PAPER *Special Issue on Statistical Modeling for Speech Processing*

Acoustic Model Adaptation Using First-Order Linear Prediction for Reverberant Speech

Tetsuya TAKIGUCHI[†], **Masafumi NISHIMURA**[††], ***and*** **Yasuo ARIKI**[†††], ***Members***

SUMMARY This paper describes a hands-free speech recognition technique based on acoustic model adaptation to reverberant speech. In hands-free speech recognition, the recognition accuracy is degraded by reverberation, since each segment of speech is affected by the reflection energy of the preceding segment. To compensate for the reflection signal we introduce a frame-by-frame adaptation method adding the reflection signal to the means of the acoustic model. The reflection signal is approximated by a first-order linear prediction from the observation signal at the preceding frame, and the linear prediction coefficient is estimated with a maximum likelihood method by using the EM algorithm, which maximizes the likelihood of the adaptation data. Its effectiveness is confirmed by word recognition experiments on reverberant speech.

key words: *acoustic model, reverberant speech, adaptation, hands-free speech recognition*

1. Introduction

In hands-free speech recognition, one of the key issues for practical use is the development of technologies that allow accurate recognition of reverberant speech. Current speech recognition systems are capable of achieving impressive performance in clean acoustic environments. However, if the user speaks at a distance from the microphone, the recognition accuracy is seriously degraded by the influence of reverberation.

Convolution distortion is usually caused by a telephone channel, microphone characteristics, reverberation, and so on. Its effect on the input speech appears as a convolution in the wave domain and is represented as a multiplication in the linear-spectral domain. Conventional normalization techniques, such as CMS (Cepstral Mean Subtraction) and RASTA, have been proposed and their effectiveness has been confirmed for a telephone channel or microphone [1][2][3] that has short impulse responses. When the length of the impulse response is shorter than the analysis window used for the spectral analysis of speech, those methods are effective. However, as the length of the impulse response for the room reverberation becomes longer than the analysis window, the performance degrades. This is because each segment of speech is affected by the reflection energy of the preceding segment in reverberant environments. To reduce the effect of the reverberation, microphone array techniques were proposed [4][5][6][7]. Array processing can offer the additional advantage of spatial processing, but microphone arrays may not be suitable in some cases because of their size and cost.

One scheme for removing the effect of the reverberation is to pass the distorted speech through a filter which exactly inverts the effect of the reverberation. But it is not easy to find the exact inverse filter (for example, see [8]). In [4][5], the dereverberation is realized using the microphone array based on some kind of inverse filtering techniques.

Techniques without microphone arrays were also proposed, e.g. [9][10][11]. If the analysis window is long relative to the length of the reverberation, the effect of the reverberation can be considered as only multiplication in the frequency domain [9]. Therefore, using the long time window, the conventional normalization techniques, such as CMS and RASTA, may reduce reverberation effects [9]. However, the problem is that such a technique will reduce the discrimination of the speech features.

In [10], the reverberation time is estimated, and then the reverberated acoustic model with the closest reverberation time is selected out of a library of offline trained reverberated acoustic models. Therefore if the mismatch of the reverberation factor between the library (database) and real test environments is large, the performance will degrade.

In [11], a new dereverberation method has been proposed. This technique transforms the reverberant signal to its direct signal based on an inverse filtering operation. This is able to effectively reduce the reverberation factor, but the operation requires many reverberant speech signals. It may not be practical to collect a large set of utterances over every environment. Therefore we propose a model adaptation method based on the conventional short-time analysis window, where a small amount of a user's reverberant speech is used.

This paper describes a model adaptation technique for reverberant speech recognition. The new technique is based on HMM composition [12] using a first-order linear prediction. In reverberant environments, the

Manuscript received June 30, 2005.
Manuscript revised June 30, 2005.
Final manuscript received June 30, 2005.

[†]The author was with the IBM Tokyo Research Laboratory, Japan. He is now with the Department of Computer and System Engineering, Kobe University, Japan.
[††]The author is with the IBM Tokyo Research Laboratory, Japan.
[†††]The author is with the Department of Computer and System Engineering, Kobe University, Japan.

speech signal is affected by the reflection energy of the preceding segment. As the model adaptation in [12] was not able to deal with the reflection signal, the recognition performance was not sufficiently improved. In this paper, to compensate for the reflection signal, we introduce a frame-by-frame adaptation method adding the reflection signal to the means of the acoustic model.

In this paper, we approximate the reflection signal of the reverberant speech by the linear prediction from the observation signal at the preceding frame. Adding the reflection signal to the means of the acoustic model, a frame-by-frame adaptation is implemented for reverberant speech. Furthermore, this paper also describes a technique to estimate the linear prediction coefficient. This method estimates the parameters of the reverberation to maximize the likelihood of the adaptation data.

