ALTERNATIVE SPEECH COMMUNICATION BASED ON CUED SPEECH

Panikos Heracleous, Noureddine Aboutabit, and Denis Beautemps

GIPSA-lab, Speech and Cognition Department, CNRS UMR 5216 / Stendhal University/UJF/INPG
961 rue de la Houille Blanche Domaine universitaire - BP 46 F - 38402 Saint Martin d'Hères cedex, France
E-mail: {Panikos.Heracleous, Noureddine.Aboutabit, Denis.Beautemps}@gipsa-lab.grenoble-inp.fr

ABSTRACT

This study focuses on alternative speech communication based on Cued Speech. Cued Speech is a visual mode of communication that uses handshapes and placements in combination with the mouth movements of speech to make the phonemes of a spoken language look different from each other and clearly understandable to deaf and hearing-impaired people. Originally, the aim of Cued Speech was to overcome the problems of lip reading and thus enable deaf children and adults to wholly understand spoken language. In this study, we investigate the use of Cued Speech not only for perception, but also for speech production in the case of speech- or hearing-impaired individuals. The proposed method is based on hidden Markov model (HMM) automatic recognition. Automatic recognition of Cued Speech can be served as an alternative speech communication method for individuals with speech- or hearing impairments. This article presents vowel and consonant, and also isolated word recognition experiments for Cued Speech for French. The results obtained are promising and comparable to the results obtained when using audio signal.

1. INTRODUCTION

Speech is often described as a unimodal process. However, it is well known that speech is multimodal in nature, and includes modalities such as audio modality, visual modality, touch modality, signal of muscles, articulatory speech, and brain activity during speech production.

The audio modality describes the audible speech used for communication between individuals without disorders in speech perception and production. In addition, audio modality includes inaudible speech produced when vocal folds are not vibrating (i.e., whisper speech), and also tissue-conducted Non-Audible Murmur speech [1]. Messages can be also transferred using information produced by lips and mouth, and by facial information. Cued Speech, a system that uses handshapes in different positions to complement lipreading, is a wholly visual speech system [2]. Speech can be perceived and interpreted by touching, as in the Tadoma technique. Tadoma is a technique that enables a person who is both blind and deaf to receive and interpret speech. In addition, speech messages can be transferred through the activation signal of muscles during speech production [3]. Invisible articulatory components i.e., tongue and velum is the motor which makes link between the different modalities. Movements of the invisible articulators during speech production also carry speech information [4, 5]. Also, some studies investigate the brain activity during speech production [6].
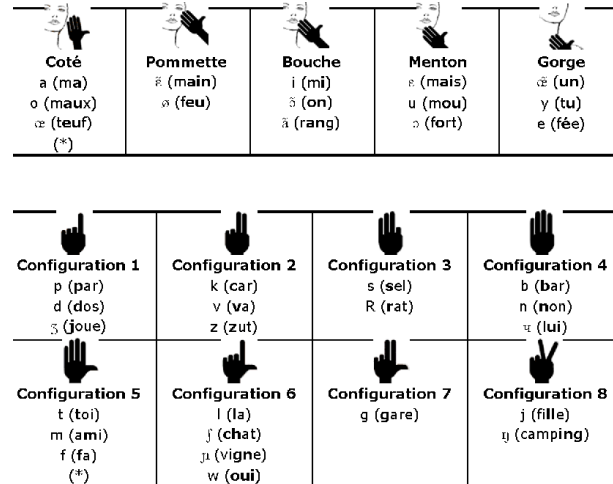
Figure 1: Hand position for vowels (top) and handshapes for consonants (bottom) in Cued Speech for French.

Alternative speech communication makes possible for individuals with impairments to communicate using their inborn abilities. A speech-impaired individual (after laryngectomy, for instance) can use gestures to transfer messages, and a blind-and-deaf individual can use touch speech modality to perceive speech. Alternative speech can be also used in education for pronunciation training for second language learners, and for perception and production rehabilitation of hearing impaired children. Since environmental noise does not affect alternative speech, it can be used when the communication occurs in noisy environments. In addition, it can be also used in situations when privacy in communication is preferable.

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. However, even with high lipreading performance, speech without knowledge of the semantic context can not be completely perceived. To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, in 1967 Cornett 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/, /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.

