone frontiers. Moreover, it is well known that the lip can anticipate the acoustic realization. Thus, in the automatic process of lip target calculation, the middle of the phoneme interval is considered as a first estimation of the instant of vocalic target. The target instant is finally obtained at the nearest instant of minimum lip velocity. In the case of important anticipation the research process is limited by the end frontier of the phone acoustic realization. Lip velocity (see Figure 3) is estimated from the lip area S parameter as the difference between two successive values normalized by the sample periodicity (20 ms). Note that S is highly correlated to the crossing of A by B (r = 0.99).

The algorithm for vocalic lip target instant detection is thus as follows: (1) calculation of the lip velocity from S parameter, (2) detection of all the local minima, (3) determination of the mid-point of the vowel from the phonetic chain (4) choice of the nearest instant of lip velocity local minimum.

From the L2 instants obtained with this method, it has been shown that vowels could be grouped into three categories in conformity with the phonetic description of the vowels (anterior non rounded vowels [a, ɛ̃, i, œ̃, e, ɛ], high and mid-high rounded [ɔ̃, y, o, ø, u] low and mid-low rounded vowels [ã , ɔ, œ]) (see Aboutabit et al., 2006 [4] for more details).

In this study, it has been demonstrated that when the CS hand position was given without error, high scores of vowel identification are obtained (89% as average recognition rate of vowel) with only one measure instant, defined by L2 instant.

5. Vowel recognition

5.1. Method

The complete vowel recognition needs to merge both manual and lip informations obtained from the two previous automatic processing. Several merging approaches can be considered. Among these methods, the classical models of audio-visual integration (Schwartz et al., 1998 [9]) are interesting but a few ones can be adapted to the Cued Speech merging case, while others do not. For example, the direct identification model (DI) seems to be not appropriate to the CS gestures fusion. One raison is that this kind of identification needs a system that is capable to merge quantitative information provided from lips (lip parameter values) and qualitative information provided from CS hand gestures (hand position and configuration). Even if a transformation in quantitative components is possible from the CS hand information, the fact to take into account components with different origin in a same vector poses the problem of their weighting. As second raison, the temporal desynchronization between lip and hand information is a serious problem for this model.

Alternatively, the separated identification model (SI) seems to be convenient. Considering this model, on one hand, at the M2 instant the CS hand position is known and defines a first group of vowels, composed by two or three vowels. On the other hand, at the corresponding L2 instant a second group of vowels (viseme) is derived from a classification of lip parameters. Then, the vowel recognition results from the intersection between these two groups of vowels. In this case, it is possible that the intersection may be empty in the case of a determinist fusion. Thus, no vowel is recognized. This non identification problem may be caused by an error on the lip decision and/or on the hand position decision. To solve this problem, instead to use one classifier of viseme in the lip process, five classifiers, one for each CS hand position could be considered. Then, as result from the lip process, five vowels are selected (i.e. one vowel for each classifier).

To reduce the number of lip classifiers from five to a single one, a fusion model derived from the SI one is considered in the following. It consists to constrain the lip decision by the hand position decision considering the advance of the M2 instant over the L2 instant in the case of vowels (Aboutabit et al., 2006 [3]). Then, between two successive M2 instants, a L2 instant is located excepted in the case of a consonant alone. At the first M2 instant the CS hand position allows to identify a first group of vowels. A simple Gaussian classification on lip parameters extracted at L2 recognizes one element among this group. Figure 7 illustrates the method.

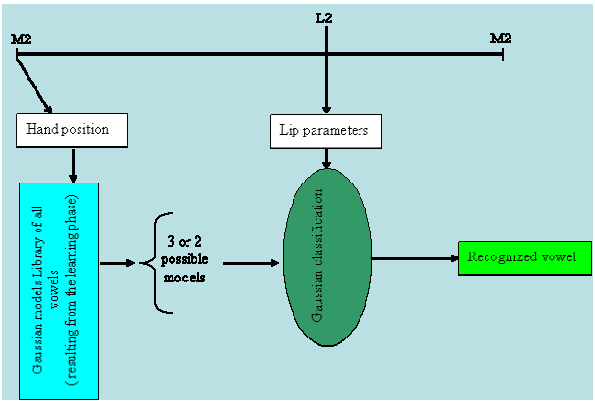


Figure 7*: Vowel identification: fusion schema of both lip and hand information.*

5.2. Results and discussion

To test this merging method, the test corpus contains 774 vowels from 120 coded and pronounced sentences. The global score of vowel recognition is 77.6%. If the CS hand position is known without error, the vowel recognition is 89% (Aboutabit et al. [4]). Then, the difference between these two scores is due to the automatic decision on the CS hand position and/or to the matching of M2 and L2.

The recognition score of 77.6% is slightly lower than the CS perceptual effectiveness score of 83.5% obtained by Nicholls and Ling (1982, [6]) in their study on the reception of CV and VC syllables with hearing-impaired children.

In the evaluation of the corpus coded by the speaker, a decoding test in reception was performed by a profoundly hearing-impaired subject practicing regularly the CS system. The test consisted in the decoding of the whole sentences of the corpus. For the 120 sentences used previously, the score of correct decoding attains 94.8% for the vowels. This latter is the reference to compare the performance of the recognition method (77.6%).

6. Conclusions and perspectives

The merging model, in which the hand is considered at the M2 instant in advance to the L2 instant of the lips, was validated. The recognition score for vowels (77.6%), compared to the different reference scores previously recalled, is promising to go further. Indeed, an improvement of the automatic segmentation of the CS hand position should improve the decision on the CS hand position and enhance the precision of the M2 instant. This allows to reduce the M2-L2 matching errors due to the imprecision on M2. In addition, concerning the lips, a better selection of the velocity local minima, should increase the recognition performances.

As a perspective, the merging model could be extended to the consonant in vocalic context. In this case, the complete [M2, L2] interval should be considered in order to take into account the complex effect of the co articulation.

7. Acknowledgements

Many thanks to Sabine Chevalier, our CS speaker, for having accepted the recording constraints. This work is supported by the French TELMA project (RNTS / ANR).

8. References

[1] Cornett, R.O. "Cued Speech", American Annals of the Deaf, 112, pp. 3-13, 1967.

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

[3] Aboutabit, N., Beautemps, D. and Besacier, L. "Hand and Lips desynchronization analysis in French Cued Speech: Automatic segmentation of Hand flow", In Proceedings of ICASSP'06, 2006.

[4] Aboutabit, N., Beautemps, D. and Besacier, L., "Vowels classification from lips: the Cued Speech production case". In Proceedings of ISSP'06, 2006.

[5] Lallouache, M.-T. " Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres," Doctoral dissertation, Institut National Polytechnique de Grenoble, Grenoble 1991.

[6] Nicholls, G., Ling, D. "Cued Speech and the reception of spoken language", Journal of Speech and Hearing Research.25, 262-269, 1982.

[7] Lamy, R., Moraru, D., Bigi, B., Besacier, L. Premiers pas du CLIPS sur les données d'évaluation ESTER. *In* Proc. of Journées d'Etude sur la Parole, Fès, Maroc, 2004.

[8] Vaufreydaz, D., Bergamini, J., Serignat, J. F., Besacier, L. & Akbar, M. A New Methodology for Speech Corpora Definition from Internet Documents. LREC2000, 2nd International Conference on Language Resources and Evaluation. Athens, Greece, pp. 423-426, 2000.

[9] Schwartz, J. L., Robert-Ribes, J. & Escudier, P. Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. Hearing by Eye II, Advances in the psychology of speechreading and auditory visual speech. Psychology Press, pp. 85-108, Hove (UK), 1998.

Heracleous, P., Beautemps, D., Aboutabit, N., 2010. Cued Speech automatic recognition in normal-hearing and deaf subjects. Speech Communication 52(6): 504-512.

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 504–512
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Cued Speech automatic recognition in normal-hearing and deaf subjects

Panikos Heracleous [a,b,*], Denis Beautemps [b], Noureddine Aboutabit [b]

[a] ATR, Intelligent Robotics and Communication Laboratories, 2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan
[b] GIPSA-lab, Speech and Cognition Department, CNRS UMR 5216/Stendhal University/UJF/INPG, 961 rue de la Houille Blanche Domaine universitaire BP 46, F-38402 Saint Martin d'Hères cedex, France

Received 31 March 2009; received in revised form 2 March 2010; accepted 3 March 2010

Abstract

This article discusses the automatic recognition of Cued Speech in French based on hidden Markov models (HMMs). Cued Speech is a visual mode which, by using hand shapes in different positions and in combination with lip patterns of speech, makes all the sounds of a spoken language clearly understandable to deaf people. The aim of Cued Speech is to overcome the problems of lipreading and thus enable deaf children and adults to understand spoken language completely. In the current study, the authors demonstrate that visible gestures are as discriminant as audible orofacial gestures. Phoneme recognition and isolated word recognition experiments have been conducted using data from a normal-hearing cuer. The results obtained were very promising, and the study has been extended by applying the proposed methods to a deaf cuer. The achieved results have not shown any significant differences compared to automatic Cued Speech recognition in a normal-hearing subject. In automatic recognition of Cued Speech, lip shape and gesture recognition are required. Moreover, the integration of the two modalities is of great importance. In this study, lip shape component is fused with hand component to realize Cued Speech recognition. Using concatenative feature fusion and multi-stream HMM decision fusion, vowel recognition, consonant recognition, and isolated word recognition experiments have been conducted. For vowel recognition, an 87.6% vowel accuracy was obtained showing a 61.3% relative improvement compared to the sole use of lip shape parameters. In the case of consonant recognition, a 78.9% accuracy was obtained showing a 56% relative improvement compared to the use of lip shape only. In addition to vowel and consonant recognition, a complete phoneme recognition experiment using concatenated feature vectors and Gaussian mixture model (GMM) discrimination was conducted, obtaining a 74.4% phoneme accuracy. Isolated word recognition experiments in both normal-hearing and deaf subjects were also conducted providing a word accuracy of 94.9% and 89%, respectively. The obtained results were compared with those obtained using audio signal, and comparable accuracies were observed.

Keywords: French Cued Speech; Hidden Markov models; Automatic recognition; Feature fusion; Multi-stream HMM decision fusion

1. Introduction

To date, visual information is widely used to improve speech perception or automatic speech recognition (lipreading) (Potamianos et al., 2003). With lipreading technique, speech can be understood by interpreting the movements of lips, face and tongue. In spoken languages, a particular facial and lip shape corresponds to a specific sound (phoneme). However, this relationship is not one-to-one and many phonemes share the same facial and lip shape (visemes). It is impossible, therefore to distinguish phonemes using visual information alone.

Without knowing the semantic context, one cannot perceive the speech thoroughly even with high lipreading performances. To date, the best lip readers are far away into reaching perfection. On average, only 40–60% of the vowels of a given language (American English) are recognized by lipreading (Montgomery and Jackson, 1983), and 32% when relating to low predicted words (Nicholls and Ling, 1982). The best result obtained amongst deaf participants was 43.6% for the average accuracy (Auer and Bernstein,

* Corresponding author at: ATR, Intelligent Robotics and Communication Laboratories, 2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan.
E-mail address: panikos@atr.jp (P. Heracleous).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.03.001

2007; Bernstein et al., 2007). The main reason for this lies in the ambiguity of the visual pattern. However, as far as the orally educated deaf people are concerned, the act of lipreading remains the main modality of perceiving speech.

To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, Cornett (1967) developed the Cued Speech system in 1967 to complement the lip information and make all phonemes of a spoken language clearly visible. As many sounds look identical on face/lips (e.g., /p/, /b/, and /m/), using hand information those sounds can be distinguished and thus make possible for deaf people to completely understand a spoken language using visual information only.

Cued Speech (also referred to as Cued Language (Fleetwood and Metzger, 1999)) uses hand shapes placed in different positions near the face along with natural speech lipreading to enhance speech perception from visual input. This is a system where the speaker faces the perceiver and moves his hand in close relation with speech. The hand, held flat and oriented so that the back of the hand faces the perceiver, is a cue that corresponds to a unique phoneme when associated with a particular lip shape. A manual cue in this system contains two components: the hand shape and the hand position relative to the face. Hand shapes distinguish among consonant phonemes whereas hand positions distinguish among vowel phonemes. A hand shape, together with a hand position, cues a syllable.

Cued Speech improves the speech perception of deaf people (Nicholls and Ling, 1982; Uchanski et al., 1994). Moreover, for deaf people who have been exposed to this mode since their youth, it offers a complete representation of the phonological system, and therefore it has a positive impact on the language development (Leybaert, 2000). Fig. 1 describes the complete system for French. In French Cued Speech, eight hand shapes in five positions are used. The system was adapted from American English to French in 1977. To date, Cued Speech has been adapted in more than 60 languages.

Another widely used communication method for deaf individuals is the Sign Language (Dreuw et al., 2007; Ong and Ranganath, 2005). Sign Language is a language with its own grammar, syntax and community; however, one must be exposed to native and/or fluent users of Sign Language to acquire it. Since the majority of children who are deaf or hard-of-hearing have hearing parents (90%), these children usually have limited access to appropriate Sign Language models. Cued Speech is a visual representation of a spoken language, and it was developed to help raise the literacy levels of deaf individuals. Cued Speech was not developed to replace Sign Language. In fact, Sign Language will be always a part of deaf community. On the other hand, Cued Speech is an alternative communication method for deaf individuals. By cueing, children who are deaf would have a way to easily acquire the native home language, read and write proficiently, and communicate more easily with hearing family members who cue them.

In the current study, the authors demonstrate that visible gestures are as discriminant as audible orofacial gestures. Phoneme recognition and isolated word recognition experiments were conducted using data from a normal-hearing cuer, and promising results were obtained. In addition, the proposed methods were applied to a deaf cuer and similar results were obtained compared with automatic recognition of Cued Speech in normal-hearing subjects.

In the first attempt for vowel recognition in Cued Speech, in Aboutabit et al. (2007) a method based on separate identification, i.e., indirect decision fusion was used

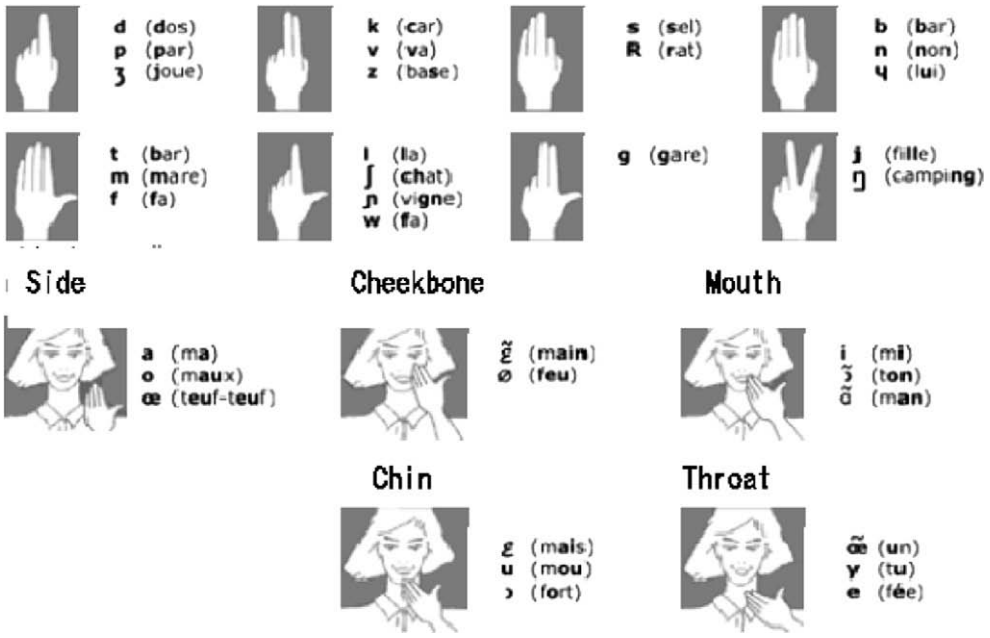


Fig. 1. Hand shapes for consonants (top) and hand position (bottom) for vowels in French Cued Speech.

and a 77.6% vowel accuracy was obtained. In this study, however, the proposed method is based on HMMs and uses concatenative feature fusion and multi-stream HMM decision fusion to integrate the components into a combined one and then perform automatic recognition. Fusion (Nefian et al., 2002; Hennecke et al., 1996; Adjoudani and Benoît, 1996) is the integration of all available single-modality streams into a combined one. In this study, lip shape and hand components are combined in order to realize automatic recognition in Cued Speech for French.

2. Methodology

2.1. Cued Speech materials

The data for vowel-, consonant-, and phoneme recognition experiments were collected from a normal-hearing cuer. The female native French speaker employed for data recording was certified in transliteration speech into Cued Speech in the French language. She regularly cues in schools. The cuer wore a helmet to keep her head in a fixed position and opaque glasses to protect her eyes against glare from the halogen floodlight. The cuer's lips were painted blue, and blue marks were marked on her glasses as reference points. These constraints were applied in recordings in order to control the data and facilitate the extraction of accurate features. The data were derived from a video recording of the cuer pronouncing and coding in Cued Speech a set of 262 French sentences. The sentences (composed of low predicted multi-syllabic words) were derived from a corpus that was dedicated to Cued Speech synthesis (Gibert et al., 2005). Each sentence was dictated by an experimenter, and was repeated two or three times (to correct the pronunciation errors) by the cuer resulting in a set of 638 sentences.

The audio part of the video recording was synchronized with the image. Fig. 2 shows the lip shape parameters used in the study. An automatic image processing method was applied to the video frames in the lip region to extract their inner and outer contours and derive the corresponding characteristic parameters: lip width (A), lip aperture (B), and lip area (S) (i.e., six parameters in all).

The process described here resulted in a set of temporally coherent signals: the 2D hand information, the lip width (A), the lip aperture (B), and the lip area (S) values for both inner and outer contours, and the corresponding acoustic signal. In addition, two supplementary parameters relative to the lip morphology were extracted: the pinching of the upper lip (Bsup) and lower (Binf) lip. As a result, a set of eight parameters in all was extracted for modeling lip shapes. For hand position modeling, the xy coordinates of two landmarks placed on the hand were used (i.e., four parameters). For hand shape modeling, the xy coordinates of the landmarks placed on the fingers were used (i.e., 10 parameters). Non visible landmarks receive default coordinates $[0, 0]$.

