

Automatic identification of vowels in the Cued Speech context

Noureddine Aboutabit, Denis Beautemps, Laurent Besacier

▶ **To cite this version:**

Noureddine Aboutabit, Denis Beautemps, Laurent Besacier. Automatic identification of vowels in the Cued Speech context. AVSP 2007 - 6th International Conference on Auditory-Visual Speech Processing, Aug 2007, Hilvarenbeek, Netherlands. pp.8. hal-00266694

HAL Id: hal-00266694
https://hal.science/hal-00266694

Submitted on 25 Mar 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Automatic identification of vowels in the Cued Speech context

Noureddine Aboutabit[1], *Denis Beautemps*[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

Noureddine.aboutabit@gipsa-lab.inpg.fr

Abstract

The phonetic translation of Cued Speech (CS) (Cornett [1]) 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. [3]) between these two flows of information. The automatic coding of CS hand positions and lip targets (Aboutabit et al. [3], Aboutabit et al. [4]) are thus a key factor in the mixing process. This contribution focuses on the identification of vowels by merging CS hand positions and vocalic lip information produced by a CS speaker. The hand flow is coded automatically as plateaus between transition phases. A plateau is defined as the interval during which the hand is maintained at a specific CS hand position. A transition is the interval during which the hand moves from a specific CS hand position to another one. The CS hand position is automatically obtained as the result of the hand 2d-coordinates Gaussian classification. The instants of reached hand targets are used as reference instants to define the interval inside which the lip target instant of the vowel is automatically detected. The lip parameters extracted at this instant are processed in a Gaussian classifier as to identify the vocalic lip feature of the vowel. The vowel is obtained as the result of the combination of the corresponding hand position and the lip feature. The global performance of the method attains 77.6% as correct identification score. This result does not take into account the CS coding errors. This result has to be compared with the global 83.5% score of speech perception by deaf people using CS (Nichols and Ling, 1982 [6].

Index Terms: Cued Speech production, lip target segmentation, vocalic lip classification, and CS gesture segmentation.

1. Introduction

The Cued Speech (CS) (Cornett, 1967 [1]) is a manual cues system used to disambiguate the lip-reading and enhance speech perception from visual input by deaf and impaired-hearing people. In this system, the speaker moves the hand in close relation with speech (see Attina et al., 2004 [2] for a detailed 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 which describes 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 Aboutabit et al., 2006 [3] for an example of a complete analysis of the hand flow).

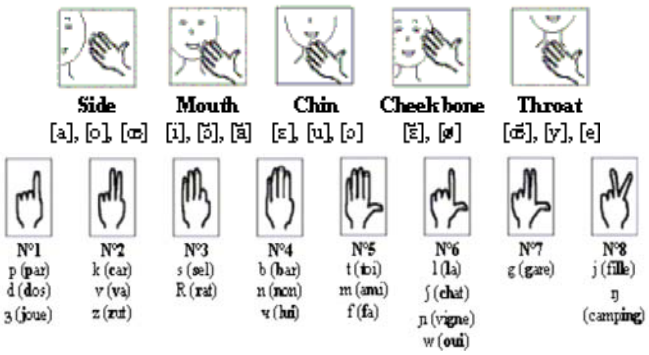


Figure 1: *CS Hand position (top) for vowels and CS handshapes (bottom) for consonants (adapted from [2]).*

This paper focuses on the automatic recognition of French vowels by combining the two flows of information (hand and lip). The first part presents a method to automatically segment the hand flow. The objective of this segmentation is to detect the CS hand position and also to identify wether the hand is maintained in a position or is in transition between two positions. The second part develops an automatic method of lip target segmentation applied to vowels. The last part discusses how both flows of information are merged to recognize vowels in a sequence. Finally, an experimental evaluation is presented to measure the performance of the merging system as well as the performance of the segmentation method for each separate modality.

2. Data

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



Figure 2: *Image of the speaker.*

The French CS speaker is a native female speaker of French, certified in French CS. She regularly translates into French CS code in a school. The recording was made in a sound-proof booth at the Parole & Cognition department of Grenoble Images Parole Signal Automatique laboratory, (GIPSA-lab, department P&C), 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. The camera in large focus was used for the hand and the face and was connected to a betacam recorder. The lips were painted in blue, and blue marks were placed on the speaker's glasses as reference points. 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).

A square paper was recorded 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. A specific image processing was applied to the Bitmap frames in the lip region to extract the inner and outer contours and to derive the corresponding characteristic parameters (Lallouache, 1991 [5]): lip width (A), lip aperture (B) and lip area (S). These parameters were converted using a pixel-to-centimeter conversion formula. Finally the parameters were low-pass filtered. 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.

The acoustic signal was automatically labeled at the phonetic level using forced alignment (see Lamy, 2004 [7] 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, 2000 [8]).

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 lip parameter values every 20 ms and the corresponding acoustic signal (figure 4).

