Sound Field Interpolation for Rotation-Invariant Multichannel Array Signal Processing

Yukoh Wakabayashi, *Member, IEEE*, Kouei Yamaoka, *Student Member, IEEE*, and Nobutaka Ono, *Senior Member, IEEE*

***Abstract*—In this paper, we present a sound field interpolation for array signal processing (ASP) that is robust to rotation of a circular microphone array (CMA), and we evaluate beamforming as one of its applications. Most ASP methods assume a time-invariant acoustic transfer system (ATS) from sources to the microphone array. This assumption makes it challenging to perform ASP in real situations where sources and the microphone array can move. Therefore, considering a time-variant ATS is an essential task for the use of ASP. In this study, we focus on one such movement, the rotation of the CMA. Our method interpolates the sound field on the circumference of a circle, where microphones are equally spaced, based on the sampling theorem on the circle. The interpolation enables us to estimate the signals at the microphone positions before the rotation. Hence, conventional ASP, which assumes a time-invariant ATS, is applicable after interpolation without modification. We developed two beamforming schemes, one for batch and one for online processing, that combine the minimum power distortionless response beamformer and sound field interpolation. We evaluated the dependences of the interpolation on frequency and rotation angle using the signal-to-error ratio. Additionally, simulation results demonstrated that the two proposed schemes improve the beamformer's performance when the CMA rotates.**

***Index Terms*—Sound field interpolation, noninteger sample shift, circular microphone array, time-variant acoustic transfer system, online array signal processing, wearable devices.**

I. Introduction

ARRAY signal processing methods remain an important research topic. Examples of its related topics include beamforming, source separation, and estimation of the direction and the time difference in source arrival. Independent low-rank matrix analysis [1] and multichannel nonnegative matrix factorization [2] are state-of-the-art source separation methods using sophisticated models. They have been extended to studies such as independent deeply learned matrix factorization [3] and an alternative update rule of the demixing matrix, that is, the so-called iterative source steering [4]. Novel approaches for beamforming have also been presented, e.g., time-frequency-bin-wise switching beamforming [5], [6] and the time-varying spatial covariance matrix (SCM) estimation [7], [8]. These advanced methods have achieved high performance through the modification of the spatial or source models or the calculation methodology. At the same time, they require a time-invariant acoustic transfer system (ATS) to maintain the performance. In other words, in these methods, it is assumed that the microphone array and the sources do not move. A change in the ATS imposes the re-estimation of the spatial filter, making real-time processing difficult. Most array signal processing (ASP) methods require statistical information such as the SCM to estimate the spatial model. Therefore, re-estimation of the filter requires a long time duration, which becomes a bottleneck for online processing in a real environment.

As mentioned above, dealing with the time-invariant ATS is one factor to be considered in the practical use of online ASP in a real environment. The problem of time-variant ATS is separated into two cases: moving sources and moving sensors. The basic approach in the former case is blockwise processing by combining the direction-of-arrival (DOA) information estimated by another module [9], [10], tracking multisources using DOA estimates, and separating the sources. However, even such an approach requires the re-estimation of the spatial filter for every block in which the DOAs change. In comparison, Taseska and Habets [11] realized online source separation by estimating the SCM sequentially with DOA information estimation. Our method adopts the latter case, but this is not a highly active area of research; the overview of ASP by Gannot et al. [12] introduces several studies in the former case, but not in the latter case. Examples of moving sensors include the situation where a robot or human wearing a microphone array on the head or a human wearing hearing aids rotates the head to listen attentively to the ambient conversation. Also, Valimaki et al. [13] have considered controlling spatial acoustic information in the virtual reality space. For such a situation, Tourbabin and Rafaely [14] interpolated the sound field with a motion compensation matrix in the spherical harmonic (SH) domain to estimate a DOA for a moving humanoid robot. They used the *Wigner's D* matrix related to spherical and symmetric rigid rotors as the rotation matrix. Corey and Singer [15] surveyed beamforming performance with the two types of deformable microphone array when these rotate, assuming that the SCM after rotation is obtained as the training data. Casebeer et al. [16] endeavored

Manuscript received 29 November 2022; revised 28 April 2023; accepted 18 May 2023. Date of publication 1 June 2023; date of current version 16 June 2023. This work was supported in part by JSPS KAKENHI under Grants JP20K19827 and JP20H00613 and in part by JST CREST under Grant JPMJCR19A3, Japan. The associate editor coordinating the review of this manuscript and approving it for publication was J. Borgstrom. *(Corresponding author: Yukoh Wakabayashi.)*

Yukoh Wakabayashi is with the Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi 441-8580, Japan (e-mail: wakayuko@cs.tut.ac.jp).

Kouei Yamaoka is with the Graduate School of Systems Design, Tokyo Metropolitan University, Hino 191-0065, Japan (e-mail: yamaoka-kouei@ed.tmu.ac.jp).

Nobutaka Ono is with the Graduate School of Systems Design, Tokyo Metropolitan University, Hino 191-0065, Japan (e-mail: onono@tmu.ac.jp).

Digital Object Identifier 10.1109/TASLP.2023.3282098

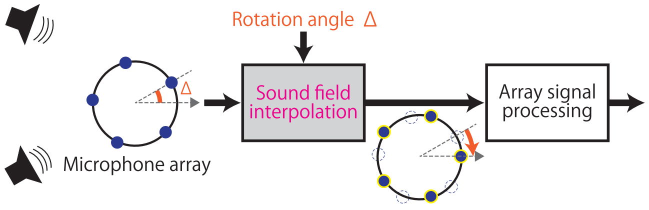

Fig. 1. Conceptual diagram of sound field interpolation.

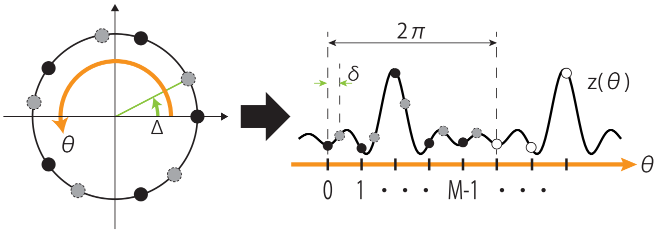

Fig. 2. Sound field interpolation in a circular microphone array with $\delta = \Delta M / 2\pi$ sample shifts of the discrete sound field.

to deal with the pose change of mobile microphone arrays and proposed time-varying SCM estimation using a recurrent neural network. In addition to processing in a dynamic scenario where sensors and sources move, various methods for dynamic scenario simulation have also been studied [17], [18].

We assume the sensor-moving situation described above. In this study, in particular, we take one of the ATS variations, the rotation of a microphone array, into consideration and propose a sound field interpolation for ASP robust to the array's rotation, using a circular microphone array (CMA). That is, even if the CMA rotates, the proposed scheme enables the time-variant ATS to be regarded virtually as a time-invariant one, thereby enabling any conventional ASP to work well. A strength of the proposed method is its independence on any particular kind of downstream ASP. The conceptual diagram of the framework is shown in Fig. 1. Interpolation of the sound field before ASP compensates for the time-invariant ATS. The proposed method utilizes the periodicity of the sound field on the circumference of a circle and the relationship between sensing the sound field with an equally spaced CMA and discretizing the sound field. These two points and the noninteger sample shift theorem of a discrete signal enable the estimation of the sound field of the position at which there are no microphones by a simple calculation. Moreover, we apply this scheme to beamforming, which is one ASP, extend it to online beamforming, and confirm its efficacy via numerical simulation.

Our study, focusing on the CMA rotation, is related to two research topics: modal ASP and interpolation. First, it is necessary to touch on modal ASP, i.e., SH-domain processing or, in the case of restriction to the two-dimensional space, circular or cylindrical harmonics (CH)-domain processing [19]. SH-domain processing decomposes the sound field into different directivities with characteristics such as monopole and dipole. Handling frequency–wavenumber spectra in each directivity enables the control of the spatial filter and the generation of the desired directivity patterns. It is often applied, for example, to beamforming [20], [21], [22], [23], [24], SCM estimation [25], sound field reproduction [26], and DOA estimation [14], [27]. In particular, the concept in [14] resembles that of our work, and they have a theoretical relationship (see the discussion in Section II-C2). Unlike the method in [14], our method works in the original signal domain, not in the SH nor CH domain. It allows us to utilize any ASP techniques as they are and make a connection with the interpolation as we describe next. This idea enables beamforming in the short-time Fourier transform (STFT) domain to naturally be extended from batch to online processing. It works seamlessly even under the CMA rotation. Also, the other related topic is interpolation. Ueno et al. [28] interpolated the sound field of the desired area by solving the optimization problem using the Helmholtz equation. Yamaoka et al. [29] interpolated the generalized cross-correlation function of two sound sources via the sinc function to obtain their time difference at the noninteger sample level. Schüldt [30] introduced the trigonometric interpolation to solve the problem of oscillation in polynomial beamforming [31] by using the symmetry and periodicity of the CMA, as in our method. These studies show that the acquisition of noninteger sample points improves the estimation accuracy. Our method applies the noninteger sample shift theorem to the sound field on a circle to achieve interpolation. It leads to a simple linear transformation for the equally spaced CMA. An expansion with unequally spaced CMA has been discussed in [32].

