Un système de détection des cibles manuelles et labiales en Langue Française Parlée Complétée

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

(1) Institut de la Communication Parlée
CNRS UMR5009 /INPG/Université Stendhal,
46 Avenue Félix Viallet, 38031 Grenoble, Cedex 1, France
(2) Communication Langagière et Interaction Personne Système
CNRS UMR 5524 /UJF/INPG 385, rue de la Bibliothèque-B.P.53 - 38041 Grenoble Cedex 9
Noureddine.Aboutabit@icp.inpg.fr

ABSTRACT

The phonetic translation of Cued Speech (CS) 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. [7]) between these two flows of information. The automatic segmentation and labelling of CS hand and lip targets addressed in this contribution are thus a key factor in the mixing process. For the hand, the method is based on the Gaussian modelling of the 2D target positions and uses a criteria of minimum of velocity for the hand displacement. The vocalic lip targets are defined at the instant of minimum of velocity of the lip inner contour area parameter, constrained by the corresponding acoustic labelling.

1. INTRODUCTION

La Langue Française Parlée Complétée (LPC) héritée du *Cued Speech* (Cornett [1], Attina et al. [2]) est un code manuel utilisé pour désambiguïser la lecture labiale et ainsi améliorer la perception de la parole par les malentendants et sourds profonds (voir Leybaert et al. [3], pour une revue complète). Avec cette méthode, le locuteur pointe des positions précises sur le côté de son visage ou à la base du cou en présentant de dos des formes de main bien définies. En Français cinq positions de la main sont utilisées pour coder les voyelles et huit formes de main sont utilisées pour les consonnes (Figure 1). Une même position de la main code plusieurs voyelles, celles pour lesquelles les formes labiales sont bien contrastées. Il en de même pour les consonnes. Ainsi l'information de la main et de la forme labiale aux lèvres permettent l'identification d'un percept unique. Enfin ce système est syllabique dans le sens où la main pointant une position et présentant une forme de main précise fournit le code de la consonne C et celui de la voyelle V pour la syllabe CV (voir Attina et al. [2] et [4] pour une étude de l'organisation temporelle de la production de ce code).

La transcription phonétique du code LPC nécessite de fusionner les informations de main et de lèvres correspondantes. Dans cette perspective, l'identification de la position LPC et de la clé digitale constituent une étape dans le processus de fusion avec les formes labiales.

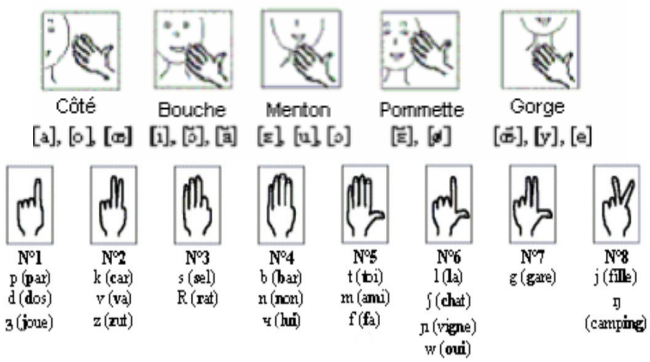

Figure 1: Positions de la main pour les voyelles et formes de mains pour les consonnes (adapté de Attina et al. [2]).

D'autre part, la segmentation temporelle du flux de la position de la main LPC est aussi un point important, compte tenu de la désynchronisation naturelle entre les flux manuel et labial (Attina et al. [2], Aboutabit et al. [7]). Cette contribution présente une méthode d'identification et de segmentation temporelle du flux manuel du code LPC ainsi que les cibles labiales repérées automatiquement dans le cas des voyelles du Français.

