SPEAKER ADAPTATION IN NOISY ENVIRONMENTS BASED ON PARAMETER ESTIMATION USING UNCERTAIN DATA

Naoto Iwahashi\textsuperscript{1} and Akihiko Kawasaki\textsuperscript{1,2}

\textsuperscript{1}Sony Computer Science Labs Inc.
Takanawa Muse Bldg. 3-14-13 Higashi-Gotanda Shinagawa-ku, Tokyo, 141-0022 Japan,
iwahashi@csl.sony.co.jp
\textsuperscript{2}Dept. Electrical, Electronics and Computer Engineering, Waseda University
Shinjuku-ku, Tokyo 169-8555 Tokyo

ABSTRACT
This paper describes new method for the speaker adaptation of HMM parameters in environments with background noise. This method is based on Bayesian estimation, and calculates the \textit{a posteriori} distribution of clean-speech HMM parameters from their \textit{a priori} distribution by using noisy speech observations. The advantage of the method is that the distribution of the noise can be taken into account in adapting clean-speech HMMs to a target speaker’s speech without noise. The results of the experiments using noninformative prior show that the recognition performance in a noise-free environment was improved by this method even when the SNR of the noisy speech data used for the adaptation was $-6$ dB.

1. INTRODUCTION
Speech recognition is degraded by various environmental factors, and should ideally be robust even when these factors occur simultaneously. Although speaker adaptation has been accomplished in noise-free environments, a speech recognition system will be more flexible if the speaker adaptation can be carried out robustly even in noisy environments.

There have been several previous studies on the simultaneous treatment of the multiple factors degrading recognition, and Gales and Young \cite{1} have proposed a method for compensating background noise and channel transfer characteristics simultaneously using HMM combination. The simultaneous treatment of background noise and speaker characteristics has not, however, been studied systematically yet. Since the background-noise characteristics in most applications can change utterance-by-utterance, the method used for speaker adaptation in an environment with background noise should improve the recognition of the target speaker’s speech even when the background-noise characteristics of the speech sample used in the recognition differ from those of the speech samples used for the adaptation. Therefore, the speaker adaptation in the background-noise environment should adapt the parameters of clean-speech HMMs using noisy speech data in such a way that the HMMs match the characteristics of the target speaker’s speech without noise. This requires the use of high performance noise-compensation methods in both the adaptation and recognition processes. Consideration of the noise distribution is particularly important in noise compensation \cite{2,3}, so the adaptation should be done while taking account of the uncertainty due to the noise distribution. We developed a new Bayesian method for speaker adaptation. It calculates the \textit{a posteriori} distribution of the parameters of clean-speech HMMs from their \textit{a priori} distribution by using the observed noisy-speech data while taking account of the uncertainty due to the noise component of noisy speech.

2. BAYESIAN ESTIMATION OF CLEAN-SPEECH HMM PARAMETERS USING NOISY SPEECH

Noisy speech includes both speech and noise components. Suppose that the information of noise distribution is given. Let $\mathcal{Y}$ be a sequence of feature vectors of noisy speech, and let $\mathcal{X}$ be a sequence of feature vectors of the component of speech in that noisy speech. Since $\mathcal{X}$ cannot be observed, we assume that it is a sequence of random variables. We assume that $p(\mathcal{X})$ have a parametric form, and its parameter vector $\lambda$ is a random variable with the \textit{a priori} distribution $p(\lambda)$. Then the \textit{a posteriori} distribution $p(\lambda|\mathcal{Y})$ can be written

\[
p(\lambda|\mathcal{Y}) = \int p(\lambda|\mathcal{X})p(\mathcal{X}|\mathcal{Y})d\mathcal{X}
= \int \frac{p(\lambda|\mathcal{X})p(\mathcal{X}|\mathcal{Y})}{\int p(\mathcal{X}|\lambda)p(\lambda)d\mathcal{X}}p(\mathcal{X}|\mathcal{Y})d\mathcal{X}
= p(\lambda) \cdot \int p(\mathcal{X}|\lambda) \frac{p(\mathcal{X}|\mathcal{Y})}{\int p(\mathcal{X}|\lambda)p(\lambda)d\mathcal{X}}d\mathcal{X}. \tag{1}
\]

