EVALUATION OF NOISY SPEECH RECOGNITION BASED ON NOISE REDUCTION AND ACOUSTIC MODEL ADAPTATION ON THE AURORA2 TASKS

M. Fujimoto and Y. Ariki
Department of Electronics and Informatics
Ryukoku University, Seta, Otsu-shi, Shiga, 520-2194, JAPAN
masa@arikilab.elec.ryukoku.ac.jp

ABSTRACT

In this paper, we have evaluated a noisy speech recognition method based on noise reduction and acoustic model adaptation, on the AURORA2 tasks.

For noise reduction method, we employed two noise reduction methods. One is an Adaptive Sub-Band Spectral Subtraction (ASBSS) method which can optimize the noise subtraction rate according to the SNR in frequency bands at each frame. The other is a Kalman filtering estimation method which re-estimates the accurate speech spectra from those estimated by ASBSS. The accurate speech spectra was estimated by combining these methods. Usually, a noise reduction method has a problem that it degrades the recognition rate because of spectral distortion caused by residual noise occurred through noise reduction and over estimation. To solve the problem in noise reduction method, adaptation of the acoustic models is employed by using an unsupervised MLLR adaptation to the spectral distortion.

In evaluation on the AURORA2 tasks, our method showed the significant improvement in recognition accuracy for both clean training condition and multi training condition.

1. INTRODUCTION

In recent years, many types of speech recognition systems have been proposed and developed toward the practical use in a 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.

Noise robust speech recognition methods are classified into two types. One reduces the noise component from noisy speech based on noise reduction techniques[1]-[4]. The other adapts itself to any kinds of noises based on model adaptation techniques[5].

In noise reduction method, Spectral Subtraction(SS)[1] has been proposed as the conventional noise reduction method. It is effective, considering that its computation amount is small. In the SS, speech spectra are estimated by adjusting the noise subtraction rate according to the SNR. Generally, SNR is defined and computed as the average over all the input speech signal. However, even if the noise is stationary, SNR varies according to speech energy. Therefore, the subtraction coefficient should be adjusted according to the segmental SNR. This method has been proposed as an Adaptive SS(ASS)[2]. Furthermore, in the SS, the subtraction coefficient is usually fixed for all the frequency bands under the assumption that the SNR is same throughout the frequency bands. However, SNR optimizes according to the spectral features such as formants of vowels. Therefore, the subtraction coefficient should be adjusted according to the sub-band SNR. This method has been proposed as Non-linear SNR(NSS)[3].

In this paper, we employ an Adaptive Sub-Band Spectral Subtraction (ASBSS) method[4] by combining ASS and NSS. Namely, it optimizes the subtraction coefficient according to SNR in frequency bands at each frame. In addition, to obtain higher noise reduction accuracy, we combined the ASBSS with a Kalman filtering estimation method.

Here, the noise reduction method has a problem that it degrades the recognition rate because of spectral distortion caused by residual noise occurred through noise reduction and over estimation. To solve this problem in noise reduction method, adaptation of the acoustic models is employed by using unsupervised MLLR adaptation[6] to spectral distortion.

Our proposed method has been evaluated on the AURORA2 tasks. In the evaluation results, the proposed method showed the significant improvement in recognition accuracy for both clean training condition and multi training condition.

2. ADAPTIVE SUB-BAND SPECTRAL SUBTRACTION

2.1. Adaptive Spectral Subtraction

Let \( X(f,i) \) denote a power spectrum of noisy speech (\( f \) denotes a channel number in FFT analysis and \( i \) denotes a frame index.), \( \hat{X}(f) \) denote an estimated power spectrum of noise and \( \hat{S}(f,i) \) denote the estimated power spectrum...
of clean speech, then SS is presented as follows:

$$\hat{S}(f, i) = \max \left[ X(f, i) - \alpha \overline{N}(f), \beta X(f, i) \right]$$  

(1)

where $\alpha$ and $\beta$ are called the subtraction coefficient and the flooring coefficient.

