Automatic recognition of French Cued Speech using multimodal fusion based on hidden Markov models

Panikos Heracleous, Noureddine Aboutabit, Denis Beautemps

Automatic recognition of French Cued Speech using multimodal fusion based on hidden Markov models

Panikos Heracleous, Noureddine Aboutabit, and Denis Beautemps

***Abstract*—In this article, automatic recognition of French Cued Speech based on hidden Markov models (HMM) is presented. Cued Speech is a visual 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 lipreading and thus enable deaf children and adults to understand full spoken language. In automatic recognition of Cued Speech, lip shape and gesture recognition are required. In addition, the integration of the two modalities is of the greatest importance. In this study, lip shape component is fused with gestures components to realize Cued Speech recognition. Using concatenative feature fusion and multi-stream HMM decision fusion, vowel recognition and consonant recognition experiments have been conducted. For vowel recognition, an 87.6% vowel accuracy was obtained showing a 61.3% relative improvement compared to the sole use of lip shape parameters. In the case of consonant recognition, a 78.9% accuracy was obtained showing a 56% relative improvement compared with the use of lip shape only. In addition to vowel and consonant recognition, a complete phoneme recognition experiment using concatenated feature vectors and Gaussian mixture model (GMM) discrimination has been conducted showing a 74.4% phoneme accuracy. The obtained results were compared to the results obtained using the audio signal showing comparable accuracies. The achieved results show the effectiveness of the proposed approaches as far as Cued Speech recognition is concerned.**

***Index Terms*—French Cued Speech, hidden Markov models, automatic, feature fusion, multi-stream HMM decision fusion.**

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.

However, even with high lip reading performances, without knowledge about the semantic context, speech cannot be thoroughly perceived. The best lip readers are far way of reaching perfection. On average, only 40 to 60% of the phonemes of a given language are recognized by lip reading [1], and 32% when relating to low predicted words [2]. The best results obtained amongst deaf participants was 43.6% for the average accuracy and 17.5% for standard deviation with regards to words [3], [4]. The main reason for this lies in the ambiguity of the visual pattern. However, as far as the orally educated deaf people are concerned, the act of lip-reading remains the main modality of perceiving speech.

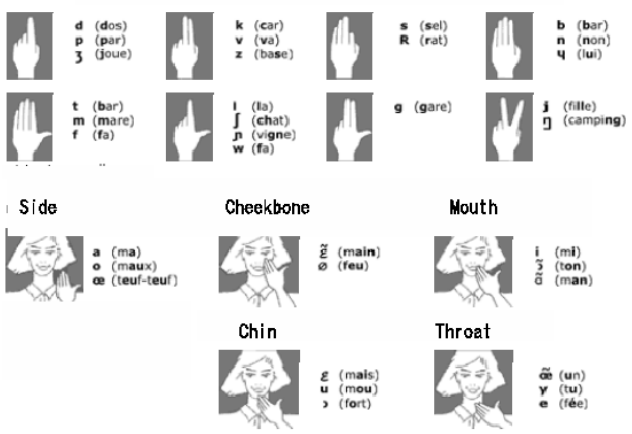

Fig. 1. Handshapes for consonants (top) and hand position (bottom) for vowels in French Cued Speech.

To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, in 1967 Cornett [5] 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 only visual information.

Cued Speech uses handshapes placed in different positions near the face in combination with natural speech lipreading to enhance speech perception from visual input. This is a system where the speaker faces the perceiver and moves his hand in close relation with speech. 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 in this system contains two components: the handshape and the hand position relative to the face. Handshapes distinguish among consonant phonemes whereas hand positions distinguish among vowel phonemes. A handshape together with a hand position cue a syllable. Cued Speech recognition requires gestures recognition and lip shape

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, Noureddine.Aboutabit, Denis.Beautemps@gipsa-lab.inpg.fr