2. LES DONNÉES

Les données proviennent de l'enregistrement vidéo d'un locuteur français prononçant et codant en LPC un ensemble de 267 phrases. Pour cette étude, un sous-ensemble de 60 phrases répétées au moins deux fois a été utilisé ce qui a conduit à 130 phrases. Le locuteur, que l'on appellera codeur dans la suite, est français, diplômé en LPC et interprète en LPC en milieu scolaire, de ce fait pratique très régulièrement. L'enregistrement a été réalisé dans la chambre sourde de l'Institut de la Communication Parlée (ICP), à l'aide de deux caméras analogiques enregistrant à 50 Hz. Le locuteur était assis et sa tête maintenue fixe par un casque solidaire du mur afin que la tête reste dans le

champ des caméras. Le locuteur portait des lunettes aveugles afin de protéger ses yeux du fort éclairage. Sur les lunettes, des pastilles bleues sont utilisées comme repère de référence pour les mesures labiales et de position de main et de doigt. Les lèvres ont été maquillées en bleu et des pastilles de couleur bleue posées sur le dos de la main et à l'extrémité des doigts afin d'extraire les paramètres labiaux et de suivre le mouvement de la main (figure 2). Une première caméra en champ large dédiée à la main et au visage a été synchronisée à une seconde caméra en zoom sur les lèvres, chacune d'entre elle étant reliée à un magnétoscope Bétacam.

Figure 2: Image du codeur avec les pastilles de couleur sur la main et les axes servant de repère (issu de Aboutabit et al. [7]).

En début d'enregistrement, un pavé de LEDs placé dans le champ des deux caméras est allumé durant 20 ms afin d'avoir un phénomène physique commun permettant de caler par la suite les étiquettes temporelles des deux enregistrements Bétacam. De plus un tableau plan quadrillé en centimètres a été enregistré par les 2 caméras en début de session afin de permettre par la suite une conversion des pixels en centimètres pour les différents paramètres extraits des images vidéo. Utilisant le poste Image-Parole de l'ICP, la partie image de la bande vidéo a été numérisée comme images Bitmap toutes les 20 ms, en synchronie avec la bande son numérisée à 22050 Hz. Le contour interne des lèvres ainsi que les paramètres labiaux correspondants (l'étirement A, l'aperture B et l'aire intérolabiale S) ont été extraits des images en utilisant le système de traitement des images maquillées de l'ICP (Lallouache [5] et Auduy [6]). Ces paramètres ont été filtrés par un filtre moyenneur. Les coordonnées 2D x et y du centre de gravité des pastilles placées sur le dos de la main ont été extraites et repérées par rapport au centre de la pastille placée sur la lunette droite.

Le signal acoustique a été ensuite automatiquement étiqueté au niveau phonétique en utilisant les outils d'alignement (une description un peu plus détaillée du système de reconnaissance automatique de la parole peut notamment être trouvée dans Lamy et al. [8]). En effet, la transcription de chaque phrase prononcée par le codeur étant connue, un dictionnaire de prononciation a été utilisé pour produire la séquence de phonèmes correspondant à chaque signal. Cette séquence est ensuite alignée avec le signal en utilisant des modèles acoustiques HMM du Français appris sur la base BRAF100 (Vaufreydaz et al. [9]). A l'issue de cette étape, un étiquetage phonétique temporel du signal acoustique est disponible, pouvant comporter un certain nombre d'erreurs dû au dictionnaire de prononciation. L'ensemble des traitements a conduit à un ensemble cohérent de signaux : les coordonnées x et y du centre de la pastille placée sur le dos de la main près des osselets, toutes les 20 ms, les valeurs des paramètres labiaux extraits des contours internes et externes, également toutes les 20 ms, ainsi que la réalisation acoustique du signal correspondant accompagnée de sa segmentation et son étiquetage phonétique, corrigé manuellement.

3. SEGMENTATION TEMPORELLE DU FLUX MANUEL

Les trajectoires x et y présentent des transitions plus ou moins rapides entre valeurs extrêmes, de durée plus ou moins longues, caractéristiques des cibles du code LPC. La segmentation temporelle consiste à déterminer automatiquement les limites entre positions cibles et transitions.

