STATE SYNCHRONOUS MODELING ON PHONE BOUNDARY FOR AUDIO VISUAL SPEECH RECOGNITION AND APPLICATION TO MUTI-VIEW FACE IMAGES

Kenichi Kumatani and Rainer Stiefelhagen

Interactive Systems Labs, University of Karlsruhe, Germany
k_kumatani@ieee.org and stiefel@ira.uka.de

ABSTRACT

Visual speech cues are known to improve the performance of automatic speech recognition (ASR). However, many researchers have used speaker's frontal pose mainly. We therefore introduce a new database for large vocabulary audio visual automatic speech recognition (AV-ASR), which contains not only frontal face images but also face images taken from different angles (multi-view face images). Another contribution of this paper is to present a new algorithm which can model audio and visual characteristics between phones. Finally we conducted large vocabulary continuous speech recognition experiments on the new database using the new algorithm. Experimental results show that the proposed AV-ASR system achieved high accuracy even if there are mismatches of the views between training and test data.

Index Terms— Audio visual automatic speech recognition, visual information, product HMM, multi-view.

1. INTRODUCTION

It is well known that humans use acoustic and visual information for speech recognition, and many researchers have shown that also automatic speech recognition systems benefit from using additional visual features, especially under noisy environments [1]-[7].

Most of the studies, however, have paid attention to using visual features extracted from frontal faces of the user. Less work has been conducted on using non-frontal faces. Yoshinaga et al., for example, conducted speech recognition experiments using profile face images [3]. However, the addressed task was limited to only four connected Japanese digits.

In our work we address the problem of audio-visual speech recognition from non-frontal faces on a large vocabulary task. In order to study the effect of different face views on AV-ASR, we constructed a new database which contains three view face images.

This paper also describes a new training method for AV-ASR, which aims at representing the relationship between audio and visual features faithfully. The new algorithm is based on product HMM approach [5]-[7]. We evaluated the new algorithm on the new database.

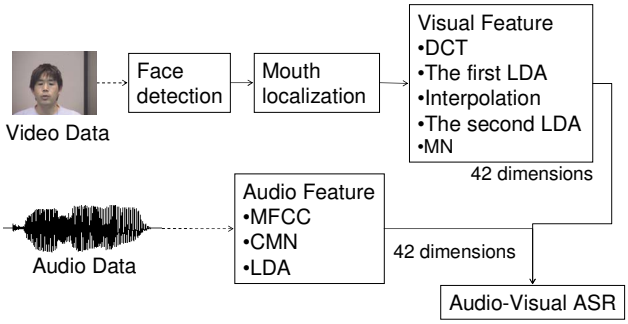

Fig. 1. Overview of our AV-ASR system

The remainder of this paper is organized as follows: Section 2 describes the outline of the AV-ASR system. Section 3 presents the detail of our mouth localization system. Section 4 presents the new training algorithm. Section 5 describes the database used in this work. Results are presented and discussed in Section 6.

2. BLOCK DIAGRAM OF AV-ASR SYSTEM

As shown in Fig. 1, an audio feature is extracted from audio data and a visual feature is extracted from video data. In the audio feature extraction step, 13 mel-frequency cepstral coefficients (MFCC) are calculated and cepstral mean normalization (CMN) is performed. After that, the feature vectors of 11 consecutive frames are concatenated into one 143 (=11x13) dimensional vector and then the dimension of the vector are reduced to 42 by linear discrimination analysis (LDA).

In the visual feature extraction step, a face is detected and a mouth region is localized by the mouth localization system, which we describe in the next section. After that, the mouth image is transformed by 2-dimensional discrete cosine transformation (DCT) and the 100 DCT coefficients with the high energy are extracted. Then, 100 coefficients are reduced to 30 by LDA. Here, the visual feature vector is

