APSIPA Transactions on Signal and Information Processing, 2024, 13, e5

Original Paper

Self-Rotation-Robust Online-Independent Vector Analysis with Sound Field Interpolation on Circular Microphone Array

Taishi Nakashima[1*], Yukoh Wakabayashi[2] and Nobutaka Ono[1]

[1] *Tokyo Metropolitan University, Tokyo, Japan*
[2] *Toyohashi University of Technology, Aichi, Japan*

ABSTRACT

In this paper, we propose an online blind source separation (BSS) method that is robust against the self-rotation of microphone arrays. Online auxiliary-function-based independent vector analysis (OIVA) is one of the promising methods for real-time BSS. One major issue of real-time BSS is robustness against the movements of sources or microphones. Parameter re-estimation is necessary if such changes occur during processing. OIVA is robust against smooth movements of sources and achieves high separation performance. However, OIVA should perform better against rapid movements of microphones. In this study, we exploit sound field interpolation (SFI) for circular microphone arrays (CMAs) with OIVA. SFI cancels out the rotation of a CMA, enabling us to apply BSS without parameter re-estimation. We propose two methods: a combination of SFI and OIVA for preprocessing and a method using parameter transformations for practical applications. Simulation experiments confirmed that SFI improves the robustness of OIVA in situations where the microphone is rotating.

*Corresponding author: Taishi Nakashima, taishi@ieee.org. This work was supported by JSPS KAKENHI Grant Number 22KJ2548 and JST CREST Grant Number JPMJCR19A3.

Received 01 September 2023; Revised 13 November 2023
ISSN 2048-7703; DOI 10.1561/116.00000163

Keywords: Blind source separation, online-independent vector analysis, circular microphone array, sound field interpolation.

1 Introduction

Blind source separation (BSS) [17] is a technique to extract source signals from their observed mixture. Popular approaches for BSS include independent vector analysis (IVA) [7, 11], auxiliary-function-based IVA (AuxIVA) [22], and their extensions [3, 12, 20]. These methods assume a time-invariant acoustic transfer system (ATS). However, in practical applications, considering time variations of the ATS, such as microphone movements, is necessary.

Multichannel acoustic signal processing techniques considering the dynamic environment have recently attracted considerable attention. Several methods using recursive parameter updates have been proposed, such as beamforming [6], direction-of-arrival tracking [32], and speaker tracking [24]. Furthermore, competitions have been organized, and datasets have been developed for speech processing in dynamic scenarios. Clarity Challenge [1] aims to improve speech intelligibility in hearing aids and includes data on the listeners' head movements. SPEAR Challenge [5] is a speech enhancement challenge for head-worn hearing devices, and extensive datasets called EasyCom [4] have been distributed, which include speaker and head movement data.

In relation to BSS, many methods have been proposed on the basis of a block batch [9, 13, 14] or online processing [10, 27] to account for environmental changes. In particular, online AuxIVA (OIVA) shows high separation performance in real-time scenarios [27]. It has also been actively researched recently for applications in hearing aids [25], joint optimization with dereverberation [28], and computationally efficient optimization [19]. OIVA estimates demixing matrices in a frame-by-frame manner and can track smooth environmental changes in ATS, such as slow movements of sources. However, rapid changes in ATS, such as the emergence of new sources or microphone movements, make online BSS difficult and thus degrade separation performance.

Several methods have been proposed to cope with such rapid changes. Sound field interpolation (SFI) for circular microphone arrays (CMAs) has been proposed to address the rotation of a CMA [30]. This method exploits the symmetry of the CMA to estimate the sound field before the rotation of the CMA by a simple linear operation. The applications of SFI to beamforming [29] and steering vector estimation [31] have also been proposed, as well as a method of self-estimating the rotation angle of a CMA [16]. We expect that the combination of OIVA and SFI will improve the robustness against the rotation of CMAs.

In this paper, we address BSS in situations where a CMA rotates. SFI cancels out the effect of the rotation, and BSS is applied in the latter stage.

As described in Section 4, a naive combination of SFI and OIVA has a problem for practical applications. In contrast, in this study, we develop a more straightforward method than this combination and demonstrate its effectiveness through experiments. Experiments show that our proposed method is significantly better than the conventional OIVA.

The rest of this paper is organized as follows. We formulate our problem in Section 2. In Section 3, we describe online BSS and SFI, and how to combine them. In Section 4, we propose a new online BSS that utilizes the information before and after the rotation of a CMA. We conducted some experiments to show the efficacy of SFI for online BSS in Section 5. In Section 6, we conclude this paper.

2 Problem Formulation

Let us consider the BSS problem with a CMA that can be horizontally rotated as shown in Figure 1. Let K and M be the numbers of sources and microphones, respectively. We assume that the observed signal $\boldsymbol{x}_{f,t}$ is in the short-time Fourier transform (STFT) domain modeled as

$$\boldsymbol{x}_{f,t} = \boldsymbol{a}_{1,f,t}s_{1,f,t} + \cdots + \boldsymbol{a}_{K,f,t}s_{K,f,t} = \sum_{k=1}^{K} \boldsymbol{a}_{k,f,t}s_{k,f,t}, \tag{1}$$

where $f = 1, \ldots, F$ denotes the frequency bin index, $t = 1, \ldots, T$ denotes the time frame index, $s_{k,f,t} \in \mathbb{C}$ $(k = 1, \ldots, K)$ denotes the kth source signal, and $\boldsymbol{a}_{k,f,t} \in \mathbb{C}^M$ denotes the steering vector of the kth source signal for each microphone. Moreover, we here set the *reference microphone* $\ell = 1$ without the loss of generality. In this case, each steering vector can be denoted as

$$\boldsymbol{a}_{k,f,t} \coloneqq \begin{bmatrix} 1 & a_{2,k,f,t} & \ldots & a_{M,k,f,t} \end{bmatrix}^{\mathsf{T}} \ (k = 1, \ldots, K). \tag{2}$$

Under this definition, $a_{m,k,f,t}$ $(m = 2, \ldots, M,\ \forall k = 1, \ldots, K)$ corresponds to the relative transfer function from the kth source to the mth microphone, and thus each source signal $s_{1,f,t}, \ldots, s_{K,f,t}$ can be regarded as the *source*

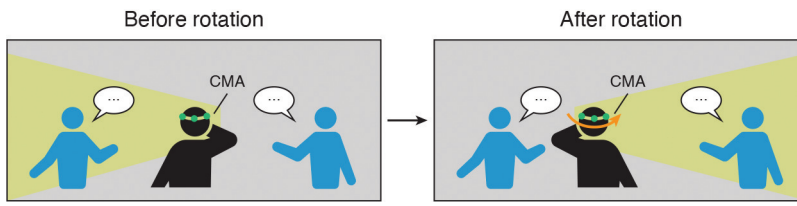


Figure 1: Overview of blind source separation problem with a CMA.

image at the reference microphone. Note that many BSS methods require a time-invariant mixing system $\boldsymbol{a}_{k,f}$ $(\forall k)$, whereas, in this paper, steering vectors are time-variant $\boldsymbol{a}_{k,f,t}$ $(\forall k)$ to account for CMA rotation. We aim to estimate source images at the reference microphone even when the CMA is rotated. In the following, we assume that the rotation angle θ_t at each frame is known using another sensor, such as an angular acceleration sensor.

Next, we consider an online BSS problem analyzed using the model defined above. We henceforth assume that the number of microphones of the CMA is equal to that of sources: $M = K$. We aim to estimate demixing matrices and signals using only the currently and previously observed signals $\boldsymbol{x}_{f,1}, \ldots, \boldsymbol{x}_{f,t}$:

$$\boldsymbol{W}_{f,t} = \begin{bmatrix} \boldsymbol{w}_{1,f,t} & \ldots & \boldsymbol{w}_{K,f,t} \end{bmatrix}^{\mathsf{H}} \in \mathbb{C}^{K \times M}, \tag{3}$$

$$\boldsymbol{y}_{f,t} = \boldsymbol{W}_{f,t} \boldsymbol{x}_{f,t} \in \mathbb{C}^{K}, \tag{4}$$

where $\boldsymbol{W}_{f,t}$ is the demixing matrix and $\boldsymbol{y}_{f,t}$ is the estimated signal.