This paper includes some of the content of the conference paper [33] in which we reported the sound field interpolation and applied it to beamforming. The contribution of this paper is that we extended beamforming to online processing. In addition, note that we modified the formulation of the sound field interpolation for an even number of microphones because we found an error in the sign in equation (3) of [33]. Moreover, we rethought the handling of the Nyquist frequency component in the sound field interpolation and conducted new experiments. The remainder of this paper is organized as follows. In Section II, we explain the idea of our sound field interpolation and formulate it. Also, we discuss several points related to the formulation. In Section III, we describe how to apply the interpolation to beamforming as an example of multichannel signal processing with batch and online methodologies. Then, in Section IV, we evaluate the performance of the sound field interpolation itself using the signal-to-error ratio (SER) as a metric and measure the accuracy of beamforming as downstream processing. Finally, we conclude this paper in Section V.

II. Sound Field Interpolation using Circular Microphone Array

A. Formulation

First, we consider a continuous sound field on the circumference of a circle, $z(\theta),\ \theta \in [0, 2\pi)$, as shown in Fig. 2. Here, θ indicates the spatial angle. Obviously, z is a periodic function with 2π, although we cannot a priori know its concrete formulation. When we set M microphones on the circle circumference at even intervals and record a sound field with them, the mth

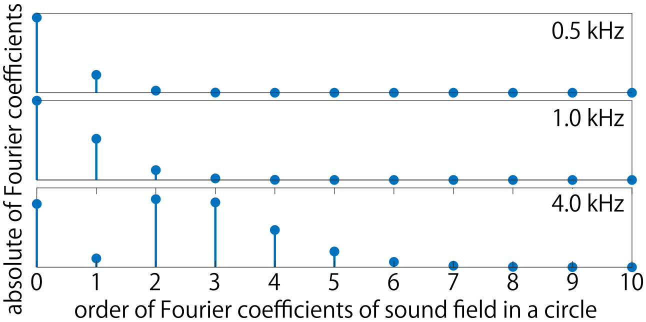

Fig. 3. Absolute value of Fourier coefficient of sound field in a circle, $|Z_k|$, assuming planar waves with 0.5 kHz, 1 kHz, and 4 kHz, where the horizontal axis indicates k.

observed signal is represented as

$$z_m = z\left(2\pi\frac{m}{M}\right), \quad m = 0, \ldots, M-1. \tag{1}$$

That is, sensing a sound field with a CMA is equal to the discretization of the sound field along the spatial angle.

Then, assuming that the maximum frequency of $z(\theta)$ is less than half of the spatial sampling frequency on the circle's circumference, $M/2\pi$, according to the sampling theorem, we can say that the discrete signal z_m can reconstruct the continuous sound field $z(\theta)$. Generally speaking, this assumption is not strictly satisfied. For example, for a planar wave $e^{j\omega t}$ coming from a direction of $0\,\mathrm{rad}$, the Fourier coefficients of (1) are represented by the Bessel function of the first kind, such as $Z_k = e^{j\omega t}\int_0^{2\pi} e^{-j(k\theta + r\omega\cos\theta/c)}d\theta$, where k is the order of the Fourier coefficient, c is the sound speed, and r is the circle's radius. As is well known, Z_k does not have finite support of k. However, for a large k, $|Z_k|$ can be sufficiently small for practical use, depending on ω. Concretely, $|Z_k|$ is almost negligible for $\omega r/c > k$, which has been discussed by Alon and Rafaely [34]. Fig. 3 shows some examples of $|Z_k|$ in the case of $r = 0.05\ \mathrm{m}$. We can see that $|Z_k|$ is close to zero for a large k. On the basis of this observation, we proceed to a discussion, assuming the sampling theorem holds. We will investigate the effect of the error of this approximation experimentally in a later section.

B. Formulation of Linear Interpolation

To formulate a sound field interpolation, we utilize the noninteger sample shift theorem in the Fourier domain. We can clearly say that the shifted signal $z_m(\delta)$ of the δ-sample corresponds to the observation with the CMA rotated $\Delta = 2\pi\delta/M\ \ \mathrm{rad}$. Specifically, when we designate the sound field with the CMA in the reference position (i.e., not rotating) as z_m, the observation of the same sound field with the $\Delta\,\mathrm{rad}$-rotated CMA is represented as $z_m(\delta) = z(2\pi m/M + \Delta)$. From the shift theorem, $z_m(\delta)$ can be expressed as

$$\begin{aligned} z_m(\delta) &= \frac{1}{M}\sum_{k=-M/2+1}^{M/2}\left(\mathscr{F}_{\mathrm{D}}\left[z_m\right]e^{j\Delta k}\right)e^{j\frac{2\pi m k}{M}} \\ &\equiv \sum_{n=0}^{M-1} z_n U_{mn}(\delta). \end{aligned} \tag{2}$$

Although the shift theorem using the discrete Fourier transform (DFT) $\mathscr{F}_{\mathrm{D}}$ does not strictly hold with a noninteger δ, we assume its satisfaction. The coefficient $U_{mn}(\delta)$ is defined as

$$\begin{aligned} &U_{mn}(\delta) \\ &= \begin{cases} \frac{1-(-1)^{n-m}e^{-j\delta\pi}}{M} + \frac{\mathrm{sinc}(ML)\cos(M+2)L\pi}{\mathrm{sinc}(2L)}, & (\text{even } M), \\ \frac{1}{M} + \frac{M-1}{M}\frac{\mathrm{sinc}\big((M-1)L\big)\cos(M+1)L\pi}{\mathrm{sinc}(2L)}, & (\text{odd } M), \end{cases} \end{aligned} \tag{3}$$

where $L = (n-m-\delta)/2\,M$ and $j = \sqrt{-1}$. The detailed derivation of (3) is provided in Appendix A. From (2), we can also represent the sound field interpolation using matrix operation with the rotation transform matrix $\boldsymbol{U}(\Delta)$ defined as

$$\begin{aligned} \boldsymbol{z}(\delta) &= \begin{bmatrix} z_0(\delta) & z_1(\delta) & \cdots & z_{M-1}(\delta) \end{bmatrix}^{\mathsf{T}} \\ &= \begin{bmatrix} U_{00}(\delta) & \cdots & U_{0(M-1)}(\delta) \\ U_{10}(\delta) & \cdots & U_{1(M-1)}(\delta) \\ \vdots & \ddots & \vdots \\ U_{(M-1)0}(\delta) & \cdots & U_{(M-1)(M-1)}(\delta) \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_{M-1} \end{bmatrix} \\ &= \boldsymbol{U}(\Delta)\boldsymbol{z}. \end{aligned} \tag{4}$$

Note that $\boldsymbol{U}(\Delta)$ is a cyclic and unitary matrix and its components do not depend on the frequency of the observation.

C. Analysis

1) Discussion of Nyquist Frequency Component: The major difference between odd and even M in (3) is whether or not we must handle the Nyquist frequency (NyqF) component. It is generally known that we cannot identify whether NyqF is positive or negative. Thus, a noninteger sample shift of an even-numbered-point signal requires handling of the shifted NyqF component. As shown in (3), the numerator of the first term when M is even, $-(-1)^{n-m}e^{-j\delta\pi}$, corresponds to the NyqF component. For example, supposing a noninteger shift of a real-valued even-numbered-point signal via (2) and (3), this complex-valued term translates the signal into the complex-valued signal, causing a contradiction. To avoid this, we neglect the negative effect by considering two alternates. One is that we substitute zero into δ of only this term. The other is that we handle only the real part of this term. We respectively call them "zero phase NyqF (ZPN)" and "real part of NyqF (ReN)" in this paper. In addition, we call the case in which the complex value of the term is directly used "CxN" to discriminate them.

2) Relationship to Spherical Harmonics Expansion: This sound field interpolation has a close relationship with the SH expansion in the two-dimensional circle circumference, i.e., the CH expansion [19]. It is particularly meaningful to compare the rotation transform matrix (4) with the rotation matrix in the previous work by Tourbabin and Rafaely [14]. They proposed a rotation matrix using *Wigner's D* matrix in the SH domain, $\boldsymbol{R}(\theta)$, as follows:

$$\boldsymbol{R}(\theta) = \mathrm{diag}(\boldsymbol{D}_0(\theta), \ldots, \boldsymbol{D}_\Psi(\theta)) \in \mathbb{C}^{(\Psi+1)^2\times(\Psi+1)^2}, \tag{5}$$

$$\boldsymbol{D}_\psi(\theta) = \mathrm{diag}(e^{-j(-\psi)\theta}, \ldots, e^{-j\psi\theta}) \in \mathbb{C}^{2\psi+1\times 2\psi+1}, \tag{6}$$

where diag$(*)$ is the diagonal matrix whose elements are $*$, and Ψ is the SH order. On the other hand, coming back to the beginning of our derivation for sound field interpolation, i.e., the sample shift in the Fourier domain,

$$z_m(\delta) = \mathscr{F}_{\mathrm{D}}{}^{-1}\left[\mathscr{F}_{\mathrm{D}}\left[z_m\right] e^{j\Delta k}\right]. \tag{7}$$

By translating (7) to the equation with the DFT matrix $\boldsymbol{F}$, we can formulate

$$\boldsymbol{z}(\delta) = \boldsymbol{F}^{-1}\boldsymbol{E}(\Delta)\boldsymbol{F}\boldsymbol{z}, \tag{8}$$

where $\boldsymbol{E}(\Delta) = \mathrm{diag}(e^{j\Delta\lceil -(M-1)/2\rceil}, \ldots, e^{j\Delta\lceil (M-1)/2\rceil})$ is the matrix representing phase rotation for the δ-sample shift, and $\lceil * \rceil$ indicates the ceiling function. By multiplying the matrix $\boldsymbol{F}$ from the left, we obtain the following equation:

$$\boldsymbol{F}\boldsymbol{z}(\delta) = \boldsymbol{E}(\Delta)\boldsymbol{F}\boldsymbol{z}. \tag{9}$$

Multiplying multichannel observation vectors, $\boldsymbol{z}(\delta)$ and $\boldsymbol{z}$, by $\boldsymbol{F}$ constitutes the CH transform of the sound field. Therefore, our sound field interpolation method is equivalent to multiplying the observation by $\boldsymbol{E}(\Delta)$ in the CH domain. Here, we raise a concrete example with the CH order $\Psi = 2$, i.e., the number of microphones $M = 5$. At this time, the rotation matrix in [14] corresponding to the second order is transcribed as

$$\mathrm{diag}(e^{2j\theta}, e^{j\theta}, 1, e^{-j\theta}, e^{-2j\theta}), \tag{10}$$

which is equal to the phase rotation matrix $\boldsymbol{E}(\Delta)$ above. In addition, (8) shows the diagonalization of the rotation transform matrix $\boldsymbol{U}(\Delta) = \boldsymbol{F}^{-1}\boldsymbol{E}(\Delta)\boldsymbol{F}$ by the DFT matrix, which implies that $\boldsymbol{U}(\Delta)$ is a cyclic matrix.

