



Hand shape Coding for HMM-based Consonant Recognition in Cued Speech for French

Nouredidine Aboutabit, Panikos Heracleous, Denis Beautemps

► To cite this version:

Nouredidine Aboutabit, Panikos Heracleous, Denis Beautemps. Hand shape Coding for HMM-based Consonant Recognition in Cued Speech for French. SPECOM 2009 - 13th International Conference on Speech and Computer (SPECOM2009), Jun 2009, St. Petersburg, France. pp.1-4. hal-00372704

HAL Id: hal-00372704

<https://hal.science/hal-00372704>

Submitted on 2 Apr 2009

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.

Hand shape Coding for HMM-based Consonant Recognition in Cued Speech for French

Noureddine Aboutabit, Panikos Heracleous, and Denis Beautemps

GIPSA-lab, Speech and Cognition Department, CNRS UMR 5216 / Stendhal University/ UJF/ INPG
961 rue de la Houille Blanche Domaine universitaire BP 46 F - 38402 Saint Martin d'Hères cedex
FRANCE

E-mail: {Noureddine.Aboutabit, Panikos.Heracleous, Denis.Beautemps}@gipsa-lab.inpg.fr

Abstract

Cued Speech (CS) is a visual communication mode that makes use of hand shapes placed in different positions near the face in combination with the natural speech lipreading, to enhance speech perception from visual input. This system is based on the motions of the speaker's hand moving in close relation with speech. In a CS system, hand shapes are designed to distinguish among consonants and hand placements are used to distinguish among vowels. Due to the CS system, both manual and lip flows produced by the CS speaker carry a part of the phonetic information. This contribution presents automatic hand shape coding of a CS video recording with 92% obtained accuracy, and multi-stream hidden Markov models (HMMs) fusion to integrate hand shape and lip shape elements into a combined component and perform automatic recognition of CS for French. Compared with using lip shape modality alone, by applying fusion the accuracy of CS consonant recognition was raised from 52.1% to 79.6%.

1. Introduction

The benefit of using visual information in speech perception, or recognition is well known (i.e. lipreading). However, even with high lipreading performances speech cannot be thoroughly perceived without knowledge about the semantic context. To date, the best lipreaders are far way of reaching perfection. On average, only 40-60% of vowels of a given language (American English) are recognized by lipreading [1], and 32% when related to low predicted words [2]. The main reason is the ambiguity of the visual pattern. On the other hand, for orally educated deaf people lipreading remains the main modality of perceiving speech. To overcome the problems of lipreading, Cornett developed in 1967 the Cued Speech system as a complement to lip information [3].

CS is a visual communication mode that uses hand shapes placed in different positions near the face in combination with natural speech lipreading to enhance speech perception from visual input. CS is a system where the speaker faces the perceiver, and moves his hand in close relation with speech [4]. The hand -held flat and oriented so that the back of the hand faces the perceiver- is a cue that corresponds to a unique phoneme when associated with a particular lip shape. A manual cue consists of two components: the hand shape and the hand placement relative to the face. Hand shapes are designed to distinguish consonant phonemes and hand placements are used to














 <p>Côté a (ma) o (maux) œ (teuf) (*)</p>	 <p>Pommette ɛ (main) ø (feu)</p>	 <p>Bouche i (mi) ɔ̃ (on) ã (rang)</p>	 <p>Menton ɛ (mais) u (mou) ɔ̃ (fort)</p>	 <p>Gorge œ (un) y (tu) e (fée)</p>
 <p>Configuration 1 p (par) d (dos) ɔ̃ (joue)</p>	 <p>Configuration 2 k (car) v (va) z (zut)</p>	 <p>Configuration 3 s (sel) R (rat)</p>	 <p>Configuration 4 b (bar) n (non) ɥ (lui)</p>	
 <p>Configuration 5 t (toi) m (ami) f (fa) (*)</p>	 <p>Configuration 6 l (la) ʃ (chat) ʒ (vigne) w (oui)</p>	 <p>Configuration 7 g (gare)</p>	 <p>Configuration 8 j (fille) ɥ (camping)</p>	

Figure 1: Hand position for vowels (top) and hand shapes for consonants (bottom) in French Cued Speech

distinguish vowel phonemes. 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. CS improves speech perception for deaf people [2, 5]. Moreover, for deaf people, who have been exposed to this method since their youth offers them a complete representation of the phonological system, and therefore it has a positive impact on language development [6].

