

HAND AND LIP DESYNCHRONIZATION ANALYSIS IN FRENCH CUED SPEECH: AUTOMATIC TEMPORAL SEGMENTATION OF HAND FLOW

Noureddine Aboutabit, Denis Beautemps, Laurent Besacier*

Institut de la Communication Parlée
CNRS UMR5009 /INPG/Université Stendhal,
46 Avenue Félix Viallet, 38031 Grenoble, Cedex 1, France
*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

In the context of Cued Speech gesture phonetic translation, the automatic recognition of lip and hand movements is a key factor. The hand and the lip parameters are not synchronized, thus the fusion of the two channels (hand and lips) needs the knowledge of the desynchronized delay. This contribution focuses on the presentation of an automatic algorithm temporal segmentation of the hand Cue information based on Gaussian modeling of the hand position and minimum of velocity. The segmentation delivers the beginning of the hand transition and the instant of attained position. The hand segmentation is used to calculate the delay between hand and lip targets, in relation with the corresponding acoustic realization in the case of French CV syllables extracted from a corpus of phrases uttered and coded by a Cued Speech speaker. This study confirms in a more complex context the importance of the instant of attained hand position as pointed out by Attina and colleagues, in terms of control and for the fusion process.

1. INTRODUCTION

Manual Cued Speech (CS) [1] is an effective method used to enhance speech perception for hearing-impaired people orally educated. CS is designed to complement speech lipreading and is based on the association of lip shapes with cues formed by the hand placed at specific locations on one side of the face. While speaking, the speaker uses one of his or her hand to point out specific positions around the mouth, facing him or herself, so that the speech reader can see the back of the hand simultaneously with the lips. The manual cues are formed along two parameters: Hand Placement and Handshape. Hand Placements code groups of vowels (*Figure 1*) whereas Handshapes allow one to distinguish among groups of consonants.

In their study of the production of CV syllables in French Cued Speech (FCS), Attina and colleagues [2] observed that, on average, the hand gesture preceded that of the lips. The hand anticipates its movement in order to reach the vocalic

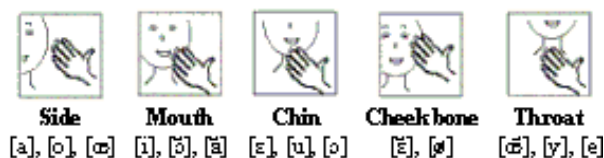


Figure 1: Hand Placement for French vowels (from [2]).

position in synchronization with the beginning of the CV syllable, i.e. in the consonant. Thus the hand is in position before the vowel being realized at the lips. In this study, the authors used a manual segmentation of the hand trajectory based on the peak of acceleration and deceleration derived from the position of the hand. Gibert and colleagues [3] also observed a similar timing of the hand with respect to the corresponding acoustic segments in a large corpus of phrases.

These previous studies on CS production reveal that the hand is the carrier gesture on which the digital gesture of the consonant is superimposed. The hand displacement between vocalic positions is speech time-locked, following the syllabic rhythm of speech. In the perspective of phonetic transcription of CS gestures in which hand and lips information have to be mixed, the detection of hand and lip targets with respect to their temporal desynchronization is a main challenge. Thus this work focuses on the automatic temporal segmentation and classification of hand positions. To work around the problem of recognition of hand shape from an image containing both the speaker's hand and face ([4]) which beyond the scope of this paper, the hand movement is followed by landmarks placed on the back of the hand and on the extremity of the fingers ([2]). The results are used to analyze the temporal desynchronization between the hand and the lips, in reference to the corresponding acoustic realization for CV syllables extracted from a corpus of phrases coded in FCS.

2. THE DATA

The data were obtained from a video recording of a FCS speaker pronouncing and coding in FCS a set of 267 phrases. For this study, a subset of 60 sentences repeated at least twice was selected, for a total of 130 phrases. The FCS speaker is a female native speaker of French, certified in FCS. She regularly translates into FCS code at school. The recording was realized in a sound-proof booth at Institut de la Communication Parlée (ICP), at 50 frames/second for the image part. The speaker was seated and head maintained fixed with a helmet in order to keep it in the camera field. The speaker was wearing opaque goggles in order 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. Blue marks were placed on the left hand, on the back and at the extremity of the fingers to independently follow the displacement of the hand and the handshape formation. Blue marks were placed on the speaker's goggles as reference points (figure 2).



