TALKER LOCALIZATION IN A REAL ACOUSTIC ENVIRONMENT BASED ON DOA ESTIMATION AND STATISTICAL SOUND SOURCE IDENTIFICATION

Takanobu Nishiura†, Satoshi Nakamura†, Kiyohiro Shikano‡

† Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Japan
‡ ATR Spoken Language Translation Research Laboratories

Graduate School of Information Science, Nara Institute of Science and Technology

ABSTRACT

For a hands-free speech interface, it is very important 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 algorithms 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 algorithm consisting of two algorithms. One is DOA (Direction Of Arrival) estimation algorithm for multiple sound source localization based on CSP (Cross-power Spectrum Phase) coefficient addition method. The other is statistical sound source identification algorithm based on GMM (Gaussian Mixture Model) for localizing the target talker position among localized multiple sound sources. In this paper, we particularly focus on the talker localization performance based on the combination of these two algorithms with a microphone array.

1. INTRODUCTION

The high-quality sound capture of distant talking speech is very important for teleconference systems or voice control systems. 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]. However, to achieve the high-quality sound capture of distant talking speech, talker localization is necessary to steer the microphone array. Conventional talker localization algorithms 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 algorithm consisting of two algorithms. One is DOA (Direction Of Arrival) estimation algorithm for multiple sound source localization based on CSP (Cross-power Spectrum Phase) analysis. The other is statistical sound source identification algorithm for localizing the target talker position among localized multiple sound sources. We have already proposed DOA estimation algorithm for localizing multiple sound sources with the CSP coefficient addition method [3] and statistical sound source identification algorithm with statistical speech and environmental sound models based on GMMs (Gaussian Mixture Models) and a microphone array [4].

In this paper, we particularly focus on the talker localization performance based on the combination of DOA estimation and statistical sound source identification algorithms with a microphone array. If the talker can be localized, then his/her speech can be recognized.

2. TALKER LOCALIZATION ALGORITHM

As shown in Figure 1, we assume that DOA1 as the desired speech comes from the right direction and DOA2 as undesired noise (non-speech) comes from the left direction. In this situation, talker localization is necessary for effectively capturing and accurately recognizing distant talking speech using a microphone array.

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

2.1. DOA estimation with CSP coefficient addition method

DOA must be estimated to automatically steer the microphone array. CSP (Cross-power Spectrum Phase) analysis is very popular method for estimating DOA. However, multiple DOA estimation with CSP analysis is very difficult because of the cross-
correlation of multiple sound signals. To overcome this problem, we proposed the CSP coefficient addition method to estimate multiple DOAs at ICASSP2000 [3]. In the environment of Figure 1, the CSP coefficients are derived from Equation (1).

\[ \text{CSF}_{m,n}(k) = \text{IDFT} \left[ \frac{\text{DFT} \left[ s_{m}(t) \right] \text{DFT} \left[ s_{n}(t) \right]^*}{|\text{DFT} \left[ s_{m}(t) \right]||\text{DFT} \left[ s_{n}(t) \right]|} \right], \quad (1) \]

where \( t \) and \( k \) are the time index, \( \text{DFT} \) (or \( \text{IDFT} \)) is the discrete Fourier transform (or the inverse discrete Fourier transform), and the symbol \(^*\) is the complex conjugate. Then, CSP coefficients are added as shown in Equation (2).

\[ \text{CSP}_{\theta}(\theta) = \sum_{n=1}^{N} \text{CSP}_{m,n}(\theta), \quad \text{subject to} \quad \theta = \cos^{-1} \left( \frac{c \cdot k / F_s}{d_0} \right), \quad (2) \]

where \( N \) is the number of additions, \( d_0 \) is the distance between two adjacent transducers, \( c \) is the sound propagation speed, and \( F_s \) is the sampling frequency. The DOA can be accurately estimated by finding the maximum values of the added CSP coefficients by Equation (3).

\[ \text{DOA}_{\theta} = \arg \max_{\theta} (\text{CSP}_{\theta}(\theta)). \quad (3) \]

2.2. 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. Multiple sound signals of estimated DOA are enhanced by the microphone array steering because the delay-and-sum beamformer can form directivity to the estimated DOAs.

2.3. Speech / non-speech 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 proposed a speech / non-speech identification algorithm that uses statistical speech and environmental sound GMMs at EUROspeech2001 [4]. The multiple sound signals enhanced with the microphone array steering are identified by Equation (4).

\[ \hat{\lambda} = \arg \max_{\lambda} P(S(w)|\lambda_s, \lambda_n), \quad (4) \]

---

**Table 1: Recording conditions**

<table>
<thead>
<tr>
<th>Microphone array</th>
<th>14 transducers, 2.83 cm spacing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frequency</td>
<td>16 kHz</td>
</tr>
<tr>
<td>Impulse response</td>
<td>RWCP-DB [7]</td>
</tr>
<tr>
<td>Reverberation time</td>
<td>0.0, 0.3, and 1.3 sec.</td>
</tr>
<tr>
<td>SNR</td>
<td>-5dB, -30dB, and clean</td>
</tr>
</tbody>
</table>

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 (4). This algorithm allows the talker to be localized among estimated DOAs.

3. EVALUATION EXPERIMENTS

We evaluate the talker localization performance and ASR performance in the noisy reverberant environments.

3.1. Experimental conditions

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 talker localization performance and ASR performance are evaluated subject to variations in the SNR (Signal to Noise Ratio) and the reverberant environment.

