MODEL ADAPTATION BASED ON HMM DECOMPOSITION FOR REVERBERANT SPEECH RECOGNITION

Tetsuya TAKIGUCHI[1] *Satoshi NAKAMURA*[1] *Qiang HUO*[2] *Kiyohiro SHIKANO*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama Ikoma, Nara, 630-01 JAPAN

[2]ATR Interpreting Telecommunications Research Labs.
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

E-mail: tetuy-t@is.aist-nara.ac.jp

ABSTRACT

The performance of a speech recognizer is degraded drastically in reverberant environments. We proposed a novel algorithm which can model an observation signal by composition of HMMs of clean speech, noise and an acoustic transfer function[1]. However, how to estimate HMM parameters of the acoustic transfer function is a remaining serious problem. In our previous paper[1], we measured real impulse responses of training positions in an experiment room. It is inconvenient and unrealistic to measure impulse responses for every possible new experiment room. This paper presents a new method to estimate HMM parameters of the acoustic transfer function from some adaptation data by using an HMM decomposition algorithm which is an inverse process of the HMM composition. Its effectiveness is confirmed by a series of speaker dependent and independent word recognition experiments on simulated distant-talking speech data.

1. INTRODUCTION

In hands-free speech recognition, one of the key issues to practical use is the development of a technology which enables accurate recognition of noisy reverberant speech. This technology will play an especially important role in the recognition of distant-talking speech. In the past few years, many works have been performed in HMMs, and their training algorithms to improve the speaker independent speech recognition accuracy. To achieve a high recognition accuracy, a user usually must equip with a close-talking microphone. If a speaker inputs his/her speech from the distance or through a telephone channel, the recognition accuracy will seriously degrade due to the influences of reverberation or telephone channel distortion and environment noise. Many methods have been proposed to cope with the problems caused by additive noise and convolutional distortion. Among them, speech enhancement and model compensation approaches are two examples. For the speech enhancement approach, spectral subtraction for additive noise and cepstral mean normalization for convolutional distortion had been proposed (e.g., [2, 3, 4]). For the model compensation approach, conventional multi-template technique, model adaptation (e.g., [9, 10]) as well as model (de-)composition methods (e.g., [1, 5, 6, 7, 8, 11, 12]) had been developed.

In our previous paper [1], we apply the HMM composition to recognition of the signal which is contaminated by not only additive noise but also reverberation. If signal sources are independent each other and additive, the HMM composition algorithm can be adopted. Noise and speech are assumed to be independent and additive in the time domain. While an acoustic transfer function and speech are convolutional in the time domain, they are assumed to be independent and additive in the cepstral domain. We showed effectiveness of the proposed method[1] through the recognition experiments for noisy reverberant speech. However, how to estimate HMM parameters of an acoustic transfer function is a remaining serious problem. In our previous paper[1], we measured real impulse responses of the training positions in the testing room. The mean vectors of an acoustic transfer function HMM are derived from measured impulse responses. However, it is inconvenient and unrealistic to measure impulse responses for the testing room every time.

This paper presents a new method to estimate HMM parameters of the acoustic transfer function based on the HMM decomposition. The proposed algorithm is obtained as the natural result of a reverse process of the HMM composition. It can be applied to estimate the parameters of the acoustic transfer function HMM efficiently by using some adaptation speech data from the user's location instead of measured impulse responses.

2. HMM COMPOSITION

This section presents an overview of the HMM composition algorithm. The observation signal in a noisy reverberant room is modeled by

$$O(t) = S(t) \cdot H(t) + N(t).$$

We assume that clean speech and noise are independent, and clean speech and an acoustic transfer function are convolutional. Since the signal processing for the speech recognition are normally based on the short time spectral analysis, we regard $O(t), S(t), N(t)$ as short time linear spectra whose analysis window starts at time t from now on. $H(t)$ is short time linear spectrum of the acoustic transfer function. It is also denoted to be a function of t because we assume that the speaker may move around in a room.

