Cued Speech Automatic Recognition in Normal Hearing and Deaf Subjects

Panikos Heracleous, Denis Beautemps, Noureddine Aboutabit

Elsevier Editorial System(tm) for Speech Communication
Manuscript Draft

Manuscript Number: SPECOM-D-09-00049R4

Title: Cued Speech Automatic Recognition in Normal-hearing and Deaf Subjects

Article Type: SI:Speech and Face-to-Face Communication

Section/Category: Special Issue Paper

Keywords: French Cued Speech; hidden Markov models; automatic recognition; feature fusion; multi-stream HMM decision fusion

Corresponding Author: Dr Panikos Heracleous, Ph.D

Corresponding Author's Institution: ATR

First Author: Panikos Heracleous, Ph.D

Order of Authors: Panikos Heracleous, Ph.D; Denis Beautemps; Noureddine Aboutabit

Manuscript Region of Origin: FRANCE

Abstract: This article discusses the automatic recognition of Cued Speech in French based on hidden Markov models (HMMs). Cued Speech is a visual mode which, by using hand shapes in different positions and in combination with lip patterns of speech, makes all the sounds of a spoken language clearly understandable to deaf people. The aim of Cued Speech is to overcome the problems of lipreading and thus enable deaf children and adults to understand spoken language completely. In the current study, the authors demonstrate that visible gestures are as discriminant as audible orofacial gestures. Phoneme recognition and isolated word recognition experiments have been conducted using data from a normal-hearing cuer. The results obtained were very promising, and the study has been extended by applying the proposed methods to a deaf cuer. The achieved results have not shown any significant differences compared to automatic Cued Speech recognition in a normal-hearing subject. In automatic recognition of Cued Speech, lip shape and gesture recognition are required. Moreover, the integration of the two modalities is of great importance. In this study, lip shape component is fused with hand component to realize Cued Speech recognition. Using concatenative feature fusion and multi-stream HMM decision fusion, vowel recognition, consonant recognition, and isolated word recognition experiments have been conducted. For vowel recognition, an 87.6\% vowel accuracy was obtained showing a 61\% 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 to 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 was conducted, obtaining a 74.4\% phoneme accuracy. Isolated word recognition experiments in both normal-hearing and deaf subjects were also conducted providing a word accuracy of 94.9\% and 89\%, respectively. The obtained results were compared with those obtained using audio signal, and comparable accuracies were observed.

Cued Speech Automatic Recognition in Normal-hearing and Deaf Subjects

Panikos Heracleous[1,2], Denis Beautemps[2], Noureddine Aboutabit[2]

[1]*ATR, Intelligent Robotics and Communication Laboratories,*
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan
[2]*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@atr.jp

Abstract

This article discusses the automatic recognition of Cued Speech in French based on hidden Markov models (HMMs). Cued Speech is a visual mode which, by using hand shapes in different positions and in combination with lip patterns of speech, makes all the sounds of a spoken language clearly understandable to deaf people. The aim of Cued Speech is to overcome the problems of lipreading and thus enable deaf children and adults to understand spoken language completely. In the current study, the authors demonstrate that visible gestures are as discriminant as audible orofacial gestures. Phoneme recognition and isolated word recognition experiments have been conducted using data from a normal-hearing cuer. The results obtained were very promising, and the study has been extended by applying the proposed methods to a deaf cuer. The achieved results have not shown any significant differences compared to automatic Cued Speech recognition in a normal-hearing subject. In automatic recognition of Cued Speech, lip shape and gesture recognition are required. Moreover, the integration of the two modalities is of great importance. In this study, lip shape component is fused with hand component to realize Cued Speech recognition. Using concatenative feature fusion and multi-stream HMM decision fusion, vowel recognition, consonant recognition, and isolated word 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 to 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 was conducted, obtaining a 74.4% phoneme accuracy. Isolated word recognition experiments in both normal-hearing and deaf subjects were also conducted providing a word accuracy of 94.9% and 89%, respectively. The obtained results were compared with those obtained using audio signal, and comparable accuracies were observed.

Key words: French Cued Speech, hidden Markov models, automatic recognition, feature fusion, multi-stream HMM decision fusion

***Revision Note**

Response to Reviewers' Comments

Ref	Ms. No. SPECOM-D-09-00049
Title	Automatic Recognition in Normal-hearing and Deaf Subjects
Authors	Panikos Heracleous, Denis Beautemps, and Noureddine Aboutabit

The authors would like to very much thank the quest editors and the reviewers for their effort to review the manuscript, and also for their very helpful and interesting comments. All the comments have been addressed.

