SPEAKER FEATURE EXTRACTION FROM PITCH INFORMATION 
BASED ON SPECTRAL SUBTRACTION FOR SPEAKER IDENTIFICATION

Yoshiroh Itoh, Jun Toyama and Masaru Shimbo
E-mail: yito@main.eng.hokudai.ac.jp
Graduate School of Engineering, Hokkaido University, Japan

ABSTRACT

Robust speaker feature extraction under noise conditions is an important issue for application of a speaker recognition system.

It is well known that LPC cepstrum, which expresses the spectral envelope, is effective for speaker recognition. This implies that the spectral rough structure is effective for speaker recognition. However, LPC cepstrum is a noise-sensitive feature.

On the other hand, spectral subtraction is an effective speech enhancement method under stationary noise conditions.

In this study, we developed a new method for feature extraction based on spectral subtraction and noise robust spectral rough structures, and we evaluated the effectiveness of the feature extraction method in speaker identification experiments.

1. INTRODUCTION

Construction of a speaker recognition system involves the following three problems:

1) Individualities in speech signal are not well known.

2) A large amount of speech data is required for learning the parameters of a recognition system, and there is therefore a great workload for the users of the recognition system.

3) Additive background noise, channel distortion by a communication lines and corruption of speech signals due to the characteristics of a microphone distort the speech source signal and this distortion decreases recognition performance.

We focused on the third problem. Some approaches for reducing the effect of background noise have been proposed. These approaches can be divided into two groups:

1) Those that attempt to modify the pattern matching method in order to account for the interfering noise.

2) Those that attempt to preprocess a corrupted waveform in such a way that the resulting parameters are close to those of clean speech. Techniques in the latter category include speech enhancement methods such as spectral subtraction [1], continuous spectral subtraction [2] and minimum mean square error [3][4].

We focus on the spectral subtraction method because the calculation cost of that method is lower than that of continuous spectral subtraction or minimum mean square error. A speaker recognition method based on spectral subtraction was proposed by Drygajlo et al.[5]. They first enhanced the observed speech signal by spectral subtraction, then divided the linear frequency axis into Bark bands, and finally extracted a feature from each frequency band. When one or more sampling points included in a Bark band were lost due to noise, they ignored any information of the band. Their feature extraction method is effective in the sense of canceling noise effect. However, the useful speaker information may be also canceled.

Spectral subtraction is an effective speech enhancement method when the speech source signal is voiced and when vowels are thought to include important information of individuality such as formant, harmonics, etc.

We propose a robust speaker feature extraction method in noisy environments based on spectral subtraction, and we present results of a speaker identification experiment showing the effectiveness of this method.

2. SPECTRAL SUBTRACTION

Spectral subtraction is an effective speech enhancement method for noisy environments. Since it is difficult to exclude the additive noise accurately, we regard the additive noise as the none speech part in the observed signal.

If a noise signal $n(k)$ is stationary, and the noise signal is added to a windowed speech signal $s(k)$, then
the observed signal forms

\[ y(k) = s(k) + n(k), \]

and taking Fourier transform gives

\[ Y(\omega) = S(\omega) + N(\omega). \]

It is difficult to estimate the \( N(\omega) \) in real time because the phases of speech and noise signal are not clear. Therefore, an averaged noise power spectrum \( |\hat{N}(\omega)|^2 \) is estimated by

\[ |\hat{N}(\omega)|^2 = \frac{1}{M} \sum_{i=0}^{M-1} |N_i(\omega)|^2, \]

where \( M \) is the number of frames. An averaged noise power spectrum \( |\hat{N}(\omega)|^2 \) is estimated from the non-speech part of an observed signal. There is a need to reduce the spectral error between the speech source signal and the estimated speech signal. The spectral error is caused by a difference between the phases of the speech source signal and additive-noise signal. The expectation value of the spectral error is zero. Therefore, to reduce the spectral error, the magnitude averaging is adapted. The observed speech power spectrum \(|Y(\omega)|^2 \) is replaced with \(|\hat{Y}(\omega)|^2 \):

\[ |\hat{Y}(\omega)|^2 = \frac{1}{M} \sum_{i=0}^{M-1} |Y_i(\omega)|^2. \]

To estimate the speech power spectrum \(|\hat{S}(\omega)|^2 \), \(|\hat{Y}(\omega)|^2 \) and \(|\hat{N}(\omega)|^2 \) are used:

\[ |\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{N}(\omega)|^2 & \text{(if } |Y(\omega)|^2 - |\hat{N}(\omega)|^2 > \epsilon) \\ \epsilon & \text{ (otherwise)} \end{cases}, \]

where \( \epsilon \) is a positive number called 'flooring value'.