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

in the cepstral domain:

$$O_{cep}(t) = \mathcal{F}^{-1}(\log(\exp(\mathcal{F}(S_{cep}(t) + H_{cep}(t))) + N(t))).$$

Here, $\mathcal{F}$, $\mathcal{F}^{-1}$ are Fourier(cosine) transform and inverse Fourier(cosine) transform, respectively. Accordingly a composed HMM of the observation signal in the cepstral domain is represented by

$$M_{O_{cep}} = \mathcal{F}^{-1}(\log(\exp(\mathcal{F}(M_{S_{cep}} \oplus M_{H_{cep}})) \oplus kM_{N_{lin}})).$$

Here, M represents an associated HMM model; *cep* and *lin* represents the cepstral domain and the linear spectral domain respectively; k is a coefficient to adjust SNR; and $\oplus$ denotes the model composition procedure. The model composition is carried out as follows: The number of states and transition probabilities of composed HMMs become a product of number of states and transition probabilities of each component model. The state observation probability density functions (PDFs) of composed HMMs are obtained by the convolution of the two associated distributions. If the distributions are Gaussians, say, $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, their convolution is still a Gaussian of $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. The HMM composition procedure is schematically summarized in Figure 1.

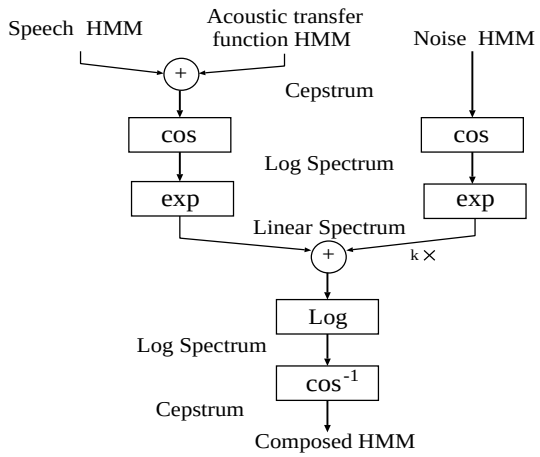

Figure 1. Block Diagram of HMM Composition

3. HMM DECOMPOSITION

If the structures of a noise HMM and an acoustic transfer function HMM are given, the parameters of the individual HMM can be estimated by a decomposition process. However, M_N will usually be obtained separately, since noise HMM parameters can be estimated accurately from the signal during noise periods. The estimation equation of the acoustic transfer function HMM is rewritten as follows in the cepstral domain:

$$M_{H_{cep}} = \mathcal{F}^{-1}(\log(\exp(\mathcal{F}(M_{O_{cep}})) \ominus k\exp(\mathcal{F}(M_{N_{cep}}))) \ominus M_{S_{cep}}.$$

Then the HMM decomposition procedure to estimate M_H is described as follows:

1. Re-estimate parameters of a composed HMM $\hat{M}_{O_{cep}}$ using adaptation data in the noisy reverberant room by ML (maximum likelihood) or MAP (maximum *a posteriori*)[13] estimation in the cepstral domain.
2. Estimate parameters of a noise HMM $\hat{M}_{N_{cep}}$ from the signal during noise periods.
3. Convert $\hat{M}_{O_{cep}}$, $\hat{M}_{N_{cep}}$ to the linear spectral domain:

$$\hat{M}_{O_{lin}} = \exp(\mathcal{F}(\hat{M}_{O_{cep}})),$$

$$\hat{M}_{N_{lin}} = \exp(\mathcal{F}(\hat{M}_{N_{cep}})).$$

4. Decompose $\hat{M}_{SH_{lin}}$ from $\hat{M}_{O_{lin}}$:

$$\hat{M}_{SH_{lin}} = \hat{M}_{O_{lin}} \ominus k\hat{M}_{N_{lin}}.$$

Here, $\ominus$ denotes deconvolution of distributions. If the distributions can be represented by Gaussian distribution, $N(\mu, \sigma^2)$ can be deconvolved into two distributions which are $N(\mu_1, \sigma_1^2)$ and $N(\mu - \mu_1, \sigma^2 - \sigma_1^2)$.

