SPEECH RECOGNITION UNDER MUSICAL ENVIRONMENTS USING KALMAN FILTER AND ITERATIVE MLLR ADAPTATION

M. Fujimoto and Y. Ariki

Department of Electronics and Informatics
Ryukoku University, Seto, Otsu-shi, Shiga, 520-2194, JAPAN
masa@artikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract

In this paper, we propose a speech recognition method under non-stationary musical environments using Kalman filtering speech signal estimation method and iterative unsupervised MLLR (Maximum Likelihood Linear Regression) adaptation. Our proposing method estimates the speech signal under non-stationary noisy environments such as musical background by applying speech state transition model to Kalman filtering estimation. The speech state transition model represents the state transition of speech component in non-stationary noisy speech and is modeled by using Taylor expansion. In this model, the state transition of noise component is estimated by using linear predictive estimation. Furthermore, to obtain higher recognition accuracy, we consider to adapt the acoustic models by using iterative unsupervised MLLR adaptation to speech spectra distorted by Kalman filtering residual noise.

In order to evaluate the proposed method, we carried out large vocabulary continuous speech recognition experiments under 3 types of music. As a result, the proposed method obtained the significant improvement in word accuracy, from 20.04% to 64.43% at 0dB SNR.

1. Introduction

In recent years, many types of speech recognition systems have been proposed and developed toward the practical use in the real world. However, most of the works recognize clean speech collected in quiet environments. For practical use it is required for recognition systems to be robust for interfering noises, especially non-stationary noises. Robust speech recognition systems are classified into two types. One adapts itself to any kinds of noises based on model adaptation techniques[1]-[3]. The other reduces the noise component from noisy speech based on noise reduction techniques[4]-[7].

Proposed method is proposed as a conventional noise reduction method. However, SS has a problem that it deteriorates the recognition rate due to spectral distortion by over or under subtraction. In addition, the SS does not consider the time varying property of noise spectra, because noise spectra is estimated as mean spectra within the time section assigned to be noise (beginning of utterance).

To solve the above mentioned problems, we propose here a speech recognition method under non-stationary musical environments using Kalman filtering speech signal estimation method and iterative unsupervised MLLR(Maximum Likelihood Linear Regression)[8] adaptation. The proposed method estimates the speech signal under non-stationary noisy environments by applying speech state transition model to fast Kalman filtering estimation. Furthermore, to obtain higher recognition accuracy, the acoustic models are adapted by using iterative unsupervised MLLR adaptation to speech spectra distorted by Kalman filtering residual noise.

In order to evaluate the proposed method, we carried out LVCSR (Large Vocabulary Continuous Speech Recognition) experiments under non-stationary musical environments. As a result, the proposed method obtained the significant improvement in word accuracy.

2. Speech State Transition Model

At the kth frame, power spectra of clean speech under noisy environments is represented as follows:

\[ S(k) = \exp \left( X(k) \right) \]

(1)

where \( X(k) \), \( S(k) \) and \( N(k) \) denote the vectors of power spectra of noisy speech, clean speech and noise at the kth frame respectively, and superscript \( d \) denotes the log-spectral domain.

In Eq (1), speech state transition from \( S(k) \) to \( S(k+1) \) is represented as follows:

\[ S(k + 1) = \exp \left( \Delta X'(k + 1) \right) - \exp \left( \Delta N'(k + 1) \right) \]

\[ S(k + 1) = \exp \left( X'(k) + \Delta X'(k) \right) - \exp \left( N'(k) + \Delta N'(k) \right) \]

(2)

where \( \Delta X'(k) = X'(k+1) - X'(k) \) and \( \Delta N'(k) = N'(k+1) - N'(k) \) respectively.

Here, by expanding Eq (2) using first order Taylor series, speech state transition can be linearized as follows:

\[ S(k + 1) \approx S(k) + \frac{\partial S(k)}{\partial X'(k)} \Delta X'(k) + \frac{\partial S(k)}{\partial N'(k)} \Delta N'(k) \]
\[ S(k) + X(k) \Delta X'(k) - N(k) \Delta N'(k) \]
\[ = S(k) + (s(k) + N(k)) \Delta X'(k) - N(k) \Delta N'(k) \]
\[ = (1 + \Delta X'(k)) S(k) + N(k) (\Delta X'(k) - \Delta N'(k)) \]
\[ = F_s S(k) + G_s W(k) \tag{3} \]
\[ \frac{\partial S(k)}{\partial X'(k)} \frac{\partial (X(k) - N(k))}{\partial X(k)} \frac{\partial X(k)}{\partial X'(k)} = X(k) \tag{4} \]
\[ \frac{\partial S(k)}{\partial N'(k)} = \frac{\partial (X(k) - N(k))}{\partial N(k)} \frac{\partial N(k)}{\partial N'(k)} = -N(k) \tag{5} \]
where \( F_s = 1 + \Delta X'(k), \ G_s = N(k) \) and \( W(k) = \Delta X'(k) - \Delta N'(k) \) respectively.