D. Versatility

Although in this paper we demonstrate only beamforming as downstream processing, (4) shows the versatility of the proposed method. That is, if the multichannel observation by the CMA is feasible, any downstream ASP is applicable, e.g., source separation, DOA estimation, and sound source localization. As one example, in [35], the self-rotation angle of the CMA is localized on the basis of our method. Moreover, another advantage is the adaptation to rapid rotation. In conventional approaches for dynamic scenarios, gradual temporal changes of environments are assumed, making it difficult for them to adapt to abrupt sensor movement. In contrast, the proposed method performs sound field interpolation by utilizing the rotation angle information at each frame, enabling it to effectively accommodate even rapid sensor movement. This advantage enables the extension from batch processing to online processing.

III. Source Enhancement With Sound Field Interpolation

A. Problem Setting

Although the sound field interpolation is available for various types of multichannel signal processing, such as blind source separation and source localization, this paper focuses on beamforming as the aim of the interpolation. We consider the following condition as described in Section I.

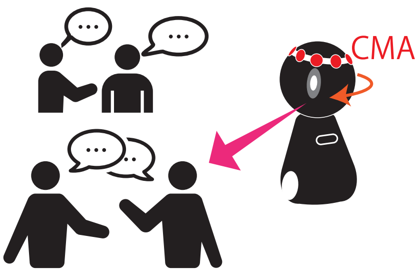

Fig. 4. A situation in which sound field interpolation may be applied to beamforming: sources are stationary, the CMA rotates, and the steering vector of a target source and the rotation angle are given.

1) Multiple sound sources do not move, and the CMA rotates. An example of this situation is when a humanoid robot or human wearing a CMA on the head rotates the head to listen to ambient conversations carefully while the speakers talk without moving, as shown in Fig. 4.
2) The steering vector of the target sound source observed by the CMA at the reference position, $\boldsymbol{a}_f$, is given, where f indicates the frequency index. The steering vector is often assumed to be obtained or estimated in advance in beamforming research [6], [36], [37].
3) The rotation angle for every time frame, θ_t, is given, where t indicates the time frame index. We can obtain the information using another module, e.g., a CMA with an acceleration sensor, such as an inertial measurement unit [38], hardware measurement by an outer camera(s), or some sensor localization scheme [35], [39], [40].

Under the above conditions, we design two methodologies for applying the sound field interpolation to beamforming. One is the use of a pre-estimated spatial filter (Section III-B), The other is online spatial filtering (Section III-C). We assume the use of the minimum power distortionless response beamformer (MPDR-BF) [41]. In the following sections, we call the observation by the CMA in the reference position without rotation the "reference observation" and assume the CMA is at the reference position at the start time.

We define the time–frequency multichannel observation by STFT as the sound field and formulate it using the spectrum x_{mtf} as follows:

$$\boldsymbol{x}_{tf} = \begin{bmatrix} x_{0tf} & \cdots & x_{(M-1)tf} \end{bmatrix}^{\mathsf{T}}, \tag{11}$$

where $m\ (0, \ldots, M-1)$, $t\ (0, \ldots, T-1)$, and $f\ (0, \ldots, F-1)$ are the microphone, time frame, and frequency bin indexes, respectively.

B. Batch Processing With Predesigned Spatial Filter

In this process, we always use a fixed spatial filter $\boldsymbol{w}_f$ that is estimated in advance. In other words, we require the reference observation with a long enough time to estimate the SCM $\boldsymbol{V}_f$ under the assumption that sound sources and a CMA do not move during all that time. Here, the MPDR-BF is formulated as

$$\boldsymbol{w}_f = \frac{\boldsymbol{V}_f^{-1}\boldsymbol{a}_f}{\boldsymbol{a}_f^{\mathsf{H}}\boldsymbol{V}_f^{-1}\boldsymbol{a}_f}, \tag{12}$$

where $\boldsymbol{V}_f = \mathbb{E}[\boldsymbol{x}_{tf}\boldsymbol{x}_{tf}^{\mathsf{H}}]$ is calculated from time-averaged $\boldsymbol{x}_{tf}\boldsymbol{x}_{tf}^{\mathsf{H}}$, assuming the ergodic process.

Before beamforming, estimating the reference observation $\hat{\boldsymbol{x}}_{\mathrm{ref},tf}$ by the following equation enables the direct use of the pre-estimated spatial filter.

$$\hat{\boldsymbol{x}}_{\mathrm{ref},tf} = \boldsymbol{U}(-\theta_t)\boldsymbol{x}_{tf}, \tag{13}$$

which means that the interpolation along the inverse rotation can put the rotated CMA back into the reference position. Then we can enhance the target source by using the filter (12) and the reference observation as usual:

$$y_{tf} = \boldsymbol{w}_f^{\mathsf{H}}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}. \tag{14}$$

Another viewpoint in the pre-estimated beamformer: Beamforming with sound field interpolation by (14) results in filtering of the interpolated observation. At the same time, (14) indicates that we can obtain a spatial filter at a CMA position different from the reference position using the interpolation, as follows:

$$y_{tf} = \boldsymbol{w}_f^{\mathsf{H}}\Big(\boldsymbol{U}(-\theta)\boldsymbol{x}_{tf}\Big) = \boldsymbol{w}_f^{\mathsf{H}}(-\theta)\boldsymbol{x}_{tf}. \tag{15}$$

The interpolated filter $\boldsymbol{w}_f(-\theta)$ is expanded as

$$\begin{aligned}\boldsymbol{w}_f(-\theta) &= \boldsymbol{U}^{\mathsf{H}}(-\theta)\boldsymbol{w}_f = \boldsymbol{U}^{\mathsf{H}}(-\theta)\frac{\boldsymbol{V}_f^{-1}\boldsymbol{a}_f}{\boldsymbol{a}_f^{\mathsf{H}}\boldsymbol{V}_f^{-1}\boldsymbol{a}_f}\\ &= \boldsymbol{U}^{\mathsf{H}}(-\theta)\frac{\boldsymbol{V}_f^{-1}\boldsymbol{U}(-\theta)\boldsymbol{a}_f(\theta)}{\boldsymbol{a}_f(\theta)^{\mathsf{H}}\boldsymbol{U}^{\mathsf{H}}(-\theta)\boldsymbol{V}_f^{-1}\boldsymbol{U}(-\theta)\boldsymbol{a}_f(\theta)}\\ &= \frac{(\boldsymbol{U}(\theta)\boldsymbol{V}_f\boldsymbol{U}^{\mathsf{H}}(\theta))^{-1}\boldsymbol{a}_f(\theta)}{\boldsymbol{a}_f^{\mathsf{H}}(\theta)(\boldsymbol{U}(\theta)\boldsymbol{V}_f\boldsymbol{U}^{\mathsf{H}}(\theta))^{-1}\boldsymbol{a}_f(\theta)}\\ &= \frac{\boldsymbol{V}_f^{-1}(\theta)\boldsymbol{a}_f(\theta)}{\boldsymbol{a}_f^{\mathsf{H}}(\theta)\boldsymbol{V}_f^{-1}(\theta)\boldsymbol{a}_f(\theta)},\end{aligned} \tag{16}$$

where $\boldsymbol{V}_f(\theta) = \boldsymbol{U}(\theta)\boldsymbol{V}_f\boldsymbol{U}^{\mathsf{H}}(\theta)$ and $\boldsymbol{a}_f(\theta) = \boldsymbol{U}(\theta)\boldsymbol{a}_f$. Note that the inversion of $\boldsymbol{U}(\theta)$ is equivalent to $\boldsymbol{U}(-\theta)$, and $\boldsymbol{U}(\theta)$ is unitary. This formula might imply that interpolation of the observation and that of the steering vector are identified theoretically. One possibility based on this discussion is that the steering vector at the reference position and the sound field interpolation may enable the estimation of a steering vector of a sound source after rotation. Although this is not the main focus of this paper, it may be a topic of future work.