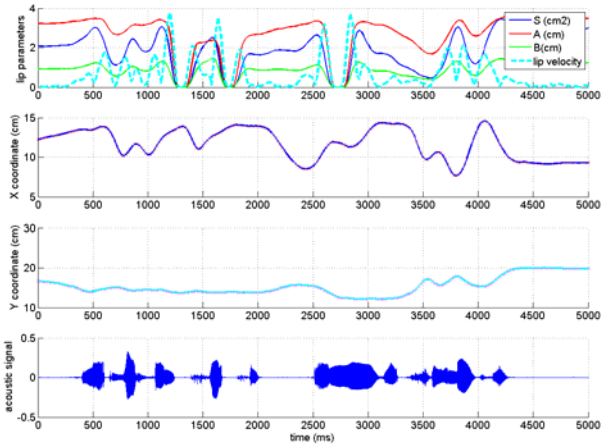


Figure 4: *Example of signals. From top to bottom: Inner lip parameters (A, B, S and lip velocity calculated from S), x and y coordinates of the reference hand landmark, the acoustic realization.*

3. Automatic 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 computed from a Gaussian model of the x-y coordinates that corresponds to 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 modelized 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 one 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 coordinate, 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 recognized hand position is to the one that gives the maximum amongst the 5 components of the final vector.

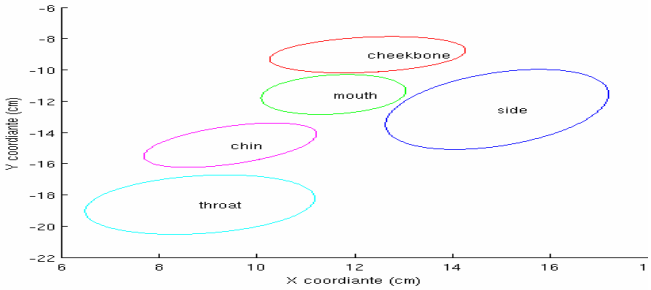


Figure 4: *2 standard deviation ellipsis around the corresponding average values of the reference hand landmark extracted from the learning data, for the 5 hand positions.*

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

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 20ms. 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 the 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.

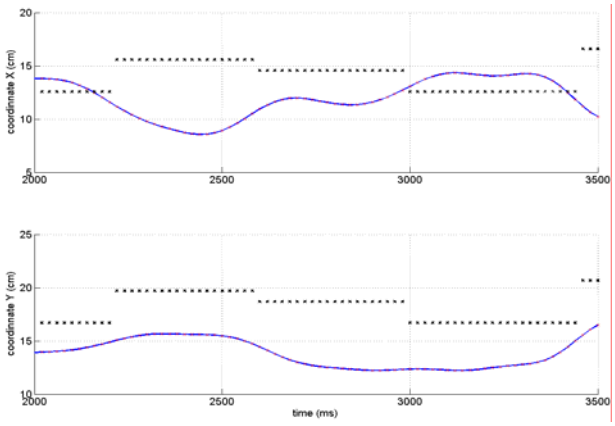


Figure 5: *Zoom of signals with hand positions plateaus delivered by the classifier.*

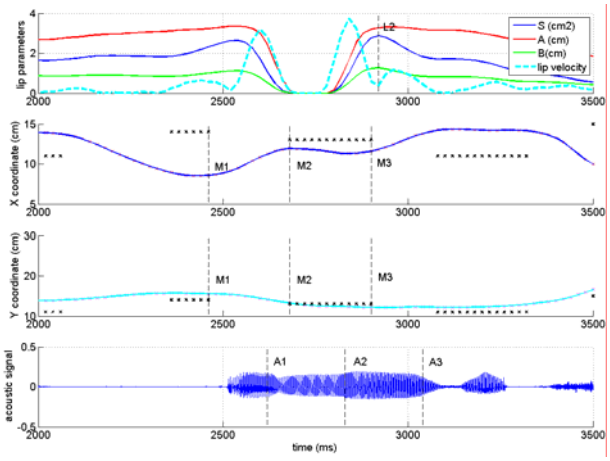


Figure 6: *Zoom of signals with M1, M2, M3, A1, A2, A3 and L2 labels.*

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. For the acoustic signal, the beginning of the acoustic realization of the consonant and of the vowel and the end of the vowel were labeled A1, A2 and A3.

4. Automatic lip target segmentation

The lips were characterized at the instant the lip target was attained. The automatic definition of this instant, labeled L2 (see figure 6 for an example), 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. In the case of important anticipation the research process is limited by the end frontier of the phone acoustic realization. Lip velocity (see Figure 3) 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 = 0.99).

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.

From the L2 instants obtained with this method, it has been shown that vowels could be grouped into three categories in conformity with the phonetic description of the vowels (anterior non rounded vowels [a, ɛ̃, i, œ̃, e, ɛ], high and mid-high rounded [ɔ̃, y, o, ø, u] low and mid-low rounded vowels [ã , ɔ, œ]) (see Aboutabit et al., 2006 [4] for more details).

In this study, it has been demonstrated that when the CS hand position was given without error, high scores of vowel identification are obtained (89% as average recognition rate of vowel) with only one measure instant, defined by L2 instant.