Figure 2: Image of the speaker with colored marks on the hand and axes in superimposition used for landmark localization.

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 have the correspondence between the time codes of the two video recordings. As 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 digitalized 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 ([5] and [6]) was applied to the Bitmap frames of the lips to extract the inner and outer contours and to derive the corresponding characteristic parameters: lip width, lip aperture and lip area. The x and y coordinates of the center of gravity of the hand landmarks were automatically extracted from the image as follows. A process based on image processing detected all marks on the image, and the knowledge of those on the back of hand and on the goggles allowed to extract the marks on the fingers. The coordinates initially in pixels were converted into centimeters using the pixel-to-centimeter conversion formula.

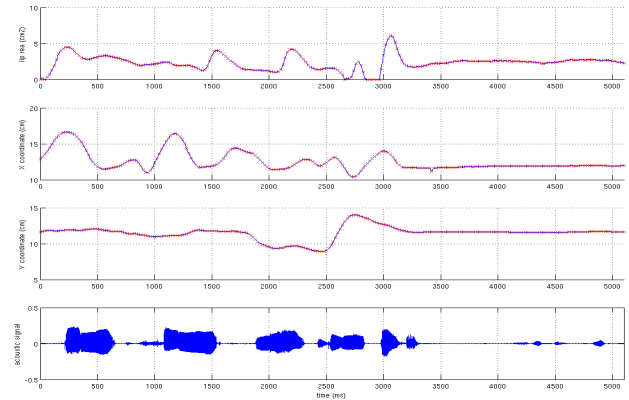


Figure 3: From top to bottom: Inner lip area, x and y coordinates of the reference hand landmark, the acoustic realization.

The whole process resulted in a set of temporally coherent signals: the x and y hand position of the reference hand landmark placed on the back of the hand near the knuckles, every 20 ms, the area lip parameter values every 20 ms and the corresponding acoustic signal.

3. HAND SEGMENTATION

The x and y trajectories are characterized by smooth deviations between local extrema corresponding to spatial FCS Hand positions. The automatic segmentation process of the hand trajectories involved automatic temporal marking of the beginning and the end of each of these segments. The first step consisted of the automatic labeling of a hand position to each frame, i.e. every 20 ms.

The method uses the likelihood of Gaussian modeling of the x-y coordinates of the center of gravity of hand landmarks. This kind of classifier was chosen for its simplicity and especially for the homogeneous dispersion of the positions (see Figure 4, the results for the reference hand landmark). Each of the five hand positions was modeled by two 2-dimensional Gaussian models built from a dictionary of 30 images manually selected in the corpus. The first one is devoted to the reference hand landmark and the second to the landmark placed at the extremity of the pointing finger. The use of the x-y coordinates of these two landmarks was needed to improve the robustness of the classifying method. For the classification phase, we consider a given frame with its x-y coordinates of

both landmarks. For each landmark x-y coordinates, a vector made of five probability densities is delivered, thanks to the five Gaussian models. The two computed vectors are combined by a scalar product in order to obtain a final vector with five components. Thus the number of the hand position is defined as the index associated to the maximum value of these 5 components of the final vector.

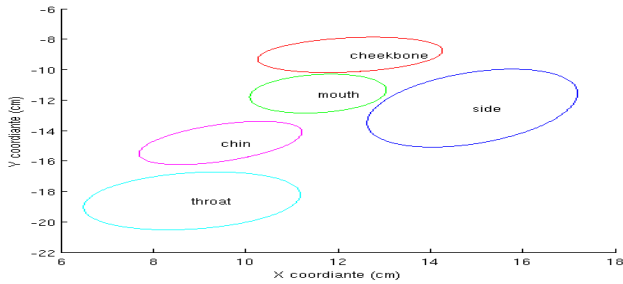


Figure 4: 2 standard deviation ellipsis around the corresponding average values of the reference hand landmark extracted from the learning data, for the 5 hand positions.

This method applied to each frame of a phrase delivers a sequence of hand position number from 1 to 5, with a set of plateaus. A plateau is defined by a set of successive identical position number.