2. HMM adaptation to reverberant speech

The observed signal is generally considered as the addition of the direct signal and the reflection signal:

$$O(\omega;n) \approx S(\omega;n)\cdot H_0(\omega) + \sum_{d=1} S(\omega;n-d)\cdot H_d(\omega) = \sum_{d=0} S(\omega;n-d)H_d(\omega) \quad (1)$$

where $O(\omega;n)$ and $S(\omega;n)$ are the linear spectrum for the observed signal and the clean speech of the frequency ω at the n-th frame. $H(\omega)$ is the reverberation factor. The reflection signal is represented by the summation over the time delay which may be longer than a phoneme interval. This is because the reflection signal can be seen as the overlapping segment from the previous segment. Figure 1 and 2 show clean speech and reverberant speech, respectively. As can be seen from these figures, each segment of reverberant speech is affected by the reflection energy of the preceding segment, and the reflection signal will be viewed as additive noise from the preceding segment of speech.

In this paper, we consider the reflection signal of the reverberant speech as additive noise and approximate it by a linear prediction from the observation signal at the preceding frame. The observed signal is therefore represented by

$$O(\omega;n) \approx S(\omega;n)\cdot H(\omega) + \alpha(\omega)\cdot O(\omega;n-1) \quad (2)$$

where

$$\alpha(\omega)\cdot O(\omega;n-1) = \sum_{d=1} S(\omega;n-d)\cdot H_d(\omega) \quad (3)$$

where $\alpha(\omega)$ is the linear prediction coefficient for the frequency ω. The observation signal $O(\omega;n-1)$ includes all reflection signals. But it includes also the direct signal. Therefore we approximate the reflection signal by a first-order linear prediction from the observation signal at the preceding frame, $O(\omega;n-1)$.

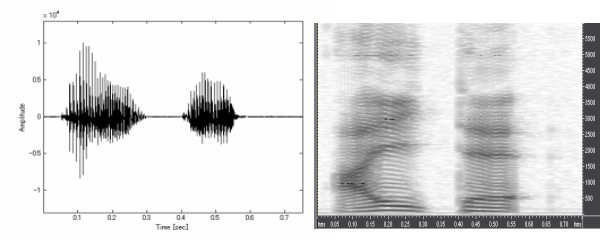


Fig. 1 Original speech: the speech waveform and spectrogram of the Japanese utterance /aite/.

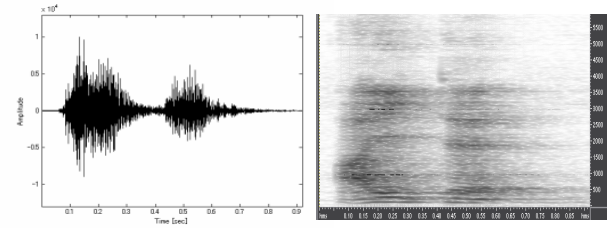


Fig. 2 Reverberant speech (reverberation time = 0.47 sec): the speech waveform and spectrogram of the Japanese utterance /aite/.

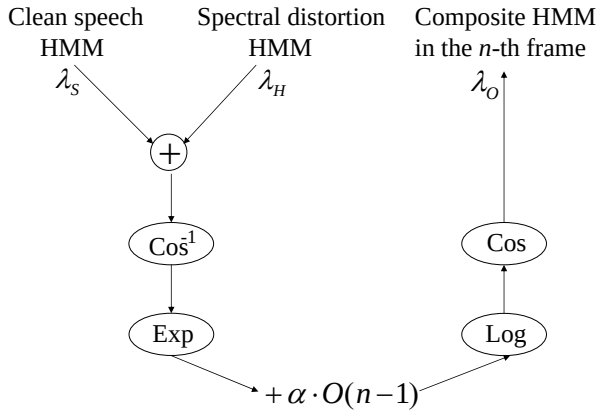


Fig. 3 Frame-by-frame adaptation using a first-order linear prediction

Adding the reflection signal to the means of the acoustic model, a frame-by-frame adaptation is implemented for reverberant speech which has the longer impulse response than the analysis window.

As shown in (2), the reverberation factor is approximated by the addition of the influence within the frame and outside of the frame. Here the former is the spectral distortion within each frame, $H(\omega)$, and the latter is the reflection signal which is approximated by a first-order linear prediction from the observation signal at the preceding frame.

Using (2), the composite HMM for reverberant speech is computed. The procedure is as follows (Figure 3).

