STATISTICAL SOUND SOURCE IDENTIFICATION IN A REAL ACOUSTIC ENVIRONMENT FOR ROBUST SPEECH RECOGNITION USING A MICROPHONE ARRAY

Takanobu Nishiura†, Satoshi Nakamura†, and Kiyohiro Shikano‡
† ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan
‡ Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

ABSTRACT
It is very important for a hands-free speech interface to capture distant talking speech with high quality. A microphone array is an ideal candidate for this purpose. However, this approach requires localizing the target talker. Conventional talker localization methods in multiple sound source environments not only have difficulty localizing the multiple sound sources accurately, but also have difficulty localizing the target talker among known multiple sound source positions. To cope with these problems, we propose a new talker localization method consisting of two algorithms. One algorithm is for multiple sound source localization based on CSP (Cross-power Spectrum Phase) analysis. The other algorithm is for sound source identification among localized multiple sound sources towards talker localization. In this paper, we particularly focus on sound source identification among localized multiple sound sources with statistical speech and environmental sound models based on GMMs (Gaussian Mixture Models) and a microphone array towards talker localization.

1. INTRODUCTION
For teleconference systems or voice control systems, the high-quality sound capture of distant talking speech is very important. However, background noise and room reverberation seriously degrade the sound capture quality in real acoustical environments. A microphone array is an ideal candidate for capturing distant talking speech. With a microphone array, the desired speech signals can be acquired selectively by steering the microphone array in the desired speech direction sensitively [1].

Accordingly, the microphone array is often used for the front-end processing of ASR (Automatic Speech Recognition) at present [2, 3]. However, to achieve the high-quality sound capture of distant talking speech, talker localization is necessary to steer the microphone array. Conventional talker localization methods in multiple sound source environments not only have difficulty localizing the multiple sound sources accurately, but also have difficulty localizing the target talker among known sound source positions.

To cope with these problems, we propose a new talker localization method consisting of two algorithms. One algorithm is for multiple sound source localization. The other algorithm is for statistical sound source identification among localized multiple sound sources towards talker localization. We have already proposed a multiple sound source localization algorithm with the CSP (Cross-Power Spectrum Phase) coefficient addition method [4]. In this paper, we particularly focus on sound source identification among localized multiple sound sources with statistical speech and environmental sound models based on GMMs (Gaussian Mixture Models) and a microphone array towards talker localization.

2. TALKER LOCALIZATION ALGORITHM
As shown in Figure 1, we assume that the desired speech comes from the front direction and undesired noise comes from the right direction. In this situation, talker localization is necessary for effectively capturing and accurately recognizing distant talking speech using a microphone array. However, conventional talker localization methods in multiple sound source environments not only have difficulty localizing multiple sound sources accurately, but also have difficulty localizing the desired talker among localized sound source positions.

Accordingly, we propose a new talker localization method as shown in Figure 2. First, multiple sound sources are localized with the CSP coefficient addition method after multiple sound signals are captured. Then, these localized sound signals are enhanced by steering the microphone array to them. Finally, the talker can be localized after identification between "speech" or "non-speech" using statistical speech and environmental sound models among the localized and enhanced multiple sound signals. The system recognizes the input from a sound source identified as being "speech".

We have already proposed a multiple sound source localization algorithm with the CSP coefficient addition method [4]. In this paper, we particularly focus on sound source identification us-
Numerous sound sources are necessary to design the speech and environmental sound models and microphone array steering among localized multiple sound sources towards talker localization. If the talker can be localized, then his/her speech can be recognized.

2.1. Microphone array steering for speech enhancement

Microphone array steering is necessary to capture distant signals effectively. In this paper, a delay-and-sum beamformer [1] is used to steer the microphone array to the desired sound direction. Localized sound signals are enhanced by the microphone array steering because the delay-and-sum beamformer can form directivity to the localized sound sources.

2.2. Statistical sound source identification based on GMMs

Multiple sound signals are captured effectively and enhanced by microphone array steering. Therefore, the talker can be localized by identifying the enhanced multiple sound signals. Until now, a speech model alone was usually used for speech / non-speech segmentation [5] or identification. However, a single speech model has problems in that it not only requires a threshold to identify between “speech” and “non-speech”, but also degrades the identification performance in noisy reverberant environments. To overcome these problems, we propose a new speech / non-speech identification algorithm that uses statistical speech and environmental sound GMMs. The multiple sound signals enhanced with the microphone array steering are identified by Equation (1).

\[ \hat{\lambda} = \arg \max_{\lambda} P(S(w)|\lambda_s, \lambda_n), \]

where \( S(w) \) is the enhanced signal with the microphone array steering (frequency domain), \( \lambda_s \) represents the statistical speech model, and \( \lambda_n \) represents the statistical environmental sound model. The enhanced signals are identified as “speech” or “non-speech” by estimating the maximum likelihood in Equation (1). This algorithm allows the talker to be localized among localized multiple sound sources.

