ADAPTATION OF MODEL PARAMETERS BY HMM DECOMPOSITION IN NOISY REVERBERANT ENVIRONMENTS

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

This paper presents a new method to estimate HMM parameters of an acoustical transfer function based on HMM decomposition in model domain. The model parameters are estimated by maximizing a likelihood of adaptation data. The proposed method is obtained as the natural result of a reverse process of the HMM composition. In our previous paper[1], we proposed a method which can model an observed signal by the composition of HMMs of clean speech, noise and an acoustical transfer function. The previously proposed method needs measurement of impulse responses. It is inconvenient and unrealistic to measure impulse responses for a new environment. The new method is able to estimate HMM parameters of the acoustical transfer function from a small amount of adaptation data. Its effectiveness is confirmed by a series of speaker dependent and independent word recognition experiments on simulated distant-talking speech.

1. INTRODUCTION

In hands-free speech recognition, a speaker inputs his/her speech from a distance. Therefore the recognition accuracy seriously degrades due to influences of reverberation 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 have 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]) have 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 the signal sources are independent each other and additive, the HMM composition method can be adopted. The noise and the speech signal are assumed to be independent and additive in the time domain. While the acoustical transfer function and the speech signal are convolutional in the time domain, they are assumed to be independent and additive in the cepstral domain. Therefore the HMM composition is applied twice in the cepstral domain and the linear spectral domain. We show effectiveness of the previously proposed method [1] through the recognition experiments for noisy reverberant speech. However, how to estimate HMM parameters of the acoustical transfer function is a remaining serious problem. The mean vectors of the acoustical transfer function HMM are derived from measured impulse responses. It is inconvenient and unrealistic to measure impulse responses for a new environment.

This paper presents a new method to estimate HMM parameters of the acoustical transfer function based on the HMM decomposition in model domain. The estimation is implemented by maximizing a likelihood of adaptation data from any user's position. In [9], an estimation method in ML is presented, where the estimation of the acoustical transfer function is implemented in the time domain. On the other hand, we estimate the acoustical transfer function in model domain. Therefore the estimation in model domain can reduce computation amount.

2. MODEL ADAPTATION BY HMM DECOMPOSITION

Model parameters are estimated in a maximum likelihood(ML) manner using the expectation maximization(EM) algorithm, where the likelihood of the observed signal is maximized. The proposed method is based on the HMM decomposition method. Therefore the estimation of the acoustical transfer function is implemented in model domain.

The observed signal in the noisy reverberant environment is represented by

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

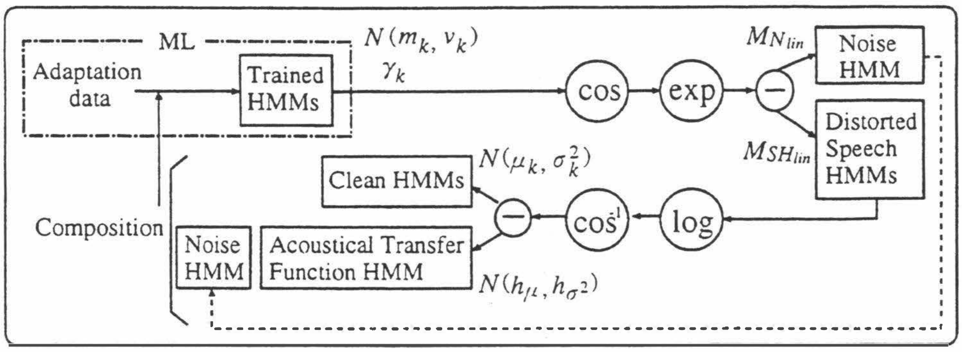
Here $\mathcal{F}$, $\mathcal{F}^{-1}$ are Fourier(cosine) transform and inverse Fourier(cosine) transform respectively. $O_{cep}(t;m)$, $S_{cep}(t;m)$ and $H_{cep}(t;m)$ are cepstra for an observed signal, a clean speech signal and an acoustical transfer function of quefrency t in m-th frame respectively; and $N(\omega;m)$ is linear spectra for a noise signal of frequency ω in m-th frame. Accordingly, the acoustical transfer function is represented by

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

The estimation equation of the acoustical transfer function HMM is rewritten by

$$\begin{aligned} M_{H_{cep}} &= \mathcal{F}^{-1}[\log\{ \exp(\mathcal{F}(M_{O_{cep}})) \\ &\quad \ominus M_{N_{lin}}\}] \ominus M_{S_{cep}}, \end{aligned}$$

where M represents the associated HMMs; the suffixes of *cep* and *lin* represent the cepstral domain and the linear spectral domain respectively. The estimation of the