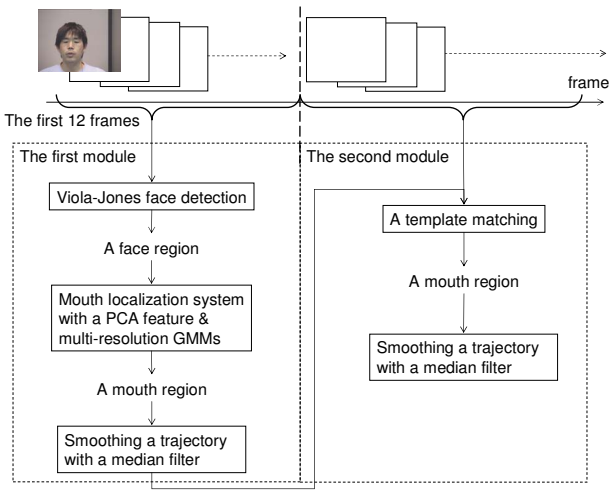

Fig. 2. Mouth region localization system

interpolated because video frame rate is less than audio one. After the interpolation, 15 consecutive frame vectors are concatenated and the dimension of the concatenated vector are reduced to 42 by LDA again in order to train the dynamics of the visual feature. And the mean is subtracted in the same manner as CMN in the audio feature extraction. Note that we reduce the dimension of a visual feature vector hierarchically.

3. MOUTH LOCALIZATION SYSTEM

Fig. 2 describes our mouth region localization system. The system has two stages (modules) in order to decrease computation time. In the first stage, a face region is localized by a Viola-Jones face detector [8], and the search area for the mouth region is limited to half of the detected lower face. Then, a feature vector is calculated by principle component analysis (PCA) for a rectangle region (analysis window). While translating and scaling the analysis window over the search area, the system searches the region which provides the maximum likelihood, given a Gaussian mixture model (GMM) which has been trained with mouth images in advance [9]. The found region is considered a mouth image. After the mouth regions are localized for 12 frames, the trajectory of mouth positions is smoothed with a median filter. Then, the mouth detector is switched to the second stage. The second module is based on a template matching approach. Templates are constructed from the mouth images localized in the first stage. In the second stage, the correlation between input and template images is calculated. If a correlation value is more than 0.9, the input image is added to a set of templates and the template with the worst score is removed. The trajectory is smoothed in the same way as the first module.

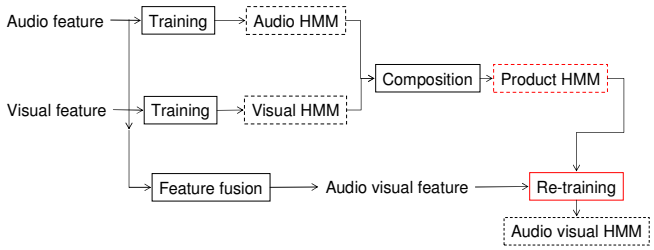

Fig. 3. Training algorithm for a product HMM

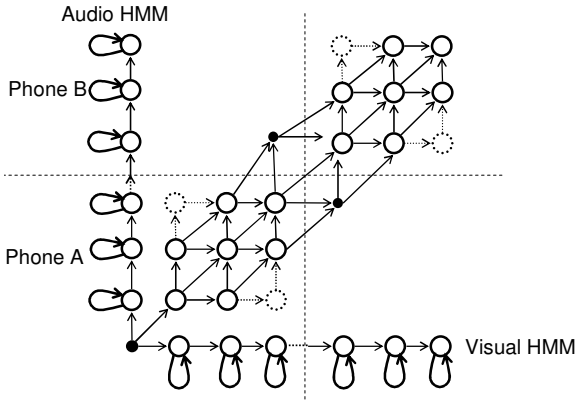

Fig. 4. The proposed topology of a product HMM

4. AUIDO-VISUAL MODELING

It has turned out in many publications that a product HMM can represent asynchronous characteristics between audio and visual events very well in AV-ASR [1][2][5]-[7]. We also showed that recognition accuracy was improved by re-training the product HMM with the audio visual features [5]. Fig. 3 depicts the block diagram of a training algorithm for the product HMM. First, audio and visual features are extracted from the input signals, respectively. Second, each feature model is built individually by the expectation-maximization (EM) algorithm [10]. Then the product HMM is combined from the audio and visual HMMs

Since a conventional product HMM forced a strict synchronization on every phoneme boundary, Nakamura et al proposed a new product HMM which could allow an asynchronous transition beyond the phoneme boundary [7]. However, their method increases the computational complexity and requires complicated implementation in the case that it was applied to the context-dependent model. Note that they used a mono-phone only in their recognition experiments. In order to overcome those problems, we propose a new product HMM which approximates their method. As shown in Fig. 4, only multiple transition paths are put on the phone boundaries. In [7], a product of states which belong to different phones was used as an output probability. Thus, the total number of such states is equal to the square of the number of tri-phones. On the other hand, since our method doesn't employ them, a low computation

load can be achieved. The proposed topology can also loosen the restriction between the phones. After the product HMM is combined, it is re-trained with audio visual features. During the re-training, the sequence of the feature is associated with the hidden state based on maximum likelihood criteria and the parameters of the states are updated. Rather than explicitly modeling an asynchronous state beyond a phone boundary, we approximate it by the states of the boundary.

