Lip shape and hand position fusion for automatic vowel recognition in Cued Speech for French

Panikos Heracleous, Noureddine Aboutabit, Denis Beautemps

Lip shape and hand position fusion for automatic vowel recognition in Cued Speech for French

Panikos Heracleous, Noureddine Aboutabit, and Denis Beautemps

***Abstract*—Cued Speech is a sound-based system, which uses handshapes in different positions and in combination with lip-patterns of speech, and 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 lip reading and thus enable deaf children and adults to wholly understand spoken language. Cued Speech recognition requires gesture recognition and lip shape recognition. In addition, the integration of the two components is of great importance. This article presents hidden Markov model (HMM)-based vowel recognition as used in Cued Speech for French. Based on concatenative feature fusion and multi-stream HMM decision fusion, lip shape and hand position components were integrated into a single component, and automatic vowel recognition was realized. In the case of multi-stream HMM decision fusion, the obtained vowel classication accuracy using lip shape and hand position information was 87.6%, showing absolute improvement of 19.6% in comparison with a use restricted only to lip parameters. The results achieved show the effectiveness of the proposed approaches to Cued Speech recognition.**

***Index Terms*—Cued Speech, HMM, vowel recognition, concatenative fusion, multi-stream HMM fusion.**

I. INTRODUCTION

TO date, in many studies dealing with speech perception or speech recognition, visual information has also been used to complement the audio information (lip reading). The visual pattern, however, is ambiguous, and on an average, only 40 to 60% of the vowels are recognized by the lip reading system for a given language (American English) [1], and the accuracies are down to 10 to 30% where words are concerned [2]. However, for the orally educated deaf or hearing-impaired people, lip reading remains the main modality of perceiving speech. Therefore, in 1967 Cornett developed in 1967 the Cued Speech system as a supplement to lip reading [3]. Cued Speech improves speech perception for the hearing-impaired people [4], and for those who are exposed to this method since their youth, offers them a complete representation of the phonological system, and thereby has exerts a positive impact on the language development [5].

Cued Speech is a visual communication system that uses handshapes placed in different positions near the face in combination with natural speech lip reading to enhance speech perception with visual input. Here, two components make up a manual cue: the handshape and the hand position relative to the face. Handshapes are designed to distinguish consonant phonemes whereas hand positions are used to distinguish vowel phonemes. Fig. 1 describes the complete system relating to the French language, which uses eight handshapes in five different positions [6].

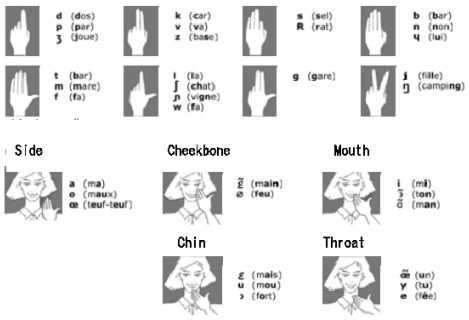

Fig. 1. Handshapes for consonants (top) and hand position (bottom) for vowels in Cued Speech for French.

Access to communication technologies has become essential for handicapped people. The TELMA project (Phone for deaf people) aims at developing an automatic translation system of acoustic speech into visual speech complete with Cued Speech and vice versa, i.e. from Cued Speech components into auditory speech [7]. This project would enable deaf users to communicate with each other and with normal-hearing people with the help of the autonomous terminal TELMA. In this system, the automatic translation of Cued Speech components into a phonetic chain is the key. The Cued Speech system allows both hand and lip ows to convey a part of the phonetic information. This means that, in order to recover the complete phonetic- and lexical information, lip- and hand components should be used jointly.

The Cued Speech paradigm requires accurate recognition of both lip shape and hand information. Fusion of lip shape and hand components is also necessary and very important. Fusion is the integration of available single modality streams into a combined stream. There were several studies in the past relating to automatic audio-visual recognition and integration of visual- and audio modalities [8]. The aim of audio-visual speech recognition is to improve the performance of a recognizer, especially in noisy environments.