(· Estimate parameter of a noise HMM using a signal during noise periods)

Figure 1. Parameter estimation by HMM decomposition

acoustical transfer function is implemented by maximizing the likelihood of the observed signal,

$$M_{H_{cep}} = \underset{M_{H_{cep}}}{\operatorname{argmax}} P(O|M_{H_{cep}}, M_{S_{cep}}, M_{N_{cep}}).$$

The decomposition of HMMs is represented by $\ominus$ operator. For example $C = A \ominus B$ means decomposition of C from A on condition that convolution of C and B equal to A. If the distributions of A and B are Gaussian, say, $N(\mu_A, \sigma_A^2)$ and $N(\mu_B, \sigma_B^2)$ respectively, the distribution of C is $N(\mu_A - \mu_B, \sigma_A^2 - \sigma_B^2)$.

The proposed method is shown in the followings. Here l is number of iteration.

1. Re-estimate parameters of composed HMMs $M_{O_{cep}}^{(l)}$ using adaptation data in the noisy reverberant environment by ML estimation in the cepstral domain.

2. Estimate parameters of a noise HMM $M_{N_{cep}}$ from the signal during noise periods.

3. Convert $M_{O_{cep}}^{(l)}$ and $M_{N_{cep}}$ to the linear spectral domain:
$$M_{O_{lin}}^{(l)} = \exp(\mathcal{F}(M_{O_{cep}}^{(l)})),$$
$$M_{N_{lin}} = \exp(\mathcal{F}(M_{N_{cep}})).$$

4. Decompose $M_{SH_{lin}}^{(l)}$ from $M_{O_{lin}}^{(l)}$:
$$M_{SH_{lin}}^{(l)} = M_{O_{lin}}^{(l)} \ominus M_{N_{lin}}.$$

5. Convert $M_{SH_{lin}}^{(l)}$ to the cepstral domain:
$$M_{SH_{cep}}^{(l)} = \mathcal{F}^{-1}(\log(M_{SH_{lin}}^{(l)})).$$

6. Estimate parameters of the acoustical transfer function, a mean and a variance, $(h_\mu^{(l)}, h_{\sigma^2}^{(l)})$. Here, to simplify a description of equations, the clean speech HMMs $M_{S_{cep}}$ is represented by tied-mixture HMMs, (μ_k, σ_k^2).
$$\Delta h_\mu^{(l)} = h_\mu^{(l)} - h_\mu^{(l-1)},$$
$$\Delta h_{\sigma^2}^{(l)} = h_{\sigma^2}^{(l)} - h_{\sigma^2}^{(l-1)},$$
$$\eta^{(l)} = \underset{(\Delta h_\mu^{(l)}, \Delta h_{\sigma^2}^{(l)})}{\operatorname{argmax}} P(O|\eta^{(l-1)}, M_{S_{cep}}).$$

$\eta^{(l)}$ is computed by maximization of the following auxiliary function,

$$\begin{aligned} &Q(\eta^{(l)}|\eta^{(l-1)}) \\ &= -\sum_t^T \sum_k^K \gamma_{t,k}^{(l)} \left\{ \frac{1}{2} \log(\sigma_k^2 + h_{\sigma^2}^{(l-1)} \right. \\ &\qquad \left. + \Delta h_{\sigma^2}^{(l)}) + \frac{(o_t - \mu_k - h_\mu^{(l-1)} - \Delta h_\mu^{(l)})^2}{2(\sigma_k^2 + h_{\sigma^2}^{(l-1)} + \Delta h_{\sigma^2}^{(l)})} \right\}, \end{aligned}$$

where o_t is the observed noisy reverberant signal at time t in the cepstrum domain. K and T are the number of Gaussian distributions and the number of total frames of adaptation data respectively. On the assumption that the variance is fixed, $\Delta h_\mu^{(l)}$ is derived from $\partial Q / \partial \Delta h_\mu^{(l)} = 0$, then given by

$$\Delta h_\mu^{(l)} = \frac{\displaystyle\sum_{k=1}^K \gamma_k^{(l)} \frac{m_k^{(l)} - \mu_k - h_\mu^{(l-1)}}{\sigma_k^2 + h_{\sigma^2}^{(l-1)}}}{\displaystyle\sum_{k=1}^K \frac{\gamma_k^{(l)}}{\sigma_k^2 + h_{\sigma^2}^{(l-1)}}},$$

where $m_k^{(l)} = \sum_t \gamma_{t,k}^{(l)} o_t / \gamma_k^{(l)}$. $m_k^{(l)}$ is the mean of adaptation data at l-th iteration by EM algorithm. $\Delta h_{\sigma^2}^{(l)}$ is derived from $\partial Q / \partial \Delta h_{\sigma^2}^{(l)} = 0$, then given by

