Multi-frame Full-rank Spatial Covariance Analysis for Underdetermined Blind Source Separation and Dereverberation

Hiroshi Sawada, *Fellow, IEEE*, Rintaro Ikeshita, *Member, IEEE*, Keisuke Kinoshita, *Senior Member, IEEE*, Tomohiro Nakatani, *Fellow, IEEE*

***Abstract*—Full-rank spatial covariance analysis (FCA) is a technique for blind source separation (BSS), and can be applied to underdetermined situations where the sources outnumber the microphones. This paper proposes multi-frame FCA as an extension of FCA to improve the BSS performance when the room reverberations are not so short that multiple time frames are needed to cover the dominant parts of the reverberations. There has already been proposed an FCA model that considers delayed source components. However, the existing FCA model does not take the correlation between different time frames into account. In contrast, our new extension models multiple time frames with multivariate Gaussian distributions of larger dimensionality than the existing FCA models, aiming to better model the source components spanning multiple time frames. We derive an expectation-maximization (EM) algorithm to optimize the model parameters. Experimental results show that the proposed multi-frame FCA performed clearly better than the existing FCA techniques in BSS tasks and also joint BSS and blind dereverberation tasks.**

***Index Terms*—Blind source separation (BSS), blind dereverberation (BD), full-rank spatial covariance analysis (FCA), expectation-maximization (EM) algorithm, multivariate complex Gaussian distribution, weighted prediction error (WPE)**

I. Introduction

Blind source separation (BSS) aims to separate N sources from the mixtures on M sensors (e.g., microphones in audio cases) without the information about the sources and the mixing situation (e.g., the directions of sources). Among the BSS techniques extensively developed in more than thirty years [3]–[9], independent component analysis (ICA) [5], [6], [10], [11] is a well-established one in which the mixing system is assumed to be invertible. Full-rank spatial covariance analysis (FCA) [12]–[15], on the other hand, is rather a new technique that models a more flexible mixing system than ICA does. The most crucial difference between ICA and FCA is that ICA can only be applied to determined ($N = M$) and over-determined ($N < M$) cases whereas FCA can also be applied to **underdetermined** cases ($N > M$).

In a real room environment, signals are mixed in a convolutive manner with reverberations. Frequency-domain approach, where we first apply a short-time Fourier transform (STFT) to the input time-domain signals, is effective for such convolutive mixtures. In the multiplicative transfer function (MTF) approximation [16], a convolution in the time domain is approximated as a multiplication in the frequency domain. Most of the existing frequency-domain BSS methods [12]–[15], [17]–[24] are based on the MTF approximation. However, in general cases where the room reverberation time is not too short, an STFT analysis window of typical length (e.g., 128 ms) cannot cover the dominant part of the reverberations, and the delayed source components are contaminated in the following time frames. To better cope with such cases, the convolutive transfer function **(CTF) approximation** [25] has been proposed to explicitly model the delayed components, and has recently been employed in multiple-speaker localization [26], speech separation and enhancement [27], and BSS [28], [29].

On the other hand, the CTF approximation is nothing special in blind dereverberation (BD) [30]–[34], for which weighted prediction error (WPE) [31]–[33] is a representative method. Moreover, there have been proposed many methods that combine BSS and BD [35]–[44]. Among them, the methods proposed in [36], [37] combine ICA and WPE, and the methods proposed in [39]–[44] combine FCA and WPE. Especially in [42]–[44], the idea of delayed source components has been proposed for the FCA part to adopt the CTF approximation. However, all these WPE related methods basically apply to determined and over-determined cases ($N \leq M$), since WPE assumes that the mixing system is invertible likewise ICA.

In this paper, we discuss how to extend the original FCA to improve the underdetermined BSS and BD performances in reverberant situations where CTF approximation is appropriate. A straightforward way is to employ the above-mentioned existing idea of delayed source components. However, this extension, which we call FCAd, does not take the correlation between different time frames into account. To better model the source components spanning multiple time frames, we newly propose multi-frame FCA mfFCA in this paper.

