AN AUDIO-VISUAL SPEECH RECOGNITION SYSTEM FOR TESTING NEW AUDIO-VISUAL DATABASES

Tsang-Long Pao
Department of Computer Science and Engineering,Tatung University, Taipei, Taiwan, R.O.C

Wen-Yuan Liao
Department of Computer Science and Engineering,Tatung University, Taipei, Taiwan, R.O.C

Keywords: Audio-visual database, Audio-visual speech recognition, Hidden Markov model.

Abstract: For past several decades, visual speech signal processing has been an attractive research topic for overcoming certain audio-only recognition problems. In recent years, there have been many automatic speech-reading systems proposed that combine audio and visual speech features. For all such systems, the objective of these audio-visual speech recognizers is to improve recognition accuracy, particularly in the difficult condition. In this paper, we will focus on visual feature extraction for the audio-visual recognition. We create a new audio-visual database which was recorded in two languages, English and Mandarin. The audio-visual recognition consists of two main steps, the feature extraction and recognition.We extract the visual motion feature of the lip using the front end processing. The Hidden Markov model (HMM) is used for the audio-visual speech recognition. We will describe our audio-visual database and use this database in our proposed system, with some preliminary experiments.

1 INTRODUCTION

Automatic speech recognition (ASR) by machine has been a goal and an attractive research area for past several decades. However, in spite of the enormous of researches, the performance of current ASR is far from the performance achieved by humans. Most previous ASR systems make use of the acoustic speech signal only and ignore the visual speech cues. They all ignore the auditory-visual nature of speech.

In recent years, there have been many automatic speech-reading systems proposed, that combine audio and visual speech features. For all such systems, the objective of these audio-visual speech recognizers is to improve recognition accuracy, particularly in difficult condition. They most concentrated on the two problems of visual feature extraction and audio-visual fusion. Thus, the audio-visual speech recognition is a work combining the disciplines of image processing, visual/speech recognition and multi-modal data integration. Recent reviews can be found in Mason, Henneke, Goldschen and Chen. In this paper, we mainly concentrate on the bimodal speech recognition.

On the audio-visual database, however, there has been some effort in creating database for the audio-visual research area, but these are almost in English or other language, such as Tulips1, AVLetters, M2VTS, CUAVE, etc. The Mandarin database is rare in comparison with other languages. In out research, we record and create a new audio-visual database of Mandarin speech.

We also focus on the visual feature extraction for the audio-visual recognition. The audio-visual recognition consists of two main steps, the feature extraction and recognition. In the proposed approach we extract the visual motion feature of the lip in the front end processing. In the post-processing, the Hidden Markov model (HMM) is used for the audio-visual speech recognition. The overall structure of the proposed system is depicted in Fig.1.

The organization of this paper is as follows. The audio-visual database format is introduced in Section 2. The extraction of the acoustic and visual features is presented in Section 3. The experimental results and conclusions are given in Section 4 and 5.

Pao T. and Liao W. (2006).
AN AUDIO-VISUAL SPEECH RECOGNITION SYSTEM FOR TESTING NEW AUDIO-VISUAL DATABASES.
In *Proceedings of the First International Conference on Computer Vision Theory and Applications*, pages 192-196
DOI: 10.5220/0001369101920196
Copyright © SciTePress

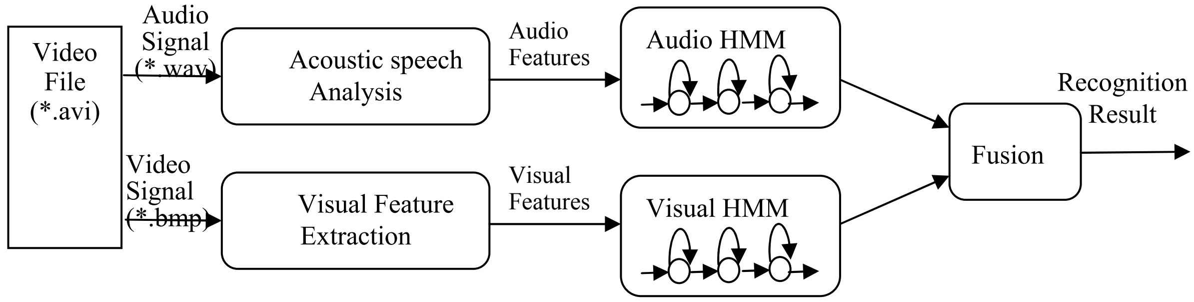

Figure 1: The overall structure of audio-visual extraction and recognition system.

2 AUDIO-VISUAL DATABASE FORMAT

Our audio-visual database consists of two major parts, one in English and one in Mandarin. The video in English was recorded from 35 speakers while the video in Mandarin was recorded from 40 speakers. The importance of the database is to allow the comparison of recognition of English speech and Mandarin speech. The video is in color with no visual aids given for lip or facial feature extraction. In both parts of database, each individual speaker was asked to speak 10 isolated English and Mandarin digits, respectively, facing a DV camera.

