HMM-based vowel and consonant automatic recognition in Cued Speech for French

— **Source link** ↗

Panikos Heracleous, Noureddine Aboutabit, Denis Beautemps

Institutions: Stendhal University

Related papers:

- Feature adaptation of hearing-impaired lip shapes: the vowel case in the Cued Speech context

- Temporal measures of hand and speech coordination during french cued speech production

- The role of rhythmic groups in the segmentation of continuous French speech

- Rapid adaptation to foreign-accented English

- Speaker-Machine Interaction in a Limited Speech Recognition System

HMM-based Vowel and Consonant Automatic Recognition in Cued Speech for French

Panikos Heracleous, Noureddine Aboutabit, Denis Beautemps

HMM-based Vowel and Consonant Automatic Recognition in Cued Speech for French

Panikos Heracleous, Noureddine Aboutabit, 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

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

Abstract—In this paper, hidden Markov models (HMM)-based vowel and consonant automatic recognition in French Cued Speech is presented. Cued Speech is a visual communication mode which uses handshapes in different positions and in combination with lip-patterns of speech, makes all the sounds of spoken language clearly understandable to deaf and hearing-impaired people. The aim of Cued Speech is to overcome the problems of lipreading and thus enable deaf children and adults to fully understand a spoken language. Previously, the authors have reported experimental results on vowel recognition in Cued Speech for French based on feature fusion and multi-stream HMM decision fusion. This study, further investigates the vowel recognition by considering vowel classes based on similarities on the lips. Also automatic consonant recognition experiments in Cued Speech for French are reported. The obtained results are promising and comparable with results obtained using audio signal.

I. INTRODUCTION

To date, visual information is widely used to improve speech perception, or automatic speech recognition (lipreading). With lipreading technique, speech can be understood by interpreting movements of lips, face and tongue. In spoken languages, a particular facial and lip shape corresponds to each sound (phoneme). However, this relationship is not one-to-one and many phonemes share the same facial and lip shape (visemes). It is impossible, therefore to distinguish phonemes using visual information alone.

Even with high lipreading performance, speech without knowledge of the semantic context can not be completely perceived. On average, only 40-60% of the vowels are recognized by lipreading system for a given language American English) [1], and only 10-30% of the words [2]. To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, in 1967 Cornett [3] developed the Cued Speech system to complement the lip information and make all phonemes of a spoken language clearly visible. As many sounds look identical on lips (e.g., /p/ and /b/), using hand information those sounds can be distinguished and thus make possible for deaf people to completely understand a spoken language using visual information alone.

Cued Speech uses handshapes placed in different positions near the face in combination with natural speech lipreading to enhance speech perception from visual input. A manual cue in this system contains two components: the handshape and the hand position relative to the face. Handshapes distinguish
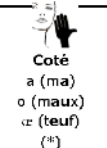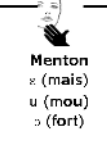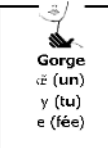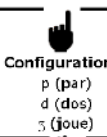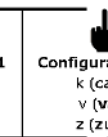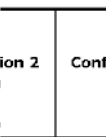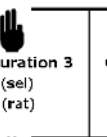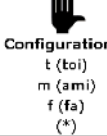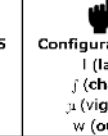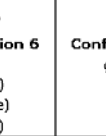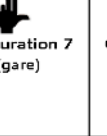

Fig. 1. Hands positions for vowels (top) and handshapes (bottom) for consonants in Cued French language.

consonant phonemes whereas hand positions distinguish vowel phonemes. A handshape together with a hand position cue a syllable. Cued Speech recognition requires gesture recognition and lip shape recognition, and also integration of the two elements. The advantage of Cued Speech is that improves speech perception to a large extent for hearing-impaired people [4]. Moreover, Cued Speech offers a complete representation of the phonological system for hearing-impaired people exposed to this method since their youth, and therefore has a positive impact on language development [5]. Fig. 1 describes the complete system for French. In French Cued Speech, eight handshapes in five positions are used. The system was adapted from American English to French in 1977.

The access to communication technologies has become essential for the handicapped people. The TELMA project (Phone for deaf people) aims to develop an automatic translation system of acoustic speech into visual speech completed with Cued Speech and vice versa, i.e. from Cued Speech components into auditory speech [6]. This project would enable deaf users to communicate with each other and with normal-hearing people through the help of the autonomous terminal TELMA.