$p(\mathcal{X}|\mathcal{Y})$ is determined by the noise distribution. Thus, this equation allows us to taking account of the informatin of the noise distribution in the estimation of \textit{a posteriori} distribution of clean-speech HMM. We call this way of an estimation noise-predictive Bayesian estimation, because the

ISCA Archive
http://www.isca-speech.org/archive
a posteriori distribution is predicted using the information of the noise distribution.

Consider a HMM with a Gaussian state observation distribution. To simplify the discussion, we assume the state-transition probabilities of the HMM are fixed and known. If we use a segmental MAP algorithm [4] in the estimation process, the estimation problem reduces to estimating the parameters of the Gaussian distributions of each state by using the set of the observation vectors allocated to that state by the state segmentation of the segmental MAP algorithm. Let \( Y = \{y_1, y_2, ..., y_n\} \) be a set of \( D_y \)-dimensional feature vectors of noisy speech, and let \( X = \{x_1, x_2, ..., x_n\} \) be a set of \( D_x \)-dimensional feature vectors of the clean-speech component of that noisy speech. Suppose that the state observation distribution is represented by

\[
p(Y|X) = \int p(X|x)p(x)p(y|x) \, dx
\]

\[
= \int p(X|x)p(y|x) \, dx
\]

\[
\approx \alpha_0 \prod_{i=1}^n \left\{ (2\pi)^{D_y/2} |V_i|^{-\frac{1}{2}} \times \exp \left( -\frac{1}{2} (y_i - \mu_i)^T V_i^{-1} (y_i - \mu_i) \right) \right\},
\]

where \( y_i \) is the \( i \)th sample in \( Y \), \( \mu_i \) and \( V_i \) are respectively the mean vector and the covariance matrix corresponding to \( y_i \), and \( \alpha_0 \) is a scaling factor. Then the a posteriori distribution is approximated as

\[
p(\mu|Y) \approx (2\pi)^{-\frac{D_x}{2}} |\Sigma_0|^{-\frac{1}{2}} \times \exp \left( -\frac{1}{2} (\mu - \tilde{\mu}_0)^T \Sigma_0^{-1} (\mu - \tilde{\mu}_0) \right),
\]

where

\[
\tilde{\Sigma}_0 = \left( \sum_{i=1}^n (V_i + \tilde{\Sigma}_0)^{-1} \right)^{-1},
\]

\[
\tilde{\mu}_0 = \tilde{\Sigma}_0 \sum_{i=1}^n (V_i + \tilde{\Sigma}_0)^{-1} (\mu_i + \Sigma_0^{-1} \mu_0)
\]

\( \tilde{\mu} = \tilde{\mu}_0 \) is the noise-predictive MAP (NP-MAP) estimate of the mean vector \( \mu \).

3. IMPLEMENTATION

The noise-predictive Bayesian learning of clean-speech HMM parameters is implemented with MFCC as the features of clean speech. A filter-bank power spectrum is used as the features of noisy speech. The calculation here is simplified by using the following approximation:

\[
p(X|Y) \approx \alpha_1 \prod_{i=1}^n p(x_i|y_i)
\]

\[
\approx \alpha_1 \prod_{i=1}^n \left\{ (2\pi)^{-D_x/2} |V_i|^{-\frac{1}{2}} \times \exp \left( -\frac{1}{2} (x_i - m_i)^T V_i^{-1} (x_i - m_i) \right) \right\},
\]

where \( m_i \) and \( V_i \) are respectively the mean vector and the covariance matrix and \( \alpha_1 \) is a scaling factor.

