SPEECH RECOGNITION FOR A DISTANT MOVING SPEAKER BASED ON HMM COMPOSITION AND SEPARATION

T. Takiguchi[†], *S. Nakamura*[‡], *K. Shikano*[‡]

[†]IBM Tokyo Research Laboratory,
1623-14, Shimotsuruma, Yamato-shi, Kanagawa, 242-8502, JAPAN
[‡]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN

ABSTRACT

This paper describes a hands-free speech recognition method based on HMM composition and separation for speech contaminated not only by additive noise but also by an acoustic transfer function. The method realizes an improved user interface such that a user is not encumbered by microphone equipment in noisy and reverberant environments. In this approach, an attempt is made to model acoustic transfer functions by means of an ergodic HMM [1]. The states of this HMM correspond to different positions of the sound source. It can represent the positions of the sound sources, even if the speaker moves. The HMM parameters of the acoustic transfer function are estimated by HMM separation [2]. The method is obtained through the reverse of the process of HMM composition, where the model parameters are estimated by maximizing the likelihood of adaptation data uttered from an unknown position. Therefore, measurement of impulse responses is not required. In this paper, we record the speech of a distant moving speaker in real environments. The results of experiments for the speech of a distant moving speaker clarified the effectiveness of HMM composition and separation.

1. INTRODUCTION

In hands-free speech recognition, one of the key issues as regards practical use is the development of a technology that allows accurate recognition of noisy and reverberant speech. Many methods have been presented for solving problems caused by additive noise and convolutional distortion in robust speech recognition. Two common examples of such methods are the speech enhancement and model compensation approaches. For the speech enhancement approach, spectral subtraction for additive noise and cepstral mean normalization for convolutional distortion have been proposed (e.g., [3, 4]). For the model compensation approach, the conventional multi-template technique, model adaptation (e.g., [5, 6]) and model (de-)composition methods (e.g., [1, 7, 8, 9, 10]) have been developed.

We applied HMM composition to the recognition of speech contaminated not only by additive noise but also by the reverberation of the room [1]. We also proposed HMM separation for estimating the HMM parameters of an acoustic transfer function [2]. The model parameters are estimated by maximizing the likelihood of adaptation data uttered from an unknown position. This paper describes the performance of the HMM composition and separation for recognition of the speech of a distant moving speaker. The speech of the distant moving speaker is recognized by using an ergodic HMM of acoustic transfer functions. Each state of the ergodic HMM of acoustic transfer functions corresponds to a position in a room, where all transitions among states are permitted. Therefore, the ergodic HMM of acoustic transfer functions is able to trace the positions of sound sources.

First, we give a brief overview of HMM composition [1]. Following this, we describe a method for estimating the HMM parameters of the acoustic transfer function, based on HMM separation [2]. We also describe the performance of HMM composition and separation for the speech of a distant moving speaker.

2. HMM COMPOSITION FOR NOISY AND REVERBERANT SPEECH

The observed speech in a noisy and reverberant room is represented by

$$O(\omega;m) = S(\omega;m) \cdot H(\omega;m) + N(\omega;m),$$

where $O(\omega;m)$, $S(\omega;m)$, $H(\omega;m)$, and $N(\omega;m)$ are short-term linear spectra for observed speech, clean speech, an acoustic transfer function, and noise in the analysis window m, respectively.

HMM composition is applicable if two stochastic information sources are additive. To apply HMM composition, the equation can be rewritten as follows:

$$O(\omega;m) = \exp(\cos(S_{cep}(t;m)+H_{cep}(t;m)))+N(\omega;m), \quad (1)$$

where $S_{cep}(t;m)$, and $H_{cep}(t;m)$ are cepstra for the clean speech, and the acoustic transfer function of quefrency t in the analysis window m. Accordingly, a composed HMM of the observed speech in the cepstral domain is represented by

$$\lambda_{O_{cep}} = \mathrm{Cos}^{-1}[\mathrm{Log}\{\mathrm{Exp}(\mathrm{Cos}(\lambda_{S_{cep}} \oplus \lambda_{H_{cep}})) \oplus k\lambda_{N_{lin}}\}],$$

where λ represents an associated HMM model, and the suffixes of *cep* and *lin* represent the cepstral domain and the linear-spectral domain, respectively. Cos, Log, and Exp are the cosine transform, logarithm transform, and exponential transform of the Gaussian *pdf*, respectively. To adjust the signal-to-noise-ratio (SNR), a coefficient, k, is used, and $\oplus$ denotes the model composition procedure.