Spectral subtraction is one of the speech enhancement methods which is effective under noisy conditions, but it also has a problem regarding residual noise, called 'musical noise'. When spectral subtraction is applied to unvoiced speech, the musical noise decreases the recognition performance significantly.

On the other hand, when spectral subtraction is applied to voiced speech, harmonics is stably extracted because the energy concentrates the harmonics. Vowels are typical voiced speech and include individuality such as formants. Therefore, we use vowels for speaker recognition.

3. FEATURE EXTRACTION

In this section, we show the speaker feature extraction method using the estimated speech power spectrum, which was introduced in section 2.

First, the estimated speech power spectrum is divided into several frequency bands. The effectiveness of dividing frequency bands is well known in speech and speaker recognition approaches, e.g. the MFCC method.

On the other hand, the effectiveness of an approach based on the LPC analysis method such as LPC cepstrum coefficients for speaker recognition have been reported. LPC cepstrum forms a spectral envelope, so that the rough spectral structure like the spectral envelope includes speaker features.

However, LPC analysis parameters are sensitive to additive noise and other disturbance, so that the spectral envelope formed by LPC cepstrum coefficients is not robust under noisy conditions.

Therefore, we form a spectral rough structure like a spectral envelope on the divided frequency bands. We propose the noise robust speaker features defined by

\[ \theta_j = \tan^{-1}\left( \min_j \left( \frac{\log(|\hat{S}(\omega_{j,\max})|^2) - \log(|\hat{S}(\omega_j)|^2)}{\omega_{j,\max} - \omega_j} \right) \right) \]

\[ b_j = |L_j(\omega_{j,\text{center}})|^2. \]

Here, \( \omega_j, \omega_{j,\max} \) and \( \omega_{j,\text{center}} \) are frequency in the \( j \)-th frequency band, the frequency that gives maximum magnitude in the \( j \)-th frequency band, and the center frequency in the \( j \)-th frequency band, respectively. \(|\hat{S}(\omega)|^2 \) is a speech power spectrum estimated by spectral subtraction, and \(|L_j(\omega)|^2 \) is the line segment in the \( j \)-th band that is drawn in each frequency band and forms a spectral rough structure (Fig. 1). Since the feature is extracted from speech harmonics in which speech energy is concentrated, it can be acquired stably under noisy conditions. Thus, we expect that when the the magnitude of additive noise is less than the speech magnitude at all of the speech harmonics, the features can be extracted stably under noisy condition and provide high speaker recognition performance.

In the speaker feature extraction method, determination of the suitable frequency bandwidth is important. When a bandwidth is too wide, the rough spectral structure will not be smooth and this causes a decrease in recognition performance. On the other hand, when a bandwidth is too narrow, the rough spectral structure cannot be formed because one frequency band must be include more than 2 speech harmonics to form a line segment. Therefore, we set the frequency bandwidth as 600 [Hz] in this study. This is because a bandwidth of 600 [Hz] is sufficient to have more than 2
harmonics for any speaker. Then, the speaker feature forms a 26-dimension (13 frequency bands) vector per frame.