$$\sum_k^K \gamma_k^{(l)} \frac{\sigma_k^2 + h_{\sigma^2}^{(l-1)} + \Delta h_{\sigma^2}^{(l)} - \phi_k^{(l)}}{(\sigma_k^2 + h_{\sigma^2}^{(l-1)} + \Delta h_{\sigma^2}^{(l)})^2} = 0,$$

$$\phi_k^{(l)} = v_k^{(l)} + m_k^{2^{(l)}} + (\mu_k + h_\mu^{(l)})(\mu_k + h_\mu^{(l)} - 2m_k^{(l)}),$$

$$v_k^{(l)} = \sum_t \gamma_{t,k}^{(l)} (o_t - m_k^{(l)})^2 / \gamma_k^{(l)},$$

where $v_k^{(l)}$ is the variance of adaptation data at l-th iteration by EM algorithm. Here we define the function

$$f(\Delta h_{\sigma^2}^{(l)}) = \frac{\sigma_k^2 + h_{\sigma^2}^{(l-1)} + \Delta h_{\sigma^2}^{(l)} - \phi_k^{(l)}}{(\sigma_k^2 + h_{\sigma^2}^{(l-1)} + \Delta h_{\sigma^2}^{(l)})^2}.$$

$\Delta h_{\sigma^2}^{(l)}$ converges to zero by EM algorithm. Therefore Taylor expansion is able to be applied to the above equation. The first order expansion is computed. Then $\Delta h_{\sigma^2}^{(l)}$ is given by

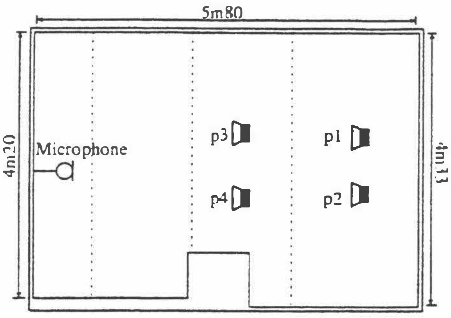


Figure 2. A top view of the experimental room

$$\Delta h_{\sigma^2}^{(l)} = \frac{\sum_{k}^{K} \gamma_k^{(l)} \left\{ \frac{1}{\sigma_k^2 + h_{\sigma^2}^{(l-1)}} - \frac{\phi_k^{(l)}}{(\sigma_k^2 + h_{\sigma^2}^{(l-1)})^2} \right\}}{\sum_{k}^{K} \gamma_k^{(l)} \left\{ \frac{1}{(\sigma_k^2 + h_{\sigma^2}^{(l-1)})^2} - \frac{2\phi_k^{(l)}}{(\sigma_k^2 + h_{\sigma^2}^{(l-1)})^3} \right\}}.$$

7. Compose three distributions according to the equation (1). [1]
8. Repeat the above procedure until the log likelihood probability converges.

The procedure is summarized in Figure 1. The mean and variance of adaptation data are represented by $N(\mu_k^{(l)}, v_k^{(l)})$ at l-th iteration by EM algorithm. The distribution is converted to the linear spectral domain. The decomposition of the distribution and noise is applied in the spectral domain. The obtained distribution is converted to the cepstral domain. Then the acoustical transfer function is decomposed from the obtained distribution in the cepstral domain.

3. EXPERIMENTS AND RESULTS

Recognition experiments are conducted to evaluate effectiveness of the proposed method. Figure 2 shows a top view of the experimental room. The sound signal is captured by using a single directional microphone. We measured 4 transfer functions corresponding to 4 sound source positions by using the method reported in [13]. The length of reverberation time is approximately 180 msec for the experiment 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. The speaker independent(SI) models are trained by using utterances from 64 speakers in the ASJ database. The speaker dependent(SD) models are trained by using 2620 words of one male speaker from the ATR database. The reverberant speech signal is simulated by linear convolution of clean speech signal and measured impulse responses from the positions $p1, \ldots, p4$. The noise signal is collected in a computer room and added to the reverberant speech signal as the SNR is 15dB.

54 context independent phone models are used. Each phoneme HMM is a left-to-right 3-state tied-mixture HMM. There are 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 acoustical transfer function for each position and noise.

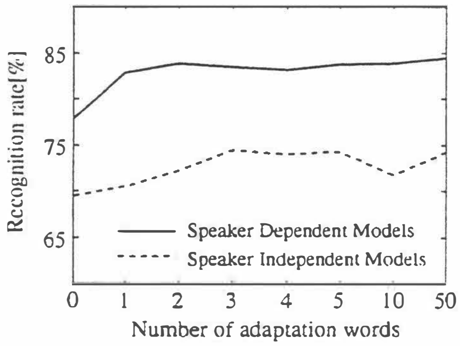


Figure 3. Recognition rates for reverberant speech.

