Noisy speech recognition using blind spatial subtraction array technique and deep bottleneck features

Norihide Kitaoka* Tomoki Hayashi* Kazuya Takeda*
*Nagoya University
E-mail: kitaoka@nagoya-u.jp Tel: +81-52-789-3626

Abstract—In this study, we investigate the effect of blind spatial subtraction arrays (BSSA) on speech recognition systems by comparing the performance of a method using Mel-Frequency Cepstral Coefficients (MFCCs) with a method using Deep Bottleneck Features (DBNF) based on Deep Neural Networks (DNN). Performance is evaluated under various conditions, including noisy, in-vehicle conditions. Although performance of the DBNF-based system was much more degraded by noise than the MFCC-based system, BSSA improved the performance of both methods greatly, especially when matched condition training of acoustic models was employed. These results show the effectiveness of BSSA for speech recognition.

I. INTRODUCTION

Speech recognition systems are now widely used in mobile phones, and speech interfaces have enabled drivers to operate complicated information technology devices, such as car navigation systems, hands-free telephones, and entertainment systems while driving. For these types of in-vehicle speech recognition applications, noise robustness is essential.

To reduce noise and enhance speech, various microphone array-based techniques have been proposed. Blind spatial subtraction arrays (BSSA) proposed in [1] have proven to be effective, and speech enhanced by BSSA is very easily understood by human listeners. Furthermore, BSSA’s computational cost is reasonable, thus it can even be used in mobile equipment. In the study cited, however, no testing was done to determine or not the enhanced speech was suitable for speech recognition.

Our aim in this research is to improve speech recognition performance for in-vehicle application. In this paper, we investigate the effect of BSSA on speech recognition performance. Even though the clarity of the output speech is higher than the original noisy speech for human perception, the enhanced speech is still heavily distorted. To test the generalization abilities of different feature extraction methods (in this case, robustness to distortion), we use not only conventional MFCCs but also DBNFs extracted using a DNN.

This paper consists of seven sections. In Sections II and III, we briefly explain BSSA and DBNF, respectively. We mention matched condition training in Section IV. Test data used in this research is described in Section V and experimental results are shown in Section VI. We conclude this paper in Section VII.

II. BLIND SPATIAL SUBTRACTION ARRAY

Independent component analysis (ICA) [2], [3] is often used to separate sound sources. ICA can estimate diffuse noise very well when the other sounds are directional. Making use of this characteristic, Blind Spatial Subtraction Arrays (BSSA) [1] estimate diffuse noise and subtract it from noisy speech in order to enhance it. The BSSA procedure is described below (see also Fig. 1):

1) Speech enhancement in primary path

Multichannel sound signals obtained by microphone array \( x(t) = (x_1(t), \ldots, x_M(t))^T \) are transformed into \( x(f, \tau) \) using a discrete Fourier transform, and filtered with delay-and-sum filter \( g_{DS}(f) \) to obtain primary path output \( y_{DS}(f, \tau) \) [4]:

\[
y_{DS}(f, \tau) = g_{DS}(f)^T x(f, \tau),
\]

\[
g_{DS}(f) = (g^{(DS)}_1, \ldots, g^{(DS)}_M)^T,
\]

\[
g^{(DS)}_m(f) = \frac{1}{M} \exp \left( -\frac{2\pi i f \sin \theta_u}{N} \right),
\]

where \( \theta_u \) indicates the direction of the target sound.