C. Online Processing of Spatial Filter

Although the batch processing described in Section III-B can be a powerful solution when the ATS, except for the CMA rotation, is stationary, it is rare to satisfy this condition strictly in the real world. Considering the situation in which the ATS changes slightly, we advocate online processing with sound field interpolation for beamforming and design an updated algorithm in this section. We expect online processing to enable dealing with the slight variation of the ATS except for the CMA rotation. We introduce a well-known smoothing (forgetting) factor [42] to update the SCM in the online processing. In addition, we use a matrix inversion lemma, the Sherman–Morrison formula, which can reduce the complexity of calculating the covariance inversion that appears in the MPDR-BF formulation.

Algorithm 1: Online Beamforming Update Algorithm With Sound Field Interpolation.

```
Input: $\boldsymbol{x}_{tf} \in \mathbb{C}^{M\times 1}$, $\boldsymbol{a}_f \in \mathbb{C}^{M\times 1}$, $\theta_t \in \mathbb{R}$
Output: $y_{tf}$
  for $f = 0 : F-1$ do
    initialize $\hat{\boldsymbol{V}}_f{}^{-1}$
  end for
  for $f = 0 : F-1$ do
    $\hat{\boldsymbol{x}}_{\mathrm{ref},tf} \leftarrow \boldsymbol{U}(-\theta_t)\boldsymbol{x}_{tf}$
    $\hat{\boldsymbol{V}}_{tf}{}^{-1} \leftarrow$
    $\frac{1}{\alpha}\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1} - \frac{\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1}}{\frac{\alpha^2}{1-\alpha} + \alpha\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}}$
    $\boldsymbol{w}_{tf} \leftarrow \frac{\boldsymbol{V}_{tf}{}^{-1}\boldsymbol{a}_f}{\boldsymbol{a}_f^{\mathsf{H}}\boldsymbol{V}_{tf}{}^{-1}\boldsymbol{a}_f}$
    $y_{tf} = \boldsymbol{w}_{tf}^{\mathsf{H}}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}$
  end for
```

Firstly, we estimate the reference observation from the observation by (13), as in the batch processing. By using the interpolated observation, we can estimate the SCM at the tth frame from that of the $(t-1)$th frame, $\hat{\boldsymbol{V}}_{(t-1)f}$, and the smoothing factor α as follows:

$$\hat{\boldsymbol{V}}_{tf} = \alpha\hat{\boldsymbol{V}}_{(t-1)f} + (1-\alpha)\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}. \tag{17}$$

Such a formulation with the smoothing factor has often been seen in various studies for online signal processing [43], [44]. Furthermore, the Sherman–Morrison formula enables the calculation of the inversion of $\hat{\boldsymbol{V}}_{tf}$ with low complexity and its application to MPDR beamforming (12) in every time frame. Its inversion is formulated as

$$\begin{aligned}&\hat{\boldsymbol{V}}_{tf}{}^{-1}\\ &= \frac{1}{\alpha}\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1} - \frac{(\alpha\hat{\boldsymbol{V}}_{(t-1)f})^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}(\alpha\hat{\boldsymbol{V}}_{(t-1)f})^{-1}}{\frac{1}{1-\alpha} + \hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}(\alpha\hat{\boldsymbol{V}}_{(t-1)f})^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}}\\ &= \frac{1}{\alpha}\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1} - \frac{\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1}}{\alpha^2/(1-\alpha) + \alpha\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\hat{\boldsymbol{V}}_{(t-1)f}{}^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}}.\end{aligned} \tag{18}$$

After that, updating the spatial filter using the inversion matrix enhances the target source online. Algorithm 1 illustrates the whole pseudocode summarizing the above formulation for framewise online processing. Note that $\hat{\boldsymbol{V}}_{tf}{}^{-1}$ can be initialized with, for example, a random matrix, the identity matrix, or the inversion of $\boldsymbol{x}_{tf}\boldsymbol{x}_{tf}^{\mathsf{H}}$ averaged in the first several time frames.

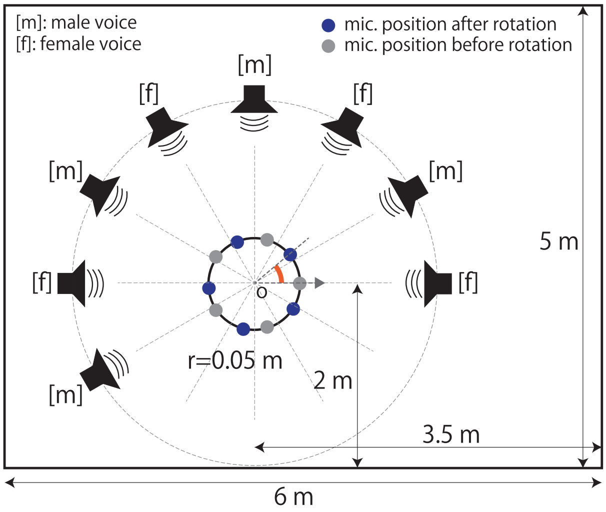

Fig. 5. Experimental environment for numerical simulation.

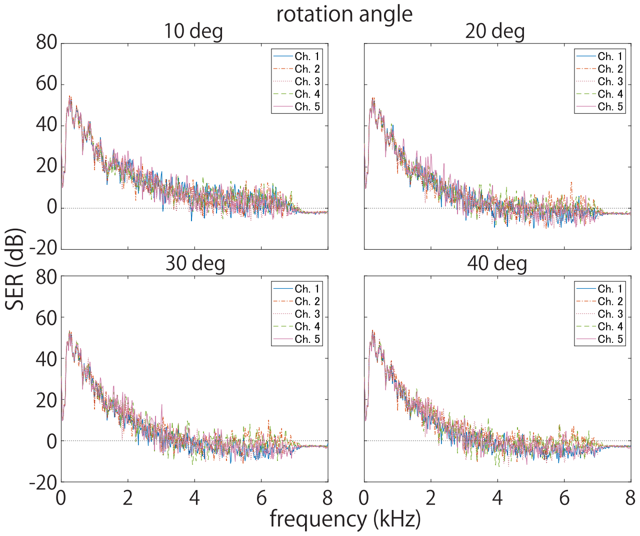

Fig. 6. Examples of SERs as a function of frequency, where the number of microphones $M = 5$.

IV. Numerical Evaluation

A. Experimental Condition

We conducted computational simulations to evaluate the performance of the proposed sound field interpolation and its influence on ASP. We used eight samples (four female and four male voices) with the sampling rate of $16\,\mathrm{kHz}$ from the SiSEC database [45] as anechoic sound sources. We generated observed signals that are convolutive mixtures of room impulse responses (RIRs) simulated by the RIR generator [46] on the basis of the image method [47]. In this environment, the reverberation time RT60 was approximately $100\,\mathrm{ms}$. We used such a small RT60 for the conceptual confirmation of the proposed method, although it is not theoretically affected by reverberation. We mixed two arbitrary sources selected from among eight sources from different directions so that the angle between two sources is 30, 60, ..., 180 deg, as shown in Fig. 5. In this manner, we simulated twelve environments (two patterns at each of six angles). We simulated the signals with the equally spaced M-channel CMA with a radius of $0.05\,\mathrm{m}$. We set the reference position of the CMA such that the first channel microphone is in the positive direction on the horizontal axis. Also, we simulated the same sound field with the CMA rotated Δ rad. Then, we estimated observation signals at the reference position from signals obtained after rotation using the rotational angle $\phi = \Delta\pi/180$ deg as a known value. For analysis, we conducted the STFT using a 1/8-shifted Hamming window with a length of $64\,\mathrm{ms}$. We performed two simulations. First, we evaluated the sound field interpolation performance from the SER defined as

$$\mathrm{SER}_{mf} = 10\log_{10}\left(\frac{\sum_t |x_{mtf}|^2}{\sum_t |\hat{x}_{mtf} - x_{mtf}|^2}\right), \qquad (19)$$

where x_{mtf} is the mth-channel STFT complex-valued spectrum at the tth time frame and fth frequency bin, and $\hat{x}_{mtf}$ is its estimate. We set the number of microphones M in the range of 3–8 and varied the rotation angle ϕ. We evaluated the case of one source, that is, we did not mix sound sources in the first simulation.

Second, we evaluated the source enhancement performance with an MPDR-BF [41] after interpolation using the signal-to-distortion ratio (SDR) and source-to-interference ratio (SIR) [48] in two ways as described in Section III. The former is batch processing described in Section III-B, and the latter is online processing in Section III-C. In the former experiment, we set $M = 5$ and ϕ as 10, 20, 30, 36, and 40 deg. The rotational angle of 36 deg corresponds to a 0.5 sample shift along the circle because the angle between the two microphones is 72 deg. In the latter experiment, we set $M = 5$ and 8, used sound sources with a long time length, and changed the CMA position twice halfway through playing the sound. The detailed setup will be explained in the following section (Section IV-E1). In either case, to estimate the filter, we used the relative transfer function (RTF) [49] calculated using RIRs from the target source to each microphone.

In addition to the above, we evaluated the robustness against microphone positions. Sound field interpolation requires an equally spaced CMA, while there are some cases where microphones are difficult to locate at equal spaces completely. In this analysis, we considered the microphone position perturbation as the difference from equally spaced positions along the same circumference, which follows the zero-mean Gaussian distribution. The details are shown in Section IV-D.

B. Interpolation Accuracy

Fig. 6 shows some examples of SERs when $M = 5$. As shown in these examples, the proposed method estimated the lower frequency component, as expected. Such a component can roughly satisfy the assumption mentioned in Section II-A. In comparison, it was not easy to estimate the higher frequency component. There also was no variance among channels.