Table 1. Recognition rates[%] with 3 adaptation words for reverberant speech.

Input	HMM-S		Adap-ML	
	SD	SI	SD	SI
Reverberant speech	77.8	69.4	83.4	74.5

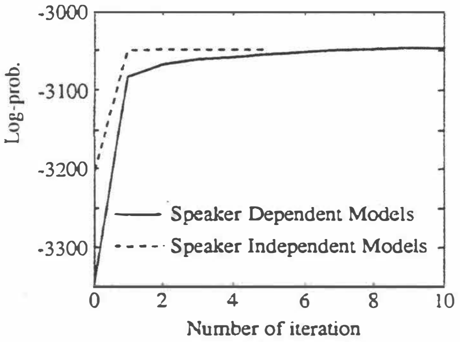


Figure 4. Convergence of the adaptation algorithm for SD and SI models.

The speech recognition is conducted to examine improvements of recognition rates for

- reverberant speech,
- noisy reverberant speech.

Results for reverberant speech are indicated in Figure 3. The Figure shows one male's recognition results averaged over four positions by using different amount of adaptation data. The recognition rates with initial HMMs(clean speech HMMs) for the SD and the SI models are 77.8% and 69.4% respectively. By using the proposed method without the procedure 2~5, the performance is improved with only a few adaptation words. Table 1 shows the recognition rates with 3 adaptation words. The recognition rates for the SD and the SI models are improved from 77.8% and 69.4% to 83.4% and 74.5% respectively. By using the known acoustical transfer function, the recognition rates is 83.5% [1]. These results also show there is no difference between the previously proposed method [1] (the known acoustical transfer function) and the proposed method(the unknown acoustical transfer function).

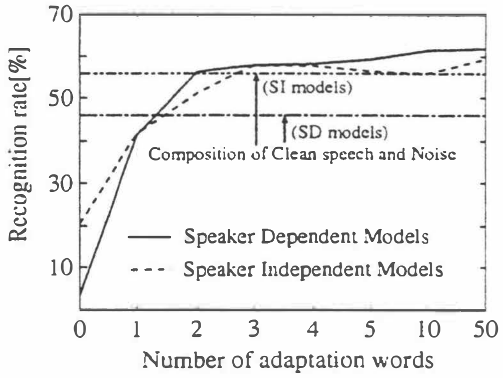


Figure 5. Recognition rates for noisy reverberant speech.

Table 2. Recognition rates[%] with 3 adaptation words for noisy reverberant speech.

Input	HMM-S		Adap-ML	
	SD	SI	SD	SI
Noisy Reverberant speech	3.5	20.4	57.9	57.9

On the other hand, the SD recognition rate using HMM trained by reverberant speech is 96.6%. The performance is supposed to be still insufficient.

Figure 4 shows the convergence property of the proposed method for reverberant speech in the SD and the SI models. The log-likelihood of one adaptation word versus number of iteration in EM algorithm is plotted. The results show that one or two iteration seem enough.

Results for noisy reverberant speech are indicated in Figure 5. The Figure shows one male's recognition results averaged over four positions by using different amount of adaptation data. The recognition rates with initial HMMs(clean speech HMMs) for the SD and the SI models are 3.5% and 20.4% respectively. By using the proposed method, the recognition rates for the SD and the SI models are improved to 57.9% and 57.9% with 3 adaptation words respectively (Table 2). In comparison with the composed HMMs of the clean speech HMMs and the noise HMM, the proposed method achieves slightly higher performance using 2∼3 adaptation words. However the more sophisticated estimation algorithm is necessary for noisy reverberant speech.

4. CONCLUSION

We have presented a new method to estimate HMM parameters of the acoustical transfer function based on the HMM decomposition. This method enables to estimate the parameters of the acoustical transfer function HMM not by the measured impulse responses but by the adaptation speech from the user's location. The experiments indicate that the proposed method improves the recognition rates for the speaker dependent models and the speaker independent models from 3.5% and 20.4% to 57.9% and 57.9% respectively with 3 adaptation words for noisy reverberant speech. The performance of the speaker independent recognition rates is supposed to be still insufficient. The further improvement of the HMM adaptation would be necessary as a future work.

REFERENCES

[1] S.Nakamura, T.Takiguchi and K.Shikano, "Noise and Room Acoustics Distorted Speech Recognition by HMM Composition", *Proc. ICASSP96*, 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] 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.