CHARACTERIZATION OF CUED SPEECH VOWELS FROM THE INNER LIP CONTOUR

Noureddine Aboutabit, Denis Beautemps, Laurent Besacier.*

Institut de la Communication Parlée
CNRS UMR5009/ INPG/ Université Stendhal, Grenoble, France
*Communication Langagière et Interaction Personne Système
CNRS UMR 5524 /UJF/INPG 385, Grenoble, France
Noureddine.Aboutabit@icp.inpg.fr

Abstract

Cued Speech (CS) is a manual code that complements lip-reading to enhance speech perception from visual input. The phonetic translation of CS gestures needs to combine the manual CS information with information from the lips, taking into account the desynchronization delay (Attina et al. [1], Aboutabit et al. [2]) between these two flows of information. This paper focuses on the analysis of the lip flow for vowels in French Cued Speech. The vocalic lip targets are defined automatically at the instant of minimum velocity of the inner lip contour area parameter, constrained by the corresponding acoustic labeling. We discuss in particular the possibility of discriminating the vowels with geometric lip parameters using the values at the instant of vocalic targets when associated to a Cued Speech hand position.

Index Terms: Cued Speech production, lip segmentation, vocalic lip classification

1. Introduction

Cued Speech (CS) [8] is a visual communication system that uses handshapes placed in different positions near the face in combination with the natural speech lip-reading to enhance speech perception from visual input. In this system, the speaker moves his or her hand in close relation with speech (see Attina et al. [1] for a precise 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, 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 [2] for an example of a complete analysis of the hand flow). This contribution focuses on the lip flow (lip parameters) and discusses lip-shape classification of vowels for each cued speech hand position. For example, considering the side position (Figure 1, upper left), the lip parameters characterizing the lip-shapes of the vowels [a], [o] and [œ] will be considered at the instant when the lip target is attained. The lip parameters are extracted from the inner lip contour since it is well known that this shape is precisely controlled in speech production, contrary to the external contour. This paper addresses the following questions: (1) Since the manual coding of CS is an artificial system that is superposed to the speech, do the 2D lip parameters extracted from the inner lip contour still characterize the vowels in the context of CS production? (2) Is the minimum velocity of the inner contour able to define the vocalic lip target? (3) Does the distribution of lip parameters inside each group of CS hand positions allows vowel discrimination?

Section 2 of this paper describes the experimental setup and data used. Sections 3 and 4 are dedicated to lip target segmentation and vocalic lip grouping respectively. Finally, section 5 discusses vowel discrimination inside each CS hand position, and section 6 concludes this work.

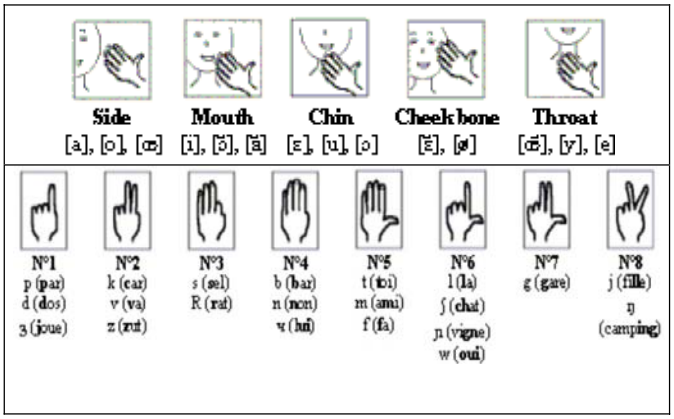

Figure 1*: CS Hand position (top) for vowels and CS handshapes (bottom) for consonants (adapted from [1]).*

2. Experimental set-up and data

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

The French CS speaker is a female native speaker of French, certified in French CS. She regularly translates into French CS code in a school setting. The recording was made in a sound-proof booth at Institut de la Communication Parlée (ICP), 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. One camera in large focus was used for the hand and the face and was connected to a betacam recorder. A second camera in zoom mode dedicated to the lips was synchronized with the first one but connected to a second betacam recorder. The lips were painted in blue, and blue marks were placed on the speaker's glasses as reference points (Figure 2).

