Lip shape and hand position fusion for automatic vowel recognition in Cued Speech for French
Panikos Heracleous, Noureddine Aboutabt, Denis Beutemps

To cite this version:

HAL Id: hal-00346166
https://hal.archives-ouvertes.fr/hal-00346166v2
Submitted on 9 Feb 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Lip shape and hand position fusion for automatic vowel recognition in Cued Speech for French
Panikos Heracleous, Noureddine Aboutabit, and Denis Beutemps

Abstract—Cued Speech is a sound-based system, which uses handshapes in different positions and in combination with lip-patterns of speech, and makes all the sounds of spoken language clearly understandable to deaf and hearing-impaired people. The aim of Cued Speech is to overcome the problems of lip reading and thus enable deaf children and adults to wholly understand spoken language. Cued Speech recognition requires gesture recognition and lip shape recognition. In addition, the integration of the two components is of great importance. This article presents hidden Markov model (HMM)-based vowel recognition as used in Cued Speech for French. Based on concatenative feature fusion and multi-stream HMM decision fusion, lip shape and hand position components were integrated into a single component, and automatic vowel recognition was realized. In the case of multi-stream HMM decision fusion, the obtained vowel classification accuracy using lip shape and hand position information was 87.6%, showing absolute improvement of 19.6% in comparison with a use restricted only to lip parameters. The results achieved show the effectiveness of the proposed approaches to Cued Speech recognition.

Index Terms—Cued Speech, HMM, vowel recognition, concatenative fusion, multi-stream HMM fusion.

I. INTRODUCTION

To date, in many studies dealing with speech perception or speech recognition, visual information has also been used to complement the audio information (lip reading). The visual pattern, however, is ambiguous, and on an average, only 40 to 60% of the vowels are recognized by the lip reading system for a given language (American English) [1], and the accuracies are down to 10 to 30% where words are concerned [2]. However, for the orally educated deaf or hearing-impaired people, lip reading remains the main modality of perceiving speech. Therefore, in 1967 Cornett developed in 1967 the Cued Speech system as a supplement to lip reading [3]. Cued Speech improves speech perception for the hearing-impaired people [4], and for those who are exposed to this method since their youth, offers them a complete representation of the phonological system, and thereby has exerts a positive impact on the language development [5].

Cued Speech is a visual communication system that uses handshapes placed in different positions near the face in combination with natural speech lip reading to enhance speech perception with visual input. Here, two components make up a manual cue: the handshape and the hand position relative to the face. Handshapes are designed to distinguish consonant phonemes whereas hand positions are used to distinguish vowel phonemes. Fig. 1 describes the complete system relating to the French language, which uses eight handshapes in five different positions [6].

Access to communication technologies has become essential for handicapped people. The TELMA project (Phone for deaf people) aims at developing an automatic translation system of acoustic speech into visual speech complete with Cued Speech and vice versa, i.e. from Cued Speech components into auditory speech [7]. This project would enable deaf users to communicate with each other and with normal-hearing people with the help of the autonomous terminal TELMA. In this system, the automatic translation of Cued Speech components into a phonetic chain is the key. The Cued Speech system allows both hand and lip ows to convey a part of the phonetic information. This means that, in order to recover the complete phonetic- and lexical information, lip- and hand components should be used jointly.

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 into a combined stream. There were several studies in the past relating to automatic audio-visual recognition and integration of visual- and audio modalities [8]. The aim of audio-visual speech recognition is to improve the performance of a recognizer, especially in noisy environments.

This article focuses on vowel recognition in Cued Speech for French (referred to also as Cued French Language [9]) based on HMMs. As far as our knowledge goes, automatic vowel recognition in Cued Speech based on HMMs is being introduced for the first time ever. Based on a review of the literature written about Cued Speech, the authors of this study...
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 with the associated phonetically labelled transcriptions. 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 modelling lip shapes. For hand position modelling, the coordinates of two landmarks placed on the hand were used (i.e., 4 parameters). Fig. 2 shows the lip shape parameters that were used.

In automatic speech recognition, a diagonal covariance matrix is often used based on the assumption that the parameters are uncorrelated. As far as lip reading is concerned, however, parameters show a strong correlation. In the framework of this study, Principal Component Analysis (PCA) was applied to decorrelate the lip shape parameters, then a diagonal covariance matrix was used. All 24 PCA lip shape components were used for HMM training. In the experiments, context-independent models with a 3-state, left-to-right HMM topology were used.

### III. Results

#### A. Vowel-viseme classification experiment

This section details the classification of vowel-visemes. A viseme [12] consists of group of phonemes that look similar on lips/mouth (e.g., /p/, /b/, and /m/).

Previous studies have shown that in the French language, 3 vowel groups reflect the most accurate phoneme-viseme mapping [11]. Table I shows the mapping of the French vowels to visemes. An experiment was conducted to evaluate the performance of the lip reading system relating to vowel-viseme recognition based on HMMs. The parameter vectors length was 24 (i.e., 8 basic lip parameters, ∆, and ∆Δ) and 32 Gaussian mixtures per state were used. The experiment resulted in a three-by-three confusion matrix and the diagonal elements (i.e., correctly classified visemes) yielded a 96.1% classification accuracy.

