ADAPTATION OF MODEL PARAMETERS TO DISTORTED SPEECH IN ADVERSE ENVIRONMENTS BY HMM DE/COMPOSITION

Satoshi NAKAMURA, Tetsuya TAKIGUCHI, Kiyohiro SHIKANO
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama Ikoma, Nara, 630-01 JAPAN
e-mail: nakamura@is.isist-nara.ac.jp

ABSTRACT
This paper presents a robust speech recognition method based on the HMM composition and decomposition for the noisy reverberant speech. The method realizes an improved user interface such as the user is not encumbered by microphone equipments. The proposed HMM composition is obtained by natural extension of the HMM composition method of an additive noise to that of the convolutional distortion of the reverberation. The HMM composition is conducted by 2 steps: 1) Composition of HMMs of a speech and acoustical transfer function in the cepstrum domain, 2) Composition of the distorted speech and noise HMMs in the linear spectral domain. Furthermore the paper proposes a novel estimation algorithm of acoustical transfer function based on HMM decomposition. The speaker dependent/independent word recognition experiments are carried out using the speech database contaminated by the additive noise and convolutional distortion of the reverberation. These results clarified the effectiveness of the proposed method.

1. INTRODUCTION
In spite of recent advances, the speech recognition technology did not reach to the practical use in the real world. The reason is that the advances are almost achieved in the error reduction of the clean speech recognition.

A key issue way to the practical use is a development of a hands-free speech recognition technology in noisy and reverberant environments. This technology will especially take an important role on recognition of distant-talking speech.

The accuracy of speaker independent speech recognition is made a remarkable progress by the arrival of stochastic modeling of speech, HMM, and its training algorithms. Although the HMM brought a high recognition accuracy, a speaker must be equipped a close-talking microphone. If the speaker inputs his speech from the distance or through a telephone channel, the accuracy will be drastically degraded by the influences of the room acoustics or telephone channel distortion and environment noises. Therefore we have to overcome the two problems such as,

- Additive noise
- Convolutional distortion

Many works are presented to solve these problems. These approach are summarized as follows:

- Speech Enhancement
- Model Adaptation

As for the speech enhancement approach, the spectral subtraction method for an additive noise and the cepstral mean normalization method for a convolutional noise had been proposed and confirmed their effectiveness[1, 2]. As for the model modification approach, the conventional multi-template approach, and model adaptation approach[7, 13] and the model (de-)composition approach[3, 4, 6, 10, 9] had been proposed. Among these approaches the HMM composition approach is the most promising, because the HMM for the noisy speech can be easily generated by composing the speech HMMs and the noise HMM which trained during noise period. The papers[4, 6] showed the composed noisy HMM outperforms very high accuracy.

In this paper, we apply the HMM composition to the recognition of the speech which is contaminated by not only an additive noise but also the room reverberation. If the components are independent each other and additive, HMM composition can be adopted. The noise and speech are independent and additive in the linear spectral domain. While the transfer function and speech are convolutional in the linear spectral domain, they are independent and additive in the cepstral domain. Thus the HMM composition is applicable for noisy and reverberant speech. Some studies have been already presented for the problem of spectral tilt compensation of speech with a noise and channel distortion[5, 8]. This paper addresses the compensation not only the spectral tilt but also room acoustical transfer function.

This paper presents the HMM composition algorithm for the noisy and reverberant speech and evaluates its effectiveness by the recognition experiments of the distant-talking speech in a noisy room, where the speech is suffered from the noise and reverberation[10]. Furthermore the paper proposes a novel estimation algorithm of the stochastic acoustical transfer function model based on HMM decomposition[11, 12].

2. HMM COMPOSITION

On the assumption that speech signal, s(t), and noise signal, n(t), are independent, the observed signal, o(t) is represented by

\[ o(t) = s(t) + n(t) \]
The conventional approach estimates noise statistics during the noise period and recognizes an input noisy speech by using the noise added reference patterns. Since the signal level is generally different between training and testing, an adjustment factor $k$ is introduced. Thus the observed signal is represented by

$$o(t) = s(t) + k \cdot n(t)$$

This relation is preserved in the linear spectral domain.

