STATE SYNCHRONOUS MODELING OF AUDIO-VISUAL INFORMATION FOR BI-MODAL SPEECH RECOGNITION

Satoshi Nakamura[1], *Ken'ichi Kumatani*[1,2], *and Satoshi Tamura*[1,3]

[1] ATR Spoken Language Translation Research Laboratories
[2] Nara Institute of Science and Technology, [3] Tokyo Institute of Technology

ABSTRACT

There have been higher demands recently for Automatic Speech Recognition (ASR) systems able to operate robustly in acoustically noisy environments. This paper proposes a method to effectively integrate audio and visual information in audio-visual (bi-modal) ASR systems. Such integration inevitably necessitates modeling of the synchronization of the audio and visual information. To address the time lag and correlation problems in individual features between speech and lip movements, we introduce a type of integrated HMM modeling of audio-visual information based on HMM composition. The proposed model can represent state synchronicity not only within a phoneme but also between phonemes. Evaluation experiments show that the proposed method improves the recognition accuracy for noisy speech.

1. INTRODUCTION

The performance of ASR systems has been drastically improved recently. However, it is well known that the performance can be seriously degraded in acoustically noisy environments. Audio-visual ASR [1, 2, 4] systems offer the possibility of improving the conventional speech recognition performance by incorporating visual information, since the speech recognition performance is always degraded in acoustically noisy environments whereas visual information is not.

Audio and visual phonetic features have different durations. In other words, there is loose synchronicity between them, for instance, a speaker opens the mouth before making an utterance, and closes it after making the utterance. Furthermore, the time lag between the movement of the mouth and the voice might be dependent on the speaker or context.

As audio-visual integration methods for ASR systems, early integration and late integration are well known [1, 2]. In the early integration scheme, a conventional HMM is trained using audio-visual data. This method, however, cannot sufficiently represent the loose synchronization between the audio and visual information. Furthermore, the visual features of the conventional HMM may end up relatively poorly trained because of mis-alignments during the model estimation caused by the segmentation of the audio features. In the late integration scheme, the audio data and visual data are processed separately to build two independndt HMMs [1, 4]. This scheme assumes complete asynchronization between the audio and visual features. In addition, it can make the best use of the audio and visual data because there is a smaller bi-modal database than the typical database for audio only. However, the audio and visual features are regarded as independent.

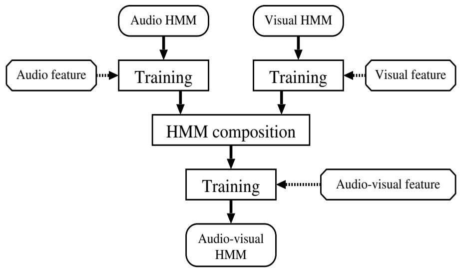

Fig. 1. Procedure Overview

In this paper, in order to model the synchronization between audio and visual features, we propose a method of state synchronous audio-visual integration based on HMM composition. The proposed model can represent synchronicity not only within a phoneme but also beyond phoneme boundaries.

2. AUDIO-VISUAL INTEGRATION BASED ON PRODUCT HMM

Figure 1 shows the outline of the acoustic model training for ASR systems in this paper. Figure 2 shows the proposed HMM topology. First, in order to create the audio and visual phoneme HMMs independently, audio features and visual features are extracted from audio data and visual data, respectively. In general, the frame rate of audio features is higher than that of visual features. Accordingly, the extracted visual features are incorporated such that the audio and visual features have the same frame rate. Second,

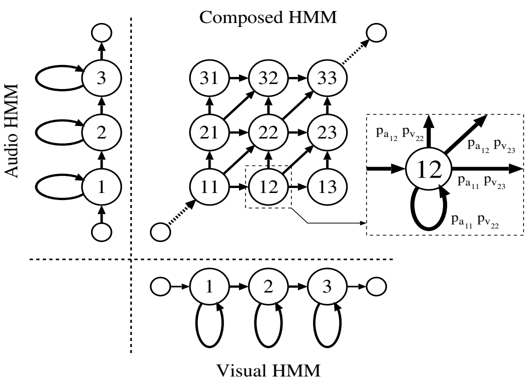

Fig. 2. Product HMM

the audio and visual features are modeled individually into two HMMs by the EM algorithm. Finally, an audio-visual phoneme HMM is composed as the product of these two HMMs based on HMM composition.

