Near Field Sound Source Localization Based on Cross-power Spectrum Phase Analysis with Multiple Microphones

Kohei Hayashida\(^1\), Masanori Morise\(^2\), Takanobu Nishiura\(^2\)

\(^1\)Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Japan
\(^2\)College of Information and Science, Ritsumeikan University, Kusatsu, Japan

Abstract

We study sound source localization in the near field. In sound source localization research, 2D-MUSIC has already been developed. However, its performance degrades in diffused noisy environments. A localization method based on CSP in the near field has already been developed. However, localization accuracy depends on the accuracy of the estimation of the time delay between paired microphones. To overcome these problems, we have developed 2D-CSP with multiple microphones for robust localization. We carried out the evaluation experiment in a conference room. The proposed method could localize sound sources more robustly than conventional methods.

Index Terms: Microphone array, Spherical wave, Sound source localization, Cross-power spectrum phase analysis, Multiple channels

1. Introduction

A microphone array is effective at capturing distant-talking speech with high quality in noisy environments. It captures the target speech by steering the directivity on the basis of the direction of the target speaker. Techniques not only to estimate the directions of sound sources in the far field, but also to localize the positions of sound sources in the near field have been developed. Localizing sound sources in the near field is especially important for developing useful speech interfaces such as speech-controlled machines, humanoid robots, acoustic surveillance systems, etc. In near-field sound source localization research, two-dimensional multiple signal classification (2D-MUSIC)\(^1\) is an effective technique that has been developed. However, its performance degrades if it is used in diffused noisy environments, if the number of localized sources is limited to M-1 (M: the number of microphones), and so on. As a result, 2D-MUSIC needs a large-scale microphone array for robust localization in real environments. Two-dimensional cross-power spectrum phase analysis with paired microphones (multiple paired 2D-CSP)\(^2\) has also been developed for sound source localization. However, its localization accuracy depends on the accuracy of the estimation of time delay between paired microphones. To overcome these problems, we developed two-dimensional cross-power spectrum phase analysis with multiple microphones (multiple channel 2D-CSP) by expanding conventional CSP for robustly localizing sound sources in the near field. This method adds a correlation matrix to conventional CSP and utilizes phase differences at the same time. We aim to improve the sound source localization performance in the near field with this multiple-channel 2D-CSP.

2. Conventional methods for sound source localization

2.1. Conventional method based on MUSIC

2.1.1. MUSIC

Multiple signal classification (MUSIC)\(^3, 4\) is a method for estimating the direction of arrival (DOA) of sound sources in the far field. It estimates the DOAs on the basis of the orthogonality between the signal subspace and the noise subspace of the input signal based on the plane wave. It generates steering vectors and scans the whole direction of interest. The steering vector \(\mathbf{d}_k(\theta)\) is derived from Eq. (1).

\[
\mathbf{d}_k(\theta) = [e^{-j\omega_k \tau_{M_1}(\theta)}, \ldots, e^{-j\omega_k \tau_{M_N}(\theta)}]^T.
\]

The symbol \(k\) denotes the frequency index, \(\theta\) the sound source direction, \(\omega_k\) the discrete angular frequency, and \(\tau_{M_i}(\theta)\) the difference of propagation time between \(M_i\) and \(M_1\) for the sound source direction \(\theta\). On the basis of the plane wave, \(\tau_{M_i}(\theta)\) is derived from Eq. (2).

\[
\tau_{M_i}(\theta) = (i-1)d\cos(\theta)/c.
\]

The symbol \(d\) denotes the interval between microphones, and \(c\) the sound propagation velocity. Figure 1(a) shows a propagation image of a sound wave on the basis of a plane wave. On the basis of the plane wave, the spatial spectrum \(P(\theta, k)\) for MUSIC can be derived from Eq. (3).

\[
P(\theta, k) = \frac{1}{|\mathbf{d}_k(\theta)\mathbf{R}_k^{\perp}|^2}.
\]

The symbol \(\mathbf{R}_k^{\perp}\) denotes the noise subspace matrix. In the matched condition of \(\theta\) and the target source direction, the denominator of Eq. (3) goes to 0 on the basis of the orthogonality and the \(P(\theta, k)\) has a maximum peak. The averaged spatial spectrum \(\bar{P}(\theta)\) is calculated in the frequency range by Eq. (4).

\[
\bar{P}(\theta) = \frac{\sum_{k=k_L}^{k_H} P(\theta, k)/(k_H - k_L)}{k_H - k_L}.
\]

The symbols, \(k_L, k_H\), are indices for the lower and the upper bounds of the frequency range. MUSIC is a method that does not depend on the energy of the sound source.

