Organisation spatio-temporelle main - lèvres - son de séquences CV en Langage Parlé Complété

Virginie Attina, Denis Beautemps, Marie-Agnès Cathiard

Institut de la Communication Parlée (ICP) - UMR 5009 CNRS / INPG / Univ. Stendhal
46, av. Félix Viallet 38031 Grenoble Cedex 1, France
Mél : attina@icp.inpg.fr, beautemps@icp.inpg.fr, cathiard@icp.inpg.fr

ABSTRACT

This study was designed to investigate the coordinations in space and time between manual and oro-facial gestures involved in "Langage Parlé Complété", an efficient method of communication by hearing-impaired people. Cued CV syllabic sequences were analysed. Results showed (i) five distinct positions for vowels and (ii) manual anticipation with respect to lip movements and sound, manual information being delivered at the beginning of a CV syllable.

1. INTRODUCTION

Les malentendants s'appuient largement sur la lecture labiale pour percevoir la parole, mais en raison de la présence de sosies labiaux, leur compréhension de la parole reste généralement assez faible (en l'absence de restes auditifs). De nombreuses études ont montré que les enfants malentendants constituent, à partir de la seule lecture labiale, des représentations phonologiques sous-spécifiées, qui entravent le développement normal du langage (pour une revue récente voir [Ley00]).

C'est en 1967 que le Dr Orin R. Cornett, physicien et vice-président du Gallaudet College à Washington, crée un système manuel complétant la lecture labiale : le Cued Speech (CS) [Cor67], introduit en France en 1977 et rebaptisé Langage Parlé Complété (LPC). Le LPC permet de lever les ambiguïtés de l'information labiale par la vision : le locuteur exécute, pendant qu'il parle, une succession de gestes de la main autour de son visage, le dos de la main étant visible pour l'interlocuteur. Chaque geste est composé de deux paramètres : la configuration des doigts (position des doigts les uns par rapport aux autres) et la position de la main autour du visage. La forme adoptée par la main code les consonnes tandis que l'endroit où se place la main code les voyelles (figure 1).

Les clés sont définies de façon à ce que les phonèmes qui peuvent être confondus sur les lèvres (par exemple : [p], [b], [m]) soient représentés manuellement par des clés différentes afin de bien les différencier, tandis que ceux qui ont des images labiales clairement distinctes (par exemple : [p], [d], [ʒ]) soient regroupés sous une même clé. Une phrase se code à partir de son découpage syllabique en une suite de CV (consonne-voyelle) ou lorsque ce n'est pas possible, de consonne ou voyelle isolées.

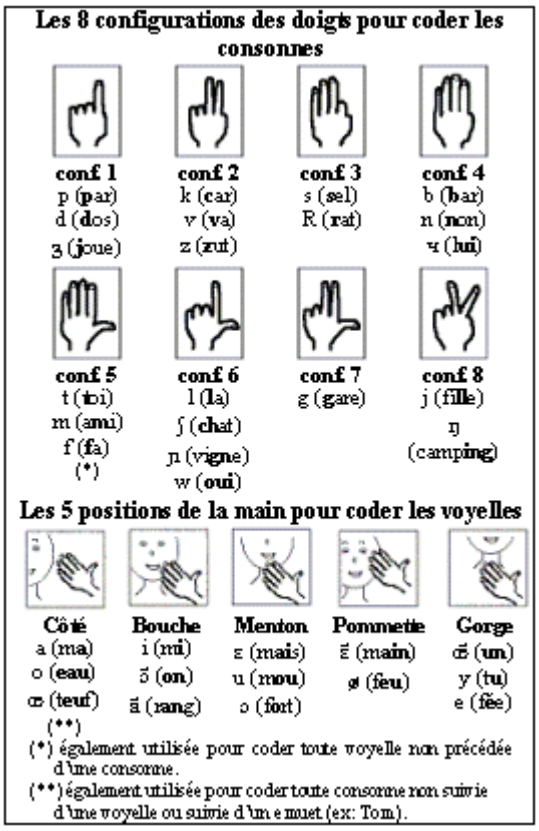

Figure1: Positions de la main et configurations des doigts utilisées en Langage Parlé Complété (en Français) pour coder les voyelles et les consonnes.