In Eq. (3)~(5), \( \frac{\partial S(k)}{\partial X'(k)} \) and \( \frac{\partial S(k)}{\partial N'(k)} \) mean that partial differentiation is applied to each element independently, under the assumption that each element in the vector is uncorrelated.

In the above description, we defined Eq. (3) as speech state transition model and applied Eq. (3) to Kalman filtering estimation to estimate speech power spectra \( S(k) \) from noisy power spectra \( X(k) \).

### 3. Kalman Filtering Estimation

#### 3.1. The State Space Model

To estimate the \( S(k) \) by using Kalman filtering estimation, we determined the state space model as follows:

\[ S(k+1) = F_s S(k) + G_s W(k) \tag{6} \]
\[ X(k) = S(k) + N(k) \tag{7} \]

In above equations, Eq. (6) corresponds to state equation, and Eq. (7) corresponds to observation equation.

#### 3.2. Kalman Filtering Algorithm

By using the state space model described in 3.1, Kalman filtering algorithm is obtained as follows:

\[ S(k) = F_{s,i-1} S(k-1) + K_i (X(k) - F_{s,i-1} S(k-1)) \tag{8} \]
\[ K_i = Q_k \left[ Q_s + \Sigma_{N(k)} \right]^{-1} \tag{9} \]
\[ Q_k = F_{s,i-1} (I - K_{i-1}) Q_{s,i-1} F_{s,i-1}^T + G_{s,i-1} \Sigma_{W(i-1)} G_{s,i-1}^T \tag{10} \]
where \( S(k) \) denotes the estimation of \( S(k) \) and \( Q_k \) denotes diagonal co-variance matrix of the estimating error respectively.

The initial values for Eq. (8)~(10) are represented as follows:

\[ S(0) = 0 \tag{11} \]
\[ Q_0 = 0 \tag{12} \]

In Eq. (10), \( \Sigma_{W(i)} \) denotes diagonal co-variance matrix of \( W(k) \). \( \Sigma_{W(i)} \) is computed by the following equation under the assumption that \( W(k) \) follows zero mean Gaussian process.

\[ \Sigma_{W(i)} = W(k) W(k)^T \tag{13} \]

On the other hand, in Eq. (9), \( \Sigma_{N(k)} \) denotes diagonal co-variance matrix of \( N(k) \). \( \Sigma_{N(k)} \) is computed by the following equation under the assumption that \( N(k) \) follows zero mean Gaussian process as well as \( W(k) \).

\[ \Sigma_{N(k)} = N(k) N(k)^T \tag{14} \]

#### 3.3. Linear Predictive Estimation for \( N(k) \)

To compute the \( \Sigma_{W(i)} \) and \( \Sigma_{N(k)} \), the value of \( N(k) \) is required. However, observable value is only \( X(k) \). Therefore, we have to estimate the value of \( N(k) \) by using 8th order linear prediction expressed as follows:

\[ N_j(k) = \begin{cases} \frac{|X_j(k)|}{\sum_{i=0}^{8} a_{ji} N_j(k-i)} & 0 \leq k < p \\ 0 & k \geq p \end{cases} \tag{15} \]

where \( j \) denotes the channel number in FFT analysis and \( a_{ij} \) denotes the linear predictive coefficient at channel \( j \).

In Eq. (15), when \( 0 \leq k < p \), \( N_j(k) \) is obtained as \( N_j(k) = X_j(k) \) under the assumption that the time section \( 0 \leq k < p \) exists where only the noise component is included as at the beginning of utterance and when \( k \geq p \), \( N_j(k) \) is estimated by the linear predictive estimation. In this paper, the number of linear predictive coefficient \( p \) was set to 14.

### 4. Experiments

LVCSR experiments were carried out for the speech signals estimated by the proposed speech signal estimation method.

#### 4.1. Experimental conditions

The experimental materials are 100 sentences spoken by 23 Japanese males. These materials are taken from the IPA Information Technology Promotion Agency (IPA), Japan-98-TestSet. The noises are non-stationary music of 3 pianos solos (Piano 1, Piano 2 and Piano 3). They are added to clean speech signal by a computer as shown in Eq. (16), changing the SNR at 3 levels: 0dB, 10dB and 20dB.

\[ x(t) = s(t) + \frac{P_{on}}{10^{SNR/10}} n(t) \tag{16} \]

where \( x(t) \), \( s(t) \) and \( n(t) \) are noisy speech, clean speech and noise respectively. \( P_{on} \) and \( P_{on} \) are RMS (Root Mean Square) power of clean speech and RMS power of noise respectively.

We carried out LVCSR using speaker independent monophone HMMs. Their structure is composed of 5 states with 3 loops and 12 mixtures for each state. They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data was taken from the database of Acoustical Society of Japan. The feature parameters are composed of 39 MFCCs with 12 MFCCs, log energy and their first and second order derivatives. CMN (Cepstral Mean Normalization) is applied to each sentence to remove the difference of input circumstances. Table 1 and 2
summarize the experimental conditions for acoustic analysis and phoneme HMM. Here, MFCC as feature parameters for LVCSR was not computed from the wave form which was reconstructed from power spectra of the speech estimated by the proposed method, but it was computed directly from estimated power spectra by using Mel Filter Bank and DCT. Then CMN is applied to each sentence as well as the training data.