The access to communication technologies has become essential for the handicapped people. This study is a part of the TELMA project (Phone for deaf people), that aims to develop an automatic translation system of acoustic speech into visual speech completed with CS and vice versa, i.e. from CS components into auditory speech [7]. This project would enable deaf users to communicate between themselves and with normal-hearing people through the help of the autonomous terminal TELMA. In this context, the automatic translation of CS components into a phonetic chain is a key issue. The CS system allows both hand and lip flows produced by the CS speaker to carry a part of phonetic information. Thus, in order to recover the complete phonetic and lexical information, lip and hand elements should be used jointly.

In the first attempt for vowel recognition in CS for French, in [8] a method based on separate identification, i.e., indirect decision fusion has been used by the authors, and 77.6% vowel ac-

This work is supported by the French TELMA project (RNTS / ANR).

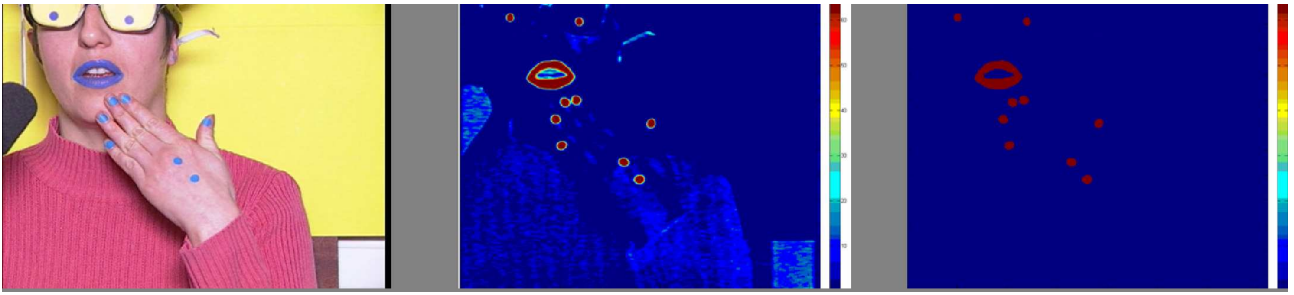


Figure 2: The three-step algorithm applied for lip shape and gesture detection based on detection of blue objects

curacy was obtained. Also the authors have presented automatic vowel recognition in CS for French based on HMMs [9]. In the current study, the method was extended to deal also with consonant recognition in CS for French. The objective is to present an innovative modeling of the CS elements (i.e. hand and lips) for automatic consonant recognition. In this article, methods for automatic coding of the CS hand shape component, and identification of fingers from each video frame are introduced. The two methods estimate the xy coordinates of the extremity of each finger of the cuer, which serve as features in hand shape modeling. Based on multi-stream HMM decision fusion [10], CS hand shape element and lip shape element were integrated into a combined one, and automatic consonant recognition in CS for French was realized.

2. Methods

A review of the literature written about hand processing shows that the hand segmentation issue -in the context of CS- has originally been investigated by researchers who have applied technologies to CS, mainly in the field of speech synthesis [11]. On the other hand, the problem of CS hand segmentation for gesture recognition has been far less studied. In this study, hand movements are derived automatically from the image processing of the cuer's video.

The French CS cuer was a female French native speaker certified in French CS. She regularly interprets French CS code in schools. The recording has been made in a sound-proof booth, and the image video recording has been set on 50 frames/second. Because of the CS system's nature (i.e., the cuer faces the perceiver), lips are characterized with the front view of the face, and, therefore, artifices have been used to mark directly the relevant information. Blue marks were placed on the speaker's glasses as reference points. Blue marks were also placed on the left hand's back and at the extremity of the fingers to follow their movements and the CS hand shape formation. These constraints were applied in recordings in order to control the data and facilitate the extraction of accurate features (see [12] for details).