This article focuses on vowel recognition in Cued Speech for French (referred to also as Cued French Language [9]) based on HMMs. As far as our knowledge goes, automatic vowel 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

The authors are with 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@gipsa-lab.inpg.fr

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

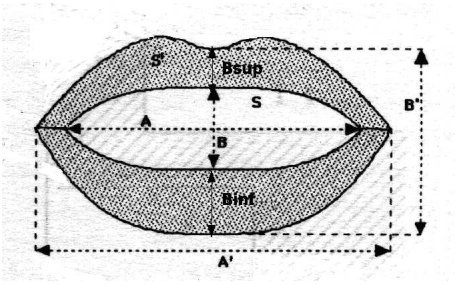

Fig. 2. Parameters used for lip shape modeling.

have not come across any published work related to automatic vowel or consonant recognition in Cued Speech for any other Cued language.

In the first attempt at vowel recognition in Cued Speech [10], a method based on separate identication, i.e., where indirect decision fusion has been used, a vowel accuracy of 77.6% was achieved. In this study however, the proposed method is based on HMMs and uses concatenative feature fusion and multi-stream HMM decision fusion aimed at integrating the two components, i.e., to combine the hand- and lip components in order to perform conventional automatic vowel recognition.

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 oodlight. 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 recording of data in order to control the data and facilitate the extraction of accurate features (see [11] 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 multisyllabic 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. The recording process resulted in a set of 638 sentences, which contained 5751 vowel instances. For the purposes of training and the tests, 3838 and 1913 vowel instances, respectively, were used. Table II shows the number of vowel instances included in the training and test sets, respectively. The training and test sets were used in all classication experiments.

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).

TABLE I
VOWEL-TO-VISEME MAPPING IN THE FRENCH LANGUAGE.

Viseme	Vowels
V1	/ɔ̃/, /y/, /o/ /ø/, /u/
V2	/a/, /ɛ̃/. /ɪ/ /œ̃/, /e/, / ɛ/
V3	/ɑ̃/, /ɔ/, /œ/

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.

In automatic speech recognition, a diagonal covariance matrix is often used based on the assumption that the parameters are uncorrelated. As far as lip reading is concerned, however, parameters show a strong correlation. In the framework of this study, Principal Component Analysis (PCA) was applied to decorrelate the lip shape parameters, then a diagonal covariance matrix was used. All 24 PCA lip shape components were used for HMM training. In the experiments, context-independent models with a 3-state, left-to-right HMM topology were used.

III. RESULTS

A. Vowel-viseme classification experiment

This section details the classication of vowel-visemes. A viseme [12] consists of group of phonemes that look similar on lips/mouth (e.g., $/p/$, $/b/$, and $/m/$).

Previous studies have shown that in the French language, 3 vowel groups reflect the most accurate phoneme-viseme mapping [11]. Table I shows the mapping of the French vowels to visemes. An experiment was conducted to evaluate the performance of the lip reading system relating to vowel-viseme recognition based on HMMs. The parameter vectors length was 24 (i.e., 8 basic lip parameters, Δ, and $\Delta\Delta$) and 32 Gaussian mixtures per state were used. The experiment resulted in a three-by-three confusion matrix and the diagonal elements (i.e., correctly classified visemes) yielded a 96.1% classication accuracy.

B. Hand position classification

Hand position is one of the two mandatory components (the other is lip shape) in respect of vowel recognition in Cued Speech. To evaluate the proposed system, an experiment was conducted relating to hand position recognition from the cuing of a Cued Speech cuer using HMMs. The parameters used for HMM modelling were the xy coordinates of the two marks placed on the cuers hand, automatically extracted from the data. The st- and second derivatives were used as

TABLE II
NUMBER OF VOWEL INSTANCES USED FOR TRAINING AND TEST.

Set	French vowels														Total
	/ø/	/y/	/o/	/ɔ̃/	/u/	/a/	/ɛ̃/	/ɪ/	/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

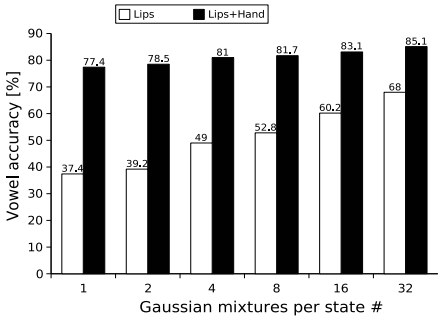

Fig. 3. Cued Speech vowel classification accuracy based on concatenative feature fusion.