<table>
<thead>
<tr>
<th>Table 1: Acoustic analysis conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frequency</td>
</tr>
<tr>
<td>Pre-emphasis</td>
</tr>
<tr>
<td>Feature parameter (Noise reduction)</td>
</tr>
<tr>
<td>Feature parameter (Recognition)</td>
</tr>
<tr>
<td>Analysis frame length</td>
</tr>
<tr>
<td>Analysis frame shift</td>
</tr>
<tr>
<td>Analysis window</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Structure of phoneme HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of states</td>
</tr>
<tr>
<td>Number of loops</td>
</tr>
<tr>
<td>Number of mixtures</td>
</tr>
<tr>
<td>Number of phonemes</td>
</tr>
<tr>
<td>Type</td>
</tr>
</tbody>
</table>

As a language model, bigram is employed in the 1st-pass of the continuous speech decoder and trigram in the 2nd-pass. It was trained using the Mainichi newspaper articles of 75 months. The number of the words in the dictionary is 20,000.

4.2. Experimental Results

Table 3, 4 and 5 show results of LVCSR. In each table, LVCSR results are evaluated by word accuracy (Acc) defined by Eq.17.

$$\text{Acc}(\%) = \frac{N - S - D - I}{N} \times 100$$ (17)

$S$ : The number of substituted words
$D$ : The number of deleted words
$I$ : The number of inserted words
$N$ : Total number of words

In each table, the proposed speech signal estimation method showed the improvement of Acc. However, the improvement was small under all the conditions. The reason why the improvement was small, can be explained that the accurate noise power spectra $N(k)$ was not obtained because linear prediction error of $N(k)$ in Eq.15 was large. In preliminary experiments, we confirmed that if true $N(k)$ is given, the improvement of Acc is significantly large (for example, Acc is improved up to approximately 78% from 36.33% for Piano1 under 0dB noisy environment.). Therefore, it is necessary to estimate $N(k)$ as accurate as possible.

In addition, in Eq.3, we applied the first order Taylor expansion to Eq.2. However, the first order Taylor expansion is rough approximation in accuracy. Therefore, approximation accuracy should be improved by using the higher order Taylor expansion.

5. Iterative MLLR Adaptation

In above experiments, the improvement of Acc was small due to the linear prediction error of $N(k)$. From this fact, it can be assumed that the spectral distortion was caused by Kalman filtering residual noise on estimated speech signal. To solve this problem, we propose here to adapt the acoustic models to spectral distortion by using iterative unsupervised MLLR (Maximum Likelihood Linear Regression) adaptation. In this paper, the adaptation materials are same as the experimental materials, and the number of the maximum iteration was set to 5.

In the unsupervised MLLR adaptation, speaker independent monophone HMM is moved toward speech signal estimated by Kalman filter. In order to make this adaptation feasible, monophone labels are required for the estimated speech signal. To obtain the monophone labels, the LVCSR is carried out to the speech signal estimated by Kalman filter. Then the MLLR is applied to the monophone HMMs by using estimated speech and the monophone labels.

6. Experimental Results with Iterative MLLR Adaptation

Table 6, 7 and 8 show the results of the LVCSR after the iterative MLLR adaptation. In each table, 'Without NR' denotes the result without noise reduction described in Sec.2, 3 and 'With NR' denotes the result with the noise reduction. In the iteration of MLLR, 'With NR' indicates the LVCSR results after iterative MLLR adaptation described in Sec.5. On the other hand, 'Without NR' indicates the LVCSR results after iterative MLLR adaptation without Kalman filter, in monophone labels acquisition as well as monophone HMMs adaptation.
In each table, Acc has almost saturated until up to 2 iterations. In each noise environment, the MLLR adaptation showed the improvement compared with the result obtained without MLLR at 0 iteration. Lower the SNR, more significant improvement was achieved. Especially, at 0dB SNR, speech recognition was improved by 20~30%. Compared the result by ‘Without NR’ and the result by ‘With NR’, they showed almost equal result at 0dB and 20dB SNR. However, it showed significant improvement at 0dB SNR (almost 30% improvement). From this fact, two conclusions are obtained.

1. The MLLR can indeed adapt the monophone HMMs to the noise environment regardless the noise component is reduced or not.

2. When the noise component is reduced, unsupervised MLLR can adapt the monophone HMMs toward estimated speech signals due to the noise reduction effect.

In our method, significant improvement was obtained at 0dB SNR, since Kalman filter reduced the non-stationary noise component and MLLR adapted to monophone HMMs toward the estimated speech signals with residual noise.

7. Conclusions

In this paper, we proposed the speech recognition method under non-stationary musical environments using Kalman filtering speech signal estimation method and iterative unsupervised MLLR adaptation and showed the significant improvement in word accuracy. In future, to improve the word accuracy under any types of non-stationary noisy environments, we are planning to develop accurate estimation method even for state transition of the noise spectra in the proposed speech signal estimation method. Furthermore, we will study more accurate adaptation method to the spectral distortion caused by residual noise.

8. References