Let us explain how the original FCA is extended to FCAd and mfFCA by referring to Fig. 1. The original FCA is based on the MTF approximation, and thus the source components $\mathbf{c}_{nt}$ are forced to contain all the components including the direct one plus early reflections from the current time frame and also the late reverberant components from the previous time frames. To avoid such contaminations, we explicitly consider the delayed source components in both extensions with

H. Sawada, R. Ikeshita, and T. Nakatani are with NTT Communication Science Laboratories, NTT Corporation. K. Kinoshita was with NTT Communication Science Laboratories, NTT Corporation. He is now with Google Inc.

Manuscript received January 19, 2023; revised July 24, 2023. Earlier versions of this work were presented at ICASSP 2022 [1] and ICA 2022 [2] as conference papers.

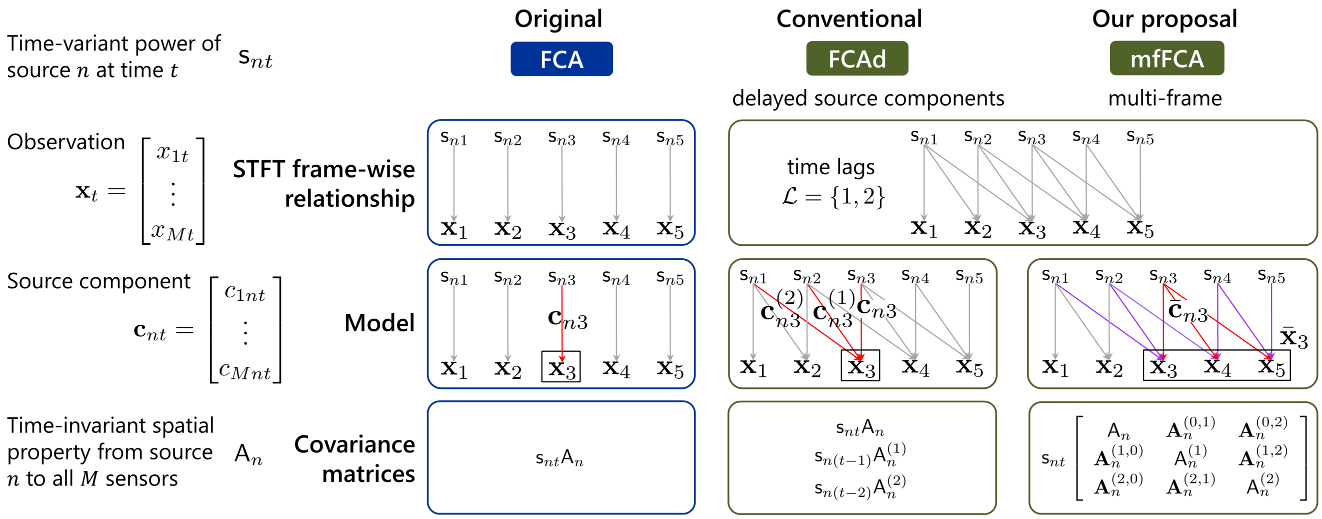

Original
FCA
Conventional
FCAd
delayed source components
Our proposal
mfFCA
multi-frame
Time-variant power of source n at time t
s_{nt}
Observation
$\mathbf{x}_t = \begin{bmatrix} x_{1t} \\ \vdots \\ x_{Mt} \end{bmatrix}$
STFT frame-wise relationship
time lags
$\mathcal{L} = \{1, 2\}$
Source component
$\mathbf{c}_{nt} = \begin{bmatrix} c_{1nt} \\ \vdots \\ c_{Mnt} \end{bmatrix}$
Model
s_{n1} s_{n2} s_{n3} s_{n4} s_{n5}
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$
$\mathbf{c}_{n3}$
$\mathbf{c}_{n3}^{(2)}$ $\mathbf{c}_{n3}^{(1)}$
$\bar{\mathbf{c}}_{n3}$
$\bar{\mathbf{x}}_3$
Time-invariant spatial property from source n to all M sensors
A_n
Covariance matrices
$\mathsf{s}_{nt}\mathsf{A}_n$
$\mathsf{s}_{nt}\mathsf{A}_n$
$\mathsf{s}_{n(t-1)}\mathsf{A}_n^{(1)}$
$\mathsf{s}_{n(t-2)}\mathsf{A}_n^{(2)}$
$\mathsf{s}_{nt} \begin{bmatrix} \mathsf{A}_n & \mathbf{A}_n^{(0,1)} & \mathbf{A}_n^{(0,2)} \\ \mathbf{A}_n^{(1,0)} & \mathsf{A}_n^{(1)} & \mathbf{A}_n^{(1,2)} \\ \mathbf{A}_n^{(2,0)} & \mathbf{A}_n^{(2,1)} & \mathsf{A}_n^{(2)} \end{bmatrix}$

Fig. 1. Illustrations of the original FCA and its conventional extension FCAd and our new extension mfFCA.

