1aSCb4. Generative approach for robust acoustic model training for blindly separated speech recognition

Yuto Dekiura, Norihide Kitaoka* and Kazuya Takeda

*Corresponding author's address: Nagoya University, Nagoya, 464-8603, Aichi, Japan, kitaoka@nagoya-u.jp

We proposed a generative approach for acoustic model training to recognize blindly separated speech using dodecahedral microphone array and frequency domain ICA. Once we measured the transfer functions between various sound source positions and microphones in the microphone array, we can generate the enough amount of speech data with distortions and residual noises by convoluting the functions to the speech data and applying the separation. Then we can make acoustic models matched to the test condition without recording huge amount of speech data using the microphone array. We evaluated the matched acoustic models by blindly separated speech recognition experiments and showed the performance improvement.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Speech recognition has become a popular technology and is often used in mobile systems such as mobile phones and car navigation systems. Speech recognizers can achieve acceptable recognition performance in “clean” environments, but their performance is severely degraded by background noise. One application which is often demanded is speech recognition of speech uttered simultaneously by multiple speakers. Speech of additional speakers is one of the most difficult noises for speech recognition systems to handle. In such cases, sound source separation is one of the possible solutions.

Blind source separation techniques such as ICA [1] and time-frequency masking [2], which employ multiple microphones, are often used for separating speech from multiple speakers. In a previous study, we proposed a dodecahedral microphone array and processing based on frequency-domain ICA, and achieved good speech separation performance [3]. The resulting speech was clear enough for human audition, but the speech recognizer could not interpret it with sufficient accuracy. Distortion and residual noises, which sound minor to human listeners, can severely reduce speech recognition performance.

To address this problem, the use of acoustic models matched to such noisy and distorted speech has proved to be effective. To make matched acoustic models, adaptation techniques are often used [4] [5]. The distortion and noises, however, result in highly non-linear transformation of the speech features, and the results vary widely, so this adaptation alone is not sufficient to achieve acceptable recognition performance. Acoustic model training using speech under matched conditions is more effective, but it requires a large amount of training data.

However, we can produce a sufficient amount of speech data by simulating sound source separation using transfer functions recorded a priori. This simulation-based speech generation enables us to train statistical acoustic models without recording large amounts of real human speech using a device.

In this paper, we propose a generative acoustic model training method, and evaluate the method with a series of speech recognition experiments. We also suggest a possible technique for recognizing partially overlapping speech uttered by multiple speakers in our conclusion.

BLIND SPEECH SEPARATION USING A DODECAHEDRAL MICROPHONE ARRAY

Figure 1 shows the dodecahedral microphone array (DHMA) device we have developed [3]. Its diameter is 8 cm and the interval between adjacent faces is 36 degrees. Microphones are installed on ten faces, excluding the top and bottom faces. We installed 6 microphones around the center of each face. The distance between the centers of the microphones on the same face is 7 mm.

The signals observed at each face have different acoustic features such as sound pressure level, arrival time, influence of diffraction waves, and so on. Therefore, we utilize these features to group the frequency components of the separated signals that are obtained, using frequency-domain ICA. We mainly use amplitude similarity for low frequency bins and phase similarity for high frequency bins. A block diagram of the speech separation procedure is shown in Figure 2.

![Distance between microphones: 7 mm Diameter: 8 cm](image)

**FIGURE 1.** Developed dodecahedral microphone array (DHMA). Ten faces, except the top and bottom, are available to install microphones, and the maximum number of microphones is 160. Six microphones are installed around the center of each face.
Our original blind source separation method worked well, and the separated speech had high clarity for human listeners. However, the speech contained residual noises and distortions that severely degraded speech recognition accuracy. The noises were non-stationary and the distortion transformed the speech non-linearly. Thus, reconstruction of the clean speech was very difficult. One solution for speech recognition in such situations is to use matched acoustic models. If we train or adapt the statistical acoustic models using speech data distorted by the separation (that is, matched speech data), the models can include variations caused by the distortion in the models and become robust to this condition.

In our case, we can measure the transfer functions from various positions to the microphones a priori, and generate simulated recordings of speech recorded by our device, using “dry” speech sources convolved with transfer functions. [You may need to explain “dry” speech sources – I had trouble finding the meaning – ed.]

Figure 3 shows the setup of the microphone and the sound sources. Twelve sound sources were located on a horizontal plane with the DHMA at a height of 130 cm from the floor. All sound sources were positioned at equal intervals. We measured the transfer functions between each microphone and the sound sources. Then we randomly selected two sound source positions and convolved the transfer functions with the dry sources to make mixed speech at the microphones. Other conditions for blind source separation using the DHMA and frequency-domain ICA are described in Table 1.