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

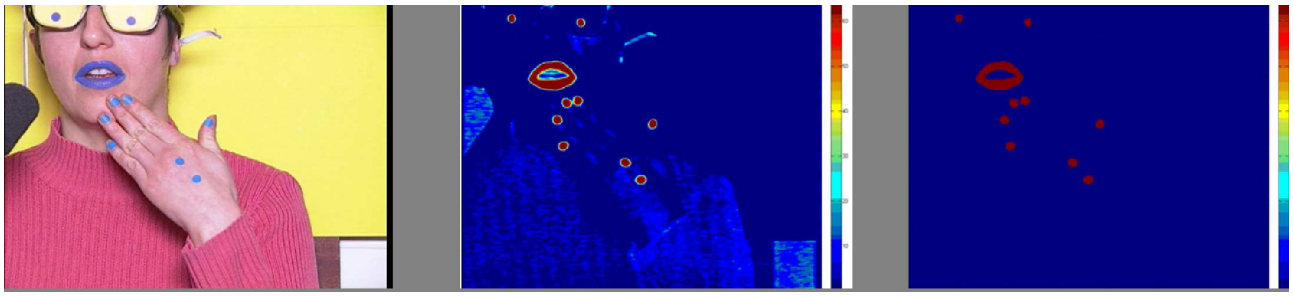

Fig. 2. The three-step algorithm applied for lip shape and gesture detection based on detection of blue objects.

recognition, and also integration of the two modalities.

Cued Speech improves speech perception for deaf people [2], [6]. Moreover, it offers to deaf people a thorough representation of the phonological system, in as much as they have been exposed to this method since their youth, and therefore it has a positive impact on the language development [7]. 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. To date, Cued Speech is adapted to more than 60 languages.

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 completed with Cued Speech and vice versa, i.e. from Cued Speech components into auditory speech [8]. This project would enable deaf users to communicate with each others and with normal-hearing people through the help of the autonomous terminal TELMA. In this context, the automatic translation of Cued Speech components into a phonetic chain is a key issue. The Cued Speech system allows both hand and lip flows to convey a part of the phonetic information. This means that, in order to recover the complete phonetic and lexical information, lip and hand modalities should be used jointly.

In a first attempt for vowel recognition in Cued Speech, in [9] a method based on separate identification, i.e., indirect decision fusion was used and 77.6% vowel accuracy was obtained. In this study, the proposed method is based on HMMs and it uses concatenative feature fusion and multi-stream HMM decision fusion to integrate the modalities into a combined one and then perform automatic recognition.

II. Feature extraction and processing

For the sake of data collection, the female native French speaker-employed was certified in French Cued Speech. The cuer wore a helmet to keep the head in a fixed position together with opaque glasses to protect her eyes against halogen floodlight. A camera with a zoom facility was connected to a betacam recorder and was used to shoot the hand and face. The lips were painted in blue, and blue marks were placed on the speaker's glasses as reference points. Since the aim of this study is the fusion of lip shape and hand position components for vowel recognition in French Cued Speech, these constraints were applied in data recording in order to control the data and to facilitate the extraction of the accurate features (see [10], [11] for details).

Using blue artifices makes easier the extraction of the lip contour and the hand features. To do this, the task needs a specific image processing. This processing is done on two main steps:

- Detection of the blue objects on the image.
- Extraction of the lip contours.

In the first 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.

In the second step, the lip contour and the blue landmarks on the back of the hand, and also at the extremities 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 audio part of the video recording was digitized at 22,050 Hz in synchrony with the image part. The acoustic signal was automatically labeled at phonetic level using forced alignment. The previously described image processing was applied to the video frames in 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).

The described process resulted in a set of temporally coherent signals: the 2-D hand information, the lip width, the lip aperture and the lip area values for both inner and outer contours, and the corresponding acoustic signal with the associated phonetic labelled transcriptions. In addition, two supplementary parameters relative to the lip morphology were extracted: the pinching of the upper and lower lips. As a result, a total set of eight parameters 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). For handshape modeling, the coordinates of five landmarks placed on the fingers were used (I.e., 10 parameters).

TABLE I
PHONEME-TO-VISEME MAPPING IN FRENCH LANGUAGE.