5. Vowel recognition

5.1. Method

The complete vowel recognition needs to merge both manual and lip informations obtained from the two previous automatic processing. Several merging approaches can be considered. Among these methods, the classical models of audio-visual integration (Schwartz et al., 1998 [9]) are interesting but a few ones can be adapted to the Cued Speech merging case, while others do not. For example, the direct identification model (DI) seems to be not appropriate to the CS gestures fusion. One raison is that this kind of identification needs a system that is capable to merge quantitative information provided from lips (lip parameter values) and qualitative information provided from CS hand gestures (hand position and configuration). Even if a transformation in quantitative components is possible from the CS hand information, the fact to take into account components with different origin in a same vector poses the problem of their weighting. As second raison, the temporal desynchronization between lip and hand information is a serious problem for this model.

Alternatively, the separated identification model (SI) seems to be convenient. Considering this model, on one hand, at the M2 instant the CS hand position is known and defines a first group of vowels, composed by two or three vowels. On the other hand, at the corresponding L2 instant a second group of vowels (viseme) is derived from a classification of lip parameters. Then, the vowel recognition results from the intersection between these two groups of vowels. In this case, it is possible that the intersection may be empty in the case of a determinist fusion. Thus, no vowel is recognized. This non identification problem may be caused by an error on the lip decision and/or on the hand position decision. To solve this problem, instead to use one classifier of viseme in the lip process, five classifiers, one for each CS hand position could be considered. Then, as result from the lip process, five vowels are selected (i.e. one vowel for each classifier).

To reduce the number of lip classifiers from five to a single one, a fusion model derived from the SI one is considered in the following. It consists to constrain the lip decision by the hand position decision considering the advance of the M2 instant over the L2 instant in the case of vowels (Aboutabit et al., 2006 [3]). Then, between two successive M2 instants, a L2 instant is located excepted in the case of a consonant alone. At the first M2 instant the CS hand position allows to identify a first group of vowels. A simple Gaussian classification on lip parameters extracted at L2 recognizes one element among this group. Figure 7 illustrates the method.

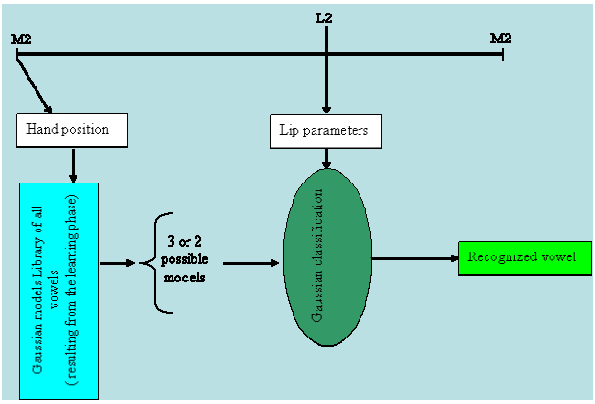


Figure 7*: Vowel identification: fusion schema of both lip and hand information.*

5.2. Results and discussion

To test this merging method, the test corpus contains 774 vowels from 120 coded and pronounced sentences. The global score of vowel recognition is 77.6%. If the CS hand position is known without error, the vowel recognition is 89% (Aboutabit et al. [4]). Then, the difference between these two scores is due to the automatic decision on the CS hand position and/or to the matching of M2 and L2.

The recognition score of 77.6% is slightly lower than the CS perceptual effectiveness score of 83.5% obtained by Nicholls and Ling (1982, [6]) in their study on the reception of CV and VC syllables with hearing-impaired children.

In the evaluation of the corpus coded by the speaker, a decoding test in reception was performed by a profoundly hearing-impaired subject practicing regularly the CS system. The test consisted in the decoding of the whole sentences of the corpus. For the 120 sentences used previously, the score of correct decoding attains 94.8% for the vowels. This latter is the reference to compare the performance of the recognition method (77.6%).

6. Conclusions and perspectives

The merging model, in which the hand is considered at the M2 instant in advance to the L2 instant of the lips, was validated. The recognition score for vowels (77.6%), compared to the different reference scores previously recalled, is promising to go further. Indeed, an improvement of the automatic segmentation of the CS hand position should improve the decision on the CS hand position and enhance the precision of the M2 instant. This allows to reduce the M2-L2 matching errors due to the imprecision on M2. In addition, concerning the lips, a better selection of the velocity local minima, should increase the recognition performances.

As a perspective, the merging model could be extended to the consonant in vocalic context. In this case, the complete [M2, L2] interval should be considered in order to take into account the complex effect of the co articulation.

7. Acknowledgements

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).

8. References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] Nicholls, G., Ling, D. "Cued Speech and the reception of spoken language", Journal of Speech and Hearing Research.25, 262-269, 1982.

[7] 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.

[8] 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.

[9] Schwartz, J. L., Robert-Ribes, J. & Escudier, P. Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. Hearing by Eye II, Advances in the psychology of speechreading and auditory visual speech. Psychology Press, pp. 85-108, Hove (UK), 1998.