La première étape a consisté à affecter à chaque couple de coordonnées x et y, c'est-à-dire toutes les 20 ms, un numéro de position de main parmi les cinq du code LPC. La méthode s'appuie sur le maximum de vraisemblance selon une modélisation gaussienne des coordonnées x et y des pastilles de la main et des doigts. Cette classification a été choisie pour sa simplicité et notamment du fait de l'homogénéité des dispersions des positions. Chacune des cinq positions a ainsi été modélisée par deux gaussiennes bi-dimensionnelles construites à partir d'un dictionnaire de 30 images cibles sélectionnées par un expert. La première modélise la position 2D de la pastille de référence du dos de la main, la seconde modélise celle de la pastille placée à l'extrémité du doigt directeur. Ce second modèle est utilisé pour pondérer le premier afin d'améliorer la robustesse de la méthode de classification.

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

former des plateaux cibles. A ce niveau de classification, il n'est pas possible de définir des transitions entre les positions cibles de la main du fait de la méthode du maximum de vraisemblance qui fournit toujours une solution, même si celle-ci est très peu probable. Une seconde étape a donc consisté à filtrer les cibles potentielles par application d'un critère non linéaire, afin d'affiner la taille des plateaux cibles. Le critère est un seuil ajouté au minimum de vitesse de déplacement de la pastille de référence du dos de la main, afin de définir l'intervalle de ralentissement caractéristique de l'atteinte d'une position cible. La vitesse *v(t)* au point de coordonnée (x(t), y(t)) du centre de gravité de la pastille de référence a été définie comme étant la distance euclidienne entre les deux points (x(t), y(t)) et (x(t+Δ), y(t+ Δ)) successifs ramenée à l'espacement temporel Δ de 20 ms. Pour le traitement d'un plateau donné de positions cibles identiques, l'instant de vitesse minimum est repéré. Afin de prendre en compte la rapidité variable de déplacement de la main dans les transitions qui influe la valeur du minimum de vitesse, le contraste de vitesse entre le pic de vitesse précédent (recherché dans l'intervalle défini par le milieu du plateau précédent et celui du plateau considéré) et le minimum de vitesse considéré est utilisé. Un pourcentage de 40 % (fixé de manière empirique) de ce contraste est retenu et ajouté au minimum de vitesse, ce qui définit une valeur seuil de la vitesse notée v_s au dessus de laquelle les points (x(t), y(t)) du plateau sont exclus de la position cible et considérés dans la transition. Inversement les points du plateau dont la vitesse v(t) est en dessous du seuil v_s sont définis comme étant dans la cible. Cette étape de filtrage permet de supprimer de faux plateaux cibles. Le résultat final (Figure 3) définit des bornes de plateau qui correspondent à l'instant d'atteinte de position cible LPC (noté M2) et l'instant de fin de tenue (M3), selon la nomenclature définie par Attina et collègues ([2]).

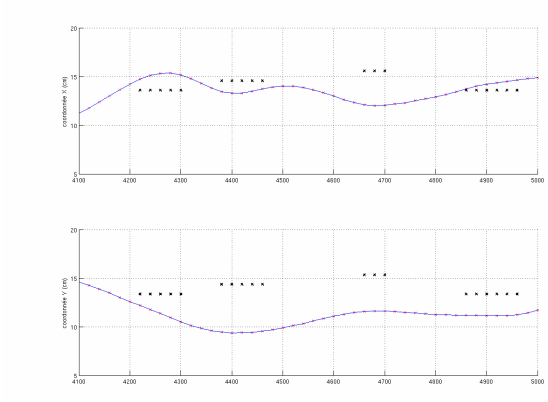

Figure 3 : Plateaux de labels fournis par la chaîne complète de traitement superposés aux trajectoires des coordonnées x (en haut) et y (en bas) du centre de gravité de la pastille de référence sur le dos de la main.