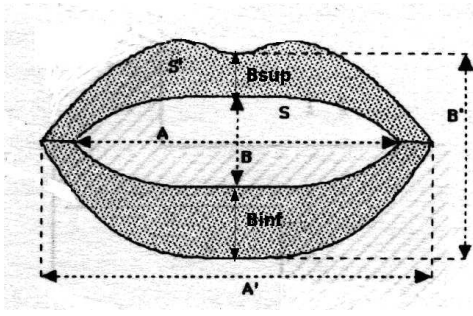Cued Speech uses handshapes placed in different po-

Figure 2: Parameters used for lip shape modeling.

sitions near the face in combination with natural speech lipreading to enhance speech perception from visual input. A manual cue in this system contains two components: the handshape and the hand position relative to the face. Hand-shapes distinguish consonants whereas hand positions distinguish vowels. A handshape together with a hand position cue a syllable. The advantage of Cued Speech is that improves speech perception to a large extent for hearing-impaired people. Moreover, Cued Speech offers a complete representation of the phonological system for hearing-impaired people exposed to this method since their youth, and therefore has a positive impact on language development [7]. Fig. 1 describes the complete system for French. In Cued French, eight handshapes in five positions are used.

The access to communication technologies has become essential for the handicapped people. The TELMA project (Phone for deaf people) aims to develop an automatic translation system of acoustic speech into visual speech completed with Cued Speech and vice versa, i.e. from Cued Speech components into auditory speech [8]. This project would enable deaf users to communicate between themselves and with normal-hearing people through the help of the autonomous terminal TELMA.

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 to a combined one. Previously, several studies have been made in automatic audiovisual recognition and integration of visual and audio modalities, aiming at improving the performance of a recognizer under noisy environments [9].

In the first attempt for vowel recognition in Cued Speech for French, in [10] a method based on separate identification, i.e., indirect decision fusion was used by the authors, and 77.6% vowel accuracy was obtained. A similar study for digit recognition in English Cued language was introduced in [11]. Previously, the authors presented a method for automatic vowel recognition in Cued Speech for French based on HMMs [12]. In the current study, vowel recognition is further investigated, and the proposed method was extended to also deal with consonant and isolated word recognition.

2. METHODS

The female native French speaker-employed for data recording was certified in transliteration speech into Cued Speech in the French language. She regularly cues in schools. The cuer wore a helmet to keep her head in a fixed position and opaque glasses to protect her eyes against glare from the halogen floodlight. The speaker'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 [13] for details).

The data were derived from a video recording of the speaker pronouncing and coding in Cued Speech a set of 262 French sentences. The sentences (composed of low predicted multi-syllabic words) were derived from a corpus that was dedicated to Cued Speech synthesis. Each sentence was dictated by an experimenter, and was repeated two- or three times (to correct errors in pronunciation of words) by the cuer resulting in a set of 638 sentences.

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

The process described here resulted in a set of temporally coherent signals: the 2-D hand information, the lip width (A), the lip aperture (B), and the lip area (S) values for both inner- and outer contours, and the corresponding acoustic signal. 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). Fig. 2 shows the lip shape parameters used in the current study.

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

For isolated word recognition experiments, 1450 isolated words were recorded employing a different cuer than that employed in the previous recordings. The vocabulary included 50 words, and each one was repeated twenty-nine times. Twenty repetitions of each word were used to train fifty, 6-state, whole word HMMs, and 9 repetitions were used for test. For lip shape and handshape modeling , the same kind of parameters as in vowel and consonant recognition experiments were used. In this data recording session, the recording constraints were eliminated by excluding the use of the helmet by the cuer.

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

Table 1: Phoneme-to-viseme mapping in the 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/ /ʀ/, /l/		

The French language includes 14 vowels and 17 consonants. Based on similarities on lips, the 31 phonemes can be grouped into 8 visemes. A viseme consists of phonemes that look similar on mouth/lips. Table 2 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 [14].

Because of the nature of Cued Speech, for vowel recognition lip shape and hand position elements were integrated into a single component using concatenative feature fusion. For consonant recognition, lip shape and handshape elements were integrated. The feature concatenation uses the concatenation of the synchronous lip shape and hand features as the joint bimodal feature vector

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

where O_t^{LH} is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand feature vector, and D the dimensionality of the joint feature vector. In 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 handshape stream was 30 (10 basic parameters, 10 Δ, and 10 $\Delta\Delta$ parameters). The dimension D of the joint lip-handshape feature vectors was, therefore 54.

3. EXPERIMENTS

3.1 Vowel-viseme classification using lip shape information

An experiment was conducted for recognition of the three vowel-visemes in order to evaluate the proposed method for

Table 2: 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

Table 3: Accuracy for three vowel groups considering similarities on lips.

	Element		
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

Table 4: Vowel and consonant overall accuracy.