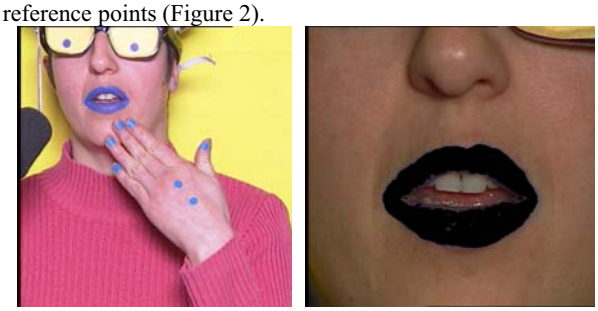

Figure 2: *Left: Image of the speaker from the camera in large focus. Right: Image of the speaker from the camera in zoom mode.*

At the beginning of the recording session, a set of LEDs was placed in the field of the two cameras and activated in order to establish the correspondence between the time codes of the two video recordings. In addition, a square paper was recorded by the two cameras 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. The image processing system developed at ICP ([3] and [4]) was applied to the Bitmap frames of the lips to extract the inner contour and to derive the corresponding characteristic parameters: lip width (A), lip aperture (B) and lip area (S).

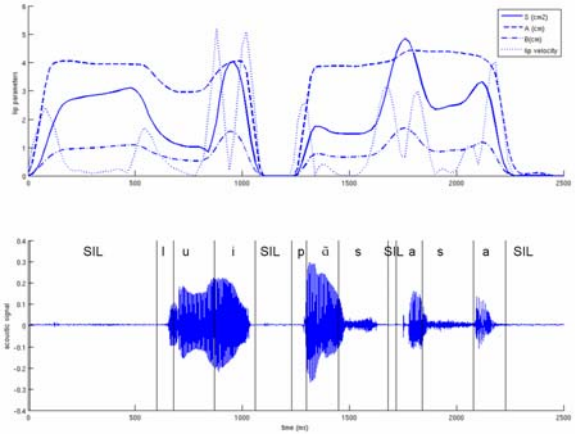

Figure 3: *Top: A (dashed line), B (dash-dot line) and S (solid line) parameters extracted from the inner contour and the corresponding lip velocity (dotted line). Bottom the corresponding acoustic realization. SIL: acoustic silence.*

These parameters were converted using a pixel-to-centimeter conversion formula. Finally the parameters were low-pass filtered. Concerning the hand movement tracking, the process is explained in detail in [2].

The acoustic signal was automatically labeled at the phonetic level using forced alignment (see [5] 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 et al. [6]).

This process resulted in a set of temporally coherent signals: the 2D hand position (see [2]) the lip width (A), the lip aperture (B) and the lip area (S) values every 20 ms, and the corresponding acoustic signal with the associated phonetic chain temporally marked. Figure 3 shows an example of the different data flows for a single sentence.

3. Lip target segmentation

The lips were characterized at the instant the lip target was attained. The automatic definition of this instant 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. Lip velocity (see Figure 3, dotted line) 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 = 0.9591).

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.

4. Vocalic lip grouping

In this section, the instant of the vocalic target is labeled manually (directly from the A, B and S lip parameters) and automatically (with the previous algorithm, see section 3) on a verified (acoustically labeled) subset of the whole recorded corpus. Table 1 shows the selected vowels and the sample size for each vowel.

Table 1: *Selected vowels and sample size by vowel.*

ø	œ	ɛ	ɔ	a	ɑ̃	e	i	œ̃	o	ɔ̃	u	ɛ̃	y
39	34	37	13	59	16	25	37	17	25	30	31	12	31

A Mahalanobis distance was computed on the basis of the A, B and S lip parameters and was used to trace the hierarchical cluster tree (dendrogram) from the vowel distribution. Figure 4 (instant of vowel targets manually labeled) and Figure 5 (instant of vowel target automatically labeled, see section 3) show the results of vowel grouping. The dendrogram consists of many U-shaped lines connecting objects (vowels or group of vowels) in a hierarchical tree. The height of each U represents the distance (using the Mahalanobis distance) between the two objects being connected. A first result is the similar grouping for both the trees, thus validating the automatic process of target instant detection. The slight differences are explained by the fact that the automatic labeling marks the instant of the beginning of the target while the manual labeling was done at the climax, which does not always correspond to the instant of the minimum lip velocity.