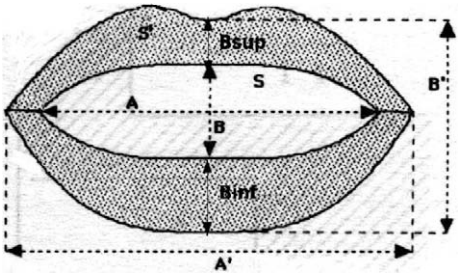


Fig. 2. Parameters used for lip shape modeling.

During the recording of Cued Speech material for isolated word recognition experiments, the conditions were different from the ones described earlier. The system was improved by excluding the use of a helmet by the cuer, enabling in this way the head movements during recording. The subject was seated on a chair in a way to avoid large movements in the third direction (i.e., towards the camera). However, the errors that might occur have not been evaluated. In addition, the landmarks placed on the cuer's fingers were of different colors in order to avoid the hand shape coding and the finger identification (cf. Section 2.3), and this helped to simplify and speed up the image processing stage. In these recording sessions, a normal-hearing cuer and a deaf cuer were employed. The corpus consisted of 1450 isolated words with each of 50 words repeated 29 times by the cuers.

2.2. Lip shape and hand components modeling

In the phoneme recognition experiments, context-independent, three-state, left-to-right, no-skip-phoneme HMMs were used. Each state was modeled with a mixture of 32 Gaussians. In addition to the basic lip and hand parameters, first- (Δ) and second-order derivatives ($\Delta\Delta$) were used as well. For training and test, 426 and 212 sentences were used, respectively. The training sentences contained 3838 vowel and 4401 consonant instances, and the test sentences contained 1913 vowel and 2155 consonant instances, respectively. Vowels and consonants were extracted automatically from the data after a forced alignment was performed using the audio signal.

For isolated word recognition experiments two HMM sets were trained (deaf and normal-hearing). Fifteen repetitions of each word were used to train 50, six-state, whole word HMMs, and 14 repetitions were used for testing. Eight and ten parameters were used for lip shape and hand shape modeling, respectively.

In automatic speech recognition, a diagonal covariance matrix is often used because of the assumption that the parameters are uncorrelated. In lipreading, however, parameters show a strong correlation. In this study, a global Principal Component Analysis (PCA) using all the training data was applied to decorrelate the lip shape parameters and then a diagonal covariance matrix was used. The test data were then projected into the PCA space.

All PCA lip shape components were used for HMM training. For training and recognition the HTK3.1 toolkit (Young et al., 2001) was used.

2.3. Ordering finger landmarks

For consonant recognition, the correct hand shape is also required. Instead of a deterministic recognition of the hand shape, a probabilistic method is used based on the *xy* coordinates of the landmarks placed on the fingers (Fig. 3). The coordinates are used as features for the hand shape modeling. During the image processing stage, the system detects the landmarks located on the cuer's fingers and their coordinates are computed. Since the landmarks are of the same color, the system cannot assign the coordinates to the appropriate finger in order to be used correctly (i.e., correct order) in the feature vectors. To do this, the hand shape is automatically recognized a-priori, and the information obtained is then used to assign the coordinates to the appropriate finger.

In French Cued Speech, recognition of the eight hand shapes is considered exceptional. In fact, a causal analysis based on some knowledge, such as the number and dispersion of fingers and also the angle between them, can distinguish those eight hand shapes. Based on the number of landmarks detected on fingers, the correct hand shape can be recognized. In Fig. 1 the hand shapes were numbered from left to right (i.e., S1–S8). The proposed algorithm to identify the Cued Speech hand shapes is as follows:

- Number of fingers on which landmarks are detected = 1, then the hand shape is S1.
- Number of fingers on which landmarks are detected = 4, then the hand shape is S4.
- Number of fingers on which landmarks are detected = 5, then the hand shape is S5.
- Number of finger on which landmarks are detected = 3, then the hand shape is S3 or S7. If the thumb finger is detected (using finger dispersion models) then the hand shape is S7, else the hand shape is S3.
- Number of finger on which landmarks are detected = 2, then the hand shape is S2 or S6 or S8. If the thumb finger is detected then the hand shape is S6, else the angle between the two finger landmarks according to the landmarks on the hand can identify if it is hand shape S2 or S8 (using a threshold).
- In any other case hand shape S0, i.e., no Cued Speech hand shape was detected.

The objective of finger identification stage (cf. Section 2.3) is to assign the computed coordinates to the correct finger and, in this way, to have the correct order in the feature vectors. The identification has been done in three steps. In the first step, all landmarks in the frame were detected. The landmarks placed on the speaker glasses and on the back of the hand were benched to have only the landmarks corresponding to the fingers. Secondly, the coordinates of these landmarks were projected on the hand axis defined by the two landmarks on the back of the hand. The third step consisted of sorting the resulted coordinates following the perpendicular axis to the hand direction from the smaller to the largest (Fig. 4). In this step, the hand shape coding was used to associate each coordinate with the corresponding finger. For example, when there were three landmarks coordinates and the hand shape number was S3, the smallest coordinate was associated with the middle finger, the middle one to the ring finger, and the biggest one to the baby finger. The coordinates of the fingers which are not present in a hand shape are replaced by a constant in order to keep the same dimension of the feature vectors in the HMM modeling.

2.4. Concatenative feature fusion

The feature concatenation uses the concatenation of the synchronous lip shape and hand features as the joint feature vector

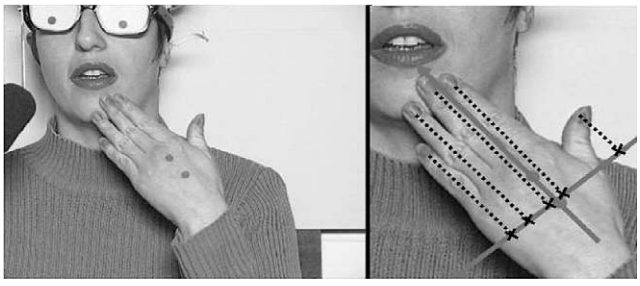


Fig. 4. Image of a Cued Speech cuer (left) and the projection method (right).

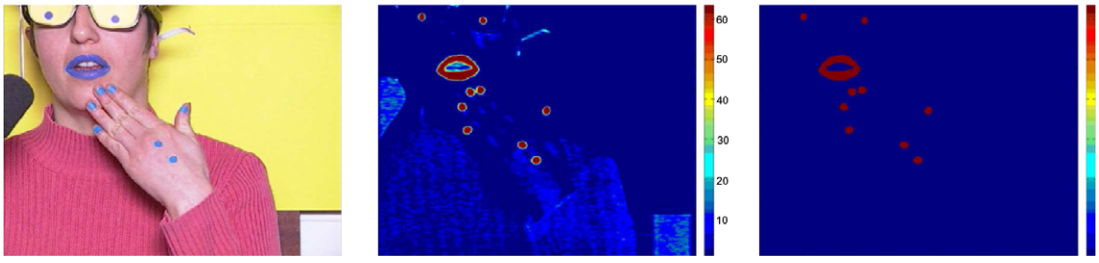


Fig. 3. The three-step algorithm applied for lip shape and gesture detection based on detection of blue objects. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$O_t^{LH} = \left[O_t^{(L)^T},\ O_t^{(H)^T} \right]^T \in R^D, \quad (1)$$

where O_t^{LH} is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand feature vector, and D the dimensionality of the joint feature vector. In vowel recognition experiments, the dimension of the lip shape stream was 24 (8 basic parameters, 8 Δ, and 8 $\Delta\Delta$ parameters) and the dimension of the hand position stream was 12 (4 basic parameters, 4 Δ, and 4 $\Delta\Delta$ parameters). The dimension D of the joint lip-hand position feature vectors was, therefore 36. In consonant recognition experiments, the dimension of the hand shape stream was 30 (10 basic parameters, 10 Δ, and 10 $\Delta\Delta$ parameters). The dimension D of the joint lip-hand shape feature vectors was, therefore 54.

2.5. Multi-stream HMM decision fusion

Decision fusion captures the reliability of each stream, by combining the likelihoods of single-modality HMM classifiers. Such an approach has been used in multi-band audio only ASR (Bourlard and Dupont, 1996) and in audio-visual speech recognition (Potamianos et al., 2003). The emission likelihood of multi-stream HMM is the product of emission likelihoods of single-modality components weighted appropriately by stream weights. Given the O joint observation vector, i.e., lip shape and hand position component, the emission probability of multi-stream HMM is given by

$$b_j(O_t) = \prod s = 1^S \left[\sum_{m=1}^{M_s} c_{jsm} N\left(O_{st}; \mu_{jsm}, \Sigma_{jsm}\right) \right]^{\lambda_s}, \quad (2)$$

where $N(O; \mu, \Sigma)$ is the value in O of a multivariate Gaussian with mean μ and covariance matrix Σ, and S the number of streams. For each stream s, M_s Gaussians in a mixture are used, with each weighted with c_{jsm}. The contribution of each stream is weighted by λ_s. In this study, we assume that the stream weights do not depend on state j and time t, as happen in the general case of multi-stream HMM decision fusion. However, two constraints were applied, namely

$$0 \leqslant \lambda_h,\ \lambda_l \leqslant 1 \quad \text{and} \quad \lambda_h + \lambda_l = 1, \quad (3)$$

where λ_h is the hand position stream weight, and λ_l is the lip shape stream weight. The HMMs were trained using maximum likelihood estimation based on the Expectation–Maximization (EM) algorithm. However, the weights cannot be obtained by maximum likelihood estimation. The weights were adjusted to 0.7 and 0.3 values, respectively. The selected weights were obtained experimentally by maximizing the accuracy on held-out data.

3. Experiments and results

3.1. Hand shape recognition

To evaluate the previously described hand shape recognition system, a set of 1009 frames was used and recognized automatically. Table 1 shows the confusion matrix of the recognized hand shapes by the automatic system. It can be seen that the automatic system recognized correctly 92.4% of the hand shapes on average. This score showed that using only the 2D coordinates of five landmarks placed at the finger extremities, the accuracy did not decrease drastically compared with the 98.8% of recognized hand shapes obtained by Gibert et al. (2005) with the use of the 3D coordinates of 50 landmarks placed on the hand and the fingers derived from a motion capture system. The most common errors can be attributed to landmark detection processing. However, in some cases, one or more landmarks were not detected because of the rotation of the hand. In some other cases, landmarks remained visible even when the fingers were bended.

Table 1
Confusion matrix of hand shape recognition evaluation (derived for Aboutabit (2007)).

	S0	S1	S2	S3	S4	S5	S6	S7	S8	%c
S0	33	2	0	0	0	0	0	0	0	94
S1	16	151	0	0	0	0	1	0	5	87
S2	1	2	93	0	0	0	0	0	6	91
S3	0	0	0	163	2	0	0	3	9	91
S4	3	0	0	0	100	0	0	3	0	94
S5	2	0	0	4	4	193	0	0	1	95
S6	0	0	0	0	0	0	124	5	0	96
S7	0	0	0	0	0	0	0	17	0	100
S8	1	05	0	2	0	0	0	0	58	95

3.2. Vowel and consonant recognition

Fig. 5 shows the vowel recognition results when concatenative feature fusion was used. As shown, by integrating hand position component with lip shape component, a vowel accuracy of 85.1% was achieved, showing a 53% relative improvement compared to the sole use of lip shape parameters.

Fig. 6 shows the results achieved for vowel recognition, when multi-stream HMM decision fusion was applied.

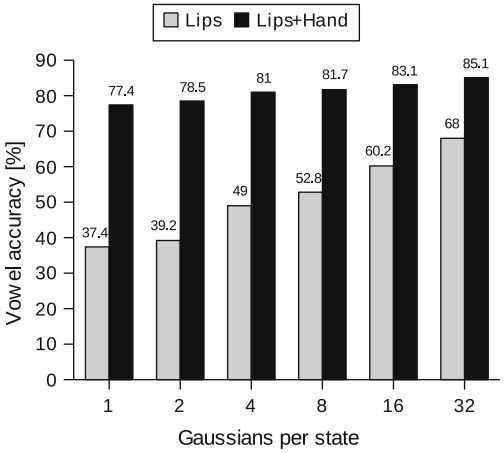


Fig. 5. Cued Speech vowel recognition using only lip and hand parameters based on concatenative feature fusion.

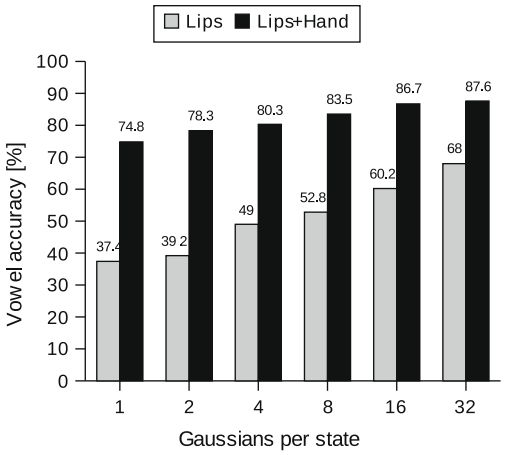


Fig. 6. Cued Speech vowel recognition using only lip and hand parameters based on multi-stream HMM decision fusion was used.

Using 32 Gaussians per state, an 87.6% vowel accuracy was obtained, showing a relative improvement of 61%. The results obtained are comparable with the results obtained for vowel recognition when using audio speech (e.g., Merkx and Miles (2005)).

The results showed that multi-stream HMM decision fusion results in better performance than a concatenative feature fusion. To decide whether the difference in performance between the two methods is statistically significant, the McNemar's test was applied (Gillick and Cox, 1989).The observed p-value was 0.001 indicating that the difference is statistically significant.

Using concatenative feature fusion, lip shape component was integrated with hand shape component and consonant recognition was conducted. For hand shape modeling, the xy coordinates of the fingers, and first- and second-order derivatives were used. In total, 30 parameters were used for hand shape modeling. For lip shape modeling, 24 parameters were used. Fig. 7 shows the obtained results in the function of Gaussians per state. It can be seen that when using 32 Gaussians per state, a consonant accuracy of 78.9% was achieved. Compared to the sole use of lip shape, a 56% relative improvement was obtained.

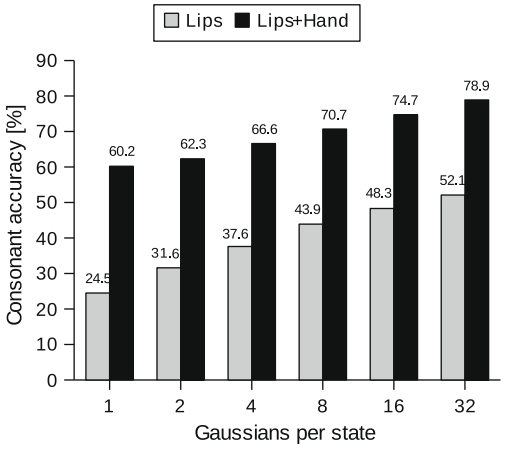


Fig. 7. Cued Speech consonant recognition using only lip and hand parameters based on concatenative feature fusion.

3.3. Phoneme recognition

In the previous sections, it was reported that different types of modeling were used for vowels and consonants. More specifically, for vowel modeling, fusion of lip shape and hand position components was used. For consonant modeling, fusion of lip shape and hand shape components was used. As a result, feature vectors in vowel and consonant recognition are of different lengths. The feature vectors for vowels have a length of 36 and the feature vectors of consonants have a length of 54. This is a limitation in using a common HMM set for phoneme recognition. To deal with this problem, three approaches were proposed in order to realize phoneme recognition in Cued Speech for French.

In the first approach, feature vectors with the same length were used. At each frame, lip shape parameters, hand position parameters, and hand shape parameters were extracted during the image processing stage. For each phoneme, all parameters were used in concatenated feature vectors with a length of 66 (i.e., eight lip shape parameters, four coordinates for hand position, 10 coordinates for hand shape, along with the first- and second-order derivatives, as well). The obtained phoneme accuracy was as low as 61.5% due to a high number of confusions between vowels and consonants. Although, phoneme recognition in Cued Speech is a difficult task, the obtained phoneme accuracy was lower than expected. To obtain a performance with higher phoneme accuracy, different approaches were also investigated.

Cued Speech phoneme recognition was further improved by applying GMM discrimination. A vowel-independent and a consonant-independent GMM models were trained using lip shape parameters only. For training the two GMMs the corresponding vowel and consonant data were used. For modeling, 64 Gaussians were used. The number of Gaussians was selected experimentally on several experiments in order to achieve the highest classification scores. Phoneme recognition was realized in a two-pass scheme. In the first pass, using the two GMMs, the nature of the input was decided. More specifically, the input and the two GMMs were matched. Based on the obtained likelihood, the input was considered to be vowel or consonant. When the likelihood of the vowel-GMM was higher than that of the consonant-GMM, the decision was made for a vowel. When the consonant-GMM provided a higher likelihood, the input was considered to be a consonant. In the case of vowel inputs, the discrimination accuracy was 86.9% and in the case of consonant inputs the discrimination accuracy was 81.5%. In the second pass, switching to the appropriate HMM set took place, and vowel or consonant recognition was realized using feature vectors corresponding to the vowel modeling or consonant modeling, respectively. The obtained phoneme accuracy

was 70.9%. The obtained accuracies were lower than the accuracies obtained in the separate vowel and consonant recognition experiments, because of the discrimination errors of the first pass. The obtained result, however, showed a relative improvement of 24% compared to the use of concatenated feature vectors.

In the third approach for phoneme recognition, instead of two, eight GMM models (i.e., each one corresponding to a viseme) were used. Three vowel-viseme GMMs and five consonant-viseme GMMs were used based on the viseme grouping. Similar to the previous experiment, phoneme recognition was performed in two-passes. In the first pass, matching between the input and the eight GMMs took place. Based on the maximum likelihood, the system switched to the corresponding vowel- or consonant-HMM set, and recognition was performed. Using eight GMMs, the discrimination accuracy was increased up to 89.3% for the vowels and up to 84.6% for the consonants. The achieved phoneme recognition was 74.4% (i.e., 80.3% vowel accuracy and 68.5% consonant accuracy). Compared to the use of two GMMs, a relative improvement of 12% was obtained. Compared with the use of the full set of concatenated parameters, a relative improvement of 33.5% was achieved.