4. LES CIBLES LABIALES DES VOYELLES

L'objectif de cette partie est de repérer les cibles labiales des voyelles contenues dans des phrases. Une solution est de s'appuyer sur l'étiquetage phonétique du signal acoustique qui fournit la segmentation des phonèmes (début et fin), l'hypothèse initiale étant que l'instant d'atteinte de cible labiale se trouve dans cet intervalle. Or, en plus de la désynchronisation possible et bien connue entre les flux auditif et visuel, se pose le problème de l'imprécision des instants de début et de fin ce qui nous a conduit à rechercher la cible labiale autour de l'instant du *milieu* de cet intervalle [début, fin] sans être contraint par les bornes. Le critère est de définir la cible labiale à l'instant de minimum local de vitesse du paramètre labial considéré, le plus proche de l'instant *milieu*. La vitesse des lèvres est estimée en calculant la distance euclidienne entre les deux points S(t) et S(t+Δ) successifs ramenée à l'espacement temporel Δ de 20 ms (S étant l'aire intérolabiale du contour interne des lèvres). Le choix du paramètre S est justifié par le fait que S est fortement corrélé au produit A×B (r=0.9591). En effet, la vitesse des lèvres peut être calculée sur deux composantes verticale (sur B) et horizontale (sur A). Le système de recherche des cibles pour les voyelles fonctionne donc en 4 étapes: (1) Calcul de la vitesse labiale sur le paramètre S, (2) Recherche de tous les minima locaux de cette vitesse, (3) Localisation de la voyelle et de l'étiquette *milieu* de l'audio, (4) Estimation de l'instant de cible labiale de la voyelle par l'instant de minimum local le plus proche de l'instant *milieu*. La Table 1 présente les valeurs des différents paramètres labiaux mesurées à l'instant de cible labiale. Les Figures 4 et 5 illustrent la Table 1 avec deux exemples de dispersion dans le plan (A, S) autour des valeurs moyennes, pour les groupes de voyelle composant respectivement la position *Menton* et la position *Côté* du code LPC.

Table 1 : Valeurs moyennes (m) et écart-types (σ) des paramètres (A, B, S) pour 652 voyelles.

	A (en cm)		B (en cm)		S (en cm^2)	
voyelle	m	σ	m	σ	m	σ
ø	3,40	0,51	0,73	0,17	1,65	0,40
œ	3,45	0,32	1,18	0,25	2,71	0,74
ɛ	4,39	0,29	1,32	0,22	3,84	0,78
ɔ	3,64	0,18	1,32	0,29	3,20	0,60
a	4,24	0,36	1,36	0,33	3,89	1,08
ã	3,87	0,20	0,97	0,21	2,40	0,56
e	4,36	0,21	1,22	0,18	3,45	0,55
i	4,33	0,27	1,34	0,21	3,84	0,75
œ̃	4,24	0,19	1,33	0,21	3,74	0,72
o	3,60	0,13	0,79	0,11	1,86	0,27
õ	3,43	0,19	0,69	0,11	1,51	0,27
u	3,10	0,57	0,61	0,15	1,31	0,40
ɛ̃	4,20	0,18	1,38	0,26	3,94	0,78

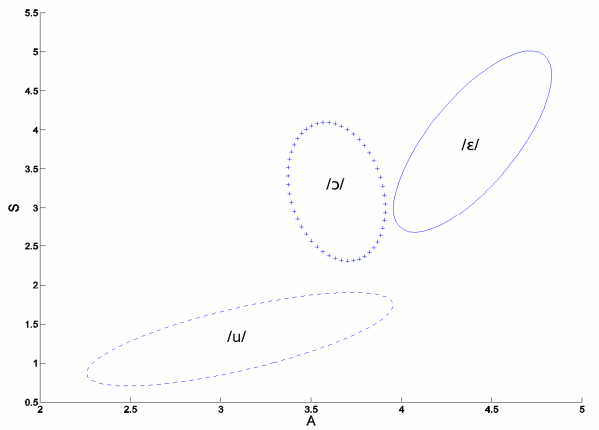

Figure 4 : Ellipses de dispersion à 1.5 écart type des voyelles composants le groupe de la position menton en code LPC dans le plan (A, S) (A en cm et S cm^2).