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

III. EXPERIMENTAL CONDITIONS

Cued Speech paradigm requires accurate recognition of both lip shape and hand gestures. Fusion of lip shape and hand modalities 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 modalities were fused. For consonant recognition, lip shape and handshape modalities were fused.

Previously, several studies have been made in automatic audio-visual recognition and integration of visual and audio modalities [12], [13], [14], [15], [16], [17] . The aim of audio-visual speech recognition is to improve the performance of a recognizer, especially under noisy environments. Although the objective is different, results obtained by the authors showed that similar approaches can also be applied for Cued Speech recognition.

In the experiments, speaker-dependent, context-independent models were used. A 3-state, left-to-right with no skip HMM topology was used. Each state was modeled with 32 Gaussian mixtures. In addition to the static lip and hand parameters, the first (Δ) and second derivatives ($\Delta\Delta$) were used, as well. For training and test 426 and 212 continuous utterances were used, respectively. The training utterances contained 3838 vowels and 4401 consonants. The test utterances contained 1913 vowels and 2155 consonants. Vowels and consonants were excised automatically 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, 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 21 consonants. In this study, only 17 consonants were considered because of the data available for the classification techique. Based on similarities on lips, the 31 phonemes can be grouped into 8 visemes. In visual domain, however phonemes articulated in the same manner are ambiguous and speech is represented by visemes. A viseme consists of phonemes that are hard to be distinguished using only lip shape information. Table I 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 [18].

IV. EXPERIMENTAL RESULTS FOR VOWEL RECOGNITION

In this section, vowel recognition experimental results in French Cued Speech are presented. Vowel recognition requires integration of lip shape and hand position components. For hand position modeling, the coordinates of two landmarks placed on the hand are used, along with the first and second derivatives.

A. Hand position recognition

Hand position is one of the two mandatory modalities (I.e., the other is lip shape) in respect of vowel recognition in Cued Speech. To evaluate the proposed system, an experiment has been conducted regarding 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 cuer's hand, automatically extracted from the data. The first and second derivatives were used as well. Considering the French Cued Speech, five positions were taken into consideration and a HMM was trained in relation to each position. The labels of the phonetic transcriptions were merged to form only five labels corresponding to the five positions. The resulting average hand position accuracy was up to 91.5%. The high recognition rate of hand position raised the hope that a complete vowel recognition system for Cued Speech would achieve high performance.

B. Vowel-viseme recognition using lip shape information

An experiment has been conducted for recognition of the three vowel-visemes in order to evaluate the proposed method for lip shape modeling and the vowel-viseme mapping. For each viseme, a HMM was trained using the previously described data. The vowel labels in the phonetic transcription were translated to include only three viseme labels. Table II shows the confusion matrix of vowel-viseme recognition. As can be seen, the vowel-visemes were recognized with high accuracy. More specifically, 96.1% viseme accuracy was achieved. Table II also shows that the partial error corresponding to each viseme is uniform and almost identical.

C. Lip shape and hand position integration using concatenative feature fusion

In this section, lip shape and hand position modalities were integrated into a single modality using concatenative feature fusion. Our aim, was to combine the two streams into a

TABLE II
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

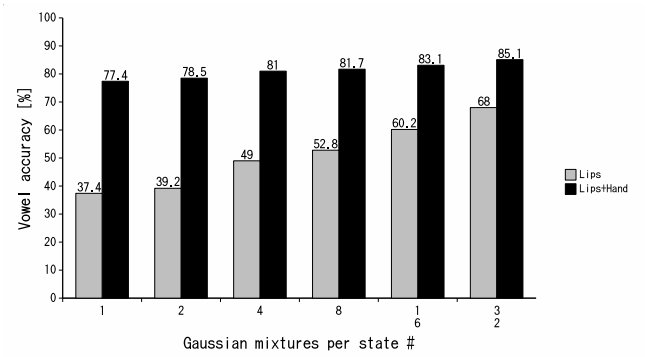

Fig. 3. Cued Speech vowel recognition using only lip and hand parameters. Vowel instances were excised automatically from the data and concatenative feature fusion was used.

