CONTINUOUS PHONEME RECOGNITION IN CUED SPEECH FOR FRENCH

Panikos Heracleous[1], *Denis Beautemps*[2], *and Norihiro Hagita*[1]

[1]ATR, Intelligent Robotics and Communication Laboratories, Japan
[2]GIPSA-lab, Speech and Cognition Department, UMR 5216, CNRS-Grenoble University, France

E-mail:panikos@atr.jp

ABSTRACT

Cued Speech is a visual communication mode, which uses hand shapes and lip shapes making all the sounds of spoken language clearly understandable to deaf and hearing-impaired people. Using Cued Speech the problems of lipreading can be overcome resulting thus in understanding of full spoken language by deaf children and adults. In automatic recognition of Cued Speech, lip shape recognition, gesture recognition, and integration of the two modalities are required. Previously, the authors have reported studies on vowel-, consonant-, and isolated word recognition in Cued Speech for French. In the current study, continuous phoneme recognition experiments are presented using data from a normal-hearing and a deaf cuer. In the case of the normal-hearing cuer, the obtained phoneme correct was 82.9%, and in the case of the deaf cuer 81.5%. The results showed, that automatic recognition of Cued Speech shows similar performance in both normal-hearing and deaf cuers.

Index Terms— Cued Speech, hidden Markov models, fusion, phoneme recognition

1. INTRODUCTION

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

Even with high lipreading performances speech cannot be thoroughly perceived without knowledge about the semantic context. To date, the best lip readers are far way of reaching perfection. On average, only 40 to 60% of the vowels of a given language (American English) are recognized by lipreading [1], and 32% when relating to low predicted words [2]. The best results obtained amongst deaf participants was 43.6% for the average accuracy [3]. 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.

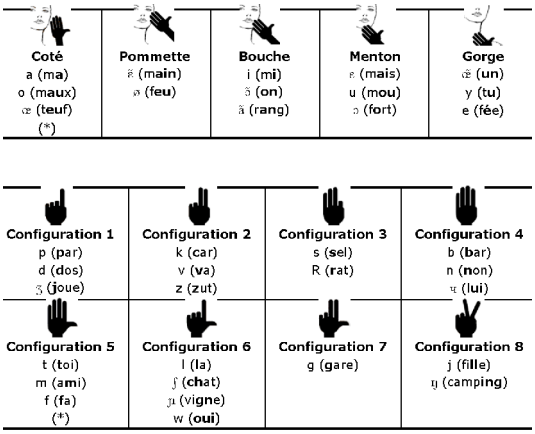

Fig. 1. hand shapes for consonants (top) and hand position (bottom) for vowels in French Cued Speech.

To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, in 1967 Cornett [4] developed the Cued Speech system to complement the lip information and make all phonemes of a spoken language clearly visible. Because many sounds look identical on lips/face (e.g., /*p*/, /*b*/, and /*m*/), using information from hand those sounds can be distinguished. As a result, deaf people can understand a particular language without any sound being necessary, but using visual information only.

Cued Speech (also referred to as Cued Language [5]) uses hand shapes placed in different positions near the face. These hand shapes are combined with lip shapes in order to produce and perceive speech from visual input alone. This is a system where the speaker faces the perceiver and moves his hand in close relation with speech. The hand is held flat and oriented so that the back of the hand faces the perceiver. When the

This work was supported by KAKENHI (21118001)

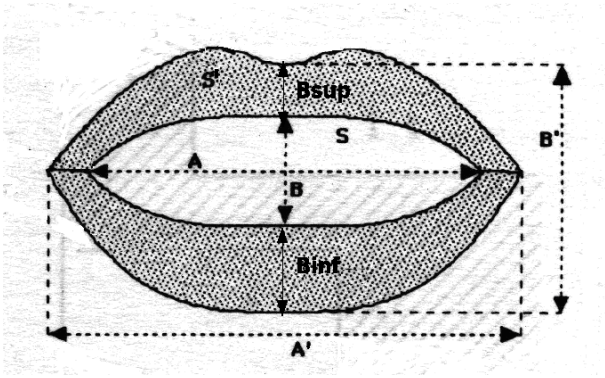

Fig. 2. *Parameters used for lips shape modeling.*

hand is associated with a particular lip shape corresponds to a phoneme. A manual cue in Cued speech system contains the hand shape component and the hand position relative to the face. hand shapes distinguish among consonant phonemes whereas hand positions distinguish among vowel phonemes. A syllable is cued by a hand shape together with a hand position.

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. The Cued Speech for French consisted of eight hand shapes in five positions,and was adapted from American English to French in 1977.