To simplify the analysis, we restricted the frequency range to 0–3 kHz in SER and averaged the SERs in decibels. Fig. 7 shows the dependence of the SER on the rotation angle, where the

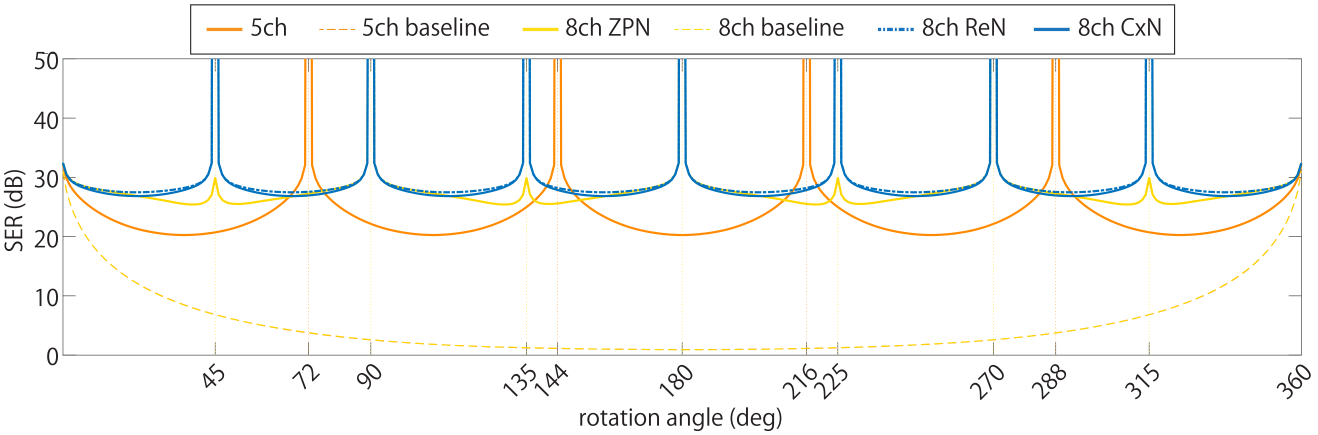

Fig. 7. Anglewise SER averaged in the frequency range up to 3 kHz and the eight environments, where "baseline" indicates that in the cases without interpolation, and ZPN, ReN, and CxN indicate the patterns when considering only the sign, only the real part, and the complex value of the NyqF component in the sound field interpolation, respectively. Note that the 5ch and 8ch baselines completely overlapped.

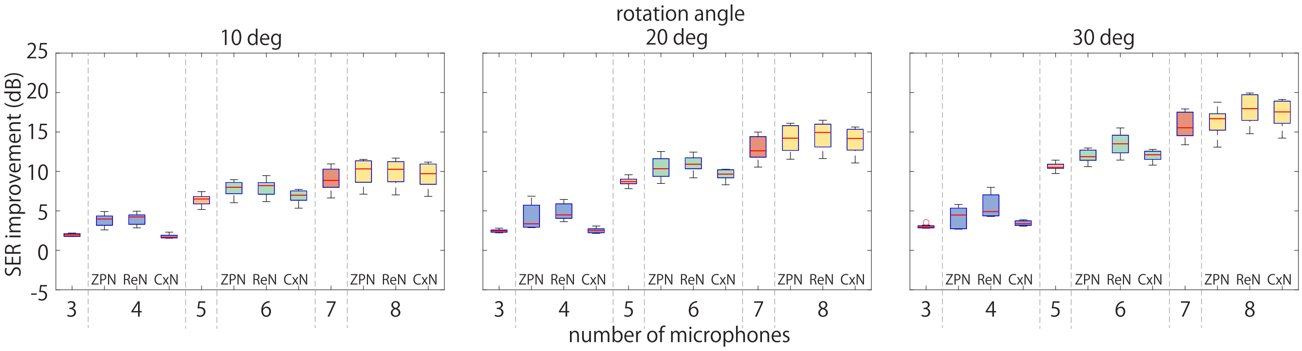

Fig. 8. Boxplots of total SER improvement in the frequency range up to 3 kHz in the eight environments relative to the cases without interpolation, where ZPN, ReN, and CxN indicate the patterns when considering only the sign, only the real part, and the complex value of the NyqF component in the sound field interpolation, respectively.

vertical axis is the average of the SERs in the eight environments with $M = 5$ and 8. Also, the baseline illustrates the SER without interpolation. As shown, we find that the behavior is periodic. This is because the CMA rotation by the angle between adjacent microphones merely shifts the microphone index, e.g., when $M = 5$ and the rotation is 72 deg, the corresponding permutation is

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 2 & 3 & 4 \end{pmatrix}. \tag{20}$$

This fact reflects that the rotation transform matrix $\boldsymbol{U}$ becomes an M-cyclic permutation matrix in such a case. Therefore, the SER becomes high for the rotation angle of $72, 144, \ldots, 288$ deg when $M = 5$ and $45, 90, \ldots, 315$ deg when $M = 8$. Moreover, the SERs with ZPN, ReN, and CxN when $M = 8$ are the same at 0, 90, 180, and 270 deg, but they are very different at 45, 135, 225, and 315 deg. We can interpret this difference using the formulation of interpolation (3). As described in Section II-C1, in the ZPN case, the term of NyqF, $\exp(j\delta\pi)$, in (3) is 1 to avoid the impact of the complex value. As a result, this procedure prevents $\boldsymbol{U}$ from being a cyclic permutation. However, even if such a procedure is done for the ZPN case, the formulation is coincident with the theoretical equation because $\exp(j\delta\pi)$ becomes 1 at 90, 180, and 270 deg. It might be unnecessary to neglect NyqF components when we consider the interpolation of the complex value such as an STFT spectrum even if M is even. Note that ZPN does not always degrade the interpolation accuracy more than CxN in any case, e.g., the SERs with ZPN in the range of 1–22 deg are slightly higher. This result implies that handling the NyqF component, even if only a little, affects the interpolation accuracy at around 0, 90, 180, and 270 deg. In comparison with ZPN and CxN, ReN has a slight superiority at any angle.

Fig. 8 shows channelwise SER improvements in the frequency range up to 3 kHz relative to the case without sound field interpolation. It is considered that using more microphones would result in better performance because the spatial sampling rate would increase. However, interestingly, these results illustrate that using more microphones does not always improve the SER, as in the case of changing the number of microphones M from 3 to 4 in the CxN case, i.e., when the NyqF component is considered. As explained in Section II-C1, in ZPN, the NyqF component does not contribute to the interpolation, except at

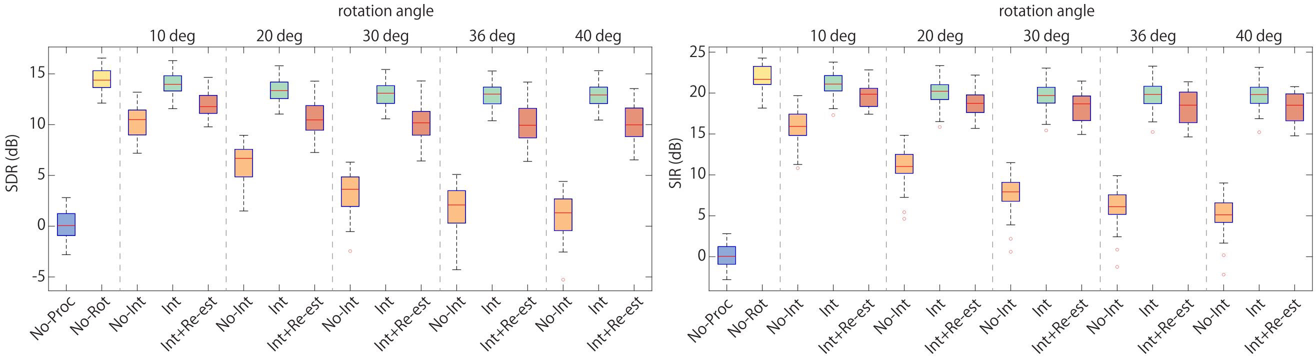

Fig. 9. Boxplots of SDR and SIR obtained by 5ch MPDR-BF for five cases: unprocessed (**No-Proc**), no rotation of CMA (**No-Rot**), without interpolation when CMA rotates (**No-Int**), with interpolation when CMA rotates (**Int**), and re-estimation of the filter after interpolation when CMA rotates (**Int+Re-est**).

the angle where $\exp(j\delta\pi) = 1$. When M is even, one Fourier component (NyqF component) is wasted for interpolation. For example, in the ZPN case with $M = 4$, one of four components, 25% of the information, is wasted. Therefore, using odd M is sometimes more efficient than using even M in our method if the available number of microphones is three or four. By contrast, ReN always improves the SER with increasing M. From these results, ReN is judged to be effective for sound field interpolation, and we use ReN in the following evaluation when M is even.

C. Source Enhancement with Batch Processing

First, we produced the MPDR-BF filter $\boldsymbol{w}$ using the RTF and the multichannel STFT spectrogram observed with the CMA at the reference location (no rotation). We applied $\boldsymbol{w}$ to the following three spectrograms: without CMA rotation (**No-Rot**), without interpolation when the CMA rotates (**No-Int**), and with interpolation when the CMA rotates (**Int**). Also, we applied another MPDR-BF calculated from the same steering vector and the interpolated spectrogram to the interpolated spectrogram (**Int+Re-est**). We can use this method to predict the performance of the online beamforming described in the next section. We also used the unprocessed case (**No-Proc**), i.e., the observation, and **No-Rot** as baselines and compared them with the other three cases. **No-Rot** shows the best performance of the MPDR-BF when the ATS does not change.