\( m_i \) and \( V_i \) are calculated using a set of the feature vectors obtained by spectral subtraction using the noisy speech observation \( y_i \), and each sample in a set of the noise samples collected during a non-speech period. The algorithm used in calculating \( m_i \), and \( V_i \) for static MFCC is summarized as follows:

1. Before each utterance, filter-bank power spectral vectors of the noise during \( M \) frames are saved into a noise buffer.

2. During the utterance, subtract each noise spectral vector in the noise buffer from the filter-bank power spectral vector \( y_i \), for the \( i \)th frame in the noisy speech input, thus obtaining \( M \) power spectral vectors.

3. Transform the \( M \) power spectral vectors into \( M \) cepstral vectors.

4. Calculate the mean vector and the covariance matrix among the \( M \) cepstrums, which results in \( m_i \) and \( V_i \) for \( i \)th frame.

In order to match the recognition process with the adaptation process in terms of noise-compensation schemes, the state observation probability of the feature vector \( y \) in noisy speech is calculated in the recognition process as

\[
p(y) \approx (2\pi)^{-\frac{D_y}{2}} |\Sigma + V|^{-\frac{1}{2}} \times \exp \left( -\frac{1}{2} (m - \tilde{\mu})^T (\Sigma + V)^{-1} (m - \tilde{\mu}) \right),
\]

where \( \tilde{\Sigma} \) and \( \Sigma \) are respectively noise-predictive MAP estimate of the mean vector and the fixed covariance matrix of the state observation distribution. \( m \) and \( V \) are calculated as in the adaptation process, and this calculation is called stochastic features extraction (SFE) [3]. The stochastic features incorporate the influence of the noise distribution effectively in the framework of spectral subtraction.

4. EXPERIMENTS

4.1. Conditions

The NP-MAP method was evaluated in a connected-digit recognition task in a car-noise environment. The context-independent digit HMMs as initial speaker-independent HMMs were trained using a noise-free connected-digit speech database consisting of the speech spoken by sixty-four speakers (thirty-two males and thirty-two females).
All state observation distributions were diagonal covariance Gaussian distributions. The features used were thirty-eight dimensional vector (thirteen order MFCC, their delta, and acceleration, without static $C_0$). Test data consisted of connected seven-digit utterances spoken by thirty-two speakers (sixteen males and sixteen females). The clean speech data and car-noise data were recorded with the same microphone. The noise samples in different periods in the noise data were added artificially to each clean speech sample by several SNRs.

The initial speaker-independent HMMs were adapted using part of test data, and the recognition performance was evaluated using the rest of the test data (200 utterances per speaker). The recognition used Viterbi decoder with full search. The noise samples used to calculate $m_i$ and $V_i$ were collected during the 2,000 msec (200 analysis frames) immediately before each utterance.

For comparison, the conventional MAP estimation method using the features obtained by nonlinear spectral subtraction [5] was also evaluated (SS-MAP). In the case of SS-MAP, the recognition of test utterances was carried out also using the features obtained by nonlinear spectral subtraction (NSS).

In all experiments, a noninformative prior is used as the a priori distribution of each mean vector of all state observation distributions. Therefore, the estimates were substantially equivalent to those based on maximum likelihood criterion. The adaptation was carried out in a supervised way.

### 4.2. Results

The sentence recognition rates obtained using the initial HMMs with the methods for compensating noise in features are listed in Table 1. We can see clearly that the recognition rates were highest with SFE in all noise environments.

The sentence recognition rates in noise-free tests using the HMMs adapted using noisy speech data are listed in Table 2. We can see that the recognition rates obtained using HMMs adapted by NP-MAP were higher than those obtained using the initial speaker-independent HMMs (96.14%) even when the SNR of the adaptation data was −6 dB. The recognition rates obtained using HMMs adapted by SS-MAP were not higher than those obtained using the initial speaker-independent HMMs when the SNR of the adaptation data was lower than 0 dB. In most of the cases, NP-MAP resulted in better recognition rates than SS-MAP. When the number of the adaptation utterances was 25, however, the recognition rates obtained with NP-MAP were lower than those obtained with SS-MAP, even when the adaptation speech data had a high SNR.