3.4. Isolated word recognition

In this section, isolated word recognition experiments both in normal-hearing and deaf subjects are presented. In these experiments, the landmarks were of different colors in order to avoid the hand shape recognition and the finger identification stage. The image processing system locates the landmarks, and the coordinates of each landmark are assigned to each finger based on the colors. Doing this, the feature vectors of hand shape contains the coordinates of the landmarks in the correct order, and the errors that occur during the hand shape coding are not accumulated into the recognition stage.

Fig. 8 shows the results obtained in the function of several Gaussians per state in the case of the normal-hearing cuer. In the case of a single Gaussian per state, using lip shape alone obtained a 56% word accuracy; however, when hand shape information was also used, a 92.8% word accuracy was obtained. The highest word accuracy when using lip shape was 72%, obtained in the case of using four Gaussians per state. In that case, the Cued Speech word accuracy using also hand information was 94.9%.

Fig. 9 shows the obtained results in the case of a deaf cuer. The results show that in the case of the deaf subject, words were better recognized when using lip shape alone compared to the normal-hearing subject. The fact that deafs rely on lipreading for speech communication may increase their ability not only for speech perception but also for speech production. The word accuracy in the case of the deaf subject was 89% compared to the 94.9% in the normal-hearing subject. The difference in performance might be because of the lower hand shape recognition in

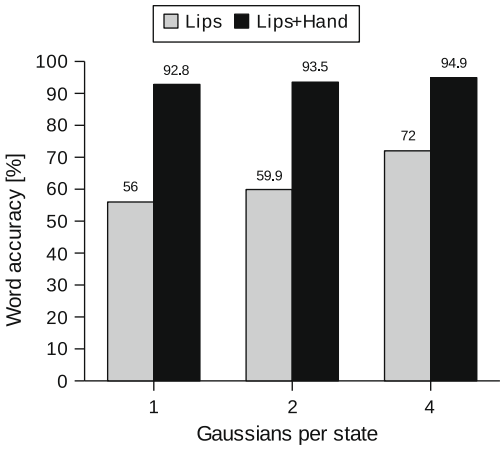


Fig. 8. Word accuracy for isolated word recognition in the case of a normal-hearing subject.

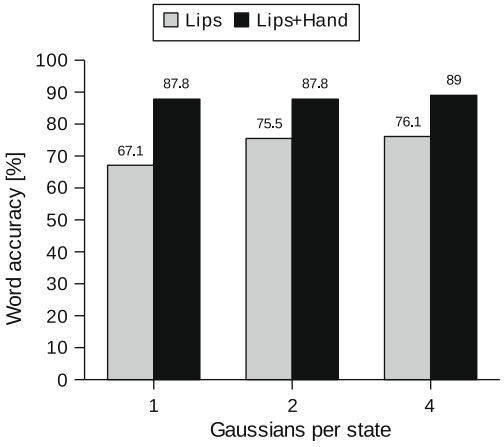


Fig. 9. Word accuracy for isolated word recognition in the case of a deaf subject.

the deaf subject. It should also be noted that the normal-hearing cuer was a professional teacher of Cued Speech. The results show that there are no additional difficulties in recognizing Cued Speech in deaf subjects, other than those appearing in normal-hearing subjects.

A multi-cuer isolated word recognition experiment was also conducted using the normal-hearing and the deaf cuers' data. The aim of this experiment is to investigate whether it is possible to train speaker-independent HMMs for Cued Speech recognition. The training data consisted of 750 words from the normal-hearing subject, and 750 words from the deaf subject. For testing 700 words from normal-hearing subject and 700 words from the deaf subject were used, respectively. Each state was modeled with a mixture of four Gaussian distributions. For lip shape and hand shape integration, the concatenative feature fusion was used.

Table 2 shows the results obtained when lip shape and hand shape features were used. The results show, that due to the large variability between the two subjects, word accuracy of cross-recognition is extremely low. On the

Table 2
Word accuracy of a multi-speaker experiment.

Test data	HMMs		
	Normal	Deaf	Normal + Deaf
Normal	94.9	0.6	92.0
Deaf	2.0	89.0	87.2

other hand, the word accuracy in normal-hearing subject when using multi-speaker HMMs was 92%, which is comparable with the 94.9% word accuracy when cuer-dependent HMMs were used. In the case of the deaf subject, the word accuracy when using multi-cuer HMMs was 87.2%, which was also comparable with the 89% word accuracy when using speaker-dependent HMMs.

The results obtained indicate that creating speaker-independent HMMs for Cued Speech recognition using a large number of subjects should not face any particular difference, other than those appear in the conventional audio speech recognition. To prove this, however, additional experiments using a large number of subjects are required.

4. Discussion

This study deals with the automatic recognition of Cued Speech in French based on HMMs. As far as our knowledge goes, automatic vowel-, consonant- and phoneme recognition in Cued Speech based on HMMs is being introduced for the first time ever by the authors of this study. Based on a review of the literature written about Cued Speech, the authors of this study have not come across any other published work related to automatic vowel- or consonant recognition in Cued Speech for any other Cued language.

The study aims at investigating the possibility of integrating lip shape and hand information in order to realize automatic recognition, and converting Cued Speech into text with high accuracy. The authors were interested in the fusion and the recognition part of the components, and details of image processing techniques are not covered by this work.

In the conducted experiments, it was assumed that lip shape and hand shape components are synchronous. Based on previous studies, however, there might be asynchrony between the two components (Aboutabit et al., 2006). Late fusion (Potamianos et al., 2003), coupled HMMs (Nefian et al., 2002) and product HMMs (Nakamura et al., 2002) would be used as possible alternatives to the state-synchronous fusion methods used in this work.

Although the results are promising, problems still persist. For example, in order to extract accurate features, some constraints were applied in recording, and the computational cost was not considered. Also, a possible asynchrony between the components should be further investigated. The current pilot study on Cued Speech recognition attempts to extend the research in areas related to deaf communities, by offering to individuals with hearing disorders additional communication alternatives. For practical use, however, many questions should be addressed and solved, such as speaker-, environment-independence, real-time processing, etc. The authors are still analyzing the remaining problems in the framework of the TELMA project.

5. Conclusion

In this article, vowel-, consonant-, and phoneme recognition experiments in Cued Speech for French were presented. To recognize Cued Speech, lip shape and hand components were integrated into a single component using concatenative feature fusion and multi-stream HMM decision fusion. The accuracies achieved were promising and comparable to those obtained when using an audio speech. Specifically, accuracy obtained was 87.8% for vowel recognition, 78.9% for consonant recognition, and 74.4% for phoneme recognition. In addition, isolated word experiments in Cued Speech in both normal-hearing and deaf subjects were also conducted obtaining a 94.9% and 89% accuracy, respectively. A multi-cuer experiment using data from both normal-hearing and deaf subject showed an 89.6% word accuracy, on average. This result indicates that training cuer-independent HMMs for Cued Speech using a large number of subjects should not face particular difficulties. Currently, additional Cued Speech data collection is in progress, in order to realize cuer-independent continuous Cued Speech recognition.

Acknowledgements

The authors would like to thank the volunteer cuers Sabine Chevalier, Myriam Diboui, and Clémentine Huriez for their time spending on Cued Speech data recording, and also for accepting the recording constraints. Also the authors would like to thank Christophe Savariaux and Coriandre Vilain for their help in the Cued Speech material recording. This work was mainly performed at GIPSA-lab, Speech and Cognition Department and was supported by the TELMA project (ANR, 2005 edition).

References

Aboutabit, N., 2007. Reconnaissance de la Langue Française Parlée Complété (LPC): Décodage phonétique des gestes main-lèvres. Ph.D. Dissertation, Institut National Polytechnique de Grenoble, Grenoble, France.

Aboutabit, N., Beautemps, D., Besacier, L., 2006. Hand and lips desynchronization analysis in French Cued Speech: automatic segmentation of hand flow. In: Proceedings of ICASSP'2006, pp. 633–636.

Aboutabit, N., Beautemps, D., Besacier, L., 2007. Automatic identification of vowels in the Cued Speech context. In: Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP).

Adjoudani, A., Benoît, C., 1996. On the integration of auditory and visual parameters in an HMM-based ASR. In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Humans and Machines. Springer, Berlin, Germany, pp. 461–471.

Auer, E.T., Bernstein, L.E., 2007. Enhanced visual speech perception in individuals with early-onset hearing impairment. Journal of Speech, Language, and Hearing 50, 1157–1165.

Bernstein, L., Auer, E., Jiang, J., 2007. Lipreading the lexicon and Cued Speech. In: la Sasso, C., Crain, K., Leybaert, J. (Eds.), Cued Speech and Cued Language for Children Who are Deaf or Hard of Hearing. Plural Inc. Press, Los Angeles, CA.

Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. In: Proceedings of International Conference on Spoken Language Processing, pp. 426–429.

Cornett, R.O., 1967. Cued Speech. American Annals of the Deaf 112, 3–13.

Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., Ney, H., 2007. Speech recognition techniques for a sign language recognition system. In: Proceedings of Interspeech, pp. 2513–2516.

Fleetwood, E., Metzger, M., 1999. Cued Language Structure: An Analysis of Cued American English Based on Linguistic Principles. Calliope Press, Silver Spring, MD (USA), ISBN 0-9654871-3-X.

Gibert, G., Bailly, G., Beautemps, D., Elisei, F., Brun, R., 2005. Analysis, synthesis of the 3D movements of the head, face and hand of a speaker using Cued Speech. Journal of Acoustical Society of America 118 (2), 1144–1153.

Gillick, L., Cox, S., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: Proceedings of ICASSP'89, pp. 532–535.

Hennecke, M.E., Stork, D.G., Prasad, K.V., 1996. Visionary speech: looking ahead to practical speechreading systems. In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Humans and Machines. Springer, Berlin, Germany, pp. 331–350.

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

Merkx, P., Miles, J., 2005. Automatic vowel classification in speech. An artificial neural network approach using cepstral feature analysis. Final Project for Math 196S, pp. 1–14.

Montgomery, A.A., Jackson, P.L., 1983. Physical characteristics of the lips underlying vowel lipreading performance. Journal of the Acoustical Society of America 73 (6), 2134–2144.

Nakamura, S., Kumatani, K., Tamura, S., 2002. Multi-modal temporal asynchronicity modeling by product HMMs for Robust. In: Proceedings of Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02), p. 305.

Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K., 2002. A coupled HMM for audio-visual speech recognition. In: Proceedings of ICASSP 2002.

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

Ong, S., Ranganath, S., 2005. Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Transactions on PAMI 27 (6), 873–891.

Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A., 2003. Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE 91 (9), 1306–1326.

Uchanski, R.M., Delhorne, L.A., Dix, A.K., Braida, L.D., Reedand, C.M., Durlach, N.I., 1994. 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.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2001. The HTK Book. Cambridge University Engineering Department.

Beautemps, D., Cathiard, M. A., Attina, V., Savariaux, C., 2012. Temporal organisation of Cued Speech Production. In Audiovisual Speech Processing, G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds), pp. 104-120.

6 Temporal organization of Cued Speech production

D. Beautemps, M.-A. Cathiard, V. Attina, and C. Savariaux

6.1 Introduction

Speech communication is multimodal by nature. It is well known that hearing people use both auditory and visual information for speech perception (Reisberg *et al.* 1987).[1] For deaf people, visual speech constitutes the main speech modality. Listeners with hearing loss who have been orally educated typically rely heavily on speechreading based on lips and facial visual information. However lipreading alone is not sufficient due to the similarity in visual lip shapes of speech units. Indeed, even the best speechreaders do not identify more than 50 percent of phonemes in nonsense syllables (Owens and Blazek 1985) or in words or sentences (Bernstein *et al.* 2000).

This chapter deals with Cued Speech, a manual augmentation for lipreading visual information. Our interest in this method was motivated by its effectiveness in allowing access to complete phonological representations of speech for deaf people, from the age of one month, access to language and eventually performance in reading and writing similar to that of hearing people. Finally with the current high level of development of cochlear implants this method helps facilitate access to the auditory modality.

A large amount of work has been devoted to the effectiveness of Cued Speech but none has investigated the motor organization of Cued Speech production, i.e. the coarticulation of Cued Speech articulators. Why might the production of an artificial system as long ago as 1967 be of interest? Apart from the clear evidence that such a coding system helps in acquiring another artificial system such as reading, Cued Speech provides a unique opportunity to study lip–hand coordination at syllable level. This contribution presents a study of the temporal organization of the manual cue in relation to the movement of the lips and the acoustic indices of the corresponding speech sound, in order to characterize the nature of the syllabic structure of Cued Speech with reference to speech coarticulation.

6.2 Overview on manual cueing

6.2.1 Cued Speech system

Cued Speech was designed to complement speechreading. Developed by Cornett (Cornett 1967; Cornett 1982), this system is based on the association of lip shapes with cues formed by the hand. While uttering, the speaker uses one hand to point out specific positions around the mouth, palm towards the speaker so that the speechreader can see the back of the hand simultaneously with the lips. The cues are formed along two parameters: hand placement and hand shape. Placements of the hand code vowels while hand shapes (or configurations) distinguish the consonants. In English, eight hand shapes and four hand placements are used to group phonemes (Figure 6.1). The primary factor in assignment of phonemes to groups associated with a single hand shape or hand placement is the visual contrast at the lips (Woodward and Barber 1960). For example, phonemes [p], [b], and [m], with identical visual shapes, are associated to different hand shapes, while phonemes easily discriminated from the lips alone are grouped in the same configuration. Each group of consonants is assigned to a hand shape. For the highest frequency group the hand shapes that require less energy to execute are chosen. The frequency of appearance of consonant clusters and the difficulties these might present in changing quickly from one hand configuration to another are also taken into account.

Vowel grouping was worked out similarly, high priority being given to the ease of cueing for diphthongs. Vowel positions are indicated with one of the fingers. The middle finger is used for all the consonant cues except those of the [d, p, ʒ], [j, tʃ], and [l, ʃ, w] groups, for which the index finger is used. An exception exists for the [j, tʃ] group: The middle finger is used as the pointer for the mouth position, while the index finger is used for the chin, throat, and side positions.

The information given by the hand is not sufficient for phoneme identification. The visible information of the lips is still essential. The identification by the lips of a group of look-alike consonants and the simultaneous identification of a group of consonants by the hand shape result in the identification of a single consonant. Thus the combination of hand shape and hand location with the information visible on the lips identifies a single consonant-vowel syllable.

The system was based on the CV syllabification of speech. The syllable strings $C(C_n)V(C_m)$, as complex as they can be, are broken down into CVs each CV being coded both by the shape of the hand for the consonant and by the place of the hand on the face side for the vowel. When a syllable consists only of a vowel, this V syllable is coded using hand shape N°5 (Figure 6.1), with the hand at the appropriate position for the vowel. If a consonant cannot be linked to a vowel, as is the case when two consonants follow each other or when a

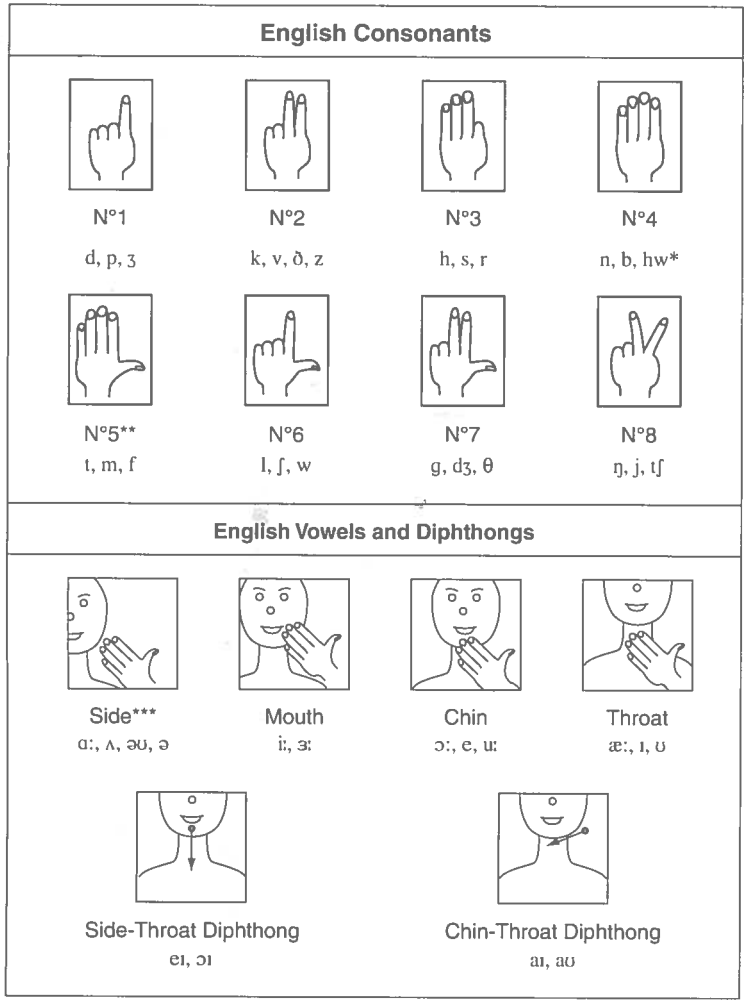


Figure 6.1 Visible cues for English consonants, vowels, and diphthongs (from Cornett 1967)
Notes: * Some teachers of Cued Speech may prefer to cue /hw/ as /h/ plus w; ** This hand shape is also used for a vowel without a preceding consonant; *** The side position is used also when a consonant is cued without a following vowel.

consonant is followed by a schwa, the hand is placed at the side position with the associated consonant hand shape. Diphthongs are considered to be pairs of vowels (VV) and are therefore cued with a shift from the position of the first vowel towards the position of the second vowel (cf. Figure 6.1).