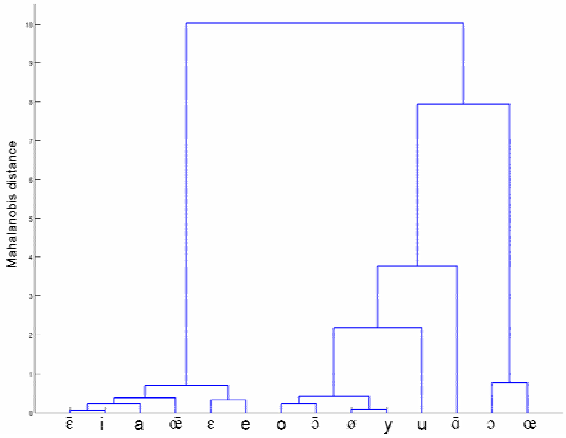

Figure 4: *Hierarchical cluster tree of the vowels labeled manually by an expert.*

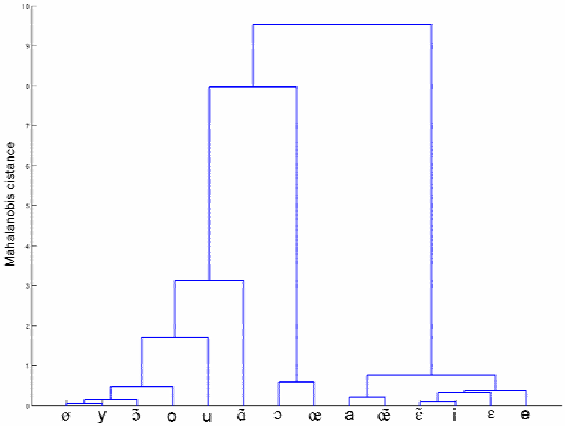

Figure 5: *Hierarchical cluster tree of the vowels labeled automatically.*

A second result is the vowel grouping into three categories (visemes, see [7]), which is in conformity with the phonetic description of the vowels (anterior non rounded vowels [ɛ̃, œ̃, a, i, e, ɛ], high and mid-high rounded [o, u, y, ɔ̃, ø, ɑ̃], low and mid-low rounded vowels [ɔ, œ]). From this, we can conclude that the extracted lip parameters from the inner contour still accurately characterize the phonetic content in the CS context of speech production.

5. Vowel discrimination inside Cued Speech hand positions

The previous grouping into three vocalic visemes is compatible with the grouping of the five CS hand positions, except for two cases. For example, the vowels [ɛ, u, ɔ] of the CS *chin* position are included in the three different visemes, and thus are well discriminated, as seen in figure 6.

The first exception corresponds to the CS mouth position with the [i, ɔ̃, ɑ̃] vowels. The [ɔ̃] and [ɑ̃] vowels are associated with the same viseme, but seem to be separated in terms of Mahalanobis distance. Figure 7 confirms this observation from the distribution in the plan (A, S) of the vowel of the corresponding CS mouth position, which shows weak covering between vowels.

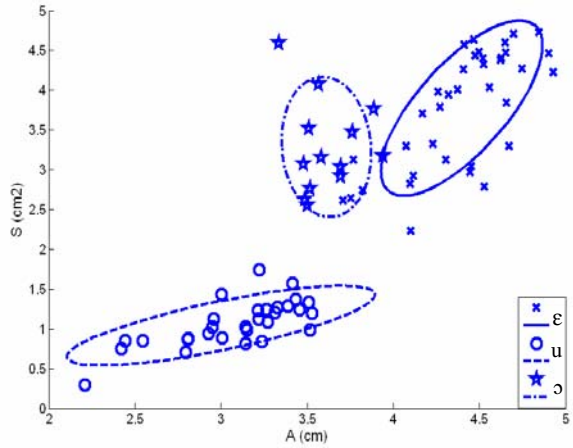

Figure 6: *French vowels with the CS chin hand position in the [A(cm), S(cm2)] plan, 1.5 standard deviation ellipsis.*

