INTERSPEECH 2009 BRIGHTON

Cued Speech Recognition for Augmentative Communication in Normal-hearing and Hearing-impaired Subjects

Panikos Heracleous, Denis Beautemps, and Noureddine Abboutabit

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

{Panikos.Heracleous, Denis.Beautemps, Noureddine.Aboutabit}@gipsa-lab.grenoble-inp.fr

10.21437/Interspeech.2009-427

Abstract

Speech is the most natural communication mean for humans. However, in situations where audio speech is not available or cannot be perceived because of disabilities or adverse environmental conditions, people may resort to alternative methods such as augmented speech. Augmented speech is audio speech supplemented or replaced by other modalities, such as audiovisual speech, or Cued Speech. Cued Speech is a visual communication mode, which uses lipreading and handshapes placed in different position to make spoken language wholly understandable to deaf individuals. The current study reports the authors' activities and progress in Cued Speech recognition for French. Previously, the authors have reported experimental results for vowel- and consonant recognition in Cued Speech for French in the case of a normal-hearing subject. The study has been extended by also employing a deaf cuer, and both cuer-dependent and multi-cuer experiments based on hidden Markov models (HMM) have been conducted.

Index Terms: Cued Speech, hidden Markov models, automatic recognition

1. Introduction

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

Even with high lipreading performances, speech cannot be thoroughly perceived without knowledge about the semantic context. To date, the best lipreaders 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 result obtained amongst deaf participants was 43.6% for the average accuracy [3, 4]. The main reason for this lies in the ambiguity of the visual pattern. However, as far as the orally educated deaf people are concerned, the act of lipreading remains the main modality of perceiving speech.

To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, in 1967 Cornett [5] developed the Cued Speech system to complement the lip information and make all phonemes of a spoken language clearly visible. As many sounds look identical on lips (e.g., /p/, /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.

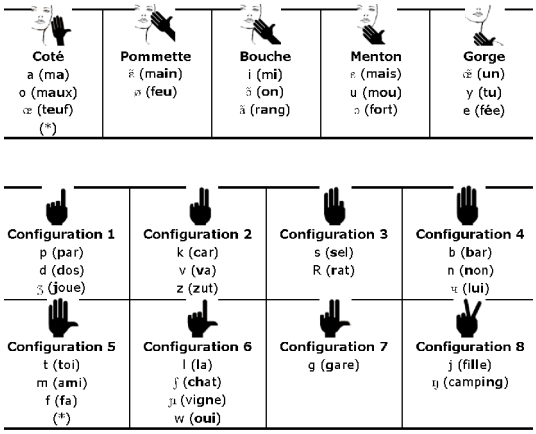

Figure 1: *Hand positions for vowels (top) and handshapes for consonants (bottom) in Cued Speech for French*

Cued Speech (also referred to as Cued Language [6]) uses handshapes placed in different positions near the face in combination with natural speech lipreading to enhance speech perception from visual input. This is a system where the speaker faces the perceiver and moves his hand in close relation with speech. The hand, held flat and oriented so that the back of the hand faces the perceiver, is a cue that corresponds to a unique phoneme when associated with a particular lip shape. A manual cue in this system contains two components: the handshape and the hand position relative to the face. Handshapes distinguish among consonant phonemes whereas hand positions distinguish among vowel phonemes. A handshape together with a hand position cue a syllable.

Cued Speech improves speech perception for deaf people [2, 7]. Moreover, for those who have been exposed to this method since their youth, offers a thorough representation of the phonological system, and therefore it has a positive impact on language development [8]. Fig. 1 describes the complete system for French. In Cued Speech for French, eight handshapes in five positions are used. The system was adapted from American English to French in 1977. To date, Cued Speech is adapted to more than 60 languages.

Another widely used communication method for deaf individuals is the Sign Language [9, 10]. Sign Language is a language with its own grammar, syntax and community; however,

(a) (b)

Figure 2: *(a) Photo of the normal-hearing cuer; (b) photo of the deaf cuer*

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.

Access to communication technologies has become essential for handicapped people. The current study is a part of the TELMA project (Phone for deaf people) aiming at developing an automatic translation system of acoustic speech into visual speech completed with Cued Speech and vice versa, i.e. from Cued Speech components into auditory speech [11]. This project would enable deaf users to communicate with each others and with normal-hearing people through the help of the autonomous terminal TELMA. In this context, the automatic translation of Cued Speech components into a phonetic chain is a key issue. The Cued Speech system allows both hand and lip flows to convey a part of the phonetic information. Therefore, in order to recover the complete phonetic and lexical information, lip and hand components should be used jointly.