Cued Speech Automatic Recognition in Normal-hearing and Deaf Subjects

Panikos Heracleous[1,2], Denis Beautemps[2], Noureddine Aboutabit[2]

[1]*ATR, Intelligent Robotics and Communication Laboratories,*
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan
[2]*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@atr.jp

Abstract

This article discusses the automatic recognition of Cued Speech in French based on hidden Markov models (HMMs). Cued Speech is a visual mode which, by using hand shapes in different positions and in combination with lip patterns of speech, makes all the sounds of a spoken language clearly understandable to deaf people. The aim of Cued Speech is to overcome the problems of lipreading and thus enable deaf children and adults to understand spoken language completely. In the current study, the authors demonstrate that visible gestures are as discriminant as audible orofacial gestures. Phoneme recognition and isolated word recognition experiments have been conducted using data from a normal-hearing cuer. The results obtained were very promising, and the study has been extended by applying the proposed methods to a deaf cuer. The achieved results have not shown any significant differences compared to automatic Cued Speech recognition in a normal-hearing subject. In automatic recognition of Cued Speech, lip shape and gesture recognition are required. Moreover, the integration of the two modalities is of great importance. In this study, lip shape component is fused with hand component to realize Cued Speech recognition. Using concatenative feature fusion and multi-stream HMM decision fusion, vowel recognition, consonant recognition, and isolated word 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 to 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 was conducted, obtaining a 74.4% phoneme accuracy. Isolated word recognition experiments in both normal-hearing and deaf subjects were also conducted providing a word accuracy of 94.9% and 89%, respectively. The obtained results were compared with those obtained using audio signal, and comparable accuracies were observed.

Key words: French Cued Speech, hidden Markov models, automatic recognition, feature fusion, multi-stream HMM decision fusion

1. Introduction

To date, visual information is widely used to improve speech perception or automatic speech recognition (lipreading) (Potamianos et al., 2003). With lipreading technique, speech can be understood by interpreting the movements of lips, face and tongue. In spoken languages, a particular facial and lip shape corresponds to a specific 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.

Without knowing the semantic context, one cannot perceive the speech thoroughly even with high lipreading performances. To date, the best lip readers are far away into reaching perfection. On average, only 40 to 60% of the vowels of a given language (American English) are recognized by lipreading (Montgomery and Jackson, 1983), and 32% when relating to low predicted words (Nicholls and Ling, 1982). The best result obtained amongst deaf participants was 43.6% for the average accuracy (Auer and Bernstein, 2007; Bernstein et al., 2007). 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 lipreading remains the main modality of perceiving speech.

To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, Cornett (Cornett, 1967) developed in 1967 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 face/lips (e.g., /*p*/, /*b*/, and /*m*/), using hand information those sounds can be distinguished and thus make possible for deaf people to completely understand a spoken language using visual information only.

Cued Speech (also referred to as Cued Language (Fleetwood and Metzger, 1998)) uses hand shapes placed in different positions near the face along 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

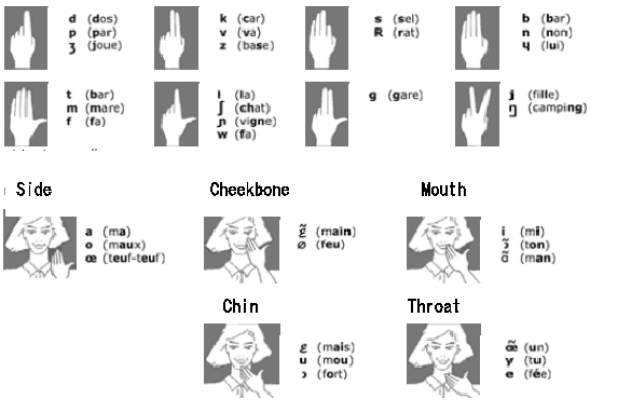

Figure 1: Hand shapes for consonants (top) and hand position (bottom) for vowels in French Cued Speech.

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 hand shape and the hand position relative to the face. Hand shapes distinguish among consonant phonemes whereas hand positions distinguish among vowel phonemes. A hand shape, together with a hand position, cues a syllable.

Cued Speech improves the speech perception of deaf people (Nicholls and Ling, 1982; Uchanski et al., 1994). Moreover, for deaf people who have been exposed to this mode since their youth, it offers a complete representation of the phonological system, and therefore it has a positive impact on the language development (Leybaert, 2000). Fig. 1 describes the complete system for French. In French Cued Speech, eight hand shapes in five positions are used. The system was adapted from American English to French in 1977. To date, Cued Speech has been adapted in more than 60 languages.