1) Compose HMMs of the clean speech and spectral distortion within each frame in the cepstral domain.

$$\mu_{\rm cep}^{(SH)} = \mu_{\rm cep}^{(S)} + \mu_{\rm cep}^{(H)}, \quad \Sigma_{\rm cep}^{(SH)} = \Sigma_{\rm cep}^{(S)} + \Sigma_{\rm cep}^{(H)} \quad (4)$$

Here the subscript cep represents the cepstral domain, $(\mu^{(S)},\ \Sigma^{(S)})$ is the mean vector and covariance matrix of the clean speech HMM, and (H) means the spectral distortion within each frame. In this paper, $\Sigma_{\rm cep}^{(H)}$ is set to zero.

2) Transform $(\mu_{\rm cep}^{(SH)},\ \Sigma_{\rm cep}^{(SH)})$ from the cepstral domain to the linear-spectral domain.

2.1) Compute the inverse cosine transform of each Gaussian probability density function (PDF) of the HMM's.

$$\mu_{\log}^{(SH)} = \Gamma^{-1}\mu_{\rm cep}^{(SH)}, \quad \Sigma_{\log}^{(SH)} = (\Gamma^{-1})^T \Sigma_{\rm cep}^{(SH)} \Gamma^{-1} \quad (5)$$

Here, Γ is a cosine transform matrix, $\mu_{\log}^{(SH)}$, and $\Sigma_{\log}^{(SH)}$ are the mean vector and covariance matrix of a Gaussian PDF in the log-power spectral domain. The transposition is denoted by "T".

2.2) Compute the exponential transform to the linear-spectral domain. The normal random vector obtained by exponential transform, $Z = \exp(Y)$, has log-normal distribution. The mean and covariance are given by

$$\mu_{\rm lin,i}^{(SH)} = \exp\left\{\mu_{{\rm log},i}^{(SH)} + \frac{\sigma_{{\rm log},ii}^{(SH)}}{2}\right\} \quad (6)$$

$$\sigma_{{\rm lin},ij}^{(SH)} = \mu_{{\rm lin},i}^{(SH)} \cdot \mu_{{\rm lin},j}^{(SH)} \cdot \exp\left\{\sigma_{{\rm log},ij}^{(SH)} - 1\right\} \quad (7)$$

Here, $\mu_{{\rm lin},i}^{(SH)}$ and $\sigma_{{\rm lin},ij}^{(SH)}$ are the i-th mean and the (i,j) element of the covariance matrix in the linear-spectral domain.

3) Frame-by-frame adaptation to the reverberant speech using the preceding frame. Add the reflection signal estimated by the linear prediction from the observation signal at the preceding frame to the means of the acoustic model.

$$\hat{\mu}_{\rm lin}^{(O)} = \mu_{\rm lin}^{(SH)} + \alpha \cdot O_{\rm lin}(n-1) \quad (8)$$

$$\hat{\sigma}_{{\rm lin},ij}^{(O)} = \sigma_{{\rm lin},ij}^{(SH)} \quad (9)$$

4) Transform $(\hat{\mu}_{\rm lin}^{(O)},\ \hat{\Sigma}_{\rm lin}^{(O)})$ from the linear-spectral domain to the cepstral domain.

4.1) Compute the log transform.

$$\hat{\mu}_{{\rm log},i}^{(O)} = \log \hat{\mu}_{{\rm lin},i}^{(O)} - \frac{1}{2}\left\{\frac{\hat{\sigma}_{{\rm lin},ii}^{(O)}}{\hat{\mu}_{{\rm lin},i}^{(O)} \cdot \hat{\mu}_{{\rm lin},i}^{(O)}} + 1\right\} \quad (10)$$

$$\hat{\sigma}_{{\rm log},ij}^{(O)} = \log\left\{\frac{\hat{\sigma}_{{\rm lin},ij}^{(O)}}{\hat{\mu}_{{\rm lin},i}^{(O)} \cdot \hat{\mu}_{{\rm lin},j}^{(O)}} + 1\right\} \quad (11)$$

4.2) Compute the cosine transform to the cepstral domain.

$$\hat{\mu}_{\rm cep}^{(O)} = \Gamma\hat{\mu}_{\log}^{(O)}, \quad \hat{\Sigma}_{\rm cep}^{(O)} = \Gamma^T \hat{\Sigma}_{\log}^{(O)} \Gamma \quad (12)$$

Given the composite HMM for the reverberant speech, a speech recognition system estimates the word string associated with the test waveform.