L'efficacité du LPC a pu être montrée en comparant chez des enfants sourds leur perception de syllabes ou de mots en lecture labiale seule et en lecture labiale avec codage LPC ([Nic82] [Uch94]). L'ajout du LPC apporte un gain significatif, puisque les taux passent de 30-40% en lecture labiale seule à 80-95% avec LPC : il est à noter que ces derniers résultats sont comparables à ceux d'enfants entendants en situation d'écoute de la parole. De plus, l'exposition précoce au LPC chez les enfants sourds favorise un bon développement des représentations phonologiques, permettant une bonne acquisition de la lecture, de l'écriture et de l'orthographe [Ley01].

Un système de génération automatique de Cued Speech (CS) en temps réel basé sur la reconnaissance automatique de parole existe actuellement pour la langue anglaise [Duc00]. L'image du visage d'un locuteur, filmé en train de parler, est retransmise – avec un délai de 2 secondes – sur un écran avec l'ajout d'une main codant

en CS. Ce système est géré par deux ordinateurs chargés du traitement, l'un s'occupant de l'image du locuteur et de l'affichage des clés, et l'autre s'occupant de transcrire automatiquement le son en séquences phonétiques et de générer les clés correspondantes. Afin d'améliorer leur système en perception, les auteurs ont défini de manière empirique des règles d'affichage des clés (notamment, ils ont estimé à partir d'une vidéo que la main devait précéder systématiquement le son de 100 ms). Les auteurs ont obtenu un taux moyen de réception du codage automatique de CS de 66 %. Ces résultats, meilleurs qu'en lecture labiale seule (35%), restent significativement moins bons qu'en codage CS manuel (environ 90%).

Dans ce travail, nous étudions le codage LPC par l'analyse de signaux biologiques extraits des mouvements d'un codeur de LPC. Ceci devrait permettre d'établir des règles de coordination main – lèvres – son, afin de réaliser une synthèse de LPC pour le français, c'est-à-dire intégrer une main de synthèse codant du LPC sur le clone parlant 3D de L'ICP [Odi01].

2. PROTOCOLE EXPERIMENTAL

Corpus

Le corpus utilisé pour cette étude a été élaboré pour : (i) caractériser spatialement les positions vocaliques de la main autour du visage et ainsi vérifier si elles sont bien distinctes comme le stipule le LPC et (ii) définir les relations temporelles entre main, lèvres et son pour des syllabes CV.

Etant donné que chaque position de la main code plusieurs voyelles et afin de ne pas allonger le corpus, nous choisissons la voyelle qui est la plus claire de chaque groupe, du point de vue de l'image labiale, c'est-à-dire [a] pour la position côté, [i] pour la position bouche, [u] pour la position menton, [ø] pour la position pommette et [e] pour la position gorge. Nous étudions les consonnes [m, t] qui sont codées par la configuration de la main correspondant également à celle des voyelles seules (configuration 5 sur la figure 1), et la consonne [p] (à la fois visible sur les lèvres et bien caractérisée sur le signal acoustique), codée par une autre configuration manuelle (configuration 1). Nous étudions donc des séquences syllabiques composées d'une suite de CV. Etant donné que nous nous intéressons aux transitions entre positions de la main, nous fixons la clé digitale en conservant la même consonne tout au long d'une séquence. Ainsi, pour une consonne donnée C, la voyelle varie de façon à explorer toutes les transitions de la main entre les 5 positions. Le corpus se compose donc de trois ensembles de séquences [$CaCV_1CV_2CV_1$], où C est soit [m], [t] ou [p] et V, [a, i, u, ø, e] (par exemple, [mamamima]), soit 20 séquences au total pour chaque contexte consonantique. La première syllabe servant de position de départ (sur le côté), nous ne traitons que les 3 dernières. En ce qui concerne l'étude spatiale, nous obtenons, pour chacune des trois conditions consonantiques [m, t, p], 12 réalisations d'une même syllabe ; ce qui nous donne au total, toutes consonnes confondues, 36 réalisations d'une même position. En condition contrôle, nous étudions des séquences [$CaV_1V_2CV_1$] (par exemple, [maaima]) qui nous permettront de voir une éventuelle variation de position due à la consonne. En ce qui concerne l'étude temporelle, nous étudions, pour chaque séquence « $S_1S_2S_3S_4$ » (où S_i est une syllabe de type CV), la syllabe S_3 grâce aux transitions de S_2 à S_3 et de S_3 à S_4 ; ce qui revient au total à 60 (20 séquences x 3 consonnes) séquences.