Unless otherwise specified, the indices f, t, and k always range from 1 to F, T, and K, respectively. We omit the bounds of sets for these indices when they span the ranges. $\{\boldsymbol{x}_{f,t}\}_{f,t}$ denotes the set of $\boldsymbol{x}_{f,t}$ for all f and t, for example.

3 Conventional Methods

3.1 Batch Auxiliary-Function-Based Independent Vector Analysis

As the basis of our work, we first summarize the *batch* AuxIVA [22]. In AuxIVA, we estimate *time-invariant* demixing matrices $\{\boldsymbol{W}_f\}_f$ using all the time frames $t = 1, \ldots, T$ by minimizing the following objective function:

$$J_f(\boldsymbol{W}_f) = \sum_{k=1}^{K} \boldsymbol{w}_{k,f}^{\mathsf{H}} \boldsymbol{V}_{k,f} \boldsymbol{w}_{k,f} - \log|\det \boldsymbol{W}_f|^2, \tag{5}$$

$$\boldsymbol{V}_{k,f} = \frac{1}{T} \sum_{t=1}^{T} \varphi(r_{k,t}) \boldsymbol{x}_{f,t} \boldsymbol{x}_{f,t}^{\mathsf{H}}, \tag{6}$$

$$r_{k,t} = \sqrt{\textstyle\sum_{f=1}^{F} |\boldsymbol{w}_{k,f}^{\mathsf{H}} \boldsymbol{x}_{f,t}|^2}, \tag{7}$$

where $\varphi\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is defined as $\varphi(r) = \psi'(r)/(2r)$, with $\psi'(r)$ being the first derivative of $\psi(r)$ for r and $\psi(r)$ is the contrast function derived from the probability density function of source signals. In this paper, we assume that $\varphi(r) = F/r^2$, which represents the time-varying Gaussian distribution [21]. $\boldsymbol{V}_{k,f}$ is the *weighted covariance matrix* of the observed signals. One popular method of minimizing the objective function (5) with respect to $\boldsymbol{W}_f$ includes

iterative projection (IP) [22]. IP cyclically updates each row vector of the demixing matrix (*demixing vector*) $\boldsymbol{w}_{k,f}^{\mathsf{H}}$ $(k = 1, \ldots, K)$ using the following update rule:

$$\boldsymbol{w}_{k,f} \leftarrow (\boldsymbol{W}_f \boldsymbol{V}_{k,f})^{-1} \boldsymbol{e}_k, \tag{8}$$

$$\boldsymbol{w}_{k,f} \leftarrow \frac{\boldsymbol{w}_{k,f}}{\sqrt{\boldsymbol{w}_{k,f}^{\mathsf{H}} \boldsymbol{V}_{k,f} \boldsymbol{w}_{k,f}}}, \tag{9}$$

where $\boldsymbol{e}_k \in \mathbb{R}^K$ is the canonical basis vector with the kth element unity. The estimated signal is estimated as $\boldsymbol{y}_{f,t} = \boldsymbol{W}_f \boldsymbol{x}_{f,t}$.

3.2 Online Auxiliary-Function-Based Independent Vector Analysis

Online AuxIVA (OIVA) [27] is an extension of batch AuxIVA to an online algorithm. In OIVA, the weighted covariance matrices are updated with the following incremental update rule as an approximation of (6):

$$\boldsymbol{V}_{k,f,t} = \alpha \boldsymbol{V}_{k,f,t-1} + (1 - \alpha) \varphi(r_{k,t}) \boldsymbol{x}_{f,t} \boldsymbol{x}_{f,t}^{\mathsf{H}}, \tag{10}$$

$$r_{k,t} = \sqrt{\textstyle\sum_{f=1}^{F} |\boldsymbol{w}_{k,f,t}^{\mathsf{H}} \boldsymbol{x}_{f,t}|^2}, \tag{11}$$

where α $(0 \leq \alpha < 1)$ is the forgetting factor. With this incremental update rule, we can directly apply IP to estimating time-varying demixing matrices $\{\boldsymbol{W}_{f,t}\}_f$ at each time frame t by simply replacing $\boldsymbol{w}_{k,f}$, $\boldsymbol{V}_{k,f}$ in (8), (9) with $\boldsymbol{w}_{k,f,t}$, $\boldsymbol{V}_{k,f,t}$.

We then estimate the signal using by (4). The scale of the output estimated signal $\boldsymbol{y}_{f,t}$ may be contaminated by the scale ambiguity problem. To restore the scale ambiguity and obtain an estimated source image at the reference microphone, we apply the following backprojection [18] for postprocessing:

$$\widehat{\boldsymbol{W}}_{f,t} \leftarrow \mathrm{diag}\left(\boldsymbol{e}_1^{\mathsf{T}} \boldsymbol{W}_{f,t}^{-1}\right) \boldsymbol{W}_{f,t}, \tag{12}$$

$$\boldsymbol{y}_{f,t} \leftarrow \widehat{\boldsymbol{W}}_{f,t} \boldsymbol{x}_{f,t}, \tag{13}$$

where $\mathrm{diag}(\cdot)$ is an operator for constructing a diagonal matrix with each of its elements equals the corresponding element of the given vector. Algorithm 1 summarizes the OIVA algorithm. The online updates of $\boldsymbol{V}_{k,f,t}$ (10) enable OIVA to track the slow movement of source signals or microphones progressively. Nevertheless, it cannot promptly adapt to rapid changes, which can happen in head rotation when wearing CMA on the head. In this research, we aim to solve this problem.

3.3 Sound Field Interpolation on Circular Microphone Array

In this subsection, we briefly review the sound field interpolation method on a CMA, initially proposed in [29, 30]. Let $x(\phi)$ be a continuous sound pressure

Algorithm 1 Online AuxIVA (OIVA)

Input: $\{\boldsymbol{x}_{f,t}\}_{f,t}$, $\{\boldsymbol{W}_{f,0}\}_{f}$, $\{\boldsymbol{V}_{k,f,0}\}_{k,f}$, α, N_{itr}
Output: $\{\boldsymbol{y}_{f,t}\}_{f,t}$

for $t = 1, \ldots, T$
  for $f = 1, \ldots, F$
    $\boldsymbol{W}_{f,t} \leftarrow \boldsymbol{W}_{f,t-1}$
  for $n_{\mathrm{itr}} = 1, \ldots, N_{\mathrm{itr}}$
    for $k = 1, \ldots, K$
      $r_{k,t} \leftarrow \sqrt{\sum_{f=1}^{F} |\boldsymbol{w}_{k,f,t}^{\mathsf{H}} \boldsymbol{x}_{f,t}|^2}$ // (11)
      for $f = 1, \ldots, F$
        $\boldsymbol{V}_{k,f,t} \leftarrow \alpha \boldsymbol{V}_{k,f,t-1} + (1-\alpha)\varphi(r_{k,t})\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}$ // (10)
        $\boldsymbol{w}_{k,f,t} \leftarrow (\boldsymbol{W}_{f,t}\boldsymbol{V}_{k,f,t})^{-1}\boldsymbol{e}_k$ // (8)
        $\boldsymbol{w}_{k,f,t} \leftarrow \frac{\boldsymbol{w}_{k,f,t}}{\sqrt{\boldsymbol{w}_{k,f,t}^{\mathsf{H}} \boldsymbol{V}_{k,f,t} \boldsymbol{w}_{k,f,t}}}$ // (9)
  for $f = 1, \ldots, F$
    $\widehat{\boldsymbol{W}}_{f,t} \leftarrow \mathrm{diag}\left(\boldsymbol{e}_1^{\mathsf{T}} \boldsymbol{W}_{f,t}^{-1}\right) \boldsymbol{W}_{f,t}$ // (12)
    $\boldsymbol{y}_{f,t} \leftarrow \widehat{\boldsymbol{W}}_{f,t}\boldsymbol{x}_{f,t}$ // (13)

on the circumference of a circle at a spatial angle ϕ ($0 \leq \phi < 2\pi$), as shown in Figure 2. Then, we observe a sound field with M microphones distributed on the circle at even intervals. The mth observed signal is denoted as

$$x_m = x\left(2\pi\frac{m}{M}\right), \ (m = 0, \ldots, M-1). \tag{14}$$

In other words, we regard the observations of the sound field with a CMA as the discretizations of that along an angle on the circumference ϕ. We assume that this spatial sampling by the CMA satisfies Shannon's sampling theorem, *i.e.*, $x(\phi)$ contains no frequency components higher than half of the sampling frequency on the circumference of the circle. If the CMA is rotated by a spatial angle $\theta \in \mathbb{R}$, we can regard its observation as the δ-sample shift of x_m where $\delta = \frac{M}{2\pi}\theta$;

$$x\left(\frac{2\pi}{M}m + \theta\right) \coloneqq x_{m+\delta}. \tag{15}$$

Next, we formulate a sound field interpolation problem with the model above. We define the M-point spatial discrete Fourier transform (DFT) of x_m