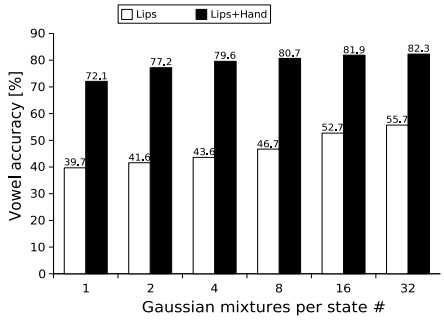

Fig. 4. Cued Speech vowel recognition accuracy in unconstrained phoneme recognition experiment based on concatenative feature fusion.

well. Considering that the study was of Cued Speech for French, five positions were taken into consideration and a HMM was trained in relation to each position. A five-by-five confusion matrix was generated and the diagonal elements (i.e., correctly classified hand positions) yielded a 91.5% classication accuracy.

C. Vowel recognition using using concatenative feature fusion

In these experiments, lip- and hand elements were integrated into a single component using concatenative feature fusion, and recognition experiments were conducted to recognize the 14 French vowels. The feature concatenation used the concatenation of the synchronous lip

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

where O_t^{LH} is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand position feature vector, and D the dimensionality of the joint feature vector. In these experiments, the dimension of the lip shape stream was 24 (8 basic parameters, 8 Δ, and 8 $\Delta\Delta$ parameters) and the dimension of the hand position stream was 12 (4 basic parameters, 4 Δ, and 4 $\Delta\Delta$ parameters). The dimension D of the joint lip-hand feature vectors was, therefore 36. This size was reasonable, and no further processing was applied to reduce the dimensionality.

Fig. 3 shows the results obtained in a vowel-classication experiment. In this case, the vowel instances were automatically extracted from the sentences using forced alignment based on the audio signal. Using 32 Gaussian mixtures per state, a 85.1% vowel classication accuracy was obtained. Compared with the sole use only lip parameters, an absolute improvement of 17.1% was obtained.

Fig. 4 shows the vowel correct (substitutions and deletions were considered) obtained in an unconstrained phoneme recognition experiment using continuous sentences, and without having recourse to a language model. In this case, the vowel correctness achieved was 82.3%. Compared with the sole use of lip parameters, an absolute improvement of 26.6% was obtained. In both cases, high recognition rates were obtained which were comparable to those obtained by automatic vowel recognition experiments using audio signals (e.g., [13]).

D. Vowel recognition using multi-stream HMM decision fusion

In these experiments, lip shape and hand position elements were integrated into a single component using multi-stream HMM decision fusion, and recognition experiments were conducted to recognize the 14 French vowels. Decision fusion captures the reliability of each stream, by combining the likelihoods of single-stream HMM classis. Such an approach has been used in multi-band audio only ASR [14] and in audio-visual speech recognition [8]. 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 position elements, the emission probability of multi-stream HMM is given by

$$b_j(O_t) = \prod_{s=1}^{S} [\sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm})]^{\lambda_s} \tag{2}$$

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}. 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,

$$0 \leq \lambda_h, \lambda_l \leq 1 \tag{3}$$

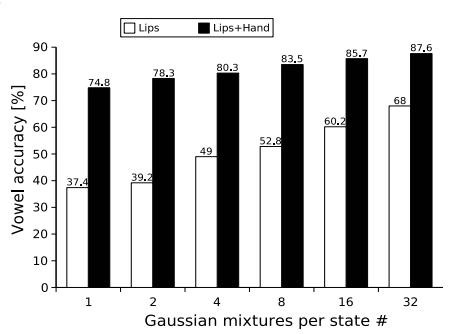

Fig. 5. Cued Speech vowel classification based on multi-stream HMM decision fusion.