2.1.2. 2D-MUSIC

On the basis of MUSIC, two-dimensional multiple signal classification (2D-MUSIC)\(^1\), an expanded version of conventional MUSIC, has been developed as a method for localizing sound
sources in the near field. It can localize target sound sources on
the basis of spherical waves. It generates location vectors and
scans whole spaces of interest. The location vector \( \mathbf{d}_k(S_x, S_y) \)
is derived from Eq. (5).

\[
\mathbf{d}_k(S_x, S_y) = \left[ e^{-j\omega_k \tau_{M_1}(S_x, S_y)}, \ldots, e^{-j\omega_k \tau_{M_N}(S_x, S_y)} \right]^T.
\]  
(5)

The symbol \((S_x, S_y)\) denotes the coordinates of a sound source, and \(\tau_{M_i}(S_x, S_y)\) the propagation time from the sound source \((S_x, S_y)\) to the microphone \(M_i\). On the basis of the spherical wave, \(\tau_{M_i}(S_x, S_y)\) is derived from Eq. (6).

\[
\tau_{M_i}(S_x, S_y) = \sqrt{(S_x - M_{ix})^2 + (S_y - M_{iy})^2} / c.
\]  
(6)

The symbol \((M_{ix}, M_{iy})\) denotes the coordinates of the microphone \(M_i\). Figure 1(b) shows an image of propagation of a sound wave based on a spherical wave. In Fig. 1(b) conditions, the spatial spectrum \(P(S_x, S_y, k)\) for 2D-MUSIC can be derived from Eq. (7).

\[
P(S_x, S_y, k) = \frac{1}{|\mathbf{d}_k^H(S_x, S_y) \mathbf{R}_x |^2}.
\]  
(7)

In the matched condition of \((S_x, S_y)\) and one for the target source location, the denominator of Eq. (7) goes to 0 based on the orthogonality and \(P(S_x, S_y, k)\) has a maximum peak. The averaged spatial spectrum \(\bar{P}(S_x, S_y)\) is calculated in the frequency range by Eq. (8).

\[
\bar{P}(S_x, S_y) = \frac{1}{k_H - k_L} \sum_{k=k_L}^{k_H} P(S_x, S_y, k)/(k_H - k_L).
\]  
(8)

2D-MUSIC is a method that does not depend on the energy of a sound source.

2.2. Conventional method based on CSP

2.2.1. CSP

Cross-power spectrum phase analysis (CSP)[5, 6] is a powerful DOA estimation technique that does not depend on the frequency characteristics of the sound sources. It estimates the CSP coefficients and the time delay of arrival (TDOA) on the basis of captured signals of a paired microphone. The CSP coefficients \(CSP(t)\) and the TDOA \(\hat{\tau}\) are derived from Eqs. (9) and (10).

\[
CSP(t) = DFT^{-1}\left[ \frac{DFT(x_i(t)) DFT(x_j(t))^*}{|DFT(x_i(t))||DFT(x_j(t))|} \right]
\]  
(9)

\[
\hat{\tau} = \text{argmax}(CSP(t)).
\]  
(10)

The symbols \(x_i(t)\) and \(x_j(t)\) denote the captured signals with a paired microphone, \(DFT\) the discrete Fourier transform, and \(DFT^{-1}\) the inverse discrete Fourier transform. The CSP method first calculates the discrete Fourier transform of captured signals \(x_i(t)\) and \(x_j(t)\) with a paired microphone, then calculates phase differences on the basis of an amplitude normalization, and finally acquires the CSP coefficients with an inverse discrete Fourier transform, as shown in Eq. (9). In contrast, the TDOA is acquired by utilizing the time lag on the basis of the maximum CSP coefficient, as shown in Eq. (10). The DOA is also derived from Eq. (11).

\[
\theta = \cos^{-1}\left( \frac{\tau_\ell}{d} \right).
\]  
(11)

The symbol \(d\) denotes the interval between the two microphones, and \(F_s\) the sampling frequency. The CSP has the maximum CSP coefficient value in a DOA with dominant power. The CSP method calculates the DOAs on the basis of the plane wave, as shown in Fig. 1(a).