The video was recorded at a resolution of 320 × 240 with the NTSC standard of 29.97 fps, using a 1-mega-pixel DV camera. The on-camera microphone was used to record resulting speeches. Lighting was controlled and a blue background was used to allow change of different backgrounds for further applications. In order to split the video into the visual part and the audio part, we developed a system to decompose the video format (*.avi) into the visual image files (*.bmp) and speech wave files (*.wav) automatically. Figure 2 shows the sample of the result from the Audio-Visual Extraction System (AVES). Some samples of extracted images from the video files of our audio-visual database are shown in Fig. 3.

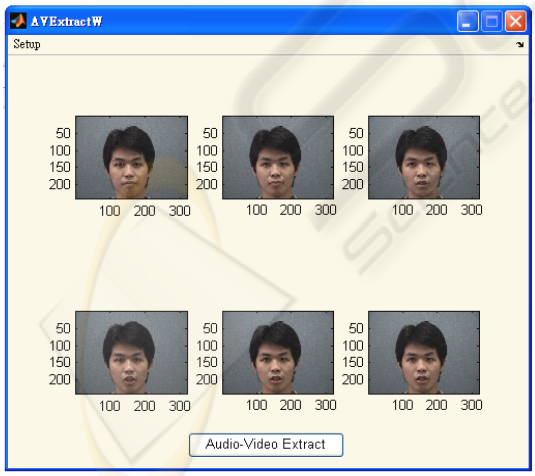

Figure 2: The sample of result performed by the Audio-Visual Extraction System (AVES).

3 FEATURE EXTRACTION

3.1 Acoustic Features

Since speech basically is a non-stationary random process, it is stochastic and its statistics are time-varying. According to the study of speech production, human speech's differences are occurred from mouth and vocal tract varying. These properties are short-time stationary and present on frequency domain. In order to recognize a speech, we do not only get the features on time domain, but also on frequency domain.

The mel-frequency cepstrum coefficient (MFCC) features have been shown to be more effective than other features. In the case of MFCCs, the windowing function is first applied to the frame before the short-time log-power spectrum is computed. Then the spectrum is smoothed by a bank of triangular filters, in which the pass-bands are laid out on a frequency scale known as mel-frequency scale. The filtering is performed by using the DFT.

In our proposed approach we extract the acoustic features by using the MFCC features, including basic and derived features.

Figure 3: The sample of images for our new audio-visual database.

3.2 Visual Features

Generally speaking, the features for visual speech information extraction from image sequences can be grouped into the following classes: geometric features, model based features, visual motion features, and image based features. In our system, the visual motion feature is used.

The motion-based feature approach assumes that visual motion during speech production contains relevant speech information. Visual motion information is likely to be robust to different skin reflectance and to different speakers. However, the algorithms usually do not calculate the actual flow field but visual flow field. A further problem consists in the extraction of related features from the flow field. However, recent research about motion-based segmentation got more performance than previous experiments. So the visual motion analysis can improve the performance of recognition.

An image is partitioned into a set of non-overlapped, fixed size, small rectangular blocks. The translation motion within each block is assumed to be uniform. This model only considers translation motion originally, but other types of motion, such as rotation and zooming, may be approximated by the piecewise translation of these small blocks.

In our system, the region of interested is first extracted from the original image. The main features, corners or edges of the mouth, are then found as the cues for motion estimation. The motion vectors computation corresponding to the feature points are performed by the block matching based motion estimation algorithm. Finally, the motion vectors for selected feature points are carried out for the feature vectors as the input of the recognition. Figure 4 shows the example of selected feature points, marked by the white circle dots, and their corresponding motion vectors.

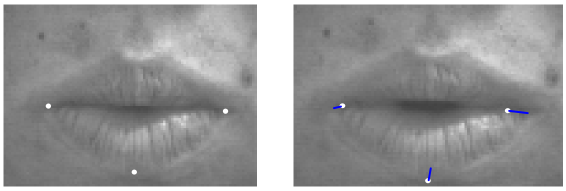

Figure 4: Motion vectors between consecutive images of lips: original images (left) and motion vectors (right).

4 PRELIMINARY EXPERIMENTS

4.1 Recognition Systems

As described in previous section, speech perception by human is a bimodal process characterized by high recognition accuracy and attractive performance degradation in the presence of distortion. As proven experimentally, acoustic and visual speeches are correlated. They show the complementarity and supplementary for speech recognition, especially under noisy conditions. However, precisely how and when humans integrate the audio and visual stimuli is still not clear.

In this research, we use the hidden Markov model (HMM) for the recognition. Our recognition experiment begins with the acoustic-only, then acoustic with noise, visual-only and finally audio-visual recognition. The acoustic and visual parameters are used to train each model, by means of HMM. This model is widely used in many ASR systems. The audio-visual recognizer uses a late-fusion system with separate audio and visual HMM-based speech recognizers.