In (2), the reflection signal is approximated by a first-order linear prediction from the observation signal at the preceding frame. We next consider about this approximation. From (1), the following equation is obtained:

$$O(\omega; n-1) \approx \sum_{d=0} S(\omega; n-1-d) \cdot H_d(\omega)$$

$$\alpha(\omega) \cdot O(\omega; n-1) \approx \sum_{d=0} S(\omega; n-1-d) \cdot \alpha(\omega) \cdot H_d(\omega) \quad (13)$$

Comparison of the above equation with (3), the following equation is finally obtained:

$$\begin{aligned} H_1 &= \alpha H_0 \\ H_2 &= \alpha H_1 = \alpha\alpha H_0 \\ H_3 &= \alpha H_2 = \alpha\alpha\alpha H_0 \\ &\cdots \end{aligned}$$

Thus in the proposed method the effect of the reverberation decreases according to the product of α, as the time delay increases.

This section has only described how to adapt the acoustic model to reverberant speech. Therefore estimation of the reverberant parameters remains a serious problem. The next section describes how to estimate the linear prediction coefficient.

3. Estimation of reverberant parameters

Estimations of the spectral distortion within each frame

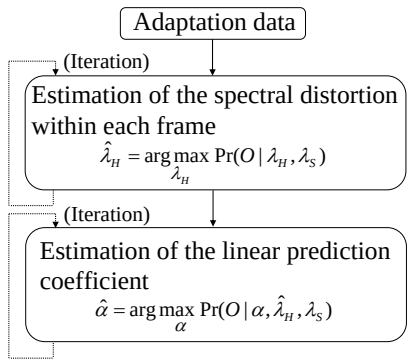


Fig. 4 Estimation of reverberant parameters using EM algorithm

and the linear prediction coefficient are performed by maximizing the likelihood of the adaptation data. First the spectral distortion is estimated using HMM separation [12] in the cepstral domain, where α is set to zero. Then the linear prediction coefficient is estimated in the linear-spectral domain. The steps to estimate the reverberant parameters are as follows (Figure 4):

1) Estimate the spectral distortion using the HMM separation [12] based on the Expectation-Maximization (EM) in the cepstral domain.

$$\hat{\lambda}_H = \underset{\lambda_H}{\operatorname{argmax}} \Pr(O|\lambda_H, \lambda_S) \tag{14}$$

Here λ denotes the set of HMM parameters. In this paper, we apply the HMM separation to only the mean vector. The re-estimation formula for $\hat{\lambda}_H$ is given by

$$\hat{\mu}^{(H)} = \frac{\sum_p^P \sum_v^{W_p} \sum_j \sum_k \sum_n^{N_{p,v}} \gamma_{p,v,j,k,n} \frac{O_{p,v,n} - \mu_{p,j,k}^{(S)}}{\Sigma_{p,j,k}^{(S)}}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}}{\Sigma_{p,j,k}^{(S)}}}$$

$$= \frac{\sum_p \sum_j \sum_k \gamma_{p,j,k} \frac{\mu_{p,j,k}^{(O')} - \mu_{p,j,k}^{(S)}}{\Sigma_{p,j,k}^{(S)}}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}}{\Sigma_{p,j,k}^{(S)}}} \tag{15}$$

$$\mu_{p,j,k}^{(O')} = \frac{\sum_v \sum_n \gamma_{p,v,j,k,n} O_{p,v,n}}{\sum_v \sum_n \gamma_{p,v,j,k,n}} \tag{16}$$

$$\gamma_{p,v,j,k,n} = \Pr(O_{p,v,n}, j, k|\lambda_H, \lambda_S) \tag{17}$$

where $\mu_{p,j,k}^{(S)}$ and $\Sigma_{p,j,k}^{(S)}$ are the means and variances corresponding to a phoneme p, state j, and mixture k in the model λ_S. Each phoneme consists of W_p adaptation data, and $\sum_v N_{p,v}$ is the total number of training frames for the phoneme p. $O_{p,v,n}$ is the n-th observation sequence in the v-th adaptation data for a phoneme p.

