Analyse de la production d'un codeur LPC sourd

Pablo Sacher, Denis Beautemps, Marie-Agnes Cathiard, Noureddine Aboutabit

Analyse de la production d'un codeur LPC sourd

Pablo Sacher[1], *Denis Beautemps*[1], *Marie-Agnès Cathiard*[2], *Noureddine Aboutabit*[1]

[1] Grenoble Image Parole Signal Automatique, département Parole & Cognition
46, av. Félix Viallet, 38031 Grenoble, Cedex 1, France

[2] Centre de Recherche sur l'Imaginaire, E.A 610, Université Stendhal Grenoble III

ABSTRACT

If the studies on Cued Speech (CS) perception are numerous, those refering to its production by a normal-hearing cuer are fewer, and almost non-existent in the case of a deaf cuer. This latter is the topic of this contribution. In an experiment of distant interaction between a deaf participant using CS and a normal-hearing participant, we analyze the distance between the vowels from the produced lip shapes. The results show at first that the set of vowels with similar lip shapes (so called visemes) are the same groups that for a normal-hearing cuer. Secondly, the lip shapes are in coherence with the CS system. Finally, the analysis of the temporal coordination between the CS hand gestures and the lips one, reveals the same scheme as observed at first by Attina [6] with normal-hearing cuers and confirmed by Aboutabit et al.[3] in a more complex speech context, i.e. the advance of the CS hand on the lip shapes. **Keywords**: Cued Speech, Shape recognition, Gaussian classification

1. Introduction

Cet article se concentre sur l'analyse de la production de la Langue Française Parlée Complétée (LPC) par une personne sourde. Par analyse nous entendons coordination temporelle des gestes labiaux et manuels du code LPC. Ce dernier est un complément manuel à la lecture labiale utilisé par les personnes sourdes de tradition oraliste. Il a pour but de désambiguïser les formes de lèvres similaires ([p, b, m] ou [i, e]) à l'aide d'un code manuel. Une position de la main code pour un groupe de voyelles, une configuration de la main pour un groupe de consonnes. L'association d'une position et d'une configuration code pour une syllabe de type Consonne-Voyelle (CV).

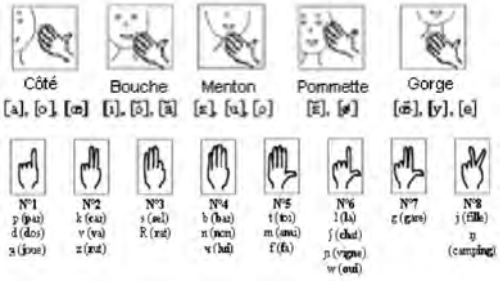

Fig. 1: Liste des positions et des configurations manuelles du code LPC, d'après Attina et al. [7]

Tout d'abord, remarquons que l'analyse effectuée entre dans le cadre du projet TELMA de téléphonie à l'usage des malentendants [1]. Ainsi, en faisant l'hypothèse forte que production et perception du code LPC sont liées, analyser la production d'un codeur sourd nous renseignerait sur les mécanismes d'intégration perceptive des informations manuelles et labiales ; ce qui aurait comme corollaire d'augmenter la pertinence et la fiabilité de la plate-forme téléphonique TELMA. C'est une des raisons pour lesquelles nous avons construit une expérimentation ayant pour objectif de recueillir des données audiovisuelles relatives à la production de code LPC par une personne sourde dans une situation d'interaction à distance.

2. Expérimentation

L'expérimentation à laquelle nous avons participé consistait à simuler une conversation téléphonique entre une personne sourde et une personne normo-entendante. Nous avons utilisé le paradigme expérimental du Magicien d'Oz afin de nous rapprocher sensiblement d'une situation d'usage courant de la plate-forme TELMA.

2.1. Paradigme Expérimental du Magicien d'Oz

Ce paradigme a été développé au début des années 80 par Kelley [2] dans le domaine de l'Interaction Homme-Machine. Son objectif fondamental est, dans un environnement expérimental, de recueillir des données les plus proches d'une situation d'usage courant. Le moyen pour y parvenir repose sur une illusion : celle du participant. Il s'agit de faire croire à ce dernier que le système est totalement opérationnel et que le but de l'expérience est de tester l'efficience et la robustesse de celui-ci. Dans ce cas, le participant agit sur le système conformément à la représentation qu'il en a, c'est-à-dire naturellement.