Previously, the authors reported experimental results on vowel- and consonant recognition in Cued Speech for French based on HMMs [12, 13], and using fusion [14, 15] to integrate the lip shape and handshape components. In the case of a normal-hearing cuer, a vowel accuracy of 87.6% and a consonant accuracy of 79.6% were achieved. In the current

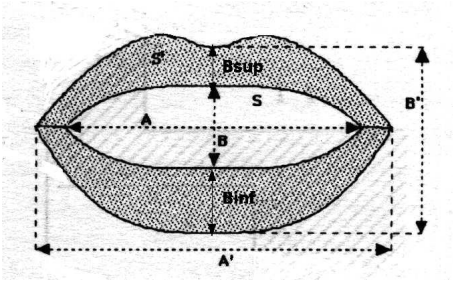

Figure 3: *Parameters used for lips shape modeling.*

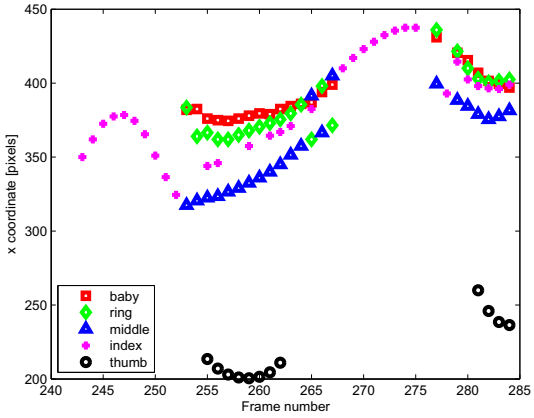

Figure 4: *Finger x-coordinates during the /VINGTHUIT/ word production by the deaf cuer.*

study, a deaf cuer was also employed, and both cuer-dependent and multi-cuer isolated word recognition experiments in Cued Speech were conducted.

2. Methodology

2.1. Cued Speech Materials

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 -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 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. Fig. 2 shows the photos of the two cuers, and also the landmarks used in feature extraction. The data were derived from a video recording of the cuers pronouncing and coding in Cued Speech a set of 50 French isolated words, each one repeated 29 times. In the case of the deaf cuer, the vocabulary was extended up to 100 words by additionally recording another 50 words. Each new word was repeated 15 times.

An automatic image processing method (see [16] for details) 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 automatic process 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. 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 the two landmarks placed on the hand were used (i.e., 4 parameters). For handshape modeling the xy coordinates of the landmarks placed on the fingers were used (i.e., 10 parameters). Fig. 3 shows the lip shape parameters used in the current study, and Fig. 4 shows the x-coordinates of the fingers dur-

ing the $/VINGTHUIT/$ word production, as tracked by the automatic image processing system.

2.2. Lip shape and handshape modeling

In the experiments, context-independent whole-word models were used. A 6-state, left-to-right with no skip HMM topology was used. Each state was modeled with 4 Gaussian distributions. In addition to the basic lip and hand parameters, the first (Δ) and second derivatives ($\Delta\Delta$) were also used. For training and test 750 and 700 words were used, 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, Principal Component Analysis (PCA) was applied to decorrelate the lip shape parameters, and then a diagonal covariance matrix was used. All 24 PCA lip shape components were used for HMM training. For training and recognition the HTK3.1 toolkit was used.

2.3. Concatenative feature fusion

The feature concatenation uses the concatenation of the synchronous lip shape and hand position features as the joint bimodal feature vector

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

where O_t^{LH} is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand feature vector, and D the dimensionality of the joint feature vector. In these experiments, the dimension of the lip shape stream was 24 (8 static parameters, 8 Δ, and 8 $\Delta\Delta$ parameters). The dimension of the handshape stream was 30, and the D dimension was, therefore, 54.

2.4. Multistream 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 [17] and in audio-visual speech recognition [18]. The emission likelihood of multistream 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 multistream 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. 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. In these experiments, the weights were adjusted to 0.4 and 0.6 values, respectively. The selected weights were obtained experimentally by maximizing the accuracy on a held-out data.

3. Experiments

In this section, cuer-dependent and multi-cuer isolated word recognition experiments in both normal-hearing and deaf subjects are presented.

3.1. Automatic recognition in the normal-hearing subject

Table 1 shows the results obtained in the case of the normal-hearing cuer. The recognition rate when using lip parameters was 69.8%, and 90.7% when using handshape parameters. When lip shape and handshape components where integrated, a 94.8% recognition rate was obtained. The recognition rate was raised to 95.2% when multistream HMM decision fusion was applied. The obtained results are promising, and show the effectiveness of integrating handshape and lip shape in order to realize automatic recognition of Cued Speech.