a set $\mathcal{L} = \{l_1, \ldots, l_L\}$ of time lags. However, how to model the delayed source components and their covariance matrices are different between FCAd and mfFCA. Let n and t be the indices of a source and a time frame, respectively. In FCAd, a direct $\mathbf{c}_{nt}$ and delayed $\mathbf{c}_{nt}^{(l)}$, $l \in \mathcal{L}$, components included in a single frame sensor observation $\mathbf{x}_t$ are modeled by mutually independent multivariate Gaussian distributions (see (5) and (18)) with covariance matrices $\mathsf{s}_{nt}\mathsf{A}_n$ and $\mathsf{s}_{n(t-l)}\mathsf{A}_n^{(l)}$, $l \in \mathcal{L}$, and the dimensionality remains M as the original FCA. In contrast, our proposed mfFCA concatenates the direct and delayed source components included in a multi-frame sensor observation $\bar{\mathbf{x}}_t$ ($\bar{\mathbf{x}}_3 = [\mathbf{x}_3^\mathsf{T}, \mathbf{x}_4^\mathsf{T}, \mathbf{x}_5^\mathsf{T}]^\mathsf{T}$ in Fig. 1) to make a multi-frame vector $\bar{\mathbf{c}}_{nt}$ ($\bar{\mathbf{c}}_{n3} = [\mathbf{c}_{n3}^\mathsf{T}, \mathbf{c}_{n4}^{(1)\mathsf{T}}, \mathbf{c}_{n5}^{(2)\mathsf{T}}]^\mathsf{T}$ in Fig. 1), which is modeled by a multivariate Gaussian distribution with a covariance matrix of dimensionality $M \times (L+1)$. Consequently, we newly model the correlation between different time frames on the block off-diagonal parts, e.g., $\mathbf{A}_n^{(0,1)}$, of the introduced larger covariance matrix. A similar multi-frame model has been proposed in [34]. However, there are many differences between the model and ours as will be explained in Section III-E.

The contributions of this paper are summarized as follows.

1) Multi-frame FCA mfFCA is proposed as a novel technique. Advancing from our previous work [1], we have refined the model, and provided the complete derivations of the related probabilistic models and the EM algorithm. A sample code is available at https://github.com/nttcslab-sp/mfFCA.
2) We have succeeded in separating reverberant sources that span multiple STFT time frames in fully blind underdetermined cases. The effectiveness of the proposed mfFCA over the conventional FCAd is clearly shown in the experiments with increased varieties of experimental setups from our previous work [1], [2].

In the rest of the paper, Section II reviews the original FCA and its conventional extension FCAd. Section III describes our proposed extension mfFCA. Section IV collects some issues that should be considered in practice. Section V explains the experimental settings and shows the results from various perspective. Section VI concludes the paper.