The output probability at state ij of the audio-visual HMM is,

$$b_{ij}(O_t) \;\;=\;\; b_i^A(O_t^A)^{\alpha_A} \times b_j^V(O_t^V)^{\alpha_V} \qquad (1)$$

which is defined as the product of the output probabilities of the audio and visual streams. Here, $b_i^A(O_t^A)^{\alpha_A}$ is the output probability of the audio feature vector at time instance t in state i, $b_j^V(O_t^V)^{\alpha_V}$ is the output probability of the visual feature vector at time instance t in state j, and α_A and α_V are the audio stream weight and visual stream weight, respectively. In a similar manner, the transition probability from state ij to state kl in the audio-visual HMM is defined as follows,

$$p_{ij,kl} \;\;=\;\; p_{a_{i,k}} \times p_{v_{j,l}} \qquad (2)$$

where $p_{a_{i,k}}$ is the transition probability from state i to state k in the audio HMM, and $p_{v_{j,l}}$ is the transition probability from state j to state l in the visual HMM. This composition is performed for all phonemes.

In the method proposed by [4], a similar composition is used for the audio and visual HMMs. However, because the audio and visual HMMs are trained individually, the dependencies between the audio and visual features are ignored. This results in the following two problems.

1. The product HMMs can not represent the loose synchronicity within phonemes.
2. The product HMMs force a strict synchronization on every phoneme boundary.

This paper proposes a new approach to solve the two problems. First, we propose the re-estimation of the product HMMs by using a small amount of audio-visual synchronous adaptation data. Second, we propose a new structure for the product HMMs. This new structure includes loose state synchronicity beyond the phoneme boundary.

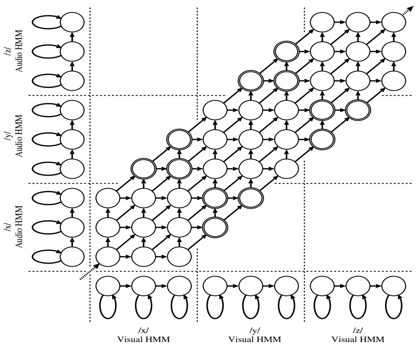

Fig. 3. Word HMMs

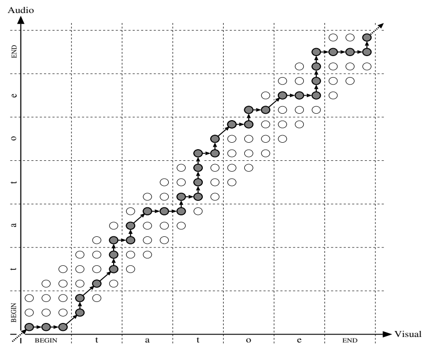

Fig. 4. State Alignment by a Word Model

2.1. State Synchronous Modeling within a Phoneme

The first problem is from the inability of the conventional product HMMs to represent loose state synchronicity within a phoneme. This problem is caused by the fact that the transition probabilities and output probabilities are obtained by the multiplication of probabilities from independent states of audio and visual HMMs. We propose new product HMMs whose parameters are re-estimated using audio-visual synchronous adaptation data [3]. The advantages of performing re-estimation are as follows.

- The re-estimation solves the state alignment problem. An inconsistent state alignment can be caused by the composition of states of two independent HMMs. These two composed states are originally aligned to different time periods based on the different HMMs.
- The re-estimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMM.

The re-estimation procedure is carried out using a small amount of audio-visual synchronous data. After the composition of two HMMs, the product HMMs can be re-estimated based on the Baum-Welch algorithm for multi-stream HMMs. Figure 5 shows results comparing audio HMMs, visual HMMs, early integration, late integration, and product HMMs with and without re-estimation [3]. The experimental conditions