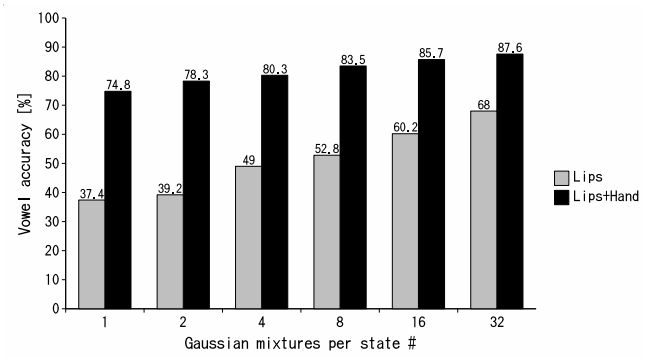

Fig. 4. Cued Speech vowel recognition using only lip and hand parameters. Vowel instances were excised automatically from the data and multi-stream HMM decision fusion was used.

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

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

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 static parameters, 8 Δ, and 8 $\Delta\Delta$ parameters) and 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. This size was reasonable, and no further processing was applied to reduce the dimensionality. Fig. 3 shows the obtained results. As can be seen, by integrating hand position component with lip shape component, a 85.1% vowel accuracy was achieved, showing a 53% relative improvement compared with the sole use of lip shape parameters.

D. Lip shape and hand position integration using multi-stream HMM decision fusion

In this experiment, lip shape and hand position modalities were integrated into a single modality using multi-stream HMM decision fusion. Decision fusion captures the reliability of each stream, by combining the likelihoods of single-modality HMM classifiers. Such an approach has been used in multi-band audio only ASR [19] and in audio-visual speech recognition [12]. The emission likelihood of multi-stream HMM is the product of emission likelihoods of single-modality components weighted appropriately by stream weights. Given the O bimodal observation vector, i.e., lip shape and hand position modality, the emission probability of multi-stream HMM is given by

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

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_{sjt}. In this study, we assume that the stream weights do not depend on state j and time t. However, two constraints were applied. Namely,

$$0 \leq \lambda_h, \lambda_l \leq 1 \quad (3)$$

and

$$\lambda_h + \lambda_l = 1 \quad (4)$$

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 to 0.7 and 0.3 values, respectively. The selected weights were obtained experimentally by maximizing the accuracy on a held-out data. Fig. **??** shows the achieved results when multi-stream HMM decision fusion was used. Using 32 Gaussian mixtures per state, an 87.6% vowel accuracy was obtained, showing a relative improvement of 61%.

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

	Modality		
Vowel group	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

E. 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 modality 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 only.

Using the 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 includes the most confusable vowels based on lip shape. Table III shows the obtained results. Using only lip shape parameters, the average vowel accuracy was 64.4% because of a high number of confusions between similar vowels in each group. On the other hand, by integrating hand position modality with lip shape modality, the average vowel accuracy raised to 91.2%, showing 75.3% relative improvement compared with using lip shape parameters only. However, the vowels belonging to the same group based on lips similarities, are distinguishable using hand position information and the confusions between them drastically decreased. Table III also shows the results when spectral parameters were used. More specifically, the acoustic signal was parameterized using 12 Mel-Frequency Cepstral Coefficients (MFCC) and 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 using the audio signal.

V. Proposed method for handshape coding

In French Cued Speech, recognition of the eight handshapes is an exceptional case of the handshape recognition. In fact, a causal analysis based on some knowledge like the number and dispersion of fingers, and also the angle between them can distinguish between those eight handshapes. Based on the number of landmarks detected on fingers, the correct handshape can be recognized. In fig. 1 handshapes were numbered from left to right (i.e, S1–S8). The proposed algorithm to identify the Cued Speech handshapes is as follows:

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

TABLE IV
CONFUSION MATRIX OF HANDSHAPE RECOGNITION EVALUATION.