Throughout this paper, we use lower bold fonts to represent vectors, e.g., $\mathbf{c}_{nt}$. Moreover, we use Sans-serif fonts to indicate positiveness or positive definiteness: the lower and upper cases represent positive scalars, e.g., s_{nt}, and Hermitian positive definite matrices, e.g., A_n, respectively. We occasionally omit the range of summation when the space is limited and the range information is clear from the context, e.g., the right-hand side of (17).

II. Full-rank Spatial Covariance Analysis (FCA)

A. Probabilistic model

Here we define the original FCA model where $n = 1, \ldots, N$ sources are mixed and observed at $m = 1, \ldots, M$ sensors at every time frame $t \in \{1, \ldots, T\}$. Let

$$\mathbf{x}_t = [x_{1t}, \ldots, x_{Mt}]^\mathsf{T} \in \mathbb{C}^M , \tag{1}$$

$$\mathbf{c}_{nt} = [c_{1nt}, \ldots, c_{Mnt}]^\mathsf{T} \in \mathbb{C}^M \tag{2}$$

be M-dimensional complex vectors representing sensor observations and source components, respectively. For model tractability, we assume the independence of the sensor observations among different time t, i.e.,

$$p(\{\mathbf{x}_t\}_{t=1}^T \mid \theta) = \textstyle\prod_{t=1}^T p(\mathbf{x}_t \mid \theta) . \tag{3}$$

Then, the probabilistic model of FCA is specified by

$$p(\mathbf{x}_t \mid \{\mathbf{c}_{nt}\}_{n=1}^N, \theta) = \mathcal{N}(\mathbf{x}_t \mid \textstyle\sum_{n=1}^N \mathbf{c}_{nt}, \beta\mathsf{I}) , \tag{4}$$

$$p(\mathbf{c}_{nt} \mid \theta) = \mathcal{N}(\mathbf{c}_{nt} \mid \mathbf{0}, \mathsf{C}_{nt}) , \quad \mathsf{C}_{nt} = \mathsf{s}_{nt}\mathsf{A}_n , \tag{5}$$

where $\mathcal{N}$ represents a multivariate complex proper and circular Gaussian distribution with a Hermitian positive covariance matrix [45]

$$\mathcal{N}(\mathbf{c} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\pi^M \det \Sigma} \exp\left[-(\mathbf{c} - \boldsymbol{\mu})^* \Sigma^{-1} (\mathbf{c} - \boldsymbol{\mu})\right] ,$$

and

$$\theta = \{\{\mathsf{s}_{nt}\}_{t=1}^T, \mathsf{A}_n\}_{n=1}^N \tag{6}$$

is a set of parameters to be optimized. The parameters s_{nt} and A_n represent the time-variant power of source n at time frame t and the time-invariant spatial property from source n to all M sensors, respectively. The sensor noise is assumed to be uncorrelated among sensors and represented by an identity matrix I and the noise power parameter β, which can be predefined, e.g., as $\beta = 10^{-3}$.

In the FCA model, the component vectors are assumed to be independent of each other:

$$p(\{\mathbf{c}_{nt}\}_{n=1}^N \mid \theta) = \textstyle\prod_{n=1}^N p(\mathbf{c}_{nt} \mid \theta)\,. \tag{7}$$

Then, the joint distribution $p(\mathbf{x}_t, \{\mathbf{c}_{nt}\}_{n=1}^N \mid \theta)$ turns out to be a zero-mean Gaussian distribution with a covariance matrix

$$\begin{bmatrix} \mathsf{X}_t & \mathsf{C}_{1t} & \cdots & \mathsf{C}_{Nt} \\ \mathsf{C}_{1t} & \mathsf{C}_{1t} & & \mathsf{0} \\ \vdots & & \ddots & \\ \mathsf{C}_{Nt} & \mathsf{0} & & \mathsf{C}_{Nt} \end{bmatrix} \tag{8}$$

with

$$\mathsf{X}_t = \textstyle\sum_{n=1}^N \mathsf{C}_{nt} + \beta\mathsf{I}\,, \tag{9}$$

as shown in Appendix A. Once the joint distribution is obtained, it is easy to derive the marginal and conditional distributions, both of which are also Gaussian distributions [46]. For example, the marginal distribution $p(\mathbf{x}_t \mid \theta)$ is given as

$$p(\mathbf{x}_t \mid \theta) = \mathcal{N}(\mathbf{x}_t \mid \mathbf{0}, \mathsf{X}_t)\,. \tag{10}$$