The HMM recognizer decodes observed speech on a trellis diagram by maximizing the log-likelihood. The decoded path will find an optimal combination of speech, noise, and the acoustic transfer function.

2.1. Modeling of the Acoustic Transfer Function

Figure 3 shows the acoustic transfer function HMM in the case of three states. Each state of the acoustic transfer function HMM corresponds to a position in a room, and all transitions among states are permitted. Therefore, the acoustic transfer function HMM is able to represent the positions of sound sources, even if the speaker moves.

3. ESTIMATION OF THE ACOUSTIC TRANSFER FUNCTION ON THE BASIS OF HMM SEPARATION

Model parameters are estimated in an ML manner by using the expectation-maximization (EM) algorithm, which maximizes the likelihood of the observed speech:

$$\hat{\lambda}_H = \underset{\lambda_H}{\mathrm{argmax}} \Pr(O|\lambda_H, \lambda_N, \lambda_S).$$

Here, λ denotes the set of HMM parameters, while the suffixes of S, N, and H denote clean speech, noise, and the acoustic transfer function.

The observed speech is now represented by equation (1). Accordingly, the acoustic transfer function is

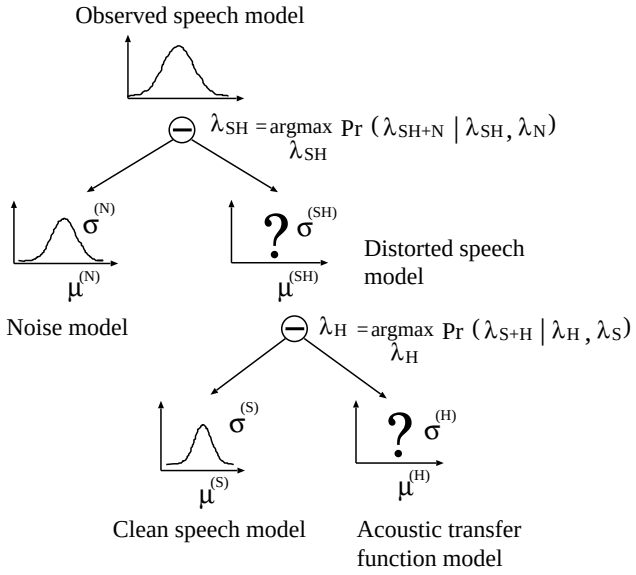

Figure 1: Illustration of model separation. The composed HMM is separated into a known HMM and an unknown HMM.

represented by

$$\begin{aligned} H_{cep}(t;m) &= \cos^{-1}[\log\{\exp(\cos(O_{cep}(t;m))) \\ &\quad -N(\omega;m)\}] - S_{cep}(t;m). \end{aligned}$$

The estimation equation of the acoustic transfer function HMM is written as

$$\lambda_{H_{cep}} = \mathrm{Cos}^{-1}[\mathrm{Log}\{\ \mathrm{Exp}(\mathrm{Cos}(\lambda_{O_{cep}})) \ominus \lambda_{N_{lin}}\}] \ominus \lambda_{S_{cep}},$$

where the separation of HMMs is represented by the $\ominus$ operator.

