Abstract: We have developed the field recording, recognition and reproduction (FIR3) system to record a sound field for later reproduction with the goal of reconstructing the sound information of a room in another space at another time. In this system, a surrounding microphone array is used to record a sound field. A method for detecting sound source positions using this microphone array is discussed in this paper. First, the microphone array properties were examined. On the basis of the results of this examination, we developed a method in which the multiple signal classification (MUSIC) algorithm and the spatial smoothing technique are integrated and named it “rearrangement and presmoothing for MUSIC” (RAP-MUSIC). Measurement in an actual room showed that, using this method, source positions in a reverberant room can be accurately detected.

Keywords: Surrounding microphone array, Sound source localization, MUSIC algorithm, Spatial smoothing technique, Delay-and-sum algorithm

PACS number: 43.60.−c, 43.60.Fg [doi:10.1250/ast.28.181]

1. INTRODUCTION

Recently, techniques have been developed with the goal of obtaining abundant sound field information to thereby enable highly realistic sound reproduction for such applications as home theater systems and cinemas. The sound produced by these systems is, however, unrealistic.

Many studies on recording a sound field for later reproduction in another space and time have been conducted. For example, binaural recording using a dummy head [1] and a method based on the Kirchhoff-Helmholtz integral equation [2] have been proposed for this purpose. Although a system of binaural recording and reproduction is easy to implement, the reproduced sound does not change adaptively in accordance with the user’s movement. This problem can be avoided using a method based on the Kirchhoff-Helmholtz integral equation. This type of method is, however, incapable of handling a case in which sound sources exist within the control region. We have therefore developed the field recording, recognition and reproduction (FIR3) system.

We employed two rooms to realize the FIR3 system. One is a room in which an array of microphones is installed 30 cm from all four walls and the ceiling using pipes. This room is called a “recording room” (Fig. 1). Figure 2 shows the arrangement of the microphones in the array. There are two narrow walls and two wide walls, represented by Wall A and Wall B, respectively, in Fig. 2. 20 microphones are installed on each of the narrow walls and there are 36 microphones on each of the wide walls. The ceiling has 45 microphones. All microphones are separated from each other by 50 cm. We designated these microphones collectively as a “surrounding microphone array.” This surrounding microphone array, therefore, includes 157 microphones. The microphones (Type 4951; Brüel & Kjær) are connected to 10 microphone amplifiers (Type 2694; Brüel & Kjær). The other room is a room having the same configuration as the recording room. In this room, speakers are set at the same positions as those of the microphones in the recording room to reproduce an identical sound field effectively. This room is therefore called the “reproduction room” (Fig. 3). The reverberation time of both rooms is about 0.15 s.

We are attempting to develop a method of recording the necessary information of the sound field in the recording room and reproduce it for a listener in the reproduction room using the FIR3 system. In the method described here, direct sound and reflected sound from walls are provided separately to the listener. The direct sound is reproduced in accordance with the position of the listener relative to the sound source and is presented to the listener through an auditory display device called the “localization...
auditory display with opened ear-canal for mixed reality” (LADOMi) [3]. This device, which has no ear muffler, as shown in Fig. 4, is suitable for use with this method because the reflected sound enters a listener’s ears with less distortion. The listener’s position is sensed using a gyro and an ultrasonic sensor (IS-900; INTER-SENSE) set in the room. The listener can, therefore, move around in this room [4]. Primary reflected sounds are presented to the listener using the surrounding speaker array. Because the properties of the recording room and the reproduction room are identical, other reflected sounds are reproduced automatically by the reflected sounds in the reproduction room.

Sound source positions and waves in the recording room are estimated from the input sound signals obtained using the surrounding microphone array. Moreover, a method of extracting the first reflected sound from each microphone input is also needed to realize the FIR$^3$ system.

In this study, as a first step toward realizing a realistic sound field reproduction system using FIR$^3$, a method of detecting sound source positions using the surrounding microphone array was investigated.

2. PROPERTIES OF THE SURROUNDING MICROPHONE ARRAY

2.1. Size of the Array

The surrounding microphone array is larger than that of commonly used microphone arrays. Considering the relation between the size of the array and the distance between the source and each microphone, we assumed that the input signals of each microphone were not plane waves, but spherical waves. In addition, we estimated the sound source positions in a three-dimensional space rather than in the directions of the sound sources.

