LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION
UNDER REAL ENVIRONMENTS USING
ADAPTIVE SUB-BAND SPECTRAL SUBTRACTION

M.Fujimoto, J.Ogata and Y.Ariki

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

ABSTRACT

In this study, we propose an Adaptive Sub-Band Spectral Subtraction (ASBSS) method which can vary noise subtraction rate according to SNR in frequency bands at each frame. In the conventional Spectral Subtraction (SS), speech spectral is estimated by adjusting noise subtraction rate according to SNR. In general, 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 rate of noise spectral should be adjusted according to the segmental SNR. This method is called Adaptive SS (ASS). Considering difference of spectral features such as vowel and consonant, the subtraction rate of noise spectral should be adjusted according to the sub-band SNR. This idea leads to the ASBSS method we propose in this paper. In order to evaluate the proposed method, we carried out Large Vocabulary Continuous Speech Recognition experiments and compared the results by our method with the conventional method in word accuracy.

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.

Robust speech recognition systems are classified into two types. One adapts itself to any kinds of noises based on model adaptation techniques. The other reduces the noise component from noisy speech based on noise reduction techniques.

Parallel Model Combination (PMC) [1, 2] has been proposed which adapts the speech recognition system to any kinds of noises. But this method has a problem that PMC needs a huge quantity of computation, if it is applied to the acoustic model which has a large number of phonemes with mixture distributions like a triphone model HMM.

On the other hand, Spectral Subtraction (SS) [3] has been proposed as a conventional noise reduction method. It is effective, considering that its computation amount is small. In the SS, speech spectral is estimated by adjusting subtraction rate of noise spectral (value of subtraction coefficient) according to SNR. In general, 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 Adaptive SS (ASS) [4].

Furthermore, in the SS, the subtraction coefficient is usually fixed for all frequency bands under the assumption that the SNR is same all over the frequency bands. However, SNR varies according to the spectral features such as vowel and consonant. Therefore, the subtraction coefficient should be adjusted according to the sub-band SNR. This method has been proposed as Non-linear SS (NSS) [5, 6].

In this study, we propose an Adaptive Sub-Band Spectral Subtraction (ASBSS) method by combining ASS and NSS. Namely it varies the subtraction coefficient according to SNR in frequency bands at each frame. In order to evaluate the proposed method, we carried out Large Vocabulary Continuous Speech Recognition (LVCSR) experiments and compared the results by our method with SS and ASS in word accuracy.

2. ADAPTIVE SPECTRAL SUBTRACTION

Let $X(f,l)$ denote power spectral of noisy speech ($f$ denote channel number in FFT analysis and $l$ denote frame index.), $\hat{S}(f,l)$ denote power spectral of clean speech and $\hat{N}(f)$ denote estimated power spectral of noise, then SS is presented as follows:

$$\hat{S}(f,l) = \max \left[ X(f,l) - \alpha \hat{N}(f), \beta X(f,l) \right]$$ (1)
where $\alpha$ and $\beta$ are called the subtraction coefficient and the flooring coefficient respectively.

In the SS, the speech spectral is estimated by adjusting subtraction coefficient $\alpha$ according to SNR. In general, 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 according to the segmental SNR $SNR(l)$ and the subtraction coefficient determination function $f$ as follows.

$$\alpha(l) = f(SNR(l))$$  

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

$$SNR(l) = \begin{cases} 
20 \log_{10} \frac{Pow_s(l)}{Pow_n} & Pov_s(l) > 0 \\
\gamma \ (=-10) & Pov_s(l) \leq 0 
\end{cases}$$  

(3)

$$Pow_x(l) = Pow_x(l) - \overline{Pow_n}$$  

(4)

If $Pow_s(l)$ has minus value, Eq.(3) can’t compute $SNR(l)$. In this case, $\gamma$ is substituted for $SNR(l)$.

Fig.1 shows the subtraction coefficient determination function $f$ as a function of segmental SNR $SNR(l)$ used in this study. In Fig.1, when $SNR(l)$ is less than 0dB, the subtraction rate is maximized as $\alpha(l) = 2.2$ and when $SNR(l)$ is more than 40dB, $\overline{Pow_n}$ is not subtracted. Here, the subtraction coefficient determination function $f$ shown in Fig.1 was designed depending on preliminary experiments. The flooring coefficient was adjusted as $\beta = 0.2$.

![Figure 1: Subtraction coefficient determination function $f$](image)

3. ADAPTIVE SUB-BAND SPECTRAL SUBTRACTION

ASS described in Sec.2 varies the subtraction coefficient at each frame, but fixes it for all frequency bands under the assumption that the SNR is same all over the frequency bands. However, SNR varies according to the spectral features such as vowel and consonant. Therefore the subtraction coefficient should be adjusted according to the subband segmental SNR. We propose this method as Adaptive Sub-Band Spectral Subtraction (ASBSS). The processing flow of ASBSS is shown in Fig.2.