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

where $O(\omega; m), S(\omega; m), N(\omega; m)$ are the short time linear spectra in analysis window $m$ respectively. The HMM composition executes addition in the HMM model domain instead of the addition in the parameter domain. Generally, parameters for speech recognition are represented by the cepstrum. The parameters have to be transformed to the linear domain as an addition of the speech and noise holds[4, 6].

As for a convolutional distortion, the observed distorted spectrum is represented by

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

where $H(\omega; m)$ is a transfer function from the sound source to the microphone. $H(\omega; m)$ is a function of time $t$ since the sound source may move. The multiplication can be converted to sum in the cepstral domain as

$$O_{cep}(t; m) = S_{cep}(t; m) + H_{cep}(t; m)$$

where $O_{cep}(t; m), H_{cep}(t; m) and S_{cep}(t; m)$ are cepstra for the observed signal, acoustic transfer function and speech signal, respectively. Therefore the observed signal is represented by

$$O(\omega; m) = \exp(F(S_{cep}(t; m) + H_{cep}(t; m)))+kN(\omega; m) \quad (1)$$

This procedure is summarized in Fig.1. The cosine transform, inverse cosine transform, exponential transform and log transform are conducted on HMM parameters.

The procedure is as follows:

1. Estimate HMMs of the speech, noise and acoustical transfer function in the cepstral domain.

2. Compose HMMs of the speech and acoustical transfer function in the cepstral domain.

$$\mu(cep, SH) = \mu(cep, S) + \mu(cep, H)$$

$$\Sigma(cep, SH) = \Sigma(cep, S) + \Sigma(cep, H)$$

Here, $\mu(cep, S), \Sigma(cep, S), \mu(cep, H), \Sigma(cep, N), \mu(cep, H)$ and $\Sigma(cep, SH)$ are a mean vector and covariance matrix of HMMs of the speech, acoustical transfer function, and composed HMMs in the cepstral domain, respectively.

3. Cosine transform of each Gaussian pdf of HMMs.

$$\mu(\log, SH) = \Gamma \cdot \mu(cep, SH)$$

$$\Sigma(\log, SH) = \Sigma(cep, SH) \cdot \Gamma^T$$

Here, $\Gamma$ is a cosine transform matrix, $\mu(\log, SH)$ and $\Sigma(\log, SH)$ are a mean vector and covariance matrix of Gaussian pdf in the log power spectral domain, respectively.

4. Exponential transform to the linear spectral domain. The normal random vectors obtained by exponential transform, $Z = \exp^2$, has log normal distribution. A mean and covariance are given by

$$\mu(\log, SH) = \exp(\mu(\log, SH), i) + \frac{\sigma^2_{log, SH}(i)}{2}$$

$$\sigma^2_{log, SH}(i,j) = \exp(\mu(\log, SH), i) \cdot \exp(\mu(\log, SH), j)$$

Here, $\mu(\log, SH)$ and $\Sigma(\log, SH)$ are a mean vector and covariance matrix in the linear power spectral domain.

5. Compose two distributions according to the equation (1).

$$\mu(\log, SHN) = \mu(\log, SH) + k \cdot \mu(\log, N)$$

$$\Sigma(\log, SHN) = \Sigma(\log, SH) + k^2 \cdot \Sigma(\log, N)$$

Here, $\mu(\log, N), \Sigma(\log, N), \mu(\log, SHN)$ and $\Sigma(\log, SHN)$ are a mean vector and covariance matrix of the noise and composed model in the linear power spectral domain, respectively.

6. Log transform of composed HMM.

$$\mu(\log, SHN, i) = \log(\mu(\log, SHN), i)$$

$$\sigma^2_{log, SHN}(i,j) = \frac{1}{2} \left\{ \frac{\sigma^2_{log, SHN}(i,j)}{\mu(\log, SHN, i) \cdot \mu(\log, SHN, j)} + 1 \right\}$$

Figure 1. Block diagram of HMM Composition
7. Inverse cosine transform to the cepstral domain.

\[ \mu_{\text{cep,SHN}} = \Gamma^{-1} \cdot \mu_{\text{log,SHN}} \]

\[ \Sigma_{\text{cep,SHN}} = \Gamma^{-1} \cdot \Sigma_{\text{log,SHN}} \cdot (\Gamma^{-1})^T \]

The HMM recognizer decodes the observed signal on a trellis diagram according to maximize the log likelihood. Decoded path will bring an optimal combination of a speech, transfer function, and noise.