This equation shows that HMM separation is applied twice to noisy and reverberant speech. First, HMM separation is applied in the linear-spectral domain to estimate the distorted-speech HMMs by ML estimation. Then, the distorted-speech HMMs are converted to the cepstral domain, and HMM separation is applied again in the cepstral domain to estimate the acoustic transfer function HMM by ML estimation. Figure 1 illustrates HMM separation.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Conditions

The recognition algorithm is based on tied-mixture diagonal covariance HMMs. Each HMM has three states and three self-loops. The models of 55 context-independent phonemes are trained by using about 9600 sentences uttered by 64 speakers, which are contained in

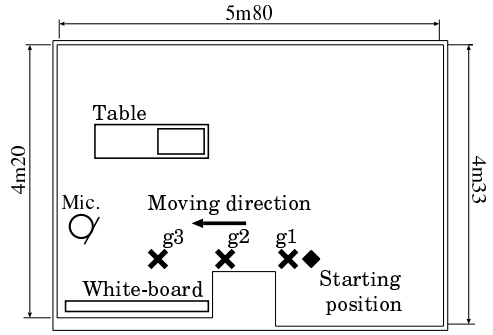

Figure 2: Recording conditions for the speech of a distant moving speaker

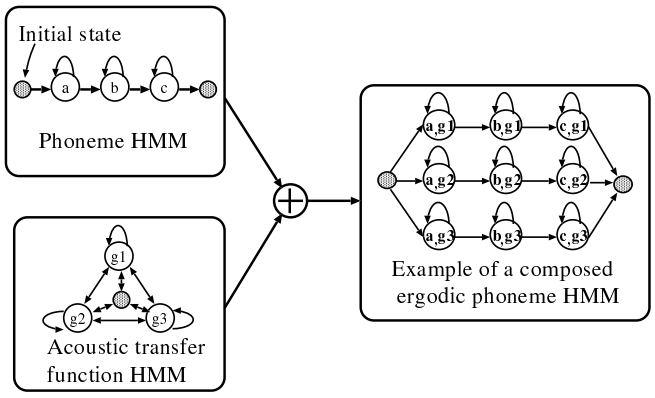

Figure 3: Example of a composed ergodic HMM in experiments with a distant moving speaker

the Acoustical Society of Japan (ASJ) continuous-speech database.

The speech signal is sampled at 12 kHz and windowed with a 32-msec Hamming window every 8 msec. Then FFT is used to calculate 16-order MFCCs (mel-frequency cepstral coefficients) and power. In recognition, the power term is not used, because it is only necessary to adjust the SNR in HMM composition. Sixteen-order MFCCs with their first-order differentials (ΔMFCC), and first-order differentials of normalized logarithmic energy (Δpower), are calculated as the observation vector of each frame. There are 256 Gaussian mixture components with diagonal covariance matrices shared by all of the models for MFCC and ΔMFCC, respectively. There are 128 Gaussian mixture components shared by all of the models for Δpower. A single Gaussian is employed to model an acoustic transfer function.

Figure 2 shows the recording conditions for the speech of the distant moving speaker. One male is walking from the "starting position" shown in figure 2. The speaker utters 31 sentences while moving. We also record the speech of a distant stationary speaker from the positions of sound sources g1, g2, and g3. One sentence is used for estimation of each acoustic transfer function. Figure 3 shows the composed ergodic HMM in experiments.

Table 1: Phrase accuracy [%] for a distant stationary speaker

Models	g1	g2	g3	Average
Clean-speech HMMs	58.1	72.6	77.7	69.5
Parallel models	67.0	76.3	86.1	76.5
Ergodic HMMs (g1, g2, g3)	66.1	73.5	87.0	75.5

Table 2: Phrase accuracy [%] for a distant moving speaker

Models	Phrase accuracy
Clean-speech HMMs	63.3
Parallel models	76.7
Ergodic HMMs (g1, g2, g3)	82.3
Ergodic HMMs (g1, g2)	78.6
Ergodic HMMs (g1, g3)	76.3
Ergodic HMMs (g2, g3)	80.0

4.2. Experimental Results

The points to be investigated are the performance of:

- Parallel models of acoustic transfer functions:

  Composed HMMs for each acoustic transfer function (each position) are individually set. Likelihood scores for their composed HMMs are calculated, and composed HMMs having the maximum likelihood are then selected.

- Ergodic models of acoustic transfer functions

A phrase recognition experiment was carried out for continuous-sentence speech, in which the sentences included 6 to 7 phrases on average. The task contained 306 phrases with a phrase perplexity of 306. The phrase accuracy for a close-talking microphone was 90.4%.