2) Compose the HMMs of the clean speech, λ_S, and the spectral distortion, $\hat{\lambda}_H$, in the cepstral domain according to (4).
3) Transform ($\hat{\mu}_{\text{cep}}^{(SH)}$, $\hat{\Sigma}_{\text{cep}}^{(SH)}$) from the cepstral domain to the linear-spectral domain.
4) Estimate the linear prediction coefficient.

$$\begin{aligned}\hat{\alpha} &= \underset{\alpha}{\operatorname{argmax}} \Pr(O|\alpha, \hat{\lambda}_H, \lambda_S) \\ &= \underset{\alpha}{\operatorname{argmax}} \Pr(O|\alpha, \hat{\lambda}_{SH})\end{aligned} \tag{18}$$

The estimation of the linear prediction coefficient is performed in a maximum likelihood fashion by using the Expectation-Maximization (EM) algorithm. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step, the following auxiliary function is computed.

$$\begin{aligned}&Q(\hat{\alpha}|\alpha) \\ &= E[\log \Pr(O, b, c|\hat{\alpha}, \hat{\lambda}_{SH})|\alpha, \hat{\lambda}_{SH}] \\ &= \sum_p \sum_v \sum_{b_{p,v}} \sum_{c_{p,v}} \frac{\Pr(O_{p,v}, b_{p,v}, c_{p,v}|\alpha, \hat{\lambda}_{SH})}{\Pr(O_{p,v}|\alpha, \hat{\lambda}_{SH})} \\ &\quad \cdot \log \Pr(O_{p,v}, b_{p,v}, c_{p,v}|\hat{\alpha}, \hat{\lambda}_{SH})\end{aligned} \tag{19}$$

Here b and c are the unobserved state sequence and the unobserved mixture component labels corresponding to the observation sequence O.

The joint probability of observing the sequences O, b, and c can be calculated as

$$\begin{aligned}&\Pr(O, b, c|\hat{\alpha}, \hat{\lambda}_{SH}) \\ &= \prod_n a_{b_{n-1}, b_n} w_{b_n, c_n} \Pr(O_n|\hat{\alpha}, \hat{\lambda}_{SH})\end{aligned} \tag{20}$$

where a is the transition probability, and w is the mixture weight. Since we consider reflection signal of the reverberant speech as additive noise and approximate it by a linear prediction from the preceding frame, the mean to mixture k in the model λ_O is derived by adding the reflection signal estimated by linear prediction from the observation signal at the preceding frame to the mean of the acoustic model $\hat{\lambda}_{SH}$. Therefore, (20) can be written as

$$\begin{aligned}&\Pr(O, b, c|\hat{\alpha}, \hat{\lambda}_{SH}) \\ &= \prod_n a_{b_{n-1}, b_n} w_{b_n, c_n} \cdot N(O; \hat{\mu}_{p,j,k_n}^{(SH)} + \hat{\alpha} O_{n-1}, \hat{\Sigma}_{p,j,k_n}^{(SH)})\end{aligned} \tag{21}$$

where $N(O; \mu, \Sigma)$ denotes the multivariate Gaussian distribution. It is straightforward to derive that [13]

$$\begin{aligned}&Q(\hat{\alpha}|\alpha) \\ &= \sum_p \sum_v \sum_i \sum_j \sum_n \\ &\quad \Pr(O_{p,v}, b_{p,v,n} = j, b_{p,v,n-1} = i|\hat{\lambda}_{SH}) \log a_{p,i,j} \\ &\quad + \sum_p \sum_v \sum_j \sum_k \sum_n \\ &\quad \Pr(O_{p,v}, b_{p,v,n} = j, c_{p,v,n} = k|\hat{\lambda}_{SH}) \log w_{p,j,k} \\ &\quad + \sum_p \sum_v \sum_j \sum_k \sum_n \\ &\quad \Pr(O_{p,v}, b_{p,v,n} = j, c_{p,v,n} = k|\hat{\lambda}_{SH}) \\ &\quad \cdot \log N(O_{p,v,n}; \hat{\mu}_{p,j,k}^{(SH)} + \hat{\alpha} O_{n-1}, \hat{\Sigma}_{p,j,k}^{(SH)})\end{aligned} \tag{22}$$

Here we focus only on the term involving ($\hat{\theta} = \{\hat{\alpha}\}$).

$$\begin{aligned}&Q_{\hat{\theta}}(\hat{\alpha}|\alpha) \\ &= \sum_p \sum_v \sum_j \sum_k \sum_n \\ &\quad \Pr(O_{p,v}, b_{p,v,n} = j, c_{p,v,n} = k|\hat{\lambda}_{SH}) \\ &\quad \cdot \log N(O_{p,v,n}; \hat{\mu}_{p,j,k}^{(SH)} + \hat{\alpha} O_{p,v,n-1}, \hat{\Sigma}_{p,j,k}^{(SH)})\end{aligned}$$

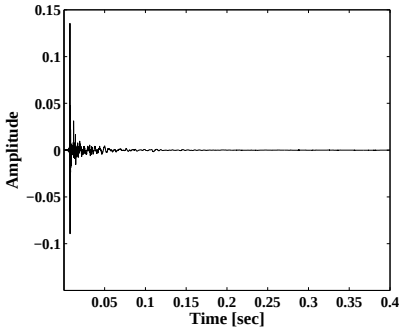


Fig. 5 Impulse response (Reverberation time: 300 msec) which is measured using the TSP (Time-Stretched Pulse) method. Reverberant speech is simulated by a linear convolution of clean speech and impulse responses.