Another widely used communication method for deaf individuals is the Sign Language (Dreuw et al., 2007; Ong and Ranganath, 2005). Sign Language is a language with its own grammar, syntax and community; however, one must be exposed to native and/or fluent users of Sign Language to acquire it. Since the majority of children who are deaf or hard-of-hearing have hearing parents (90%), these children usually have limited access to appropriate Sign Language models. Cued Speech is a visual representation of a spoken language, and it was developed to help raise the literacy levels of deaf individuals. Cued Speech was not developed to replace Sign Language. In fact, Sign Language will be always a part of deaf community. On the other hand, Cued Speech is an alternative communication method for deaf individuals. By cueing, children who are deaf would have a way to easily acquire the native home language, read and write proficiently, and communicate more easily with hearing family members who cue them.

In the current study, the authors demonstrate that visible gestures are as discriminant as audible orofacial gestures. Phoneme recognition and isolated word recognition experiments were conducted using data from a normal-hearing cuer, and promising results were obtained. In addition, the proposed methods were applied to a deaf cuer and similar results were obtained compared with automatic recognition of Cued Speech in normal-hearing subjects.

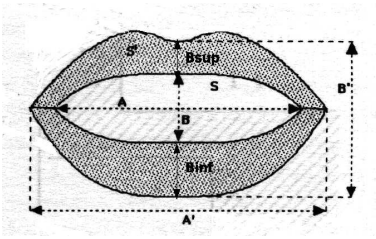

Figure 2: Parameters used for lip shape modeling.

In the first attempt for vowel recognition in Cued Speech, in (Aboutabit et al., 2007) a method based on separate identification, i.e., indirect decision fusion was used and a 77.6% vowel accuracy was obtained. In this study, however, the proposed method is based on HMMs and uses concatenative feature fusion and multi-stream HMM decision fusion to integrate the components into a combined one and then perform automatic recognition. Fusion (Nefian et al., 2002; Hennecke et al., 1996; Adjoudani and Benoît, 1996) is the integration of all available single modality streams into a combined one. In this study, lip shape and hand components are combined in order to realize automatic recognition in Cued Speech for French.

2. Methodology

2.1. Cued Speech Materials

The data for vowel-, consonant-, and phoneme recognition experiments were collected from a normal-hearing cuer. 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. The cuer's 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 (Aboutabit et al., view) for details). The data were derived from a video recording of the cuer 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 (Gibert et al., 2005). Each sentence was dictated by an experimenter, and was repeated two or three times (to correct the pronunciation errors) by the cuer resulting in a set of 638 sentences.

The audio part of the video recording was synchronized with the image. Fig. 2 shows the lip shape parameters used in the study. An automatic image processing method was applied to the video frames in the lip region to extract their inner and outer

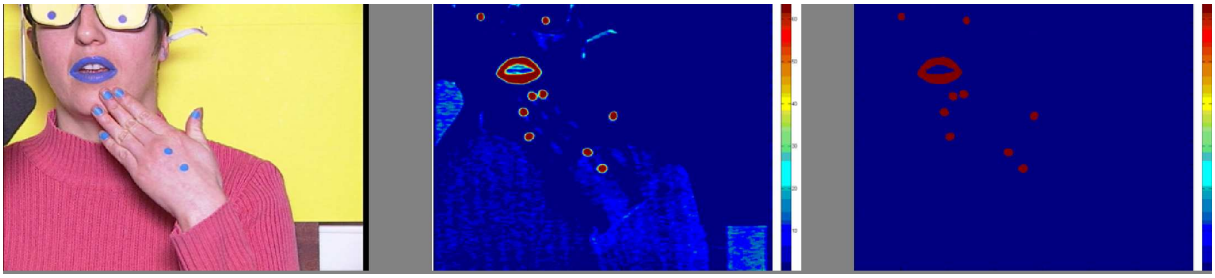

Figure 3: The three-step algorithm applied for lip shape and gesture detection based on detection of blue objects.

contours and 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 2D 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. 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 modeling lip shapes. For hand position modeling, the *xy* coordinates of two landmarks placed on the hand were used (i.e., 4 parameters). For hand shape modeling, the *xy* coordinates of the landmarks placed on the fingers were used (i.e., 10 parameters). Non visible landmarks receive default coordinates [0,0].

During the recording of Cued Speech material for isolated word recognition experiments, the conditions were different from the ones described earlier. The system was improved by excluding the use of a helmet by the cuer, enabling in this way the head movements during recording. The subject was seated on a chair in a way to avoid large movements in the third direction (i.e. towards the camera). However, the errors that might occur have not been evaluated. In addition, the landmarks placed on the cuer's fingers were of different colors in order to avoid the hand shape coding and the finger identification (cf. section "ordering finger landmarks)", and this helped to simplify and speed up the image processing stage. In these recording sessions, a normal-hearing cuer and a deaf cuer were employed. The corpus consisted of 1450 isolated words with each of 50 words repeated 29 times by the cuers.