	S0	S1	S2	S3	S4	S5	S6	S7	S8	%c
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

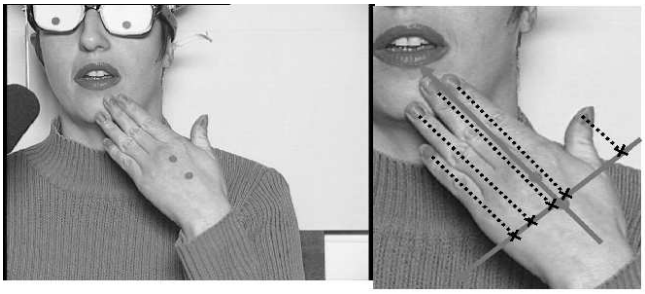

Fig. 5. Image fof a Cued Speech cuer (left) and the projection method (right).

A. Evaluation and results for handshape coding

The previous method was applied to each sequence of the corpus. To evaluate the previous handshape recognition system, a set of 1009 frames were used and recognized automatically. Table IV shows the confusion matrix of the recognized handshapes by the automatic system. As can be seen, the automatic system recognized correctly 92% of the handshapes. This score justified the choice of the authors, and showed that using only 5 landmarks placed at the finger extremities the accuracy did not decrease drastically compared with the 98,78% of recognized handshapes obtained by [20] system based on 50 tags. The most general errors 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. In some other cases, landmarks remain visible even when the fingers are bended.

B. Finger identification

The objective of this section was to identify fingers in each frame of the recording. The identification has been done in three steps. In the first step, all landmarks in the frame were detected. The landmarks placed on the speaker glasses and on the back of the hand are 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 consisted of sorting the resulted coordinates following the perpendicular axis to the hand direction from the smaller to the largest (Fig. 5). In this step, the handshape coding was used to associate each coordinate with the corresponding finger. For example, when there were three landmarks coordinates and the handshape number was S3, the smallest coordinate was associated with the middle finger, the middle one to the ring finger, and the biggest one to the baby finger.

VI. Experimental results for consonant recognition

In this section, automatic consonant recognition in French Cued Speech is presented. For consonant recognition, fusion of lip shape and handshape components are required. For handshape modeling, the coordinates of the five landmarks placed on the fingers were used.

TABLE V
CONFUSION MATRIX OF CONSONANT-VISEME RECOGNITION USING LIP SHAPE INFORMATION ONLY.

	C1	C2	C3	C3	C4	%correct	%error
C1	310	0	1	0	0	99.7	0.0
C2	0	129	7	0	0	94.9	0.3
C3	18	16	649	26	81	82.2	7.0
C4	1	2	24	93	9	72.1	1.8
C5	2	2	89	28	541	81.7	6.0

A. Consonant-viseme recognition using lip shape information

Similarly to vowel-viseme recognition, an experiment has been conducted for consonant-viseme recognition in order to evaluate the performance of the lipreading system. According to the Table I, in the French language consonants are grouped into five visemes. For each viseme, a HMM was trained using the previously described data. The consonant labels in the phonetic transcription were translated to include only five viseme labels. Table V shows the confusion matrix of consonant-viseme recognition. As can be seen, the consonant-visemes were recognized with a 84.9% viseme accuracy. Table V also shows that visemes articulated at front are recognized with higher accuracy compared to visemes articulated at back. The consonant-viseme accuracy was lower compared to vowel-viseme accuracy. The reason is that the consonants have less visual information compared to the vowels.

B. Consonant recognition based on concatenative feature fusion

Using concatenative feature fusion, lip shape modality was integrated with handshape modality and consonant recognition has been conducted. For handshape modeling, the xy coordinates of fingers were used and first and second derivatives, as well. In total 30 parameters were used for handshape modeling. For lip shape modeling 24 parameters were used i.e., 8 static, 8 Δ and 8 $\Delta\Delta$ lip shape parameters. Fig. 6 shows the obtained results in the function of Gaussian mixtures per state. As can be seen, using 32 mixtures per state a consonant accuracy of 78.9% was achieved. Compared with the sole use of lip shape, a 56% relative improvement was obtained.