The sentence recognition rates obtained when the SNR of the recognition test data was the same as that of the adaptation data are listed in Table 3. In this case, the recognition rates obtained with NP-MAP were lower than those obtained with SS-MAP. The recognition rates obtained, with 200 adaptation utterances, when SS-MAP and NP-MAP were used under other combinations of adaptation data SNR and recognition data SNR are listed in Tables 4 and 5. When clean speech was recognized, the recognition rates obtained with NP-MAP were higher than those obtained with SS-MAP. In the other cases, however, the recognition rates obtained with NP-MAP were lower than obtained with SS-MAP.
Table 4: Recognition performance of SS-MAP with 200 adaptation utterances

<table>
<thead>
<tr>
<th>adaptation SNR (dB)</th>
<th>recognition data SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>∞</td>
<td>99.14 89.83 82.1 67.84 47.91</td>
</tr>
<tr>
<td>2.5</td>
<td>98.02 98.89 97.89 94.38 85.8</td>
</tr>
<tr>
<td>0</td>
<td>97.22 98.69 98.25 95.75 88.38</td>
</tr>
<tr>
<td>-3</td>
<td>96.08 98.64 98.28 96.47 90.27</td>
</tr>
<tr>
<td>-6</td>
<td>93.6 98.28 97.61 96.22 91.22</td>
</tr>
</tbody>
</table>

Table 5: Recognition performance of NP-MAP with 200 adaptation utterances

<table>
<thead>
<tr>
<th>adaptation SNR (dB)</th>
<th>recognition data SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>∞</td>
<td>99.03 88.29 82.71 70.63 53.18</td>
</tr>
<tr>
<td>2.5</td>
<td>98.11 97.25 95.61 91.77 82.76</td>
</tr>
<tr>
<td>0</td>
<td>97.75 97.25 96.28 92.83 84.68</td>
</tr>
<tr>
<td>-3</td>
<td>97.66 97.61 96.39 93.91 87.01</td>
</tr>
<tr>
<td>-6</td>
<td>97.13 97.41 96.19 94.36 88.8</td>
</tr>
</tbody>
</table>

5. DISCUSSION

By the results listed in Table 2, it was confirmed that NP-MAP could adapt the parameters of clean-speech HMMs using a target speaker’s noisy speech data to match the characteristics of the target speaker’s speech without noise. By comparing with the results obtained with SS-MAP, in NP-MAP the method for taking account of noise distribution proved to be effective.

NP-MAP needed more training data than SS-MAP, probably because NP-MAP learns the parameters while taking into account the uncertainty of the observations. Although it must be taken into account if the learning is to be robust, this uncertainty might make the learning slow. When the number of adaptation utterances is small, the use of an appropriate a priori distribution might improve the performance.

When noisy speech was recognized by using the HMMs adapted using the noisy speech, the recognition rates obtained with NP-MAP were lower than those obtained with SS-MAP. The recognition rates became higher, however, as the number of adaptation utterances increased. The recognition rates obtained with either NP-MAP or SS-MAP might be improved by using more adaptation utterances.

The approximation of the function specified in Eq. 2 allows us to evaluate the function at a low computational cost. A closer approximation of the a posteriori distribution, however, would improve the performance.

6. CONCLUSION

This paper described a method using a target speaker’s noisy speech data to adapt the parameters of clean-speech HMM so that the target speaker’s speech is recognized better in noise-free environments. The effectiveness of our new Bayesian method was confirmed experimentally, and our future work will include the study of ways to further improve the recognition performance by optimizing the a priori distribution of the HMM parameters.

7. ACKNOWLEDGEMENTS

We thank Dr. Watanabe in ATR Human Information Processing Research Labs for helpful comments and discussion, and Prof. Kobayashi in Waseda University for supporting this research.

8. REFERENCES