2.2. Lip shape and hand components modeling

In the phoneme recognition experiments, context-independent, 3-state, left-to-right, no-skip-phoneme HMMs were used. Each state was modeled with a mixture of 32 Gaussians. In addition to the basic lip and hand parameters, first- (Δ) and second-order derivatives ($\Delta\Delta$) were used as well. For training and test, 426 and 212 sentences were used, respectively. The training sentences contained 3838 vowel and 4401 consonant instances, and the test sentences contained 1913 vowel and 2155 consonant instances, respectively. Vowels and consonants were extracted automatically from the data after a forced alignment was performed using the audio signal.

For isolated word recognition experiments two HMM sets were trained (deaf and normal-hearing). Fifteen repetitions of each word were used to train 50, 6-state, whole word HMMs, and 14 repetitions were used for testing. Eight and ten parameters were used for lip shape and hand shape modeling, respectively.

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, a global Principal Component Analysis (PCA) using all the training data was applied to decorrelate the lip shape parameters and then a diagonal covariance matrix was used. The test data were then projected into the PCA space. All PCA lip shape components were used for HMM training. For training and recognition the HTK3.1 toolkit (Young et al., 2001) was used.

2.3. Ordering finger landmarks

For consonant recognition, the correct hand shape is also required. Instead of a deterministic recognition of the hand shape, a probabilistic method is used based on the *xy* coordinates of the landmarks placed on the fingers (Fig. 3). The coordinates are used as features for the hand shape modeling. During the image processing stage, the system detects the landmarks located on the cuer's fingers and their coordinates are computed. Since the landmarks are of the same color, the system cannot assign the coordinates to the appropriate finger in order to be used correctly (i.e. correct order) in the feature vectors. To do this, the hand shape is automatically recognized a-priori, and the information obtained is then used to assign the coordinates to the appropriate finger.

In French Cued Speech, recognition of the eight hand shapes is considered exceptional. 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 those eight hand shapes. Based on the number of landmarks detected on fingers, the correct hand shape can be recognized. In Fig. 1 the hand shapes were numbered from left to right (i.e, S1-S8). The proposed algorithm to identify the Cued Speech hand shapes is as follows:

- Number of fingers on which landmarks are detected = 1, then the hand shape is S1.

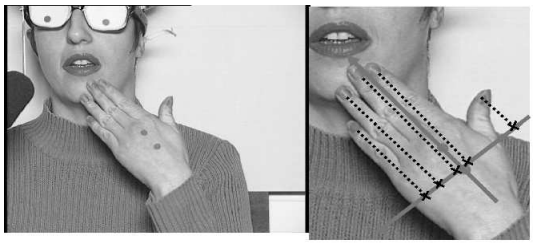

Figure 4: Image of a Cued Speech cuer (left) and the projection method (right).

- Number of fingers on which landmarks are detected = 4, then the hand shape is S4.
- Number of fingers on which landmarks are detected = 5, then the hand shape is S5.
- Number of finger on which landmarks are detected = 3, then the hand shape is S3 or S7. If the thumb finger is detected (using finger dispersion models) then the hand shape is S7, else the hand shape is S3.
- Number of finger on which landmarks are detected = 2, then the hand shape is S2 or S6 or S8. If the thumb finger is detected then the hand shape is 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 objective of finger identification stage (cf. section "ordering finger landmarks) is to assign the computed coordinates to the correct finger and, in this way, to have the correct order in the feature vectors. 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 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 consisted of sorting the 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 with the corresponding finger. For example, when there were three landmarks coordinates and the hand shape 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. The coordinates of the fingers which are not present in a hand shape are replaced by a constant in order to keep the same dimension of the feature vectors in the HMM modeling.

2.4. Concatenative Feature Fusion

The feature concatenation uses the concatenation of the synchronous lip shape and hand features as the joint 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 feature vector, and D the dimensionality of the joint feature vector. In vowel recognition 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 position feature vectors was, therefore 36. In consonant recognition experiments, the dimension of the hand shape stream was 30 (10 basic parameters, 10 Δ, and 10 $\Delta\Delta$ parameters). The dimension D of the joint lip-hand shape feature vectors was, therefore 54.

2.5. 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 (Bourlard and Dupont, 1996) and in audio-visual speech recognition (Potamianos et al., 2003). 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 joint observation vector, i.e., lip shape and hand position component, the emission probability of multi-stream HMM is given by

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

where $N(O; \mu, \Sigma)$ is the value in O of a multivariate Gaussian with mean μ and covariance matrix Σ, and S the number of streams. 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, as happen in the general case of multi-stream HMM decision fusion. However, two constraints were applied, namely

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

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. The weights were adjusted to 0.7 and 0.3 values, respectively. The selected weights were obtained experimentally by maximizing the accuracy on held-out data.