### TABLE 1. Experimental conditions for blind source separation

<table>
<thead>
<tr>
<th>Condition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frequency</td>
<td>40kHz</td>
</tr>
<tr>
<td>Frame length</td>
<td>1024 pt (64 msec)</td>
</tr>
<tr>
<td>Frame shift</td>
<td>256 pt (16 msec)</td>
</tr>
<tr>
<td>Window function</td>
<td>Hanning window</td>
</tr>
<tr>
<td>Number of FFT points</td>
<td>1024 pt</td>
</tr>
</tbody>
</table>
Speech Recognition Experiments

Experimental Conditions

We used HMM (Hidden Markov Model) acoustic models trained using JNAS [6], a Japanese read speech database consisting of 37,804 utterances made by 130 male and 130 female speakers, as a baseline. We used left-to-right HMMs with 3 states with emission probability density functions per model. Each emission probability function had 32 Gaussians. Feature vectors for recognition consisted of 12 dimensional MFCCs ($C_1 - C_{12}$) with power, their delta and delta delta. We tested MLLR (Maximum Likelihood Linear Regression) adaptation, MAP (Maximum A posteriori) adaptation, and ML (Maximum Likelihood) training using generated speech. In the case of adaptation, we used 1,000 or 5,000 generated utterances as adaptation data. We used 36,179 or 73,811 generated utterances for matched training.

We used a Julius decoder [7] for speech recognition along with a trigram language model consisting of approx. 60,000 words trained with 75 months of newspaper texts.

Test data consisted of 200 utterances spoken by 23 male and 23 female speakers. We selected two utterances randomly from the 200 utterances and convolved them with transfer functions. We then used them for two source positions selected randomly from the twelve possible positions, and added them to make mixed speech. Finally we separated them to make a test set.

Results

A comparison of clean models, models made using adaptation methods, and models made using generative training is shown in Table 2. Clean models recognized clean utterances with high accuracy, but the recognition rate of separated utterances was very low. We could not improve recognition performance with MLLR adaptation because MLLR adaptation assumes that distortion can be compensated for by using linear transformations of the model parameters. The distortion caused by the separation is highly non-linear, so this assumption is not true. MAP adaptation was able to improve the recognition rate to some extent. Of the approaches tested, matched training obtained the best performance, and our generative approach enabled us to achieve this result. From these results, we can see that generative training allows us to make robust statistical acoustic models because sufficient amounts of training data can be generated without the high cost of recording real human utterances.
TABLE 2. Recognition results of separated speech. Clean models, adapted models, and trained models are compared. #utterances indicates the number of utterances used for adaptation/training [%].

<table>
<thead>
<tr>
<th>Acoustic Models (# of utterances)</th>
<th>Clean models (1,000)</th>
<th>MLLR (3,000)</th>
<th>MAP (1,000)</th>
<th>MAP (5,000)</th>
<th>Matched training (36,179)</th>
<th>Matched training (73,811)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean utterances</td>
<td>Correct</td>
<td>95.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>93.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Separated utterances</td>
<td>Correct</td>
<td>36.7</td>
<td>23.3</td>
<td>22.3</td>
<td>42.8</td>
<td>46.5</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>-44.4</td>
<td>-32.8</td>
<td>-31.1</td>
<td>-19.7</td>
<td>-23.0</td>
</tr>
</tbody>
</table>

CONCLUSION AND FUTURE WORK

We proposed a generative approach to acoustic model training to recognize blindly separated speech using a dodecahedral microphone array and frequency domain ICA. Once we measured the transfer functions between various sound source positions and the microphones in the array, we could generate a sufficient amount of speech data with distortion and residual noise by convolving the functions to the speech data and applying separation. We could then make acoustic models matched to the test conditions without recording huge amounts of speech data using the microphone array. We evaluated the matched acoustic models with blindly separated speech recognition experiments, and achieved a substantial improvement in performance.

In the real world, when recording dialogs (for example, meetings), utterances often partially overlap each other. This means that only the parts of the utterances without overlaps can be recorded by the microphone cleanly. In Table 2, we saw that acoustic models matched to the separated speech obtained lower recognition result of clean speech than the clean models. This result implies that clean speech and overlapping speech should be recognized by clean models and matched models, respectively. To achieve this, we will need to develop a way to detect overlapping speech segments. We would then separate only the overlapping sections, concatenate the separated sections with the clean sections (to which we do not apply separation), and recognize the concatenated speech using acoustic model switching during decoding. Multipath HMM [8] is one possible method to achieve this.

ACKNOWLEDGMENTS

This work was partially supported by the Core Research for Evolutional Science and Technology (CREST) Program of the Japan Science and Technology Agency (JST).

REFERENCES