3. ESTIMATION OF ACOUSTICAL TRANSFER FUNCTION BASED ON HMM DECOMPOSITION

Estimation of 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 section presents a new method to estimate HMM parameters of the acoustical transfer function based on the HMM decomposition in the model domain. The estimation is implemented by maximizing a likelihood of adaptation data from any user’s position. In [7], 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 the model domain. Therefore the estimation in the model domain can reduce computation amount.

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 the model domain.

The observed signal in the noisy reverberant environment is represented by

\[ O_{\text{cep}}(t; m) = \mathcal{F}^{-1} \left[ \log \left\{ \exp (\mathcal{F}(S_{\text{cep}}(t; m)) + H_{\text{cep}}(t; m)) \right\} + N(\omega; m) \right] \].

Here \( \mathcal{F} \), \( \mathcal{F}^{-1} \) are Fourier (cosine) transform and inverse Fourier (cosine) transform respectively. \( O_{\text{cep}}(t; m), S_{\text{cep}}(t; m) \) and \( H_{\text{cep}}(t; m) \) are cepstra for an observed signal, a clean speech signal and an acoustical transfer function respectively of the observed signal \( \omega \) in the cepstral domain. Accordingly, the acoustical transfer function is represented by

\[ H_{\text{cep}}(t; m) = \mathcal{F}^{-1} \left[ \log \left\{ \exp (\mathcal{F}(O_{\text{cep}}(t; m))) + N(\omega; m) \right\} - S_{\text{cep}}(t; m) \right]. \]

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

\[ M_{\text{Hcep}} = \mathcal{F}^{-1} \left[ \log \left\{ \exp (\mathcal{F}(O_{\text{cep}}(t; \gamma))) + M_{\text{Ncep}} \right\} \right] \Theta M_{\text{Scep}} \]

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

\[ M_{\text{Hcep}} = \arg \max_{M_{\text{Hcep}}} P(O|M_{\text{Hcep}}, M_{\text{Scep}}, M_{\text{Ncep}}). \]

The decomposition of HMMs is represented by \( \Theta \) operator. The proposed HMM decomposition method is applied twice to noisy reverberant speech. Firstly, the HMM decomposition method is applied in the linear spectral domain to estimate the distorted speech HMMs by ML estimation. The obtained distorted speech HMMs are converted to the cepstral domain. Then the HMM decomposition method is applied again in the cepstral domain to estimate the acoustical transfer function HMM by ML estimation.

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

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

2. Estimate parameters of a noise HMM \( M_{\text{Ncep}} \) from the signal during noise periods.

3. Convert \( M_{\text{Ocep}}^{(l)} \) and \( M_{\text{Ncep}} \) to the linear spectral domain:

\[ M_{\text{Ocep}}^{(l)} = \exp (\mathcal{F}(M_{\text{Ocep}}^{(l)})), \quad M_{\text{Ncep}} = \exp (\mathcal{F}(M_{\text{Ncep}})). \]

4. Decompose \( M_{\text{SHlin}}^{(l)} \) from \( M_{\text{Ocep}}^{(l)} \):

\[ M_{\text{SHlin}}^{(l)} = \arg \max_{M_{\text{SHlin}}} P(O|M_{\text{SHlin}}, M_{\text{Ncep}}) \]

\[ M_{\text{SHlin}}^{(l)} = \arg \max_{M_{\text{SHlin}}} P(O|M_{\text{SHlin}}, M_{\text{Ncep}}) \]

\[ N(m_k^{(l)}, v_k^{(l)}) = N(m_k - \mu_{\text{noise}}, \sigma_k^{2}) \]

where \( m_k^{(l)} = \sum_t \gamma_k^{(l)} \alpha_t / \gamma_k^{(l)}, \quad v_k^{(l)} = \sum_t \gamma_k^{(l)} \alpha_t - m_k^{(l)} F / \gamma_k^{(l)} \), \( m_k \), \( F \), \( \gamma_k^{(l)} \) and \( v_k^{(l)} \) are the mean and the variance of adaptation data at \( l \)-th iteration by EM algorithm.