2) Noise estimation in reference path

Using ICA, the observed signal is separated by applying matrix \( W_{ICA}(f) \) to obtain a \( k \)-dimensional vector containing separated signal \( o(f, \tau) \):

\[
o(f, \tau) = W_{ICA}(f)x(f, \tau).
\]

We then remove estimated target signal \( o_U(f, \tau) \) and estimate noise signal \( q(f, \tau) \):

\[
q(f, \tau) = (o_1(f, \tau), \ldots, o_{U-1}(f, \tau), 0, o_{U+1}(f, \tau), \ldots, o_K(f, \tau))^T.
\]

We apply the “projection back” method to remove amplitude ambiguity [5]. Finally, the delay-and-sum method is applied to the signal to obtain estimated noise signal \( z(f, \tau) \):

\[
z(f, \tau) = g_{DS}^T(f)q(f, \tau).
\]

3) Noise reduction

The estimated noise signal obtained from the reference path is subtracted from the enhanced target speech signal obtained from the primary path in the spectral domain, according to the spectral subtraction method proposed in
### III. BOTTLENECK FEATURE EXTRACTION USING DNN

Deep Learning [8] has attracted much attention in the field of pattern recognition research, and Deep Neural Networks (DNN) which have been trained using deep learning have achieved surprising performance. Speech recognition performance is also drastically improved by using DNNs [9] and thus many researchers are exploring the application of DNN to speech recognition.

Roughly speaking, there are two ways DNN can be applied to improve speech recognition: one is to calculate HMM (Hidden Markov Model) state output probabilities directly; the other is to extract speech features. We use DNN for the latter application, and extract features known as Deep Bottleneck Features (DBNF).

#### A. Structure of DNN

The structure of the DNN in the training phase is illustrated in Fig. 2(a). Input into the DNN consists of 12 dimensional MFCCs, log power, and their $\Delta$ and $\Delta \Delta$. Five preceding and five succeeding frame vectors are also input, thus the total number of input nodes of the DNN is 429 ($=39 \times 11$). Each dimension of the input is normalized, with the mean becoming zero and the standard deviation becoming one. Each output node corresponds to a Japanese phoneme. The activation function used for hidden layers is a sigmoid function, and that for the output layer is a soft-max function. After applying pre-training and fine-tuning, two layers are removed as shown in Fig. 2(b), and the network is used as a feature extractor. (The output nodes do not use the sigmoid function when extracting features.)

Other settings of the DNN structure are shown in Table II.
TABLE II
PARAMETERS OF DNN STRUCTURE

<table>
<thead>
<tr>
<th></th>
<th>Training phase</th>
<th>Test phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layers</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Elements in each hidden layer</td>
<td>2048</td>
<td></td>
</tr>
<tr>
<td>Dim. of extracted feature vector</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Elements in output layer</td>
<td>42</td>
<td></td>
</tr>
</tbody>
</table>

TABLE III
PARAMETERS FOR DNN TRAINING

<table>
<thead>
<tr>
<th></th>
<th>Pre-training</th>
<th>Fine-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB-RBM</td>
<td>0.0025</td>
<td>0.005</td>
</tr>
<tr>
<td>BB-RBM</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.00001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Weight cost</td>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>Momentum</td>
<td>256</td>
<td>50</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>10</td>
<td>50</td>
</tr>
</tbody>
</table>

B. Training of DNN

Training the DNN consists of two stages: pre-training and fine-tuning.

During the pre-training, we used a Gaussian-Bernoulli-Restricted Boltzmann Machine (GB-RBM) for the first layer and Bernoulli-Bernoulli-RBMs (BB-RBMs) for the other layers. The output layer is initialized randomly without pre-training.

We used frame-wise correspondences between speech data and phonemes to perform back-propagation as fine-tuning. Because setting the appropriate learning rate is very critical for fine-tuning, after each training iteration, we checked frame-wise discrimination performance on a developmental data set. We retained the learning rate when performance improved, but when performance declined we discarded the result of the iteration and halved the learning rate.

The other parameters for DNN training are shown in Table III.

IV. MATCHED CONDITION TRAINING

Matched condition training occurs when HMM acoustic models are trained using speech recorded under the same conditions as during the test phase [10], and this has shown to be effective. Although BSSA is effective at reducing noise, it may result in adding other types of distortion to speech, such as musical noise. To address these other kinds of distortion, we adopted matched condition training. Transmission channel effect and additive noise are applied to training data, and this noisy data is then processed using BSSA. This processed speech is used for training.