Finally, in the adaptation of Cued Speech to other languages (more than 50 in Cornett 1988), the criterion of compatibility with the English version was given a higher priority than phoneme frequency of the considered language. An additional position next to the cheekbone is needed for coding all vowels used in French, German, Italian, and Spanish. In German, some hand shapes code consonant clusters directly (as it is the case for the frequently encountered [ʃt], [ʃp], [tʃ], and [ʃv] clusters) to avoid affecting speech rhythm, a problem that would occur with frequent consecutive hand shape modifications (Pierre Lutz, personal communication).

6.2.2 *Perceptual effectiveness of manual cueing*

The perceptual effectiveness of Cued Speech has been evaluated in many studies. Nicholls and Ling (1982) presented eighteen profoundly hearing-impaired children with CV or VC syllables made of twenty-eight English consonants combined with the vowels [i, a, u] in seven conditions, with auditory, lipreading and manual cue presentations combined. A similar test was conducted with familiar monosyllabic nouns inserted in sentences. Under audition (A) alone, subjects correctly identified 2.3% syllables, whereas scores in lipreading (L), audition + lipreading (AL), manual cues alone (C) and audition + manual cues (AC) reached 30 to 39% without significant differences. Higher scores were obtained with lipreading + manual cues (LC = 83.5%) and audition + lipreading + manual cues (ALC = 80.4%). This last result was also found for the test sentences where the mean scores for key words reached more than 90% in the LC and ALC conditions.

Uchanski *et al.* (1994) confirmed the effectiveness of Cued Speech for the identification of various conversational materials (sentences with high or low predictability). The highly trained subjects obtained mean scores varying from 78% to 97% with Cued Speech against 21% to 62% with lipreading alone.

For French, Alégria *et al.* (1992) tested deaf children who had been exposed to Cued Speech early (before the age of three) both at home and at school. They compared these early-exposed children with children exposed late from the age of six and only at school. The subjects exposed early and intensively to Cued Speech were better lipreaders and better Cued Speech readers in identifying words and pseudo words. It seems that early exposure to Cued Speech allows children to develop more accurate phonological representations (Leybaert 2000). Thereafter their reading and writing skills progress in a similar way to those of hearing children since Cued Speech early-exposed deaf children can use precise grapheme to phoneme correspondences (Leybaert 1996).

Finally, the studies on working memory of Cued Speech deaf children reveal that they use a phonological loop probably based on the visual components of

Cued Speech: mouth shapes, hand shapes, and hand placements (Leybaert and Lechat 2001).

6.2.3 *Phonological representations in Cued Speech*

Fleetwood and Metzger (1998, p. 29) proposed the term *cuem*, which 'refers to an articulatory system that employs non-manual signals (NMS) found on the mouth and the hand shapes and hand placements of Cued Speech to produce visibly discrete symbols that represent phonemic (and tonemic) values'. Neither the production nor the reception of acoustic information or of speech is implied in the term '*cuem*'. The authors maintain that Cued Speech can be delivered without production of an acoustic speech signal. This is the usual situation in an interpreting task where the Cued Speech speaker translates silently into cues for deaf people as the hearing speaker is talking. The authors also refer to the studies of Nicholls (1979) and Nicholls and Ling (1982), which claim that the acoustic signal is not necessary in Cued Speech. Nicholls and Ling (1982) found no advantage of audition for syllable identification; the score obtained in the Cued Speech presentation (manual cues alone; C = 36%) was not significantly different from the Audition + Cued Speech score (AC = 39%). Similarly, there was no difference between the lipreading + Cued Speech condition (LC = 83.5%) and audition + lipreading + Cued Speech condition (ALC = 80.4%). The pattern of results was quite different for key words; a better score was recorded for the AC condition (59.2% for low predictability sentences and 68.8% for high predictability) than for the C condition (respectively, 42.9% and 50.0%); in LC and ALC, key word scores were similar, around 96%, revealing a ceiling effect. The advantage of the AC condition for key words in sentences was explained as the use of supra-segmental information. Nicholls and Ling (1982) concluded that speech information in Cued Speech can be perceived through vision alone. Thus Fleetwood and Metzger (1998) proposed that the phonological representations underlying the perception of Cued Speech be defined only by the mouth shapes, hand shapes and hand positions (Fleetwood and Metzger 1998).

However we think this position is perhaps too restrictive. In their taxonomy of tactile speech perception methods, Oerlemans and Blamey (1998) proposed to distinguish between the speech-based and language-based tactile codes. The code was considered speech-based when the user had direct access to the articulatory gestures, as in the Tadoma method (Reed *et al.* 1985), where the blind-deaf user directly touches the vocal tract of the speaker, placing a hand on the talker's face. In contrast, the tactile version of Sign Language was classified as language-based. If the same taxonomy for visual perception is used, speech-based and language-based methods can be distinguished. In our view, Cued Speech is clearly a speech-based code, since the visual lip and mouth information

directly results from the articulatory gestures. The fact that the emission of sound is not necessary for the production or reception of Cued Speech does not mean that the code is purely visual.

We maintain that Cued Speech is speech-based in the sense that articulatory gestures are recovered from the visual modality. As we will show, these visual lip cues are highly dependent on the speech flow for their temporal time-course.

6.2.4 *Face and hand coordination for Cued Speech*

The fact that manual cues must be associated with lip shapes to be effective for speech perception reveals a real coordination between hand and mouth. As yet no fundamental study has been devoted to the analysis of the skilled production of Cued Speech gestures, i.e. the temporal organization existing between lip movements and hand gestures in relation to the acoustic realization.[2] Except for a theoretical aside by Cornett pointing out some consonant clusters where speech should be delayed to leave the hand enough time to reach the correct position (Cornett 1967, p. 9), the problems of cue presentation timing are only incidentally touched on in the course of technological investigations.[3]

In the Cornett Autocuer system (Cornett 1988), cues are defined from the sound recognition of the pronounced word and are displayed on one group of LEDs on glasses worn by the speechreader. The whole process involves a delay of 150 to 200 ms for the cue display, compared to the production time of the corresponding sound. This system, designed for isolated words, attained 82% correct identification.

In the system for the automatic generation of Cued Speech developed by Duchnowski *et al.* (2000) for American English the cues are presented with the help of pre-recorded hands, and rules for temporal coordination with sound are proposed. This system uses a phonetic recognizer of audio speech to obtain a list of phones which are then converted to a time-marked stream of cue codes. The appropriate cues are visually displayed by superimposing hand shapes on the video signal of the speaker's face. The display is presented with a delay of two seconds, a delay that is necessary to correctly identify the cue (since the cue can only be determined at the end of each CV syllable). The superimposed hand shapes are always digitized images of a real hand. Scores of correct word identification reached a mean value of 66% and were higher than the 35% obtained with speechreading alone but they were still under the 90% level obtained with Manual Cued Speech. This 66% mean score was obtained for the more efficient display, called 'synchronous', in which 100 ms were allocated to the hand target position and 150 ms to the transition between two positions. In this 'synchronous' display, the time at which cues were displayed was advanced by 100 ms relative to the start time determined by the recognizer; i.e. for stop consonants, the detected instant of acoustic silence (Duchnowski, personal communication). This advance was fixed empirically by the authors.

In these investigations, the time of cue presentation is related only to the corresponding acoustic events: there is no discussion of the relation between cue presentation and lip motion. However it is well known that lip gesture can anticipate acoustic realization (Perkell 1990; Abry *et al.* 1996, for French). In the Autocuer system, the cue presentation is automatically later than lip motion. The impact of this delay was not evaluated and the identification scores were still high for isolated words. On the other hand, the closer timing of the hand to the acoustic realization is a key factor for the improvement of the Duchnowski *et al.* (2000) system. It should be stressed that this latter system functions with continuous speech and uses hand cues; thus it is closer to the natural Cued Speech conditions than the Autocuer.

6.3 First results on Cued Speech production

It has been mentioned that the Cued Speech system is based on CV syllabic organization, the hand giving information on both the consonant and the vowel. The shifting of the hand between two hand positions corresponds to the vocalic transition and the hand shape (or finger configuration) constitutes the consonant information. The main objective of this section is to determine precisely how the hand gesture *co-produces* the consonantal and vocalic information. In short, is the temporal organization of vocalic and consonant hand gestures similar to the organization of speech, as revealed by the classical model of coarticulation (Öhman 1967b)?

To this end we will examine a comparative study of the temporal organization of manual cues with lip and acoustic gestures. The temporal organization of Cued Speech articulators is analysed from a recording of a Cued Speech speaker. The time-course of the lip parameter and the hand x y coordinates are investigated in relation to acoustic events. The occurrence of hand shape formation is measured in relation to hand position.

6.3.1 *The Cued Speech speaker*

The Cued Speech speaker is a thirty-six-year-old French female who has been using Cued Speech at home with her hearing-impaired child for eight years. She qualified in Cued Speech for French in 1996 and regularly translates into Cued Speech code at school.

6.3.2 *Audiovisual data*

The different parameters involved in the analysis were derived from the processing of an audiovisual recording of the Cued Speech speaker. The recording

was made in a soundproof booth, at 50 frames per second. A first camera in wide focus was used for the hand and the face. A second one in zoom mode dedicated to the lips was synchronized with the first one. The lips were made up in blue. Coloured marks were placed on the hand for tracking hand movement. A second experiment was devoted to the analysis of hand shape formation. In this investigation the Cued Speech speaker was wearing a data glove with two sensors for each of the five fingers covering the first and second articulation with an additional sensor between the fingers. The sensor raw data has a linear relationship to the deviation angle between two segments of a finger articulation. The hand position is located with the use of coloured landmarks placed on the glove. In both experiments, the subject wore opaque goggles to protect her eyes against the halogen spotlight and her head was maintained in a fixed position with a helmet. Blue marks were placed on the speaker's goggles as reference points.

Two Betacam recorders had to be synchronized. At the beginning of the recording session a push button was activated, switching on the set of LEDs (placed in the field of the two cameras) during the first A-frame instant of the video image. This enabled the correspondence between the time codes of the two cameras to be calculated. The audio line was digitized in synchrony with the video image. When the data glove was used a system for synchronization with the audio part was needed. In this system an audio signal was released at the thumb and index finger contact and recorded on the audio line of the video tape. Finger contact resulted in a plateau on the raw data from the glove sensors measuring the movement of the two fingers which allowed synchronization of the data glove with the audio recording. The delay between the time codes of the two cameras was calculated using the first system.

The image processing-based automatic extraction system developed at ICP (Lallouache 1991) provided a set of lip parameters every 20 ms. We chose to explore the temporal evolution of the between-lip area (S), which is a good parameter for characterizing sounds at both the acoustic and articulatory levels. In synchrony with lip area parameter and audio signal, the x and y coordinates of the hand landmark placed near the wrist were extracted. The onset and offset of hand and lip gesture transitions were manually labelled at the acceleration peaks (Schmidt 1988; Perkell 1990).[4] On the audio signal, the onsets and offsets of the acoustic realization for consonants and the vowels were also labelled.

These two experiments had complementary objectives. The first explored the movement of the hand from one hand position to another, i.e. the carrier gesture of Cued Speech. Because the hand shape was fixed, interference with hand shape formation was avoided. The second experiment tested the timing of the production of hand shape formation in relation to hand position.

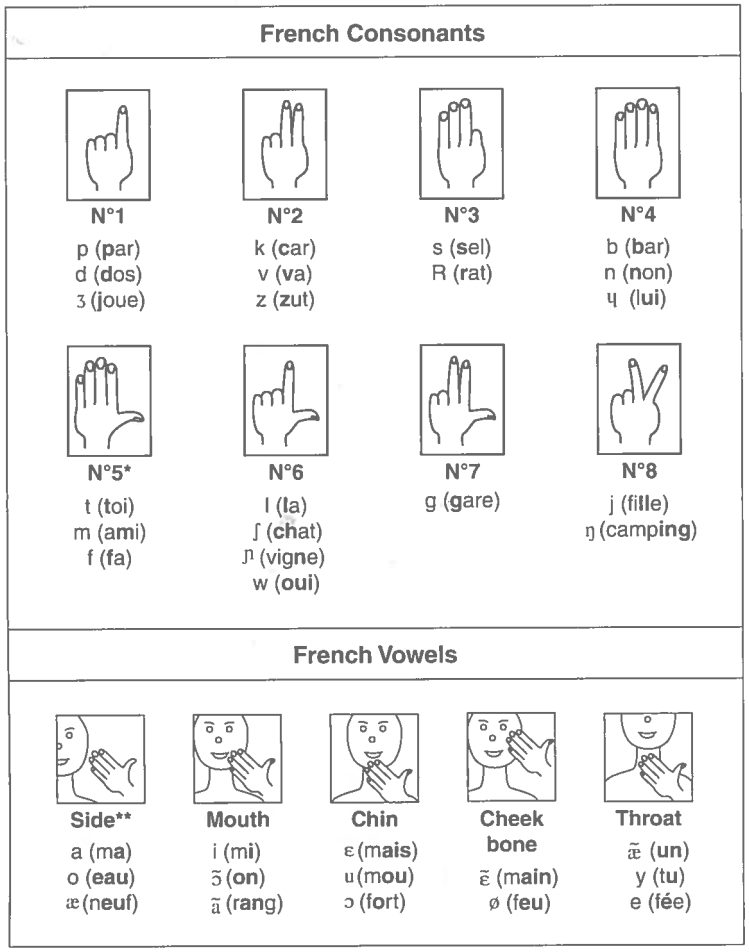


Figure 6.2 Hand placements and hand shapes used in French.
Notes: * This hand shape is also used for a vowel not preceded by a consonant. ** This position is also used when a consonant is isolated or followed by a schwa.

6.3.3 *Experiment 1: Hand displacement*

6.3.3.1 Corpus Displacement of the hand was analysed with [$CaCV_1CV_2CV_1$] sequences made up of [m, p, t] consonants for C combined with the vowels [a, i, u, ø, e] for V_1 and V_2, i.e., the vowel with the best visibility for each of the five hand positions of the French code (Figure 6.2).

The choice of consonants was fixed according to their labial or acoustic characteristics: [m, p] present a typical bilabial occlusion that appears on the lip video signal as a null lip area, and [p, t] are marked by a clear silent period. The hand shape was fixed during the production of the whole sequence: [m] and [t] are coded with the same hand shape as isolated vowels are (hand shape N°5), while [p] is associated with hand shape N°1. The whole corpus contained twenty sequences, such as [mamamima], for each of the three consonants. A control condition with no consonant for the second (S_2) and third (S_3) syllables was also used, i.e., [$maV_1V_2mV_1$], made up of the vowels [a, i, u, ø, e] for V_1 and V_2 (e.g., [maaima]). We thus obtained twenty additional sequences. For each of the eighty sequences the analysis was carried out on [CV_2] or [V_2] in the absence of a consonant (i.e. on transitions from the S_2 syllable towards S_3 and from S_3 towards S_4), in order to avoid the biases inherent at the beginning of the gesture.

Consider, for example, the [pupøpu] $S_2S_3S_4$ sequence (from the whole [papupøpu] $S_1S_2S_3S_4$ sequence) in Figure 6.3. The following events were determined for the hand trajectory:

- M1 is the beginning of the hand gesture (determined by acceleration peak) towards the position corresponding to S3;
- M2 is the hand position target reached (coding S_3). It is determined by peak deceleration and maintained until M3, the instant of peak acceleration and the time at which the hand begins the gesture towards the following position for S_4 coding;
- M4 corresponds to the S_4 hand target reached. In the case of non-concordance of acceleration events on x and y, the first M1 and M3 and the last M2 and M4 points were considered. The hand target is defined as a time when the hand reaches the target both in x and y, i.e. between the end of the transition and the beginning of the transition towards the following target.
- For lip area, L1 marks the beginning of the vowel gesture. This was easily detectable for sequences with [p] and [m] consonants, since L1 was coincident with the end of the lip closure phase. We used the beginning of the acoustical silence to determine L1 in the case of sequences with [t]. L2 is the lip target instant labelled at the end of the lip transition towards the maximal lip-opening target (in the case of absence of a lip vocalic plateau the acceleration peak coincided with the maximal lip value).
- For the corresponding acoustic signal A1 marks the beginning of the consonant of the S_3 syllable.

6.3.3.2 Results For this analysis we took into account only the transitions from the S_2 syllable towards S_3 and from S_3 towards S_4. In order to evaluate the coordination between lip, hand, and sound, we determined different duration

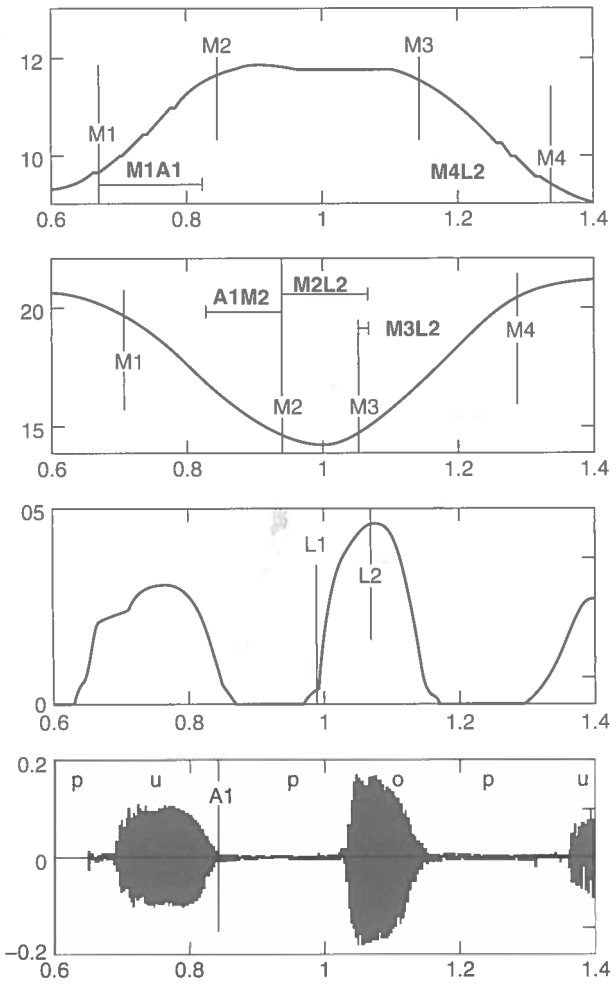


Figure 6.3 Speech vs. lips and hand motion for the [pupøpu] sequence. From top to bottom: horizontal x (cm) and vertical y (cm) hand motion paths are shown in the top two panes (an increase in x means that the hand moves from the face to the right side, an increase in y means the hand moves towards the bottom of the face); the two bottom panes contain the lip area (cm^2) time-course and the corresponding audio signal.

intervals. From the events labelled on each signal, we located the following intervals:

- M1A1 corresponds to the interval between the beginning of the manual gesture for S_3 and the acoustic consonant closure;

- A1M2 is the interval between the acoustic consonant closure and the onset of the hand target;
- M2L2 is the interval between the onset of the hand target and the onset of the lip target of the vowel of S_3;
- M3L2 is the interval between the lip target and the beginning of the following hand Cued Speech gesture.

