ACOUSTIC MODEL ADAPTATION USING FIRST ORDER PREDICTION FOR REVERBERANT SPEECH

Tetsuya Takiguchi and Masafumi Nishimura

IBM Research, Tokyo Research Laboratory,
1623-14, Shimotsuruma, Yamato-shi, Kanagawa, 242-8502, JAPAN
{takigu, nisimura}@jp.ibm.com

ABSTRACT

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 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.

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. Thus approach without microphone arrays are also proposed, e.g. [8][9].

This paper describes a model adaptation technique for reverberant speech recognition. The new technique is based on HMM composition [10] using a first-order linear prediction. In this paper, we approximate the reflection signal of the reverberant speech by the linear prediction from 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

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

$$\hat{O}(\omega;t) \approx S(\omega;t) \cdot H(\omega) + \alpha(\omega) \cdot O(\omega;t-1) \qquad (1)$$

where $O(\omega;t)$ and $S(\omega;t)$ are the linear spectrum for the observed signal and the clean speech of the frequency ω at the t-th frame, $H(\omega)$ is the spectral distortion within each frame, and $\alpha(\omega)$ is the linear prediction coefficient for the frequency ω.

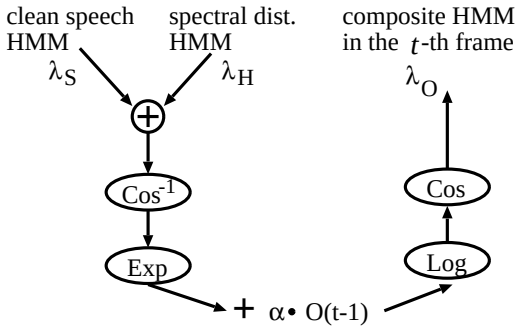

Figure 1: Frame-by-frame adaptation using a first-order linear prediction

Using Equation (1), the composite HMM for reverberant speech is computed. The procedure is as follows (Figure 1).

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

$$\mu_{\rm cep}^{(SH)} = \mu_{\rm cep}^{(S)} + \mu_{\rm cep}^{(H)}, \quad \Sigma_{\rm cep}^{(SH)} = \Sigma_{\rm cep}^{(S)} + \Sigma_{\rm cep}^{(H)} \quad (2)$$

Here the subscript cep represents the cepstral domain, $(\mu^{(S)},\ \Sigma^{(S)})$ is the means and variances of the clean speech HMM, and (H) means the spectral distortion within each frame.

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

3) Frame-by-frame adaptation to the reverberant speech using the preceding frame.

3.1) Add the reflection signal estimated by the linear prediction from the preceding frame to the means of the acoustic model.

$$\hat{\mu}_{\rm lin}^{(O)} = \mu_{\rm lin}^{(SH)} + \alpha \cdot O_{\rm lin}(t-1) \quad (3)$$

Here the subscript lin represents the linear-spectral domain.

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

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

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 and the spectral distortion within each frame.

3. ESTIMATION OF REVERBERANT PARAMETERS

Estimations of the spectral distortion within each frame and the linear prediction coefficient are performed by maximizing the likelihood of the adaptation data. First the spectral distortion is estimated using HMM separation [10] 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 2):

1) Estimate the spectral distortion using the HMM separation [10] in the cepstral domain.

$$\hat{\lambda}_H = \underset{\lambda_H}{\rm argmax} \Pr(O|\lambda_H, \lambda_S) \quad (4)$$

Here λ denotes the set of HMM parameters.

2) Compose the HMMs of the clean speech, λ_S, and the spectral distortion, $\hat{\lambda}_H$, in the cepstral domain according to Equation (2).

3) Transform $(\hat{\mu}_{\rm cep}^{(SH)},\ \hat{\Sigma}_{\rm cep}^{(SH)})$ from the cepstral domain to the linear-spectral domain.

4) Estimate the linear prediction coefficient.

$$\begin{aligned} \hat{\alpha} &= \underset{\alpha}{\rm argmax} \Pr(O|\alpha, \hat{\lambda}_H, \lambda_S) \\ &= \underset{\alpha}{\rm argmax} \Pr(O|\alpha, \hat{\lambda}_{SH}) \end{aligned} \quad (5)$$

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.

$$Q(\hat{\alpha}|\alpha) = E[\log \Pr(O|\hat{\alpha}, \hat{\lambda}_{SH_{\rm lin}})|\alpha, \hat{\lambda}_{SH_{\rm lin}}] \quad (6)$$