At this step of classification, it is not possible to distinguish the transitions between attained hand positions. Thus, a second step was needed to refine this result. Its principle was based on the use of the velocity minimum applied to the x-y coordinated of the reference hand landmark. The velocity was defined as the Euclidean distance between two successive (x-y) points temporally spaced by 20 ms. Inside each plateau, the value of the velocity minimum is detected. In addition, the value of the velocity maximum is detected between the middle of the previous plateau and the middle of the considered one. The contrast is calculated as the difference between these two extreme values. A percentage (40%) of this contrast is added to the minimum value in order to define a threshold value. Thus for the considered plateau, the positions for which the velocity is lower than the threshold value are considered in the target hand position. In other hand, the positions for which the velocity is higher than threshold value are considered in the transition. Finally, incorrect plateau detections (see comparison between Figure 5 and Figure 6) are considered as points in a transition.

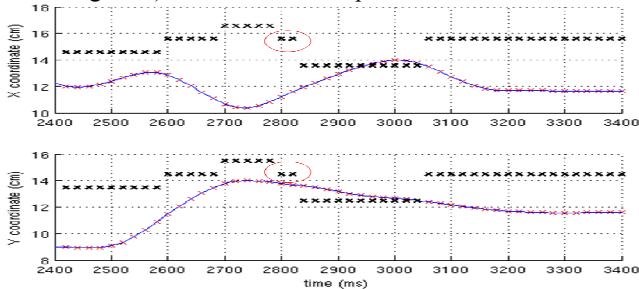


Figure 5: Zoom of signals with hand positions plateaus delivered by the classifier.

Following the nomenclature of Attina and colleagues, the extremity of the plateaus delimiting the attained hand position and the transitions were automatically labeled M1,

M2 and M3 for the onset of the transition, the onset and the end of the reached hand position, respectively.

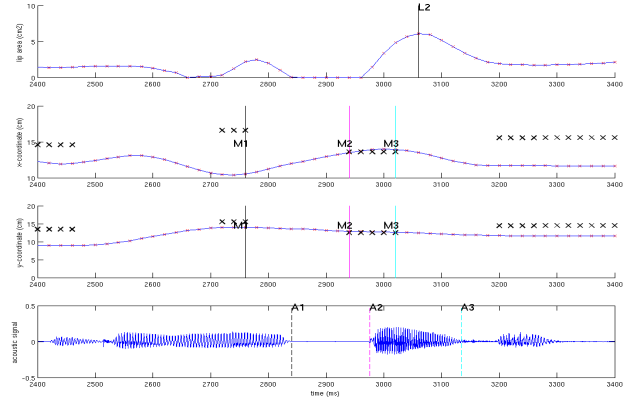


Figure 6: Zoom of signals with M1, M2, M3, A1, A2, A3 and L2 labels.

4. ANALYSIS OF THE TEMPORAL ORGANIZATION

The automatic detection of the M1, M2 and M3 labels was applied to all the sequences of the corpus. For this part, 57 CV syllables made of [p, t, k, b, m] for the consonant C and [a, i, u, y, e, ε, œ, ø, o, ɔ] for the vowels V were extracted from the sequences. The CV syllables at the beginning and at the end of the sequences, as well syllables preceded or followed by a prosodic pause were not considered, in order to avoid the specific cases due to the initialization of the hand gesture, and the relaxing of the hand. The beginning of the acoustic realization of the consonant and of the vowel and the end of the vowel were labeled A1, A2, A3 manually, respectively. In addition, the instant of vocalic realization at the lips was labeled L2. From these labels, a set of interval durations were computed: M1A1, A1M2, M1M2, M2M3, A1A3, A1A2, M2L2 and M3L2. All durations were calculated as arithmetic differences, i.e. the second label minus the first one (e.g., A1-M1 in ms for A1M1) (see Table 1).

Intervals	Mean values (ms)	Standard deviation (ms)
Syllable CV (A1A3)	284.10	74.58
Consonant (A1A2)	142.51	34.95
M1M2	215.09	64.17
M2M3	94.74	69.23
M1A1	151.79	86.97
M1M3	309.83	100.57
A1M2	63.30	71.57
M2L2	144.17	80.68
M3L2	49.43	96.65

Table 1: Average values and standard deviation of the durations for the different intervals.

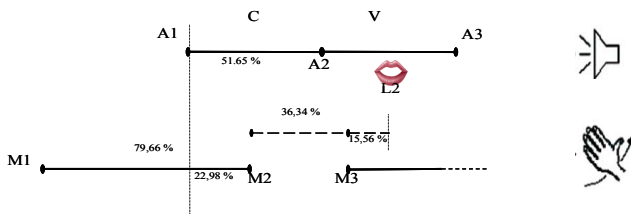


Figure 7: Temporal pattern for coordination between sound, lips and hand placement for Cued Speech production. The durations of each interval as a percentage of the A1A3 duration.

