Adaptation de la production labiale d'un participant sourd et classification : le cas des voyelles en contexte du code LPC.

Noureddine Aboutabit, Denis Beautemps, Olivier Mathieu, Laurent Besacier

Adaptation de la production labiale d'un participant sourd et classification : le cas des voyelles en contexte du code LPC

Noureddine Aboutabit[1], Denis Beautemps[1], Olivier Mathieu[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

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. This contribution focuses on the lip flow modeling in the case of French vowels. Previously, classification models have been developed for a professional normal hearing CS speaker (Aboutabit et al., [7]). These models are used as a reference. Now, we process the case of a deaf CS speaker and discuss the possibilities of classification. The best performance (92,8%) is obtained with the adaptation of the deaf data to the reference models.

Keywords: Lipreading, Lip Modeling, Vowel Classification, Cued Speech.

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 est 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 Aboutabit et al., [4] pour une étude de l'organisation temporelle de la production de ce code). Dans un système de communication entre des personnes normo entendantes et des personnes malentendantes, la transcription phonétique du code LPC nécessite de fusionner l'information issue des gestes de main et de lèvres. Du fait de la conception du système LPC, les deux flux labial et manuel portent chacun une partie de l'information.

Cette contribution est centrée sur le traitement du flux labial dans le cas des voyelles. Il a été démontré que la classification des voyelles par position LPC permet d'obtenir un taux de reconnaissance de 89% en utilisant seulement trois paramètres du contour interne prises à l'instant d'atteinte de la cible labiale de la voyelle (Aboutabit et al., [7]). Dans cette étude, les données proviennent d'un enregistrement d'un codeur professionnel normo entendant sous certaines conditions (lèvres maquillées en bleu, tête fixée par un casque). Cependant, que devient la classification des voyelles produites par un codeur sourd en contexte du code LPC ? Et comment, en s'appuyant sur la modélisation des données du codeur professionnel, considéré comme une référence, remédier à la variabilité inter-codeurs ? Cette contribution se propose d'apporter quelques éléments de réponse à ces questions importantes.

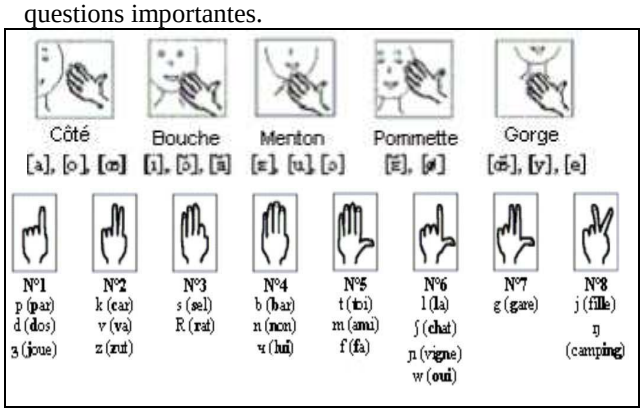

Figure 1 : positions de la main pour les voyelles et formes de mains pour les consonnes (adapté de Attina et al. [2]).

2. LE MATÉRIEL EXPERIMENTAL

Les données utilisées dans cette expérience proviennent d'un enregistrement d'une jeune femme sourde profonde codant le code LPC et participant à une expérimentation d'une conversation téléphonique avec un normo entendant. Ce participant, que l'on appellera codeur sourd par la suite, pratique le code LPC quotidiennement notamment pour communiquer avec d'autres personnes sourdes. A son insu, l'expérimentation utilise le paradigme de Magicien d'Oz afin de placer le codeur sourd dans une situation d'usage réel d'un service de téléphonie: ainsi le codeur sourd croit que la personne normo-entendante se trouve