5. Convert $\hat{M}_{SH_{lin}}$ to the cepstral domain:

$$\hat{M}_{SH_{cep}} = \mathcal{F}^{-1}(\log(\hat{M}_{SH_{lin}})).$$

6. Decompose $\bar{M}_{H_{cep}}$ from $\hat{M}_{SH_{cep}}$:

$$\bar{M}_{H_{cep}} = \hat{M}_{SH_{cep}} \ominus M_{S_{cep}}.$$

Here, $\bar{M}_{H_{cep}}$ is averaged over all distributions, states and phone models based on the assumption that $\bar{M}_{H_{cep}}$ is the same over the adaptation speech. It is also assumed that $\bar{M}_{H_{cep}}$ is a one-state HMM having a single Gaussian with a diagonal covariance matrix. A tied-mixture HMM is used to model each speech unit in our experiments. The means $\bar{\mu}$ and variances $\bar{\sigma}^2$ of $\bar{M}_{H_{cep}}$ will be calculated as follows:

$$\begin{aligned}
\bar{\mu} &= \sum_{l=1}^{L}\sum_{s=1}^{S}\gamma_s^{(l)}\left\{\sum_{m=1}^{M}\lambda_{s,m}^{(l)}(\hat{\mu}_{s,m} - \mu_{s,m})\right\} \\
&= \sum_{l=1}^{L}\sum_{s=1}^{S}\gamma_s^{(l)}\left\{\sum_{m=1}^{M}\lambda_{s,m}^{(l)}\mu_{s,m}^{'}\right\} \\
&= \sum_{l=1}^{L}\sum_{s=1}^{S}\gamma_s^{(l)}\bar{\mu}_s^{(l)}, \\
\bar{\sigma}^2 &= \sum_{l=1}^{L}\sum_{s=1}^{S}\gamma_s^{(l)}\left\{\sum_{m=1}^{M}[\lambda_{s,m}^{(l)}(\hat{\sigma}_{s,m}^2 - \sigma_{s,m}^2) \right. \\
&\quad \left. + \lambda_{s,m}^{(l)}(\mu_{s,m}^{'} - \bar{\mu}_s^{(l)})^2]\right\},
\end{aligned}$$

where $\mu_{s,m}^{'} = \hat{\mu}_{s,m} - \mu_{s,m}$ and $\bar{\mu}_s^{(l)} = \sum_{m=1}^{M}\lambda_{s,m}^{(l)}\mu_{s,m}^{'}$. $(\hat{\mu}_{s,m}, \hat{\sigma}_{s,m}^2)$ and $(\mu_{s,m}, \sigma_{s,m}^2)$ are (mean, variance) of mth distributions of $\hat{M}_{SH_{cep}}$ and $M_{S_{cep}}$ respectively. Here $(\mu_{s,m}, \sigma_{s,m}^2) = (\mu_{s',m}, \sigma_{s',m}^2), s \neq s'$, since $\hat{M}_{SH_{cep}}$ and $M_{S_{cep}}$ are tied mixture HMMs. $\lambda_{s,m}^{(l)}$ is the mixture coefficient of mth distribution of sth state of lth speech unit which keeps fixed during adaptation.

L is the number of speech units existed in adaptation data and S is the number of state of each speech unit. $\gamma_s^{(l)}$ is the weighting coefficient which is a ratio of the total number of frames belonging to state s of speech unit l over the total number of frames of the adaptation data.

4. EXPERIMENTS AND RESULTS

Recognition experiments are conducted to evaluate effectiveness of the proposed method. In this study, as a first step, we focus on room reverberation distortion only and examine the decomposition of

$$\bar{M}_{H_{cep}} = \hat{M}_{SH_{cep}} \ominus M_{S_{cep}}.$$

The decomposition of $\hat{M}_{SH_{lin}}$ from $\hat{M}_{O_{lin}}$ can be dealt with separately. Figure 2 shows a top view of the experimental room. The sound signal is captured by using a single omni-directional microphone. We measured 9 transfer functions corresponding to 9 sound source positions by using the method reported in [14]. The length of reverberation time is approximately 180 msec for the experiment room.