3. Experiments and results

3.1. Hand shape recognition

To evaluate the previously described hand shape recognition system, a set of 1009 frames was used and recognized automatically. Table 1 shows the confusion matrix of the recognized hand shapes by the automatic system. It can be seen that the automatic system recognized correctly 92.4% of the hand shapes on average. This score showed that using only the 2D coordinates of 5 landmarks placed at the finger extremities, the accuracy did not decrease drastically compared with the 98.8% of

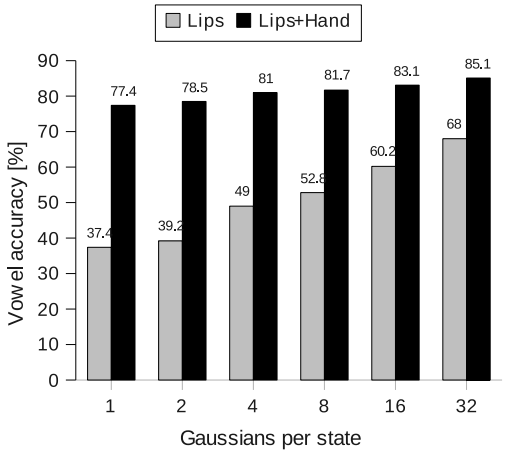

Figure 5: Cued Speech vowel recognition using only lip and hand parameters based on concatenative feature fusion.

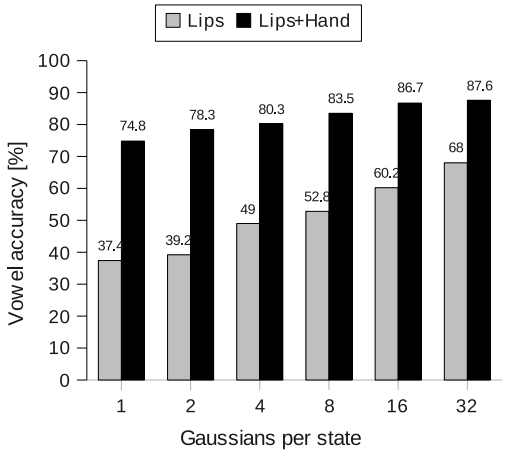

Figure 6: Cued Speech vowel recognition using only lip and hand parameters based on multi-stream HMM decision fusion was used.

recognized hand shapes obtained by (Gibert et al., 2005) with the use of the 3D coordinates of 50 landmarks placed on the hand and the fingers derived from a motion capture system. The most common errors can be attributed to landmark detection processing. However, in some cases, one or more landmarks were not detected because of the rotation of the hand. In some other cases, landmarks remained visible even when the fingers were bended.

3.2. Vowel and consonant recognition

Fig. 5 shows the vowel recognition results when concatenative feature fusion was used. As shown, by integrating hand position component with lip shape component, a vowel accuracy of 85.1% was achieved, showing a 53% relative improvement compared to the sole use of lip shape parameters.

Fig. 6 shows the results achieved for vowel recognition, when multi-stream HMM decision fusion was applied. Using 32 Gaussians per state, an 87.6% vowel accuracy was obtained, showing a relative improvement of 61%. The results obtained are comparable with the results obtained for vowel recognition when using audio speech (e.g., (Merkx and Miles, 2005)).

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 significant, the McNemar's test was applied (Gillick and Cox, 1989).The observed p-value was 0.001 indicating that the difference is statistically significant.

Table 1: Confusion matrix of hand shape recognition evaluation (derived for (Aboutabit, 2007)

	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

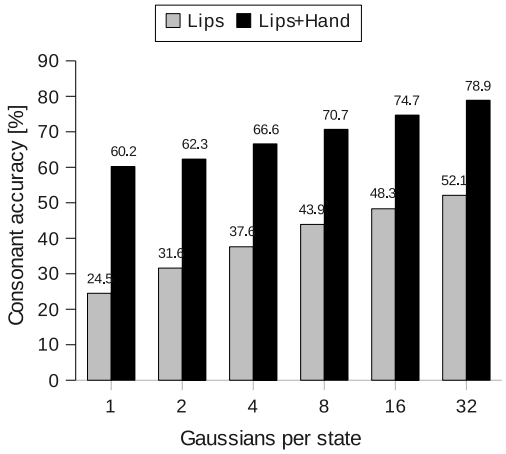

Figure 7: Cued Speech consonant recognition using only lip and hand parameters based on concatenative feature fusion.