B. Objective function and EM algorithm

The parameters in θ are optimized according to observations $\{\mathbf{x}_t\}_{t=1}^T$ by maximizing the objective function, which is the log-likelihood $\ln p(\{\mathbf{x}_t\}_{t=1}^T \mid \theta)$ with the assumption (3)

$$\textstyle\sum_{t=1}^T \ln p(\mathbf{x}_t \mid \theta)\,. \tag{11}$$

The objective function can be locally maximized [12], [13] by the EM algorithm [47] from some initial values of parameters. In the **E-step**, we calculate the conditional distributions of the component vector $\mathbf{c}_{nt}$ as

$$p(\mathbf{c}_{nt} \mid \mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{c}_{nt} \mid \boldsymbol{\mu}_{nt}, \Sigma_{nt})\,, \tag{12}$$

$$\boldsymbol{\mu}_{nt} = \mathsf{C}_{nt}\mathsf{X}_t^{-1}\mathbf{x}_t\,,\ \ \Sigma_{nt} = \mathsf{C}_{nt} - \mathsf{C}_{nt}\mathsf{X}_t^{-1}\mathsf{C}_{nt}\,. \tag{13}$$

In the **M-step**, we optimize the parameters by maximizing the so-called $\mathcal{Q}$ function (62) derived in Appendix B as

$$\mathsf{s}_{nt} \leftarrow \frac{1}{M}\mathrm{tr}\left(\mathsf{A}_n^{-1}\widetilde{\mathsf{C}}_{nt}\right)\,, \tag{14}$$

$$\mathsf{A}_n \leftarrow \frac{1}{T}\sum_{t=1}^T \mathsf{s}_{nt}^{-1}\widetilde{\mathsf{C}}_{nt}\,, \tag{15}$$

where tr calculates the trace of a matrix, and

$$\widetilde{\mathsf{C}}_{nt} = \boldsymbol{\mu}_{nt}\boldsymbol{\mu}_{nt}^* + \Sigma_{nt} \tag{16}$$

is derived in (63). Throughout this paper, we use $\cdot^*$ notation for representing a conjugate transpose operation.

Once the parameters θ are optimized, we obtain separated signals $\mathbf{y}_{nt}$ simply by getting $\boldsymbol{\mu}_{nt}$ from (13), as the result of multichannel Wiener filter.

C. Considering delayed source components

In this Subsection, we introduce time lags l and delayed source components $\mathbf{c}_{nt}^{(l)}$ to explicitly model the room reverberations that an STFT frame cannot cover. Then, we explain a conventional FCA extension FCAd whose key ideas have already been proposed in [42]–[44].

As shown in the FCAd column of Fig. 1, a delayed source component $\mathbf{c}_{nt}^{(l)}$ comes from the source n emitted at time frame $t-l$ and observed at time frame t through the time lag l. Let $\mathcal{L} = \{l_1, \ldots, l_L\}$ be the set of time lags. For notational convenience, let $\mathbf{c}_{nt}^{(0)} = \mathbf{c}_{nt}$, $\mathsf{A}_n^{(0)} = \mathsf{A}_n$, and $\mathcal{L}_0 = \{0\} \cup \mathcal{L}$. Then, the FCAd model is described as

$$p(\mathbf{x}_t \mid \{\{\mathbf{c}_{nt}^{(l)}\}_{l\in\mathcal{L}_0}\}_{n=1}^N, \theta) = \mathcal{N}(\mathbf{x}_t \mid \textstyle\sum_n\sum_l \mathbf{c}_{nt}^{(l)}, \beta\mathsf{I})\,, \tag{17}$$