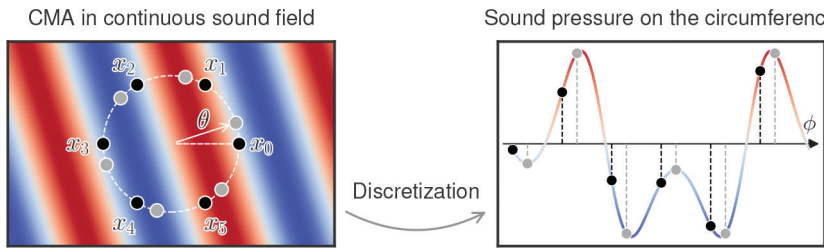


Figure 2: Concept of sound field interpolation on a CMA. The left figure shows the CMA in a continuous sound field (six microphones in this example). Each black dot is a microphone at the reference position, and each gray dot is a microphone at the rotated position. Background colors indicate the sound pressure of a plane wave. The right side shows the sound pressure along the circumference.

and its inverse transform as

$$\mathcal{F}_K[x_m] = \sum_{m\in\mathcal{K}} x_m z^{-mk} \coloneqq X_k, \tag{16}$$

$$\mathcal{F}_K^{-1}[X_k] = \frac{1}{M}\sum_{k\in\mathcal{K}} X_k z^{mk}, \tag{17}$$

where $z \coloneqq \exp(j\frac{2\pi}{M})$ is a twiddle factor of the M-point DFT, j is the imaginary unit, and $\mathcal{K}$ is an index set defined as

$$\mathcal{K} = \begin{cases} \{-\frac{K}{2}+1, -\frac{K}{2}+2, \ldots, \frac{K}{2}\} & (K \text{ is even}), \\ \{-\frac{K-1}{2}, -\frac{K-1}{2}+1, \ldots, \frac{K-1}{2}\} & (K \text{ is odd}). \end{cases} \tag{18}$$

As is well known, for the shift theorem of DFT, the following equation is satisfied for any integers d:

$$\mathcal{F}_K[x_{m+d}] = X_k z^{dk}. \tag{19}$$

Although (19) does not hold strictly for real numbers, we assume that the following equation holds approximately for a real number δ:

$$\mathcal{F}_K[x_{m+\delta}] = X_k z^{\delta k}. \tag{20}$$

From these assumptions, we have

$$x_{m+\delta} = \mathcal{F}_K^{-1}\left[X_k z^{\delta k}\right], \tag{21}$$

$$= \frac{1}{M}\sum_{k\in\mathcal{K}} \left(X_k z^{\delta k}\right) z^{mk}, \tag{22}$$

$$= \frac{1}{M}\sum_{k\in\mathcal{K}} \left(\sum_{n=0}^{M-1} x_n z^{-nk}\right)\left(z^{(\delta+m)k}\right), \tag{23}$$

$$= \sum_{n=0}^{M-1} x_n \left(\frac{1}{M} \sum_{k \in \mathcal{K}} z^{(\delta+m-n)k} \right) \coloneqq \sum_{n=0}^{M-1} x_n u_{m,n}(\delta). \tag{24}$$

The coefficient $u_{m,n}(\delta)$ is calculated as

$$u_{m,n}(\theta) = \begin{cases} \dfrac{z^{-L(\frac{M}{2}-1)}}{M} \dfrac{1-z^{LM}}{1-z^{L}} \\ \dfrac{z^{-L\frac{K-1}{2}}}{M} \dfrac{1-z^{LM}}{1-z^{L}} \end{cases} = \begin{cases} \dfrac{\mathrm{sinc}(L)}{\mathrm{sinc}(L/M)} z^{\frac{L}{2}} & (M \text{ is even}), \\ \dfrac{\mathrm{sinc}(L)}{\mathrm{sinc}(L/M)} & (M \text{ is odd}), \end{cases} \tag{25}$$

where $L = \delta + m - n$ $(m, n = 0, \ldots, M-1)$ [30]. (25) is an alternative expression for (3) in [30]. See Section II of [30] for the detailed derivation.

The above relationship also holds in the frequency domain. Let us define the following vector:

$$\boldsymbol{x} = \begin{bmatrix} x_0 & \ldots & x_{M-1} \end{bmatrix}^{\mathsf{T}}, \tag{26}$$

$$\boldsymbol{x}_\delta = \begin{bmatrix} x_{0+\delta} & \ldots & x_{(M-1)+\delta} \end{bmatrix}^{\mathsf{T}}. \tag{27}$$

From (24), we have

$$\boldsymbol{x}_\delta = \begin{bmatrix} u_{0,0}(\theta) & \ldots & u_{0,M-1}(\theta) \\ \vdots & \ddots & \vdots \\ u_{M-1,0}(\theta) & \ldots & u_{M-1,M-1}(\theta) \end{bmatrix} \boldsymbol{x} \coloneqq \boldsymbol{U}(\theta)\boldsymbol{x}, \tag{28}$$

where $\boldsymbol{U}(\theta) \in \mathbb{C}^{M \times M}$ is the *rotation matrix*. By definition, $\boldsymbol{U}(\theta)$ is obviously a unitary matrix: $\boldsymbol{U}^{-1}(\theta) = \boldsymbol{U}^{\mathsf{H}}(\theta)$. Next, let $\boldsymbol{X}$ be its DFT defined as $\begin{bmatrix} X_0 & \ldots & X_{M-1} \end{bmatrix}^{\mathsf{T}} = \boldsymbol{F}\boldsymbol{x}$, where $\boldsymbol{F}$ is a K-point DFT matrix. By using these expressions, we can diagonalize $\boldsymbol{U}(\theta)$ as $\boldsymbol{F}^{\mathsf{H}}\boldsymbol{U}(\theta)\boldsymbol{F}$ since $\boldsymbol{U}(\theta)$ is a unitary matrix. Therefore, the relationship between $\boldsymbol{X}$ and $\boldsymbol{X}_\delta$ can be expressed as $\boldsymbol{X}_\delta = \boldsymbol{U}(\theta)\boldsymbol{X}$, because $\boldsymbol{x}_\delta = \boldsymbol{F}^{\mathsf{H}}\boldsymbol{U}(\theta)\boldsymbol{F}\boldsymbol{x}$.

In the STFT domain, we consider a situation where a CMA is rotated by degree θ_t at the time frame t, and let $\widetilde{\boldsymbol{x}}_{f,t}$ be the observed signal recorded without CMA rotation (*reference position*). By using the expression above, we assume that the observed signal with CMA rotation $\boldsymbol{x}_{f,t}$ is expressed as the following linear approximation:

$$\widetilde{\boldsymbol{x}}_{f,t} = \boldsymbol{U}^{-1}(\theta_t)\boldsymbol{x}_{f,t}. \tag{29}$$

As mentioned in the previous section, note that θ_t must be known using another sensor, such as an angular acceleration sensor, or estimated from the acoustic observation itself [16].

4 Proposed Method

4.1 SFI-based OIVA with Transformation of Latest Observation

In the beamforming that is robust against the self-rotation proposed in [30], the signals observed by the CMA with angle θ_t are transformed frame by frame to what would have been observed at the reference position, namely, at the angle $\theta = 0$, using SFI, and then the beamforming is applied to the transformed signals.