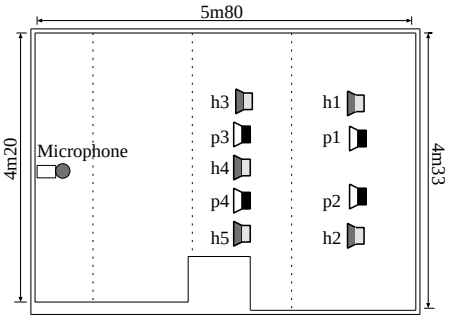

Figure 2. A top view of the experimental room

Two speech corpora are used for evaluation. One is the A-set of the ATR Japanese speech database. The other is the ASJ continuous speech database. The former contains word utterances and the latter contains sentence utterances, both spoken by announcers. The speaker independent (SI) model is trained by using utterances from 64 speakers in the ASJ database. The speaker dependent (SD) model is trained by using 2620 words of two male and one female speakers from the ATR database, respectively. 500 words for testing are different from those used in SD training. The adaptation words are also selected from those used in training, and excluded from testing set. Each set of the adaptation word consists of 50 words. The test and adaptation data are simulated by linear convolution of clean speech signal and measured impulse responses from the positions $p1,\ldots,p4$.

54 context independent phone models are used. Each phoneme HMM is a left-to-right 3-state tied-mixture HMM. There are in total 256 Gaussian mixture components with diagonal covariance matrices. Each feature vector consists of 16 mel-frequency cepstral coefficients (MFCCs). A single Gaussian PDF is used to model an acoustic transfer function for each position. A series of comparative experiments are conducted to examine:

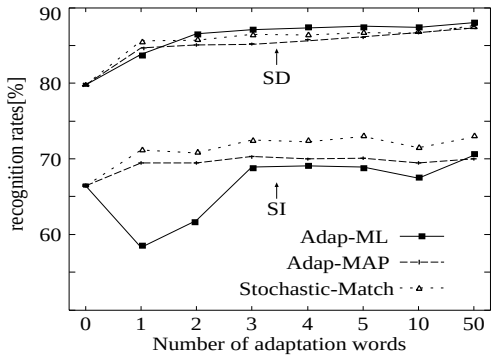

Figure 3. SD and SI word recognition rates[%] by HMM adaptation

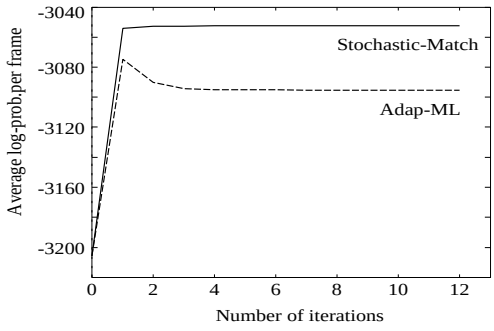

Figure 4. Convergence of the adaptation algorithms for SI seed model

- how the proposed methods work in both the SD and the SI recognition of reverberant speech.
- performance of the composed model adaptations which include ML and MAP reestimation.
- performance differences between the proposed methods and other two popular techniques, namely ML stochastic matching(SM)[9] and cepstral mean subtraction (CMS) [3].

Figure 3 shows 500-word recognition results averaged over two male and one female speakers. 'Adap-ML' refers to the results by using the proposed method where composed-model adaptation is carried out via ML reestimation, while 'Adap-MAP' is that of its MAP counterpart. The averaged SD and SI recognition rates with clean speech HMMs are 79.8% and 66.5%, respectively. The 'Adap-ML' and 'Adap-MAP' improves the SD recognition rate to 87.6% and 86.2%, and the SI recognition rate to 68.9% and 70.1% by using 5 adaptation words in the average of the three speakers, respectively. The result also shows that the 'Adap-MAP' method is able to rapidly adapt the model parameters of the acoustic transfer function HMM by MAP estimation, whereas in the SD recognition, 'Adap-ML' outperforms 'Adap-MAP' when more adaptation data become available. This is because with the SD seed model, we can get a more accurate alignment for SD adaptation data. Furthermore, 'Stochastic-Match' refers to the result by using

Table 1. Comparison of several methods

-	SD	SI
Clean	79.8%(77.8%)	66.5%
Adap-ML	87.6%(84.3%)	68.9%
Adap-MAP	86.2%(83.1%)	70.1%
Ergodic-CHMM	- (86.2%)	-
CMS	- (75.2%)	-
Stochastic-Match	86.8%(83.5%)	73.0%

( ) indicates result for one speaker.

the SM method[9]. The experimental results show that the recognition performance of the 'Adap-ML' is slightly better than (or no big difference from) that of the SM method in the SD seed model case, whereas the SM method achieves a better performance than the proposed methods in the SI seed model case. One possible explanation for the latter observation is that the assumption of the proposed methods that each state of the re-trained HMM corresponds to the associated state of the SI HMM is too fragile in the SI case.