dans un autre bâtiment en liaison avec un service de codage LPC situé dans un centre de France Télécom et que l'image vidéo du traducteur LPC lui est retransmis sur son terminal par une liaison de type visiophonie. De même, le codeur sourd croit que ses gestes LPC sont automatiquement reconnus et transformés en son de parole transmis à la personne normo-entendante. Alors qu'en réalité, la personne normo-entendante et le traducteur LPC, tous deux complices de l'expérience se trouvent dans des pièces contiguës de celle où se trouve le codeur sourd, un réseau multimédia permettant la transmission de la parole audio-visuelle. Dans ces conditions expérimentales, et de manière similaire à l'étude du codeur professionnel (Aboutabit et al., [7]), l'information des lèvres et de la main du codeur sourd était marquée par des artifices (voir Figure 2). Par contre, pour être conforme à une situation d'usage réel, un certain nombre de contraintes expérimentales ont pu être levées : ainsi le codeur sourd était libre de ses mouvements (tête non fixée) et du choix de son lexique dans un cadre de communication (réservation d'un voyage auprès d'une agence et prise de rendez-vous auprès d'un secrétariat médical). Un éclairage indirect a pu être utilisé permettant de conserver la possibilité d'extraire les contours marqués avec l'avantage d'éviter au codeur sourd d'avoir à supporter une protection oculaire. L'image vidéo et la parole sonore du codeur sourd ont été ainsi enregistrées. La figure 2 illustre les conditions expérimentales.

Le corpus obtenu dans ces conditions comporte 1026 voyelles du Français (table 1). Utilisant le poste Image-Parole du département Parole & Cognition de GIPSA-lab, les images des bandes vidéo de l'enregistrement ont été numérisées comme images Bitmap toutes les 20 ms, en synchronie avec la bande son numérisée à 44100 Hz.

Figure 2 : images des codeurs. A gauche : le codeur normo-entendant de l'étude précédente ([7]) ; A droite : le codeur sourd analysé dans cette étude.

L'information labiale est extraite directement des images maquillées à l'aide d'un traitement qui localise d'abord les contours interne et externe des lèvres, et ensuite détermine les évolutions temporelles des paramètres A, B et S (respectivement étirement, aperture et aire intérolabiale). 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. [5]). 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. [6]). 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 (voir figure 3) : les valeurs des paramètres labiaux extraits du contour interne, toutes les 20 ms, et la réalisation acoustique du signal correspondant accompagnée de sa segmentation et de son étiquetage phonétique, corrigé manuellement.

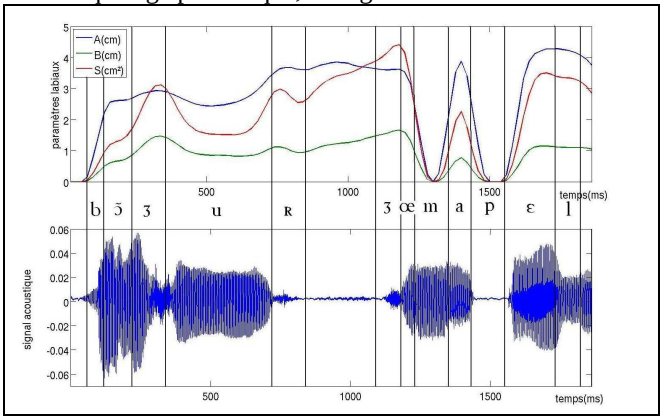

Figure 3 : les paramètres du contour interne des lèvres (A, B et S) et la réalisation acoustique.

3. MODÉLISATION

3.1 Localisation des 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). La recherche des instants de cibles labiales des voyelles, que l'on notera dorénavant L2, est appliquée sur toutes les séquences du corpus. La table 1 présente les effectifs par voyelle.

Table 1 : effectifs par voyelle.

Voyelle	a	o	œ	ɛ̃	ø	i	ɑ̃	ɔ̃	ɛ	u	ɔ	y	e
effectif	200	63	19	40	46	162	59	57	118	58	29	45	130

3.2 Comparaison des formes labiales avec la référence

A l'instant L2, les valeurs des paramètres A, B et S du contour interne des lèvres sont extraits pour chacune des voyelles. La figure 4 présente la distribution des trois visèmes de voyelles dans le plan (A, S) pour les deux codeurs, permettant de montrer la différence entre les valeurs obtenues pour le codeur sourd et celles pour le codeur professionnel du fait de la variabilité de la géométrie labiale entre les deux.