Fig. 9 shows that small changes in ATS greatly affect the ASP performance, and the proposed method (**Int**) outperforms the case without interpolation (**No-Int**) and comes close to the best performance (**No-Rot**). The degradation in the **No-Int** case along the rotational angle resembles the SER curve of "5ch baseline" in the range of 0–40 deg in Fig. 7. This shows that the distortion of the observation itself directly affects the SDR and SIR. One of the reasons why the **Int+Re-est** case did not perform worse than the **Int** case is the mismatch between the SCM estimated from the interpolated spectrogram and the pre-estimated steering vector. The mismatch interferes with the production of the spatial filter that supresses the interfering signal and degrades the enhancement performance. Because such a mismatch might occur in the case of the extension of online processing, we expect its performance also to be similar to the performance of online processing. The proposed method could not estimate the high-frequency component, as shown in SER results (Fig. 6), but significantly improved the ASP performance, especially in batch processing.

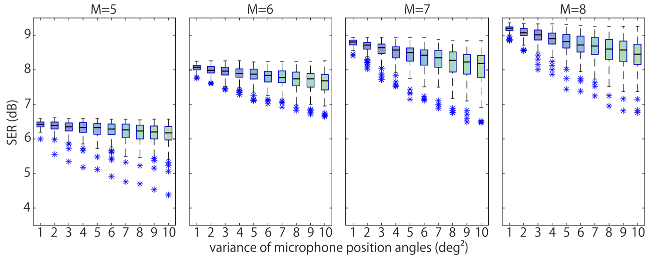

Fig. 10. Effect of SER from microphone position perturbation when the rotation angle is 10 deg, where the horizontal axis indicates the variance of the zero-mean Gaussian, and a blue-colored $*$ shows an outlier defined as a value exceeding 1.5 times the interquartile range away from the bottom or top of the box.

D. Robustness Evaluation

In this evaluation, we gave the microphone position perturbation as the difference from equally spaced microphone position and confirmed the robustness of sound field interpolation and downstream ASP against the perturbation. In this experiment, we assumed that the position error of each microphone, ϵ_m, follows the zero-mean Gaussian distribution, $\epsilon_m \sim \mathcal{N}(0, \varsigma^2)$. We conducted 100 trials by varying the variance ς^2 from 1 to 10 deg^2; for example, $\varsigma^2 = 10$ means that the perturbation within 3ς (≈ 10 deg) occupied by 99.7% from the property of Gaussian distribution. Fig. 10 shows the boxplots of SERs obtained with $M = 5, 6, 7, 8$ when a sound source is present, and the rotation angle is 10 deg. The horizontal axis is ς^2. As shown in the figure, a perturbation of several degrees affects the performance, although the degradation of SER was within approximately 1 dB when position errors were small. We can also find that the larger the number of microphones, M, the easier for SER to degrade. The conceivable reasons for this observation include the fact that a smaller angle between adjacent microphones is sensitive to perturbation, or the larger number of microphones results in a larger total amount of perturbation. Table I shows the SDR variations when M s are 5 and 8, and two sound sources are present. The values 1.0, 2.0, and 3.0 indicate the standard deviation ς. The same trend as interpolation performance variation was

TABLE I
EFFECT OF SDR FROM MICROPHONE POSITION PERTURBATION WHEN THE ROTATION ANGLE IS 10 DEG, WHERE THE VALUES 1.0, 2.0, AND 3.0 INDICATE THE STANDARD DEVIATION OF THE ZERO-MEAN GAUSSIAN

	Standard deviation of microphone position angle (deg)		
	1.0	2.0	3.0
$M = 5$	13.87 ± 1.20	13.68 ± 1.24	13.39 ± 1.34
$M = 8$	14.18 ± 0.88	13.85 ± 0.96	13.37 ± 1.16

obtained, that is, the larger the M, the easier for SDR to degrade. These results illustrate that if the difference from equally spaced positions is small, e.g., within 1 or 2 deg, such perturbation does not significantly degrade the performance of interpolation and downstream ASP. When larger errors are considered, larger degradation can be expected. In such cases, other approaches are necessary, for example, interpolation using unequally spaced CMAs [32].

E. Online Beamforming

1) Setup: The experimental conditions are almost the same as described in Section IV-C. The different points are as follows: we used two source signals with the time length of $40\,\mathrm{s}$ and simulated the impulse response with the reverberant time of RT60=$330\,\mathrm{ms}$, which corresponds to a typical office room, in addition to $100\,\mathrm{ms}$. The position of the two sources was at the angles of $\pi/6$ and $4\pi/6$ rad with the same alignment as shown in Fig. 5. We set the frame length of 4096 samples ($256\,\mathrm{ms}$). We also used the 1 s-wise SDR, i.e., the segmental SDR, to evaluate the source enhancement performance and confirmed the effect of changing the smoothing factor α. We initialized the inversion of the SCM ${\boldsymbol{V}_{tf}}^{-1}$ as the inversion of $\boldsymbol{x}_{tf}\boldsymbol{x}_{tf}^{\mathsf{H}}$ averaged over the first 10 frames ($\approx 320\,\mathrm{ms}$). The CMA rotated twice; the first rotation of θ_1 deg occurred at 10 s, and the second one of θ_2 deg at 25 s. We also considered the two cases of $\theta_1 = 30$ and $\theta_2 = 60$ and $\theta_1 = 72$ and $\theta_2 = 0$. We generated such observations by concatenating the observations of the three different microphone arrays in the simulation. Note that 0 deg means the reference position. In addition, we designed the direction-dependent covariance (**DDC**) method as an additional baseline of the online approach. In **DDC**, the SCM is reset at the time of rotation and preserved. Then, when the CMA position moves to a revisited angle, the preserved SCM is used immediately. In other words, **DDC** is a block processing based on angle, not time, and has an anglewise dictionary for SCM.

2) Segmental SDR: Fig. 11 shows segmental SDRs in the two patterns of CMA rotation. Here, we used the smoothing factor α of 0.99 that produced the highest segmental SDR in a preliminary experiment. As shown in the top four figures ($0 \Rightarrow 72 \Rightarrow 0$ deg), **Int-online** outperforms **No-Int-online** and **No-Int-DDC-online** and is close to **No-Rot** because the interpolation is completely achieved at the rotation of 72 deg when $M = 5$, as described in the previous section. When $M = 8$, although the interpolation is not completely achieved, high performance is maintained because more microphones are used. This tendency is the same regardless of reverberation time. The difference between **Int** and **Int-online** is only the time length of observation for estimating the spatial filter; **Int-online** uses only information in the previous time frame, and therefore, it degrades SDR slightly more than **Int**. In comparison with them, the bottom four graphs show different trends for the reverberation time and the number of microphones. In the case of the rotation of $0 \Rightarrow 30 \Rightarrow 60$ deg, **Int-online** improves SDR in RT60=$100\,\mathrm{ms}$, although it cannot improve SDR in RT60 = 330 ms because of incomplete interpolation, compared with the case where interpolation is completely achieved above. The estimation accuracy of the spatial filter is also one possible reason for degradation. However, this possibility is denied by the result that **No-Rot** works well and **Int** does not. A possible reason is that room reflection sound causes the assumption of a planer wave in sound field interpolation to be invalid. This degradation is solved by using more microphones, whereas a more severe reverberation environment will need more microphones to achieve source enhancement.

Moreover, the comparison between **No-Int-online** and **No-Int-DDC-online** is interesting. Their rough trends in SDR are almost the same, that is, both SDRs remain degraded during rotation in any scenario. Considering that the difference between the two methods is only the update of the SCM, this demonstrates that even if the SCM is updated immediately, the mismatch of the SCM and the steering vector always occurs as long as the steering vector is fixed (as explained in problem setting 2) in Section III-A). In contrast, the difference is the speed of improving SDR. In the $0 \Rightarrow 72 \Rightarrow 0$ scenario, after the second rotation, which is the same as the initial position, the SCM is updated correctly after several frames, and the mismatch is resolved. Therefore, **No-Int-DDC-online** reaches the same SDR as **Int-online**, more rapidly than **No-Int-online** owing to the immediate reset of internal statistics. In comparison, the mismatch is always not resolved in the $0 \Rightarrow 36 \Rightarrow 60$ scenario, and SDR does not improve. These observations indicate the necessity of steering vector estimation for managing microphone movement situations without interpolation.

F. Discussion

In this study, we assume that there are no rigid bodies, such as reflectors, in the CMA during the evaluation. Our approach is based solely on the Fourier series expansion of the sound field on a circle. Although a rigid body may exist in the CMA, the smoothness of the sound field on the circle might not be significantly affected, particularly if the rigid body exhibits rotational symmetry, such as a sphere or a head. Therefore, our approach is expected to work well in such cases. Furthermore, we confirm that the interpolation technique proposed in this study works well, even in the presence of such rigid bodies within the CMA, in an ongoing work of realizing a real-world system of self-rotation angle estimation [35].