2.2. Participants et stimuli

Nous avons effectué l'expérimentation sur 4 participants maquillés et sur lesquels des pastilles colorées ont été posées (nous reviendrons plus tard sur l'intérêt de ces deux artifices). Cependant, pour le moment, nous n'avons exploité les données que sur l'un d'entre eux. Il s'agit d'une jeune femme sourde profonde, qui, pour communiquer avec sa soeur atteinte, elle aussi, de surdité, utilise quotidiennement

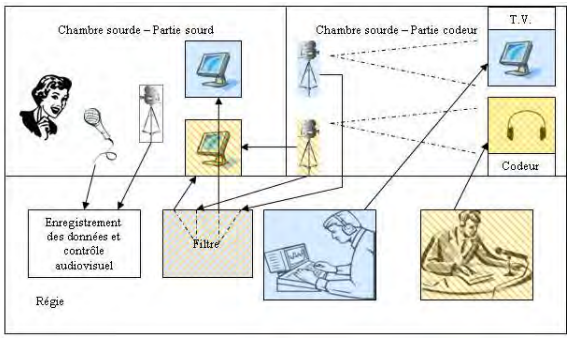

Fig. 2: Schéma de l'expérimentation.

le code LPC. Comme nous l'avons précisé dans l'introduction de cette partie, l'expérimentation consistait à simuler une conversation téléphonique entre un sourd et un normo-entendant en créant une illusion chez le participant sourd. Afin de conforter cette illusion, les dialogues étaient de deux types. Les premiers étaient préalablement rédigés (dans ce cas, le participant croyait construire une base d'apprentissage pour le système) ; les seconds (scénarios libres) étaient exempts de presque toutes contraintes dans le sens où seuls les thèmes des dialogues étaient imposés ; le participant disposant de toute sa liberté dans le choix du lexique, des constructions syntaxiques, des hésitations, etc.

Fig. 3: Repères visuels posés sur les participants.

2.3. Types de données

Concernant les données, le principal objectif de leur traitement était de dégager de le patron de coordination main-lèvres-son d'un codeur LPC sourd. Les informations pour réaliser cette tâche sont de deux natures différentes : les informations visuelles d'une part (manuelles et labiales) et d'autre part les informations sonores. Pour cela, nous avons recueilli des données audiovisuelles à l'aide d'un enregistreur Betacam dont la partie image était échantillonnée à 50 Hz. Le système CAPTURE, interne au laboratoire, a permis de numériser les données audiovisuelles (dans un fichier WAV échantillonné à 44100Hz pour le son ; sous la forme d'une série d'images BITMAP codées dans le système RGB pour la vidéo.)

Premier traitement Le premier traitement effectué sur les données extraites a été la segmentation du son en unités phonémiques. Nous avons ensuite aligné les étiquettes phonétiques sur le son de façon automatique et forcée (voir Lamy et al. [5]). De la sorte, nous avons pu disposer des instants A1, A2 et A3 correspondant respectivement, dans le cadre d'une syllabe de type CV, au début acoustique de la consonne, au début de la voyelle et à la fin de la voyelle.

Deuxième traitement Le second traitement que nous avons pratiqué était la reconnaissance automatique des contours internes et externes des lèvres - d'où l'intérêt de maquiller les lèvres du participant en bleu, une couleur très peu présente sur le visage d'un humain. Pour cela, nous avons utilisé des algorithmes développés au GIPSA-Lab (Aboutabit et al. [4]) avec comme objectif de déterminer les variations temporelles des trois paramètres A, B et S (respectivement l'étirement, l'aperture et l'aire intérolabiale).