All intervals were computed as arithmetic differences, i.e. the second label minus the first. For example, M1A1 = A1 – M1 (ms). For sequences without a consonant in S_2S_3, such as [maaima], mean values of 183 ms were obtained for the M1A1 interval and 84 ms for the A1M2 interval, the A1 instant corresponding to the onset of the glottal stop that the speaker inserted between the production of the two consecutive vowels. The hand target is clearly in advance of the lip area target (M2L2 = 73 ms). The following hand gesture begins after the lip target (M3L2 = –84 ms).

For sequences with consonants, such as [mamamima], a mean value of 239 ms was obtained for the M1A1 interval. This differed significantly from the consonant acoustical beginning. The A1M2 interval reached a mean value of 37 ms. The hand target was therefore reached during the acoustic realization of the consonant in a quasi-synchronization with the acoustic closure event. The lip target was usually reached after the corresponding hand target since a mean value of 256 ms for M2L2 was obtained. Finally the hand movement towards the following syllable placement began, on average 51 ms before the peak of the vowel lip target (M3L2 = 51 ms).

In conclusion, the hand gesture begins before the acoustical onset of the CV syllable (183 ms and 239 ms) and reaches the hand position largely before the lip target, in fact, during the consonant.

6.3.4 *Experiment 2: Hand shape formation*

This experiment examined the association between hand shape formation and consonant information. The corpus was selected so as to have only one finger component per consonant hand shape transition in each sequence. For example, the transition from [p] to [k], i.e., from hand shape N°1 to hand shape N°2 (Figure 6.2), is effected by the extension of the middle finger. Thus the modification of the hand shape required only one main sensor of the data glove. This choice was made to simplify data reading.

6.3.4.1 Corpus Hand shape formation was analysed for two kinds of sequences:

- (i) [mVC_1VC_2V] sequences with the same vowel (V = [a] or [ɛ]) were designed to investigate consonant variation. The C_1 and C_2 consonants were [p] and [k], [s] and [b], or [b] and [m]. This choice resulted in hand

Figure 6.4 Cues for the [mabuma] sequence.

shape modification at fixed hand placement (for example, the [mapaka] sequence is coded at the side position with the appropriate hand shape modifications). Ten repetitions of each sequence were recorded. The analysis focused on the C_1V syllable, resulting in 60 syllables (10 repetitions × 3 consonant groups × 2 vowels).

- (ii) [$mV_1C_1V_2C_2V_1$] sequences varied both vowel and consonant, thus involving both hand shape modification and hand placement transitions. The C_1 and C_2 consonants were [p] and [k], [ʃ] and [g], [s] and [b], or [b] and [m]. The V1 and V2 vowels were [a] and [u], [a] and [e], or [u] and [e]. Thus, for the [mabuma] sequence (see Figure 6.4) coding implicates a transition of the hand from the side position towards the chin and then back to the side position, while the hand shape changes from the N°5 to N° 4 configuration and back to the N°5. The change from 5 to 4 is realized with the thumb facing towards the palm. Five repetitions of each sequence were recorded. The analysis focused on the C_1V_2 syllable, resulting in 60 syllables (5 repetitions × 4 consonant groups × 3 vowel groups). Since an error occurred in the recording for a realization of a [mubemu] sequence, 59 sequences were considered for this corpus.
- In all sequences (with vowel-not-changed and vowel-changed), the beginning of the consonant (A1) is labelled on the acoustic signal. The beginning of the finger gesture is marked at the D1 maximum point of acceleration and the end is marked at the D2 deceleration point of the corresponding raw data trajectory. Similarly for sequences with hand movement from one hand position to another (case of vowel-changed sequences), the hand trajectory was marked by M1 and M2 (Figure 6.5).

6.3.4.2 Results It should be remembered that the analysis focused only on the second syllable. In order to evaluate the coordination between sound, finger, and hand different duration intervals were derived from the events labelled on each signal. For all the sequences:

- D1A1 is the interval between the beginning of the finger gesture and the beginning of the corresponding acoustic consonant;
- A1D2 corresponds to the interval between the beginning of the acoustic consonant and the end of the digit movement.

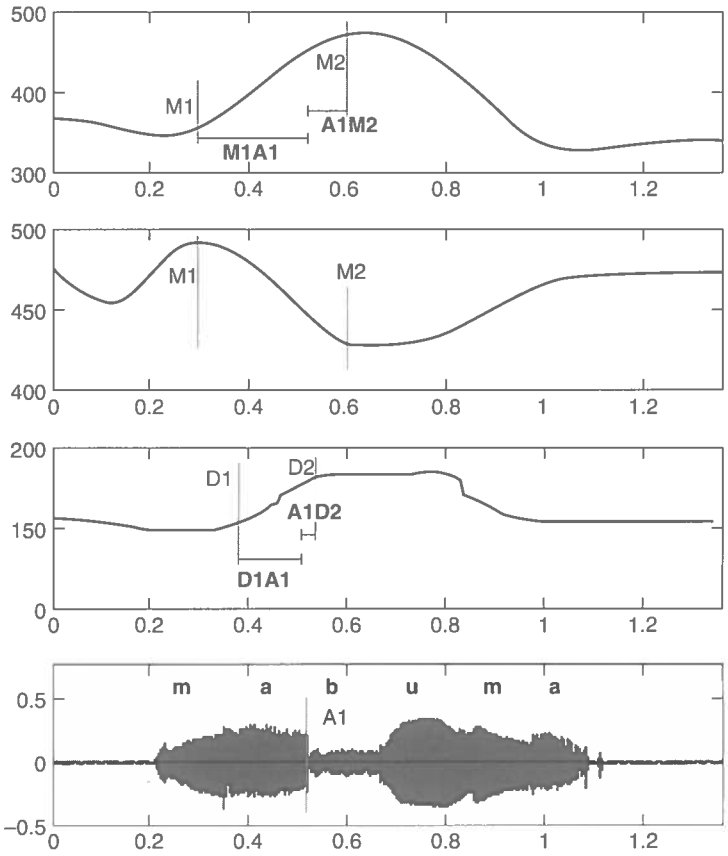


Figure 6.5 Speech vs. lips and hand motion for the [mabuma] sequence. From top to bottom: Horizontal x (cm) and vertical y (cm) hand motion paths are shown in the top two frames (an increase in x means moving the hand from the face to the right side, an increase in y means moving the hand towards the bottom of the face); the bottom two frames contain the temporal deviation of the raw data of the thumb first articulation glove sensor and the corresponding acoustic signal.

In addition, for vowel-changed sequences:

- M1A1 is the interval between the beginning of the hand movement and the beginning of the acoustic consonant;
- A1M2 corresponds to the interval between the acoustic consonantal beginning and the end of the hand gesture.

As in the first experiment, all intervals were computed as arithmetic differences, i.e., the second label minus the first label; for example, D1A1 = A1 – D1 (ms).

For the vowel-not-changed sequences (sequences with only hand shape change, the hand placement being maintained), we obtained mean values of 124 ms for the D1A1 interval and 46.5 ms for the A1D2 interval. Thus the beginning of the finger gesture precedes the acoustic onset of the consonant. The finger finishes its movement just after the beginning of the acoustic realization of the consonant.

For the vowel-changed sequences (both hand shape and hand placement change), mean values of 171 ms for the D1A1 interval and −3 ms for the A1D2 interval were obtained. Thus, for the finger gesture relative to the sound, we observed the same pattern as in the previous result. As regards the hand gesture, mean values of 205 ms for the M1A1 interval and 33 ms for the A1M2 interval were obtained. The hand gesture begins before the finger gesture and consequently well before the onset of the acoustic consonant. The hand target is reached at the beginning of the acoustic realization of the consonant. Finally if we compare duration for hand shape formation in reference to hand transition between two hand placements, we note that the consonant finger gesture is encapsulated in the hand transition.

6.3.5 *Summary of the two experiments*

There is a noticeable convergence in the results of the two experiments. To summarize, for hand position, it was observed that

- the movement of the hand towards its position begins about 200 ms before the acoustic beginning of the CV syllable. This implies that the gesture begins during the preceding syllable, i.e. during the preceding vowel;
- the hand target is attained at the beginning of the acoustic consonant onset;
- this hand target is therefore reached on average 250 ms before the vowel lip target.

These three results reveal the *anticipatory* gesture of the hand motion relative to the lips as the hand placement gesture covers the duration of the whole syllable, with a temporal advance over the vocalic speech gesture.

Finally, it was observed from the data glove that the hand shape is completely formed at the instant when the hand target position is reached. In addition it was noticed that the hand shape formation gesture uses a large part of the hand transition duration.

6.4 General discussion

6.4.1 *Cued Speech co-production*

The consideration of the two Cued Speech components within the framework of speech control has a bearing on the future elaboration of a quantitative control

model for Cued Speech production. For transmitting consonant information, the control type is figural, i.e., a postural control of the hand configuration (finger configuration). The type of control for transmitting the vowel information is a goal-directed movement performed by the wrist and carried by the arm. These two controls are linked by an in-phase locking. On the other hand, for speech, there are three types of control:

- (i) The mandibular open-close oscillation is the control of a cycle, self-initiated and self-paced (MacNeilage 1998; Abry *et al.* 2002). This is the control of the carrier of speech, the *proximal* control that produces the syllabic rhythm.
- (ii) Following Öhman (1966; see also Vilain *et al.* 2000), the vowel gesture is produced by *global* control of the whole vocal tract – from the glottis to the lips –, i.e., a figural or postural motor control type.
- (iii) The consonant gesture is produced by the control of contact and pressure performed *locally* along the vocal tract.

The carried articulators (tongue and lower lip) together with their coordinated partners (upper lip, velum, and larynx) are involved in these two distal (global and local) controls.

The mandibular and vowel controls are coupled by in-phase locking. Consonantal control is typically in-phase with the vowel for the initial consonant of a CV syllable. But it can be out-of-phase for the coda consonant in a CVC syllable. Finally consonant gestures in clusters within the onset or the coda can be in-phase (e.g., [psa] or [aps]) or out-of-phase ([spa] or [asp]).

As for speech, Cued Speech vowels and consonants depend on the wrist-arm *carrier* gesture, which is analogous to the mandibular rhythm. The control of the vowel *carried* gesture is a goal-directed movement, which aims at local placement of the hand around the face. On the other hand, the consonant *carried* gesture is a postural (figural) one. Thus the two types of control in Cued Speech are inversely distributed in comparison to speech: the configuration of global control of the speech vowel corresponds to a local control in Cued Speech, whereas the local control for the speech consonant corresponds to a global control in Cued Speech.

Once speech rhythm has been converted into Cued Speech rhythm (that is a general CV syllabification with some cluster specificities as in German), the two carriers (mandible and wrist) can be examined with respect to their temporal coordination, i.e., phasing. This CV re-syllabification means that every consonantal Cued Speech gesture will be in phase with its vocalic one, which is not always the case in speech for languages that have more than just CVs. Unlike speech the Cued Speech consonant gesture never hides the beginning of the in-phase vocalic gesture (Öhman's model). As for the phasing of the two carried vowel gestures, our experiments made clear that the Cued Speech vowel gesture did anticipate the speech vowel gesture.

6.4.2 Towards a topsy-turvy vision of Cued Speech

The coordination obtained between hand, lips, and sound confirms, in our opinion, the in-principle validity of the advance (lead) of the hand on the sound, programmed as an empirical rule by Duchnowski *et al.* (2000) for their automatic Cued Speech display. Of course the range of this anticipatory behaviour will vary with different speakers, rates, etc., and should be examined by subsequent articulatory studies.

These considerations result in quite a rather upside-down vision of the Cued Speech landscape. The *in-principle* advance of the hand over the lips (and on sound) is crucial for the question of the integration of manual and lip information. Currently Cued Speech has been designed as an augmentation for lip disambiguation. A general pattern seems to appear from our data on the temporal organization of hand and lip gestures in the production of successive CV sequences. The hand attains the vowel placement at the beginning of the CV syllable and moves from that position towards a new one even before the peak acoustic realization of the vowel and before the corresponding vocalic lip target is reached. It seems therefore that production control imposes its temporal organization on the perceptual processing of Cued Speech. This organization leads us to think that the hand placement first gives a set of possibilities for the vowel then the lips determine a unique solution. This hypothesis has been successfully tested within the framework of gating experiments for phoneme identification where recognition of CV syllables has been evaluated across the time course of available online information resulting from the coordination of hand and lip motion (see Cathiard *et al.* 2004; Troille *et al.* 2007; Troille 2009; Troille *et al.* 2010). These studies demonstrated the ability of deaf subjects to recover the anticipatory behaviour of the hand in their Cued Speech perception.

6.5 Acknowledgments

Many thanks to Martine Marthouret, speech therapist at Grenoble Hospital, for helpful discussions; to Mrs G. Brunnel, the Cued Speech speaker, for enduring the recording conditions, and C. Abry and J. L. Schwartz for their stimulating suggestions. This work has been supported by the Remediation Action of the French Research Ministry 'Programme Cognitique', a 'Jeune équipe' project of the CNRS (French National Research Centre) and a BDI grant from CNRS.

Ming, Z., Beautemps, D., Feng, G., 2013. GMM Mapping Of Visual Features of Cued Speech From Speech Spectral Features. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP), 2013.

GMM Mapping Of Visual Features of Cued Speech From Speech Spectral Features

Zuheng Ming, Denis Beautemps, Gang Feng

GMM Mapping Of Visual Features of Cued Speech From Speech Spectral Features

Zuheng Ming[1,2], *Denis Beautemps*[1,2], *Gang Feng*[1,2]

[1] Univ. Grenoble Alpes, GIPSA-lab, F-38402 Saint Martin d'Hères Cedex
[2] CNRS, UMR 5216

Denis.Beautemps@gipsa-lab.grenoble-inp.fr

Abstract

In this paper, we present a statistical method based on GMM modeling to map the acoustic speech spectral features to visual features of Cued Speech in the regression criterion of Minimum Mean-Square Error (MMSE) in a low signal level which is innovative and different with the classic text-to-visual approach. Two different training methods for GMM, namely Expecting-Maximization (EM) approach and supervised training method were discussed respectively. In comparison with the GMM based mapping modeling we first present the results with the use of a Multiple-Linear Regression (MLR) model also at the low signal level and study the limitation of the approach. The experimental results demonstrate that the GMM based mapping method can significantly improve the mapping performance compared with the MLR mapping model especially in the sense of the weak linear correlation between the target and the predictor such as the hand positions of Cued Speech and the acoustic speech spectral features.

Index Terms: Cued Speech, LSP, MFCC, GMM mapping.

1. Introduction

The framework of this paper is speech communication for deaf orally-educated people. Speech is concerned here in its multimodal dimensions and in the context of automatic processing. Indeed, the benefit of visual information for speech perception (called "lip-reading") is widely admitted. However, even with high lip reading performances, without knowledge about the semantic context, speech cannot be thoroughly perceived. The best lip readers scarcely reach perfection. On average, only 40 to 60% of the phonemes of a given language are recognized by lip reading ([1]), and 32% when relating to low predicted words ([2]) with the best results obtained amongst deaf participants - 43.6% for the average accuracy and 17.5% for standard deviation with regards to words ([3], [4]), with an advantage of deaf women with 33.3 % for females but only 23.5 % for males (see the study on word perception of [5]). The main reason for this lies in the ambiguity of the visual pattern. However, as far as the orally educated deaf people are concerned, the act of lip-reading remains the main modality of perceiving speech. This led Cornett ([6]) to develop the Cued Speech system (CS) as a complement to lip information. CS is a visual communication system that makes use of hand shapes placed in different positions near the face in combination with the natural speech lip-reading to enhance speech perception from visual input. This is a system (See Figure 1) where the speaker, facing the perceiver, moves his hand in close relation with speech (See [7] for a detailed study on CS temporal organization in French language).

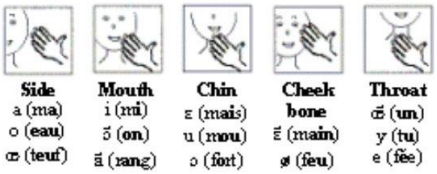


Figure 1: *Hand placements for coding vowels in French Cued Speech (from [7]).*

CS is largely improving speech perception for deaf people ([2]), relating to the identification of American-English syllables; and ([8]), relating to the identification of sentences in American-English language (scores between 78 and 97%). Moreover, CS offers to deaf people a thorough representation of the phonological system, inasmuch as they have been exposed to this method since their youth, and therefore it has a positive impact on the language development (see [9] for CS in French language).