To make OIVA robust against self-rotation, we first consider a similar approach to [30] in this subsection. In this method, we simply apply the transformation using SFI to the latest observation, *i.e.*,

$$\widetilde{\boldsymbol{x}}_{f,t} \leftarrow \boldsymbol{U}^{\mathsf{H}}(\theta_t)\boldsymbol{x}_{f,t}, \tag{30}$$

and the online update of the weighted covariance matrix $\boldsymbol{V}_{k,f,t}$ is performed using the transformed signal $\widetilde{\boldsymbol{x}}_{f,t}$ such as

$$\boldsymbol{V}_{k,f,t} \leftarrow \alpha \boldsymbol{V}_{k,f,t-1} + (1-\alpha)\varphi(r_{k,t})\widetilde{\boldsymbol{x}}_{f,t}\widetilde{\boldsymbol{x}}_{f,t}^{\mathsf{H}}. \tag{31}$$

The demixing matrices $\boldsymbol{W}_{f,t}$ are estimated by using these weighted covariance matrices $\boldsymbol{V}_{k,f,t}$ similarly to OIVA. Algorithm 2 summarizes OIVA using this approach, and Figure 3 shows the system diagram.

Assuming that a CMA rotates only once at a specific time frame t, we analyze how the proposed method, as given in (31), differs from the original formula in (10). In (10), the value of $\boldsymbol{V}_{k,f,t}$ is updated online. However, since the angle of the CMA differs between the time frames $t-1$ and t, the blending of $\boldsymbol{V}_{k,f,t-1}$, which holds information before the rotation, and $\boldsymbol{x}_{f,t}$, which holds information after the rotation, leads to an inaccurate estimation of the demixing matrix $\boldsymbol{W}_{f,t}$. If the CMA does not rotate further after the time frame t, the observation information before the rotation is gradually diminished, and $\boldsymbol{V}_{k,f,t}$ and $\boldsymbol{W}_{f,t}$ are expected to slowly converge to their values associated with the post-rotation position of CMA. However, this convergence is expected to take time. In contrast, in (31), all observed signals are translated into the observation at the reference position. Thus, even if the CMA rotates at the time frame t, there is no substantial discrepancy between $\boldsymbol{V}_{k,f,t-1}$ and $\widetilde{\boldsymbol{x}}_{f,t}$, and the effect of the rotation on the estimation of the demixing matrix $\boldsymbol{W}_{f,t}$ is expected to be markedly reduced.

Note that the separated signals obtained by this approach are the source images at the reference microphone of the CMA with the reference angle. Since the position of the reference microphone is fixed in the space regardless of the CMA rotation, this causes problems in some practical applications. For example, in hearing aid applications or virtual/augmented reality (VR/AR) applications with a head-mounted display, it is desired to present a separated

signal with the source localization sensation. For this purpose, we should estimate the source images on the reference microphone of the CMA rotated together with the head.

Algorithm 2 SFI-based OIVA using Transformation of Latest Observation (`SFIIVA-O`)

Input: $\{\boldsymbol{x}_{f,t}\}_{f,t}$, $\{\boldsymbol{W}_{f,0}\}_{f}$, $\{\boldsymbol{V}_{k,f,0}\}_{k,f}$, α, N_{itr}, $\{\theta_t\}_t$
Output: $\{\boldsymbol{y}_{f,t}\}_{f,t}$

for $t = 1, \ldots, T$
  for $f = 1, \ldots, F$
    $\widetilde{\boldsymbol{x}}_{f,t} \leftarrow \boldsymbol{U}^{\mathsf{H}}(\theta_t)\boldsymbol{x}_{f,t}$ // (30)
    $\boldsymbol{W}_{f,t} \leftarrow \boldsymbol{W}_{f,t-1}$
  for $n_{\mathrm{itr}} = 1, \ldots, N_{\mathrm{itr}}$
    for $k = 1, \ldots, K$
      $r_{k,t} \leftarrow \sqrt{\sum_{f=1}^{F} |\boldsymbol{w}_{k,f,t}^{\mathsf{H}} \widetilde{\boldsymbol{x}}_{f,t}|^2}$ // (11)
      for $f = 1, \ldots, F$
        $\boldsymbol{V}_{k,f,t} \leftarrow \alpha \boldsymbol{V}_{k,f,t-1} + (1-\alpha)\varphi(r_{k,t})\widetilde{\boldsymbol{x}}_{f,t}\widetilde{\boldsymbol{x}}_{f,t}^{\mathsf{H}}$ // (31)
        $\boldsymbol{w}_{k,f,t} \leftarrow (\boldsymbol{W}_{f,t}\boldsymbol{V}_{k,f,t})^{-1}\boldsymbol{e}_k$ // (8)
        $\boldsymbol{w}_{k,f,t} \leftarrow \frac{\boldsymbol{w}_{k,f,t}}{\sqrt{\boldsymbol{w}_{k,f,t}^{\mathsf{H}}\boldsymbol{V}_{k,f,t}\boldsymbol{w}_{k,f,t}}}$ // (9)
  for $f = 1, \ldots, F$
    $\widehat{\boldsymbol{W}}_{f,t} \leftarrow \mathrm{diag}\left(\boldsymbol{e}_1^{\mathsf{T}}\boldsymbol{W}_{f,t}^{-1}\right)\boldsymbol{W}_{f,t}$ // (12)
    $\boldsymbol{y}_{f,t} \leftarrow \widehat{\boldsymbol{W}}_{f,t}\widetilde{\boldsymbol{x}}_{f,t}$ // (13)

4.2 SFI-based OIVA with Transformation of Pre-update Demixing and Weighted Covariance Matrices

In this subsection, we propose another online BSS that is robust against CMA rotation and outputs the source image at the latest reference microphone position frame by frame. As discussed in the previous section, the mismatch between $\boldsymbol{V}_{k,f,t-1}$ and $\boldsymbol{x}_{f,t}$ causes a problem. In the previous approach, we transform $\boldsymbol{x}_{f,t}$ to what it should be at the reference position of the CMA. In this subsection, we consider another approach: transforming $\boldsymbol{V}_{k,f,t-1}$ to what it should be at the latest position of the CMA.

First, we discuss the parameter transformation in a *time-invariant* case for simplicity. In this case, demixing matrices $\boldsymbol{W}_f$ and weighted covariance matrices $\boldsymbol{V}_{k,f}$ are independent of the time frame t. Let $\boldsymbol{x}_{f,t}$ and $\widetilde{\boldsymbol{x}}_{f,t}$ be observations if the angle of the CMA is fixed at θ_1 and θ_2, respectively. If the approximation of SFI is satisfied, $\boldsymbol{x}_{f,t}$ and $\widetilde{\boldsymbol{x}}_{f,t}$ hold the following relation:

$$\widetilde{\boldsymbol{x}}_{f,t} = \boldsymbol{U}(\Delta\theta)\boldsymbol{x}_{f,t}, \tag{32}$$

where $\Delta\theta = \theta_2 - \theta_1$. From this, we have the estimated signal $\boldsymbol{y}_{f,t}$ as

$$\begin{aligned}\boldsymbol{y}_{f,t} &= \boldsymbol{W}_f\boldsymbol{x}_{f,t}, &(33)\\ &= \boldsymbol{W}_f\boldsymbol{U}^{\mathsf{H}}(\Delta\theta)\widetilde{\boldsymbol{x}}_{f,t} \coloneqq \widetilde{\boldsymbol{W}}_f\widetilde{\boldsymbol{x}}_{f,t}, &(34)\end{aligned}$$

where $\boldsymbol{W}_f$ and $\widetilde{\boldsymbol{W}}_f$ are time-invariant demixing matrices for $\boldsymbol{x}_{f,t}$ and $\widetilde{\boldsymbol{x}}_{f,t}$, respectively. Therefore,

$$\widetilde{\boldsymbol{W}}_f = \boldsymbol{W}_f\boldsymbol{U}^{\mathsf{H}}(\Delta\theta). \tag{35}$$