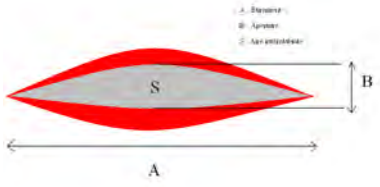

Fig. 4: Paramètres A, B et S du contour interne des lèvres

Troisième traitement Le dernier traitement consistait à reconnaître les pastilles collées sur le front et les doigts du participant, afin d'obtenir les variations de leurs mouvement au cours du temps, en utilisant les mêmes techniques que celles évoquées dans le paragraphe précédent. Le principe de détection des positions manuelles repose sur cinq modèles gaussiens (un par position). Autrement dit, après une phase d'apprentissage, nous avons calculé les probabilités $P_i(x, y)$ pour le point de coordonnées (x, y) d'appartenir à la position i avec $i = 1; 2; 3; 4; 5$, la position choisie étant celle pour laquelle $P_i(x, y)$ est maximale. La phase d'apprentissage nous a permis, sur la base d'une sélection d'images choisies par un expert, de déterminer les moyennes et écarts-types associés à chaque position, données nécessaires à la construction des cinq modèles.

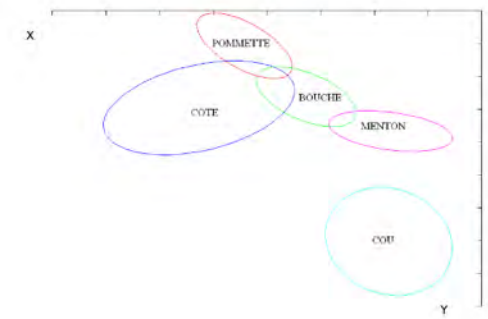

Fig. 5: Ellipses de dispersion à 2 écarts-types pour les cinq positions.

3. Détermination et classification automatique des cibles labiales des voyelles

Le seul des trois paramètres labiaux dont nous nous sommes servi pour déterminer les cibles vocaliques est l'aire intérolabiale S, car le calcul de ce paramètre fait intervenir les deux autres. Le

principe que nous avons appliqué est le suivant (Aboutabit et al. [4]) : lorsqu'une cible vocalique est atteinte, le paramètre labial présente un minimum de vitesse, c'est-à-dire que la tangente à la pente caractérisant les variations de ce paramètre au cours du temps est horizontale. Nous disposons à présent de tous les minima de vitesse de S. Ensuite nous nous sommes servis de l'instant A2 du début acoustique de la voyelle et avons cherché le minimum le plus proche de cet instant (sauf dans le cas de voyelles arrondies où nous avons cherché le dernier minimum avant A2 afin de prendre en considération les phénomènes d'anticipation). Ainsi, nous avons pu obtenir l'instant L2 de réalisation labiale de la voyelle.

Une fois les instants L2 déterminés, il est possible de catégoriser les voyelles suivant leurs paramètres articulatoire A, B et S. Pour y parvenir, nous avons utilisé une analyse de variance multivariée (MANOVA) de 14 VI (les voyelles) et 3 VD (les trois paramètres articulatoires précités) sur 1022 voyelles, afin de tester l'existence de différences significatives entre les moyennes des 14 voyelles de la langue française. Si une telle différence existe entre deux moyennes, il devient possible d'en calculer une distance. La catégorisation repose sur l'association des deux groupes de voyelles les plus proches au sens de la distance utilisée (dans notre cas la distance de Mahalanobis). Ces deux groupes forment alors un *cluster* et l'on peut calculer les distances entre ce *cluster* et tous les autres groupes afin d'en créer un nouveau. Ce principe incrémental peut se représenter sous la forme d'un regroupement hiérarchique appelé dendrogramme.

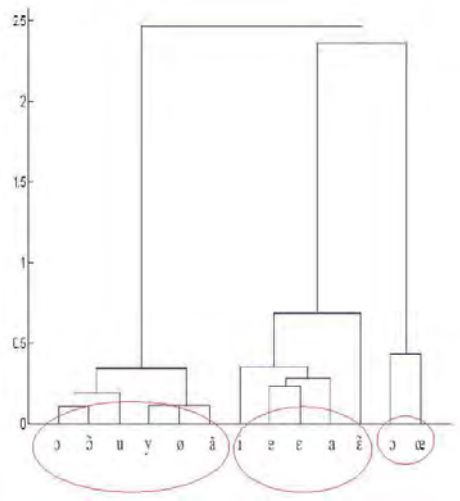

Fig. 6: Classification automatique des 1022 voyelles du corpus.