$$p(\mathbf{c}_{nt}^{(l)} \mid \theta) = \mathcal{N}(\mathbf{c}_{nt}^{(l)} \mid \mathbf{0}, \mathsf{C}_{nt}^{(l)})\,,\ \ \mathsf{C}_{nt}^{(l)} = \mathsf{s}_{n(t-l)}\mathsf{A}_n^{(l)}\,, \tag{18}$$

with a new set of parameters

$$\theta = \{\{\mathsf{s}_{nt}\}_{t=1}^T, \{\mathsf{A}_n^{(l)}\}_{l\in\mathcal{L}_0}\}_{n=1}^N\,, \tag{19}$$

in contrast to the original FCA model (4)–(6). The newly introduced parameters $\mathsf{A}_n^{(l)}, l \in \mathcal{L}$ encode the spatial property of source n affecting to all M sensors with time lag l. By the introduction of delayed source components $\mathbf{c}_{nt}^{(l)}$ with $l \in \mathcal{L}$, the physical meaning of $\mathbf{c}_{nt}$ has been changed. In the original FCA, $\mathbf{c}_{nt}$ is forced to contain the late reverberant components from the previous time frames which are not explicitly modeled. On the other hand in FCAd, and also in mfFCA whose details will be explained in Section III, $\mathbf{c}_{nt}$ contains only the direct component plus early reflections, and not forced to contain the late reverberations any more.

Although we omit the detailed derivations, the specific form of the joint distribution $p(\mathbf{x}_t, \{\{\mathbf{c}_{nt}^{(l)}\}_{l\in\mathcal{L}_0}\}_{n=1}^N \mid \theta)$ can be obtained as in the case (8) of FCA. Consequently, the marginal distribution $p(\mathbf{x}_t \mid \theta)$ is a zero-mean Gaussian distribution with a covariance matrix

$$\mathsf{X}_t = \textstyle\sum_{n=1}^N \sum_{l\in\mathcal{L}_0} \mathsf{C}_{nt}^{(l)} + \beta\mathsf{I}\,. \tag{20}$$

The parameters in θ are optimized by the EM algorithm as in the case of FCA. In the **E-step**, we calculate the conditional distributions of the component vector $\mathbf{c}_{nt}^{(l)}$ as

$$p(\mathbf{c}_{nt}^{(l)} \mid \mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{c}_{nt}^{(l)} \mid \boldsymbol{\mu}_{nt}^{(l)}, \Sigma_{nt}^{(l)})\,, \tag{21}$$

$$\boldsymbol{\mu}_{nt}^{(l)} = \mathsf{C}_{nt}^{(l)}\mathsf{X}_t^{-1}\mathbf{x}_t\,,\ \ \Sigma_{nt}^{(l)} = \mathsf{C}_{nt}^{(l)} - \mathsf{C}_{nt}^{(l)}\mathsf{X}_t^{-1}\mathsf{C}_{nt}^{(l)}\,. \tag{22}$$

In the **M-step**, we optimize the parameters as

$$\mathsf{s}_{nt} \leftarrow \frac{1}{M(L+1)}\sum_{l\in\mathcal{L}_0} \mathrm{tr}\left[\left(\mathsf{A}_n^{(l)}\right)^{-1}\widetilde{\mathsf{C}}_{nt}^{(l)}\right]\,, \tag{23}$$

$$\mathsf{A}_n^{(l)} \leftarrow \frac{1}{T}\sum_{t=1}^T \mathsf{s}_{nt}^{-1}\widetilde{\mathsf{C}}_{nt}^{(l)}\,, \tag{24}$$

where

$$\widetilde{\mathsf{C}}_{nt}^{(l)} = \boldsymbol{\mu}_{nt}^{(l)}\boldsymbol{\mu}_{nt}^{(l)*} + \Sigma_{nt}^{(l)}\,. \tag{25}$$

After the parameters θ are optimized, we obtain separated signals $\mathbf{y}_{nt}$ by accumulating all the components originate from source n and observed at time frame t as

$$\mathbf{y}_{nt} = \boldsymbol{\mu}_{nt}^{(0)} + \textstyle\sum_{i=1}^L \boldsymbol{\mu}_{n(t-l_i)}^{(l_i)}\,. \tag{26}$$

To perform dereverberation further, we obtain dereverberated separated signals $\mathbf{d}_{nt}$ as

$$\mathbf{d}_{nt} = \boldsymbol{\mu}_{nt}^{(0)} \tag{27}$$

by eliminating the delayed components, i.e., the second term of (26).