Similarly, the time-invariant weighted covariance matrix $\boldsymbol{V}_{k,f}$ estimated using $\boldsymbol{x}_{f,t}$ is calculated as

$$\boldsymbol{V}_{k,f} = \mathbb{E}_t\left[\varphi(r_{k,t})\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right], \tag{36}$$

where $\mathbb{E}_t\left[\boldsymbol{A}_t\right]$ is the expectation operation for the random variable matrix $\boldsymbol{A}_t$ with respect to the time frame t. We rewrite the weighted covariance matrix $\boldsymbol{V}_{k,f}$ of $\boldsymbol{x}_{f,t}$ using $\widetilde{\boldsymbol{V}}_{k,f}$ as

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{k,f} &= \mathbb{E}_t\left[\varphi(r_{k,t})\widetilde{\boldsymbol{x}}_{f,t}\widetilde{\boldsymbol{x}}_{f,t}^{\mathsf{H}}\right], &(37)\\ &= \mathbb{E}_t\left[\varphi(r_{k,t})\boldsymbol{U}(\Delta\theta)\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\boldsymbol{U}^{\mathsf{H}}(\Delta\theta)\right], &(38)\\ &= \boldsymbol{U}(\Delta\theta)\,\mathbb{E}_t\left[\varphi(r_{k,t})\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right]\boldsymbol{U}^{\mathsf{H}}(\Delta\theta), &(39)\\ &= \boldsymbol{U}(\Delta\theta)\boldsymbol{V}_{k,f}\boldsymbol{U}^{\mathsf{H}}(\Delta\theta). &(40)\end{aligned}$$

Although $\boldsymbol{V}_{k,f,t}$ and $\boldsymbol{W}_{f,t}$ are estimated online in OIVA, we use these variable transformations written in (35) and (40) for the preprocessing of their update to compensate for the CMA rotation such as

$$\begin{aligned}\widetilde{\boldsymbol{W}}_{f,t-1} &\leftarrow \boldsymbol{W}_{f,t-1}\boldsymbol{U}^{\mathsf{H}}(\Delta\theta_t), &(41)\\ \widetilde{\boldsymbol{V}}_{k,f,t-1} &\leftarrow \boldsymbol{U}(\Delta\theta_t)\boldsymbol{V}_{k,f,t-1}\boldsymbol{U}^{\mathsf{H}}(\Delta\theta_t), &(42)\end{aligned}$$

where $\Delta\theta_t = \theta_t - \theta_{t-1}$. This method preserves the observation as is and transforms the intermediate variables as $\boldsymbol{W}_{f,t-1}$ and $\boldsymbol{V}_{k,f,t-1}$ instead; thus, the source image at the latest CMA position can be estimated, which is an advantage of this algorithm. Furthermore, the angle of the CMA is supposed

Algorithm 3 SFI-based OIVA with Transformation of Pre-update Demixing and Weighted Covariance Matrices (`SFIIVA-M`)

Input: $\{\boldsymbol{x}_{f,t}\}_{f,t}$, $\{\boldsymbol{W}_{f,0}\}_f$, $\{\boldsymbol{V}_{k,f,0}\}_{k,f}$, α, N_{itr}, $\{\theta_t\}_t$
Output: $\{\boldsymbol{y}_{f,t}\}_{f,t}$

for $t = 1, \ldots, T$
  $\Delta\theta_t \leftarrow \theta_t - \theta_{t-1}$
  for $f = 1, \ldots, F$
    $\boldsymbol{W}_{f,t} \leftarrow \boldsymbol{W}_{f,t-1}\boldsymbol{U}^{\mathsf{H}}(\Delta\theta_t)$ // (41)
    for $k = 1, \ldots, K$
      $\widetilde{\boldsymbol{V}}_{k,f,t-1} \leftarrow \boldsymbol{U}(\Delta\theta_t)\boldsymbol{V}_{k,f,t-1}\boldsymbol{U}^{\mathsf{H}}(\Delta\theta_t)$ // (42)
  for $n_{\mathrm{itr}} = 1, \ldots, N_{\mathrm{itr}}$
    for $k = 1, \ldots, K$
      $r_{k,t} \leftarrow \sqrt{\sum_{f=1}^{F} |\boldsymbol{w}_{k,f,t}^{\mathsf{H}}\boldsymbol{x}_{f,t}|^2}$ // (11)
      for $f = 1, \ldots, F$
        $\boldsymbol{V}_{k,f,t} \leftarrow \alpha\widetilde{\boldsymbol{V}}_{k,f,t-1} + (1-\alpha)\varphi(r_{k,t})\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}$
        $\boldsymbol{w}_{k,f,t} \leftarrow \left(\boldsymbol{W}_{f,t}\boldsymbol{V}_{k,f,t}\right)^{-1}\boldsymbol{e}_k$ // (8)
        $\boldsymbol{w}_{k,f,t} \leftarrow \frac{\boldsymbol{w}_{k,f,t}}{\sqrt{\boldsymbol{w}_{k,f,t}^{\mathsf{H}}\boldsymbol{V}_{k,f,t}\boldsymbol{w}_{k,f,t}}}$ // (9)
  for $f = 1, \ldots, F$
    $\widehat{\boldsymbol{W}}_{f,t} \leftarrow \mathrm{diag}\left(\boldsymbol{e}_1^{\mathsf{T}}\boldsymbol{W}_{f,t}^{-1}\right)\boldsymbol{W}_{f,t}$ // (12)
    $\boldsymbol{y}_{f,t} \leftarrow \widehat{\boldsymbol{W}}_{f,t}\boldsymbol{x}_{f,t}$ // (13)

to be measured by integrating angular accelerometers, which can introduce bias errors. In this case, the method described in the previous subsection continues to have errors in the transformation of the observations to those at the reference position. On the other hand, in the method described in this subsection, the error included in the transformation of $\boldsymbol{W}_{f,t}$ and $\boldsymbol{V}_{k,f,t}$ should disappear with online updates. This is another advantage of this approach. We will show this in an experiment to be described later. The entire algorithm is summarized as Algorithm 3, and Figure 4 shows the system diagram of our approach.

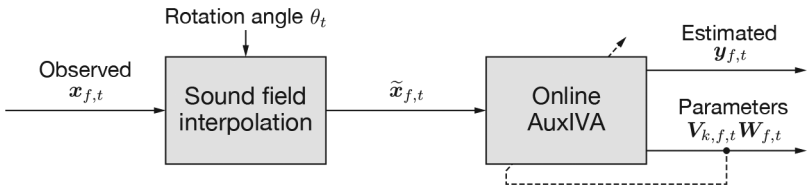


Figure 3: System diagram of proposed `SFIIVA-O`.

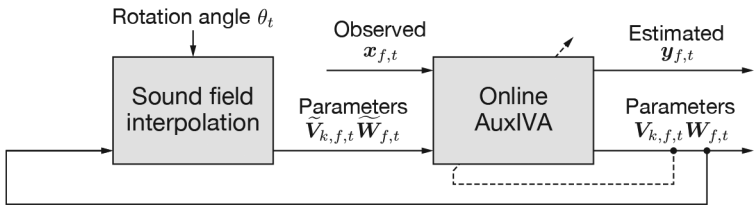


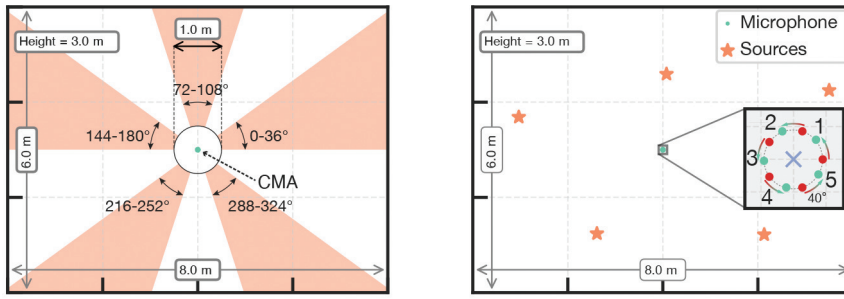
Figure 4: System diagram of proposed `SFIIVA-M`.

5 Experimental Validation

In this section, we confirm with our experimental results how much SFI contributes to BSS and discuss the difference in the location of the source image due to the SFI. Henceforth, we assume that the CMA was instantaneously rotated at the angle θ at the time frame τ.