Possible sensor movements in 3D space include three translations along each axis and three rotations: "pitch", "yaw", and "roll". In this study, we focus on "yaw", which corresponds to a head rotation along the horizontal plane due to two factors: 1) it involves rapid motion and 2) the head may not immediately return to its original position. The translational motion of the

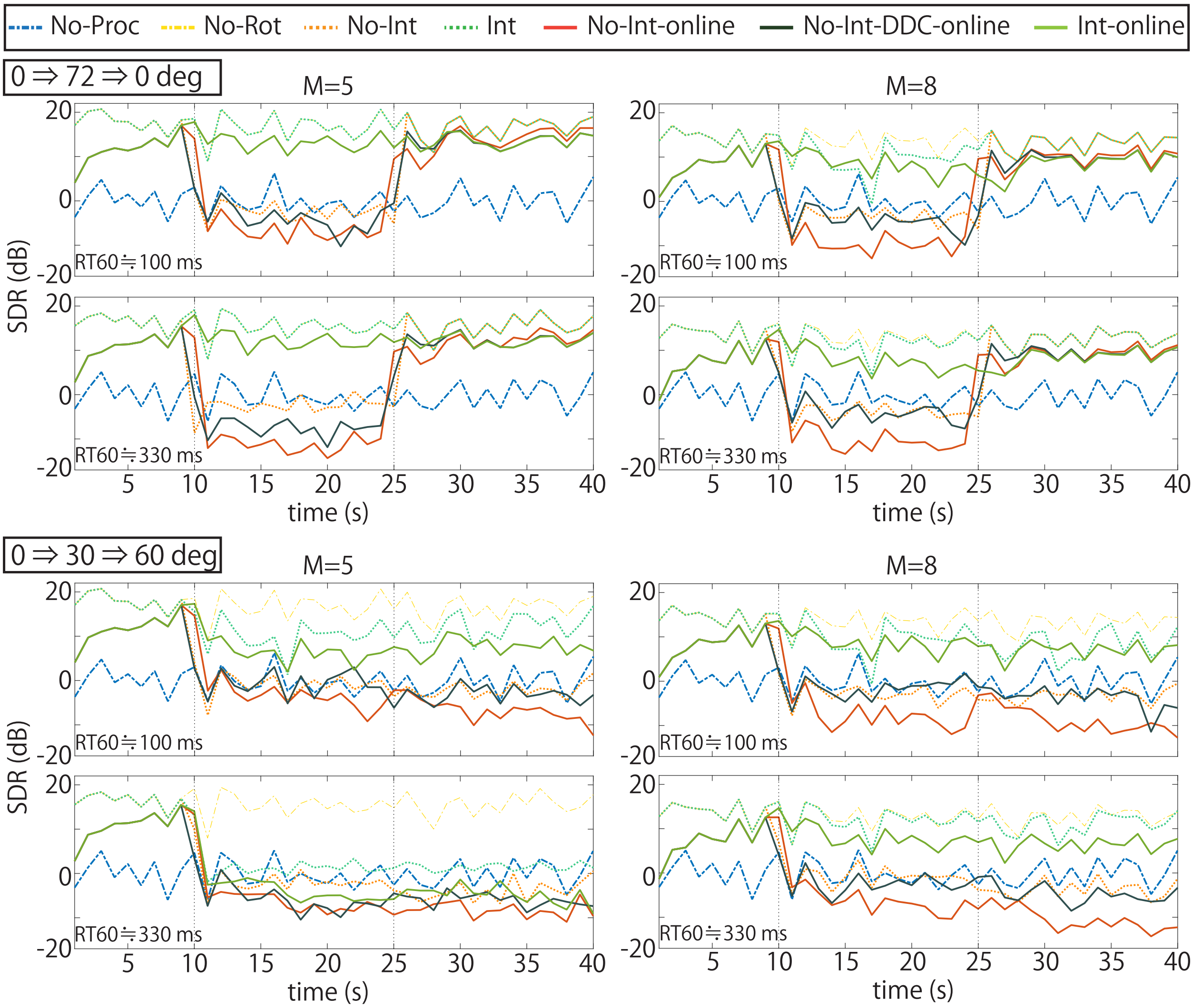

Fig. 11. Segmental SDR every 1 s with $M = 5$ and 8 and RT60 $= 100$ and 330 ms, where the two vertical dashed lines indicate the time points when the rotation occurred with the two patterns: $0 \Rightarrow 72 \Rightarrow 0$ deg and $0 \Rightarrow 30 \Rightarrow 60$ deg. **No-Proc** shows the mixture itself, **No-Rot** shows the case where rotation does not occur, **No-Int** and **Int** respectively show batch processing without/with interpolation, **DDC** indicates that the SCM is updated in direction-dependent manner, and **-online** shows online processing with the smoothing factor α of 0.99.

head along the three (but usually two) axes tends to be slower, as it occurs during activities such as walking. Additionally, "pitch" and "roll" (tilting and nodding of the head) typically result in the head rapidly returning to its initial position. Therefore, we consider that the interpolation against "yaw" is more important than that against "pitch" and "roll". A spherical microphone array might be required to manage all types of motion, which will be considered in our future studies.

V. Conclusion

We presented a new framework of beamforming robust to CMA rotation using a sound field interpolation method and applied it to batch and online beamforming. The interpolation method could virtually regard the time-variant ATS as a time-invariant one by using the periodicity of the sound field observed using the CMA and its noninteger sample shift. Experimental results illustrated that our simple method could estimate the lower band spectrum and assist the ASP, even when the CMA rotates. Future work includes improving the estimation accuracy of the higher frequency component by another approach, conducting additional experiments with different ASP methods, e.g., source separation and source localization, and confirming the performance with a real device in real environments. Although our proposed method uses the characteristics of CMA, the extension to an arbitrary array configuration could also be an important future work.

Acknowledgment

The authors would like to thank S. Luan and Prof. Toda of Nagoya University, who pointed out a misformulation.

Appendix
Derivation of Sound Field Interpolation

We consider the case where M is even. Let the Fourier coefficient of z_m $(m = 0, \ldots, M-1)$ be Z_k, $(k = -M/2 + 1, \ldots, M/2)$. Then, we formulate the sample shift using the shift theorem as follows:

$$\begin{aligned} z_m(\delta) &= z\left(2\pi\frac{m}{M}+\Delta\right) = \mathscr{F}_{\mathrm{D}}^{-1}\left[\mathscr{F}_{\mathrm{D}}\left[z_m\right]e^{j\Delta k}\right] \\ &= \frac{1}{M}\sum_{k=-M/2+1}^{M/2} Z_k e^{j\Delta k} W^{-mk}, \end{aligned} \tag{21}$$

where $W = \exp(j2\pi/M)$ is the twiddle factor. We separate the summation with respect to spatial frequency k into the terms of direct current, Nyquist frequency, and negative and positive frequency elements.

$$\begin{aligned} z_m(\delta) = \frac{1}{M}\Bigg(& Z_0 + \sum_{k=1}^{M/2-1} Z_k e^{j\Delta k} W^{-mk} \\ & + Z_{\frac{M}{2}} e^{j\Delta\frac{M}{2}} W^{-m\frac{M}{2}} \\ & + \sum_{k=-M/2+1}^{-1} Z_k e^{j\Delta k} W^{-mk}\Bigg). \end{aligned} \tag{22}$$

By using $Z_k = \sum_{n=0}^{M-1} z_n W^{nk}$, we can rewrite (22) as follows:

$$\begin{aligned} z_m(\delta) = \frac{1}{M}\Bigg\{ & \sum_{n=0}^{M-1} z_n + \sum_{k=1}^{M/2-1}\sum_{n=0}^{M-1} z_n W^{nk} e^{j\Delta k} W^{-mk} \\ & + \sum_{n=0}^{M-1} z_n W^{n\frac{M}{2}} e^{j\Delta\frac{M}{2}} W^{-m\frac{M}{2}} \\ & + \sum_{k=1}^{M/2-1} \overline{\left(\sum_{n=0}^{M-1} z_n W^{nk} e^{j\Delta k}\right) W^{mk}} \Bigg\}, \end{aligned} \tag{23}$$

where the overline $\overline{*}$ shows the conjugate of $*$. By factoring out the common element $\sum_{n=0}^{M-1} z_n$, we can summarize (23) as

$$\begin{aligned} z_m(\delta) = \frac{1}{M}\sum_{n=0}^{M-1} z_n \Bigg\{ & 1 + W^{(n-m-\delta)\frac{M}{2}} \\ & + \sum_{k=1}^{M/2-1} W^{(n-m-\delta)k} + \sum_{k=1}^{M/2-1} W^{-(n-m-\delta)k} \Bigg\}. \end{aligned} \tag{24}$$

When $\delta \notin \mathbb{Z}$, let L be $n-m-\delta$ and the term in the brackets and $1/M$ be $U_{mn}(\delta)$. Then, $U_{mn}(\delta)$ is formulated as

$$\begin{aligned} & U_{mn}(\delta) \\ &= \frac{1}{M}\left(1 + W^{L\frac{M}{2}} + \sum_{k=1}^{M/2-1} W^{Lk} + \sum_{k=1}^{M/2-1} W^{-Lk}\right) \\ &= \frac{1}{M}\left(1 - W^{-L\frac{M}{2}} + W^{L}\frac{1-W^{L\frac{M}{2}}}{1-W^{L}} + W^{-L}\frac{1-W^{-L\frac{M}{2}}}{1-W^{-L}}\right) \\ &= \frac{1}{M}\left(1-e^{jL\pi}\right) + \frac{\mathrm{sinc}\Big(L/2\Big)}{\mathrm{sinc}\Big(L/M\Big)} \cdot \cos\left(\frac{M+2}{2\,M}L\pi\right). \end{aligned} \tag{25}$$

Just as in this case, we can also formulate $U_{mn}(\delta)$ when M is odd. Note that the sign of the second term $e^{jL\pi}$ is **negative** although it was positive in our previous work [33]. We sincerely apologize for this notation error.