![Figure 2: Processing flow of ASBSS](image)

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

1. In order to obtain sub-band wave forms, pass the noisy speech into the Mel scale band pass filter with $N$ channels.
2. Estimate the sub-band segmental SNR $SNR(k,l)$ ($k$ denote the channel number of sub-band) at each sub-band by using RMS power of noisy speech and RMS power of noise.

3. Determine the subtraction coefficient $\alpha(k,l)$ in each sub-band according to $SNR(k,l)$.

4. Estimate the speech spectral by subtracting the noise spectral according to $\alpha(k,l)$ from the noisy speech spectral.

### 4. EXPERIMENTS

LVCSR experiment was carried out for the signals estimated by ASBSS. As a comparison, LVCSR by SS and ASS was also carried out.

#### 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, Japan)-98-TestSet. The noises are car noise, exhibition hall noise and speech like noise. They are added to clean speech signal by a computer as shown in Eq.(6), changing the SNR at 3 levels; 0dB, 10dB and 20dB.

$$x(t) = s(t) + \frac{Pow_s}{10^{SNR/20}}Pow_n n(t)$$

where $x(t)$, $s(t)$ and $n(t)$ are noisy speech, clean speech and noise respectively. $Pow_s$ and $Pow_n$ are RMS power of clean speech and RMS power of noise respectively.

We carried out LVCSR using speaker independent word interval triphone 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 39 MFCCs with 12 MFCCs, log energy and their first and second order derivatives. Cepstral Mean Normalization(CMN) is applied to each sentence to remove the difference of input circumstances. Table 1, 2 show the experimental conditions for acoustic analysis and HMM.

Here, MFCC as feature parameters for LVCSR was not computed from clean speech wave form reconstructed from power spectral of clean speech estimated by SS, ASS and ASBSS, but it was computed directly from estimated power spectral by using Mel Filter Bank and DCT. Then CMN is applied to each sentence as well as the training data.

#### 4.2. EXPERIMENTAL RESULTS

Table 3, 4 and 5 show results of LVCSR. In each table, ‘No NR’ denotes the result without noise reduction. In each method, upper shows word correct rate (Cor) and lower shows word accuracy (Acc). They are defined by Eq.(7) and Eq.(8).

<table>
<thead>
<tr>
<th>Table 1: Acoustic analysis conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 2: Structure of HMM</strong></td>
</tr>
<tr>
<td><strong>Table 3: Result of car noise mixed (%)</strong></td>
</tr>
<tr>
<td>(Upper: Cor, Lower: Acc)</td>
</tr>
</tbody>
</table>
Table 4: Result of speech like noise mixed(%) 
(Upper: Corr, Lower: Ace)

<table>
<thead>
<tr>
<th>SNR</th>
<th>20dB</th>
<th>10dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>No NR</td>
<td>94.61</td>
<td>80.97</td>
<td>39.27</td>
</tr>
<tr>
<td>SS</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASS</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 2)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 4)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 8)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 16)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
</tbody>
</table>

Table 5: Result of exhibition hall noise mixed(%) 
(Upper: Corr, Lower: Ace)

<table>
<thead>
<tr>
<th>SNR</th>
<th>20dB</th>
<th>10dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>No NR</td>
<td>95.61</td>
<td>88.91</td>
<td>59.27</td>
</tr>
<tr>
<td>SS</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASS</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 2)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 4)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 8)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
<tr>
<td>ASBSS (N = 16)</td>
<td>92.33</td>
<td>85.34</td>
<td>37.16</td>
</tr>
</tbody>
</table>

\[
Corr(\%) = \frac{N - S - D}{N} \times 100 \quad (7)
\]

\[
Ace(\%) = \frac{N - S - D - I}{N} \times 100 \quad (8)
\]

\[S\] : 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, comparing with SS and ASS, the proposed method showed improvement of Ace at all conditions. However, comparing with NoNR, improvement of Ace was small in speech like noise and exhibition hall noise mixed speech at 0dB SNR. On the other hand, Corr was improved in speech like noise and exhibition hall noise mixed speech at 0dB SNR. From this fact, it can be assumed that the number of substituted and deleted words has decreased and the number of inserted words has increased in speech like noise and exhibition hall noise mixed speech at 0dB SNR. As the reason why the number of inserted words has increased, it can be assumed that good estimation accuracy of \( \mathcal{N}(f) \) was not obtained because speech like noise and exhibition hall noise used in this study have the large time variation of power spectral. From this fact, it can be assumed that the residual of time variation of power spectral has affected phoneme recognition result. Therefore, \( \mathcal{N}(f) \) should be estimated considering time variation of noise spectral. Furthermore, it can be required that the LVCSR system should reject the error words based on a confidence measure.

5. CONCLUSIONS

This paper proposed Adaptive Sub-Band Spectral Subtraction which varies noise subtraction rate according to SNR in frequency bands at each frame and showed the improvement of word accuracy. In future, to improve the word accuracy under the any types of noises which have large time variation, we will study a noise reduction method considering time variation of noise spectral.

6. REFERENCES