III. Multi-frame FCA

To better model the source component $\mathbf{c}_{nt}$ with its delayed versions $\{\mathbf{c}_{n(t+l)}^{(l)}\}_{l\in\mathcal{L}}$, we propose a new extension of FCA, multi-frame FCA (mfFCA), that considers the correlation between, e.g., $\mathbf{c}_{nt}$ and $\mathbf{c}_{n(t+1)}^{(1)}$, both of which originate from the same source n at the same time frame t but are observed at adjacent two frames.

A. Multi-frame vectors

Let us introduce multi-frame vectors for the sensor observations and the source components

$$\bar{\mathbf{x}}_t = [\mathbf{x}_t^\mathsf{T}, \mathbf{x}_{t+l_1}^\mathsf{T}, \dots, \mathbf{x}_{t+l_L}^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{M(L+1)}\,, \tag{28}$$

$$\bar{\mathbf{c}}_{nt} = [\mathbf{c}_{nt}^\mathsf{T}, \mathbf{c}_{n(t+l_1)}^{(l_1)\mathsf{T}}, \dots, \mathbf{c}_{n(t+l_L)}^{(l_L)\mathsf{T}}]^\mathsf{T} \in \mathbb{C}^{M(L+1)} \tag{29}$$

for a set $\mathcal{L} = \{l_1, \dots, l_L\}$ of time lags, respectively. The mfFCA column of Fig. 1 shows $\bar{\mathbf{x}}_3$ and $\bar{\mathbf{c}}_{n3}$ with $\mathcal{L} = \{1, 2\}$ as examples. In the rest of this subsection, we detail the source component vector $\bar{\mathbf{c}}_{nt}$. The complete description of the sensor observation vector $\bar{\mathbf{x}}_t$ will be given in the next subsection.

The multi-frame component vector $\bar{\mathbf{c}}_{nt}$ is defined in (29) by concatenating all the delayed versions originating from the same source n with the same time instant t. To reflect this, we assume that $\bar{\mathbf{c}}_{nt}$ follows a zero-mean Gaussian distribution with covariance matrix

$$\bar{\mathsf{C}}_{nt} = \mathsf{s}_{nt}\bar{\mathsf{A}}_n\,, \quad \bar{\mathsf{A}}_n = \begin{bmatrix} \mathsf{A}_n^{(0)} & \dots & \mathbf{A}_n^{(0,l_L)} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_n^{(l_L,0)} & \dots & \mathsf{A}_n^{(l_L)} \end{bmatrix}. \tag{30}$$

$\bar{\mathsf{A}}_n$ is of size $M(L+1) \times M(L+1)$ and encodes the time-invariant spatial property from source n to all M sensors with all the considered time lags $\mathcal{L}_0 = \{0\} \cup \mathcal{L}$ including 0-lag. We have already seen the block diagonal submatrices $\mathsf{A}_n^{(0)}, \dots, \mathsf{A}_n^{(l_L)}$ in (5) and (18). The newly introduced block off-diagonal submatrices $\mathbf{A}_n^{(l,l')}$ satisfy the conjugate transpose relationship $\left(\mathbf{A}_n^{(l,l')}\right)^* = \mathbf{A}_n^{(l',l)}$. They represent the correlation of the source components $\mathbf{c}_{n(t+l)}^{(l)}$ and $\mathbf{c}_{n(t+l')}^{(l')}$ that originate from the same source n and time instant t but observed at different time frames according to the time lags l and l'. The existing FCA model and its extension, i.e., FCAd with (17) and (18), do not consider such block off-diagonal submatrices, and ignore such aforementioned frame-wise correlations based on how sources propagate to all sensors with reverberations.