The data have been derived from a video recording of the speaker pronouncing and coding in French CS a set of 124 sentences. The sentences -composed of low predicted multi-syllabic words- have been derived from the [11] corpus, which was dedicated to CS synthesis. The audio part of the video recording has been digitized at 22,050 kHz in synchrony with the image part.

The image processing for lip contour and hand features extraction has been done in two main steps:

- Detection of the blue objects on the image

In this step, the blue objects on each image were detected using a two-step algorithm. Firstly, the gray level image was subtracted from the blue component of the RGB image. Then, a threshold was applied to the resulted image to obtain a bi-chromatic image (i.e., each pixel having a value higher than this threshold was addressed by value 255, otherwise was addressed by 0). Fig. 2 shows the image processing steps in this stage.

- Extraction of the lip contours

In the second step, the lip contour and the blue landmarks on the back of the hand, and also at the extremity of fingers, were marked and extracted using a coloring blob algorithm applied to the bi-chromatic image. This kind of algorithm detects all regions on the image as connected components, and for each one attributes a number.

The image processing method described here was applied to the video frames on the lip region to extract the inner- and outer contours, and to derive the corresponding characteristic parameters: lip width, lip aperture and lip area (i.e., six parameters in total). In addition, two supplementary parameters relative to the lip morphology were extracted: the pinching of the upper and lower lips. As a result, eight parameters were extracted for modeling lip shapes. For hand shape modeling, the xy coordinates of the landmarks placed on the extremity of the fingers were used (i.e., 10 parameters).

In the automatic consonant recognition experiments for CS, 17 context-independent, 3-state, left-to-right with no skip HMMs were used. In addition to the basic lip and hand parameters, the first (Δ) and second derivatives ($\Delta\Delta$) were also used. For training, 4401 consonants, and for test 2155 consonants were used, respectively. Consonant instances were extracted automatically after forced alignment was performed (i.e., using the acoustic signal).

In addition to the visual models, acoustic HMMs were also trained using the audio signal. The parameters used were 12 Mel-Frequency Cepstral Coefficients (MFCC) along with the first and second derivatives (i.e., 36 parameters in total).

3. Results

3.1. Hand shape coding

In French Cued Speech, the recognition of the eight hand shapes is an exceptional case of the hand shape recognition. In fact, a causal analysis based on some knowledge -such as the number and dispersion of fingers, and also the angle between them- can distinguish between those eight hand shapes. Based on the number of landmarks detected on fingers, the correct hand shape can

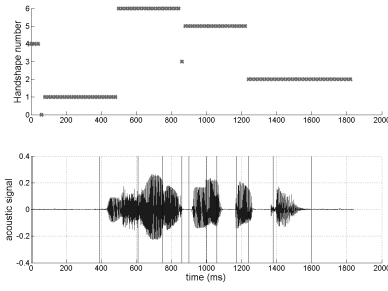


Figure 3: Hand shape plateaus delivered by the automatic system according to the acoustic signal

be recognized. Fig. 1 shows the hand shapes numbered from left to right (i.e., S1–S8). The proposed algorithm to identify the Cued Speech hand shapes is as follows:

- Total number of fingers on which landmark was detected: 1, then hand shape S1.
- Total number of fingers on which landmark was detected: 4, then hand shape S4.
- Total number of fingers on which landmark was detected: 5, then hand shape S5.
- Total number of fingers on which landmark was detected: 3, then hand shapes S3 or S7. If the thumb finger is detected (using finger dispersion models) then hand shape S7, else hand shape S3.
- Total number of fingers on which landmark was detected: 2, then hand shapes S2 or S6 or S8. If the thumb finger is detected then hand shape S6, else the angle between the two finger landmarks according to the landmarks on the hand can identify if it is hand shape S2 or S8 (using a threshold).
- In any other case hand shape S0, i.e., no Cued Speech hand shape was detected.

The previous method was applied to each sequence of the corpus. Figure 3 shows an example of a recognized hand shape sequence.

To evaluate the hand shape recognition system, a set of 1009 frame were used and recognized automatically. Table 1 shows the confusion matrix of the recognized hand shapes by the automatic system. As can be seen, the system recognized correctly 92% of the hand shapes (i.e., percentage of diagonal elements). This score justify the choice of the authors, and shows that using only 5 landmarks placed at the finger extremities, the accuracy did not decrease drastically compared with the 98.8% of recognized hand shapes obtained by [11] system based on 50 tags. The errors occurred can be attributed to landmark detection processing. However, in some cases one or more landmarks are not detected due to the rotation of the hand, and in some other cases, landmarks remain visible even when the fingers are bended.