In the first attempt for vowel recognition in Cued Speech

Set	French vowels														Total
	/ø/	/y/	/o/	/ɔ̃/	/u/	/a/	/ɛ̃/	/i/	/œ̃/	/e/	/ɛ/	/ɑ̃/	/œ/	/ɔ/	
Training	400	293	221	148	197	691	121	500	132	403	265	245	84	138	3838
Test	208	134	103	72	91	347	59	242	66	208	150	126	38	69	1913
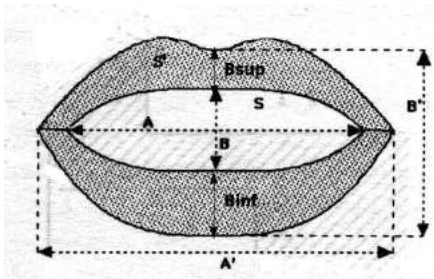
Fig. 2.   Parameters used for lip shape modeling.

Consonants		Vowels	
Viseme	Phonemes	Viseme	Phonemes
C1	/p/, /b/, /m/	V1	/ɔ̃/, /y/, /o/ /ø/, /u/
C2	/f/, /v/	V2	/a/, /ɛ̃/. /i/ /œ̃/, /e/, / ɛ/
C3	/t/, /d/, /s/ /z/, /n/, /ɲ/	V3	/ɑ̃/, /ɔ/, /œ/
C4	/ʃ/, /ʒ/		
C5	/k/ , /g/ /ʀ/, /l/		

[7], a method based on separate identification, i.e., indirect decision fusion was used and 75% vowel accuracy was obtained. In this study, however, the proposed method is based on HMMs and it uses feature fusion to integrate lip shape and hand elements into a combined component, and then perform automatic recognition. Previously, the authors reported automatic vowel recognition in Cued Speech for French using feature fusion and multi-stream HMM decision fusion with very promising results [8]. In this study, further investigations on vowel recognition using lip similarities are presented. In addition, the proposed method was extended to deal also with consonant recognition in Cued Speech for French. As far as our knowledge goes, automatic vowel and consonant recognition in Cued Speech based on HMMs is being introduced for the first time ever. Based on a review of the literature written about Cued Speech, the authors of this study have not come across any published work related to automatic vowel or consonant recognition in Cued Speech for any other Cued language.

II. METHODS

The female native French speaker-employed for data recording was certified in transliteration speech into Cued Speech in the French language. She regularly cues in schools. The cuer wore a helmet to keep her head in a fixed position and opaque glasses to protect her eyes against glare from the halogen floodlight. A camera with a zoom facility used to shoot the hand and face was connected to a betacam recorder. The speakers lips were painted blue, and blue marks were marked on her glasses as reference points. These constraints were applied in recordings in order to control the data and facilitate the extraction of accurate features (see [7], [9] for details).

The data were derived from a video recording of the speaker pronouncing and coding in Cued Speech a set of 262 French sentences. The sentences (composed of low predicted multi-syllabic words) were derived from a corpus that was dedicated to Cued Speech synthesis. Each sentence was dictated by an experimenter, and was repeated two- or three times (to correct errors in pronunciation of words) by the cuer. Table I shows the number of vowel instances included in the training and test sets, respectively.

The audio part of the video recording was synchronized with the image. Using forced alignment, the acoustic signal was automatically labelled at the phonetic level. An automatic image processing method was applied to the video frames in the lip region to extract their inner- and outer contours and to derive the corresponding characteristic parameters: lip width (A), lip aperture (B), and lip area (S) (i.e., six parameters in all).

The process described here resulted in a set of temporally coherent signals: the 2-D hand information, the lip width (A), the lip aperture (B), and the lip area (S) values for both inner- and outer contours, and the corresponding acoustic signal with the associated phonetically labelled transcriptions. In addition, two supplementary parameters relative to the lip morphology were extracted: the pinching of the upper lip (Bsup) and lower (Binf) lip. As a result, a set of eight parameters in all was extracted for modelling lip shapes. For hand position modelling, the coordinates of two landmarks placed on the hand were used (i.e., 4 parameters). Fig. 2 shows the lip shape parameters that were used.

The Cued Speech paradigm requires accurate recognition of both lip shape and hand information. Fusion of lip shape and hand elements is also necessary and very important. Fusion is the integration of available single modality streams to a combined one. In this work, lip shape, hand position, and handshape streams are available. For vowel recognition, lip shape and hand position elements were fused. For consonant recognition, lip shape and handshape elements were fused.

Previously, several studies have been made in automatic audio-visual recognition and integration of visual and audio

modalities [10], [11]. The aim of audio-visual speech recognition is to improve the performance of a recognizer, especially under noisy environments.