Figure 7 presents the temporal position of the intervals with their durations relative to the CV syllable duration A1A3. The first observation is the large anticipation of the hand movement, i.e. the transition towards the position starts largely before the beginning of the acoustic realization of the consonant (M1A1). The hand position is attained in the first part of the consonant (A1M2), thus largely before the vowel realization at the lips (M2L2). The hand leaves the position before the vocalic realization at the lips (M3L2 > 0). These results are compatible with the results of Attina ([7], [8]). However, a difference is noticed in A1M2 (22.98 % vs. 10 % for the same subject, but a wide variation 6 to 18 % for the other subjects). This difference should be explained by the definition of M2, based here on a threshold applied to the hand velocity. Moreover, the CV syllables are in the context of phrases in which more complex syllables (CCV, VCV...) succeed one another. This may also explain the difference observed in M2M3 (36% vs. 64%): the hand position is maintained for a shorter duration, thus explaining the difference in M3L2 (15% vs. -6%).

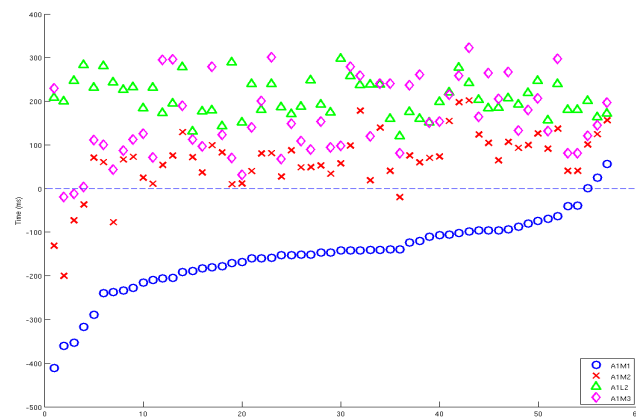


Figure 8: Temporal positions of the labels, relatively to A1 sorted by the increased values of M1A1.

The aim point of this study is the importance of the M2 instant as a main contact point for the coordination of the CS gesture M1M2 and M2M3. Indeed, in the comparison of A1M2, M2L2, M3L2, and M1A1 the variance of A1M2 is the lowest (see Table 1), thus revealing a more precise contact of the hand position onset with the beginning of the

CV syllable, in terms of control. On the other hand, the comparison between M3L2 and M2L2 (96.65 ms vs. 80.68 ms, for the standard deviations, see Table 1 and Figure 8) shows a larger variability for M3 in relation to L2, this seems to show that M3L2 is not the major control relation. Note that M1A1 has a larger variation than A1M2, which confirms our observation on M3, since M3 is not only the end of the maintained hand position but also defines the beginning of a transition towards the following hand position, thus as for M1.

5. CONCLUSION

This study confirms in a more complex context the importance of the M2 instant pointed out by Attina and colleagues. The M2 is defined at the instant at which the hand attain position and digital cue completely formed. At this instant, the consonant information contained in the CS hand is known. Thus, in the perspective of the fusion of lip and hand flows, the hand defines the instant at which the lips should be processed to specify uniquely the consonant. On the other hand, since L2 follows M3, the complete identification of the vowel needs to take into account the delay between the information given in advance by the hand (M2M3) and that delivered at the lips (L2). This latter point needs to be specifically processed in the future. Finally, the principle of the algorithm of the hand temporal segmentation will be applied for automatic lip temporal segmentation.

6. REFERENCES

- [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] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, R. Brun, "Analysis and synthesis of the 3D movements of the head, face and hand of a speech," Journal of the Acoustical Society of America, 118 (2) pp 1144-1153, 2005.
- [4] T. Burger, A. Caplier et S. Mancini, "Cued Speech Hand Gesture Recognition Tool," in proceedings of EUSIPCO'05, Antalya, Turkey, 4-8 septembre 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] V. Attina, "La Langue française Parlée Complétée (LPC): Production et Perception," Ph.D. Thesis, Institut National Polytechnique de Grenoble, Grenoble, 2005.
- [8] V. Attina, M.-A. Cathiard, D. Beautemps, (accepted) "Temporal measures of hand and speech coordination during French Cued Speech production," lecture Notes in Artificial Intelligence, LNAI/LNCS, Springer Verlag.