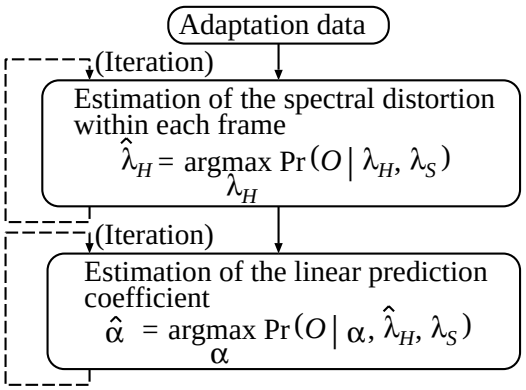

Figure 2: Estimation of reverberant parameters using EM algorithm

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

$$
\begin{aligned}
& Q_{\hat{\theta}}(\hat{\alpha}|\alpha) \\
&= -\sum_p \sum_j \sum_k \sum_t \gamma_{p,j,k}(t) \cdot \left[\frac{1}{2} \log(2\pi)^D \hat{\Sigma}^{(SH)}_{p,j,k} \right. \\
&\left. + \frac{\{O_p(t) - \hat{\mu}^{(SH)}_{p,j,k} - \hat{\alpha} \cdot O_p(t-1)\}^T \{O_p(t) - \hat{\mu}^{(SH)}_{p,j,k} - \hat{\alpha} \cdot O_p(t-1)\}}{2\hat{\Sigma}^{(SH)}_{p,j,k}} \right] \quad (7)
\end{aligned}
$$

$$
\gamma_{p,j,k}(t) = \Pr(s_p(t) = j, m_p(t) = k | O_p, \hat{\lambda}_{SH}, \alpha) \quad (8)
$$

Here $\hat{\mu}^{(SH)}_{p,j,k}$ and $\hat{\Sigma}^{(SH)}_{p,j,k}$ are the mean and variance corresponding to a phoneme p, state j, and mixture k in the model $\hat{\lambda}_{SH_{\rm lin}}$, O_p is the observation sequence (adaptation data) for a phoneme p, and D is the dimension of the adaptation vector $O_p(t)$. 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 time t 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

$$
\hat{\alpha} = \frac{\sum_p \sum_j \sum_k \sum_t \gamma_{p,j,k}(t) \frac{O_p(t-1)\{O_p(t) - \hat{\mu}^{(SH)}_{p,j,k}\}}{\hat{\Sigma}^{(SH)}_{p,j,k}}}{\sum_p \sum_j \sum_k \sum_t \gamma_{p,j,k}(t) \frac{O^2_p(t-1)}{\hat{\Sigma}^{(SH)}_{p,j,k}}} . \quad (9)
$$

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 [11]. The length of the impulse response was 300 msec. The distance to the microphone was 2 m. 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 Equation (3).

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. Also, a single Gaussian is employed to model the spectral distortion within each frame. The tests were carried out on 500-word recognition tasks, and one male spoke the 500 words. The test speaker uttered 10 words as adaptation data, different from those used in the training and testing.

Table 1: Word-recognition rates for reverberant speech

method	CMS	model adap.		matched
spectral distortion compensation	○	○	○	○
additive reflection compensation	×	×	○	○
recognition rate	86.0%	91.2%	94.0%	96.4%

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 a recognition rate of 86.0%. 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 Equation (2) improved the recognition rate to 91.2%, and a further improvement was also obtained by the adaptation with additive reflection compensation using Equation (3). 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 a slight degradation in performance.

Figure 3 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 3, the EM algorithm converges within several iterations.

Figure 4 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 $F[o(t)]/F[w(t)]$, where $w(t)$ is the measured impulse response and $F[*]$ is the Fourier transform. The performance of both approach with no mismatch between the adaptation and testing positions is very good. Here the term "adaptation position" is the position where the test speaker uttered 10 words as the adaptation data for the model adaptation approach and the position where we measured the impulse response for the inverse filtering. As

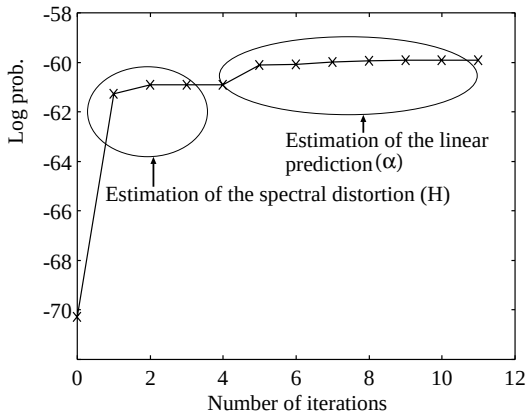

Figure 3: Convergence of the EM algorithm

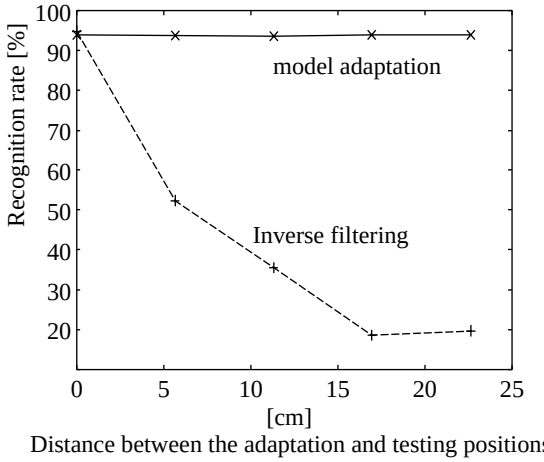

Figure 4: Comparison of the performance of model adaptation and inverse filtering

the mismatch of the positions becomes large, the performance of the inverse filtering is decreased. For the model adaptation the performance is not decreased.

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 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.

6. 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] C. Avendano, S. Tivrewala, and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," Eurospeech, pp. 1107-1110, 1997.

[9] 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.

[10] 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.

[11] 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.