2.2. Spacing between Microphones

In detecting sound source positions using a microphone array, spatial aliasing occurs at more than twice the frequency that conforms to the minimal distance between the microphones of the array, when a method based on spatial Fourier transformation, such as the delay-and-sum
beam-forming method, is used [5]. When the surrounding microphone array of FIR is used, sidelobes appear at frequencies greater than 340 Hz, because all microphones are separated from each other by a spacing of 50 cm.

We therefore decided to use the multiple signal classification (MUSIC) method [6], which is based on the eigenspatial of a spatial correlation matrix composed of input signals, to detect sound source positions.

In MUSIC, the source positions are extracted from the spatial correlation matrix $R = E[x(\omega, t)x^*(\omega, t)]$ composed of input signals $x(\omega, t)$. We extract the base vector, $V$, of the orthogonal complements from $R$. The extracted $V$ is orthogonal to the transfer function calculated from the distance between each microphone. From the base vector, $V$, and a steering vector, $a = [a_0, a_2, \ldots, a_n]^T$, which is defined as

$$a_i(\omega, x, y, z) = \frac{1}{r_i} e^{j\omega z_i},$$

using the distances, $r_i$, between microphones $i$ and the search point, the value of $P$ is calculated as

$$P(x, y, z) = \frac{1}{\|V^*a(x, y, z)\|^2}. \tag{2}$$

It is apparent that $P$ has a peak because $a$ is orthogonal to $V$ if $a$ fits the source position.

In this equation, $(x, y, z)$ are coordinates, $\omega$ is the angular frequency, $t$ is time, and $c$ is the acoustic velocity.

### 2.3. Arrangement of Microphones

Three basic patterns of arrangement can be considered for detecting three-dimensional sound source positions using microphones. One is a two-dimensional arrangement that uses only one wall or the ceiling (Fig. 5). The others are three-dimensional arrangements that include more than one wall and the ceiling (Figs. 6 and 7). In these figures, “○” indicates a microphone and “●” indicates a source. Experiments on detecting sound source positions under these arrangements were carried out to clarify the relationship between the arrangement of the microphones and the accuracy of estimation.

For each arrangement, 45 microphones were used. An experiment was carried out to detect two source positions. Two equilateral 32-hedron point source loudspeakers, each 7 cm in diameter, were used as sources. Source signals were excerpts from Japanese word lists (FW03) compiled by NTT-AT. Source 1 was a male voice (/a/i/kya/ku), and source 2 was a female voice (/i/chi/yu/u). The presentation levels of sources 1 and 2 were 65.3 and 50.4 dB ($L_{eq}$), respectively, measured at a distance of 1 m from each source. The experimental conditions are shown in Table 1.

The index $k$, defined as

$$k = \frac{P \text{ at each search position}}{P \text{ at the source position}}, \tag{3}$$

would allow for a more accurate measurement of the source positions.

### Table 1  Experimental condition.

<table>
<thead>
<tr>
<th>Sampling frequency</th>
<th>48,000 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time length of input signal</td>
<td>1 s</td>
</tr>
<tr>
<td>Window function</td>
<td>Hamming window</td>
</tr>
<tr>
<td>Window length</td>
<td>512 points</td>
</tr>
<tr>
<td>Frequency band</td>
<td>500 to 8,000 Hz</td>
</tr>
</tbody>
</table>

Fig. 5 Two-dimensional arrangement using the ceiling (pattern 1).

Fig. 6 Three-dimensional arrangement using a wall and the ceiling (pattern 2).

Fig. 7 Three-dimensional arrangement using two walls and the ceiling (pattern 3).
was used to evaluate the degree of the peak of $P$. In this case, $k$ should be unity when the positions of the search point and each source are the same. Moreover, when the position of the search point is elsewhere, $k$ should be approximately zero.

Results are shown in Figs. 8–10. These figures show the change of $k$ in each pattern by fixing two of the coordinates ($x$, $y$, or $z$) at the source position and changing the other coordinate. In the three figures, "*" indicates the correct source position. These results show that sound source positions were estimated correctly in patterns 2 and 3. In pattern 1, on the other hand, estimation of the sound source position along the $z$-axis was erroneous, but it was correct along the other axes.

This fact is explainable by the characteristics of the steering vector, $a$, which is calculated from the distance, $r_i$, between each microphone and the search point. In pattern 1, when the search point moves a short distance along the $z$-axis, all $r_i$ are lengthened or shrunk with the same sign, resulting in $a(x, y, z + dz) = ca(x, y, z)$, where $c$ is a constant number. In pattern 2 and pattern 3, on the other hand, when the search point moves along the $z$-axis, each $r_i$ is lengthened or shrunk with the anticlastic sign, resulting in $a(x, y, z + dz) \neq ca(x, y, z)$ [7]. Consequently, results showed that with a three-dimensional arrangement, sound source positions can be more accurately estimated than with a two-dimensional arrangement.