Using concatenative feature fusion, lip shape component was integrated with hand shape component and consonant recognition was conducted. For hand shape modeling, the *xy* coordinates of the fingers, and first- and second-order derivatives were used. In total, 30 parameters were used for hand shape modeling. For lip shape modeling, 24 parameters were used. Fig. 7 shows the obtained results in the function of Gaussians per state. It can be seen that when using 32 Gaussians per state, a consonant accuracy of 78.9% was achieved. Compared to the sole use of lip shape, a 56% relative improvement was obtained.

3.3. Phoneme recognition

In the previous sections, it was reported that different types of modeling were used for vowels and consonants. More specifically, for vowel modeling, fusion of lip shape and hand position components was used. For consonant modeling, fusion of lip shape and hand shape components was used. As a result, feature vectors in vowel and consonant recognition are of different

lengths. The feature vectors for vowels have a length of 36 and the feature vectors of consonants have a length of 54. This is a limitation in using a common HMM set for phoneme recognition. To deal with this problem, three approaches were proposed in order to realize phoneme recognition in Cued Speech for French.

In the first approach, feature vectors with the same length were used. At each frame, lip shape parameters, hand position parameters, and hand shape parameters were extracted during the image processing stage. For each phoneme, all parameters were used in concatenated feature vectors with a length of 66 (i.e. 8 lip shape parameters, 4 coordinates for hand position, 10 coordinates for hand shape, along with the first- and second-order derivatives, as well). The obtained phoneme accuracy was as low as 61.5% due to a high number of confusions between vowels and consonants. Although, phoneme recognition in Cued Speech is a difficult task, the obtained phoneme accuracy was lower than expected. To obtain a performance with higher phoneme accuracy, different approaches were also investigated.

Cued Speech phoneme recognition was further improved by applying GMM discrimination. A vowel-independent and a consonant-independent GMM models were trained using lip shape parameters only. For training the two GMMs the corresponding vowel and consonant data were used. For modeling, 64 Gaussians were used. The number of Gaussians was selected experimentally on several experiments in order to achieve the highest classification scores. Phoneme recognition was realized in a two-pass scheme. In the first pass, using the two GMMs, the nature of the input was decided. More specifically, the input and the two GMMs were matched. Based on the obtained likelihood, the input was considered to be vowel or consonant. When the likelihood of the vowel-GMM was higher than that of the consonant-GMM, the decision was made for a vowel. When the consonant-GMM provided a higher likelihood, the input was 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 using feature vectors corresponding to the vowel modeling or consonant modeling, respectively. The obtained phoneme accuracy was 70.9%. 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. The obtained result, however, showed a relative improvement of 24% compared to the use of concatenated feature vectors.

In the third appproach for phoneme recognition, instead of two, eight GMM models (i.e. each one corresponding to a viseme) were used. Three vowel-viseme GMMs and five consonant-viseme GMMs were used based on the viseme grouping. Similar to the previous experiment, phoneme recognition was performed in two passes. In the first pass, matching between the input and the eight GMMs 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 to the use of two GMMs, a relative improvement of 12% was obtained. Compared with the use of the full set of concatenated parameters, a relative improvement of 33.5% was achieved.

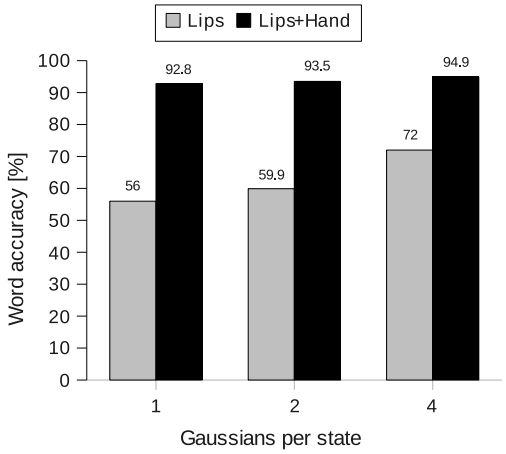

Figure 8: Word accuracy for isolated word recognition in the case of a normal-hearing subject.

3.4. Isolated word recognition

In this section, isolated word recognition experiments both in normal-hearing and deaf subjects are presented. In these experiments, the landmarks were of different colors in order to avoid the hand shape recognition and the finger identification stage. The image processing system locates the landmarks, and the coordinates of each landmark are assigned to each finger based on the colors. Doing this, the feature vectors of hand shape contains the coordinates of the landmarks in the correct order, and the errors that occur during the hand shape coding are not accumulated into the recognition stage.