Another widely used communication method for deaf individuals is the Sign Language [8, 9]. 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 for 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 more easily communicate with hearing family members who cue.

Previously, the authors presented vowel- [10], consonant- [11], and isolated word recognition [12] in Cued Speech for French. In the current study, continuous phoneme recognition experiments are introduced using data from a normal-hearing and a deaf cuer. Continuous phoneme recognition is a more complex recognition task, and high recognition rates will be an evidence that continuous Cued Speech recognition with a larger vocabulary is possible. Also, this study aims at investigating the differences between normal-hearing and deaf cuers concerning automatic Cued Speech recognition, and to examine the possible cuer variability.

2. METHODS

2.1. Data and feature extraction

In the data recording, a deaf and a normal-hearing female cuers were employed. The normal-hearing cuer regularly cues in several schools, and was certified in transliteration speech into Cued Speech in the French language. The deaf speaker was also speech-impaired, and it was very difficult to understand his speech. She uses Cued Speech to mainly communicate with her family members and other Cued Speech users.

A camera with a zoom facility used to shoot the hand and face was connected to a betacam recorder. The speakers' lips were painted blue, and color marks were placed on the speakers' fingers. These constraints were applied in recordings in order to control the data and facilitate the extraction of accurate features. The data were derived from a video recording of the cuers pronouncing and coding in Cued Speech a set of 50 French isolated words and short phrases, each one repeated 29 times.

In previous studies, the tracking of hand positions and hand shapes was based on a video processing technique based on blue color [10, 11]. In this study, however, the tracking method was modified and improved by using landmarks with different colours placed on the fingers. The new method resulted in a faster and more accurate image processing stage.

The audio part of the video and the image were recorded in synchronously. 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). 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 shape statistical modeling the coordinates of the landmarks placed on the fingers were used (i.e., 10 parameters). The lip shape parameters used in the current study are shown in the Figure 2.

2.2. Concatenative feature fusion

Cued Speech automatic recognition requires combined automatic recognition of lip shape and hand shape. To avoid the deterministic hand shape recognition which may cause unrecoverable errors in image processing, the proposed method is based on the tracking and extraction of the xy coordinates each time frame, and on the use those values as features in the

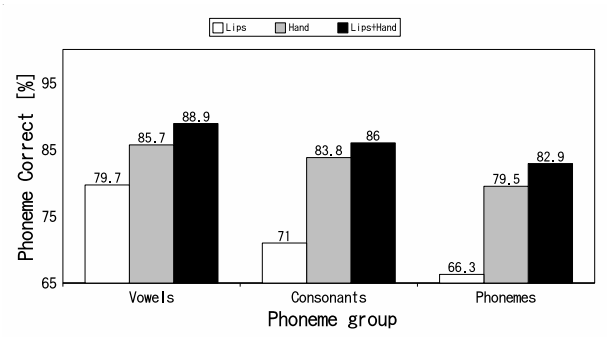

Fig. 3. Cued Speech recognition (Normal-hearing)

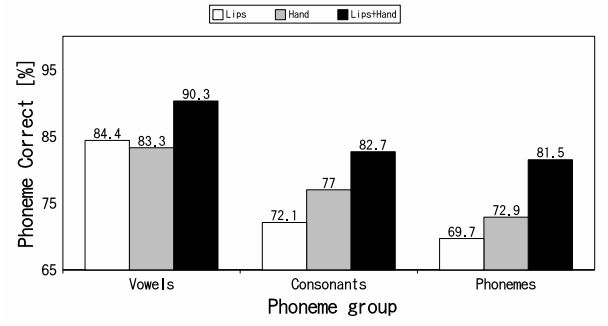

Fig. 4. Cued Speech recognition (Deaf)

HMM modeling. Feature concatenation was used to integrate the lip shape and hand shape components [13]. 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. The lip shape feature vectors were of length 24, the hand shape feature vectors were of length 30, and the joint lip shape-hand shape feature vectors we of length 54.

2.3. Statistical modeling

The statistical models were thirty-one context-independent, 3-state, left-to-right with no skip monophone HMMs. For modeling each state, a mixture of 16 distributions was used, respectively. For training and test 5,294 and 5,264 phones were used. For training, 2,248 vowel and 3,046 consonant instances were used. For testing, 2,233 vowel and 3,031 consonant instances.