$$= -\sum_p \sum_v \sum_j \sum_k \sum_n \gamma_{p,v,j,k,n} \cdot \left[\frac{1}{2} \log(2\pi)^D \hat{\Sigma}^{(SH)}_{p,j,k} + \frac{\{O_{p,v,n} - \hat{\mu}^{(SH)}_{p,j,k} - \hat{\alpha} \cdot O_{p,v,n-1}\}^T \{O_{p,v,n} - \hat{\mu}^{(SH)}_{p,j,k} - \hat{\alpha} \cdot O_{p,v,n-1}\}}{2\hat{\Sigma}^{(SH)}_{p,j,k}} \right] \quad (23)$$

Here D is the dimension of the adaptation vector $O_{p,v,n}$. In this work, we assume that the alignment for the adaptation data in the linear-spectral domain is the same as that in the cepstral domain. Therefore the probability, γ, of being in state j and mixture k at frame n is computed in the cepstral domain.

The maximization step (M-step) in the EM algorithm becomes "max $Q_{\hat{\theta}}(\hat{\alpha}|\alpha)$". The re-estimation formula can be therefore derived from knowing that $\partial Q(\hat{\alpha}|\alpha)/\partial\hat{\alpha} = 0$ as

$$\hat{\alpha} = \frac{\sum_p \sum_v \sum_j \sum_k \sum_n \gamma_{p,v,j,k,n} \frac{O_{p,v,n-1}\{O_{p,v,n} - \hat{\mu}^{(SH)}_{p,j,k}\}}{\hat{\Sigma}^{(SH)}_{p,j,k}}}{\sum_p \sum_v \sum_j \sum_k \sum_n \gamma_{p,v,j,k,n} \frac{O^2_{p,v,n-1}}{\hat{\Sigma}^{(SH)}_{p,j,k}}} \quad (24)$$

4. Experiments

4.1 Experimental conditions

The new adaptation technique was evaluated on distant-talking speech recognition tasks. Reverberant speech was simulated by a linear convolution of clean speech and impulse responses. The impulse responses were taken from the RWCP sound scene database [14][15]. The reverberation time was 300 msec (Figure 5). The distance to the microphone was about 2 m. The size of the recording room was about 6.7 m × 4.2 m (width × depth). The speech signal was sampled at 12 kHz and windowed with a 32-msec Hamming window every 8 msec. Then FFT is used to compute 16-order MFCCs (mel-frequency cepstral coefficients) and the power. In recognition, the power term is not used, because it is only necessary to adjust the power of the clean speech model in (8).

The models of 55 context-independent phonemes were trained by using 2,620 words in the ATR Japanese speech database for the speaker-dependent HMM. Each HMM has three states and three self-loops, and each state has four Gaussian mixture components. The tests were carried out on 1000-word recognition tasks, and three males spoke the 1000 words. Each test speaker uttered 10 words as adaptation data, different from those used in the training and testing.

To evaluate the proposed method for the mismatch between the adaptation and testing positions and compare it with the inverse filtering method, we used four impulse responses. Figure 7 shows the mismatch conditions. We used the measured impulse response to calculate the inverse filter.

Table 1 Word-recognition rates for reverberant speech

method	CMS	model adap.		matched
spectral distortion compensation	○	○	○	-
additive reflection compensation	×	×	○	-
speaker1	78.5%	80.3%	88.8%	94.2%
speaker2	88.1%	90.1%	91.1%	96.2%
speaker3	73.3%	79.8%	85.4%	93.9%
average	80.0%	83.4%	88.4%	94.8%

4.2 Experimental results

Table 1 shows the recognition rates for reverberant speech. In the CMS-based testing case, the phoneme HMMs are trained by using the CMS-processed clean-speech data. Subtraction of each cepstral mean value from each set of test data gives an average recognition rate of 80.0%. The result clearly shows that the simple CMS technique does not work well. As can be seen from this table, the use of the model adaptation achieves good performance, comparable with that of CMS in the reverberant environment. The use of the model adaptation without the additive reflection compensation using only (4) improved the recognition rate to 83.4%, and a further improvement was also obtained by the adaptation with additive reflection compensation using (8). However, comparing the result of the model adaptation with that of the matched model which was trained by using reverberant speech (2,620 words) shows degradation in performance.