5. Convert \( M_{\text{SHlin}}^{(l)} \) to the cepstral domain:

\[ M_{\text{SHcep}}^{(l)} = \mathcal{F}^{-1} (\log(M_{\text{SHlin}}^{(l)})). \]

6. Estimate parameters of the acoustical transfer function, a mean and a variance, \( (h_\mu^{(l)}, h_\sigma^{(l)}) \). Here, to simplify a description of equations, the clean speech HMMs \( M_{\text{Scep}} \) is represented by tied-mixture HMMs, \( (\mu_k, \sigma_k^2) \).

\[ \Delta h_\mu^{(l)} = h_\mu^{(l)} - h_\mu^{(l-1)}, \quad \Delta h_\sigma^{(l)} = h_\sigma^{(l)} - h_\sigma^{(l-1)}, \]

\[ \eta^{(l)} = \arg \max_{(\Delta h_\mu^{(l)}, \Delta h_\sigma^{(l)})} P(O|\eta^{(l-1)}, M_{\text{Scep}}). \]

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

\[ Q(\eta^{(l)}|\eta^{(l-1)}). \]
\[ E \text{stimnate parameter of a noise HMM using a signal during noise periods) } \]

Figure 2. Parameter estimation by HMM decomposition

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

where \( \gamma_k^{(l)} \) is the distorted 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, \( \Delta h_\sigma^{(l)} \) is derived from \( \partial Q/\partial \Delta h_\sigma^{(l)} = 0 \), then given by

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

\( \Delta h_\sigma^{(l)} \) is derived from \( \partial Q/\partial \Delta h_\sigma^{(l)} = 0 \), then given by

\[
\sum_{k=1}^{K} \gamma_k^{(l)} \left\{ \frac{1}{\sigma_k^{(l-1)}} - \frac{\phi_k^{(l)}}{(\sigma_k^{(l-1)} + h_\sigma^{(l-1)} + \Delta h_\sigma^{(l)})^2} \right\} = 0,
\]

\( \phi_k^{(l)} = m_k^{(l)} + m_k^{(l-1)} + (\mu_k^{(l)} + \mu_k^{(l-1)}) - 2m_k^{(l)}, \)

Here we define the function

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

\( \Delta h_\sigma^{(l)} \) 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 \( \Delta h_\sigma^{(l)} \) is given by

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

\( \Delta h_\sigma^{(l)} \) 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 \( \Delta h_\sigma^{(l)} \) is given by

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

7. Compose three distributions according to the equation (2), [10]

8. Repeat the above procedure until the log likelihood probability converges.

The procedure is summarized in Figure 2. The mean and variance of adaptation data are represented by \( N(\mu_k^{(l)}, \sigma_k^{(l)}) \) 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.

4. EXPERIMENTS

The speech recognition experiments are carried out to investigate the effectiveness of the proposed method. The evaluation of the length of the impulse response and of unknown testing sound source positions is also conducted.

We conducted recognition experiments of the degraded speech uttered in a noisy room. Fig.3 shows the room used in the experiment. We measured 9 transfer functions from 9 positions to the microphone. The former five positions, \( h_1, \ldots, h_5 \) are used for the model composition and the latter four positions, \( p_1, \ldots, p_4 \) are used for the recognition tests. The transfer functions are measured by the sweep method. The length of the original impulse response was 180msec(\( L=180\text{msec} \)). The test data are simulated by the linear convolution of speech corpus and the measured transfer function.

Figure 3. Room environment
Speech signal is sampled in 12kHz and windowed by 32msec Hamming window every 8msec. Then FFT is used to calculate 16-order MFCCs and a power. In the recognition, a power term is not used because it is only necessary to adjust the SNR in the HMM composition.

Speech corpus for evaluation is ATR Japanese speech database Set-A and ASJ continuous speech database. This database contains word and sentence utterances by announcers. The recognition algorithm is based on 256 tied-mixture diagonal covariance HMMs. The HMM has 5 states to calculate 16-order MFCCs and a power. In the recognition, the SNR is adjusted in the HMM composition.