Sur le dendrogramme représentant la catégorisation des cibles vocaliques, on observe trois groupes hautement différenciés correspondant, à une exception près, aux trois groupes de voyelles protruses, étirées et ouvertes. En effet, le participant articule les voyelles /ɑ̃/ comme s'il s'agissait de /õ/, ainsi, à la place de la chaîne /ɛ g z a k t œ m ɑ̃/ il prononce /ɛ g z a k t œ m õ/. Malgré cette exception, le résultat général, qui n'avait jamais été observé pour une personne sourde, vérifie expérimentalement le principe de construction du code LPC.

4. Extraction des instants M1, M2 et M3

Une fois les positions déterminées, nous avons appliqué, pour la main, le même critère du minimum de vitesse que précédemment en vue d'obtenir les instants M1 (début de la transition d'une position vers une autre), M2 (atteinte de la position cible) et M3 (début de la transition vers la position suivante).

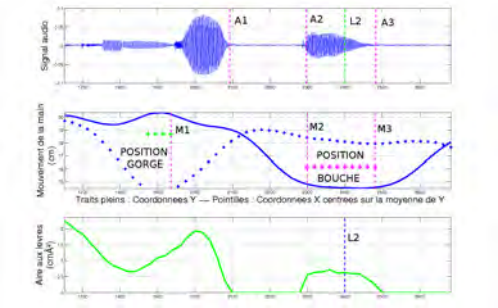

Fig. 7: Instants A1, A2, A3, M1, M2, M3 et L2 pour la syllabe CV /põ/.

Nous disposons à présent de tous les instants nécessaires pour établir le patron de coordination main-lèvres-son du codeur LPC sourd. Dans un premier temps, nous avons déterminé certaines durées, afin d'observer la distribution des différents instants. Nous avons donc calculé A1M1 ($A1M1 = M1 - A1$), A1M2, A1M3 et A1L2 et représenté ces durées dans l'ordre croissant de A1M1. Remarquons que cette analyse a été effectuée sur un jeu de 60 chaînes phonémiques de type CV_1CV_2 (où V_1 et V_2 sont des voyelles de deux groupes différents) extraits d'un scénario libre afin, d'une part, de s'assurer que la main effectue une transition entre deux positions distinctes et, d'autre part, pour utiliser les données relevant d'une situation d'usage courant.

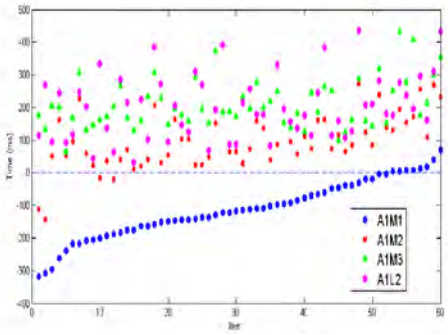

Fig. 8: Distribution des différents instants dans l'ordre croissant de A1.

D'un point de vue qualitatif, le positionnement des différents événements ne semble pas très différent de celui de Attina et al.[7] ou Aboutabit et al. [3]. On observe en effet que la transition manuelle (M1) débute avant le début de la réalisation acoustique de la

consonne (A1) - dans 86 % des cas pour notre analyse. De plus, les événements M2 semblent suivre une pente relativement similaire aux instants M1 ; phénomène que l'on pourrait interpréter comme l'indicateur d'une durée de transition assez stable. Enfin, à première vue, les instants M3 et L2 ne semblent pas dépendre du début acoustique de la consonne. Ces informations sont similaires à celles obtenues par les auteurs sus-mentionnés. Nous allons à présent dégager le patron de coordination main-lèvres-son du codeur LPC sourd.

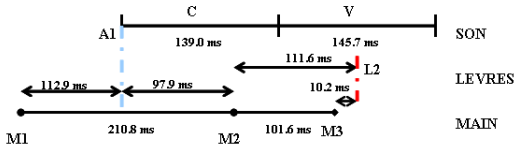

Fig. 9: Patron temporel de coordination main-lèvres-son du participant sourd.