2.2.1. Speech and environmental sound database

Numerous sound sources are necessary to design the speech and environmental sound GMMs. Therefore, we use the ATR speech Database (ATR-DB) [6] to design the speech model and the RWCP (Real World Computing Partnership) sound scene database (RWCP-DB) [7, 8] which includes various environmental sounds to design the non-speech model. RWCP-DB also includes numerous impulse responses measured in various acoustical environments. These impulse responses are used to conduct evaluation experiments in various acoustical environments.

3. EVALUATION EXPERIMENTS

We first evaluate the basic performance of statistical sound modeling though environmental sound recognition. Then, the sound source identification performance is evaluated in distant-talking robust speech recognition.

3.1. Preliminary experiment: environmental sound recognition

Environmental sound recognition experiments are carried out with 20 samples for 92 kinds of environmental sounds and a single transducer in a clean environment. The feature vectors are MFCC, \( \Delta \)MFCC, and \( \Delta^2 \)power. As a result, the environmental sound recognition performance is an average rate of 88.7% in multiple occurrence environments of the same sounds. This result confirms that the statistical modeling is very effective not only for speech recognition but also for environmental sound recognition.

3.2. Experimental conditions

The sound source identification performance is evaluated with known multiple sound source positions. Figure 3 shows the experimental environment. The desired signal comes from the front direction and white Gaussian noise comes from the right direction. The distance between the sound source and the microphone array is two meters. In this situation, the statistical sound source identification performance and ASR performance are evaluated subject to variations in the SNR (Signal to Noise Ratio) and the environment.

Table 1 shows data recording conditions and Table 2 shows experimental conditions for statistical sound source identification. We evaluate the statistical sound source identification performance using a single transducer and a microphone array, subject to SNR of -5dB, -30dB, and clean, and the reverberation times are \( T_{60} \) of 0.0, 0.3, and 1.3 sec. We also evaluate the ASR performance with the experimental conditions for ASR which are shown in Table 2. In this paper, we evaluate the statistical sound source identification
Table 2: Experimental conditions for sound source identification

<table>
<thead>
<tr>
<th>Feature</th>
<th>Speech DB</th>
<th>Non-speech DB</th>
<th>Speech model training</th>
<th>Non-speech model training</th>
<th>Test data (Open)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame length</td>
<td>32 msec.</td>
<td>8 msec.</td>
<td>150 words × 16 subjects (8 female and 8 male)</td>
<td>92 sounds × 20 sets</td>
<td>Speech: 216 words × 2 subjects (1 female and 1 male)</td>
</tr>
<tr>
<td>Frame interval</td>
<td>8 msec.</td>
<td></td>
<td>150 words × 16 subjects (8 female and 8 male)</td>
<td>92 sounds × 20 sets</td>
<td>Speech: 216 words × 2 subjects (1 female and 1 male)</td>
</tr>
<tr>
<td>Feature vector</td>
<td>MFCC (16 orders, 4 mixtures), ∆MFCC (16 orders, 4 mixtures), ∆power (1 order, 2 mixtures)</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
</tr>
<tr>
<td>Number of models</td>
<td>Speech: 1 model</td>
<td>Non-speech: 1 model</td>
<td>Speech: 1 model</td>
<td>Non-speech: 1 model</td>
<td>Speech: 1 model</td>
</tr>
</tbody>
</table>

Table 3: Experimental conditions for ASR

<table>
<thead>
<tr>
<th>Feature</th>
<th>Speech DB</th>
<th>Non-speech DB</th>
<th>Speech model training</th>
<th>Non-speech model training</th>
<th>Test data (open)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame length</td>
<td>25 msec.</td>
<td>10 msec.</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>216 words × 2 subjects(1 female and 1 male)</td>
</tr>
<tr>
<td>Frame interval</td>
<td>25 msec.</td>
<td>10 msec.</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>216 words × 2 subjects(1 female and 1 male)</td>
</tr>
<tr>
<td>Feature vector</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
<td>MFCC, ∆MFCC, ∆power (+ CMS [9])</td>
</tr>
<tr>
<td>Test data (open)</td>
<td>Speech: 216 words × 2 subjects (1 female and 1 male)</td>
<td>Non-speech: 216 words × 2 sets (1 female and 1 male)</td>
<td>Speech: 216 words × 2 subjects (1 female and 1 male)</td>
<td>Non-speech: 216 words × 2 sets (1 female and 1 male)</td>
<td>Speech: 216 words × 2 subjects (1 female and 1 male)</td>
</tr>
</tbody>
</table>