#### B. Hand position classification

Hand position is one of the two mandatory components (the other is lip shape) in respect of vowel recognition in Cued Speech. To evaluate the proposed system, an experiment was conducted relating to hand position recognition from the curing of a Cued Speech cuer using HMMs. The parameters used for HMM modelling were the xy coordinates of the two marks placed on the cuer’s hand, automatically extracted from the data. The st- and second derivatives were used as
well. Considering that the study was of Cued Speech for French, five positions were taken into consideration and a HMM was trained in relation to each position. A five-by-five confusion matrix was generated and the diagonal elements (i.e., correctly classified hand positions) yielded a 91.5% classification accuracy.

C. Vowel recognition using using concatenative feature fusion

In these experiments, lip- and hand position elements were integrated into a single component using concatenative feature fusion, and recognition experiments were conducted to recognize the 14 French vowels. The feature concatenation used the concatenation of the synchronous lip

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

where $O_t^{LH}$ is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand position feature vector, and $D$ the dimensionality of the joint feature vector. In these experiments, the dimension of the lip shape stream was 24 (8 basic parameters, 8 $\Delta$ and 8 $\Delta\Delta$ parameters) and the dimension of the hand position stream was 12 (4 basic parameters, 4 $\Delta$, and 4 $\Delta\Delta$ parameters). The dimension $D$ of the joint lip-hand feature vectors was, therefore, 36. This size was reasonable, and no further processing was applied to reduce the dimensionality.

Fig. 3 shows the results obtained in a vowel-classification experiment. In this case, the vowel instances were automatically extracted from the sentences using forced alignment based on the audio signal. Using 32 Gaussian mixtures per state, a 85.1% vowel classification accuracy was obtained. Compared with the sole use only lip parameters, an absolute improvement of 17.1% was obtained.

Fig. 4 shows the vowel correct (substitutions and deletions were considered) obtained in an unconstrained phoneme recognition experiment using continuous sentences, and without having recourse to a language model. In this case, the vowel correctness achieved was 82.3%. Compared with the sole use of lip parameters, an absolute improvement of 26.6% was obtained. In both cases, high recognition rates were obtained which were comparable to those obtained by automatic vowel recognition experiments using audio signals (e.g., [13]).

D. Vowel recognition using multi-stream HMM decision fusion

In these experiments, lip shape and hand position elements were integrated into a single component using multi-stream HMM decision fusion, and recognition experiments were conducted to recognize the 14 French vowels. Decision fusion captures the reliability of each stream, by combining the likelihoods of single-stream HMM classes. Such an approach has been used in multi-band audio only ASR [14] and in audio-visual speech recognition [8]. The emission likelihood of multi-stream HMM is the product of emission likelihoods of single-stream components weighted appropriately by stream weights. Given the $O$ combined observation vector, i.e., lip shape and hand position elements, the emission probability of multi-stream HMM is given by

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

where $N(O; \mu, \Sigma)$ is the value in $O$ of a multivariate Gaussian with mean $\mu$ and covariance matrix $\Sigma$. 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 $\lambda_s$. In this study, we assume that the stream weights do not depend on state $j$ and time $t$. However, two constraints were applied. Namely,

$$0 \leq \lambda_h, \lambda_l \leq 1$$
where $\lambda_h$ is the hand position stream weight, and $\lambda_l$ 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 experimentally to 0.7 and 0.3 values, respectively. The selected weights were obtained by maximizing the accuracy on several experiments.

Fig. 5 shows the classification accuracies obtained when vowel instances were automatically extracted from the sentences. Using 32 Gaussian mixtures per state, an 87.6% vowel classification accuracy was obtained. Compared with the sole use of lip parameters, an absolute improvement of 19.6% was obtained. Fig. 6 shows the obtained results in an unconstrained phoneme recognition experiment. In this case, the vowel correct was 84.1%. Compared with the sole use of lip parameters, an absolute improvement of 28.4% was obtained.

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 McNemars test was applied [15]. The observed p-value was 0.001 indicating that the difference is statistically significant.

IV. Conclusions and Future Work

This article details a study on automatic vowel recognition in Cued Speech based on HMMs. Using integrated lip shape and hand position parameters, and based on concatenate feature fusion and multi-stream HMM decision fusion promising vowel classification accuracies of 85.1% and 87.6%, respectively, were achieved. The results obtained showed that fusion methods applied in audio-visual speech recognition could also be applied in Cued Speech recognition with great effectiveness. In the present case, vowel recognition was presented. Consonant recognition requires a different approach from that used in the vowel recognition. More specifically, in consonant recognition handshape coding and recognition are required. We are currently working on the consonant recognition problem and preliminary results are promising.

References