3. ESTIMATION OF SOUND-SOURCE POSITIONS IN A REVERBERANT SOUND FIELD

3.1. Spatial Smoothing Technique

The reflected sound negatively affects the estimation of the sound source positions [8]. The spatial smoothing technique is a method for reducing this negative effect [9]. The phase relation of the reflected sound waves differs from that at each receiving point. Therefore, by averaging the spatial correlation matrices, which are calculated at the receiving points and which move in parallel, the cross-correlation values between the direct sound and the reflected sound decrease. Practically, when applying the spatial smoothing technique to direction-of-arrival sound estimation using the MUSIC method, we divide the array
into $N$ subarrays instead of moving them (Fig. 11). Then, we calculate the spatial correlation matrices, $R_n$, which are calculated for each subarray (Fig. 12), and execute the MUSIC method using the correlation matrix averaged using

$$ \bar{R} = \sum_{n=1}^{N} R_n. $$

(4)

This method is based on the fact that in direction-of-arrival sound estimation, the steering vector of each subarray is the same in each direction if the shapes of all subarrays are congruent. For an irregularly arranged array, moreover, some methods have been proposed in which actually obtained signals are converted into signals obtained at a virtual equally spaced discrete linear array by phase correction (Fig. 13). We assume that the distance between a search point and a real channel is $r_i$, that the distance between a search point and a virtual channel is $r_i'$, and that the signal at a real channel at time $t$ is represented as $x(t)$. In that case, a signal at a virtual channel with distance $r_i'$ at time $t$ is obtained as

$$ x'(t) = \frac{r}{r'} x(t) e^{-j \omega t \frac{r_i}{r_i'}}. $$

(5)

Using this conversion, the steering vectors of all subarrays become identical. Hence, we can apply the spatial smoothing technique using the averaged spatial correlation matrix of each virtual channel set.

However, by moving the search point to another point, the distance between the new search point and each real channel changes at random. We therefore cannot use the existing virtual channel sets and must regenerate the virtual channels, because the steering vector of each subarray is the same as that when moving a search point. Hence, to apply the spatial smoothing technique to the estimation problem of three-dimensional sound source positions, we must regenerate the virtual channels using Eq. (5) and the spatial correlation matrix at each search point.

In studies, by Friedlander and Weiss, and Wax and Sheinvaid [10,11], because the same transformation matrix was used for generating common virtual channels, the

---

3.2. Introduction of the Spatial Smoothing Technique for a Spherical Model

In this section, we propose a method of applying the spatial smoothing technique to the estimation of three-dimensional sound source positions. The steering vector of each subarray must be the same at each search point to apply the spatial smoothing technique. First, we consider the case when the search point is at a certain point. We create virtual channels at the positions at which the distance between the search point and each subarray is the same, and convert the signals at the real channels into the signals at the virtual channels by both amplitude and phase correction (Fig. 13). We assume that the distance between a search point and a real channel is $r_i$, that the distance between a search point and a virtual channel is $r_i'$, and that the signal at a real channel at time $t$ is represented as $x(t)$. In that case, a signal at a virtual channel with distance $r_i'$ at time $t$ is obtained as

$$ x'(t) = \frac{r}{r'} x(t) e^{-j \omega t \frac{r_i}{r_i'}}. $$

(5)

Using this conversion, the steering vectors of all subarrays become identical. Hence, we can apply the spatial smoothing technique using the averaged spatial correlation matrix of each virtual channel set.

However, by moving the search point to another point, the distance between the new search point and each real channel changes at random. We therefore cannot use the existing virtual channel sets and must regenerate the virtual channels, because the steering vector of each subarray is the same as that when moving a search point. Hence, to apply the spatial smoothing technique to the estimation problem of three-dimensional sound source positions, we must regenerate the virtual channels using Eq. (5) and the spatial correlation matrix at each search point.

