Cued Speech: A visual communication mode for the Deaf society

Panikos Heracleous, Denis Beautemps

Cued Speech: A visual communication mode for the deaf society

Panikos Heracleous[1,2] and Denis Beautemps[1]
[1] *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
[2] *ATR, Intelligent Robotics and Communication Laboratories*

E-mail: panikos@atr.jp, Denis.Beautemps@gipsa-lab.grenoble-inp.fr

Abstract: 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 individuals. The aim of Cued Speech is to overcome the problems of lip reading and thus enable deaf persons to wholly understand spoken language. In this study, automatic phoneme recognition in Cued Speech for French based on hidden Markov model (HMMs) is introduced. The phoneme correct for a normal-hearing cuer was 82.9%, and for a deaf 81.5%. The results also showed, that creating cuer-independent HMMs should not face any specific difficulties, other than those occured in audio speech recognition.

References

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *in Proc. of the IEEE*, vol. 91, Issue 9, pp. 1306–1326, 2003.

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

[3] 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. 31(1), pp.20–41, 1994.

[4] 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 15, pp. 339-342, 2009.

[5] P. Heracleous, N. Aboutabit, and D. Beautemps, "Vowel and Consonant Automatic Recognition in Cued Speech for French", *in Proc. of IEEE VECIMS'09*, pp. 33-37, 2009.

[6] P. Heracleous, D. Beautemps, and N. Aboutabit, "Cued Speech Recognition for Augmentative Communication in Normal-hearing and Hearing-impaired Subjects", *in Proc. of Interspeech2009*, pp. 1383-1386, 2009.

[7] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech

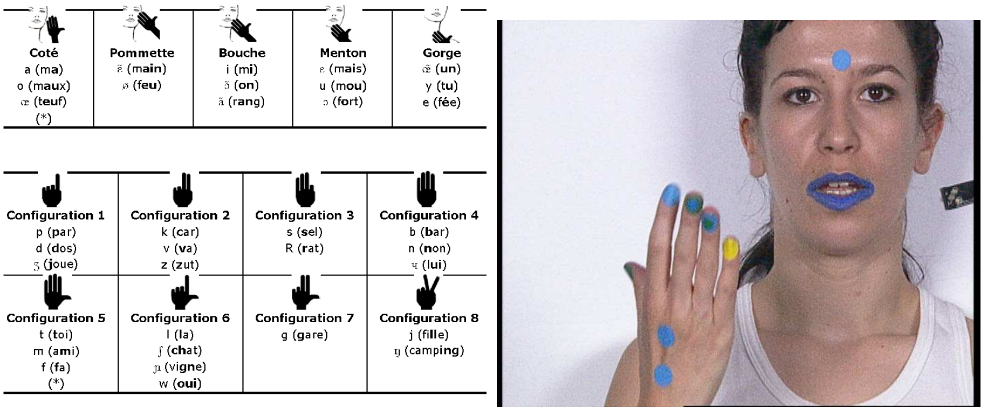
Recognition Techniques for a Sign Language Recognition System" *in Pro. of Interspeech*, pp. 2513-2516, 2007.

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

Keywords: French Cued Speech, HMM, vowel recognition, consonant recognition, concatenative feature fusion.

1 Introduction

To date, visual information has been widely used to improve speech perception, or automatic speech recognition (lipreading) [1]. 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 [2]. 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.



Coté	Pommette	Bouche	Menton	Gorge
a (ma) o (maux) œ (teuf) (*)	ɛ̃ (main) ø (feu)	i (mi) ɔ̃ (on) ɑ̃ (rang)	ɛ (mais) u (mou) ɔ (fort)	œ̃ (un) y (tu) e (fée)

Configuration 1	Configuration 2	Configuration 3	Configuration 4
p (par) d (dos) ʒ (joue)	k (car) v (va) z (zut)	s (sel) R (rat)	b (bar) n (non) ɥ (lui)
Configuration 5	**Configuration 6**	**Configuration 7**	**Configuration 8**
t (toi) m (ami) f (fa) (*)	l (la) ʃ (chat) ɲ (vigne) w (oui)	g (gare)	j (fille) ŋ (camping)

(a) (b)

Fig. 1: (a) Hand positions for vowels (top) and handshapes for consonants (bottom). (b) A normal-hearing cuer and the colored landmarks used for features extraction.

Cued Speech uses handshapes placed in different positions 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. Handshapes 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 [3]. Fig. **1**a describes the complete system for French. In Cued French, eight handshapes in five positions are used.

Another widely used communication method for deaf individuals is the Sign Language [7, 8]. 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.

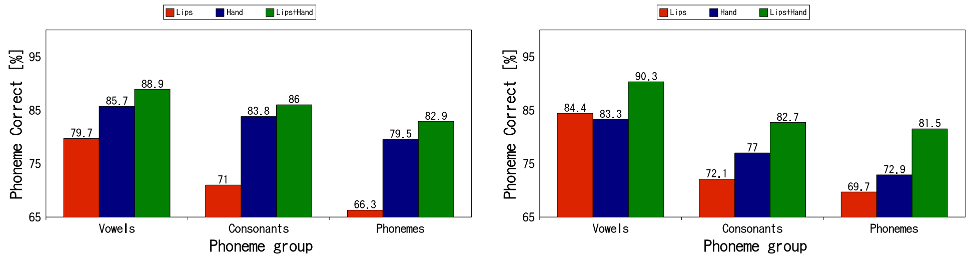
Previously the authors presented vowel- [4], consonant- [5] and isolated word recognition [6] in Cued Speech for French based on HMMs . In the current study, continuous phoneme recognition is introduced using data from a deaf and normal-hearing cuer. The aim is to investigate the possible differences between normal-hearing and deaf cuer concerning the Cued Speech automatic recognition. Also, the authors are interested in further improving the system to also deal with continuous Cued Speech recognition.