5. THE DATABASE

This section describes the specification of the new database we recorded. Fig. 5 describes the layout of equipments at the recording. Three pan-tilt-zoom (PTZ) cameras are set at different angles for a subject. A microphone array is positioned in front of a speaker and cross talking microphone is put on speaker's ear. Three kinds of video data and two kinds of audio data are recorded. The cameras and microphones are connected to different computers. Audio and video data streams are synchronized with network time protocol (NTP). Fig. 6 shows the sample images which are 0 and 45 angle face, respectively. The speakers utter English alpha-numeric strings and English sentences extracted from TIMIT database. 39 male and 9 female are recorded. Most of speakers are non native of English. The image format follows PAL (interlace) and the data are saved in JPEG.

6. EXPERIMENT

6.1. Experimental conditions

Table 2 describes the parameters of AV-ASR system. The subjects in test data are not included in training data. We evaluated the performance of AV-ASR by using the frontal and 45 angle face images. However recognition experiments for profile faces are not conducted in this paper. To train the mouth localization system, we used the IBM ViaVoiceTM Audio-Visual Database [1,2] as well as our new database. Our AVASR system is based on JANUS speech recognition toolkit developed at UKA [11].

6.2. Results and discussions

We evaluated the performance of AV-ASR with a left-to-right 3 state HMM, the conventional product HMM [6] and the proposed HMM, respectively. Fig 6 shows word error rates (WER) for the three approaches on speech recorded with a close talking microphone (CTM). Here, the frontal face view is used for recognition. As a baseline, the performance of audio only ASR is also presented. In Fig 6, labels under the x-axis present audio SNR conditions where speech is corrupted with white Gaussian noise. The results

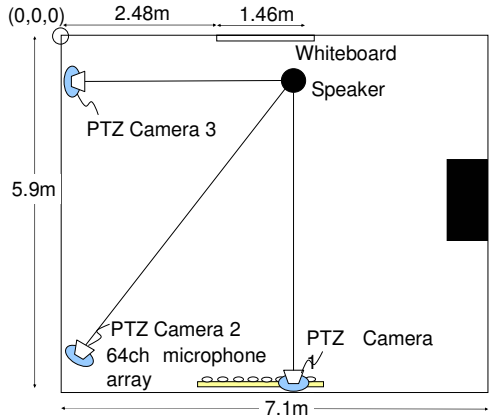

Fig. 5. Layout of equipments at the recording

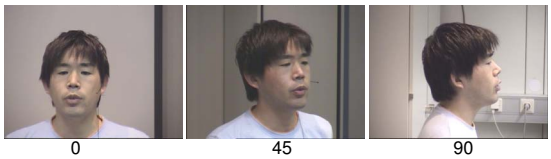

Fig. 6. Sample images

show that the proposed model has the best performance in almost all conditions although the conventional product HMM is slightly better (0.1%) in acoustically clean environment. This is because the proposed algorithm can approximate asynchronous transition beyond the phoneme boundary very well.

We also analyzed the effect of different facial views on WER. Fig. 7 describes WERs when training and test data are frontal faces (Test00-Train00), when a 45 degree's face view is tested with the frontal-face models (Test45-Train00) and when both test images and the trained visual models belong to the 45-degree face views (Test45-Train45), respectively. In Fig. 7, the bars above 'FA' presents WERs of speech recorded with a microphone array. Surprisingly even if the view of test data is mismatched to that of training data (Test45-Train00), AV-ASR could improve the WER. However, the performance of AV-ASR for 45 degree's face is worse than that for a frontal face.

7. CONCLUSIONS

In this work, we presented a novel method for modeling state transitions between product HMMs for audio-visual automatic speech recognition (AV-ASR). The proposed method is an extension of the approach proposed by Nakamura[7], and it introduces significant computational advantages, while preserving recognition accuracy.