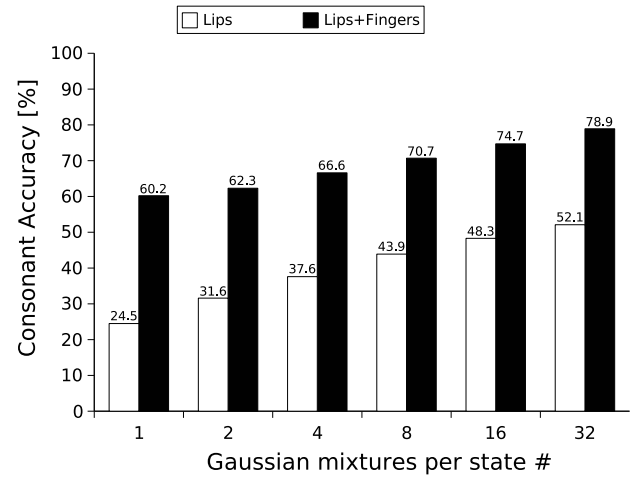

Fig. 6. Cued Speech consonant recognition using only lip and hand parameters. Consonant instances were excised automatically from the data and concatenative feature fusion was used.

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

	Modality		
Consonant group	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

C. Consonant recognition considering similarities on lips

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 using lip shape parameters only. More specifically, 65% relative improvement was obtained when handshape modality was also used. Table VI also shows the results obtained using the audio signal. It can be seen, that in the cases of $Group2$ and $Group4$, Cued Speech accuracy and accuracy obtained using the audio signal are similar. In the case of the other three groups, accuracies obtained using the audio signal are higher. A possible reason might be the errors occured in handshape recognition. However, results are still comparable and promising also in French Cued Speech consonant recognition.

VII. COMPARING NORMAL SPEECH AND CUED SPEECH RECOGNITION IN THE PRESENCE OF NOISE

In this section, experimental results for comparing normal speech and Cued Speech recognition performances in the presence of noise are presented. Office noise was superimposed on clean normal speech test data at different Signal-to-Noise Ratio (SNR) levels. The audio HMMs were trained using clean speech. Fig. 7 shows the achieved results in the case of vowel recognition. The results show that up to 30dB SNR level, Cued Speech showed higher performance compared to normal speech. In the case of clean test data, normal speech showed slightly higher vowel accuracy.

Fig. 8 shows the obtained results in the case of consonant recognition. Similar to vowel recognition, Cued Speech performed better up to 30dB SNR. In the case of clean test data, normal speech showed higher performancee compared to Cued Speech.

Both in vowel and consonant Cued Speech recognition, results were promising and comparable to normal speech recognition. Moreover, in noisy environment Cued Speech has the additional advantage of robustness against noise.

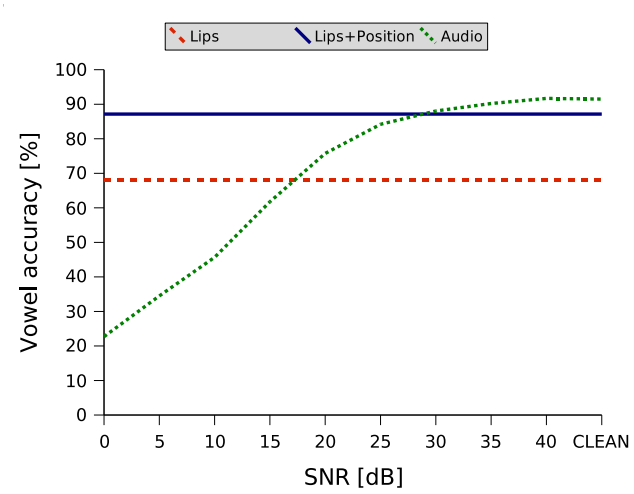

Fig. 7. Vowel recognition using lips, audio, and Cued Speech modalities.

