Improvement of Acoustic Model for Hands-Free Speech Recognition Using Spatial Subtraction Array

Ayase Takagi†, Yoshimitsu Mori†, Yu Takahashi†, Hiroshi Saruwatari† and Kiyohiro Shikano†

†Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN
Phone:+81-743-72-5287 FAX:+81-743-72-3321
Email: ayase-t@is.naist.jp

Abstract

Noise and reverberation adaptation techniques are essential to realize robust speech recognition in actual environments. In this paper, first we describe a new noise-robust speech recognition system combined with our previously proposed spatial subtraction array (SSA) which is effective for hands-free speech recognition. Next, in the SSA-based speech recognition system, we newly introduce an SSA-matched acoustic model, and assess the model’s portability and robustness in various environments. SSA-matched acoustic model can be applied to compensation of the effects due to noise, reverberation and distortion via signal processing of SSA. The experimental results show that our proposed SSA-matched model marks highest recognition performance and is still available among environment changes.

1. Introduction

Recent progress in speech recognition technology has been encouraging various practical speech applications such as robot interface. However, required equipment of a hands-on microphone or a head set microphone to users spoils its convenience and naturalness. Recognition of distance talking by far microphone may ease this problem. Therefore, Hands-free speech system is demanded.

A serious problem of hands-free speech recognition in real environments is recognition-accuracy degradation due to the reverberation effect and background noise. Therefore, the speech recognition performance is often degraded significantly. A possible solution is to address the problem the noise suppression by microphone array signal processing. Many types of microphone arrays, e.g., Delay-and-Sum (DS) and Griffith-Jim adaptive array (GJ), have been proposed in the past research. Although GJ can achieve a superior performance relatively to others, GJ requires a huge amount of calculations for learning adaptive FIR-filter of thousands or millions of taps.

In order to resolve this problem, we newly propose an acoustic model for a spatial subtraction array (SSA) and its improvement method to achieve robust hands-free speech recognition under noisy environments. In the proposed method, the noise reduction is achieved by subtracting the estimated noise spectrum from target speech spectrum to be enhanced in the mel-scale filter bank domain. Moreover, since the proposed method is performed in the mel-scale filter bank domain, the transform into mel-frequency cepstrum coefficient (MFCC) become easier, which reduces the amount of calculation in SSA compared to that of GJ. The experimental results obtained under a real environment reveal that word accuracy of the proposed method is greater than those DS and GJ. In addition we propose SSA-matched model. The report that improved the model who used known noise excellence or excellence of the reverberation used. The conventional researches have introduced a simple acoustic model which was trained under noise-less and distortion-less conditions. Signal-to-noise ratio (SNR) is improved when we use a spectrum subtraction, but a speech spectrum after processing changes from a original speech greatly. Therefore we can not perform an accurate speech recognition by using this acoustic model. We create acoustic model that a noise, influence of the reverberation and a signal distortion to occur when we handled it in SSA. We assume actual environments, and we tested it with plural acoustic models.
2. Proposed method

2.1. SSA

Figure 1 shows a configuration of the proposed SSA. SSA includes mel-scale filter bank analysis, and outputs mel-frequency cepstrum coefficient (MFCC). The triangular window \( W(k; \ell) \) to perform mel-scale filter bank analysis is designated as follows:

\[
W(k; \ell) = \begin{cases} 
  k - k_{lo}(\ell) & (k_{lo}(\ell) \leq k \leq k_{c}(\ell)), \\
  k_{c}(\ell) + k_{hi}(\ell) - k & (k_{hi}(\ell) \leq k \leq k_{hi}(\ell)), \\
  0 & \text{otherwise},
\end{cases}
\]

where \( k_{lo}(\ell), k_{c}(\ell) \) and \( k_{hi}(\ell) \) are the lower, center, and higher frequency bins of each triangle window respectively. They satisfy the relation among adjacent windows as

\[
k_{c}(\ell) = k_{hi}(\ell - 1) - k_{lo}(\ell + 1),
\]

Moreover, \( k_{c}(\ell) \) is arranged in regular intervals on mel-frequency domain. Mel-scale frequency \( Mel_{k}(\ell) \) for \( k_{c}(\ell) \) is calculated as

\[
Mel_{k}(\ell) = 2595 \log_{10} \left\{ 1 + \frac{k_{c}(\ell) \cdot f_s}{700 \cdot K} \right\},
\]

where \( f_s \) is the sampling frequency and \( M \) is the DFT size.