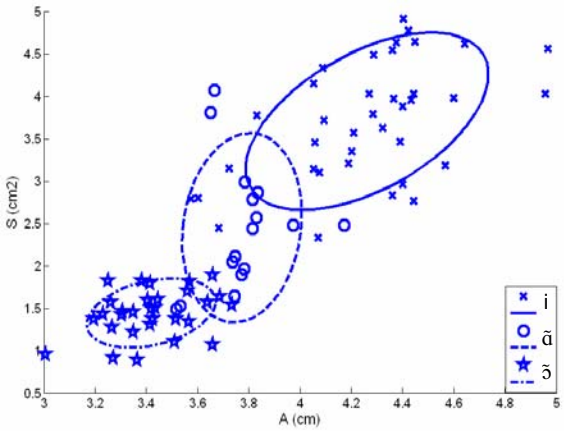

Figure 7: *French vowels with the CS mouth hand position in the [A(cm), S(cm2)] plan, 1.5 standard deviation ellipsis.*

For the second exception, i.e. the CS throat position, the [œ̃] and [e] vowels are also in the same viseme group but in contrast to the previous case, their distribution is still very closed in the (A, S) and (A, B) planes (see Figure 8 for the (A, S) plane). For this case, the discrimination between [œ̃] and [e] from the lips might be slightly tricky, even with the CS hand position information. The [œ̃] realizations are not sufficiently opened. This observation can be explained by the fact that the CS speaker does not seem to differentiate [œ̃] from [ɛ̃] with the lips, even though these two phonemes are cued with two different CS hand positions. Indeed the CS speaker produced similar realizations of lip shapes for [œ̃] and [ɛ̃], as shown by the small distance between the two corresponding distributions (see Figure 5, for example). The ambiguity is maintained due to the coding choice of the CS speaker. In this case, the complete discrimination needs a higher level of processing.

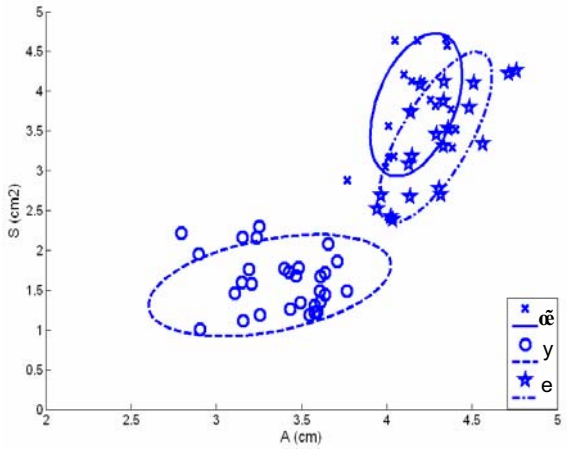

Figure 8: *French vowels with the CS throat hand position in the [A(cm), S(cm2)] plane, 1.5 standard deviation ellipsis.*

6. Conclusion and perspectives

The inner lip contour extracted from the front view of the CS speaker still characterizes lip shapes in the context of the CS production. Indeed, the parameters derived from the inner contour accurately report the phonetic features of visemes. More precisely, when the CS hand position is unknown, the viseme [ɛ̃, œ̃, a, i, e, ɛ] cannot be discriminated. But thanks to the CS system, these vowels are coded with the five different CS hand positions, thus allowing the discrimination, (except for [œ̃] and [e] for the reasons explained in section 5). This observation also applies to the [o, u, y, ɔ̃, ø, ɑ̃] and [ɔ, œ] groups of visemes. The process of automatically defining the instant of vocalic lip target from the lip velocity is efficient when the acoustic labeling is correct. In the perspective of phonetic transcription of CS gestures (hand and lips), an automatic classification based on simple Gaussian modeling of the lip parameters should be efficient for the vowels.

This approach needs to be extended to the case of consonants. It will probably be necessary to take into account vocalic context, since it is well known that the consonant realization is influenced by neighboring vowels.

7. Acknowledgments

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).

7. References

[1] 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.

[2] 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.

[3] 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.

[4] Audouy, M. “Logiciel du traitement d’images vidéo pour la détermination de mouvements des lèvres,” Projet de fin d’études, ENSIMA Grenoble, 2000.

[5] 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.

[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 Resources and Evaluation. Athens, Greece, pp. 423-426, 2000.

[7] Benoit, C., Lallouache, T., Mohamadi, T., and Abry C. “A set of French visemes for visual French speech synthesis”, in G. Bailly, C. Benoît and T.R. Sawallis (Editors). Talking Machines: Theories, Models and Designs (pp. 485-504). Amsterdam: Elsevier SC. Publishers, 1992.

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