Locuteur

La locutrice, âgée de 36 ans et titulaire du diplôme de codeuse LPC depuis 1996, pratique le LPC quotidiennement depuis plus de 8 ans, en raison de la surdité de son enfant.

Matériel et procédure

L'enregistrement a été effectué au moyen du poste Visage-Parole de l'ICP [Lal91] par deux caméras en vue de face, l'une filmant le visage du locuteur et sa main en vue d'ensemble, l'autre filmant ses lèvres en gros plan. Les lèvres du sujet ont été maquillées en bleu (figure 2) afin de récupérer avec précision les contours des lèvres, et des pastilles colorées ont été placées sur le dos de sa main droite pour récupérer les mouvements de sa main – la pastille bleue, plus proche du poignet, étant celle retenue pour notre étude – et sur le bout de l'index pour récupérer la position des cibles. Enfin, le sujet portait des lunettes aveugles dont le centre était repéré par une pastille de couleur servant par la suite de point de référence. La tête était maintenue par un système de casque fixe afin de limiter toute mobilité.

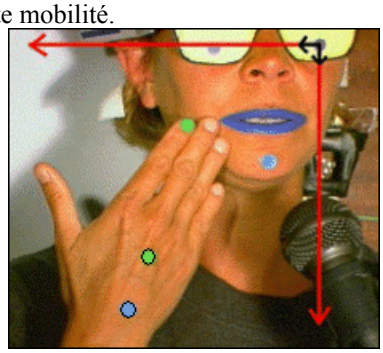

Figure 2. Photo du locuteur-codeur avec superposition du repère (en rouge) et des vecteurs directeurs (en noir) x et y choisis pour le traitement des pastilles de la main.

Traitement des données

Le son a été échantillonné à une fréquence de 22050 Hz. Les images correspondant aux séquences ont été numérisées à une fréquence d'échantillonnage de 25 Hz et détramées pour le traitement (ce qui nous permet d'avoir une information toutes les 20 ms). Grâce au système de traitement des images TACLE ([Aud00]), nous avons pu mesurer pour chaque séquence le décours

temporel de l'aire intérolabiale (S) avec un point toutes les 20 ms. Nous avons également mis au point un logiciel de suivi de pastilles colorées qui nous délivre, en synchronie avec le signal de S et le son, les trajectoires des coordonnées en x et en y du barycentre de la pastille étudiée (se trouvant sur le dos de la main près du poignet) mesurées en référence avec le centre de la lunette droite.