Usually, in automatic speech recognition, a diagonal covariance matrix is used. This is because of the assumption

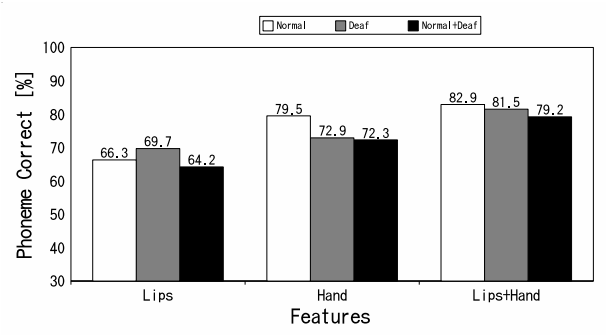

Fig. 5. Comparison between results of normal-hearing and deaf cuers.

that the features are uncorrelated. In automatic lipreading, however features 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 project into the PCA space, and all PCA lip shape components were used for statistical model (i.e., HMMs) training. For training and recognition the HTK3.1 toolkit was used.

3. RESULTS

Figure 3 shows the phoneme correct (i.e., deletions and substitutions are considered) in the case of lip shape, hand shape, and Cued speech recognition for the normal-hearing cuer. The results show, that using fusion to integrate the hand component with the lip shape component, the accuracy was drastically increased. Specifically, a vowel correctness of 88.9%, a consonant correctness of 86%, and a phoneme correctness of 82.9% were obtained. The results also show, that for vowel recognition a higher recognition rate was achieved compared to the consonant recognition. Many of the consonants have limited visual information on lips (e.g., $/k/$, $/g/$) resulting in more confusions between them.

The results achieved using data from the deaf cuer are shown in the Figure 4. As is shown, also in this case the performance significantly increased, when lip shape and hand shape components were integrated. In the case of the deaf cuer, the vowel correctness was 90.3%, the consonant correctness 82.7%, and the phoneme correctness 81.5%.

Figure 5 shows a comparison between the results obtained using data from the deaf and the normal-hearing cuers. As it is shown, the obtained results are very closely comparable. In the case of hand shape recognition, the normal-hearing cuer shows a higher performance. A possible reason may be the fact, that the normal-hearing cuer was a professional teacher of Cued Speech. The deaf cuer, however, shows a higher performance in automatic lip shape recognition. The

Table 1. *Phoneme correct for a multi-cuer experiment.*

Test set	HMMs		
	Deaf	Normal	Deaf + Normal
Deaf	81.5	-	77.0
Normal	-	82.9	79.3
Deaf + Normal	-	-	78.2

fact that the deafs rely on lipreading for speech communication might increase their ability not only for speech perception, but also for speech speech production by lips/face. In the case of the deaf cuer, higher lip shape recognition rate was achieved, even though the deaf cuer was also speech-impaired and the intelligibility of her speech was very low. Also, Cued Speech automatic recognition achieved very similar phoneme rates in both cuers. Further analysis and investigations of the deafs speech production mechanism appear to be necessary in order to better understand and explain these observations.

Another experiment was conducted using a common HMM test trained with data from both the normal-hearing and the deaf cuers. Table 1 shows the achieved results. It is shown, that when a common HMM set was used, the performance in both cases did not decrease drastically. The results obtained provide an indication, that HMMs can capture the variability of different cuers. Multi-cuer Cued Speech recognition should, therefore, be possible, facing similar difficulties as in audio automatic speech recognition.

4. CONCLUSIONS AND FUTURE WORK

In the current study, unconstrained phoneme recognition in Cued Speech for French is presented. Cuer-dependent experiments were conducted using data from a deaf and a normal-hearing cuer with promising results. In the case of the deaf cuer an 81.5%, and in the case of the normal-hearing cuer an 82.9% phoneme correct were obtained. Additional experiment was also conducted using a common HMM test trained with data from both the normal-hearing and the deaf cuers. The multi-cuer performance indicates that HMMs can capture the variability in different cuers, and, therefore, training accurate cuer-independent HMMs should be possible.

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

[5] E. Fleetwood and 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*, 1998.

[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] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," *In Proceedings of Interspeech*, pp. 2513–2516, 2007.

[9] S. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Trans. PAMI*, vol. vol. 27, no. 6, pp. 873891, 2005.

[10] P. Heracleous, N. Aboutabit, and D. Beautemps, "Lip shape and hand location fusion for vowel recognition in cued speech for french," *IEEE Signal Processing Letters*, vol. 16, Issue 5, pp. 339–342, 2009.

[11] P. Heracleous, N. Aboutabit, and D. Beautemps, "Vowel and consonant automatic recognition in cued speech for french," *in Proceedings of IEEE VECIMS'09*, pp. 33–37, 2009.

[12] P. Heracleous, D. Beautemps, and N. Aboutabit, "Cued speech recognition for augmentative communication in normal-hearing and hearing-impaired subjects," *in Proceedings of Interspeech'09*, pp. 1383–1386, 2009.

[13] 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.