References

[1] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[2] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[3] N. Makishima et al., "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, Oct. 2019.

[4] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 236–240.

[5] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7908–7912.

[6] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3461–3475, 2021.

[7] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6855–6859.

[8] A. H. Moore, S. Hafezi, R. R. Vos, P. A. Naylor, and M. Brookes, "A compact noise covariance matrix model for MVDR beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2049–2061, 2022.

[9] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 281–295, Feb. 2018.

[10] S. M. Naqvi, M. Yu, and J. A. Chambers, "Multimodal blind source separation for moving sources based on robust beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 241–244.

[11] M. Taseska and E. A. P. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 657–670, Mar. 2018.

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[13] V. Valimaki, A. Franck, J. Ramo, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 92–99, Mar. 2015.

[14] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 2046–2058, Nov. 2015.

[15] R. M. Corey and A. C. Singer, "Motion-tolerant beamforming with deformable microphone arrays," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, 2019, pp. 115–119.

[16] J. Casebeer, J. Donley, D. Wong, B. Xu, and A. Kumar, "NICE-Beam: Neural integrated covariance estimators for time-varying beamformers," 2021, *arXiv:2112.04613*.

[17] A. H. Moore, R. R. Vos, P. A. Naylor, and M. Brookes, "Processing pipelines for efficient, physically-accurate simulation of microphone array signals in dynamic sound scenes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 965–969.

[18] G. Grimm, M. M. Hendrikse, and V. Hohmann, "Interactive rendering of dynamic virtual audio-visual environments for "subject-in-the-loop" experiments," *J. Acoustical Soc. America*, vol. 146, no. 4, p. 2801, 2019.

[19] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition, Ser. Lecture Notes in Control and Information Sciences*. Berlin, Germany: Springer, 2007.

[20] D. P. Jarrett, E. A. P. Habets, J. Benesty, and P. A. Naylor, "A tradeoff beamformer for noise reduction in the spherical harmonic domain," in *Proc. IEEE 10th Int. Workshop Acoustic Signal Enhancement*, 2012, pp. 1–4.

[21] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "Spherical harmonic domain noise reduction using an MVDR beamformer and DOA-based second-order statistics estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 654–658.

[22] Q. Huang, L. Zhang, and Y. Fang, "Performance analysis of low-complexity MVDR beamformer in spherical harmonics domain," *Signal Process.*, vol. 153, pp. 153–163, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168418302482

[23] G. Huang, J. Chen, and J. Benesty, "Insights into frequency-invariant beamforming with concentric circular microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2305–2318, Dec. 2018.

[24] S. Yan, "Robust time-domain broadband modal beamforming for circular arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 3, pp. 1783–1794, Jun. 2020.

[25] A. H. Moore, W. Xue, P. A. Naylor, and M. Brookes, "Noise covariance matrix estimation for rotating microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 519–530, Mar. 2019.

[26] T. Okamoto, "Close-talking recording with planarly distributed microphones," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 4470–4474.

[27] A. M. Torres, M. Cobos, B. Pueo, and J. J. Lopez, "Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays," *J. Acoustical Soc. America*, vol. 132, no. 3, pp. 1511–1520, 2012.

[28] N. Ueno, S. Koyama, and H. Saruwatari, "Kernel ridge regression with constraint of helmholtz equation for sound field interpolation," in *Proc. IEEE 16th Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 1–440.

[29] K. Yamaoka, R. Scheibler, N. Ono, and Y. Wakabayashi, "Sub-sample time delay estimation via auxiliary-function-based iterative updates," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, 2019, pp. 130–134.

[30] C. Schüldt, "Trigonometric interpolation beamforming for a circular microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 431–435.

[31] H. Barfuss, M. Bachmann, C. Huemmer, and W. Kellermann, "Exploiting microphone array symmetry for robust two-dimensional polynomial beamforming," in *Proc. IEEE 16th Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 486–490.

[32] S. Luan, Y. Wakabayashi, and T. Toda, "Modified sound field interpolation method for rotation-robust beamforming with unequally spaced circular microphone array," in *Proc. IEEE 30th Eur. Signal Process. Conf.*, 2022, pp. 344–348.

[33] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Rotation-robust beamforming based on sound field interpolation with regularly circular microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 771–775.

[34] D. L. Alon and B. Rafaely, "Spatial aliasing-cancellation for circular microphone arrays," in *Proc. IEEE 4th Joint Workshop Hands-free Speech Commun. Microphone Arrays*, 2014, pp. 137–141.

[35] L. Guansan, Y. Wakabayashi, K. Yamaoka, and N. Ono, "Self-rotation angle estimation of circular microphone array based on sound field interpolation," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 1016–1020.

[36] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *Proc. IEEE 16th Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 146–150.

[37] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[38] A. H. Moore, L. Lightburn, W. Xue, P. A. Naylor, and M. Brookes, "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. IEEE 16th Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 461–465.

[39] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of Ad-Hoc arrays using time difference of arrivals," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 1018–1033, Feb. 2016.

[40] R. C. Felsheim, A. Brendel, P. A. Naylor, and W. Kellermann, "Head orientation estimation from multiple microphone arrays," in *Proc. IEEE 28th Eur. Signal Process. Conf.*, 2021, pp. 491–495.

[41] H. L. V. Trees *Optimum Array Processing*. Hoboken, NJ, USA: Wiley, 2002.

[42] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[43] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. IEEE 4th Joint Workshop Hands-free Speech Commun. Microphone Arrays*, 2014, pp. 107–111.

[44] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*. 2017, pp. 216–220.

[45] S. Araki et al., "The 2011 signal separation evaluation campaign (SiSEC2011): - audio source separation," in *Proc. 10th. Int. Conf., LVA/ICA*, 2012, pp. 414–422.

[46] E. A. P. Habets, "Room impulse response (RIR) generator," 2008. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

[47] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. America*, vol. 65, pp. 943–950, 1979.

[48] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[49] D. Simon, K. Walter, M. Shoji, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

Yukoh Wakabayashi received the B.E. and M.E. degrees from Osaka University, Osaka, Japan, in 2008 and 2010, respectively, and the Ph.D. degree from Ritsumeikan University, Shiga, Japan, in 2017. In 2010, he joined Rohm Incorporate, Kyoto, Japan. From 2012 to 2014, he was an Assistant Researcher with Kyoto University, Kyoto. From 2018 to 2020, he was an Affiliate Assistant Professor with Ritsumeikan University, Kyoto. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Japan, and the Faculty of Systems Design, Tokyo Metropolitan University, Hino, Japan. His research interests include acoustic signal processing, speech phase processing, array signal processing, and speaker diarization. From 2016 to 2017, he was the recipient of the JSPS Research Fellowship for Young Scientists DC2. He is a Member of the Institute of Electrical and Electronics Engineers, Institute of Electronics, Information and Communication Engineers, and Acoustical Society of Japan.

Kouei Yamaoka (Student Member, IEEE) received the B.Sc. in information engineering and M.E. degrees in engineering from the University of Tsukuba, Tsukuba, Japan, in 2017 and 2019, respectively. He is currently working toward the Ph.D. degree with Tokyo Metropolitan University, Hino, Japan. His research interests include acoustic signal processing, signal enhancement, source localization, and asynchronous distributed microphone array. He is a Member of the Acoustical Society of Japan.

Nobutaka Ono (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo, Tokyo, Japan, in 1996, 1998, and 2001, respectively.

He was a Research Associate with the University of Tokyo in 2001, and became a Lecturer in 2005. He was also an Associate Professor with the National Institute of Informatics, Tokyo, Japan, in April 2011, and became a Professor in 2017. In 2017, he was with Tokyo Metropolitan University, Hino, Japan. He is the author or co-author of more than 280 articles in international journal papers and peer-reviewed conference proceedings. His research interests include acoustic signal processing, especially microphone array processing, source localization and separation, machine learning, and optimization algorithms. He was a Tutorial Speaker with ISMIR 2010 and ICASSP 2018.

Dr. Ono is a Senior Member of IEEE Signal Processing Society and a Member of the Acoustical Society of Japan (ASJ), Institute of Electronics, Information and Communications Engineers, Information Processing Society of Japan, and Society of Instrument and Control Engineers (SICE), Tokyo, Japan. He was the Chair of Signal Separation Evaluation Campaign evaluation committee in 2013 and 2015, a Technical Program Chair of IWAENC 2018, General Chair of DCASE 2020 workshop, and Member of IEEE Audio and Acoustic Signal Processing Technical Committee from 2014 to 2019. From 2012 to 2015, he was an Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is currently the Vice Chair of IEEE Signal Processing Society Tokyo Joint Chapter. He was the recipient of the Awaya Award from ASJ in 2007, Igarashi Award at the Sensor Symposium from IEEJ in 2004, Best Paper Award at IEEE ISIE in 2008, the Measurement Division Best Paper Award from SICE in 2013, Best Paper Award in IEEE IS3C in 2014, Excellent Paper award in IIHMSP in 2014, Unsupervised Learning ICA pioneer Award from SPIE, DSS in 2015, Sato Paper Award from ASJ in 2000 and 2018, two TAF Telecom System Technology Awards in 2018, and Best Paper Award in APSIPA ASC in 2018.