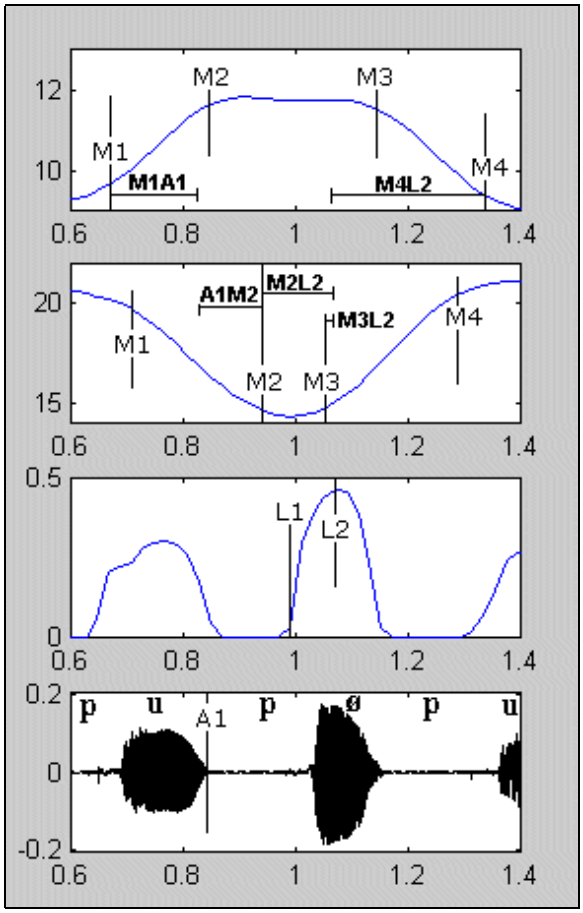

Figure 3. De haut en bas : pour la portion [pupøpu] extraite de la séquence [papupøpu] au cours du temps, trajectoire de la coordonnée x (en cm) du geste de la main (quand x augmente, la main s'éloigne de la pastille de référence), trajectoire de la coordonnée y du geste de la main (quand y augmente, la main descend) (cf. repère sur figure 2), décours temporel de l'aire aux lèvres et signal acoustique correspondant. Sur chaque signal, instants repérés sur l'ensemble des séquences et écarts temporels analysés.

Sur chacune des trajectoires (figure 3), le début des transitions a été repéré par la position du pic d'accélération et la fin par le pic de décélération. Nous obtenons ainsi, pour le signal de la main, les repères suivants sur une séquence « $S_1S_2S_3S_4$ » : M1 est le démarrage du geste LPC vers la position correspondant à S_3, M2 l'atteinte de la position cible (codant S_3) qui sera tenue jusqu'à M3, date à laquelle la main démarre le geste vers la position suivante qui code S_4, M4 correspondant à l'atteinte de cette cible. Il est à noter que, si ces repères sont différents pour la trajectoire en x et en y, on ne garde que le repère le plus précoce pour M1 et M3 et le repère le plus tardif pour M2 et M4. Notons cependant que la cible LPC a été repérée de façon à se trouver à la fois en x et en y dans l'intervalle entre l'instant d'atteinte de cible et l'instant de début de geste LPC. Pour les lèvres, nous repérons L1 pour le démarrage du geste labial vocalique de S_3 et L2 pour l'atteinte de la cible. Pour le signal acoustique, nous avons A1 qui indique le début de la consonne de la syllabe S_3.

Pour l'analyse des cibles spatiales, nous avons traité les images correspondant aux cibles LPC (étiquetées sur la pastille du dos de la main) afin d'obtenir la position du doigt qui pointait sur la cible spatiale (plus précisément le bout du doigt de l'index pour le contexte [p] et le majeur pour les autres contextes consonantiques).

3. RESULTATS

Analyse des cibles spatiales

Les coordonnées x et y des cibles ont été traitées séparément par une ANOVA à 2 facteurs (position de la main*consonnes) qui montre un effet de la position (pour x : $F(4, 220)=925$, $p<0.00001$ et pour y : $F(4, 220)=2123$, $p<0.00001$), un effet de la consonne (pour x : $F(3, 220)=3.3$, $p=0.02$ et pour y : $F(3, 220)=5.7$, $p=0.0008$), et un effet d'interaction (pour x : $F(12, 220)=2.6$, $p=0.003$ et pour y : $F(12, 220)=3.6$, $p=0.00006$). Nous retiendrons principalement que, durant la production de LPC, les 5 positions de la main autour du visage sont bien distinctes entre elles (figure 4).