2.2.2. 2D-CSP with multiple paired microphones

Using CSP, two-dimensional cross-power spectrum phase analysis with multiple paired microphones (multiple paired 2D-CSP)[2], an expanded version of conventional CSP has already been developed as a method for localizing sound sources in the near field. The multiple paired 2D-CSP first estimates the CSP coefficients and the TDOA \(\hat{\tau}\) derived from Eqs. (9) and (10). These are exactly the same steps as those of the CSP in Section 2.2.1. In addition, the TDOA \(\tau_i(S_x, S_y)\) from sound source \((S_x, S_y)\) to paired microphone \(l\) is derived from Eq. (12) on the basis of the spherical wave, with a sound propagation time \(\tau_{M_l}(S_x, S_y)\) derived from Eq. (6).

\[
\tau_i(S_x, S_y) = \tau_{M_l}(S_x, S_y) - \tau_{M_j}(S_x, S_y).
\]  
(12)

The symbol \(l\) denotes a paired microphone that consists of each microphone \(M_i, M_j\). The multiple paired 2D-CSP localizes a sound source utilizing the TDOA \(\hat{\tau}\) derived from Eqs. (9) and (10) and the TDOA \(\tau_{M_i}(S_x, S_y)\) derived from Eq. (12). The spatial coefficient \(P(S_x, S_y)\) for multiple paired 2D-CSP is derived from Eq. (13).

\[
P(S_x, S_y) = \sum_{i=1}^{m} (\tau_i(S_x, S_y) - \hat{\tau})^2.
\]  
(13)

The symbol \(m\) denotes the number of paired microphones. If \(\hat{\tau}\) agrees with \(\tau_i(S_x, S_y)\) in Eq. (13), \(P(S_x, S_y)\) goes to 0. When \(P(S_x, S_y)\) is at its minimum value, \((S_x, S_y)\) is estimated to be a sound source location. When only one pair of microphones is used, \(P(S_x, S_y)\) is 0 for some coordinates. Figure 2 shows the results of localizing sources with Eq. (13) on the basis of two sets of paired microphones, consisting of three microphones. The sound source is localized with the intersection of TDOAs on the basis of each paired microphone. In this paper, we assume that the sound source only appears in front of the microphone array. Therefore, only two paired microphones that each consists of three microphones are necessary for the multiple paired 2D-CSP. However, the localization accuracy depends on the estimation accuracy for the time delay between paired microphones. In other words, the localization accuracy depends on the sampling frequency and the interval between microphones.
3. 2D-CSP with multiple microphones for the proposed method

In this section, we describe the two-dimensional cross-power spectrum phase analysis with multiple microphones (multiple channel 2D-CSP) by expanding conventional CSP for robustly localizing sound sources in the near field. A correlation between the input signals is calculated by correlation matrix. An estimation method for the time delay of arrival that utilizes the correlation matrix on the basis of an amplitude normalization has already been developed [7]. However, this method assumes a plane wave and has to recalculate a correlation matrix at every scanning point. We developed a multiple channel 2D-CSP that utilizes the unique correlation matrix \( R \) under the spherical wave assumption. The correlation matrix \( R \) is calculated by Eq. (14).

\[
R = \begin{bmatrix}
R_{11} & \ldots & R_{1N} \\
\vdots & \ddots & \vdots \\
R_{N1} & \ldots & R_{NN}
\end{bmatrix},
\]

(14)

Each component of the correlation matrix is calculated on the basis of an amplitude normalization by Eq. (15).

\[
r_{ij} = \frac{\text{DFT}(x_i(t))\text{DFT}(x_j(t))^*}{|\text{DFT}(x_i(t))||\text{DFT}(x_j(t))|},
\]

(15)

where \( S \) is the location vector calculated by Eq. (5). In the matched condition of \((S_x, S_y)\) and one for the target source location, \( P(S_x, S_y, k) \) has a maximum peak. The averaged spatial spectrum is calculated in the frequency range by Eq. (8).