Fig. 8 shows the results obtained in the function of several Gaussians per state in the case of the normal-hearing cuer. In the case of a single Gaussian per state, using lip shape alone obtained a 56% word accuracy; however, when hand shape information was also used, a 92.8% word accuracy was obtained. The highest word accuracy when using lip shape was 72%, obtained in the case of using 4 Gaussians per state. In that case, the Cued Speech word accuracy using also hand information was 94.9%.

Fig. 9 shows the obtained results in the case of a deaf cuer. The results show that in the case of the deaf subject, words were better recognized when using lip shape alone compared to the normal-hearing subject. The fact that deafs rely on lipreading for speech communication may increase their ability not only for speech perception but also for speech production. The word accuracy in the case of the deaf subject was 89% compared to the 94.9% in the normal-hearing subject. The difference in performance might be because of the lower hand shape recognition

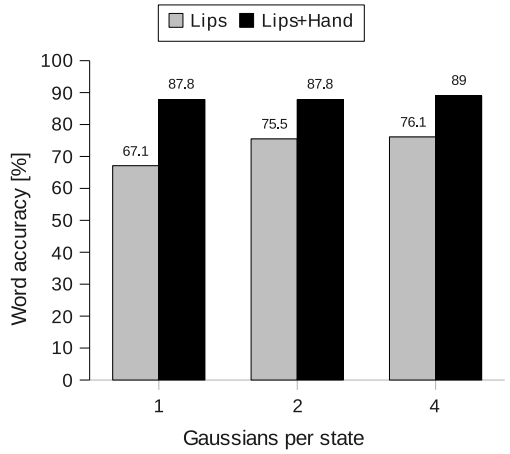

Figure 9: Word accuracy for isolated word recognition in the case of a deaf subject.

in the deaf subject. It should also be noted that the normal-hearing cuer was a professional teacher of Cued Speech. The results show that there are no additional difficulties in recognizing Cued Speech in deaf subjects, other than those appearing in normal-hearing subjects.

A multi-cuer isolated word recognition experiment was also conducted using the normal-hearing and the deaf cuers' data. The aim of this experiment is to investigate whether it is possible to train speaker-independent HMMs for Cued Speech recognition. The training data consisted of 750 words from the normal-hearing subject, and 750 words from the deaf subject. For testing 700 words from normal-hearing subject and 700 words from the deaf subject were used, respectively. Each state was modeled with a mixture of 4 Gaussian distributions. For lip shape and hand shape integration, the concatenative feature fusion was used.

Table 2 shows the results obtained when lip shape and hand shape features were used. The results show, that due to the large variability between the two subjects, word accuracy of cross-recognition is extremely low. On the other hand, the word accuracy in normal-hearing subject when using multi-speaker HMMs was 92%, which is comparable with the 94.9% word accuracy when cuer-dependent HMMs were used. In the case of the deaf subject, the word accuracy when using multi-cuer HMMs was 87.2%, which was also comparable with the 89% word accuracy when using speaker-dependent HMMs.

The results obtained indicate that creating speaker-independent HMMs for Cued Speech recognition using a large number of subjects should not face any particular difference, other than those appear in the conventional audio speech recognition. To prove this, however, additional experiments using a large number of subjects are required.

4. Discussion

This study deals with the automatic recognition of Cued Speech in French based on HMMs. As far as our knowledge goes, automatic vowel-, consonant- and phoneme recognition

Table 2: Word accuracy of a multi-speaker experiment

	HMMs		
Test data	Normal	Deaf	Normal+Deaf
Normal	94.9	0.6	92.0
Deaf	2.0	89.0	87.2

in Cued Speech based on HMMs is being introduced for the first time ever by the authors of this study. Based on a review of the literature written about Cued Speech, the authors of this study have not come across any other published work related to automatic vowel- or consonant recognition in Cued Speech for any other Cued language.

The study aims at investigating the possibility of integrating lip shape and hand information in order to realize automatic recognition, and converting Cued Speech into text with high accuracy. The authors were interested in the fusion and the recognition part of the components, and details of image processing techniques are not covered by this work.

In the conducted experiments, it was assumed that lip shape and hand shape components are synchronous. Based on previous studies, however, there might be asynchrony between the two components (Aboutabit et al., 2006). Late fusion (Potamianos et al., 2003), coupled HMMs (Nefian et al., 2002) and product HMMs (Nakamura et al., 2002) would be used as possible alternatives to the state-synchronous fusion methods used in this work.