3.3. Experimental results

Figure 4(a)(b) show experimental results using a single transducer and a microphone array that steers the directivity to the known desired sound source position. In these figures, the bar graphs represent sound source identification rates (SIR), and the line graphs represent word recognition rates (WRR).

First, we focus the bar graphs in Figure 4(a)(b). In these figures, by comparing the results using the single transducer and using the microphone array steering, we can confirm that the microphone array steering results give a higher sound source identification performance than the single transducer results especially in lower SNR environments. We therefore confirm that the proposed algorithm can achieve a higher sound source identification performance by using the microphone array steering.

Second, we describe the robustness against reverberation on the source sound identification. In Figure 4(b), the source sound identification performance using the microphone array steering is almost the same in each reverberant environment while the performance tends to decline slightly in the lower SNR and higher reverberant environments. With these results, we confirm that the proposed algorithm can distinguish "speech" or "non-speech" accurately in higher reverberant environments.

Next, we compare the proposed method with a conventional method using only speech GMM. Statistical sound source identification was carried out with the conventional method by distinguishing "speech" or "non-speech" using a threshold. Figure 5 shows the threshold estimation. As shown in the figure, we calculate accumulated likelihood histograms with training data. Then, we estimate the threshold for the conventional method by finding the equal probability point with the accumulated likelihood histograms. Figure 6 shows results of the proposed method and the conventional method with microphone array steering in $T_{160} = 0.3$sec environments. The sound source identification rate (SIR) is only about 70% with the conventional method in the higher SNR environments, although the identification performance improves where the SNR is higher. However, the sound source identification rate is more than 90% with the proposed method not only in the higher SNR environments but also in the lower SNR environments. The performance of the conventional method using only speech GMM depends a lot on the threshold. However, the proposed method using speech and environmental sound GMMs can distinguish "speech" or "non-speech" accurately because it uses the difference of two GMM’s likelihoods.

We also evaluate the relationship of the number of Gaussian mixtures for feature vectors MFCC and ∆MFCC and the sound source identification rate. Figure 7 shows the results. In the figure, we confirm that the sound source identification performance is almost the same with more than four mixtures, while the performance degrades with less than two mixtures. Therefore, the proposed method may be able to distinguish "speech" or "non-speech" even if speech and environmental sound GMMs consist of few mixtures.
Finally, we focus on the line graphs showing the word recognition rates (WRR) in Figure 4(a)(b). In these figures, by comparing the results using the single transducer and using the microphone array steering, we can confirm that the microphone array steering results give a higher ASR performance especially in lower SNR environments than the single transducer results. As an example, we explain the ASR performance in the SNR = 10 dB environment. In the $T_{60} = 0.0$ sec. and SNR = 10 dB environment, WRR is 58.3% for the single transducer. However, WRR improves from 58.3% to 92.6% using the microphone array. In addition, in the $T_{60} = 1.3$ sec. and SNR = 20 dB environment, WRR is 37.1% for the single transducer. However, WRR improves from 37.1% to 64.3% using the microphone array. This confirms that the proposed algorithm using the microphone array results in a higher ASR performance than that using the single transducer not only in anechoic environments but also in reverberant environments.

According to the above evaluation experiments, we confirm that the talker can be localized accurately by sound source identification using statistical speech and environmental sound GMMs and microphone array steering among known multiple sound sources. We also confirm that the talker's speech can be recognized robustly with the microphone array in noisy reverberant environments.

### 4. CONCLUSION

In this paper, we propose a new talker localization algorithm and focus on sound source identification among localized multiple sound sources with statistical speech and environmental sound models based on GMMs towards talker localization. In evaluation experimental results, we confirm that the talker is localized accurately by the proposed algorithm with known sound source positions in reverberant noisy environments. In addition, the talker's speech is recognized robustly with the microphone array in noisy reverberant environments. In future works, we will evaluate the proposed algorithm by using multiple sound source positions localized by the CSP coefficient addition method.

### 5. REFERENCES