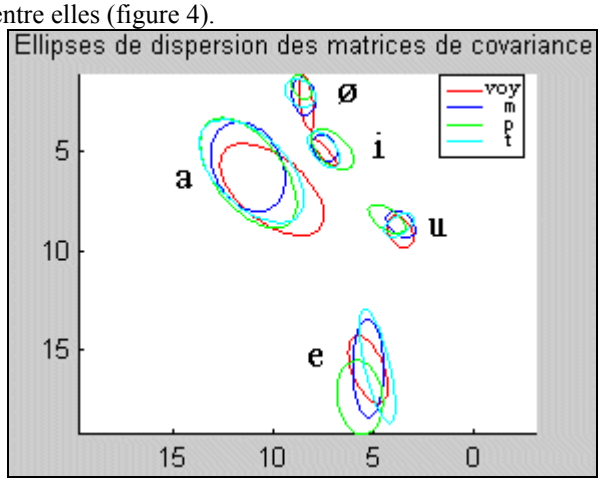

Figure 4. Ellipses de dispersion à 2 écart-types autour de la moyenne pour les positions spatiales LPC (en cm) de la main repérées par la pastille sur le doigt dans les différents contextes syllabiques.

Analyse temporelle du codage LPC de séquences CV

A partir des instants repérés sur les différents signaux (figure 3), les écarts temporels suivants ont été calculés : (i) entre le début de la consonne et le début du geste LPC correspondant (M1A1), (ii) entre le début de la consonne et l'atteinte de la cible LPC correspondante (A1M2), (iii) entre l'atteinte de la cible LPC et l'atteinte de la cible articulatoire vocalique correspondante (M2L2), (iv) entre

l'atteinte de la cible articulatoire vocalique et le début du geste LPC suivant (M3L2) et (v) entre l'atteinte de la cible articulatoire vocalique et l'atteinte de la cible LPC suivante (M4L2).

Les résultats pour M1A1 montrent que le geste en direction de la cible spatiale (côté, pommette, bouche, menton ou gorge) démarre toujours avant le début de la consonne correspondante avec une avance moyenne de 239 ms. Plus précisément, selon la consonne, la main peut être plus ou moins en avance de la syllabe à coder : en moyenne 302 ms avant [m], 237 ms avant [t] et 178 ms avant [p] (effet de contexte significatif $F(2, 57)=14.65$ avec $p<0.00001$; avance significativement différente de 0 avec pour chaque contexte respectif, $t=17.8$, $t=12.6$, $t=15.2$ pour $\nu=19$ ddl, valeur théorique du $t_{0.01}=2.86$). En revanche, les résultats pour A1M2 montrent que la main atteint sa position cible en moyenne 37 ms après le début de la consonne ; plus précisément, pour les contextes consonantiques [p, t], la position cible LPC est atteinte en moyenne 63 ms après le début de la consonne (avance significative : $t=5.5$ et $t=3.1$) ; par contre, pour la consonne [m], le geste LPC se termine en moyenne au moment où la consonne débute ($t=0.93$). De façon plus générale, pour une syllabe CV donnée, la main atteint sa position cible LPC pendant la réalisation de la consonne.