4.2 Experimental Results

In this experiment, we use our database as the input data. The database used here contains the digits 0 to 9 in English and in Mandarin by 20 speakers, 16 males and 4 females. There are a total of 400 utterances. In the training phase, the 300 utterances of the database containing English and Mandarin of digits 0-9 from all speakers are used as the training set. After we train the model, the other 100 utterances are used as the testing set in testing phase.

The video stream is a sequence of images with resolution of 100 × 75 pixel from the database. Before we compute the motion estimation, some techniques are applied to the images in order to

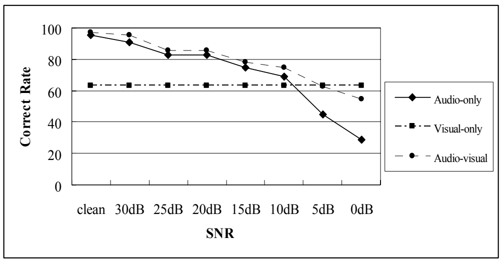

Figure 5: Comparison of recognition rate using different features at various noise levels for digits in English.

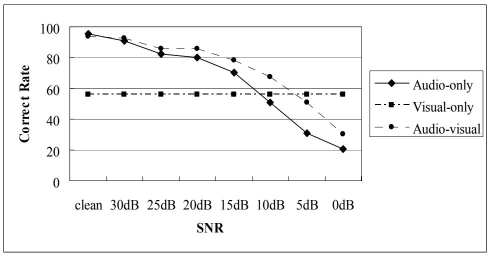

Figure 6: Comparison of recognition rate using different features at various noise levels for digits in Mandarin.

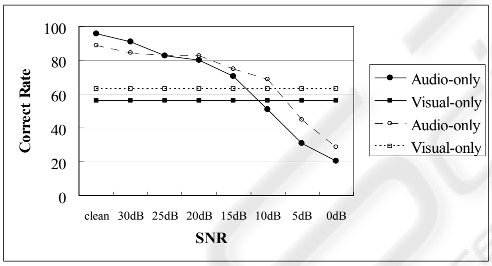

Figure 7: Comparison of the English and Mandarin digits recognition rate at various features and noise levels. (Solid lines are for English, and dashed line are for Mandarin).

make the computation convenient and increase the precision of the motion estimation. In our system, original 100 × 75 pixel image is extracted as a 100 × 72 pixel window around the center of mass for computational convenience.

Speech noise is added by using random noise at various SNRs. The system is trained on clean speech and tested under noisy conditions. The HMM-based recognizer implemented in MATLAB on Pentium-IV PC, using 4 states for Mandarin digits and 5 states for English digits. The time for feature extraction is under 6 seconds, and HMM parameter training spends about 10 seconds.

Initial results for the clean speech are good. Figure 5 and 6 show the results for English and Mandarin digits recognition, respectively. Figure 7 shows the comparison for English and Mandarin digits recognition at audio-only and visual-only feature situation. This result shows that the Mandarin digits recognition gets higher correct rate under audio-only features than English digits, the English digits recognition gets better correct rate under visual-only features than Mandarin digits.

5 CONCLUSIONS

In this paper, our focus is to construct a new audio-visual database and a lip-motion based feature extractor for the recognition system with a HMM based recognizer. The experimental results show a comparison between English and Mandarin speech recognition, and the improvement of using both audio and visual feathers.

The results for our proposed approach at the various SNRs for the speech show that the method including the visual or lip features produces a better performance than using the audio-only features. In the future, we will try to improve the overall recognition rate using other more robust features and recognizers.

REFERENCES

T. Chen, “ Audio-visual speech processing,” in *IEEE Signal Processing Magazine*, Jan. 2001

T. Chen and R. Rao, “Audiovisual interaction in multimedia communication,” in *ICASSP*, vol. 1. Munich, Apr. 1997, pp. 179-182.

C. C. Chibelushi, F. Deravi, and J. S. D. Mason, “A review of speech-based bimodal recognition,” in *IEEE Trans. Multimedia*, vol. 4, Feb. 2002, pp. 23-37.

M. N. Kaynak, Q. Zhi, etc, “Analysis of Lip Geometric Features for Audio-Visual Speech Recognition,” *IEEE Transaction on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 34, July 2004, pp. 564-570.

J. Luettin and G. Potamianos and C. Neti, “Asynchronous stream modeling for large vocabulary audio-visual speech recognition,” *2001*.

I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” in *IEEE Trans. pattern analysis and machine intelligence*, vol. 24, 2002, pp. 198-213.

S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," in *IEEE Trans. Neural Networks*, vol.13, July 2002, pp. 854-866.

G. Poamianos, etc,"Recent Advances in the Automatic Recognition of Audiovisual Speech" in *Proceeding of the IEEE,* Vol. 91, No. 9, September 2003.

SciTePress
Science and Technology Publications