It is assumed that the feature vectors have a multivariate Gaussian distribution. The likelihood $p_{im}(x_{im} | \lambda_i)$ of observed vector $x_{im}$ is calculated by

$$p_{im}(x_{im} | \lambda_i) = \frac{1}{\sqrt{2\pi|\Sigma_i|}} \exp \left\{ -\frac{1}{2}(x_{im} - \mu_i)^T \Sigma_i^{-1}(x_{im} - \mu_i) \right\}.$$  

Where suffixes $i$, $m$ and $j$ are the speaker identification number, the frame number of input speech for recognition, and the number of a divided frequency bands, respectively. $\Sigma_i$ is the covariance matrix of $\theta_{ij}$ and $b_{ij}$. To save calculation cost for recognition, only the covariances of the inside of the frequency band are considered, and those between other frequency bands are not considered. Thus, $\Sigma_i$ (D×D matrix) and $\mu_i$ (D-dimension) are formed as follows:

$$\Sigma_i = \begin{pmatrix} \Sigma_{ij}^{(i)} \\ \vdots \\ \Sigma_{ij}^{(i)} \end{pmatrix},$$

$$S_j^{(i)} = \begin{pmatrix} \sigma_{\theta_{ij}}^{2(i)} & \sigma_{\theta b_{ij}}^{(i)} \\ \sigma_{\theta b_{ij}}^{(i)} & \sigma_{b_{ij}}^{2(i)} \end{pmatrix},$$

and

$$\mu_i = (\tilde{\theta}_1^{(i)} \tilde{b}_1^{(i)} \cdots \tilde{\theta}_D^{(i)} \tilde{b}_D^{(i)}).$$

$\tilde{\theta}_j^{(i)}$, $\tilde{b}_j^{(i)}$, $\sigma_{\theta_j}^{2(i)}$, $\sigma_{b_j}^{2(i)}$ and $\sigma_{\theta b_j}^{(i)}$ are the mean of $\theta_j^{(i)}$, the mean of $b_j^{(i)}$, the variance of $\theta_j^{(i)}$, the variance of $b_j^{(i)}$ and the covariance of $\theta_j^{(i)}$ and $b_j^{(i)}$, respectively.

4. SPEAKER IDENTIFICATION EXPERIMENT

In order to evaluate the effectiveness of the feature extraction method described in section 3, we performed a speaker identification experiment in various noisy environments. The experimental conditions are shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Experimental conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>speech database</strong></td>
</tr>
<tr>
<td><strong>noise database</strong></td>
</tr>
<tr>
<td><strong>sampling rate</strong></td>
</tr>
<tr>
<td><strong>frame length</strong></td>
</tr>
<tr>
<td><strong>frame shift</strong></td>
</tr>
<tr>
<td><strong>speaker set</strong></td>
</tr>
</tbody>
</table>

We used speech sequence combinations of five vowels of isolated utterances (\(a/\), \(i/\), \(u/\), \(e/\), \(o/\)) for every speaker. Each vowel was uttered twice at different five times of three months apart (so that uttered ten times for each vowel).

An experiment was performed under five environment conditions and three SNR conditions (12, 6 and 0 [dB]) for each noise environment. The noise environments were an elevator hall in a hospital, inside a telephone booth, a crowded place, inside a booth of an exhibition hall, and an artificial white noise environment. We also performed a speaker identification experiment based on Drygajlo et al.’s feature extraction for comparison of the results.

The results are shown in Table 2. The proposed speaker feature extraction method is effective under
conditions in which the noise sources are conversations of other speakers, e.g. hospital, telephone booth, crowd and exhibition hall. Under those conditions, the speech harmonics are not lost due to the noise for all frequency channels. However, low SNR conditions make the recognition performance worse under some of the effective noise conditions ('crowd' and 'exhibition hall'). This was because the noise conditions contained in high-frequency elements and the low SNR condition amplified the high-frequency energy. Thus, some of the speech harmonics contained in high frequency channels were lost. Unfortunately, the white noise condition provides much worse recognition performance, because almost all of the speech harmonics were lost.

5. CONCLUSION

We attempted to extract speaker features for speaker recognition using the spectral rough structure under noise conditions. The proposed speaker feature extraction method was found to be effective under conditions in which the noise sources are conversations of other speakers. However, under noise conditions in which energy is concentrated in high-frequency bands (e.g. white noise), the proposed method did not work well. Further work is needed to improve the method of feature extraction to overcome the problem of high-frequency noise.

References