In the SS, the speech spectrum is estimated by adjusting subtraction coefficient $\alpha$ according to the SNR. Generally, SNR is defined and computed as the average over all the input speech signal. However, even the noise is stationary, SNR varies according to clean speech energy. Therefore, $\alpha$ should be determined by the segmental SNR $SNR(i)$ and the subtraction coefficient determination function $g$ as follows.

$$\alpha(i) = g(SNR(i))$$  

(2)

Let $Pow_s(i)$ denote the short time RMS (Root Mean Square) power of noisy speech, $\overline{Pow_n}$ denote the estimated short time RMS power of noise and $Pow_s(i)$ denote the estimated short time RMS power of clean speech, then $SNR(i)$ is estimated as follows.

$$SNR(i) = \begin{cases} 
20 \log_{10} \frac{Pow_s(i)}{\overline{Pow_n}}, & Pow_s(i) > 0 \\
\gamma (-10) \left( \frac{Pow_s(i)}{\overline{Pow_n}} \right) \leq 0 
\end{cases}$$  

(3)

$$Pow_s(i) = Pow_s(i) - \overline{Pow_n}$$  

(4)

If $Pow_s(i)$ has minus value, Eq.(3) cannot compute $SNR(i)$. In this case, $\gamma$ is substituted for $SNR(i)$.

Fig.1 shows the subtraction coefficient determination function $g(SNR(i))$ as a function of segmental SNR $SNR(i)$ used in this study. In Fig.1, when $SNR(i)$ is less than 0dB, the subtraction rate is maximized to $\alpha(i) = 2.0$ and when $SNR(i)$ is more than 30dB, $\overline{N}(f)$ is not subtracted. Here, the subtraction coefficient determination function $g(SNR(i))$ shown in Fig.1 was designed depending on preliminary experiments. The flooring coefficient was adjusted to $\beta = 0.2$.

![Fig. 1. Subtraction coefficient determination function $g(SNR(i))$](image)

**2.2. Expansion to Adaptive Sub-Band Spectral Subtraction**

ASS described in Sec.2.1 determines the subtraction coefficient $\alpha(i)$ at each frame, but fixes it for all frequency bands under the assumption that the SNR is same throughout the frequency bands. However, SNR varies even in frequency according to the spectral features such as vowels and consonants. Therefore the subtraction coefficient should be adjusted by the sub-band segmental SNR. We propose this method as Adaptive Sub-Band Spectral Subtraction (ASBSS). The processing flow of ASBSS is shown in Fig.2.

![Fig. 2. Processing flow of ASBSS](image)

In Fig.2, the detailed algorithm is described as follows.

1. Pass the noisy algorithm into the Mel scale band pass filter with 8 channels, to obtain sub-band wave forms.

2. Estimate the sub-band segmental SNR $SNR(k, i)(k)$ denotes the channel number of sub-band.) at each sub-band using RMS powers of noisy speech and the estimated noise.

3. Determine the subtraction coefficient $\alpha(k, i)$ in each sub-band based on the $SNR(k, i)$.

4. Estimate the speech spectrum by subtracting the noise spectrum computed by using $\alpha(k, i)$ from the noisy speech spectrum.

**3. KALMAN FILTERING ESTIMATION**

In this section, a Kalman filter based re-estimation method of speech power spectra is described.
3.1. The State Space Model
To re-estimate the speech power spectrum \( S(f, i) \) by using the Kalman filtering estimation, we determined the state space model as follows:

\[
\hat{S}(f, i + 1) = F_{f,i} \hat{S}(f, i) \quad (5)
\]
\[
F_{f,i} = \frac{\hat{S}(f, i + 1)}{S(f, i)} \quad (6)
\]
\[
X^l(f, i) = \log \left( \hat{S}(f, i) + \hat{N}(f, i) \right)
\]
\[
= \hat{S}(f, i) + \log \left( 1 + \frac{\hat{N}(f, i)}{\hat{S}(f, i)} \right)
\]
\[
= \hat{S}(f, i) + V(f, i) \quad (7)
\]
\[
\hat{N}(f, i) = X(f, i) - \hat{S}(f, i) \quad (8)
\]
\[
V(f, i) = \log \left( 1 + \frac{\hat{N}(f, i)}{\hat{S}(f, i)} \right) \quad (9)
\]

where \( \hat{S}(f, i) \) denotes the estimation of \( S(f, i) \), \( \hat{S}(f, i) \) denotes the speech power spectrum estimated by ASBSS and superscript \( l \) denotes the log-spectral domain. Eq.(5) corresponds to the 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:

\[
\hat{S}_i^l = F_{i-1} \hat{S}_{i-1}^l + K_i (X_i^l - F_{i-1} \hat{S}_{i-1}^l) \quad (10)
\]
\[
K_i = \frac{Q_i}{Q_i + \Sigma V_i} \quad (11)
\]
\[
Q_i = F_{i-1} (I - K_{i-1}) Q_{i-1} F_{i-1}^T \quad (12)
\]
\[
\hat{S}_i^l = \left( \hat{S}_i^l(0,0), \ldots, \hat{S}_i^l(N-1,0) \right)^T \quad (13)
\]
\[
X_i^l = \left( X_i^l(0,0), \ldots, X_i^l(N-1,0) \right)^T \quad (14)
\]
\[
F_i = \text{diag}(F_{0,0}, \ldots, F_{N-1,i}) \quad (15)
\]

where \( N \) denotes the number of channels in FFT analysis and \( \Sigma V_i \) denotes the diagonal co-variance matrix of the estimation error.