Figure 6 shows the convergence properties of the model adaptation. In this figure, the log-likelihood versus the number of iterations in the EM algorithm is plotted. As can be seen from Figure 6, the EM algorithm converges within several iterations.

Figure 10 shows a comparison of the performance of the model adaptation and the inverse filtering. The inverse filtering requires the measurement of the impulse response from the position of the sound source to the microphone, and its inverse is used to dereverberate the speech signal according to

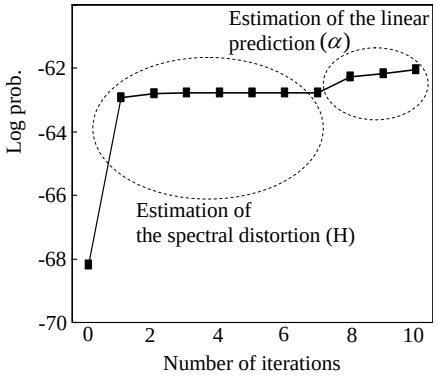


Fig. 6 Convergence of the EM algorithm

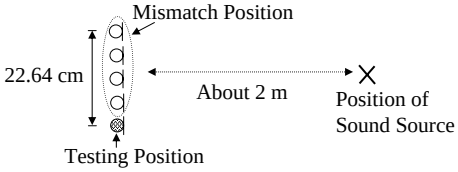


Fig. 7 Experimental condition for inverse filtering.

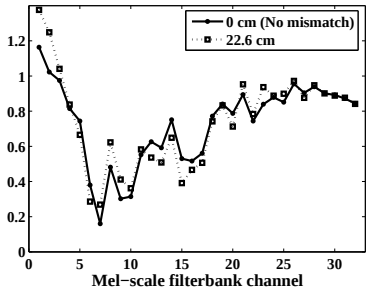


Fig. 8 Estimated linear prediction coefficient (α).

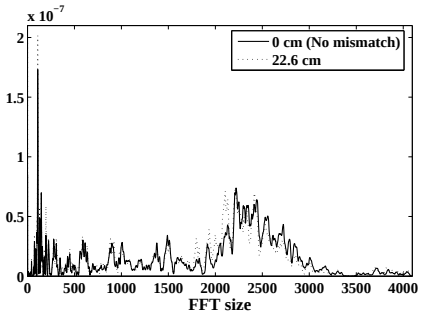


Fig. 9 Power-spectrum of impulse response.

$$\hat{S}(\omega) = F[o(t)]/F[w(t)] \tag{25}$$

$$\hat{s}(t) = F^{-1}[\hat{S}(\omega)] \tag{26}$$

where $w(t)$ is the measured impulse response, $F[*]$ is the one-time Fourier transform, and $\hat{S}(\omega)$ is the complex spectrum of the estimated clean speech. In this

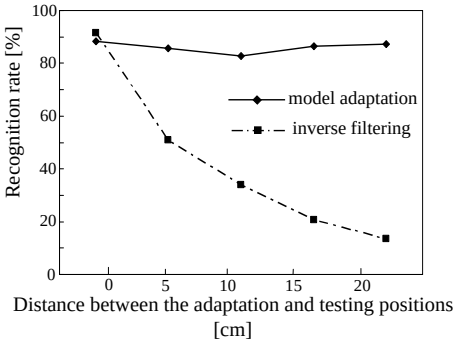


Fig. 10 Comparison of the performance of model adaptation and inverse filtering

experiment, from $\hat{s}(t)$, 16-order MFCCs are computed. Figure 7 shows the experimental condition for the inverse filtering, where the microphone position for the adaptation data and the inverse filtering is changed, and that for the testing data is fixed (The position of the test speaker is also fixed). Figure 8 and 9 show the estimated linear prediction coefficient and the power-spectrum of the impulse responses in the "no-mismatch" case and the "mismatch-distance = 22.64 cm" case. The differences shown may cause degradation of speech recognition.

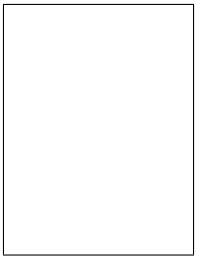
As shown in Figure 10, the performance of both approach with no mismatch between the adaptation and testing positions is very good. As the mismatch of the positions becomes large, the performance of the inverse filtering is decreased. For the model adaptation the performance is not decreased. The result shows that the model adaptation is a robust technology to the mismatch between the adaptation and testing positions.