In the experiments, context-independent models were used. A 3-state, left-to-right no skip HMM topology was used. Each state was modeled with 32 Gaussian mixtures. In addition to the basic lip and hand parameters, the first (Δ) and second derivatives ($\Delta\Delta$) were also used. For training and test 426 and 212 sentences were used, respectively. The training utterances contained 3838 vowels and 4401 consonants. The test utterances contained 1913 vowels and 2155 consonants, respectively. Vowels and consonants were automatically extracted after forced alignment -using the audio signal- was performed.

In automatic speech recognition, a diagonal covariance matrix is often used because of the assumption that the parameters are uncorrelated. In lipreading, however parameters show a strong correlation. In this study, Principal Component Analysis (PCA) was applied to decorrelate the lip shape parameters and then a diagonal covariance matrix was used. All 24 PCA lip shape components were used for HMM training. For training and recognition the HTK3.1 toolkit was used.

French language includes 14 vowels and 17 consonants. Based on similarities on the lips, the 31 phonemes can be grouped into 8 visemes. A viseme consists of group of phonemes that look similar on lips/mouth (e.g., $/p/$, $/b/$, and $/m/$). Table II shows the mapping of French phonemes to visemes. Based on previous works, five consonant visemes and three vowel visemes reflect the most appropriate phoneme-to-viseme mapping [12].

To realize automatic vowel and consonant recognition in Cued Speech, it is necessary to integrate lip shape and hand elements. In this study, lip shape and hand elements were integrated into a single component using concatenative feature fusion. Our aim, was to combine the two streams into a bimodal one, and to use the joint lip-hand feature vectors in the HMM system in order to realize vowel Cued Speech recognition. The feature concatenation uses the concatenation of the synchronous lip shape and hand position features as the joint bimodal feature vector

$$O_t^{LH} = [O_t^{(L)^T}, O_t^{(H)^T}]^T \in R^D \qquad (1)$$

where O_t^{LH} is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand feature vector, and D the dimensionality of the joint feature vector. In these experiments, the dimension of the lip shape stream was 24 (8 static parameters, 8 Δ, and 8 $\Delta\Delta$ parameters). In vowel recognition experiments, the dimension of the hand position stream was 12 (4 static parameters, 4 Δ, and 4 $\Delta\Delta$ parameters). The dimension D of the joint lip-hand feature vectors was, therefore 36. In consonant recognition the dimension of the handshape stream was 30, and the D dimension was, therefore, 54.

TABLE III
CONFUSION MATRIX OF VOWEL-VISEME RECOGNITION USING LIP SHAPE INFORMATION ONLY.

	V1	V2	V3	%correct	%error
V1	583	3	20	96.2	1.2
V2	33	1039	23	97.2	1.5
V3	11	11	209	90.5	1.2

III. RESULTS

A. Vowel recognition using concatenative feature fusion

1) Vowel recognition considering similarities on lips: In this experiment, the similarities on lips of the French vowels were considered to show how the integration of hand position element improves the recognition accuracy when vowels belonging to the same class were recognized. As previously described, vowels which show similarities on lips (visemes) cannot be recognized accurately using lip shape information alone.

Using the previously described phoneme-to-viseme mapping, the 14 vowels were classified into three groups, and three separate HMM sets were trained using the appropriate training data. In this way, each group included the most confusable vowels based on lip shape. Table IV shows the obtained results. Using lip shape parameters only, the average vowel accuracy was 64.4% because of the high number of confusions between similar vowels in each group. On the other hand, by integrating also hand position element with lip shape element, the average vowel accuracy was raised to 91.2%, showing 75.3% relative improvement compared with using lip shape parameters alone. However, the vowels belonging to the same group based on lips similarities are distinguishable using hand position information. As a result the confusions between them drastically decreased resulting in increase of recognition accuracy. Table IV also shows the results when auditory parameters were used. More specifically, the acoustic signal was parameterized using 12 Mel-Frequency Cepstral Coefficients (MFCC) and the first and second derivatives, as well. As can be seen, the performance in the case of Cued Speech is very similar to the performance obtained when using the acoustic signal.

2) Experimental results for overall vowel recognition: In this experiment, all vowels were recognized using a common HMM set. Table V shows the obtained results. Compared with using lip shape parameters alone, a 63.3% relative improvement was obtained when fusion was applied. It is also shown that results of vowel recognition in Cued Speech and those

TABLE IV
ACCURACY FOR THREE VOWEL GROUPS CONSIDERING SIMILARITIES ON LIPS.