The initial values for Eq.(10)~(12) are represented as follows:

\[
\hat{S}_0^l = \left( \hat{S}_0^l(0,0), \ldots, \hat{S}_0^l(N-1,0) \right)^T \quad (16)
\]
\[
Q_0 = \eta \cdot I \quad (\eta = 0.001) \quad (17)
\]

In Eq.(11), \( \Sigma V_i \) denotes diagonal co-variance matrix of \( \Sigma V_i \) is computed by the following equation under the assumption that \( \Sigma V_i \) follows zero mean Gaussian process.

\[
\Sigma V_i = V_i V_i^T \quad (18)
\]

4. UNSUPERVISED MLLR ADAPTATION
Usually, the noise reduction method has a problem that it degrades the recognition rate because of spectral distortion caused by the residual noise occurred through the noise reduction and over estimation. To solve this problem in the noise reduction method, adaptation of the acoustic models is employed by using (on-line) unsupervised MLLR adaptation\[6\] to the spectral distortion.

In the unsupervised MLLR adaptation, speaker independent digit HMM is adapted to speech signal estimated by the noise reduction method. To make this adaptation feasible, digit labels are required for the estimated speech signal. To obtain these labels, the speech recognition is carried out using digit HMMs before adaptation and dictionary, to the signal estimated by the noise reduction method. Then the MLLR is applied to the HMMs using the estimated speech and the labels. In this paper, the adaptation material is an input speech and the number of normal distribution clusters included in the digit HMMs was set to 1.

5. EXPERIMENTS

Our proposed noisy speech recognition method has been evaluated on the AURORA2 database by using both standard recognition back-end prepared by the AURORA2 database and complex recognition back-end[7, 8] prepared by Asela Gunawardana(Micorsoft Corp.). In all the experiments, the condition of feature extraction was kept same. In each back-end, we evaluated following two methods.

Method(1) : ASBSS + MLLR
Method(2) : ASBSS + Kalman filter + MLLR

5.1. Experimental Results by Standard Back-end
Recognition results by standard back-end are shown in Table 1. and 2..

In the clean training condition, the results by method(1) showed the significant relative improvement. On the other hand, in the multi training condition, the results by method(2) showed the significant relative improvement. Especially, in the clean training condition, relative improvement of method(2) was small at low SNR condition. The reason that the relative improvement by method(2) was large in the multi training condition though small in the clean training condition can be explained by the following two reasons.

(1) In the Kalman filtering estimation, the estimation accuracy of \( S(f, i) \) depends on the transition coefficient \( F_{f,i} \). As shown in Eq.(6), the reliability of \( F_{f,i} \) depends on the estimation accuracy of the ASBSS, since it is computed by using log-power spectra \( \hat{S}_i^l(f, i + 1) \) and \( \hat{S}_i^l(f, i) \) estimated by ASBSS. However, at low SNR condition, the reliability of \( F_{f,i} \) degrades, because the estimation accuracy of ASBSS degrades at low SNR condition. From this fact, it can be assumed that the spectral distortion
is occurred by significant estimation error in Kalman filter caused by reliability degradation of $F_{f,i}$.

(II): In this evaluation, the HMMs are trained by using training materials after noise reduction at both clean and multi training condition. In the multi training condition, the spectral distortion at low SNR condition is reflected on the HMMs because the HMMs are trained by using training materials distorted by applying noise reduction to noisy training materials. However, in the clean training condition, the spectral distortion at low SNR condition is not reflected on the HMMs because the prepared training materials are clean only.

From these facts, method(2) showed the significant relative improvement in the multi training condition.

5.2. Experimental Results by Complex Back-end
Recognition results by the complex back-end are shown in Table 3. and 4...

In each table, comparing with the results in Sec. 5.1, the results by the complex back-end showed the significant relative improvement. Furthermore, comparing the results in method(1) with method(2), it can be said that the tendency is same as the results in Sec. 5.1.

6. CONCLUSIONS
In this paper, we proposed a noisy speech recognition method based on noise reduction and acoustic model adaptation and evaluated it on the AURORA2 tasks. As the evaluation results, our method showed the significant improvement in recognition accuracy both for the clean training condition and the multi training condition. In future, to improve the recognition accuracy in clean training condition, we will study more accurate noise reduction method.

7. REFERENCES