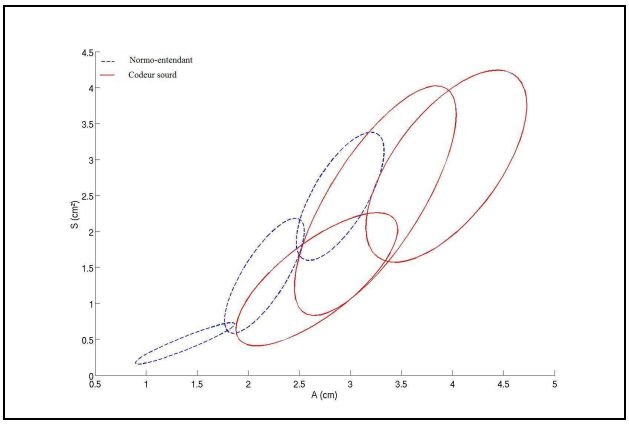

Figure 4 : les ellipses de dispersion à 1,5 écart-type autour des moyennes de chacun des 3 visèmes dans le plan (A, S) pour le codeur normo-endendant et le codeur sourd.

3.3 Classification des voyelles

Conformément à la modélisation proposée dans Aboutabit et al., [7], un classifieur gaussien tridimensionnel des paramètres labiaux est considéré pour chaque position LPC de la main. Ainsi, deux approches sont comparées : Avec et sans apprentissage. Dans la première, le corpus est divisé en deux de telle sorte que la première moitié sert à l'estimation des paramètres des classifieurs (moyennes et matrices de covariances) et la seconde moitié est utilisée pour l'évaluation. Dans la seconde approche, les modèles gaussiens du codeur de référence sont utilisés tel quels pour la classification des voyelles du codeur sourd, nécessitant une phase d'adaptation des données. Pour l'adaptation, nous avons considéré deux approches. La première consiste en une translation des moyennes des données du codeur sourd vers les moyennes du codeur de référence en laissant les écarts-types inchangés (voir équation a). La seconde complète la translation par une phase de réduction des écarts types vers la référence (voir équation b).

$$(a)\ \tilde{X}_S = X_S + (m_{NE} - m_S) \quad (b)\ \tilde{X}_S = (X_S - m_S)\frac{\sigma_{NE}}{\sigma_S} + m_{NE}$$

Avec : $\tilde{X}_S$: Paramètre labial normalisé ; X_S : Paramètre labial ; m_S , m_{NE} : valeurs moyennes pour le codeur sourd et de référence respectivement. ; σ_S , σ_{NE} : écarts-types pour le codeur sourd et de référence respectivement.

Par ailleurs, nous avons appliqué ces deux approches d'adaptation sur trois regroupements différents des voyelles. Pour le premier regroupement (R1), toutes les voyelles sont considérées en ne formant qu'un seul groupe. Dans ce cas, une seule opération de normalisation est effectuée sur l'ensemble des données. Dans le deuxième type de regroupement (R2), les voyelles sont considérées en trois groupes de visèmes définis pour le codeur de référence : les voyelles non arrondies [a, ɛ̃, i, œ̃, e, ɛ], les voyelles arrondies [ɔ̃, y, o, ø, u] et les voyelles semi-arrondies [ɑ̃ , ɔ, œ] (Aboutabit et al., [7]). Enfin, dans le troisième regroupement (R3), les voyelles sont considérées séparément (13 groupes, voir table 1).

4. RÉSULTATS ET DISCUSSION

4.1 Classification avec apprentissage

Rappelons qu'ici, un classifieur gaussien des paramètres labiaux est considéré pour chaque position LPC de la main. Le taux global de reconnaissance des voyelles est de 82,5%. La figure 4 illustre les taux pour chaque voyelle en fonction de la position LPC de la main. Ce résultat est tout de même comparable au score de 89 % obtenu pour le codeur de référence.

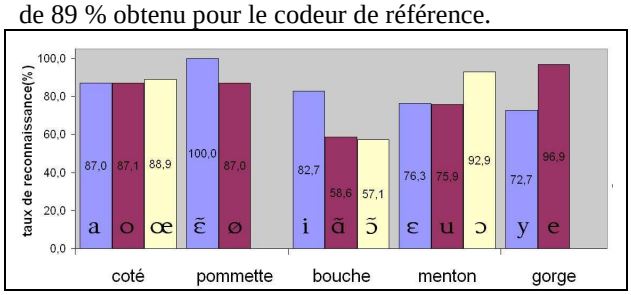

Figure 4: classification des voyelles par position LPC.