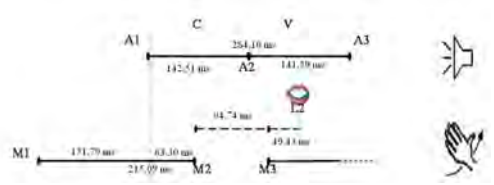

Fig. 10: Patron temporel de coordination main-lèvres-son d'un codeur normo-entendant. Aboutabit et al. [3]

Le patron du haut indique que la main débute sa transition en moyenne 112,9 ms avant le début acoustique de la consonne et qu'elle atteint sa cible au $\frac{2}{3}$ de la consonne, soit toujours bien avant la cible labiale (atteinte, elle, aux environs du milieu de la voyelle). La main reste en position 101,6 ms puis repart vers une nouvelle position 10,2 ms avant la cible labiale - ce qui signifierait qu'une fois l'information labiale rendue complètement disponible, l'information manuelle ne le serait plus. Toutefois, cette information est à nuancer, car 48,3% des valeurs de M3L2 sont négatives (dans 48,3% des cas M3 > L2, la main débute sa transition *après* la cible labiale). Ceci peut s'expliquer par le fait que dans le cas de voyelles arrondies, la réalisation labiale de la voyelle anticipe sur sa réalisation acoustique. On retiendra que la main débute sa transition vers une nouvelle position aux alentours de la cible vocalique, et donc que l'information manuelle de position est communiquée avant l'information labiale. En conséquence, pour ce codeur sourd (et pour tous les codeurs professionnels étudiés par Attina [6] et Aboutabit et al. [3]) il existe une désynchronisation entre les informations manuelles et labiales.

5. Conclusion

L'objectif de cette analyse était d'étudier la production d'un codeur LPC sourd par l'extraction de son patron temporel de coordination main-lèvres-son. Il est très important de constater qu'il est semblable à ceux dégagés par Attina [7] et Aboutabit et al. [3]. En effet, *a priori*, il ne parait pas plus incohérent de supposer l'existence de différences entre les mécanismes de production du code par une personne sourde et une personne normo-entendante, que de supposer une identité entre lesdits mécanismes. On observe, cependant, une anticipation du geste manuel sur le geste labial lors de la production de code LPC par une personne sourde dans une situation d'interaction à distance - précisons que ce patron est issu de données recueillies sur une seule personne, ce qui ne permet en aucun cas de conclure sur l'universalité de ce patron. Ainsi, il serait profitable, d'une part, de mener des recherches complémentaires, en analysant la production d'autres codeurs sourds, et d'autre part, de rechercher des procédures expérimentales permettant de comparer les mécanismes de production du code (entre des codeurs sourds et normo-entendant).

Remerciements

Nous tenons à remercier Juliette Huriez pour sa participation à l'expérience. Ce travail s'inscrit dans le cadre du projet TELMA de Téléphonie pour Malentendants (ANR / RNTS).

Références

[1] Beautemps, Girin, Aboutabit, Bailly, Besacier, Breton, Burger, Caplier, Cathiard, Chêne, Clarke, Elisei, Govokhina, Le, Marthouret, Mancini, Mathieu, Perret, Rivet, Sacher, Savariaux, Schmerber, Sérignat, Tribout, and Vidal. Telma : Telephony for the hearing-impaired people. from models to user tests. In *ASSISTH'*, 2007.

[2] Kelley J-F. An empirical methodology for writing user-friendly natural language computer applications. In *ACM CHI 83 Human Factors in Computing Systems Conference*, pages 193–196, 1983.

[3] Aboutabit N., Beautemps D., and Besacier L. Hand and lips desynchronization analysis in french cued speech : Automatic segmentation of hand flow. In *ICASSP06*, 2006.

[4] Aboutabit N., Beautemps D., and Besacier L. Vowel classification from lips : the cued speech production case. In *International Seminar on Speech Production (ISSP)*, pages 127–134, 2006.

[5] Lamy R., Moraru D., Bigi D., and Besacier L. Premiers pas du clips sur les données d'évaluation ester. In *Journées d'études sur la parole, Fès, Maroc*, 2004.

[6] Attina V. *La Langue française Parlée Complétée : production et perception.* PhD thesis, Institut National Polytechnique de Grenoble, 2005.

[7] 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*, 44 :197–214, 2004.