Although the results are promising, problems still persist. For example, in order to extract accurate features, some constraints were applied in recording, and the computational cost was not considered. Also, a possible asynchrony between the components should be further investigated. The current pilot study on Cued Speech recognition attempts to extend the research in areas related to deaf communities, by offering to individuals with hearing disorders additional communication alternatives. For practical use, however, many questions should be addressed and solved, such as speaker-, environment-independence, real-time processing, etc. The authors are still analyzing the remaining problems in the framework of the TELMA project.

5. Conclusion

In this article, vowel-, consonant-, and phoneme recognition experiments in Cued Speech for French were presented. To recognize Cued Speech, lip shape and hand components were integrated into a single component using concatenative feature fusion and multi-stream HMM decision fusion. The accuracies achieved were promising and comparable to those obtained when using an audio speech. Specifically, accuracy obtained was 87.8% for vowel recognition, 78.9% for consonant recognition, and 74.4% for phoneme recognition. In addition, isolated word experiments in Cued Speech in both normal-hearing and deaf subjects were also conducted obtaining a 94.9% and 89% accuracy, respectively. A multi-cuer experiment using data

from both normal-hearing and deaf subject showed an 89.6% word accuracy, on average. This result indicates that training cuer-independent HMMs for Cued Speech using a large number of subjects should not face particular difficulties. Currently, additional Cued Speech data collection is in progress, in order to realize cuer-independent continuous Cued Speech recognition.

Acknowledgments

The authors would like to thank the volunteer cuers Sabine Chevalier, Myriam Diboui, and Clémentine Huriez for their time spending on Cued Speech data recording, and also for accepting the recording constraints. Also the authors would like to thank Christophe Savariaux and Coriandre Vilain for their help in the Cued Speech material recording. This work was mainly performed at GIPSA-lab, Speech and Cognition Department and was supported by the TELMA project (ANR, 2005 edition).

References

G. Potamianos, C. Neti, G. Gravier, A. Garg, A. Senior, Recent advances in the automatic recognition of audiovisual speech, in Proceedings of the IEEE 91, Issue 9 (2003) 1306–1326.

A. A. Montgomery, P. L. Jackson, Physical characteristics of the lips underlying vowel lipreading performance, Journal of the Acoustical Society of America 73 (6) (1983) 2134–2144.

G. Nicholls, D. Ling, Cued speech and the reception of spoken language, Journal of Speech and Hearing Research 25 (1982) 262–269.

E. T. Auer, L. E. Bernstein, Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment, Journal of Speech, Language, and Hearing 50 (2007) 1157–1165.

L. Bernstein, E. Auer, 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 .

R. O. Cornett, Cued Speech, American Annals of the Deaf 112 (1967) 3–13.

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

R. M. Uchanski, L. A. Delhorne, A. K. Dix, L. D. Braida, C. M. Reedand, 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) (1994) 20–41.

J. Leybaert, Phonology acquired through the eyes and spelling in deaf children, Journal of Experimental Child Psychology 75 (2000) 291–318.

P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, H. Ney, Speech Recognition Techniques for a Sign Language Recognition System, In Proceedings of Interspeech (2007) 2513–2516.

S. Ong, S. Ranganath, Automatic sign language analysis: A survey and the future beyond lexical meaning, IEEE Trans. PAMI vol. 27, no. 6 (2005) 873891.

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

A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, K. Murphy, A COUPLED HMM FOR AUDIO-VISUAL SPEECH RECOGNITION, in Proceedings of ICASSP 2002 .

M. E. Hennecke, D. G. Stork, 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 (1996) 331349.

A. Adjoudani, 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 (1996) 461471.

N. Aboutabit, D. Beautemps, L. Besacier, Lips and Hand Modeling for Recognition of the Cued Speech Gestures: The French Vowel Case, Speech Communication .

G. Gibert, G. Bailly, D. Beautemps, F. Elisei, 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) (2005) 1144–1153.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book, Cambridge University Engineering Department .

H. Bourlard, 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 (1996) 426–429.

N. Aboutabit, Reconnaissance de la Langue Française Parlée Complété (LPC): Décodage phonétique des gestes main-lèvres, PhD dissertation (2007), Institut National Polytechnique de Grenoble, Grenoble, France .

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

L. Gillick, S. Cox, SOME STATISTICAL ISSUES IN THE COMPARISON OF SPEECH RECOGNITION ALGORITHMS, in Proceedings of ICASSP89 (1989) 532–535.

N. Aboutabit, D. Beautemps, , L. Besacier, Hand and Lips desynchronization analysis in French Cued Speech : Automatic segmentation of Hand flow, in Proceedings of ICASSP2006 (2006) 633–636.

S. Nakamura, K. Kumatani, S. Tamura, Multi-Modal Temporal Asynchronicity Modeling by Product HMMs for Robust, in Proceedings of Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02) (2002) 305.