Table 1 shows data recording conditions and Table 2 shows experimental conditions for statistical sound source identification towards talker localization. We evaluate the talker localization performance, subject to SNR of -5dB, -30dB, and clean, and the reverberation times were \( T_{60} = 0.0, 0.3, \text{and } 1.3 \text{ sec.} \). We also evaluate the ASR performance with the experimental conditions for ASR which are shown in Table 2. In this paper, we evaluate the talker localization performance with 616 sounds consisting of speech (216 words \( \times \) 2 subjects (1 female and 1 male)) and environmental sounds (92 sounds \( \times \) 2 sets). The ASR performance is also evaluated with speech (216 words \( \times \) 2 subjects). Equation (5) shows a definition of the Sound source Identification Rate (SIR).

\[ \text{SIR} = \frac{\sum_{n=0}^{N} I_{\text{cor}}[n]}{N}, \quad I_{\text{cor}}[n] = \begin{cases} 1 & \tilde{Q}[n] = Q[n] \\ 0 & \tilde{Q}[n] \neq Q[n], \end{cases} \quad (5) \]

where \( N \) is the number of all sounds, \( Q[n] \) is the correct answer, and \( \tilde{Q}[n] \) is the sound source identification result. The ASR performance is also evaluated by the Word Recognition Rate (WRR),
transducer and using the microphone array steering, we can con­
firm that the proposed algorithm can distinguish "speech" or "non-speech" ac­
tion especially in lower SNR environments. We therefore confirm that the
sound source identification performance with estimated DOAs arrive at almost the same performance level as the correct desired sound DOAs. On the other hand, when the SNR decreases more and more, the estimated undesired noise DOAs arrive at almost the same performance level as the correct undesired noise DOA. These facts occur because the CSP coefficients calculated by the CSP coefficient addition method depend on the signal energy when two DOAs are esti­
mated. We also evaluated the DOA estimation performance in $T_{[60]} = 0.0$ sec. and 1.3 sec. reverberation time environments. As a result, we could confirm that the performance is almost the same tendency as those in $T_{[60]} = 0.3$ sec. environment.

Next, we evaluated the talker localization and ASR performance with estimated DOAs. Figure 6 shows experimental results obtained using a microphone array that was used to steer the directivities to the estimated DOAs. In this figure, the bars represent SIR, and the lines represent WRR. We focus on the bars showing the SIR in Figure 4(b) and in Figure 6. In Figure 4(b) and Figure 6, by comparing the results in included in microphone array elements. Figure 5 shows the DOA estimation results in $T_{[60]} = 0.3$ sec. reverberation time environment. Average values and standard deviations were calculated to evaluate the results. In addition, the accuracy of the estimated DOAs was calculated by Equation (6).

\[
\text{Accuracy} = \frac{1}{N} \sum_{n=0}^{N} I[n] = \begin{cases} 
1 & \|D[n] - D[n]\| \leq Er \\
0 & \|D[n] - D[n]\| > Er,
\end{cases}
\]

Figure 4: Talker localization and ASR performance with known sound source positions.

Table 2: Experimental conditions for statistical sound source identification towards talker localization

| Frame length | 32 msec. (Hamming window) |
| Frame interval | 8 msec. |
| Feature vector | MFCC (16 orders, 4 mixtures), \(\Delta\text{MFCC}\) (16 orders, 4 mixtures), \(\Delta\)power (1 order, 2 mixtures) |
| Number of models | Speech: 1 model, Non-speech: 1 model |
| Speech model training | 150 words \(\times\) 16 subjects (8 females and 8 males) |
| Non-speech DB | RWCP-DB [7] |
| Non-speech model training | 92 sounds \(\times\) 20 sets |
| Test data (Open) | Speech: 216 words \(\times\) 2 subjects (1 female and 1 male), Non-speech: 92 sounds \(\times\) 2 sets |

Table 3: Experimental conditions for ASR

| Frame length | 25 msec. (Hamming window) |
| Frame interval | 10 msec. |
| Feature vector | MFCC, \(\Delta\text{MFCC}\), \(\Delta\)power (+ CMS) |
| Test data (Open) | 216 words \(\times\) 2 subjects (1 female and 1 male) |

3.2. Talker localization results with known source positions

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

First, we focus on the bars showing the SIR 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 sound source identification algorithm can achieve a higher sound source identification performance by using the microphone array steering. We also describe the robustness against reverberation on the sound source identification. In Figure 4(b), the sound source 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 focus on the line graphs showing the 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.

3.3. Talker localization results with estimated DOAs

An evaluation experiment using a microphone array was conducted to steer the directivities to the estimated DOAs. First, the DOAs were estimated by the CSP coefficient addition method. In this evaluation experiment, the DOAs were estimated by two times addition of the CSP coefficients with 4 transducers which are in...
In this paper, we propose a new talker localization algorithm based on the combination of DOA estimation algorithm with CSP coefficient addition method and sound source identification algorithm with statistical speech and environmental sound GMMs. In evaluation experimental results, we confirm that the talker is localized accurately by the proposed algorithm in reverberant noisy environments. In future work, we will work for improving the ASR performance in lower SNR and higher reverberant environment with more sharper beamformer such as multiple beamformer [8].

Acknowledgement: This work was mainly supported by “Auditory Brain” project of CREST fund by JST.

5. REFERENCES