As we have seen from this short review, the Cued Speech method offers a real advantage for complete speech perception. Nowadays, one of the important challenges is the question of speech communication between normal hearing people who do not practice CS but produce acoustic speech and deaf people with no auditory rests who use lip-reading completed by CS code for speech perception. To solve this question, one can use a human translator. Another solution is based on the development of automatic translation systems. This paper is a contribution to this topic. In a more general framework, two sources of information could contribute to this translation operation: (i) The *a priori* knowledge of the phonetic, phonologic and linguistic constraints; (ii) the *a priori* knowledge of the correlations between the different vocal activity: neuronal and neuro-muscular activities, articulatory movements, aerodynamic parameters, vocal tract geometry, face deformation and acoustic sound. Different methods allow their modelling and their optimal merging with the input signals and the output ones. On an axis ordering the methods in function of their dependence upon the used language, one can find at the two extremities: (i) the method using the phonetic level of interface, combining speech recognition and speech synthesis to take into account speech phonology organization. Note that the recognition and synthesis processing can call on very various modelling techniques. If the phonetic models based on Hidden Markov Models (HMM) are the basis of the main recognition systems, synthesis based on concatenative of multi-parametered units of various lengths is still very popular. Note the increasing interest of synthesis by trajectory models based on HMMs ([10], [11]) allowing the jointed learning of the recognition and synthesis systems; (ii) The methods using the correlation between signals without the help of the phonetic level but using various mapping techniques. These techniques capture the correlations between

input and output samples using Vector Quantification or Gaussian Mixture Model ([12], [13]). For Cued Speech, the classic method to convert audio speech to CS components consists of coupling a recognition system to a text-to-visual speech synthesizer ([7], [14], [15], [16]). The link between the two systems requires at least the phonetic high level. Before this work, no studies aimed at using the very low signal level. This work is a contribution to this challenge in the case of oral French vowels. A new approach based on the mapping of speech spectral parameters with the visual components made of CS and lip parameters is proposed. In this context the objective of the mapping process is to deliver visual parameters that can be used as target parameters for visual speech synthesis. In this paper, we explore the GMM-based mapping. Berthommier ([17]) applied a similar procedure to estimate the DCT coefficients of the lip region in the objective of speech enhancement of noisy signal. The present paper deals with normal speech in the case of speech supplemented by CS. In the following, we will start by defining the audio and visual parameters which will be taken into account in the mapping process. Then we will first present the results with the linear approach and then the improvement obtained with the use of multiple GMMs.

2. Experimental set-up and spectral, lip and Cued Speech material

2.1. Database recording

The data have been derived from a video recording of a speaker pronouncing and coding in CS a set of 50 isolated French words. The words were made of 32 digits (from 0 to 31), 12 months and 6 more ordinary words. Each word was presented once on a monitor placed in front of the speaker, in a random order. The corpus has been uttered 10 times. The speaker is a female native speaker of French graduated in CS. The recording has been made in a sound-proof booth and the image video recording rate was set on 25 image/second. The speaker was seated in front of a microphone and a camera connected to a Betacam recorder. Landmarks were placed between eyebrows and at the extremity of the fingers to further extraction of the coordinates used as Cued Speech hand parameters. In addition, a square paper was recorded for pixel-to-centimeter conversion.

The video recording has been done with the PAL format, thus saved as numerical Bitmap RGB images made of the interlaced half-frames of the video (respectively even and odd lines). Each image was de-interlaced into two half-frames and the missing lines of the each half-frame were filled by linear interpolation, as to obtain two de-interlaced full frames corresponding to two recordings separated with 20 ms.

2.2. Extraction of lip and hand visual features

These frames constitute the set of images at the rate of 50 Hz that we will refer to in the following. For its part, the audio of the recording was digitalized at 44100 Hz and re-sampled at 16000 Hz. For each word, the coordinates of the inner contour of the lips have been manually selected on the corresponding images and converted into centimeters with the use of the pixel-to-centimeter conversion equation. Finally, the following geometric lip features were derived (following [18]): the lip width (A), the lip aperture (B) and the lip area (S) respectively. The work presented in this paper focuses on vowels. The database has thus been made by the vowels extracted from this set of material. The audio signal was used as to first locate the vowels inside the isolated words, then to derive the corresponding video frames. Thereafter the t_0 instants in which the lips were at the corresponding target were precisely defined from the analysis of the subset of video frames. 16 LSP coefficients were derived from the audio on the basis of a 20 ms Hamming window centered on t_0 together with 16 MFCC coefficients calculated on the basis of a 32 ms Hamming window. In addition 4 formants were derived from the spectral envelop (obtained with the LSP coefficients). Since the Cued Speech hand position target are often not synchronous with variation of lips or speech, the t_1 instants for Cued Speech hand target are selected separately by analysis of the subset of video frames. Then the hand features defined as the relative (x,y) coordinates of the fingertip of the middle finger (or index finger if middle finger is missing) in reference to the landmark between eyebrows were extracted. The whole of these processing thus made it possible to constitute a database made of 1371occurrences of the 10 French vowels (table 1). In the next, the accuracy of the mapping methods will be measured using a 1/5 cross-validation test. For that, the 1371 occurrences were divided into 5 partitions made of approximate 275 elements for each partition. Finally, 4 principal components (derived from a PCA Analysis) of the 4 formants, 16 principal components of the LSPs, 16 principal components of the MFCCs, the totally 32 principal components of the set of the LSP and MFCC coefficients, the (x, y) coordinates of the hand and the (A, B, S) lip parameters were derived for each element of each of the 5 partitions.

Table 1. *List of the ten French vowels with their occurrence*

Vowels	[i]	[e]	[ɛ]	[a]	[y]	[ø]	[œ]	[ɔ]	[o]	[u]
Occurrence	236	255	231	168	37	80	137	83	40	104

3. The Multiple-linear regression based mapping modeling

In this section, the objective is to predict the (A, B, S) lip parameters and the (x,y) hand coordinates with the corresponding spectral parameters (their principal components F_i) using the multiple-linear modeling.

3.1. The used method

The set of F_i was ordered in function of their "prediction power" using their ρ correlation coefficient with the parameter to be predicted. The F_i predictors were then sorted following the decreasing values of their ρ^2 as to obtain the sorted $F = [F_1, F_2, ..., F_p]$, $(1 \leq p \leq 32)$. In the following, the method is illustrated with the lip parameter B defined as the target, after being centered. B was submitted to a linear regression with the first predictor F_1. The linear coefficient k_1 was obtained as to minimize the residual error between the real values of the target and the predicted ones in the sense of least square error. The residual error was then submitted to a linear regression with the second predictor F_2 and so on until the p order. Finally, the estimation equation at the order p is the following:

$$\widehat{B} = k_1 F_1 + k_2 F_2 + ... + k_p F_p + \overline{B} \qquad (1)$$

$$\boldsymbol{k} = (F^T F)^{-1} F^T (B - \overline{B}) \quad (2)$$

Where, $\boldsymbol{k} = [k_1, k_2, ..., k_p]^T$. As mentioned before, the 5 partitions of the database was used to evaluate the accuracy of the mapping. One of the partitions was reserved for testing by turns, while the other 4 partitions were used for the training by applying the estimation equation (1). The residual variance as the complement of the explained variance of the considered lip parameter was calculated. Finally, the average residual variance was calculated over the 5 combinations of the training and testing partitions for evaluating the model.

3.2. Results for the lip parameters

In the following, the average residual variance calculated over the 5 combinations of the training partitions is considered. Figure 2 plots the average residual variance of lip parameter B in function of the number of predictors. From the figure, it can be first observed that the residual variance decreases in function of the number of used predictors. One can then notice that the residual variance remains high with the use of formants (around 39 % of the initial variance). This is probably due to a lack of dimensions. Indeed the 16 MFCC and LSP coefficients improve very significantly the performances of the prediction (the residual variances are 25% and 18% respectively). The MFCCs allow a quicker decrease while the LSP coefficients attain a lower residual variance. Finally, the prediction based on the mixture of the MFCC and LSP has the advantage of the quick decrease property of the MFCCs and the low residual of the LSP. This mixture of MFCCs and LSP is thus considered as the best parameters for this prediction even if the final error is still relatively high (around 14 % on the training database). The prediction performance of the other two lip parameters A and S are similar to situation of the B. These results will be used as a reference for the following, in particular for the choice of the set of pertinent predictors. Finally, we obtained very similar results with the test data.

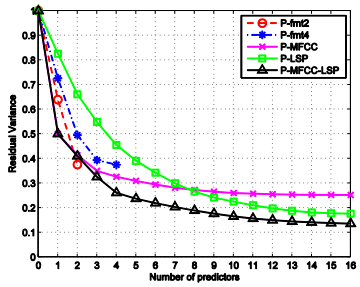


Figure 2: *The average residual variance of the lip parameters B over the 5 combinations of the training partitions, in function of the number of predictors (based on 2 formants (P-fmt2), 4 formants (P-fmt4), MFCC, LSP and the mixture of MFCC-LSP).*

3.3. Results for the hand parameters

The same method of analysis was applied for predicting the (x,y) coordinates of the hand. The final value of the residual variance reaches 39 % for x and 29 % for y, even in the case of the best predictors made up of the whole of the LSP parameters and MFCCs (see Figure 3). This high value of the final residual variance is explained by the low values of the correlations coefficient between the predictors and the target (0.43 and 0.42 respectively with x and y). This weak linear correlation between the spectral parameters and the (x, y) coordinates of the hand positions probably gives rise to the limit of the linear method.

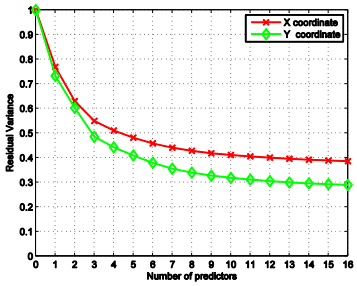


Figure 3*: The average residual variance of the hand coordinates on the 5 combinations of the training partitions with the first best predictors of the whole set of MFCC-LSP in function of the number of predictors.*

In order to check this assumption, the Cued Speech (x, y) coordinates have been re-organized in a coherent way with the French vocalic triangle defined by the formant space consisted of the first two formants, given the strong linear correlation between formants and the spectral parameters LSP (The maximum linear correlation coefficients between the spectral parameters LSP and the first two formants are 0.96 and 0.87 respectively). Therefore a great fall of the residual variances could be obtained which finally reach 7.85% and 7.08% respectively for the redistributed x and y coordinates.

4. The GMM based mapping model

4.1. The Method

In this section, the principal components of the set of 16 LSP and 16 MFCC coefficients constitute the source vector $\mathbf{x}$ with dimension p ($1 \leq p \leq 32$) and the lip or hand parameters are the target vector $\mathbf{y}$. In reference to the equation (3) (see also [19], [20]), the estimator (in the sense of MMSE) of the parameter $\mathbf{y}$ has a linear regression form of observation $\mathbf{x}$ weighted by the a posteriori conditional probability of component c_i:

$$\widehat{\mathbf{y}} = F(\mathbf{x}) = \sum_{i=1}^{m} (W_i \mathbf{x} + b_i) P(c_i|\mathbf{x}) \quad (3)$$

Where, $P(c_i|\mathbf{x})$ is the a posteriori conditional probability that observation $\mathbf{x}$ is generated by the c_i component/Gaussian with the mean vector $\boldsymbol{\mu}_i^{\mathrm{x}}$ and covariance matrix $\boldsymbol{\Sigma}_i^{\mathrm{XX}}$, W_i and b_i being the transform and the bias matrices respectively associated to component c_i.

$$b_i = \boldsymbol{\mu}_i^{\mathrm{Y}} - \boldsymbol{\Sigma}_i^{\mathrm{YX}} (\boldsymbol{\Sigma}_i^{\mathrm{XX}})^{-1} \boldsymbol{\mu}_i^{\mathrm{X}} \quad (4)$$

$$W_i = \boldsymbol{\Sigma}_i^{\mathrm{YX}} (\boldsymbol{\Sigma}_i^{\mathrm{XX}})^{-1} \quad (5)$$

$$P(c_i|\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i^{\mathrm{X}}, \boldsymbol{\Sigma}_i^{\mathrm{XX}})}{\sum_{j=1}^{m} \alpha_j N(\mathbf{x}; \boldsymbol{\mu}_j^{\mathbf{x}}, \boldsymbol{\Sigma}_j^{\mathrm{XX}})} \quad (6)$$

Where, α_i is the weighting coefficient of the Gaussian model, the sum of all the coefficients is 1; $\boldsymbol{\Sigma}_i^{\mathrm{YX}}$ is the matrix of covariance between $\mathbf{x}$ and $\mathbf{y}$ and calculated on the component

c_i ($1 \leq i \leq m$) subset of data and $\boldsymbol{\mu}_i^{\mathrm{Y}}$ is the mean vector of the target vector on this same subset. Note that when the number of the Gaussian equals to one, namely m=1, the GMM based mapping model in the sense of the MMSE regression criteria corresponds exactly to the multiple-linear regression model presented in the previous section. Finally, the spectral principal components are sorted following the decreasing order of their explanation variance of the estimated parameter. Then the first p principal components of the spectral parameters as source vectors are according to this order. The parameters of GMM such as mean vectors $\boldsymbol{\mu}_i^{\mathrm{X}}$, $\boldsymbol{\mu}_i^{\mathrm{Y}}$ and the covariance matrices $\boldsymbol{\Sigma}_i^{\mathrm{XX}}$ and $\boldsymbol{\Sigma}_i^{\mathrm{YX}}$ are determined during the GMM training processing. Two different training methods for GMM are discussed in our work. EM is the most used unsupervised training methods for GMM training, which trains the model without any label information of the elements and the data cluster automatically given the initialization parameters by the iterative procedure until the condition of convergence is satisfied [21]. In contrast with the unsupervised training method, a supervised training method is introduced to train the GMM based on the a priori information: the visemes of vowels (see Table 2) which is a speech presentation in the visual domain for the lips or the different Cued Speech five hand positions (see Figure 1) for hand respectively. Thus 3 Gaussian components of GMM corresponding to the three vowel visemes are trained for lips and 5 Gaussian components corresponding to the five hand positions defined in CS are trained for hand.

Table 2. Visemes of French vowels

Visemes	Phonemes of vowels
V1	[a],[i],[e],[ɛ]
V2	[y],[o],[u],[ø]
V3	[ɔ],[œ]

We use the joint source and target vectors defined in equation (7) rather than the source vectors only to train the GMM both in the EM and supervised training methods, which is more robust for small amounts spcifically since the joint density should lead to a more judicious clustering for the regression problem ([20]).

$$\mathbf{z} = [\mathbf{x}, \mathbf{y}] \in R^N \tag{7}$$

Where, N is the dimension of the joint vector $\mathbf{z}$. Once the GMM is trained, the parameters of GMM are fixed.

4.2. Results

The residual variances of the estimated lip and hand parameters obtained by the supervised training GMM decrease significantly compared to the multiple-linear model (see Figure 4). The residual variances decreased with the increment of the dimension of the source vector, i.e. the number of predictors. Finally, the residual variance reaches to around 7% for lip parameters (A, B and S) and 3% for hand coordinates (x, y) respectively by using 16-dimensions source vector.

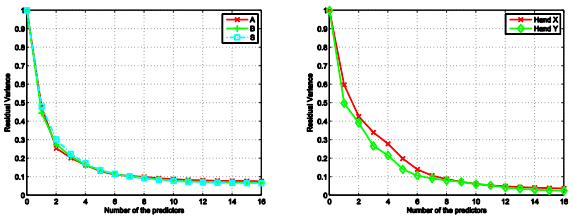


Figure 4: *The average residual variance of the lip parameters and the hand coordinates on the training data in function of the dimension of the source vector. On the left column, the number of the Gaussians m=3 for the lips parameters and on the right column, m=5 for the hand parameters.*

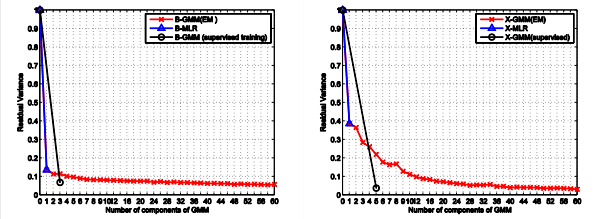


Figure 5: *The average residual variance of the lip parameters B (on left) and the hand coordinates x (on right) on training data in function of the number of Gaussians in GMM. The red line with crosses corresponds to the EM training GMM, the black line with squares corresponds to the supervised training GMM, the blue line with triangles corresponds to the multiple-linear regression model.*

Figure 5 compares the residual variance obtained by the multiple-linear regression model, the supervised training GMM mapping model and the EM trained GMM mapping model. The results show that the residual variance tends asymptotically to a limit value when the number of the Gaussians increases for both lip and hand parameters estimation in the case of EM trained GMM. And the residual variances obtained by the multiple-linear model (denoted by the blue triangle) are completely equal to the ones obtained by the uni-Gaussian GMM model. The supervised training GMM mapping model shows the competitive performance compared to the EM trained GMM model in terms of number of Gaussians. That is to say the supervised training GMM mapping model is more efficient than the one trained by EM method. In addition, the supervised training GMM model also shows good robustness in the evaluation procedure (i.e. test procedure) where it also retains the superior performance (see Table 3).

Table 3. Evaluation results of the lip and hand parameter mapping in terms of the 3 different models.

(%)	MLR	GMM (supervised)	GMM (EM)*
A	17%	9%	7% (38 comp.)
B	12%	9%	9% (40 comp.)
S	14%	8%	8% (40 comp.)

(%)	MLR	GMM (supervised)	GMM (EM) *
X	43%	8%	8% (60 comp.)
Y	31%	4%	5% (54 comp.)

* GMM (EM) shows the minimum residual variance and the number of the Gaussians with which the best results were obtained.

5. Discussion

These successive maps processing of the acoustic spectrum to the lips parameters and the hand position showed that it is much more difficult to estimate the hand position than lips parameters. Several different approaches, from the direct multi-linear method to the sophisticate GMM-based regression method, have been employed to this problem. Actually the source of the difficulty is that there is no relation between the hand position and the spectral parameters unlike the case of the lips as a vocal articulator with corresponding acoustic consequences. More specifically, there are two key points of the meaning of "no relation": (1) there is no structural topological relation between the acoustic space and the hand position space. That is to say, the two closed vowels in the acoustic space may be very far in the hand position space, such as the vowel [e] and [i]. On the contrary, two far vowels in the acoustic space may be corresponding to the same hand position, such as vowel [a] and [o]. It indicates that the two spaces have totally different topology structure since the hand position is determined by the rules of CS but not the acoustic parameters. This is the real reason why a large residual variance was obtained with the multi-linear approach; (2) there is no relation of the variance within group between the acoustic space and the hand position space. That is to say, the tiny variation of the sound will not consequently change the hand position of the speaker. Indeed the hand position around the center within group is random from person to person. Thus it is impossible to establish a global linear relation by the multi-linear model or even a local linear relation by the GMM to predict the movement of the hand around the center within group from the acoustic spectrum.