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

lip shape modeling and also the vowel-viseme mapping. For each viseme, a 3-state HMM was trained using the previously described training data. The vowel labels in the phonetic transcription were translated to include only three viseme labels. Table 2 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 classification accuracy was achieved (i.e., percentage of diagonal elements). Table 2 also shows that the partial error corresponding to each viseme is uniform and almost equal.

3.2 Vowel recognition based on concatenative feature fusion

Based on concatenative feature fusion, experiments were conducted for vowel recognition considering the similarities on lips. The aim was to investigate how the integration of hand position element improves the recognition accuracy when vowels belonging to the same class (i.e., viseme) were recognized. As previously described, vowels which show similarities on lips (visemes) cannot be recognized accurately using lip shape information alone.

Using the previously described phoneme-to-viseme mapping, the 14 vowels were classified into three groups and three separate HMM sets were trained using the appropriate training data. In this way, each group included the most confusable vowels based on lip shape. Table 3 shows the obtained results. Using lip shape parameters only, the average vowel accuracy was 64.4% because of the high number of confusions between similar vowels in each group. On the other hand, by integrating also hand position element with lip shape element, the average vowel accuracy was raised to 91.2%, showing a 75.3% relative improvement compared with using lip shape parameters alone. In this case, however, vowels belonging to the same group -based on lips similarities- are distinguishable using hand position information. As a result the confusions between them drastically decreased, which increases the recognition accuracy. Table 3 also shows the results when auditory 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, in the case of Cued Speech the performance is very close to the

Table 5: Accuracy for five consonant groups considering similarities on lips.

Consonant group	Element		
	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

Figure 3: Word accuracy for isolated word recognition.

Figure 4: Word accuracy for isolated word recognition in the function of acoustic noise.

performance obtained when using the acoustic signal.

Table 4 shows the obtained results for the overall vowel recognition using a common HMM set. Compared with using lip shape parameters alone, when fusion was applied a 63.3% relative improvement was obtained. It is also shown, that results of vowel recognition in Cued Speech and those obtained using the acoustic signal do not show significant differences.

3.3 Consonant recognition based on concatenative feature fusion

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

In a way similar to vowel recognition, the consonants were classified into groups based on lip shape similarities, and separate HMM sets were trained. Five HMM sets were trained corresponding to the five consonant groups. Table 5 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, a 65.2% relative improvement was obtained when handshape element was also used. Table 5 also shows the results obtained when using the acoustic signal. It can be seen, that accuracies obtained using the acoustic signal are higher. A possible reason might be the errors occurred in handshape recognition. However, results are still comparable and promising also in consonant recognition in Cued Speech for French.

Table 4 shows the obtained overall results for consonant recognition. It is shown, that compared with using lip shape parameters alone, when fusion was applied a 56% relative improvement was obtained.

3.4 Isolated word recognition based on concatenative feature fusion

In this section, isolated word recognition experiments in Cued Speech for French are presented. Fig. 3 shows the results obtained in the function of several Gaussians per state. In the case of a single Gaussian per state, using lip shape alone a 53.3% word accuracy was obtained. When, however, handshape information was also used, a 95.8% word accuracy was obtained, showing a 91% relative improvement. The highest word accuracy when using lip shape was 74%,
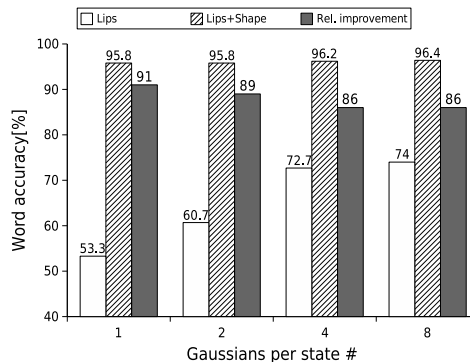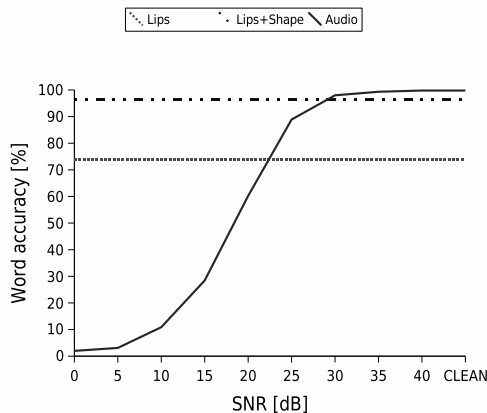
obtained in the case of using 8 Gaussians per state. In that case, the Cued Speech word accuracy using also hand information was 95.8%.