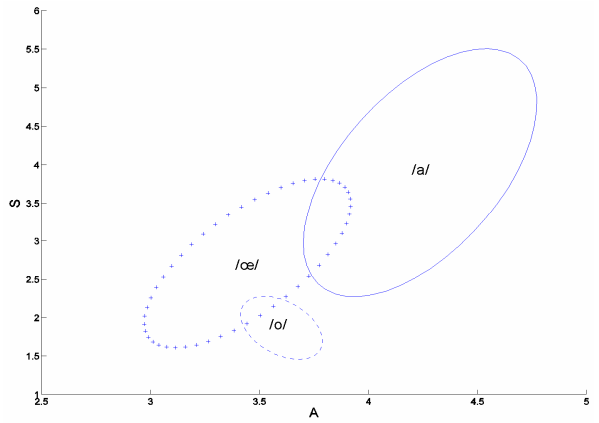

Figure 5 : Ellipses de dispersion à 1.5 écart type des voyelles composants le groupe de la position coté en code LPC dans le plan (A, S) (A en cm et S cm^2).

Ces figures montrent clairement trois groupes de voyelle distincts pour les positions du code LPC, illustrant l'efficacité des paramètres labiaux de contour interne pour identifier les voyelles étant connu la position LPC de la main. Il est à noter que ces données sont obtenues en contexte phonétique complexe (phrases), ce qui implique des effets de coarticulation et de réduction vocalique ayant une incidence sur la dispersion et l'écart entre les moyennes des groupes et donc des différences pour la discrimination par rapport à des cibles mesurées sur des voyelles isolées en contexte tenu. Enfin, on observe bien la complémentarité main-lèvres qui est la caractéristique majeure du code LPC.

5. PERSPECTIVES

Le principe général présenté dans cette contribution utilise les minima locaux de vitesse pour définir indépendamment la position de la main LPC et la cible labiale dans le cas des voyelles. Dans une perspective de fusion des flux manuel et labial, il reste à étendre la méthode au traitement des consonnes, voire des syllabes de type CV ou de structure plus complexe. Dès à présent on peut prévoir que le traitement des consonnes nécessitera de prendre en compte le contexte vocalique et donc une partie ou l'ensemble de la trajectoire entre la consonne et la voyelle par exemple. Dans cette perspective, la méthode de repérage des cibles labiales en utilisant le minimum de vitesse reste applicable pour la consonne et permettra de définir l'intervalle à considérer dans le traitement de la trajectoire.

6. REMERCIEMENTS

Nous tenons à remercier Sabine Chevalier, codeuse en LPC, qui a été le sujet enregistré pour cette étude et qui a bien voulu supporter les conditions d'enregistrement. Ce travail est soutenu par le BQR de l'Institut National Polytechnique de Grenoble et par le réseau RNTS.

BIBLIOGRAPHIE

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

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

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

[4] Attina, V., Organisation temporelle de la production et de la perception du Langage Parlé Complété. PhD Thesis. Institut National Polytechnique: Grenoble – France, 2005.

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

[6] M. Audouy, "logiciel du traitement d'images vidéo pour la détermination de mouvements des lèvres," Projet de fin d'études, ENSIMA Grenoble, 2000.

[7] N. Aboutabit, D. Beautemps, L. Besacier, "Hand and Lips desynchronization analysis in French Cued Speech: Automatic segmentation of Hand flow". In Proc. of ICASP, accepté.

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

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