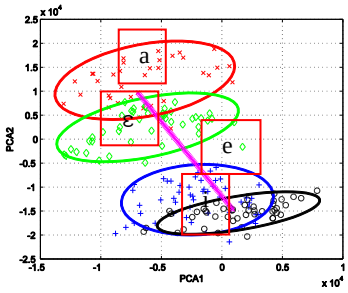


Figure 6. *The linear interpolation in the acoustic space between vowels [a] and [i].*

In order to have a further understanding and comparison of the different mapping approaches, a continuous transition has been achieved by a linear interpolation in the acoustic spectral parameter (i.e. MFCC+LSP) between the vowels [a] and [i]. The 16-dimension acoustic spectral parameters were projected onto their first two PCA components to verify the continuous linear transition in the acoustic space (see Figure 6). In fact there are many ways to go to vowel [i] from vowel [a] in the acoustic space in function of the choice on different starts and ends, but here only one of them is presented as an example to show the corresponding transition obtained by the different mapping models. The corresponding transitions of the estimated lip parameters and hand positions are shown in Figure 7 and Figure 8. For the multi-linear method, the figures present a reasonable linear relation between the linear interpolation spectral parameters and the estimated hand position or lip parameter. For the GMM based mapping method (in the MMSE regression criterion, with 5 components corresponding to the five hand position in CS for hand position estimation and 10 components corresponding to the ten vowels for lip parameter estimation), the four stable phases during the transition both for the hand position and lips are corresponding to the passing vowels ([a],[ɛ],[e],[i]) during the spectral parameters changing linearly from vowel [a] to [i] in the acoustic space. With the four stable phases, the GMM-based mapping method shows classification-like property which helps the model to decrease significantly the residual variance in comparing with the multi-linear model. However, unlike the GMM-based classification method which cannot project the variance of the source data at all, the GMM-based mapping method can still reflect the linear relation locally in the region of the phase as shown in the figures. Due to the strong linear correlation between the acoustic spectrum and the lips parameter, the transition of lips shown in Figure 7 is different as the case of the hand. The order of the phases change in coherence with the lip parameter B and the multi-linear model performs well passing close to the centers of phases corresponding to the different vowels. The local linear regression of GMM-based mapping method is effective and varies in the right direction, however in the case of the hand position estimation the local regression is weak and even in the wrong direction (such as the local regression on the phase of vowel [a] in Figure 6) due to there is "no relation" between the hand position and the acoustic spectral parameters. With the effective local regression, the GMM-based mapping method can improve the estimation performance in comparison with the multi-linear model.

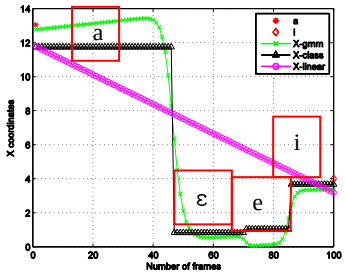


Figure 7: *The dynamic transition of X coordinates of hand position by interpolation between the vowel [a] and [i]. Results with the multi-linear mapping (X-linear), the GMM mapping (X-gmm), and the gaussian classifier applied to the spectral space (X-class).*

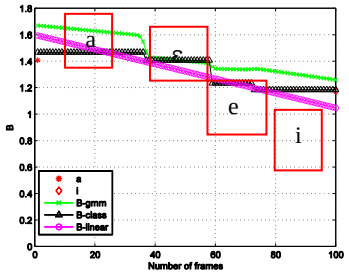


Figure 8: *The transition of lip parameter B by interpolation between the vowel [a] and [i]. Results with the multi-linear mapping (B-linear), the GMM mapping (B-gmm), and the gaussian classifier applied to the spectral space (B-class).*

In the case of the hand positions estimation, the residual variance of the GMM-based mapping approach decreases

significantly comparing to the linear approach by the class-like property but results in the phase changing rapidly. In the case of lip parameters estimation, the local regression of the GMM-based mapping approach is more effective and the phase changing is more gradual meanwhile the multi-linear approach also performs well due to the strong linear correlation between the acoustic spectral parameters and lips parameters.

With both the properties of the classifier and the regression estimator, the GMM-based mapping method decreases the residual variance of the estimated value significantly comparing with the multi-linear model. These properties are well presented in the case of the lip parameters estimation in which the local regression and classification methods perform well. However, when the relation is weak or even no relation between the source and target data, the local regression will degenerate or even meaningless such as the case of the hand position estimation. At this time, the GMM-based mapping method more tends to the classification method, thus the residual variance may be no longer appropriate for evaluating the model since the errors probably cannot be understood at all even if the model has a small residual variance. The cognitive effect of the human being in a perception task may be an alternative evaluation criterion. But note that in the GMM-based mapping method, there is one point essentially different with the classification method that is the contributions of all the components are always considered and weighted to produce the final results. From this aspect, GMM-based mapping method will effect better than the binary classification method in the mapping problem.

6. Conclusion

This paper discusses the relations between the speech spectral space and the visual space of speech and Cued Speech. This program started with the case of oral French vowels. The multiple-linear regression model as a simple case of the GMM modeling has been first used to convert the spectral parameters towards the lip parameters as well as the hand parameters of the Cued Speech. The results show that the best predictors are 16 principal components derived from the 16 LSP and 16 MFCC coefficients. The linear approach showed its limit in the case of the manual setting hand component of the Cued Speech. Two types of GMM based model have been introduced to solve the mapping problem. Since the GMM based model explore the regression relationship between the source and target vectors based on the Gaussians locally and precisely rather than the rough regression based on the global set as in the multiple-linear model, the results obtained by the GMM based model were improved significantly with an explanation of 93% for lip and 96% for hand components of the original variance for the best results. In addition, the supervised training GMM shows a high efficiency and good robustness benefiting from the a priori phonetic information in comparison with the EM training GMM which may be affected more by the outliers due to the important dependence on the data itself. For the future, these results have to be evaluated in perception with deaf persons using supplemented visual speech synthesizers.

7. Acknowledgements

The authors wish to thank the speaker to have accepted the constraints of recording and Thomas Hueber for the fruitful discussions on the method. This work is supported by the National Agency of French Research through the TELMA and PLASMODY projects.

8. References

[1] Montgomery, A. A., Jackson, P. L., 1983. Physical characteristics of the lips underlying vowel lipreading performance. Journal of the Acoustical Society of America 73(6),

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

[3] Auer, E.T. Bernstein, L.E., 2007. Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. Journal of Speech, Language, and Hearing Research, Vol.50, pp. 1157-1165.

[4] Bernstein, L.E., Auer, E.T., Jr, & Jiang, J. (2010). "Lipreading, the lexicon, and Cued Speech", in C. la Sasso, J. Leybaert, K. Crain (Eds.), Cued Speech for the Natural Acquisition of English, Reading, and Academic Achievement. Oxford University Press.

[5] Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Deguine, O. & Barone, P. 2009. Improvement in speech-reading ability by auditory training: Evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. Neuropsychologia, 47, 972-979.

[6] Cornett, R. O. (1967). "Cued Speech," American Annals of the Deaf ,112, 3-13, 1967.

[7] Attina, V., Beautemps, D., Cathiard, M. A. & Odisio, M. (2004). "A pilot study of temporal organization in Cued Speech production of French syllables: rules for Cued Speech synthesizer," Speech Communication, vol. 44, pp. 197-214.

[8] Uchanski, R.M., Delhorne, L.A., Dix, A.K., Braida, L.D., Reed, C.M., Durlach, N.I., 1994. 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.

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

[10] Tokuda, K., T. Yoshimura, et al. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey.

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

[12] Toda, T., A. W. Black, et al. (2004). Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. International Speech Synthesis Workshop, Pittsburgh, PA.

[13] Uto, Y., Y. Nankaku, et al. (2006). Voice conversion based on mixtures of factor analyzers. InterSpeech, Pittsburgh, PE.

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

[15] Gibert, G., Bailly, G., Beautemps, D., Elisei, F., & Brun, R. (2005). "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using Cued Speech," J. Acoust. Soc. Am., vol. 118(2), pp. 1144–1153.

[16] 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: Telephony for the Hearing-Impaired People, From Models to User Tests," In Proceedings of. ASSISTH 2007, pp. 201–208

[17] Berthommier F. (2003). Audiovisual Speech Enhancement Based on the Association between Speech Enveloppe and Video Features. In Proceedings of Eurospeech'2003.

[18] Lallouache, M.T. (1991). "Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres," Ph.D. Thesis, Institut National Polytechnique de Grenoble, 1991.

[19] Kain, A. (2001). High-resolution voice transformation (PhD, OGI School of Science & Engineering, Oregon Health & Science University)..

[20] Hueber, T., Benaroya, E.L, Denby, B., Chollet, G. (2011). "Statistical Mapping between Articulatory and Acoustic Data for an Ultrasound-based Silent Speech Interface", Proceedings of Interspeech, pp. 593-596, Firenze, Italia.

[21] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. New York: Springer.

Lucie Scarbel, Denis Beautemps, Jean-Luc Schwartz, S. Schmerber, Marc Sato. (2014). L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles. Journées d'Etudes sur la Parole (JEP 2014), June 2014, Le Mans, France.

L'ombre d'un doute ? Interactions perceptivo-motrices lors de taches de close-shadowing auditive et audio-visuelles

Lucie Scarbel, Denis Beautemps, Jean-Luc Schwartz, S. Schmerber, Marc Sato

L'ombre d'un doute?
Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles

Lucie Scarbel[1], Denis Beautemps [1], Jean-Luc Schwartz [1],
Sébastien Schmerber[2], Marc Sato [1]
(1) GIPSA-LAB, Département Parole & Cognition ; CNRS UMR 5216 - Université de Grenoble-Alpes ;
(2) Service ORL du CHU Grenoble
lucie.scarbel@gipsa-lab.grenoble-inp.fr

RESUME

Un argument classique en faveur des théories motrices de la perception de la parole provient du paradigme de « close-shadowing » (répétition rapide). Le fait que cette tâche de close-shadowing entraîne des réponses orales bien plus rapides qu'en réponses manuelles suggère en effet un codage des représentations perceptives dans un format moteur, compatible avec une réponse orale. Un autre argument est apporté par les interactions audio-visuelles lors de la perception de parole, souvent interprétées en référence à un couplage fonctionnel entre audition, vision et motricité. Dans cette étude, nous avons combiné ces deux paradigmes de manière à tester si la modalité visuelle pouvait induire des réponses motrices plus rapides lors d'une tâche de close-shadowing. Pour ce faire, différentes tâches de catégorisation orale et manuelle de stimuli de parole présentés auditivement ou audio-visuellement, en présence ou non d'un bruit blanc, ont été réalisées. De manière générale, les réponses orales ont été plus rapides que les réponses manuelles, mais aussi moins précises, notamment dans le bruit, ce qui suggère que la représentation motrice induite par la stimulation pourrait être peu précise dans un premier niveau de traitement. En présence d'un bruit acoustique, la modalité audiovisuelle s'est avérée à la fois plus rapide et plus précise que la modalité auditive. Aucune interaction entre le mode de réponse et la modalité de présentation des stimuli n'a cependant été observée. Nous interprétons l'ensemble de ces résultats dans un cadre théorique proposant l'existence de boucles perceptivo-motrices, dans lesquelles les entrées auditives et visuelles seraient intégrées et reliées à la génération interne de représentations motrices préalablement au processus final de décision.

ABSTRACT

The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close shadowing

One classical argument in favor of a functional role of the motor system in speech perception comes from the close shadowing task in which a subject has to identify and to repeat as quickly as possible an auditory speech stimulus. The fact that close shadowing can occur very rapidly and much faster than manual identification of the speech target is taken to suggest that perceptually-induced speech representations are already shaped in a motor-compatible format. Another argument is provided by audiovisual interactions often interpreted as referring to a multisensory-motor framework. In this study, we attempted to

combine these two paradigms by testing whether the visual modality could speed motor response in a close-shadowing task. To this aim, both oral and manual responses were evaluated during the perception of auditory and audio-visual speech stimuli, clear or embedded in white noise. Overall, oral responses were much faster than manual ones, but it also appeared that they were less accurate in noise, which suggests that motor representations evoked by the speech input could be rough at a first processing stage. In the presence of acoustic noise, the audiovisual modality led to both faster and more accurate responses than the auditory modality. No interaction was however observed between modality and response. Altogether, these results are interpreted within a two-stage sensory-motor framework, in which the auditory and visual streams are integrated together and with internally generated motor representations before a final decision may be available.

MOTS-CLES : perception de la parole, production de la parole, perception de la parole audio-visuelle, close-shadowing, interaction sensorimotrice

KEYWORDS: speech perception, speech production, audio visual speech perception, close-shadowing, sensory motor interaction

1 Introduction

Un débat classique dans le domaine de la perception de la parole concerne l'implication du système moteur et du lien fonctionnel entre représentations auditives et motrices. Les théories auditives de la perception de parole réfutent l'implication du système moteur et proposent l'existence de processus auditifs de décodage phonétique à partir du signal acoustique (Diehl et al., 2004). A contrario, Liberman et al. (1985) et Fowler (1986) supposent que pour percevoir la parole, nous utilisons des représentations procédurales motrices basées sur notre expérience de locuteur. Enfin, les théories perceptivo-motrices postulent que les représentations motrices soient utilisées en lien avec les représentations auditives dans le traitement et décodage des informations phonétiques (Skipper et al., 2007, Schwartz et al. 2012).

Dans une récente revue sur les théories motrices, Galantucci et al. (2006) rappellent les principaux arguments expérimentaux et en mentionnent notamment deux qui fournissent le cœur du présent travail. Un premier argument en faveur d'un couplage perceptivo-moteur en parole provient du paradigme de close-shadowing. Ce paradigme (Porter et al., 1980 ; Fowler et al., 2003) consiste à répéter le plus rapidement possible des stimuli de parole. L'analyse des réponses orales permet de mesurer la rapidité et le taux de répétitions correctes. Galantucci (2006) compare ces résultats avec les temps de réaction obtenus par Luce (1986) dans une tâche de réponse manuelle dans laquelle les sujets utilisaient une touche. Les résultats montrent une augmentation des temps de réaction dans le cas de la réponse manuelle. Cette différence ne pouvant être expliquée par la difficulté du choix des touches dans la tâche de décision manuelle, Galantucci et al. l'interprètent par le fait que si percevoir la parole c'est percevoir des gestes, alors la perception des gestes préparerait la réponse orale et la rendrait ainsi plus rapide.

Le second paradigme est celui de la multisensorialité. De nombreuses études ont montré que l'entrée visuelle améliore la compréhension de parole, tant pour les sujets malentendants que normo-entendants. Sumby & Pollack (1954) ont été parmi les premiers à démontrer l'apport de la modalité visuelle pour percevoir et comprendre la parole dans des conditions

bruitées. L'apport de la modalité visuelle a été également démontré dans des études de répétition (shadowing) où la tâche consistait à répéter oralement les stimuli, et à mettre ainsi en évidence une meilleure compréhension par des sujets de matériaux linguistiques complexes ou produits dans une langue étrangère ou avec accent étranger (Reisberg, 1987 ; Davis and Kim, 2001). Cependant, ces tâches de répétition se sont faites sans pression de temps (shadowing et non close-shadowing). Or la tâche de répétition rapide fournit une fenêtre riche sur la dynamique temporelle du processus de décision. A l'inverse les expériences de close-shadowing n'ont jamais incorporé la modalité visuelle, se privant d'une connaissance sur le rôle des interactions audiovisuelles en lien avec les relations perceptuo-motrices. La présente étude se propose précisément d'étudier, pour la première fois, quel est l'apport de la modalité visuelle dans une tâche de close-shadowing.

Cette étude est composée de deux expériences, toutes deux focalisées sur une évaluation conjointe de la précision et de la rapidité de réponses orales ou manuelles à des stimuli auditifs ou audio-visuels. Les deux expériences ont été réalisées sur des stimuli de parole non-lexicaux (logatomes), présentés sans bruit dans la première expérience (Expérience A) ou avec bruit acoustique dans la seconde expérience (Expérience B). Les hypothèses sous-jacentes sont que (1) les réponses orales devraient être plus rapides que les réponses manuelles, en accord avec les études précédentes sur le close-shadowing, et que (2) les réponses aux stimuli audio-visuels devraient être plus rapides et plus précises que celles aux stimuli auditifs, au moins dans l'Expérience B impliquant des stimuli bruités. Une question supplémentaire concerne la possibilité d'une interaction entre ces deux effets, qui permettrait d'évaluer si l'effet de la vision est différent entre une modalité de réponse (orale) et l'autre (manuelle).

2 Méthodologie

2.1 Participants

Deux groupes de quinze et quatorze adultes sains, de langue maternelle française, ont participé aux Expériences A et B (Expérience A: 10 femmes; moyenne d'âge: 29 ans, entre 20 et 38 ans - Expérience B: 11 femmes; moyenne d'âge: 24 ans, entre 19-34 ans). Tous les participants ont une vision normale ou corrigée à la normale et ont rapporté n'avoir jamais eu des troubles moteurs, de la parole ou de l'audition.

2.2 Procédure expérimentale

Chaque expérience consistait en deux tâches de catégorisation : une tâche de close-shadowing où les réponses étaient données oralement, en répétant le plus vite possible la séquence présentée, et une tâche de décision manuelle, où les réponses étaient données manuellement, en appuyant le plus vite possible sur la touche appropriée. Les stimuli à catégoriser correspondaient aux séquences /apa/, /ata/ et /aka/ (voir ci-dessous). Les participants étaient informés qu'on allait leur présenter des séquences /apa/, /ata/ ou /aka/, soit de manière auditive soit de manière audio-visuelle. Dans la tâche de close-shadowing, on leur demandait de catégoriser et répéter chaque séquence le plus vite possible. Pour ce faire, ils devaient produire la voyelle initiale /a/ puis répéter immédiatement la syllabe CV perçue (/pa/, /ta/ ou /ka/). Lors de la tâche de décision manuelle, les participants devaient catégoriser chaque énoncé en appuyant le plus vite

possible avec leur main dominante sur une des trois touches correspondant respectivement à /apa/, /ata/ ou /aka/. L'ordre des touches était contrebalancé entre les participants. Pour chaque tâche (avec réponses manuelles ou orales) et chaque modalité (auditive ou audio-visuelle), 16 répétitions de chacune des séquences /apa/, /ata/ et /aka/ étaient présentées de manière randomisée. Les ordres de présentation des tâches et des modalités étaient contrebalancés entre les participants. Les deux expériences ont été réalisées dans une chambre sourde. Les participants étaient assis en face d'un ordinateur à une distance d'approximativement 50 cm. Les stimuli acoustiques étaient présentés à un niveau sonore confortable, celui-ci étant le même pour tous les participants. Le logiciel Presentation (Neurobehavioral Systems, Albany, CA) a été utilisé pour contrôler la présentation des stimuli et pour enregistrer les réponses manuelles. Toutes les productions des participants ont été enregistrées grâce à un microphone AKG 1000S pour les analyses offline, avec un système assurant la synchronisation entre les stimuli présentés et la réponse des participants. Une courte session d'entrainement précédait chaque tâche. La durée totale de chaque expérience était d'environ 30 minutes.