En ce qui concerne la cible labiale vocalique (M2L2), elle est atteinte largement après la cible LPC, soit 256 ms en moyenne : en moyenne 307 ms après pour le contexte [m], 255 ms pour le contexte [t] et 204 ms pour le contexte [p] (effet de contexte significatif $F(2, 57)=6.12$ avec $p=0.004$; retard significativement différent de 0 avec pour chaque contexte respectif, $t=13.8$, $t=11.3$, $t=11.9$). En se référant à cette même cible vocalique, les résultats pour M3L2 montrent que le geste manuel pour coder la syllabe suivante démarre en moyenne 51 ms avant l'atteinte de cette cible, quelle que soit la condition consonantique (effet de contexte non significatif $F(2, 57)=0.89$ avec $p=0.4$; avance significativement différente de 0 pour [m, t, p], $t=4.6$, $t=3.3$, $t=3.8$). Ainsi, la main quitte la position cible LPC de la syllabe en cours pour atteindre la suivante, avant que la cible labiale de la syllabe en cours ne soit atteinte. Enfin, les résultats pour M4L2 montrent que la cible LPC correspondant à la syllabe suivante est atteinte en moyenne 231 ms après la cible labiale vocalique (effet de contexte non significatif $F(2, 57)=0.48$ avec $p=0.6$; valeur significativement différente de 0 pour [m, t, p], $t=13.8$, $t=11.6$, $t=26.9$), soit pendant la consonne de la syllabe suivante (ce qui correspond à ce que nous obtenions déjà pour la syllabe précédente par A1M2).

4. CONCLUSION

En conclusion de cette étude, nous constatons que les voyelles sont codées par 5 positions spatiales bien distinctes les unes des autres ; durant la production de LPC, les voyelles seraient donc bien désambiguïsées grâce à l'utilisation de ces positions bien différenciées. Que pouvons-nous dire de l'organisation temporelle main-lèvres au cours de syllabes CV ? Etant donné que nous n'avons pas varié dans nos séquences la clé digitale – correspondant au codage de la consonne –, nous ne pouvons pas indiquer avec précision à quel moment le passage d'une clé à l'autre se produit au cours du geste transitionnel de la main d'une position à l'autre autour du visage. Mais il est clair que, dans la mesure où la position de la main est atteinte en tout début de syllabe, la mise en place de la clé consonantique sera elle aussi forcément anticipée par rapport à la réalisation de la consonne. Le décalage observé avec avance de la main sur les lèvres – de plus de 200 ms en moyenne d'après notre corpus – nous permet d'avancer que, contrairement à ce qui est dit classiquement dans la littérature, ce n'est peut-être pas la main qui désambiguïse la forme labiale mais la main qui spécifie dans un premier temps un ensemble d'unités possibles, la forme labiale restreignant ensuite le choix à une seule unité.

BIBLIOGRAPHIE

[Aud00] Audouy M. (2000), "Logiciel de traitement d'images vidéo pour la détermination de mouvements des lèvres", Projet de fin d'études, option génie logiciel, ENSIMA Grenoble.

[Cor67] Cornett RO. (1967), "Cued Speech", American Annals of the Deaf, Vol. 112, pp. 3-13.

[Duc00] Duchnowski P., Lum DS., Krause JC., Sexton MG., Bratakos MS. & Braida LD. (2000), "Development of speechreading supplements based on automatic speech recognition", IEEE Transactions on Biomedical Engineering, Vol.47 (4), pp. 487-96.

[Lal91] Lallouache M. T. (1991), « Un poste visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres », PhDThesis, INP Grenoble.

[Ley00] Leybaert J. (2000), "Phonology acquired through the eyes and spelling in deaf children", J. of Experimental Child Psychology, Vol. 75, pp. 291-318.

[Nic82] Nicholls G. & Ling D. (1982), "Cued Speech and the reception of spoken language", JSHR, Vol. 25, pp. 262-269.

[Odi01] Odisio M., Elisei F., Bailly G., Badin, P. (2001), "Clones parlants 3D vidéo-réalistes: application à l'analyse de messages audio-visuels", Actes des 7èmes Journées d'Études et d'Échange "Compression et représentation des signaux audiovisuels", France, pp. 141-144.

[Uch94] Uchanski R., Delhorne L., Dix A., Braida L., Reed C. & Durlach N. (1994), "Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of cued speech", Journal of Rehabilitation Research and Development, Vol. 31, pp. 20-41.

Remerciements à Mme M. Marthouret, orthophoniste au CHU de Grenoble, pour ses conseils, à Mme G. Brunel, notre codeuse LPC, pour avoir accepté les contraintes de l'enregistrement, à C. Savariaux pour son aide technique.