Table 1: *Recognition rates for a 50-word vocabulary in the case of the normal-hearing cuer.*

Fusion	**Component**		
	Lips	Hand	Lips + Hand
Feature	69.8	90.7	94.8
Multistream	69.8	90.7	95.2

3.2. Automatic recognition in the deaf subject

Table 2 shows the results obtained in the case of the deaf cuer. In this experiment, the same vocabulary as in the normal-hearing cuer's case was used. The recognition rate was 76.1% when using lip shape parameters, and 76.3% when handshape parameters were used. When lip shape and handshape components were fused, the recognition rate was 89% in the case of using feature fusion, and 92% in the case of multistream HMM decision fusion, respectively. The results show a variability between the two cuers. Concerning the differences in handshape recognition, a possible reason might be the fact that the normal-hearing subject is a professional teacher of Cued Speech, and therefore cues more accurately. However, the recognition rates when fusion was applied are still closely comparable.

Table 2: *Recognition rates for a 50-word vocabulary in the case of the deaf cuer.*

Fusion	**Component**		
	Lips	Hand	Lips + Hand
Feature	76.1	76.3	89.0
Multistream	76.1	76.3	92.0

Table 3 shows the results obtained, when a 100-word vocabulary was used. By increasing the vocabulary, the recognition rates remained almost unchanged. In fact, a modest increase in the recognition rate can be observed. It should be noted, however, that in this pilot study the vocabularies used are still small. It may be happened, that in the case of larger vocabularies significant differences will be obtained.

Table 3: *Recognition rates for a 100-word vocabulary in the case of the deaf cuer.*

Fusion	Component		
	Lips	Hand	Lips + Hand
Feature	68.4	77.0	89.9
Multistream	68.4	77.0	92.9

3.3. Multi-cuer automatic recognition

To investigate whether it is possible to train cuer-independent HMMs using a large number of subjects, an experiment based on concatenative feature fusion was conducted using data from both the normal-hearing and the deaf subjects. The vocabulary was 100 words, and for training and test 1900 and 1750 words were used, respectively. Table 4 shows the achieved results. When using multi-cuer HMMs, the recognition rate for the deaf cuer was 88.6%, slightly lower as compared to the 89.9% recognition rate when deaf HMMs were used. In the case of the normal-hearing cuer, the recognition rates remained equal. The results obtained indicate that there should not be particular difficulties in creating cuer-independent HMMs using a larger number of subjects.

Table 4: *Recognition rates for a multi-speaker experiment using a 100-word vocabulary.*

Test set	HMMs		
	Deaf	Normal	Deaf + Normal
Deaf	89.9	-	88.6
Normal	-	94.8	94.8
Deaf + Normal	-	-	91.1

4. Conclusions

In this study, isolated word recognition experiments in Cued Speech for French in both normal-hearing and deaf subjects were presented. Using feature fusion and multistream HMM decision fusion, lip shape and handshape components were integrated into a combined component, and automatic recognition experiments were conducted. In the case of a normal-hearing cuer, a 95.2% recognition rate, and in the case of a deaf cuer, a recognition rate of 92.0% was obtained. In addition to cuer-dependent experiments, a multi-cuer experiment was also conducted showing a 91.1% recognition rate. The results obtained are promising, and show the effectiveness of the proposed methods in the automatic recognition of Cued Speech for French using visual information alone. Currently, data recording from additional cuers is in progress in order to evaluate the proposed methods using a larger number of subjects in the framework of the TELMA project.

5. 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. This work is supported by the TELMA project (ANR, 2005 edition).

6. References

[1] A. A. Montgomery and P. L. Jackson, "Physical characteristics of the lips underlying vowel lipreading performance," *Journal of the Acoustical Society of America*, vol. 73 (6), pp. 2134–2144, 1983.

[2] G. Nicholls and D. Ling, "Cued speech and the reception of spoken language," *Journal of Speech and Hearing Research*, vol. 25, pp. 262–269, 1982.

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

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] S. Nakamura, K. Kumatani, and S. Tamura, "Multi-modal temporal asynchronicity modeling by product hmms for robust," *in Proceedings of Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)*, p. 305, 2002.

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

[16] N. Aboutabit, D. Beautemps, and L. Besacier, "Lips and hand modeling for recognition of the cued speech gestures: The french vowel case," *Speech Communication*, 2008, (accepted with revision).

[17] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," *in Proceedings of International Conference on Spoken Language Processing*, pp. 426–429, 1996.

[18] 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.