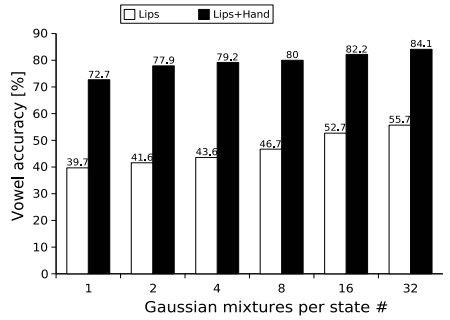

Fig. 6. Cued Speech vowel recognition using in an unconstrained phoneme recognition experiment based on multi-stream HMM decision fusion.

and

$$\lambda_h + \lambda_l = 1 \tag{4}$$

where λ_h is the hand position 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 0.7 and 0.3 values, respectively. The selected weights were obtained by maximizing the accuracy on several experiments.

Fig. 5 shows the classification accuracies obtained when vowel instances were automatically extracted from the sentences. Using 32 Gaussian mixtures per state, an 87.6% vowel classification accuracy was obtained. Compared with the sole use of lip parameters, an absolute improvement of 19.6% was obtained. Fig. 6 shows the obtained results in an unconstrained phoneme recognition experiment. In this case, the vowel correct was 84.1%. Compared with the sole use of lip parameters, an absolute improvement of 28.4% was obtained.

The results showed that multi-stream HMM decision fusion results in better performance than a concatenative feature fusion. To decide whether the difference in performance between the two methods is statistically signicant, the McNemars test was applied [15]. The observed p-value was 0.001 indicating that the difference is statistically signicant.

IV. Conclusions and future work

This article details a study on automatic vowel recognition in Cued Speech based on HMMs. Using integrated lip shape and hand position parameters, and based on concatenative feature fusion and multi-stream HMM decision fusion promising vowel classication accuracies of 85.1% and 87.6%, respectively, were achieved. The results obtained showed that fusion methods applied in audio-visual speech recognition could also be applied in Cued Speech recognition with great effectiveness. In the present case, vowel recognition was presented. Consonant recognition requires a different approach from that used in the vowel recognition. More specically, in consonant recognition handshape coding and recognition are required. We are currently working on the consonant recognition problem and preliminary results are 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, 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, 25:262-269, 1982.

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

[4] R. M. Uchanski, L. A. Delhorne, A. K. Dix, C. M Reed, L. D. Braida, and N. I. Durlach, *Automatic Speech Recognition to Aid the Hearing Impaired: Current Prospects for the Automatic Generation of Cued Speech*, Journal of Rehabilitation Research and Development, Vol. 31, 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] F. Destombes, *Aides manuelles à la lecture labiale et perspectives d'aides automatiques. "Le Projet VIDVOX"*, Centre Scientifique IBM-France, pp. 35–36, 1982.

[7] D. Beautemps, L. Girin, N. Aboutabit, G. Bailly, L. Besacier, G. Breton, T. Burger and 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] 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.

[9] E. Fleetwood, and M. Metzger, M. Cued Language Structure: An Analysis of Cued American English Based on Linguistic Principles, Calliope Press, Silver Spring, MD (USA), ISBN 0-9654871-3-X, 1998.

[10] N. Aboutabit, D. Beautemps, and L. Besacier, *Automatic identification of vowels in the Cued Speech context*, in Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP), 2007.

[11] N. Aboutabit, D. Beautemps, and L. Besacier, *Lips and Hand Modeling for Recognition of the Cued Speech Gestures: The French Vowel Case*, Speech Communication, (to appear).

[12] E. Owens, and B. Blazek, *Visemes observed by hearing-impaired and normal- hearing adult viewers*, Journal of Speech and Hearing Research Vol.28, pp. 381-393, 1985.

[13] P. Merkx, J. Miles, *Automatic Vowel Classification in Speech. An Artificial Neural Network Approach Using Cepstral Feature Analysis* Final Project for Math 196S, pp.1–14, 2005.

[14] H. Bourlard, and S. Dupont, *A new ASR approach based on independent processing and recombination of partial frequency bands*, in Proceedings of International Conference on Spoken Language Processing, pp 426-429, 1996.

[15] L. Gillick, and S. Cox, *SOME STATISTICAL ISSUES IN THE COMPARISON OF SPEECH RECOGNITION ALGORITHMS*, in Proceedings of ICASSP89, pp. 532-535, 1989.