4. Experiment

4.1. Experimental Conditions

We carried out the evaluation experiments in a conference room. Table 1 shows the experimental conditions. Figure 3 (a) shows the placement of the sound sources and three microphones in the conference room. There were 132 different sound source positions, each located 0.2 [m] from another. We used a mouth simulator to simulate the radiation characteristics of someone speaking. Figure 3 (b) shows the placement of noise sources and the three microphones. We used two loudspeakers to create the diffused noise environment by emitting toward the floor, and the other, placed 2.4 [m] off the ground, also did so by emitting toward the ceiling. The frequency ranges for estimation were under 3 [kHz] and under 7 [kHz]. The dominant frequency of the human voice is under 3 [kHz]. Seven [kHz] is the neighborhood of Nyquist frequency for estimation.

<table>
<thead>
<tr>
<th>Environments</th>
<th>Conference room (( T_{360} = 0.4 ) [s])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of mics.</td>
<td>3 mics.</td>
</tr>
<tr>
<td>Apartment of mics.</td>
<td>0.15, 0.3 [m]</td>
</tr>
<tr>
<td>Sound sources</td>
<td>Speech signals (2 [speaker] * 10 [word])</td>
</tr>
<tr>
<td>Noise source</td>
<td>White noise</td>
</tr>
<tr>
<td>Frequency range for estimation</td>
<td>0 ... 3, 0 ... 7 [kHz]</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>16 [kHz]</td>
</tr>
<tr>
<td>Quantization</td>
<td>16 [bit]</td>
</tr>
<tr>
<td>Microphone</td>
<td>HOSHIDEN, KUC-1333</td>
</tr>
<tr>
<td>Mic. amplifier</td>
<td>Thinknet, MA-2016</td>
</tr>
<tr>
<td>Mouth Simulator</td>
<td>Bruel &amp; Kjaer, Type 4227</td>
</tr>
<tr>
<td>Loudspeaker</td>
<td>MITSUBISHI, DIATONE DS-7</td>
</tr>
<tr>
<td>Speaker amplifier</td>
<td>BOSE, 1705II</td>
</tr>
<tr>
<td>A/D, D/A</td>
<td>Inrevium, TD-BD-16ADUSB</td>
</tr>
</tbody>
</table>

4.2. Experimental Results

We evaluated the localization accuracy by the estimation accuracy for direction and range. Figure 4 shows the estimation accuracy for direction. The tolerances were 5, 10, and 15 [degree]. The estimation accuracy for the direction of the proposed multiple channel 2D-CSP improved 45 % compared to that of 2D-MUSIC and 15 % compared to that of multiple paired 2D-CSP, as shown in Fig. 4(c). Figure 5 also shows the estimation accuracy for range. The tolerances were 10–50 % of the distance between the microphone array and the sound source. As a result of Fig. 5(a) in 20 % tolerance, the estimation accuracy for the range of the proposed method improved 30 % compared to that of 2D-MUSIC and 10 % compared to that of multiple paired 2D-CSP. The results, indicated that the proposed method performed much better than conventional 2D-MUSIC and mul-
tiple paired 2D-CSP, especially in the diffused noisy environment. Because the accuracy of the proposed method was better than that of multiple paired 2D-CSP, we concluded that the effectiveness of utilizing phase differences at the same time was effective. When the frequency range changed from 3 to 7 [kHz] in the diffused noisy environment, the estimation accuracy decreased. The dominant frequency of the human voice is under 3 [kHz]. In contrast, white noise has the same power in all frequency bands. Therefore, the errors increased because the white noise has more power than a voice at over 3 [kHz].

The results indicated that the proposed method is much better than conventional methods in real environments. By estimating the direction and the range of a sound source with less error, the microphone array can capture speech with high quality. In addition, the estimation accuracy for not only the direction but also the range improved. Therefore, the method may be able to be applied to security systems that use sound source location to control the direction and the diameter of the camera.

5. Conclusions

A sound source must be accurately localized to capture distant-talking speech with a microphone array. We developed a two-dimensional cross-power spectrum phase analysis with multiple microphones (multiple channel 2D-CSP) by expanding conventional CSP for robustly localizing sound sources in the near field. Evaluation experiments in diffused noisy environments showed that the proposed method could localize a sound source more robustly than conventional methods. In future work, we will evaluate the proposed method with various microphone locations and multiple sound sources.

6. Acknowledgements

This work was partly supported by Global-COE and Grants-in-Aid for Scientific Research funded by The Japanese Ministry of Education, Culture, Sports, Science and Technology.

7. References