In the proposed SSA, noise reduction is achieved by subtracting the estimates noise power spectrum from the target speech power spectrum to be enhanced in the mel-scale filter bank domain. This offers a realization of error-robust noise reduction with few computational complexities because the parameters are optimized in the small number of mel-scale filter banks. This procedure is given by

\[
m(l, \tau) = \sum_{k = k_{lo}(\ell)}^{k_{hi}(\ell)} W(k, \ell) |Y_{DS}(k, \tau)|^2 - \alpha(\tau) \cdot \beta \cdot |Y_{NBF}(k, \tau)|^2, \quad \text{(if } |Y_{DS}(k, \tau)|^2 - \alpha(\tau) \cdot \beta \cdot |Y_{NBF}(k, \tau)|^2 \geq 0),
\]

where \( k \) is the frequency bins, \( k_{lo}(\ell) \) and \( k_{hi}(\ell) \) are the lower and higher frequency bins of each triangle windows respectively, \( \tau \) is the number of filter, \( m(l, \tau) \) is the output from the mel-scale filter bank, \( W(k, \ell) \) is the triangular window to perform mel-scale filter bank analysis. \( Y_{DS}(k) \) is the output signal from DS, i.e., the partly enhanced speech signal, and \( Y_{NBF}(k) \) is the output signal from NBF in which the directional null steers in DOA of the user, i.e., the estimated noise signal. The system switches in two equations depending on the conditions in (1).

\[
m(l) = \begin{cases} 
  k - k_{lo}(\ell) & (k_{lo}(\ell) \leq k \leq k_{c}(\ell)), \\
  k_{c}(\ell) + k_{hi}(\ell) - k & (k_{hi}(\ell) \leq k \leq k_{hi}(\ell)), \\
  0 & \text{otherwise},
\end{cases}
\]

where \( k_{lo}(\ell), k_{c}(\ell) \) and \( k_{hi}(\ell) \) are the lower, center, and higher frequency bins of each triangle window respectively, \( k_{hi}(\ell) \) is arranged in regular intervals on mel-frequency domain.

Moreover, \( k_{c}(\ell) \) is calculated as

\[
Mel_{k}(\ell) = 2595 \log_{10} \left\{ 1 + \frac{k_{c}(\ell) \cdot f_s}{700 \cdot K} \right\},
\]

where \( f_s \) is the sampling frequency and \( M \) is the DFT size.

In the proposed SSA, noise reduction is achieved by subtracting the estimates noise power spectrum from the target speech power spectrum to be enhanced in the mel-scale filter bank domain. This offers a realization of error-robust noise reduction with few computational complexities because the parameters are optimized in the small number of mel-scale filter banks. This procedure is given by

\[
m(l, \tau) = \sum_{k = k_{lo}(\ell)}^{k_{hi}(\ell)} W(k, \ell) |Y_{DS}(k, \tau)|^2 - \alpha(\tau) \cdot \beta \cdot |Y_{NBF}(k, \tau)|^2, \quad \text{(if } |Y_{DS}(k, \tau)|^2 - \alpha(\tau) \cdot \beta \cdot |Y_{NBF}(k, \tau)|^2 \geq 0),
\]

where \( k \) is the frequency bins, \( k_{lo}(\ell) \) and \( k_{hi}(\ell) \) are the lower and higher frequency bins of each triangle windows respectively, \( \tau \) is the number of filter, \( m(l, \tau) \) is the output from the mel-scale filter bank, \( W(k, \ell) \) is the triangular window to perform mel-scale filter bank analysis. \( Y_{DS}(k) \) is the output signal from DS, i.e., the partly enhanced speech signal, and \( Y_{NBF}(k) \) is the output signal from NBF in which the directional null steers in DOA of the user, i.e., the estimated noise signal. The system switches in two equations depending on the conditions in (1).

\[
m(l) = \begin{cases} 
  k - k_{lo}(\ell) & (k_{lo}(\ell) \leq k \leq k_{c}(\ell)), \\
  k_{c}(\ell) + k_{hi}(\ell) - k & (k_{hi}(\ell) \leq k \leq k_{hi}(\ell)), \\
  0 & \text{otherwise},
\end{cases}
\]

where \( k_{lo}(\ell), k_{c}(\ell) \) and \( k_{hi}(\ell) \) are the lower, center, and higher frequency bins of each triangle window respectively, \( k_{hi}(\ell) \) is arranged in regular intervals on mel-frequency domain.

Moreover, \( k_{c}(\ell) \) is calculated as

\[
Mel_{k}(\ell) = 2595 \log_{10} \left\{ 1 + \frac{k_{c}(\ell) \cdot f_s}{700 \cdot K} \right\},
\]

where \( f_s \) is the sampling frequency and \( M \) is the DFT size.

2.2. Acoustic model improvement for SSA

In the speech recognizer part, the conventional researches have introduced a simple acoustic model which was trained under noise-less and distortion-less conditions, so called clean model. However there exist the following problems especially in SSA applications:

(a) SSA's noise reduction performance is limited to some extent, and consequently there are still some residual noise components in the SSA's output.