In studies, by Friedlander and Weiss, and Wax and Sheinvaid [10,11], because the same transformation matrix was used for generating common virtual channels, the
The correlation matrix $\mathbf{R}$ was evaluated to examine the effect of the spatial smoothing technique in the MUSIC algorithm. If the spatial smoothing technique is not applied, the element $R(p, q)$ of the correlation matrix $\mathbf{R}$ is given as

$$R(p, q) = E[x_p(t)x_q(t)^*]$$

$$= \frac{1}{r_{Sp}r_{Sq}} P_S e^{-j\omega(t_{Sp} - t_{Sq})} + \frac{1}{r_{Sp}r_{Cq}} E[S(t)C(t)^*] e^{-j\omega(t_{Sp} - t_{Cq})} + \frac{1}{r_{Cp}r_{Sq}} E[C(t)S(t)^*] e^{-j\omega(t_{Cp} - t_{Sq})} + \frac{1}{r_{Cp}r_{Cq}} P_C e^{-j\omega(t_{Cp} - t_{Cq})},$$

where

$$P_S = E[S(t)S(t)^*], \quad P_C = E[C(t)C(t)^*].$$

Reflected sounds can be regarded as early reflected sounds and reverberation. Early reflected sounds correlate very closely with direct sounds although reverberation does not [13]. If $C(t)$ is the early reflected sound of $S(t)$, $C(t)$ resembles $S(t)$. In this case, because the second to fourth terms affect the first term, which is formed by direct sound, the estimation accuracy of the sound source positions worsens. Hence, the second to fourth terms can be regarded as error terms of the first term in a reverberant environment. It is therefore desirable that these terms should be approximately zero.

If the spatial smoothing technique is applied, on the other hand, $R(p, q)$ at subarray $n$ ($n = 1, 2, \cdots, N$) is given as

$$R'_n(p, q) = E[x'_p+n-1(t)x'_q+n-1(t)^*]$$

$$= \frac{1}{r'_{Sp+n-1}r'_{Sq+n-1}} P_S e^{-j\omega(t_{Sp+n-1} - t_{Sq+n-1})} + \frac{r'_{Sp+n-1}r'_{Sq+n-1}}{r'_{Sp+n-1}r'_{Cq+n-1}} E[S(t)C(t)^*] e^{-j\omega(t_{Sp+n-1} - t_{Cq+n-1})} + \frac{r'_{Sp+n-1}r'_{Cp+n-1}}{r'_{Sp+n-1}r'_{Sq+n-1}} E[C(t)S(t)^*] e^{-j\omega(t_{Cp+n-1} - t_{Cq+n-1})} + \frac{r'_{Sp+n-1}r'_{Cp+n-1}}{r'_{Cp+n-1}r'_{Sq+n-1}} P_C e^{-j\omega(t_{Cp+n-1} - t_{Cq+n-1})}.$$
+ \frac{1}{r_{Sp}^2 r_{Sq}^2} E[S(t)C(t)^*] e^{-j\phi(t_{\text{d}} - t_{\text{S}})} \cdot \frac{1}{N} \sum_{n=1}^{N} E_q(n)
+ \frac{1}{r_{Sp}^2} \quad E[C(t)S(t)^*] e^{-j\phi(t_{\text{d}} - t_{\text{S}})} \cdot \frac{1}{N} \sum_{n=1}^{N} E_p(n)
+ \frac{1}{r_{Sp}^2} P Ce^{-j\phi(t_{\text{d}} - t_{\text{S}})} \cdot \frac{1}{N^2} \sum_{n=1}^{N} E_q(n) \sum_{n=1}^{N} E_p(n). 
(13)

The second to fourth terms of Eq. (13) are small when \( N \) is sufficiently large. A comparison of Eq. (8) with Eq. (13) shows that the spatial smoothing technique reduces the influence of reflected waves because of the smaller values of in the second to fourth terms.

Using this method, if the positions of the search point and source position are identical, the effect indicated by Eq. (13) occurs, and it is expected that the value of \( P \) and source position are identical, the effect indicated by in the second to fourth terms.

\( \text{influence of reflected waves because of the smaller values} \)

Therefore, this method can be used to estimate the sound source positions more accurately.

3.3. RAP-MUSIC

The method described above reduces the effect of reflected sound by averaging the correlation matrices while keeping the phase of all the direct sounds the same. We therefore decided to average the signals at virtual channels after synchronizing the phases and then calculate the correlation matrix instead of averaging the correlation matrices. This method corresponds to the delay-and-sum algorithm. We applied the delay-and-sum algorithm to the signals at the virtual channels of subarrays before applying the MUSIC algorithm. In this case, from Eq. (7), the delay-and-sum signals are given as

\[
\bar{x}(t) = \frac{1}{N} \sum_{n=1}^{N} x(t)
= \frac{1}{r_{Si}} S(t) e^{-j\phi_{\text{d}}} + \frac{1}{N} \sum_{n=1}^{N} \frac{1}{r_{Si} r_{Ci+n-1}} C(t) e^{-j\phi(t_{\text{d}} t_{\text{n}} - t_{\text{d}} t_{\text{n}} + t_{\text{d}}).} 
(14)
\]