Fig. 4 shows a comparison between audio recognition and Cued Speech recognition in the function of several Signal-to-Noise (SNR) levels of acoustic noise. Office noise at different levels was superimposed onto clean test audio data and recognition experiments were conducted using clean HMMs. For the analysis of the audio signal, 12 MFCC along with the first and second derivatives were used. Each state was modeled with 8 Gaussian mixture components. It is shown, that up to 30 dB SNR, the Cued Speech recognition performs better compared with the audio recognition. The maximum word accuracy of audio recognition was 98% in the clean case, slightly higher compared with the 96.4% word accuracy of the Cued Speech recognition.

To show the effect of integrating hand information in isolated word recognition, PCA analysis was performed using the means of the fourth state of each word-HMM. Fig. 5 shows a visualization of the words in the 3-D PCA space using lip shape alone, and Fig. 6 shows the visualization of the same words when handshape was also integrated. It can be seen, that using also hand information the words are much better discriminated. As a result, the word accuracy was significantly increased.
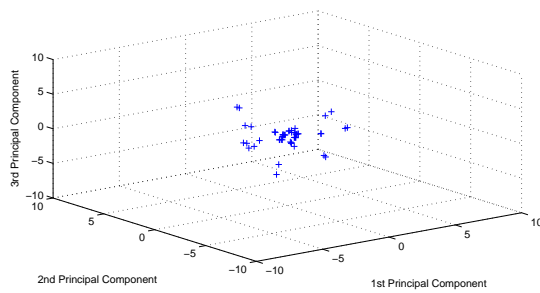
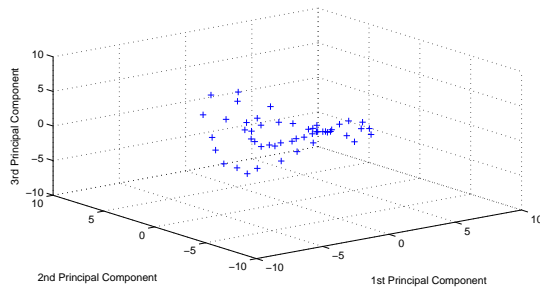Figure 5: Location of the words in the 3-D PCA space when using lip shape parameters

Figure 6: Location of the words in the 3-D PCA space when using fused lip shape and handshape parameters

4. CONCLUSION

In this paper, vowel, consonant, and isolated word recognition in Cued Speech for French was presented in order to investigate an alternative speech communication method for individuals with speech- or hearing impairments. . Using concatenative feature fusion, lip shape and hand information was integrated and automatic recognition experiments were conducted with very promising accuracies. The final aim of this study is to automatically recognize Cued Speech for French using visual information alone, and then convert it to a different modalities enabling human-human and human-machine interaction for people with disorders using their existing abilities. Currently, collection of additional data is in progress in order to realize automatic recognition for continuous Cued Speech recognition using larger vocabularies.

5. ACKNOWLEDGMENTS

REFERENCES

[1] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Non-audible murmur (nam) recognition using a stethoscopic nam microphone," *in Proceedings of Interspeech2004-ICSLP*, pp. 1469–1472, 2004.

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

[3] S.C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," *in Proceedings of Interspeech2006-ICSLP*, pp. 573–576, 2006.

[4] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data," *In Vth Conference on Articulated Motion and Deformable Objects (AMDO 2008, LNCS 5098) (F.J. Perales & R.B. Fisher, Eds.), Berlin, Heidelberg, Germany: Springer Verlag*, p. 132143, 2008.

[5] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Phone recognition from ultrasound and optical video sequences for a silent speech interface," *in Proceedings of Interspeech*, pp. 2032–2035, 2008.

[6] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Muller, V. Kunzmann, F. Losch, and G. Curio, "The berlin brain-computer interface: Eeg-based communication without subject training," *English IEEE Transcations on Neural Systems and Rehabilitation Engineering*, 2006.

[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] 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.

[10] N. Aboutabit, D. Beautemps, and L. Besacier, "Automatic identification of vowels in the cued speech context," *in Proceedings of AVSP'07*, 2007.

[11] S. Argyropoulos, D. Tzovaras, and M. G. Strintzis, "Multimodal fusion for cued speech language recognition," *in Proceedings of EUSIPCO*, pp. 1289–1293, 2007.

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

[13] N. Aboutabit, D. Beautemps, and L. Besacier, "Lips and hand modeling for recognition of the cued speech gestures: The french vowel case," *Speech Communication*, 2009, (to appear).

[14] 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.