In Figure 4, we compare the convergence property of the proposed 'Adap-ML' method and the SM method in the SI seed model case. The average log-likelihood per frame of one adaptation word versus iteration number of EM algorithm is plotted. The results show that one or two iterations seem enough for both algorithms.

Table 1 summarizes the performance comparison of several methods for the SD and the SI recognition experiments. In this table, 'Clean' is the result by using the models trained on clean speech. 'Ergodic-CHMM' is the result by using the previously proposed [1] composed model of a clean speech HMM and an ergodic acoustic transfer function HMM constructed from 5 training positions, $h_1,\ldots,h_5$. 'CMS' is the result of the cepstral mean subtraction [3]. In this case, the SD seed model is trained by using CMS-processed clean speech data. The results in Table 1 show that our proposed 'Adap-ML' method improves the SD recognition rate from 79.8% to 87.6% by using 5 adaptation words without any measurement of impulse responses. The performance for one speaker is only slightly (1.9%) worse than that of an ergodic transfer function HMM where real impulse responses are used. The result also clearly shows that the simple CMS technique does not work in reverberant speech recognition, especially with such a long reverberation time as 180 msec in this study. On the other hand, both the proposed methods and the SM method are able to improve the performance somehow.

5. DISCUSSION AND CONCLUSION

We have presented a new method to estimate HMM parameters of an acoustic transfer function based on the HMM decomposition. These methods enable to estimate the parameters of the acoustic transfer function HMM not by measured impulse responses but by the adaptation data from the user's location. The experimental results indicate that the proposed methods improve the SD recognition rate from 79.8% to 87.6%, and the SI recognition rate from 66.5% to 70.1% for reverberant speech by using 5 adaptation words in the average of the three speakers. In comparison with the ML stochastic matching method(SM)[9], the recognition rates of the proposed 'Adap-ML' method is slightly higher than that of the SM method for the SD model, whereas it is worse than the SM method for the SI model.

As future works, we need a model compensation procedure that acts over much longer intervals than the traditional assumptions of the short-time stationarity of the speech signal. When more than one sources of distortion exist, e.g., both additive and convolutional distortions, some new theoretical frameworks are required to directly take into account the nonlinear interaction between different types of distortions. When the distortion sources are non-stationary, e.g., a moving speaker and a non-stationary ambient noise, some adaptive compensation techniques are needed. To enhance the efficacy and the effectiveness of the compensation, those techniques are mostly wanted to be able to better characterize the distribution of the possible distortion types, and use this distribution to choose the appropriate compensation model. We are working along these lines of thoughts.

REFERENCES

[1] S. Nakamura, T. Takiguchi, K. Shikano, "Noise and room acoustics distorted speech recognition by HMM composition,", *Proc. ICASSP-96*, 1996, pp.69-72.

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

[3] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, " *J. Acoust. Soc. Amer.*, Vol. 55, pp.1304-1312, 1974.

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

[5] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise,", *Proc. ICASSP-90*, 1990, pp.845-848.

[6] M. J. F.Gales and S. J. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," *Proc. ICASSP-92*, 1992, pp.233-236.

[7] M. J. F. Gales, S. J. Young, "PMC for speech recognition in additive and convolutional noise," CUED-F-INFENG-TR154, 1993.

[8] F. Martin, K. Shikano and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models,", *Proc. EUROSPEECH-93*, 1993, pp.1031-1034.

[9] A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching," *Proc. ICASSP-95*, 1995, pp.121-124.

[10] V. Abrash, A. Sankar, H. Franco and M. Cohen, "Acoustic adaptation using transformations of HMM parameters," *Proc. ICASSP-96*, 1996, pp.729-732.

[11] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition," *Proc. ICASSP-95*, 1995, pp.129-132.

[12] Y.Minami and S. Furui, "Adaptation method based on HMM composition and EM algorithm," *Proc. ICASSP96*, 1996, pp.327-330.

[13] C.-H. Lee, C.-H. Lin and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. on Signal Processing*, Vol. 39, No. 4, pp.806-814, 1991.

[14] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Amer.*, Vol. 97, No. 2, pp.1119-1123, 1995.