5.1 Setup

Simulation experiments with large random synthesized datasets were conducted to evaluate the performance of BSS under a situation where a CMA rotates. The datasets consisted of 100 samples, and each sample simulated observed signals with five source signals with a five-channel CMA using the image source method [2]. As the source signals, we used the speech signals of five speakers (`jvs001`, `jvs002`, ..., `jvs005`) from the JVS dataset [26]. In this experiment, the utterances included in `parallel100` were randomly concatenated for each speaker. The length was 60 s, and the sampling frequency was resampled from 24 kHz to 16 kHz. The reverberation time was approximately 100 ms. A CMA with $K = 5$ channels and a radius of 2 cm was placed at the center of the room. Each source was randomly placed at least 1 m from the center of the CMA and within an angle ranging from $\frac{360^\circ}{K}k$ to $\frac{360^\circ}{K}(k-1) + \frac{180^\circ}{K}$ $(k = 1, \ldots, K)$. Figure 5 illustrates the range of sources' placements and its example. To simulate the instant rotation of the CMA, the source images were generated when the CMA was rotated 40° counter-clockwise from the horizontal axis and was joined together at 0–30 s and 30–60 s intervals. Note that the rotation



(a) Five sources are randomly placed in each orange-shaded area.

(b) Example of sources and microphones placement.

Figure 5: Room layout. The center of CMA is placed at (4.0 m, 3.0 m) in the room with the radius of 2 cm. CMA was rotated 40° counter-clockwise at 30 s. In the zoomed plot of (b), red dots represent the microphones at the position before rotation, green dots represent the position after rotation, and the cross lines in the middle represent the center of CMA.

angle and time are given in an oracle manner. STFT was performed on the observed signals with a frame length of 4096 points, a shift length of 2048 points, and a Hamming window. The scale of the output signal was restored by backprojection as shown in Algorithm 1, with the reference microphone as microphone 1. The separation performance was evaluated by the scale-invariant signal-to-distortion ratio (SI-SDR) [15] and its improvement (SI-SDR improvement; SI-SDRi). The reference signal for the SI-SDR was the source image recorded with a rotated CMA.

The initial values for demixing and weighted covariance matrices were $\boldsymbol{W}_{f,0} = \boldsymbol{I}$ $(\forall f)$ and $\boldsymbol{V}_{k,f,0} = 10^{-3} \times \boldsymbol{I}$ $(\forall k, f)$, respectively, where $\boldsymbol{I}$ is the K-dimensional identity matrix. We set the number of iterations in each time frame in Algorithm 1 N_{itr} to 5. We ran experiments using various forgetting factors $\alpha = 0.9$, 0.95, 0.98, and 0.99, which were chosen so that the approximate number of frames $\frac{1}{1-\alpha} = 10$, 20, 50, and 100, respectively. To increase numerous stability for the implementation of OIVA, we applied the following ad-hoc normalization $\boldsymbol{V}_{k,f,t} \leftarrow \boldsymbol{V}_{k,f,t} + 10^{-3} \times \boldsymbol{I}$ after updates using (10). We compared the following four methods:

`Naive` OIVA described in Section 3.2 as a baseline.

`Reset` Re-initialize the weighted covariance and demixing matrices when the CMA was rotated: $\boldsymbol{W}_{f,\tau-1} \leftarrow \boldsymbol{I}$, $\boldsymbol{V}_{k,f,\tau-1} \leftarrow \varepsilon \boldsymbol{I}$.

`SFIIVA-O` Apply SFI to the latest observation and estimate the source image at the fixed position (Algorithm 2).

`SFIIVA-M` Apply SFI to the weighted covariance and demixing matrices as parameter transformation and estimate the source image at the position rotated with CMA (Algorithm 3).

5.2 Source Separation Performance

5.2.1 Noiseless Environments

As an initial experiment to verify the efficacy of the proposed methods, we first investigated source separation performance under a noiseless environment. Figure 6 shows the SI-SDRi every 1 s averaged over samples and channels. As shown in Figure 6, the result of forgetting factor α of 0.9 showed faster tracking to the CMA rotation than the other results. In OIVA, setting the forgetting factor α to a smaller value results in faster convergence of the demixing matrices and quicker adaptation to the CMA rotation. However, the influence of older data diminishes more rapidly, and the approximate number of frames $\frac{1}{1-\alpha}$ is shortened, which can degrade the separation performance. Therefore, in this case, we believe that the final performance immediately saturated and thus showed lower performance with no significant difference between the four methods.

Figure 7 shows the SI-SDRi averaged over samples and channels immediately after CMA rotation and after a sufficient time has elapsed. Overall, `Naive` and `Reset` showed a performance drop immediately after CMA rotation and

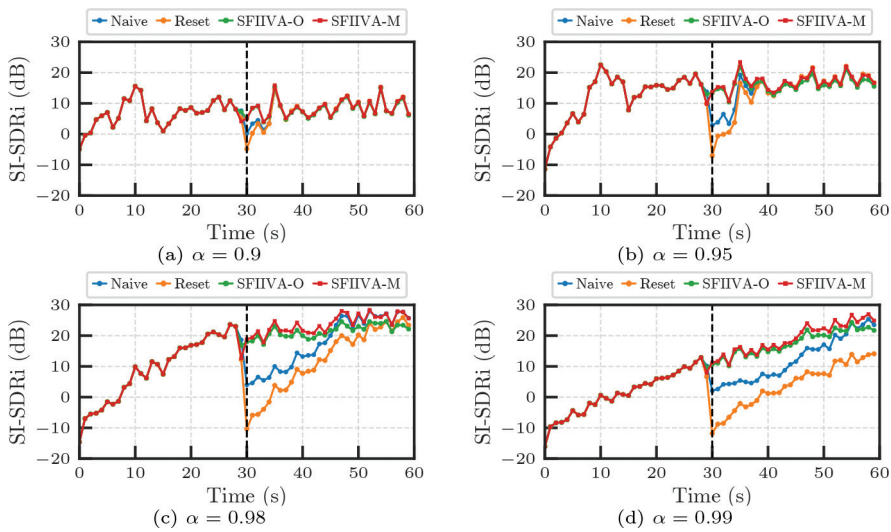


Figure 6: SI-SDR improvements (SI-SDRi) every 1 s under noiseless environments.

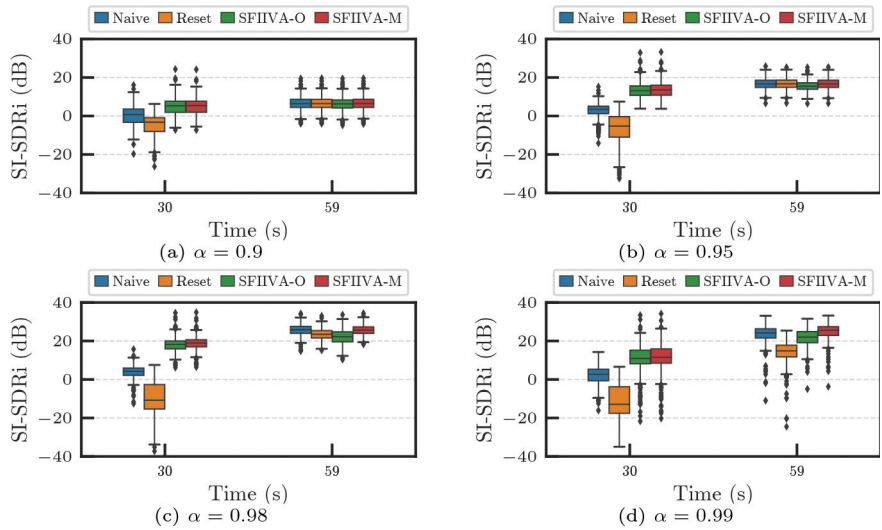


Figure 7: Box-and-whisker plots of SI-SDR improvements (SI-SDRi) immediately after CMA rotation and after a sufficient time has elapsed, corresponding 30 s and 59 s in Figure 6. The whiskers show the minima and maxima of each distribution, except for points determined to be outliers.

improved again with time. `Reset` has the lowest performance before CMA rotation among all forgetting factors, which worsens as the forgetting factor α approaches 1. In contrast, the proposed `SFIIVA-M` and `SFIIVA-O` methods performed better immediately after the CMA rotation than the others. Between `SFIIVA-M` and `SFIIVA-O`, `SFIIVA-M` performed slightly better, which may be due to the fact that `SFIIVA-O` estimates the source image at the reference microphone, whereas the reference signal for SI-SDR evaluation was the source image at the reference position.