2 Methods

In the data recording, a deaf and a normal-hearing female cuers were employed. The normal-hearing cuer was certified in transliteration speech into Cued Speech in the French language. She regularly cues in schools. The deaf speaker, who was also speech-impaired, uses Cued Speech to communicate with her family's members.

A camera with a zoom facility used to shoot the hand and face was connected to a betacam recorder. The cuers' lips were painted blue, and color marks were placed on the cuers' 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 (i.e., 1450 words in total).

In previous studies (e.g., [4, 5] the authors used a video processing technique based on blue color in order to track the hand positions and handshapes. In this study, landmarks with different colors were placed on the fingers resulting in a faster and more accurate image processing stage. 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). 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 handshape modeling the xy coordinates of the landmarks placed on the fingers were used (i.e., 10 parameters).



(a) (b)

Fig. 2: Phoneme correct in the case of a normal-hearing (a) and a deaf (b) cuer

In Cued Speech recognition, lip shape and handshape joint recognition is required. To avoid the deterministic handshape recognition which may cause unrecoverable errors in image processing stage, the proposed method tracks and extracts the xy coordinates of the landmars at each time frame, and uses those values as features in the HMM modeling. Feature concatenation was used to integrate the lip shape and handshape components [1]. The feature concatenation uses the concatenation of the synchronous lip shape and hand features as the joint feature vector

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

where O_t^{LH} is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand feature vector, and D the dimensionality of the joint feature vector. The dimension of the lip shape stream was 24 (8 basic parameters, 8 Δ, and 8 $\Delta\Delta$ parameters). 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-hand shape feature vectors was, therefore 54.

Thirty-one context-independent, 3-state, left-to-right with no skip monophone HMMs were used. The observations in each state were modeled with a mixture of 16 Gaussians. For training and test 5294 and 5264 phones were used, respectively. The training set contained 2248 vowel and 3046 consonant instances. The test set contained 2233 vowel and 3031 consonant instances. In the experiments, no language model was used.

3 Experimental Results

Figure **2**a shows the phoneme correct (i.e., deletions and substitutions were considered) in the case of lip shape, handshape, and Cued speech recognition for the normal-hearing cuer. It is shown, that when hand component was also fused with the lip shape component, the accuracy was significantly 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 vowel recognition performs better compared with the consonant recognition. A possible explanation might be the lower lip shape recognition in the case of the consonants, because many of the consonants have limited visual information on lips (e.g., /k/, /g/). Figure **2**b shows the results obtained when using data from the deaf cuer. As it is shown, also in this case the performance significantly increased, when lip shape and handshape 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%.

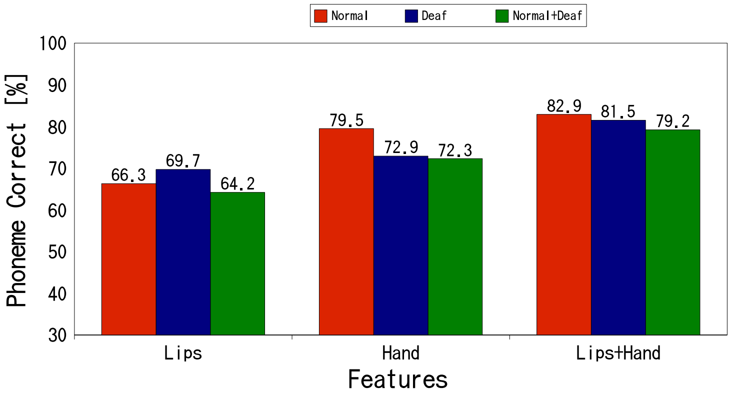


Fig. 3: Comparison between results of normal-hearing and deaf cuers.

Figure **3** shows a comparison between the results obtained using data from the deaf and the normal-hearing cuers. It is shown, that 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. On the other hand, the deaf cuer shows a higher performance in lip shape recognition. The 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. Note, that the deaf cuer was also speech-impaired and the intelligibility of her speech was very low. Even though, her lip shape automatic recognition showed higher performance compared to the normal-hearing cuer. Also, Cued Speech automatic recognition achieved very similar phoneme rates in both cuers. Further analysis of the deafs speech production mechanism appear to be necessary in order to better understand and explain these observations.

In addition to cuer-dependent experiments, an experiment was conducted using a common HMM set trained with data from both the normal-hearing and the deaf cuers. Using the common HMM set and normal-hearing test data, a 79.3% phoneme correct was obtained. When deaf test data were used, a 77% phoneme correct was achieved. The results provide an indication, that HMMs can capture the variability of different cuers. Multi-cuer Cued Speech recognition should, therefore, be possible, facing in fact similar difficulties as in audio automatic speech recognition.

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 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. Although the results are promising, problems still persist. Namely, 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 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.

5 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. In addition to cuer-dependent experiments, an experiment using a common HMM set was also conducted. The multi-cuer performance indicates that HMMs can capture the variability in different cuers, and, therefore, training accurate cuer-independent HMMs should be possible. Currently, additional data collection is in progress in order to build a cuer-independent continuous Cued Speech recognition system.

6 Acknowledgments

The authors would like to thank the volunteer cuers 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.