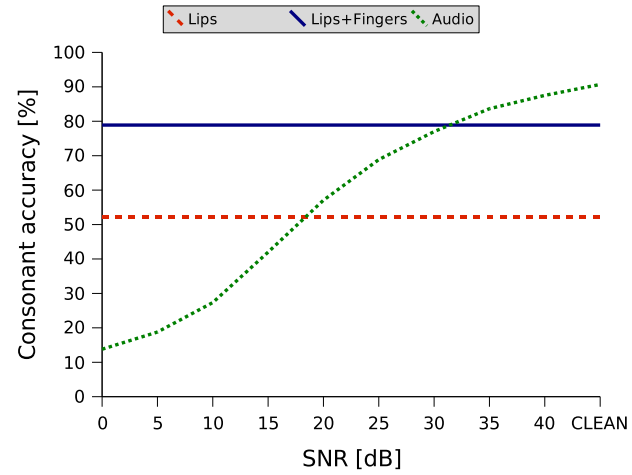

Fig. 8. Consonant recognition using lips, audio, and Cued Speech modalities.

VIII. Experimental results for phoneme recognition

In this section, experiments for a complete phoneme recognition in French Cued Speech are presented. In the previous section, vowel and consonant recognition experiments were presented. In those cases, the recognizer's grammar contained only vowels or consonants, respectively. In a complete phoneme recognition, however, the grammar should contain both vowels and consonants.

In the previous sections, it was also reported that different modeling was used for vowels and consonants. More specifically, for vowel modeling fusion of lip shape and hand position components was used. On the other hand, for consonant modeling fusion of lip shape and handshape components was used. As a result, feature vectors in vowel and consonant recognition are of different length. Feature vectors for vowels have length of 36, and feature vectors of consonants have length 54. This is a limitation in using a common HMM set for phoneme recognition. To deal with this problem, three approaches are proposed in order to realize phoneme recognition in French Cued Speech.

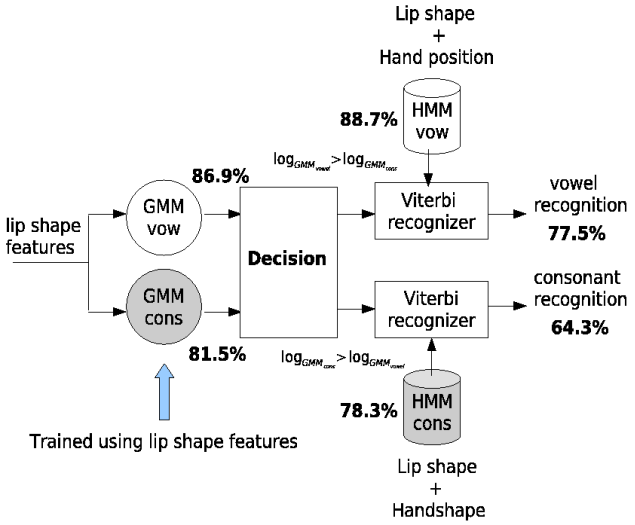

Fig. 9. Phoneme recognition based on a vowel and on consonant GMM models.

A. Experiment using concatenated feature vectors

In this experiment, feature vectors with same length were used. At each frame, lip shape parameters, hand position parameters, and handshape parameters were extracted during the image processing stage. When separate vowel or consonant recognition was realized, only the appropriate parameters were used. In this experiment, for each phoneme all parameters were used in concatenated feature vectors with length of 66 (i.e. 8 lip shape parameters, 4 coordinates for hand position, 10 coordinates forhandshape, and first and second derivatives, as well). The obtained phoneme accuracy was as low as 61.5%. Although, phoneme recognition in French Cued Speech is a difficult task, the obtained phoneme accuracy was lower than expected. Therefore, to obtain higher performance different approaches were also investigated.

B. Experiment using a vowel and a consonant Gaussian mixture models (GMM)

In this section, Cued Speech phoneme recognition based on GMM discrimination is presented as it is shown in Fig. 9. A vowel-independent and a consonant-independent 64-mixture GMM was trained, using lip shape parameters only. For the two GMMs training the corresponding vowel and consonant data were used.