The proposed method is evaluated on a novel database for large vocabulary AV-ASR, which contains not only frontal face images but also face images taken from different angles (multi-view face images). The experimental results showed

Table 2. Configuration of AV-ASR system

System Parameters	Value
Training data	42 subjects x 3 sets
Test data	6 subjects x 3 sets
The number of hidden states	300 (for each stream)
The number of mixtures	16 (for audio stream) 15 (for visual stream)
The phone model	Tri-phone
Vocabulary size	11133 words
Language model	Trigram (linear interpolation)

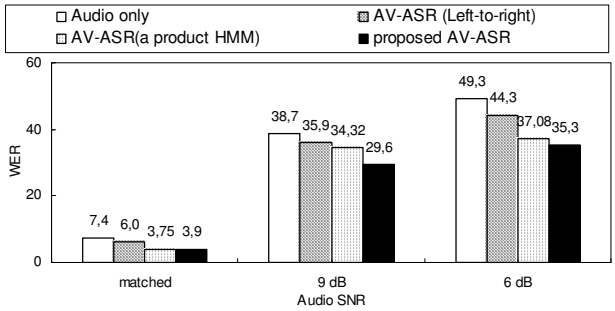

Fig. 6. Word error rate versus AV-ASR system

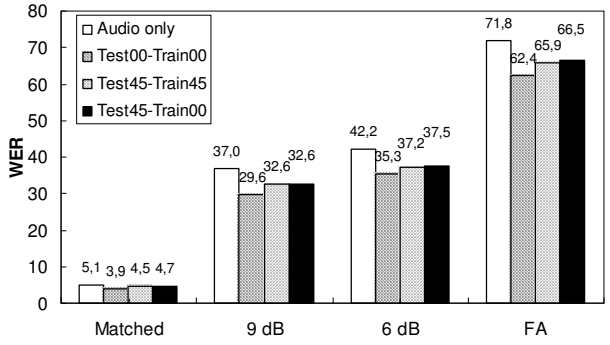

Fig. 7. Word error rate analysis for face views

that the new method could significantly improve the recognition performance. We also analyzed the effect of the facial views on the performance of AV-ASR. From the results, we can conclude that an ASR system can be improved even if there is a mismatch between facial training and test images.

Our work aims at improving AV-ASR performance under varying head orientations. In the future we will explore the use of visual models trained from different facial views for unconstrained AV-ASR.

8. ACKNOWLEDGEMENT

This work was sponsored by the European Union under the integrated project CHIL, Computers in the Human Interaction Loop (http://chil.server.de).

9. REFERENCES

[1] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A.Mashari and J. Zhou, "Audio-Visual Speech Recognition," in *Final Workshop 2000 Report*, Baltimore, , MD: Center for Language and Speech Processing, The Johns Hopkins University, Oct. 2000

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview", Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press (In Press), 2004.

[3] T. Yoshinaga, S. Tamura, K. Iwano and S. Furui, "Audio-visual speech recognition using lip movement extracted from side-face images", In Proc. AVSP2003, St Jorioz, France, pp.117-120, Sep. 2003.

[4] A. Garg, G. Potamianos, C. Neti and T.S. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition", In Proc. ICASSP2003, Hong Kong, vol. I, pp. 24-27, Apr. 2003.

[5] K. Kumatani and S. Nakamura, "Audio-visual speech recognition based on optimized product HMMs and GMM based-MCE-GPD stream weight estimation," *IEICE Trans. Inf. & Syst.*, Vol. E86-D, No. 3, pp. 454-463, Mar. 2003.

[6] M.J. Tomlinson, M.J. Russell and N.M. Brooke, "'Integrating audio and visual information to provide highly robust speech recognition"', In Proc ICASSP-96, Vol. 2, pp.821-824 May 1996.

[7] S. Nakamura, K. Kumatani and S. Tamura, "State synchronous modeling of audio-visual information for bi-modal speech recognition", In Proc. ASRU2001, Trento, Italy, pp.409-412, Dec. 2001.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", IEEE CVPR, pp. 511-518, 2001

[9] K. Kumatani and R. Stiefelhagen, "Mouth Region Localization Method based on Gaussian Mixture Model", In Proc. The Int. Workshop on Intelligent Computing in Pattern Analysis/Synthesis 2006 (IWICPAS2006), Xi'an, China, pp. 115-124, Aug. 2006.

[10] X.D. Huang, Y. Ariki and N.A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh Information Technology Series. EDINBURGH, 1990.

[11] H. Soltau, F. Metze, C. Fuegen, A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment", In Proc. ASRU2001, Trento, Italy, pp.214-217, Dec. 2001.