Therefore, the correlation matrix is given as

\[
\hat{R}(p, q) = E[x_q^*(t) \bar{x}_p(t)']
= \frac{1}{r_{Sp}^2 r_{Sq}^2} P Se^{-j\phi(t_{\text{d}} - t_{\text{S}})}
+ \frac{1}{r_{Sp}^2} E[S(t)C(t)^*] e^{-j\phi(t_{\text{d}} - t_{\text{S}})} \cdot \frac{1}{N} \sum_{n=1}^{N} E_q(n)
+ \frac{1}{r_{Sp}^2} E[C(t)S(t)^*] e^{-j\phi(t_{\text{d}} - t_{\text{S}})} \cdot \frac{1}{N} \sum_{n=1}^{N} E_p(n)
+ \frac{1}{r_{Sp}^2} P Ce^{-j\phi(t_{\text{d}} - t_{\text{S}})} \cdot \frac{1}{N^2} \sum_{n=1}^{N} E_q(n) \sum_{n=1}^{N} E_p(n).
(15)
\]

We regard the second to fourth terms of Eq. (15) as the error terms. Therefore, because of the factor \( 1/N^2 \) of the fourth term, the effect of reduced sound can be reduced more efficiently by the proposed method than by the original spatial smoothing technique.

The same method of generating virtual channels from real channels by amplitude and phase correction for beam-forming has been previously employed by Zheng et al. [14]. In their report, however, the effect of reducing the effect of reflected waves by averaging was not discussed in theoretical terms.

We applied the MUSIC algorithm using the signals averaged at each virtual channel and calculated this process at each search point. We therefore named this method “rearrangement and presmoothing for MUSIC” (RAP-MUSIC).

4. PERFORMANCE EVALUATION

4.1. Experimental Conditions

We detected sound source positions by applying the spatial smoothing technique and RAP-MUSIC using the experimental data presented in Sect. 2.3. We generated virtual channels from four real channels at their median point (Fig. 14). The number of virtual channels was 32 for patterns 1 and 2 (Figs. 15–16) and 30 for pattern 3 (Fig. 17). In these figures, “.” indicates a real channel, and “o” indicates a virtual channel.

4.2. Experimental Results

Results for each method are shown in Figs. 18–23, which compare their performances for each pattern. The vertical axis shows \( k \), as defined in Eq. (3). These graphs show the estimation results of only the five largest values of \( k \). In these graphs, shaded bars correspond to the correct positions.

These results clarified that the estimation of correct source positions was possible using the spatial smoothing

---

**Fig. 14** Keeping the phase of each signal at the virtual channel.
Fig. 15 Arrangement of real channels and virtual channels of pattern 1 (32 ch).

Fig. 16 Arrangement of real channels and virtual channels of pattern 2 (32 ch).

Fig. 17 Arrangement of real channels and virtual channels of pattern 3 (30 ch).

Fig. 18 Comparison of the performances of the three methods for source 1 (pattern 1).

Fig. 19 Comparison of the performances of the three methods for source 2 (pattern 1).

Fig. 20 Comparison of the performances of the three methods for source 1 (pattern 2).

Fig. 21 Comparison of the performances of the three methods for source 2 (pattern 2).

Fig. 22 Comparison of the performances of the three methods for source 1 (pattern 3).

Fig. 23 Comparison of the performances of the three methods for source 2 (pattern 3).
technique and RAP-MUSIC. Moreover, only RAP-MUSIC succeeded in estimating correct positions in the case of the two-dimensional arrangement (pattern 1), whereas the MUSIC algorithm and the spatial smoothing technique failed. Accordingly, RAP-MUSIC is superior to MUSIC and the spatial smoothing technique in a reverberant environment under the tested conditions.

5. CONCLUSION

In this study, an effective method of estimating sound source positions using the surrounding microphone array of the FIR³ system was developed. Experiments showed that a three-dimensional arrangement of arrays is more effective than a two-dimensional arrangement. In addition, RAP-MUSIC, based on MUSIC and the spatial smoothing technique, was proposed and employed to reduce the influence of reverberant waves.

ACKNOWLEDGEMENTS

We wish to thank Prof. Yōiti Suzuki for his help in conducting this study.

REFERENCES