5. Summary

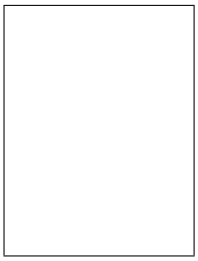
This paper has described an acoustic model adaptation technique for reverberant speech recognition. In this paper, we assume that the influence of the reverberation contributes as the spectral distortion within each frame and as additive noise, which is approximated by a first-order linear prediction from the observation signal at the preceding frame. The linear prediction coefficient is estimated using the EM algorithm from a small amount of a user's speech. Adding the reflection signal to the means of the acoustic model, a frame-by-frame adaptation is implemented for reverberant speech. The new adaptation technique was evaluated on distant-talking speech recognition tasks. The experimental results show that the use of the model adaptation achieves good performance in comparison to that of CMS, and the model adaptation is robust to the mismatch between the adaptation and testing positions in comparison with the inverse filtering approach.

References

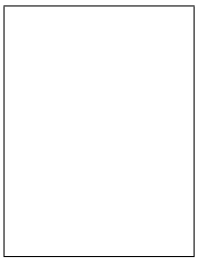
[1] J. Chang and V. Zue, "A Study of Speech Recognition System Robustness to Microphone Variations: Experiments in Phonetic Classification," ICSLP, pp. 995-998, 1994.
[2] M. G. Rahim and B.-H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," IEEE Trans. on SAP, Vol. 4, No. 1, pp. 19-30, 1996.
[3] H. Hermansky and N. Morgan, "RASTA Processing of Speech," IEEE Trans. on SAP, Vol. 2, No. 4, pp. 578-589, 1994.
[4] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," IEEE Trans. on ASSP, Vol. 36, No. 2, 1988.
[5] H. Wang and F. Itakura, "An Approach of Dereverberation using Multi-Microphone Sub-Band Envelope Estimation," ICASSP, pp. 953-956, 1991.
[6] Q.-G. Liu, B. Champagne, and P. Kabal, "A microphone array processing technique for speech enhancement in a reverberant space," Speech Communication 18, pp. 317-334, 1996.
[7] P. W. Shields and D. R. Campbell, "Intelligibility improvements obtained by an enhancement method applied to speech corrupted by noise and reverberation," Speech Communication, 25, pp. 165-175, 1998.
[8] M. Tohyama, H. Suzuki and Y. Ando, "The Nature and Technology of Acoustic Space," ACADEMIC PRESS, 1995.
[9] C. Avendano, S. Tivrewala, and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," Eurospeech, pp. 1107-1110, 1997.
[10] L. Couvreur and C. Couvreur, "Robust automatic speech recognition in reverberant environments by model selection," International Workshop on Hands-Free Speech Communication, pp. 147-150, 2001.
[11] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," ICASSP pp. 92-95, 2003.
[12] T. Takiguchi, S. Nakamura, and K. Shikano, "HMM-Separation-Based Speech Recognition for a Distant Moving Speaker," IEEE Trans. on SAP, Vol. 9, No. 2, pp. 127-140, 2001.
[13] B.-H. Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains," AT&T Tech. J., Vol. 64, No. 6, pp. 1235-1249, 1985
[14] S. Nakamura, "Acoustic sound database collected for hands-free speech recognition and sound scene understanding," International Workshop on Hands-Free Speech Communication, pp. 43-46, 2001.
[15] http://tosa.mri.co.jp/sounddb/micarray/indexe.htm



Tetsuya Takiguchi received the B.S. degree in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and the M.E. and Dr. Eng. degrees in information science from Nara Institute of Science and Technology, Nara, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Kanagawa, Japan. He is currently a Lecturer with Kobe University. His research interests include robust speech recognition, auditory scene analysis, and microphone arrays. He received the Awaya Award from the Acoustical Society of Japan in 2002. He is a member of the IEEE, the Information Processing Society of Japan, and the Acoustical Society of Japan.



Masafumi Nishimura received his B.E. and M.E. degrees in Biophysical Engineering from Osaka University in 1981 and 1983, and his Dr. Eng. degree from Toyohashi University of Technology in 1998. In 1983 he joined the Tokyo Research Laboratory of IBM Japan, working on speech and language processing. He is currently a group leader of the speech technology group of TRL. He received the SIG Research Award from the Information Processing Society of Japan in 1998, and the Outstanding Technological Development in Acoustics Prize from the Acoustical Society of Japan in 1999. He is a member of the IPSJ and the ASJ.



Yasuo Ariki received his B.E., M.E. and Ph.D. in information science from Kyoto University in 1974, 1976 and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990, and stayed at Edinburgh University as visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. Since 2003 he has been a professor at Kobe University. He is mainly engaged in speech and image recognition and interested in information retrieval and database. He is a member of IEEE, IPSJ, JSAI, ITE and IIEEJ.