Vowel group	Modality		
	Lips	Lips+Position	Audio
Group1	60.2	94.0	96.7
Group2	58.6	85.9	90.9
Group3	74.3	93.7	93.8
Average	64.4	91.2	93.8

TABLE V
VOWEL AND CONSONANT OVERALL ACCURACY.

Phonemes	Modality		
	Lips	Lips+Hand	Audio
Vowels	59.4	85.1	91.5
Consonants	52.1	78.9	90.7
Average	55.8	82.0	91.1

TABLE VI
ACCURACY FOR FIVE CONSONANT GROUPS CONSIDERING SIMILARITIES ON LIPS.

Consonant group	Modality		
	Lips	Lips+Shape	Audio
Group1	66.2	88.1	98.2
Group2	83.1	96.3	98.7
Group3	50.5	81.1	97.2
Group4	82.9	93.0	98.6
Group5	59.1	86.6	96.7
Average	68.4	89.0	97.9

obtained using the acoustic signal do not show significant differences.

B. Consonant recognition based on concatenative feature fusion

In this section, experiments for consonant recognition in Cued Speech for French are introduced. Using concatenative feature fusion, lip shape element was integrated with hand-shape element, and consonant recognition was conducted. For handshape modeling, the xy coordinates of landmarks placed on the fingers and the first and second derivatives were used, as well. In total 30 parameters were used for handshape modeling.

In a way similar to vowel recognition, the consonants were classified into groups based on lip shape similarities, and separate HMM sets were trained. Five HMM sets were trained corresponding to the five consonant groups. Table VI shows the obtained results. It can be seen, that using lip shape and handshape information, significant improvements in accuracy were obtained compared with the sole use of lip shape parameters. More specifically, a 65.2% relative improvement was obtained when handshape element was also used. Table VI also shows the results obtained using the acoustic signal. It can be seen, that in the cases of $Group2$ and $Group4$, Cued Speech accuracy and accuracy obtained using the acoustic signal are not very much different. In the case of the other three groups, accuracies obtained using acoustic signal are significantly higher. A possible reason might be the errors occured in handshape recognition, which takes place prior to consonant recognition. The results, however, are still comparable and promising in both vowel and consonant recognition in Cued Speech for French. Table V shows the obtained overall results for consonant recognition. It can be seen, that compared with using lip shape parameters only, a 56% relative improvement was obtained when fusion was applied.

IV. CONCLUSION

In this paper, vowel and consonant recognition in Cued Speech for French was presented. According to the Cued Speech original system, for vowel recognition lip shape and hand position elements were integrated, and automatic recognition was realized. In the case of consonant recognition, lip shape element and hand shape element were fused. Compared with the sole use of lip shape parameters, by integrating hand information, a 60% promising average relative improvement was obtained. The obtained results are comparable to results achieved using audio signal, even with using only visual information in Cued Speech recognition. Currently, collection of additional data is in progress in order to realize automatic Cued Speech recognition for larger vocabularies. Also Cued Speech data from two deaf cuers have been recorded, and experiments on Cued Speech recognition for deaf individuals are making progress. Preliminary results on isolated word recognition for a specific task in Cued Speech for French are very promising.

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] R. M. Uchanski, L. A. Delhorne, A. K. Dix, L. D Braida, 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. vol. 31(1), pp. 20–41, 1994.

[5] J. Leybaert, "Phonology acquired through the eyes and spelling in deaf children," *Journal of Experimental Child Psychology*, vol. 75, pp. 291–318, 2000.

[6] 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, V. 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.

[7] N. Aboutabit, D. Beautemps, and L. Besacier, "Lips and hand modeling for recognition of the cued speech gestures: The french vowel case," *Speech Communication*, 2008, (accepted with revision).

[8] P. Heracleous, N. Abboudatit, and D. Beautemps, "Lip shape and hand position fusion for automatic vowel recognition in cued speech for french," *IEEE Signal Processing Letters*, 2009 (to appear).

[9] N. Aboutabit, D. Beautemps, , and L. Besacier, "Hand and lips desynchronization analysis in french cued speech : Automatic segmentation of hand flow," *in Proceedings of ICASSP2006*, pp. 633–636, 2006.

[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] S. Nakamura, K. Kumatani, and S. Tamura, "Multi-modal temporal asynchronicity modeling by product hmms for robust," *in Proceedings of Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)*, p. 305, 2002.

[12] N. Aboutabit, D. Beautemps, J. Clarke, and L. Besacier, "A hmm recognition of consonant-vowel syllabes from lip contours: the cued speech case," *in Proceedings of Interspeech*, pp. 646–649, 2007.