La position LPC « bouche » est la catégorie pour laquelle le taux est tiré vers le bas, notamment pour les voyelles [ɑ̃] et [ɔ̃] qui sont difficilement différentiées par les formes labiales produites par le codeur sourd, comme démontré par Sacher et al. ([8]) sur le même

sujet.

4.2 Classification sans apprentissage

Deux grandes tendances apparaissent. L'utilisation de la translation seule est moins efficace qu'avec l'ajout de la réduction des écart-types. En effet quelque soit le regroupement, le taux global est inférieur au score de 82,5% obtenu précédemment. D'autre part, l'adaptation avec le regroupement R1 donne un score quasi-identique dans les deux cas d'adaptation, qui reste nettement en deçà des deux autres. Seuls les cas R2 et R3 donnent des taux comparables voire supérieurs au score de 82,5 %. Enfin, il est à noter que les scores R2 et R3 sont très proches ce qui donne finalement une prime à la condition R2 puisque dans le meilleur des cas (Translation + réduction), seulement six coefficients (3 moyennes et 3 écart-types) sont à appliquer pour l'adaptation, en comparaison de 26 pour la condition R3.

Table 3 : taux de reconnaissance en fonction du type d'adaptation et selon le niveau de regroupement.

	Translation			*Translation + réduction*		
Voyelle	*R1*	*R2*	*R3*	*R1*	*R2*	*R3*
a	83,0	82,0	83,0	83,5	89,0	93,0
o	63,5	85,7	84,1	60,3	95,2	93,7
œ	0,0	31,6	10,5	5,3	84,2	84,2
ɛ̃	90,0	90,0	92,5	90,0	90,0	90,0
ø	60,9	73,9	76,1	54,3	100	100
i	88,9	91,4	90,1	90,1	90,1	93,2
ɑ̃	61,0	27,1	40,7	52,5	83,1	83,1
ɔ̃	5,3	28,1	59,6	5,3	96,5	86,0
ɛ	66,1	85,6	89,8	64,4	94,9	94,9
u	69,0	87,9	89,7	67,2	96,6	98,3
ɔ	48,3	69,0	82,8	55,2	100	100
y	26,7	53,3	31,1	22,2	91,1	95,6
e	96,2	98,5	96,9	95,4	98,5	100
% global	**70,4**	**77,8**	**79,8**	**69,4**	**92,8**	**93,9**

5. CONCLUSION

Les formes labiales produites par un codeur LPC sourd dans le cas des voyelles peuvent aussi être classifiées par un simple outil de classification gaussienne. Ceci dit, les meilleurs résultats sont obtenus dans le cas de l'adaptation des données vers une référence. Le meilleur des cas (92,8 % de la condition R2, pour l'adaptation « Translation + réduction ») donne une performance identique (voire supérieure) à celle du codeur de référence. Ce résultat est d'autant plus appréciable, que le codeur testé ici est sourd et que les contraintes expérimentales ont été allégées par rapport au codeur de référence (tête libre, parole spontanée). Même si un seul sujet a pu être testé, cette étude montre que l'idée de modéliser finement un codeur de référence et d'adapter tout autre codeur sur cette référence semble être une démarche fructueuse. Cette contribution ouvre la voix à une extension de cette démarche vers la classification de logatomes plus complexes tels que des syllabes de type Consonne-Voyelle.

6. REMERCIEMENTS

Nous tenons à remercier Sabine Chevalier et Juliette Huriez, les codeurs LPC, pour avoir supporté les conditions expérimentales. Ce travail est soutenu par le projet TELMA (ANR/ RNTS) (Beautemps et al., [9]).

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

[5] 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.

[6] 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 Ressources and Evaluation. Athens, Greece, pp. 423-426, 2000.

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

[8] Sacher, P., Beautemps, D., Cathiard, M.-A., Aboutabit, N., "Analyse de la production d'un codeur LPC sourd". Actes des JEP2008.

[9] 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. and Vidal, S., "TELMA : Telephony for the Hearing-Impaired People. From Models to User Tests". Actes de ASSISTH'2007, Toulouse, France, 2007.