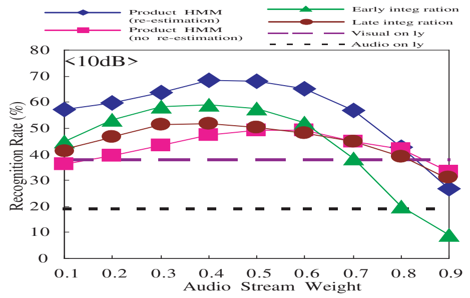

Fig. 5. Results of Product HMMs

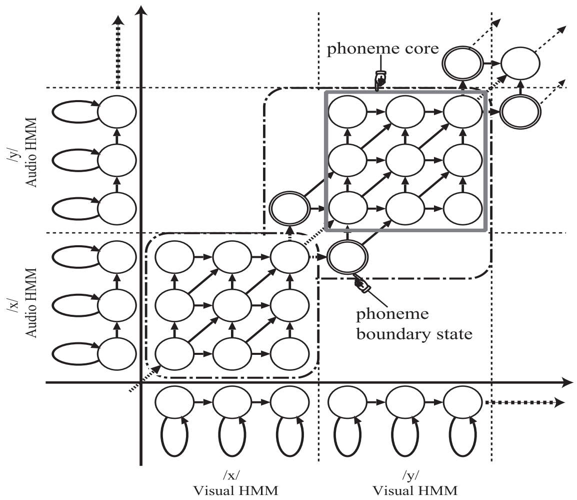

Fig. 6. Pseudo-biphone product HMMs

are the same as those in a later section except that the audio HMMs are trained using clean speech data. The figure shows that the product HMMs with re-estimation achieve the best performance, while the product HMMs without re-estimation are worse than those of the early and late integration schemes.

2.2. State Synchronous Modeling Beyond The Phoneme Boundary

The second problem is that the conventional product HMMs force a strict synchronization on every phoneme boundary. This is because the speech organs normally move earlier than the speech to be produced. Sometimes, the speech organs are already articulated in the previous audio phoneme utterance. Accordingly, we have to consider state synchronous modeling beyond the phoneme boundary.

Figure 3 shows a structure of word HMMs designed to investigate the asynchronicity on a phoneme boundary. The word HMMs have the same number of core states excluding extra asynchronous states on the phoneme boundaries, as indicated by the double circles in the figure. In the word HMMs, the core states are the same as the phoneme HMMs while the model parameters for the extra states are only re-estimated using the word utterance. Figure 4 shows a case of state alignment between input speech and states of word HMMs. Here, one can see that the input speech is aligned to the extra states, which represent state asynchronicity beyond the phoneme boundary. However, the word HMMs can not be applied to large vocaburaly speech recognition. It is necessary to represent this asynchronicity in the framework of phoneme-based speech recognition.

Considering this fact, we propose new product HMMs that include extra asynchronous states on phoneme boundaries as indicated in Fig. 6. The core states of the phoneme HMMs are the same as those of context independent phoneme product HMMs. In addition, the new product HMMs have two extra HMM states aiming to work similarly to the word HMMs. The first extra state is composed of the initial audio state and final visual state of the preceding phoneme HMM. The second extra state is composed of the initial visual state and final audio state of the preceding phoneme HMM. Since these extra states are dependent on the preceding phoneme, they can only be re-estimated in a manner similar to the biphone HMMs. Therefore, we call these HMM pseudo-biphone product HMMs. The proposed HMMs can tolerate one state asynchronicity beyond a phoneme boundary.

3. EVALUATION EXPERIMENTS

The audio signal is sampled at 12 kHz (down-sampled) and analyzed with a frame length of 32 msec every 8 msec. The audio features are 16-dimensional MFCC and 16-dimensional delta MFCC. On the other hand, the visual image signal is sampled at 30 Hz with 256 gray scale levels from RGB. Then, the image level and location are normalized by a histogram and template matching. Next, the normalized images are analyzed by two-dimensional FFT to extract 6x6 log power 2-D spectra for audio-visual ASR. Finally, 35-dimensional 2D log power spectra and their delta features are extracted. For each modality, the basic coefficients and the delta coefficients are collectively merged into one stream. Since the frame rate of the video images is 1/30, we insert the same images so as to synchronize the face image frame rate to the audio speech frame rate. For the HMMs, we use a two-mixture Gaussian distribution and assign three states for the audio stream and two states for the visual stream in the late integration HMMs and the baseline product HMMs. In this research, we perform word recognition evaluations using a bi-modal database [1]. We use 4740 words for HMM training and two sets of 200 words for testing. These 200 words are different from the words used in the training.