V. TEST DATA

A. Creating test data

To simulate utterances made inside a vehicle, microphones were place on the steering column cover, arranged as shown in Table I. A dummy head was positioned on the driver’s seat and TSP signals were emitted from its mouth, in order to measure the transmission characteristics from the mouth to the microphones. (see Fig. 3). In-vehicle noises were recorded separately while the car was being driven on city roads, using the same microphones.

Test data consisted of 200 utterances (from 23 males and 23 females) selected from the JNAS Japanese newspaper read speech corpus [11]. We prepared five test sets: (1) the “dry” source data itself (clean); (2) data convolved with transmission characteristics (IR); (3) data to which in-vehicle noises were added after being convolved with transmission characteristics (IR+noise); (4) data obtained by processing the data in (3) using BSSA (IR+noise+BSSA); and (5) data obtained by processing the data in (4) and also the HMMs used in the recognition were trained using the training data processed in the same manner as the data in (4) (IR+noise+BSSA+training).

B. Effect of BSSA

Fig. 4 shows examples of simulated utterances processed as described in (1), (3) and (4) in the previous section. The figure clearly shows the effect of BSSA when used for noise suppression; the signal-to-noise ratio is clearly improved, but distortion, in the form of musical noise generated by spectral subtraction, remains.

VI. SPEECH RECOGNITION EXPERIMENTS

A. Experimental conditions

We performed large vocabulary continuous speech recognition experiments to compare speech recognition performance. Table IV shows the experimental conditions. We used 12 dimensional MFCCs and log power, and their \( \Delta \), \( \Delta \Delta \) (a total 39 dimensions) and compared them with DBNFs. The DBNFs had 40 dimensions as shown in Table II. All of the HMM acoustic models, except the one formulated under condition (5) as described in Section V-A, were trained using features extracted from clean data.
Fig. 4. Examples of test data (from the top: with transmission characteristics; +noise; +BSSA)

**B. Results**

Table V shows the speech recognition results in regards to correct rate (Corr.) and accuracy (Acc.).

Recognition performance of about 95% was obtained for clean speech (“Clean” in Table V), but transmission channel characteristics (+IR) degraded performance. Additionally, additive noises (+noise) severely reduced recognition rates, to the point that the recognizer could no longer recognize the noisy speech. The degradation when using DBNF was much larger than when using MFCC, proving DBNF’s lack of robustness for transmission characteristics and additive car noise. The discriminative nature of DBNF training might be a reason for the large amount of degradation which occurred under mismatched conditions.

Recognition performance for the speech enhanced using BSSA (+BSSA) improved sharply when using both MFCC and DBNF. The improvement for DBNF was larger than for MFCC, leading us to conclude that DBNF is more robust to the musical noise generated by BSSA than MFCC. The results using HMMs trained with BSSA-processed speech were also much improved. The result by DBNF was inferior to MFCC, which should be further investigated.

**VII. Conclusion**

In this paper we introduced an efficient and effective noise reduction method using a blind spatial subtraction array (BSSA) as a preprocessing technique for speech recognition. We combined this BSSA preprocessing with both a conventional MFCC extractor and with a deep neural network-based feature extractor. These combinations both worked well under noisy, in-vehicle conditions when speech was processed using a BSSA, and especially well when combined with matched condition training. We plan to do further testing real, in-vehicle utterances recorded in a running vehicle in the near future.

**ACKNOWLEDGEMENTS**

We would like to thank Prof. Kiyohiro Shikano, Prof. Hiroshi Saruwatari, and Dr. Ryoichi Miyazaki for providing
us their BSSA software. This work was partially supported by the Core Research of Evolutional Science and Technology (CREST) Program of the Japan Science and Technology Agency (JST).

REFERENCES