Table 1 shows the phrase accuracy for a distant stationary speaker. The phrase accuracy with clean-speech HMMs was 69.5%. Next, we compose the clean-speech HMMs and each of the acoustic transfer function HMMs, g1, g2, and g3. The performance of the

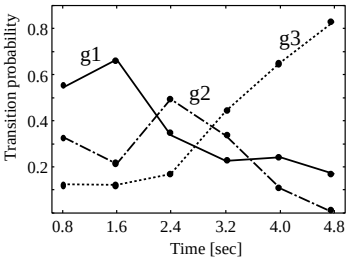

Figure 4: Estimated transition probabilities from the initial state to each state.

parallel models, where composed HMMs having maximum likelihood are selected, is 76.5% on average. The performance of the composed ergodic HMMs (shown in figure 3) is 75.5% on average. Comparison of this result with that for the parallel model shows a difference in performance of 1.0%. This is because all transition probabilities of acoustic transfer functions in the ergodic HMM are set equally, and a wrong path might be chosen. This table also indicates that the closest position, g3, results in the best performance. The greater the distance between the microphone and the positions of sound sources, the more the phrase accuracy will be decreased. This is because the SNR decreases.

Table 2 shows the phrase accuracy for the distant moving speaker. The phrase accuracy with clean-speech HMMs is 63.3%. The performance of the parallel models, where composed HMMs having maximum likelihood are selected, is 76.7%. In comparison with the case of the distant stationary speaker, the performance for the distant moving speaker is slightly better, because there were few speech data to be recorded while the distant moving speaker was in the vicinity of g1. The performance with the ergodic HMMs of acoustic transfer functions at g1, g2, and g3 is improved to 82.3%. These experimental results show the effectiveness of the ergodic HMMs for recognition of the speech of the distant moving speaker. Figure 4 shows the estimated transition probabilities from the initial state to each state: these are estimated by maximizing the likelihood of one sentence of testing data every 0.8 msec. As the testing speaker is walking from position g1 to position g3, the transition probability from the initial state to position g1 is highest in the first interval. The more time elapses, the more the transition probability to position g3 increases.

5. CONCLUSION

This paper has detailed a robust speech recognition technique for acoustic model adaptation based on HMM composition and separation in noisy and reverberant environments, where a user speaks from a distance of 0.5 m – 3.0 m. The aim of the HMM composition and separation methods is to estimate the model parameters so as to adapt the model to a target environment by using a small amount of a user's speech. In this approach, an attempt is made to model the acoustic transfer functions by means of an ergodic HMM. The states of the acoustic transfer function HMM correspond to different sound source positions. This HMM can represent the positions of sound sources, even if the speaker moves.

This paper investigated the performance of HMM composition and separation for recognition of speech of a distant moving speaker. Such speech is recognized by using an ergodic HMM of acoustic transfer functions. The experimental results show that the ergodic HMM can improve the performance of speech recognition for a distant moving speaker. In future work, we will investigate how to choose the number of states in the ergodic HMM.

6. REFERENCES

[1] S. Nakamura, T. Takiguchi, and K. Shikano, "Noise and room acoustics distorted speech recognition by HMM composition," in *Proc. ICASSP*, pp. 69-72, 1996.

[2] T. Takiguchi, S. Nakamura, Q. Huo, and K. Shikano, "Adaptation of model parameters by HMM decomposition in noisy reverberant environments," in *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels,* pp. 155-158, 1997.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE, ASSP-27*, No.2, 1979.

[4] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Ph.D Dissertation, ECE Department, CMU, Sep. 1990.

[5] A. Sankar and C-H. Lee, "Robust speech recognition based on stochastic matching," in *Proc. ICASSP,* pp. 121-124, 1995.

[6] V. Abrash, A. Sankar, H. Franco, and M. Cohen, "Acoustic adaptation using transformations of HMM parameters," in *Proc. ICASSP,* pp. 729-7, 1996.

[7] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP,* pp. 845-848, 1990.

[8] M. J. F. Gales and S. J. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. ICASSP,* pp. 233-236, 1992.

[9] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models," in *Proc. EUROSPEECH93*, pp. 1031-1034, 1993.

[10] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition," in *Proc. ICASSP*, pp. 129-132, 1995.