(b) SSA involves highly nonlinear signal processing, and this leads to creation of sound distortion in the enhanced target speech. The above-mentioned deviations from ideal clean conditions give us a mismatch of acoustic model in the speech recognizer, and this yields heavy deterioration in word recognition score. To solve this problem, in this work we propose to construct a noise- and distortion-specific acoustic model which reflects prospective conditions in the actual use of SSA. The training of the acoustic model is carried out by considering the characteristics on the sound distortion and noise residuals, which can be done with handling of SSA in an input speech. This strategy can really make distortions of SSA used at the time of recognition agree with the acoustic model.
trained in advance. Figure 2 depicts an overview of the proposed method for creating SSA-matched acoustic model. A flow of concrete processing is as follows:

1. We convolute an impulse response measured beforehand by microphone array in a speech database.
2. In a speech database made with 1, we superimpose excellence of an exterior noise of a fixed quantity.
3. An SSA dispense for 2. And it is an expression by a case(1). We perform known noise excellence after subtraction we am similar, and to be able to put [ ].
4. We learn it with EM algorithm and make an SSA matched acoustic model.
5. We give SSA for an input evaluation speech and perform known noise superimposition excellence after spectral subtraction by a case. We do speech recognition it with 4. SSA-matched models.

### Table 1: Experimental conditions

<table>
<thead>
<tr>
<th>Database</th>
<th>JNAS [ ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>speakers</td>
<td>20-k</td>
</tr>
<tr>
<td>Task</td>
<td>newspaper dictation</td>
</tr>
<tr>
<td>User angle</td>
<td>0°</td>
</tr>
<tr>
<td>Noise</td>
<td>cleaner</td>
</tr>
<tr>
<td>Noise angle</td>
<td>60°</td>
</tr>
<tr>
<td>Noise superimposition</td>
<td>10 dB</td>
</tr>
<tr>
<td>Acoustic model</td>
<td>PTM [ ]</td>
</tr>
<tr>
<td></td>
<td>(2000 stats, 64 mixture size)</td>
</tr>
<tr>
<td></td>
<td>clean, matched, clean &amp; reverberation, matched &amp; reverberation</td>
</tr>
<tr>
<td>Decoder</td>
<td>Julius ver.3.4.2 [ ]</td>
</tr>
<tr>
<td>Filter size</td>
<td>32 ms (512 taps)</td>
</tr>
<tr>
<td>Frame size</td>
<td>25 ms (400 samples)</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>16 kHz</td>
</tr>
<tr>
<td>Known noise</td>
<td>Office room noise</td>
</tr>
<tr>
<td>Amount of superimposition</td>
<td>30 dB</td>
</tr>
<tr>
<td>( \beta ) (SSA)</td>
<td>0.5, 1.0, 2.0</td>
</tr>
<tr>
<td>( \gamma ) (SSA)</td>
<td>0.2</td>
</tr>
</tbody>
</table>

### 3. Experiments and results

In this section, we evaluate our proposed noise and reverberation robust speech recognition.

#### 3.1. Conditions

Figures 3 and 4 show layouts of the reverberant rooms used in the experiment and Table 1 shows the experimental conditions. In the experiment, we use the following signals as testing data: the original speech convoluted with the impulse responses which are recorded in the actual environment, and added with exterior noise and office noise which is included in the actual environment. In this paper, we compare clean model, reverberation and known noise matched model, reverberation and exterior noise matched model, Delay-and-Sum (DS) matched model and SSA matched model. We construct each model under the conditions of 260-ms reverberation and 200-ms reverberation. Regarding the recorder, JULIUS is used. We use a Phonetic Tied Mixture[2] model from JNAS database. The evaluation task is the JNAS newspaper dictation task with 20k vocabulary size. The baseline speaker independent acoustic models are trained from 260 training speakers’ data in JNAS speech database. PTM, phonetic tied mixture models. are used. The PTM training speech database includes 260 speakers (39000 sentence utterances in total).
The test set consists of another 46 speakers from JNAS. Each test speaker utters 4 or 5 newspaper article sentences (200 test sentence utterances in total). The distance between the microphone array and the loudspeakers is 1.0m. The experiment conditions are summarized in Table 1.

3.2. Results

First compare same environments acoustic models that clean model, reverberation and known noise matched model, reverberation and exterior noise matched model, Delay-and-Sum (DS) matched model and SSA matched model on the basis of word accuracy scores. Figure 5 shows all acoustic models and the SSA in word accuracy score results. The SSA-matched acoustic model shows a higher recognition performance than the other models which include no (or less) considerations of residual noise and distortion effects. From the results, it is speculated that the word accuracy depends on quantity of characteristics to be considered when we really use SSA for speech recognition processing, and our proposed acoustic model with SSA is well matched to it. The difference by a subtraction parameter of SSA was not seen very much. Figure 6 shows mismatched environments model and the SSA in word accuracy score results. An SSA-matched model led the best result. In addition, the SSA-matched acoustic model is still a good model for recognition even if acoustical environments change. The difference by a subtraction parameter of SSA was not seen very much.

4. Conclusion

In this paper, to address the acoustic model problem in nonlinear array signal processing, SSA, we proposed to construct SSA-matched acoustic model. We showed the experimental evaluation of our model, and revealed an effectiveness of SSA-matched acoustic model. SSA-matched acoustic model provided a higher word accuracy score than the other conventional models. We also show the robustness against difference among environment changes.

Acknowledgments

Part of this work is supported by MEXT (Ministry of Education, Culture, Sports, Science and Technology) e-Society leading project.

References