2.3 Stimuli

Les séquences /apa/ /ata/ et /aka/, produites dans une chambre sourde par un homme de langue maternelle française ont été enregistrées audio-visuellement au moyen d'un microphone AKG 1000S (44.1 kHz) et d'une caméra haute qualité au format PAL placée en face du locuteur (images détramées de 572 par 520 pixels, 50 Hz). Le corpus a été enregistré avec pour objectif d'obtenir pour chaque séquence 4 productions distinctes impliquant quatre durées de la voyelle initiale /a/ (0.5s, 1s, 1.5s et 2 s) dans le but de réduire toute prédiction temporelle de la séquence à catégoriser). Les durées des 12 stimuli ainsi sélectionnés ont été égalisées (3 séquences x 4 productions). Les stimuli étaient présentés sans bruit additionnel dans l'Expérience A, un bruit blanc (filtré à -6 dB/oct) a été ajouté à chaque stimulus dans l'Expérience B (rapport signal sur bruit de -3 dB).

2.4 Analyses acoustiques

Afin de calculer les temps de réaction (RTs) et la proportion de réponses correctes dans la tâche à réponses orales, des analyses acoustiques de la production des participants ont été réalisées en utilisant le logiciel Praat (Boersma et Weenink, 2013). Les temps de réaction ont été calculés uniquement pour les réponses correctes : les omissions ou tout autre type d'erreurs (c'est-à-dire le remplacement d'une consonne par une autre ou la production de deux consonnes ou de deux syllabes pour un même stimulus dans la tâche de close-shadowing) ont été exclues. Les temps de réaction pour les réponses orales ont été mesurés entre le début du burst consonantique du stimulus et de la réponse.

2.5 Analyse des données

Dans chaque expérience, la proportion de réponses correctes et la médiane des temps de réaction étaient déterminées individuellement pour chaque participant, chaque tâche et chaque modalité, séparément pour /apa/, /ata/ et /aka/. Deux ANOVA à mesures répétées ont été réalisées sur ces données, avec le groupe (Expérience A avec stimuli non-bruités vs. Expérience B avec stimuli bruités) comme variable inter-sujets et la tâche (close shadowing

vs. décision manuelle), la modalité (audio vs. audio-visuelle), et le type de stimulus (/apa/ vs /ata/ vs /aka/) comme variables intra-sujets.

3 Résultats

Pour toutes les analyses suivantes, le niveau de significativité était fixé à $p = .05$ et corrigé en cas de violation de l'hypothèse de sphéricité. Les analyses post-hoc ont été effectuées en utilisant des tests de Bonferroni.

3.1 Temps de réaction

Comme attendu, l'effet principal du groupe est significatif ($F(1,27)=24,38$; $p<0.001$), avec des temps de réaction pour les stimuli non bruités de l'expérience A plus courts que ceux des stimuli bruités de l'expérience B (351 ms vs 484 ms). Les effets principaux de la tâche ($F(1,27)=151,70$; $p<0.001$) et de la modalité ($F(1,27)=14,79$; $p<0.001$) sont aussi significatifs. Pour la tâche, les réponses orales étaient plus rapides que les réponses manuelles (286ms vs 545ms). Par rapport à la modalité, les temps de réaction étaient plus courts dans la modalité audio-visuelle par rapport à la modalité auditive (405 ms vs. 425 ms). Une interaction significative entre groupe et modalité ($F(1,27)=21,74$; $p<0.001$) montre que l'effet bénéfique de la présentation audio-visuelle est présent avec les stimuli bruités dans l'expérience B (461 ms vs. 507 ms) mais non avec les stimuli non-bruités de l'expérience A (354 ms vs. 349 ms). Par contre, l'interaction entre modalité et réponse n'est pas significative.

Ces effets semblent être dépendants des syllabes perçues. Notamment, une interaction à trois facteurs 'tâche x modalité x syllabe' a été trouvée ($F(2,54)=6,49$; $p<0.005$). Dans la modalité auditive, aucune différence significative des temps de réaction n'a été observée entre les syllabes à la fois pour les réponses orales et pour les réponses manuelles. Par contre, dans la modalité audio-visuelle, les temps de réaction pour les réponses orales étaient plus rapides pour la syllabe /pa/ par rapport aux syllabes /ta/ et /ka/, alors que les temps de réaction pour les réponses manuelles étaient plus rapides pour /pa/ par rapport à /ka/ et pour /ka/ par rapport à /ta/.

Ainsi, globalement, on obtient un patron de résultats attendus : des temps de réponse plus courts en réponse orale, un effet du bruit ralentissant les temps de réponse, une accélération de la réponse en modalité audiovisuelle par rapport à la modalité auditive en présence de bruit. Il n'apparaît cependant pas d'interaction entre modalité et type de réponse.

3.2 Proportion de réponses correctes

L'effet principal du groupe est significatif ($F(1,27)=266,28$; $p<0.001$) avec une proportion de réponses correctes plus élevée pour les stimuli non-bruités de l'expérience A (95%) par rapport aux stimuli bruités de l'expérience B (61%). D'autres effets principaux ont été significatifs, à la fois pour la tâche ($F1(1,27)=69,40$; $p<0.001$) et pour la modalité ($F(1,27)=52,39$; $p<0.001$). Concernant la tâche, une baisse importante des réponses correctes a été observée pour les réponses orales par rapport aux réponses manuelles (73% vs. 85%). Comme l'interaction significative du groupe par la tâche ($F(1,27)=38,67$; $p<0.001$) l'indique, cet effet n'apparait que pour les stimuli bruités de l'expérience B (71%

vs. 50%) alors qu'aucune différence n'a été observée entre les réponses orales et manuelles pour les stimuli non-bruités de l'expérience A (93% vs. 98%). Concernant la modalité, la modalité audio-visuelle apporte plus de réponses correctes que la modalité auditive (82% vs. 75%). Par contre, comme l'indique l'interaction significative 'groupe x modalité' ($F(1,27)=72,36$; $p<0.001$) aucune différence n'apparaît entre les deux modalités avec les stimuli non-bruités de l'expérience A (96% vs. 95%) alors qu'avec les stimuli bruités de l'expérience B, la modalité audio-visuelle apporte plus de réponses correctes (68%) que la modalité auditive (53%).

Là encore, les résultats dépendent de la syllabe présentée. Si les 3 syllabes sont parfaitement identifiées dans l'Expérience A en l'absence de bruit, quelle que soit la tâche et la modalité, en condition bruitée (Expérience B) la syllabe « pa » apparaît la plus saillante à la fois auditivement et visuellement. Si on obtient ici encore une confirmation d'un patron attendu (réponses plus précises en l'absence de bruit et, dans le cas de stimuli bruités, en présence de la modalité visuelle), un résultat fort et inattendu doit être relevé : le fait que la réponse orale dégrade la précision de la réponse en cas de stimuli bruités (Expérience B). Une nouvelle fois, il n'apparaît pas d'interaction entre modalité et type de réponse.

Tableau 1: Moyenne des temps de réaction (en ms.) et des % de réussite

modalité	A	A	A	A	A	A	AV	AV	AV	AV	AV	AV
mode	oral	oral	oral	manuel	manuel	manuel	oral	oral	oral	manuel	manuel	manuel
syllabe	ka	pa	ta	ka	pa	ta	ka	pa	ta	ka	pa	ta
RTs sans bruit	250	208	259	471	442	465	261	197	268	474	416	506
RTs avec bruit	348	335	373	614	666	632	277	296	318	531	653	601
% sans bruit	90%	99%	93%	98%	99%	98%	89%	100%	88%	98%	96%	97%
% avec bruit	56%	36%	40%	67%	42%	79%	89%	44%	38%	96%	58%	85%

4 Discussion

4.1 Effet de la tâche : mode de réponse oral ou manuel

Dans la condition non bruitée (Expérience A), les réponses orales sont plus rapide que les réponses manuelles (240ms vs. 462ms), mais aussi marginalement moins précises (93% vs. 98%, effet non significatif). Les temps de réaction sont en adéquation avec ceux obtenus par Fowler et al. (2003) et par Porter & Castellanos (1980). Ces auteurs interprètent la rapidité de la réponse dans le mode oral en référence avec les théories motrices. Le système orofacial serait ainsi favorisé pour répondre de manière très rapide, puisque le percept serait déjà en adéquation avec le format moteur ; le système manuel, nécessitant une étape transitoire entre la décision et l'action, serait ralenti d'autant. Cependant, les données de la condition bruitée (Expérience B) apportent des précisions importantes et inattendues sur ce raisonnement. En effet, alors que les temps de réaction restent plus courts dans la tâche de close-shadowing (334ms vs. 633ms), la précision dans la tâche de réponse orale diminue considérablement par rapport à la tâche de décision manuelle (50% vs. 71%).

Ces nouvelles données nécessitent de modifier l'interprétation des tenants des théories motrices jusqu'à un certain point. Nous allons proposer une tentative d'explication dans le

cadre du modèle proposé par Skipper et al. (2007) pour intégrer les interactions perceptuo-motrices dans la perception de parole. Ces auteurs proposent un modèle inspiré de « l'analyse par la synthèse » (Stevens & Halle 1967 ; Bever & Poeppel, 2010). Le modèle de Skipper et al. implique une boucle corticale entre les aires auditives et motrices. Après un stade initial de traitement auditif dans le cortex temporal (cortex auditif primaire, secondaire et aires associatives : stade 1), le cortex frontal générerait des hypothèses phonémiques associées avec les buts articulatoires puis des commandes motrices correspondant à cette prédiction initiale (Pars opercularis, cortex pré moteur ventral et cortex moteur primaire : stade 2), afin d'émettre des copies d'efférence qui seraient ensuite renvoyées dans le cortex auditif afin d'être comparées avec l'input auditif (stade 3). Ce modèle peut être utilisé comme une base pour tenter d'interpréter nos propres données. Pour ce faire, nous supposons que les réponses manuelles et orales sont générées à deux stades différents dans cette boucle de traitement. Les réponses orales seraient générées au stade 2, en accord avec les postulats de Porter & Castellanos ou de Fowler et al. Quand l'information provenant du cortex auditif aurait généré des commandes motrices dans le cortex moteur, le système orofacial, pré-activé depuis le début de l'expérience de close-shadowing pour permettre aux participants de répondre le plus vite possible, générerait une réponse orale produite par ces commandes motrices. Les réponses orales étant traitées à un stade précoce, elles seraient donc plus rapides mais sont aussi, selon nos résultats, moins précises, ce qui correspond au modèle de Skipper et al. qui considère qu'il ne s'agirait que d'une première hypothèse de réponse qui devraient être affinées à un stade ultérieur. Au stade 2, par contre, le système manuel ne reçoit pas de stimulations spécifiques permettant de générer une réponse. Par contre, au stade suivant (stade 3), le transfert de l'information auditive au cortex auditif, grâce à la copie d'efférence, fournit, en intégrant cette hypothèse motrice avec l'entrée acoustique, une information plus précise qui peut alors être transférée au système manuel pour réponse. Dans ce raisonnement, les réponses orales et manuelles seraient émises à deux instants différents du processus d'analyse par synthèse. Par conséquent, les temps de réaction pour les réponses manuelles seraient plus lents que ceux des réponses orales, mais les réponses seraient plus précises puisque, contrairement aux réponses orales, dans la tâche de décision manuelle les prédictions auraient été confirmées et ajustées avec le cortex auditif avant la décision finale envoyée aux commandes motrices manuelles pour appui de la touche adéquate. Bien sûr cette explication est probablement trop simple pour tenir compte de tous les aspects de nos données. Néanmoins, l'aspect crucial de nos résultats est qu'une hypothèse de pur processus d'identification motrice, certes compatible avec des temps de réaction plus courts en réponse orale, ne semble pas pouvoir rendre compte de la diminution de la précision de réponse orale pour des stimuli bruités. Ces résultats semblent donc réfuter une version stricte des théories motrices au profit de théories perceptuo-motrices de la perception de parole telles que celle de Skipper et al. (2007).

4.2 Effets de la modalité : auditive vs. audio-visuelle

Les effets de la modalité n'apparaissent dans notre étude que dans l'Expérience B avec les stimuli bruités. Dans la modalité auditive, les temps de réaction sont alors plus longs que dans la modalité audio-visuelle, et les proportions de réponses correctes sont plus faibles. Pris ensemble, ces résultats montrent un bénéfice clair de l'apport de la modalité visuelle à l'input auditif, ce qui est en accord avec toutes les études depuis Sumby et Pollack (1954).

Dans notre étude, l'avantage audio-visuel apparaît essentiellement pour la syllabe /pa/ ce qui est classique et en lien avec la haute visibilité des mouvements des lèvres associées à la bilabiale /p/, et le fort degré de confusion entre les mouvements visuels associés à /t/ ou /k/, généralement considérés comme appartenant à la même classe visémique. Ces effets de la modalité ne sont pas présents dans l'Expérience A avec des stimuli non-bruités, très probablement parce que les temps de réaction dans la modalité auditive sont déjà trop courts et les proportions de réponses correctes trop élevées pour être améliorés par l'entrée visuelle (effet plafond).

Un point intéressant est qu'il n'y a pas d'interaction significative entre la modalité et la tâche c'est-à-dire que la diminution des temps de réaction et l'amélioration des proportions de réponses correctes de la modalité auditive à la modalité audiovisuelle sont similaires dans les tâches orales ou manuelles. Nous allons là encore tenter d'interpréter cette absence d'interaction dans le cadre du modèle proposé par Skipper et collègues. Dans ce modèle, les informations auditives et visuelles, après prétraitement dans les aires auditives et visuelles, convergeraient dans une aire multi-sensorielle dans le cortex temporal postéro-supérieur (stade 1). Ensuite, dans le cas d'un input multi-sensoriel, les premières hypothèses seraient donc plutôt multi-sensorielles plutôt qu'uniquement auditives. A partir de là comme précédemment génération d'une hypothèse phonémique associée avec des buts articulatoires puis des commandes motrices orofaciales (stade 2), et émission d'une copie d'efférence fournissant une prédiction multisensorielle comparée avec l'input (stade 3). Dans notre étude, les interactions audio-visuelles du stade 1 affineraient les procédés sensori-moteurs et produiraient des hypothèses phonémiques plus rapides et plus précises au stade 2, qui est le stade où, dans notre interprétation, les réponses orales seraient générées. Ensuite le même gain en rapidité et en précision serait rétro-propagé vers les aires auditives pour génération des réponses manuelles (stade 3). Il n'y aurait ainsi pas de raison d'attendre des différences de gain visuel entre les tâches orales ou manuelles, le gain étant essentiellement déterminé dès le stade 1 dans le modèle.

5 Conclusion

En résumé, les résultats de la présente étude suggèrent que les réponses manuelles et orales sont générées à deux stades différents dans la chaine de perception de la parole. Dans la théorie du modèle « d'analyse par synthèse », les réponses manuelles serait fournies seulement à la fin de la boucle complète, incluant un processus feedworward de génération de prédictions phonémiques associées avec les buts articulatoires puis des commandes motrices émettant en feedback des hypothèses multi-sensorielles comparées au flux d'évènements multi-sensoriels. Contrairement aux réponses manuelles, les réponses orales seraient produites à un stade plus précoce (fin du processus feedforward) où les commandes motrices sont générées, produisant des réponses plus rapides mais moins précises. L'apport visuel améliorerait la rapidité et la précision pour les phonèmes suffisamment visibles (comme par exemple /p/) et ce lorsque le système auditif est en difficulté (conditions adverses, par exemple en présence de bruit). Bien évidemment, d'autres interprétations ou d'autres théories pourraient être confrontées aux présentes données expérimentales. Mais globalement, l'ensemble des résultats de cette étude semble nécessiter une théorie perceptuo-motrice de la perception de parole dans laquelle les flux auditifs et visuels sont intégrés à des représentations motrices auto-générées, avant d'aboutir à une décision finale.

Remerciements

Les auteurs tiennent à remercier l'ANR Plasmody qui a permis de financer cette étude, ainsi que les participants qui se sont portés volontaires pour les deux expériences.

Bibliographie

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics,* 4.2-.3, 174-200.

Davis, C. & Kim, J. (2001), 'Repeating and remembering foreign language words: Implications for language teaching system', *Artificial Intelligence Review*, vol 16, no 1 , pp 37 - 47.

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149-179.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, 49, 296–314.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3–28.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13, 361-377.

Liberman, A. M., & Mattingly, I. G. (1985). "The motor theory of speech perception revised". *Cognition* 21 (1): 1–36.

Luce RD. (1986). *Response times: Their role in inferring elementary mental organization.* Oxford University Press; New York.

Porter, R., & Castellanos, F. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America,* 67, 1349–1356.

Reisberg, D, McLean, J. & Goldfield, A. (1987). Easy to Hear to Understand: A Lip-Reading Advantage with Intact Auditory Stimuli. *In Dodd, B. and Cambell, R. (eds.) Hearing by Eye: The Psychology of Lip-Reading*, 97-113. London: Lawrence Erlbaum.

Schwartz, J.L., Basirat, A., Ménard, L., & Sato, M. (2010). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics,* 25, 336-354.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*. 17 : 2387-2399.

Stevens, K.N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. *In: Wathem-Dunn, W. (ed), Models for the Perception of Speech and Visual Form. Cambridge, MA: MIT Press.*

Sumby W.H., & Pollack I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26 (2), 212-215.