3.2. Finger identification

The objective of this section was to identify fingers in each frame of the recording in order to assign the extracted parameters to the correct fingers. The identification has been done in three steps. In the first step, all landmarks in the frame were



Figure 4: French Cued Speech cueer and landmarks used for hand shape coding (left) and finger identification (right)

detected. The landmarks placed on the speaker glasses and on the back of the hand were benched to have only the landmarks corresponding to the fingers. Secondly, the coordinates of these landmarks were projected on the hand axis defined by the two landmarks on the back of the hand. The third step consists of sorting resulted coordinates following the perpendicular axis to the hand direction from the smaller to the largest (Fig. 4). In this step, the hand shape coding was used to associate each coordinate to the corresponding finger. For example, when there are three landmarks coordinates and the hand shape number is three, the smaller coordinate is associated to the middle finger, the second to the ring finger and the biggest one to the baby finger.

3.3. Consonant recognition based on multi-stream HMM decision fusion

In this experiment, lip shape and hand shape elements were integrated into a single component using multi-stream HMM decision fusion, and recognition experiments were conducted to recognize the 17 French consonants. Decision fusion captures the reliability of each stream, by combining the likelihoods of single-stream HMM classifiers. Such an approach has been used in multi-band audio only ASR [13] and in audio-visual speech recognition [10]. The emission likelihood of multi-stream HMM is the product of emission likelihoods of single-stream components weighted appropriately by stream weights. Given the O combined observation vector, i.e., lip shape and hand shape elements, the emission probability of multi-stream HMM is given by

$$b_j(O_t) = \prod_{s=1}^S [\sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm})]^{\lambda_s} \quad (1)$$

where $N(O; \mu, \Sigma)$ is the value in O of a multivariate Gaussian with mean μ and covariance matrix Σ . For each stream s , M_s Gaussians in a mixture are used, with each weighted with c_{jsm} .

Table 1: Confusion matrix of hand shape recognition evaluation

	S0	S1	S2	S3	S4	S5	S6	S7	S8	
S0	33	2	0	0	0	0	0	0	0	94
S1	16	151	0	0	0	0	1	0	5	87
S2	1	2	93	0	0	0	0	0	6	91
S3	0	0	0	163	2	0	0	3	9	91
S4	3	0	0	0	100	0	0	3	0	94
S5	2	0	0	4	4	193	0	0	1	95
S6	0	0	0	0	0	0	124	5	0	96
S7	0	0	0	0	0	0	0	17	0	100
S8	1	05	0	2	0	0	0	0	58	95

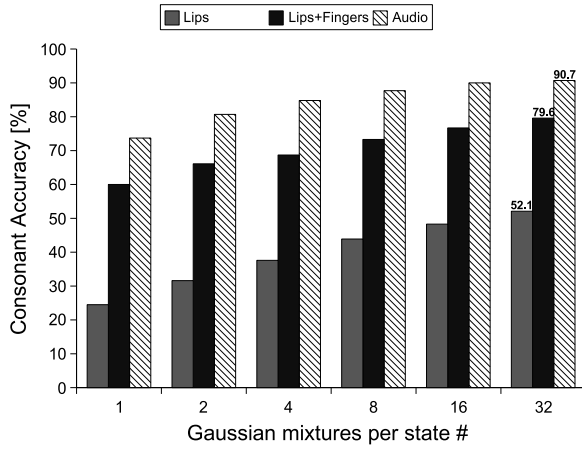


Figure 5: Accuracy of consonant recognition based on multi-stream HMM decision fusion

The contribution of each stream is weighted by λ_s . In this study, we assume that the stream weights do not depend on state j and time t . However, two constraints were applied. Namely,

$$0 \leq \lambda_h, \lambda_l \leq 1 \quad (2)$$

and

$$\lambda_h + \lambda_l = 1 \quad (3)$$

where λ_h is the hand shape stream weight, and λ_l is the lip shape stream weight. The HMMs were trained using maximum likelihood estimation based on the Expectation-Maximization (EM) algorithm. However, the weights cannot be obtained by maximum likelihood estimation. In these experiments, the weights were adjusted experimentally to 0.6 and 0.4 values, respectively. The selected weights were obtained by maximizing the accuracy on several experiments.