For the noisy distorted speech, the improvement is obtained from 4.8% to 67.2%. SD recognition rate for known training positions using HMM trained by the noisy distorted speech is 89.7%.

Next, we evaluate the performance of the model for the unknown testing sound source positions. Table 2 shows average SD recognition rates for the known training and unknown testing positions. It is confirmed that the degradation between the training and testing sound source positions is relatively small for all composed HMMs.

The spectral analysis for speech recognition is based on short time windowing. The multiplication of short time signal spectra and the transfer function is equivalent to the convolution in time domain. However, an actual distorted signal should be made by the linear convolution.

Since the proposed HMM composition of the signal and room acoustical transfer function only realizes the periodic convolution, the composed HMM can't model an actual distorted signal accurately. The difference of using periodic and linear convolution will be large according to the length of impulse response. We have also examined the effect of a length of the impulse response. The several distorted signals are artificially made by L=180, 100 and 32msec impulse responses by cutting out the original one. Table 3 shows the results by HMM-SH. Although the simulated impulse response is not exist in real room, it is observed that the effects decrease according to the length of the impulse response.

### Table 1. Model specification and speaker dependent/independent recognition rates [%]

<table>
<thead>
<tr>
<th>Input</th>
<th>HMM-S</th>
<th>HMM-SN</th>
<th>HMM-SH(μ)</th>
<th>HMM-SH(μ, β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise Composition</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Acoustics Composition</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Distorted Speech</td>
<td>78.5/-</td>
<td>87.2/-</td>
<td>84.0/-</td>
<td>67.2/57.2</td>
</tr>
<tr>
<td>Noisy Distorted Speech</td>
<td>4.5/18.7</td>
<td>59.5/53.5</td>
<td>-</td>
<td>55.2/45.4</td>
</tr>
</tbody>
</table>

### Table 2. Recognition rates at known/unknown positions [%]

<table>
<thead>
<tr>
<th>Input</th>
<th>HMM-S</th>
<th>HMM-SH(μ)</th>
<th>HMM-SH(μ, β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distorted Speech</td>
<td>78.5/77.8</td>
<td>87.2/86.2</td>
<td>84.0/83.7</td>
</tr>
</tbody>
</table>

### Table 3. SD recognition rates vs. the length of impulse response [%]

<table>
<thead>
<tr>
<th>Input</th>
<th>180msec</th>
<th>100msec</th>
<th>32msec</th>
<th>clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM-SH</td>
<td>87.2</td>
<td>94.0</td>
<td>96.2</td>
<td>96.6</td>
</tr>
</tbody>
</table>
test and speaker independent test, respectively. Figures show improvements of the recognition rates by HMM decomposition according to the increase of the number of adaptation words. The figures also show the performances by HMM composition of clean speech HMM and the measured acoustical transfer function HMM. The accuracy by the matched model shows the upper bound when the HMMs are trained by the noisy reverberant speech data in the same real environments. The results are summarized as follows;

- The HMM decomposition estimates an acoustical transfer function HMM accurately by using only a few adaptation words. The accuracy can be improved up to the rate by HMM composition with the measured impulse response.
- The HMM decomposition of not only mean vectors but also covariance matrices improves the accuracy.
- There still exists the significant difference of the performance between by the matched model and by HMM composition or HMM decomposition.

The cepstral mean normalization, CMN, is also applied to the speaker dependent test. The speech HMMs are trained by using CMN-processed clean speech data, where the cepstral mean is estimated for each word. The recognition accuracy is 53.8%. This result suggests the CMN has the problems for normalization of noisy reverberant speech.

5. CONCLUSION

This paper presents a robust speech recognition method based on the HMM composition and decomposition for the noisy reverberant speech. The proposed HMM composition is obtained by natural extension of the HMM composition method of an additive noise to that of the convolutional distortion of the reverberation. The experiments indicate that the proposed method improves the speaker dependent 500 Japanese word recognition rate from 4.8% to 67.2% for the noisy distorted speech. Furthermore the paper proposes a novel estimation algorithm of acoustical transfer function based on HMM decomposition. The experiments indicate that the proposed HMM decomposition algorithm is able to achieve almost the same performance as that by HMM composition with the measured impulse response.

REFERENCES