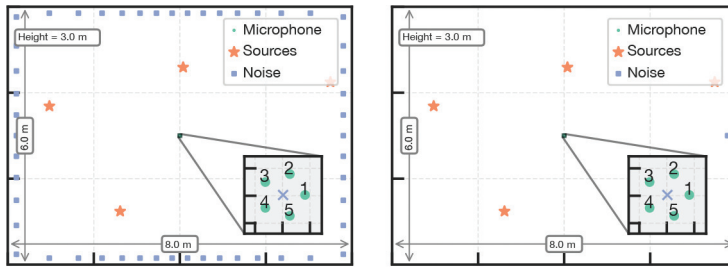
5.2.2 Noisy Environments

For a realistic simulation, we conducted experiments under the following two types of noisy environments:

`babble` Diffuse babble noise consisting of 50 speech signals.

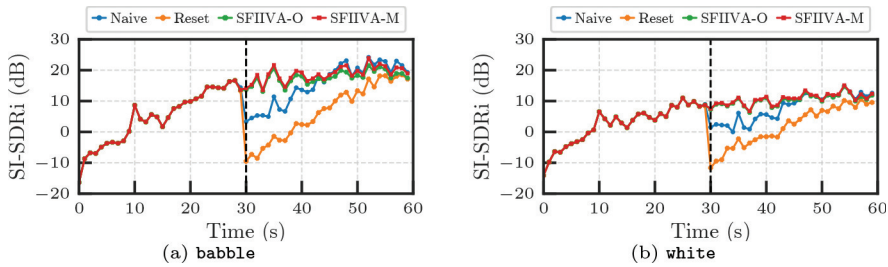
`white` Single interference source consisting of white noise.

Figure 8 shows examples of the layout of the sources and microphones in noisy environments. Noise signals in `babble` were selected from the speech signals of the JVS dataset that were different from those used for the source signals.



(a) `babble` (b) `white`

Figure 8: Examples of room layout of noisy environments. The center of CMA is placed at (4.0 m, 3.0 m) in the room with the radius of 2 cm. CMA was rotated 40° counter-clockwise at 30 s.



(a) `babble` (b) `white`

Figure 9: SI-SDR improvements (SI-SDRi) every 1 s with two kinds of noise.

The signal-to-noise ratio (SNR) is defined as

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_{k=1}^{K} \sigma_k^2}{\sigma_{\mathrm{n}}^2}, \tag{43}$$

where σ_k^2 and σ_{n}^2 are the variance of the kth source signal and noise signals at the reference microphone, respectively. After convolving room impulse responses, the variance of noise signals was scaled so that SNR = 20 dB. In this experiment, we set the forgetting number α to 0.98. Other conditions are the same as in the noiseless case.

Figure 9 shows the SI-SDRi every 1 s averaged over samples and channels under noisy environments. As shown in the figure, performance between methods was consistent with the noiseless case. Overall, the final separation performance is lower than that of the noiseless case (Figure 6(c)). The separation performance of `white` was about 10 dB lower than that of `babble`. This is because `white` has a flat spectrum, and it can be quite larger than that of speech in the high-frequency range that is not included in `babble` and has a negative impact on the separation performance.

Table 1: Average SI-SDRi (dB) immediately after CMA rotation and after a sufficient time has elapsed. The length of simulated speech signals is 60 s, and the CMA was instantaneously rotated at 30 s. The true accurate measurement of CMA rotation was at 40° and the inaccurate measurement was at 60°. Forgetting factor α was set to 0.98.

(a) Accurate measurement.			(b) Inaccurate measurement.		
	Time (s)			Time (s)	
Method	30	59	Method	30	59
SFIIVA-O	17.99	22.13	SFIIVA-O	9.92	20.26
SFIIVA-M	18.80	**25.62**	SFIIVA-M	9.98	**25.63**

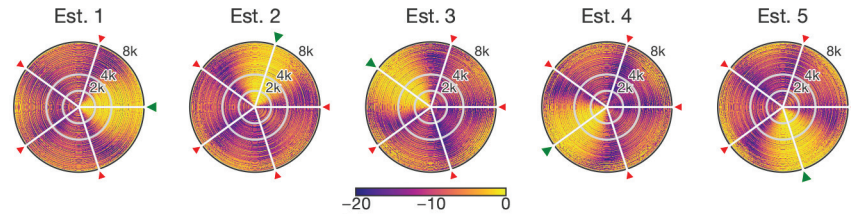
5.3 Robustness against Estimation Error

To examine the robustness against errors in measuring the angle of CMA rotation, we compared results when the measured angles used for `SFIIVA-O` and `SFIIVA-M` differed from the true angle. As discussed in Section 4.2, we expected errors to remain when there is an error in the measured angle since `SFIIVA-O` transforms the observed signal every frame. In contrast, `SFIIVA-M` performs the transformation only at the time of CMA rotation, so the effect of the measuring error should decrease with time.

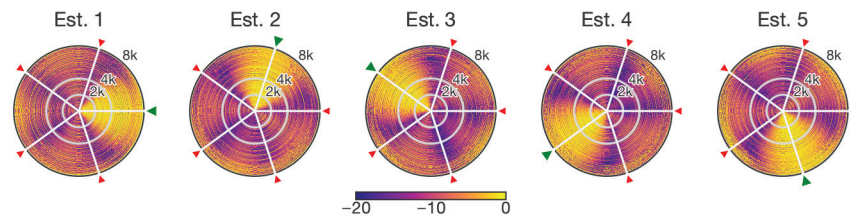
Table 1 shows the SI-SDRi (dB) averaged over samples and channels immediately after CMA rotation and after a sufficient time has elapsed with measurement errors. The forgetting factor α was fixed to 0.98 in this experiment. As expected, the SI-SDRi of both methods dropped by about 8 dB 30 s immediately after CMA rotation, when the measured angle was 60°, which is different from the true angle of 40°. On the other hand, by comparing the performance at 59 s, which is sufficient time after CMA rotation, the performance of `SFIIVA-M` was determined to be about 5 dB higher than which is that of `SFIIVA-O`.

5.4 Difference in Beam Pattern

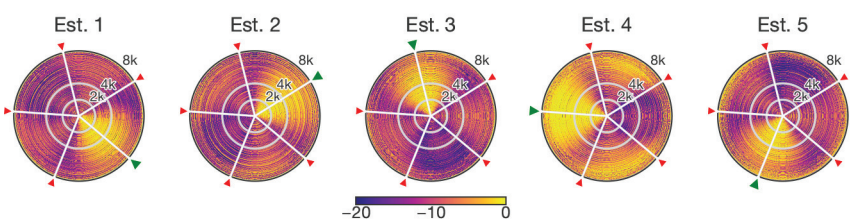
In this subsection, we selected one example to focus on the difference between the two proposed methods `SFIIVA-O` and `SFIIVA-M`. Figure 10 shows the beam patterns of the demixing matrices calculated by each method. Figure 10(a) is the result calculated before the rotation. Figures 10(b) and 10(c) are the results with the demixing matrices calculated using `SFIIVA-O` and `SFIIVA-M`, respectively, after the rotation. Dark regions in each plot represent nulls in a certain direction. In the case of OIVA, the desired result is that the direction of the estimated source is toward the brighter regions, and that of the others is toward the darker region. As shown in Figures 10 and 10(a), the results were similar to each other, and only the results in Figure 10(c) were different from



(a) Before rotation.



(b) `SFIIVA-O` after rotation.



(c) `SFIIVA-M` after rotation.