Figures 7-9 show word accuracies for acoustic SNR=15, 0, and -5dB. We compared the processed product HMMs

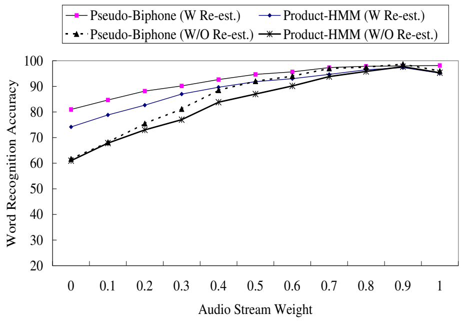

Fig. 7 Word Accuracy (SNR=15dB)

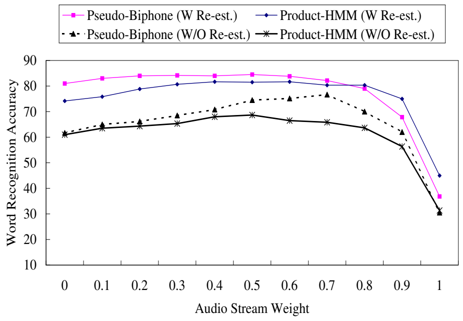

Fig. 8 Word Accuracy (SNR=0dB)

without re-estimation (Product-HMM(W/O Re-est.)), the proposed product HMMs with re-estimation (Product-HMM(W Re-est.)), the proposed pseudo-biphone product HMMs without re-estimation (Pseudo-Biphon(W/O Re-est.)), and the proposed pseudo-biphone product HMMs with re-estimation (Pseudo-Biphon(W Re-est.)). White noise was used to reduce the acoustic SNR in this experiment. The audio HMMs were trained using the SNR=15dB data. The results can be summarized as follows:

- The re-estimation of the product HMMs is quite effective to improve the performance. The re-estimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMMs. The re-estimation also produces a consistent state alignment to the input multiple modalities.
- The state synchronous modeling beyond the phoneme boundary results in significant improvements to the product HMMs. This result indicates the importance of considering the loose synchronicity of speech and speech organs over the phoneme boundary.
- The optimal stream weights for the maximum performance vary according to each method and acoustic SNR. Further investigations are necessary to adjust the optimal weights for the modalities.

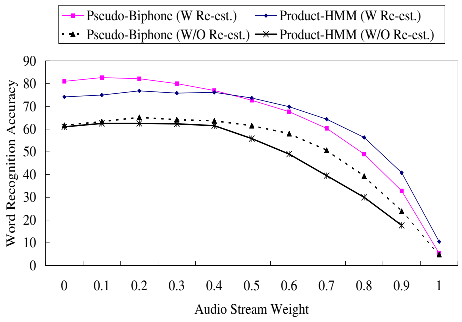

Fig. 9 Word Accuracy (SNR=-5dB)

4. CONCLUSION

This paper proposes a new HMM structure to effectively integrate audio and visual information in audio-visual (bimodal) systems. Our state synchronous modeling of audio-visual information is based on the product HMM. The proposed model can represent synchronicity not only within a phoneme but also between phonemes. Evaluation experiments show that the re-estimation of the model parameters using audio-visual synchronous data further improves the product HMMs. In addition, pseudo-biphone HMMs that introduce two extra asynchronous states are shown to improve the bimodal speech recognition accuracy. As future work, we are now working on the optimal weighting of the modalities according to the reliability in the environment.

5. ACKNOWLEDGEMENTS

The authors thank Prof. S. Furui of the Tokyo Institute of Technology and K. Shikano of the Nara Institute of Science and Technology for giving us the opportunity to conduct this study collaboratively.

6. REFERENCES

[1] Satoshi Nakamura, Ron Nagai and Kiyohiro Shikano, "Improved bimodal speech recognition using tied-mixture HMMs and 5000 word Audio-Visual Synchronous database", Proc. Eurospeech, Rhodes, pp. 1623-1626, 1997.

[2] Satoshi Nakamura, Hidetoshi Ito and Kiyohiro Shikano, "Stream weight optimization of speech and lip image sequence for Audio-Visual speech recognition", Proc. ICSLP2000, Vol. 3, pp. 20-23, 2000.

[3] Kenichi Kumatani, Satoshi Nakamura and Kiyohiro Shikano, "An Adaptive Integration Method Based on Product HMM for Bi-Modal Speech Recognition", HSC2001 (International Workshop on Hands-Free Speech Communication) pp. 195-198

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