Phoneme recognition was realized in a two-pass scheme. In the first pass, using the two GMMs decision has been taken about the nature of the input, i.e., vowel of consonant. More specifically, matching was performed between the input and the two GMMs. Based on the obtained likelihood, the input was considered to be vowel or consonant. For instance, when the likelihood of vowel-GMM was higher than that of the consonant-GMM, a decision was made for a vowel. On the other hand, when consonant-GMM provided the higher

likelihood, the input considered to be a consonant. In the case of vowel inputs, the discrimination accuracy was 86.9%, and in the case of consonant inputs the discrimination accuracy was 81.5%.

In the second pass, switching to the appropriate HMM set took place and vowel or consonant recognition was realized, respectively using feature vectors corresponding to vowel modeling, or consonant modeling. The obtained phoneme accuracy was 70.9% (I.e. 77.5% vowel accuracy and 64.3% consonant accuracy). The obtained accuracies were lower than the accuracies obtained in the separate vowel and consonant recognition experiments, because of the discrimination errors of the first pass. However, the obtained result was still promising and it showed a relative improvement of 24% compared to the use of concatenated feature vectors.

C. Experiment using eight viseme Gaussian mixture models

In order to further improve the discrimination accuracy of the first pass, an approach was applied which used eight GMM models, instead of two. More specifically, three vowel-viseme GMMs and five consonant-viseme GMMs were used in this experiment using the appropriate phoneme groups. Similarly to the previous experiment, phoneme recognition was performed in two passes. In the first pass, matching between the input and the eight visemes took place. Based on the maximum likelihood, the system switched to the corresponding vowel or consonant HMM set, and recognition was performed.

Using eight GMMs, the discrimination accuracy was increased up to 89.3% for the vowels, and up to 84.6% for the consonants. The achieved phoneme recognition was 74.4% (I.e., 80.3% vowel accuracy and 68.5% consonant accuracy). Compared with the use of two GMMs, a relative improvement of 12% was obtained. Compared with the use of concatenated full set of parameters, a relative improvement of 33.5% was achieved.

IX. Conclusion

In this paper, automatic recognition of French Cued Speech was presented based on concatenative feature fusion and multi-stream HMM decision fusion. For vowel recognition, lip shape and hand position modalities were integrated and automatic recognition was realized. In the case of consonant recognition, lip shape modality and hand shape modality were fused. Compared with using lip shape only, by integrating hand information a 60% promising average relative improvement was obtained. A complete phoneme recognition experiment was also conducted, showing a 74.4% accuracy. The results were promising and comparable to those obtained using audio signal. However, problems still remain. In this study, synchrony between lip shape and hand modalities was assumed. In the future, possible asynchrony between the two modalities will be also considered. Currently, collection of additional data is in progress in order to realize automatic Cued Speech recognition for extended tasks.

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] E. T. Auer and L. E. Bernstein, "Enhanced visual speech perception in individuals with early-onset hearing impairment," *Journal of Speech, Language, and Hearing*, vol. 50, pp. 1157–1165, 2007.

[4] L. Bernstein, E. Auer, and J. Jiang, "Lipreading, the lexicon, and Cued Speech," *In C. la Sasso and K. Crain and J. Leybaert (Eds.), Cued Speech and Cued Language for Children who are Deaf or Hard of Hearing, Los Angeles, CA: Plural Inc. Press*, In Press.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[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*, 2008, (accepted with revision).

[12] 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.

[13] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," *in Proceedings of ICASSP 2002*, 2002.

[14] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," *in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer*, p. 331349, 1996.

[15] 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.

[16] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," *in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany:Springer*, p. 461471, 1996.

[17] T. Chen, "Audiovisual speech processing. lip reading and lip synchronization," *IEEE Signal Processing Mag.*, vol. vol. 18, pp. 921, 2001.

[18] 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.

[19] 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.

[20] 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. vol. 118(2), pp. 1144–1153, 2005.