Figure 10: Beam patterns of demixing matrices. The five plots are the beam patterns of the frequency-wise demixing matrices calculated by each method. The radial direction of each plot represents the frequency, and the tangential direction represents the angle from the center of the CMA. The light and dark colors represent the gain in decibels. The five triangles in each plot indicate the true direction of the source, with green representing the target source and red the interference source.

the each other. The dark regions in Figures 10(a) and 10(b) were nearly in the desired directions since `SFIIVA-O` approximately canceled the rotation of the CMA and thus updated the demixing matrices in the same direction even after the rotation. For `SFIIVA-M`, the bright areas should be in the direction viewed from the position after the rotation, *i.e.*, 40° less anti-clockwise than the previous result in this case. The nulls may not be precisely formed in some frequency bands where the source signal is inherently weak because it does not affect the separated signal very much. Looking at the high-frequency bands of each beam pattern, there are several cases where bright regions are visible outside of the target signal, such as Est. 4 in Figure 10(c). We suppose this result might happen since the 4th estimated signal in high-frequency bands was accidentally weak.

6 Conclusion

In this study, sound field interpolation (SFI) for an equally spaced circular microphone array (CMA) was applied to online auxiliary-function-based independent vector analysis (OIVA). We have proposed the following two new methods: a simple combination of SFI and OIVA, and a practical method based on parameter transformations. Simulation experiments have confirmed that SFI improved the robustness of OIVA against CMA rotation. Future work includes combining this method with self-rotation angle estimation such as [16], under-determined BSS methods such as [33, 34], over-determined BSS methods such as [8, 23], and extending it to real-time processing.

Biographies

Taishi Nakashima received his B.E. in Engineering from Osaka University, Osaka, Japan, in 2019 and his M.S. in Computer Science from Tokyo Metropolitan University, Tokyo, Japan, in 2021. He is pursuing a Ph.D. at Tokyo Metropolitan University and has received the JSPS Research Fellowship (DC1) in April 2021. He is an esteemed Student Member of the Acoustical Society of Japan (ASJ) and the IEEE Signal Processing Society (SPS). He received the 24th Best Student Presentation Award of ASJ, the 16th IEEE SPS Japan Student Conference Paper Award in 2022, and the Top 3% Recognition at ICASSP 2023. His research interests primarily focus on blind source separation and acoustic signal processing.

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

Nobutaka Ono received his B.E., M.S., and Ph.D. degrees from The University of Tokyo, Japan, in 1996, 1998, and 2001, respectively. He became a

research associate in 2001 and a lecturer in 2005 at The University of Tokyo. He moved to the National Institute of Informatics in 2011 as an associate professor and then to Tokyo Metropolitan University in 2017 as a full professor. His research interests include acoustic signal processing, especially microphone array processing, source localization and separation, machine learning, and optimization algorithms. He is a member of IEEE, EURASIP, APSIPA, IPSJ, IEICE, and ASJ. He was a member of the IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee from 2014 to 2019. He served as Associate Editor of IEEE Transactions on Audio, Speech, and Language Processing from 2012 to 2015. He received the best paper award at APSIPA ASC in 2018 and 2021 and Sadaoki Furui Prize Paper Award from APSIPA in 2021.

References

[1] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwińska, and Z. Tu, "The 2nd Clarity Enhancement Challenge for Hearing Aid Speech Intelligibility Enhancement: Overview and Outcomes," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, DOI: 10.1109/ICASSP49357.2023.10094918.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), 1979, 943–50, DOI: 10.1121/1.382599.

[3] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, 68, 2020, 3545–58, DOI: 10.1109/TSP.2020.3000199.

[4] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments," 2021, DOI: 10.48550/arXiv.2107.04174, arXiv: 2107.04174v2 [cs.SD].

[5] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An introduction to the speech enhancement for augmented reality (SPEAR) challenge," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, DOI: 10.1109/IWAENC53105.2022.9914721.

[6] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 25(4), 2017, 780–93, DOI: 10.1109/taslp.2017.2665341.

[7] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. Independent Component Analysis and Blind Signal Separation*, March 2006, 601–8, DOI: 10.1007/11679363_75.

[8] R. Ikeshita, T. Nakatani, and S. Araki, "Overdetermined independent vector analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 591–5, DOI: 10.1109/ICASSP40776.2020.9053790.

[9] J. Janský, Z. Koldovský, J. Málek, T. Kounovský, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1), 2022, 1, DOI: 10.1186/s13636-021-00231-6.

[10] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. Circuits Syst. I*, 57(7), 2010, 1431–8, DOI: 10.1109/TCSI.2010.2048777.

[11] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 15(1), 2006, 70–9, DOI: 10.1109/TASL.2006.872618.

[12] 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, Language Process.*, 24(9), 2016, 1622–37, DOI: 10.1109/TASLP.2016.2577880.

[13] Z. Koldovský, V. Kautský, P. Tichavský, J. Čmejla, and J. Málek, "Dynamic independent component/vector analysis: time-variant linear mixtures separable by time-invariant beamformers," *IEEE Trans. Signal Process.*, 69, 2021, 2158–73, DOI: 10.1109/TSP.2021.3068626.

[14] Z. Koldovský, J. Málek, and J. Janský, "Extraction of independent vector component from underdetermined mixtures through block-wise determined modeling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, 7903–7, DOI: 10.1109/ICASSP.2019.8683431.

[15] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR — half-baked or well done?" In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, 626–30, DOI: 10.1109/ICASSP.2019.8683855.

[16] G. Lian, Y. Wakabayashi, T. Nakashima, and N. Ono, "Self-rotation angle estimation of circular microphone array based on sound field interpolation," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, December 2021, 1016–20.

[17] S. Makino, ed., *Audio Source Separation*, Springer International Publishing, 2018, DOI: 10.007/978-3-319-73031-8.

[18] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, 41(1-4), 2001, 1–24, DOI: 10.1016/S0925-2312(00)00345-3.

[19] T. Nakashima, R. Ikeshita, N. Ono, S. Araki, and T. Nakatani, "Fast online source steering algorithm for tracking single moving source using online independent vector analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, DOI: 10.1109/ICASSP49357.2023.10094962.

[20] A. A. Nugraha, K. Sekiguchi, M. Fontaine, Y. Bando, and K. Yoshii, "Flow-based independent vector analysis for blind source separation," *IEEE Signal Processing Letters*, 27, 2020, 2173–7, DOI: 10.1109/LSP.2020.3039944.

[21] N. Ono, "Auxiliary-function based independent vector analysis with power of vector-norm type weighting functions," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, December 2012, 1–4.

[22] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2011, 189–92, DOI: 10.1109/ASPAA.2011.6082320.

[23] R. Scheibler and N. Ono, "Independent Vector Analysis with More Microphones Than Sources," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, 185–9, DOI: 10.1109/WASPAA.2019.8937080.

[24] O. Schwartz and S. Gannot, "A recursive expectation-maximization algorithm for speaker tracking and separation," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1), 2021, DOI: 10.1186/s13636-021-00228-1.

[25] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, 216–20, DOI: 10.1109/ICASSP.2017.7952149.

[26] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," 2019, DOI: 10.48550/arXiv.1908.06248, arXiv: 1908.06248 [cs.SD].

[27] 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. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, 107–11.

[28] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, 506–10, DOI: 10.1109/ICASSP39728.2021.9413700.

[29] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Rotation-robust beamforming based on sound field interpolation with regularly circular microphone array," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, 771–5, DOI: 10.1109/ICASSP39728.2021.9414196.

[30] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Sound field interpolation for rotation-invariant multichannel array signal processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 31, 2023, 2286–98, DOI: 10.1109/TASLP.2023.3282098.

[31] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Steering vector estimation of moving source using sound field interpolation in the circumference," in *Proc. Autumn Meeting of the Acoustical Society of Japan*, in Japanese, September 2021, 293–4.

[32] K. Weisberg, S. Gannot, and O. Schwartz, "An online multiple-speaker DOA tracking using the Cappé-Moulines recursive expectation-maximization algorithm," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, 656–60, DOI: 10.1109/ICASSP.2019.8682659.

[33] K. Yamaoka, L. Li, N. Ono, S. Makino, and T. Yamada, "CNN-based virtual microphone signal estimation for MPDR beamforming in underdetermined situations," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2019, DOI: 10.23919/EUSIPCO.2019.8903040.

[34] K. Yamaoka, N. Ono, and S. Makino, "Time-Frequency-Bin-Wise Linear Combination of Beamformers for Distortionless Signal Enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 29, 2021, 3461–75, DOI: 10.1109/TASLP.2021.3126950.