Fig. 5 shows the results obtained in the function of several Gaussian mixtures per state. As it can be seen, using fusion of lip shape and hand shape elements, significant improvement in accuracy was obtained compared with using lip shape only. In the case of 32 Gaussian mixtures per state, 57% relative improvement was achieved. The result was very promising and showed the effectiveness of the hand shape coding and fusion method of the two elements. Fig. 5 shows also the results obtained when using the acoustic signal. The errors in hand shape recognition may be one of the reasons of the lower accuracy of CS recognition, compared with the audio recognition accuracy. However, the results are still comparable, showing high consonant recognition accuracy using visual information alone.

4. Conclusion

In this paper, automatic hand shape recognition and finger identification in the context of CS for French was presented. Using only five landmarks placed on the cuer's fingers, and extracted automatically from the video recording, a 92% hands hape recognition accuracy was obtained. Hand shape and lip shape modalities were integrated using multi-stream HMM decision fusion, extending our previously presented studies done in vowel recognition in CS, also to automatic consonant recognition. When fusion was applied, the obtained consonant accu-

racy was raised from 52.1% to 79.6%, showing a 57% relative improvement compared with using lip shape parameters only.

5. References

- [1] A. A. Montgomery and P. L. Jackson, "Physical characteristics of the lips underlying vowel lipreading performance," *Journal of the Acoustical Society of America*, vol. 73 (6), pp. 2134–2144, 1983.
- [2] G. Nicholls and D. Ling, "Cued speech and the reception of spoken language," *Journal of Speech and Hearing Research*, vol. 25, pp. 262–269, 1982.
- [3] R. O. Cornett, "Cued speech," *American Annals of the Deaf*, vol. 112, pp. 3–13, 1967.
- [4] 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*, vol. 44, pp. 197–214, 2004.
- [5] R. M. Uchanski, L. A. Delhorne, A. K. Dix, L. D Braid, C. M. Reedand, and N. I. Durlach, "Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech," *Journal of Rehabilitation Research and Development*, vol. 31(1), pp. 20–41, 1994.
- [6] J. Leybaert, "Phonology acquired through the eyes and spelling in deaf children," *Journal of Experimental Child Psychology*, vol. 75, pp. 291–318, 2000.
- [7] D. Beautemps, L. Girin, N. Aboutabit, G. Bailly, L. Besacier, G. Breton, T. Burger, A. Caplier, M. A. Cathiard, D. Chene, J. Clarke, F. Elisei, O. Govokhina, V. B. Le, M. Marthouret, S. Mancini, Y. Mathieu, P. Perret, B. Rivet, P. Sacher, C. Savariaux, S. Schmerber, J. F. Serignat, M. Tribout, and S. Vidal, "Telma: Telephony for the hearing-impaired people. from models to user tests," in *Proceedings of ASSISTH'2007*, pp. 201–208, 2007.
- [8] N. Aboutabit, D. Beautemps, and L. Besacier, "Automatic identification of vowels in the cued speech context," in *Proceedings of AVSP'07*, 2007.
- [9] P. Heracleous, N. Aboutabit, and D. Beautemps, "Lip shape and hand position fusion for automatic vowel recognition in cued speech for french," *IEEE Signal Processing Letters*, 2009 (in Press).
- [10] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," in *Proceedings of the IEEE*, vol. 91, Issue 9, pp. 1306–1326, 2003.
- [11] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, , and R. Brun, "Analysis and synthesis of the 3d movements of the head, face and hand of a speaker using cued speech," *Journal of Acoustical Society of America*, vol. 118(2), pp. 1144–1153, 2005.
- [12] N. Aboutabit, D. Beautemps, and L. Besacier, "Lips and hand modeling for recognition of the cued speech gestures: The french vowel case," *Speech Communication*, 2009, (to appear).
- [13] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," in *Proceedings of ICSLP*, pp. 426–429, 1996.