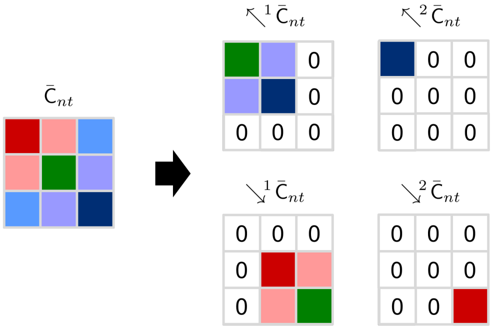

$\bar{\mathsf{C}}_{nt}$
$\nwarrow^1 \bar{\mathsf{C}}_{nt}$
$\nwarrow^2 \bar{\mathsf{C}}_{nt}$
$\searrow^1 \bar{\mathsf{C}}_{nt}$
$\searrow^2 \bar{\mathsf{C}}_{nt}$
0

Fig. 2. Diagonal shift operators $\nwarrow^i$ and $\searrow^i$ applied to a covariance matrix $\bar{\mathsf{C}}_{nt}$ of size $M(L+1) \times M(L+1)$. Each small box represents an $M \times M$ submatrix.

B. Probabilistic model

In this subsection, we develop the probabilistic model of mfFCA. We assume the independence of the multi-frame observations among different time t, i.e.,

$$p(\{\bar{\mathbf{x}}_t\}_{t=1}^{T-l_L} \mid \theta) = \textstyle\prod_{t=1}^{T-l_L} p(\bar{\mathbf{x}}_t \mid \theta)\,, \tag{31}$$

similarly with (3). Then, we will discuss the relationship between $\bar{\mathbf{c}}_{nt}$ and $\bar{\mathbf{x}}_t$. In the existing FCA and FCAd, the relationship between $\mathbf{c}_{nt}^{(l)}$ and $\mathbf{x}_t$ is clear as the simple summation models (4) and (17). This simplicity is represented in Fig. 1 as all the incoming arrows into $\mathbf{x}_3$ are related to $\mathbf{c}_{n3}^{(l)}$, $l = 0, 1, 2$ that are colored in red. On the other hand, $\bar{\mathbf{x}}_3$ of mfFCA has incoming purple arrows in addition to the red arrows that correspond to $\bar{\mathbf{c}}_{n3}$. We thus model such purple arrows by an $M(L+1) \times M(L+1)$ covariance matrix

$$\bar{\mathsf{D}}_t = \textstyle\sum_{n=1}^N \sum_{i=1}^L \left(\nwarrow^i \bar{\mathsf{C}}_{n(t-l_i)} + \searrow^i \bar{\mathsf{C}}_{n(t+l_i)} \right) \tag{32}$$

where $\nwarrow^i$ and $\searrow^i$ are newly introduced operators that diagonally shift the submatrices of size $M \times M$ as shown in Fig. 2. With this notation, the summation of the purple arrows coming into $\bar{\mathbf{x}}_3$ shown in Fig. 1 is given by $\bar{\mathsf{D}}_3 = \sum_n \left(\nwarrow^2 \bar{\mathsf{C}}_{n1} + \nwarrow^1 \bar{\mathsf{C}}_{n2} + \searrow^1 \bar{\mathsf{C}}_{n4} + \searrow^2 \bar{\mathsf{C}}_{n5} \right)$.

To summarize the discussion of this Section so far, we specify the probabilistic models of $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{c}}_{nt}$ as:

$$p(\bar{\mathbf{x}}_t \mid \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N, \theta) = \mathcal{N}(\bar{\mathbf{x}}_t \mid \textstyle\sum_{n=1}^N \bar{\mathbf{c}}_{nt}, \bar{\mathsf{D}}_t + \beta\bar{\mathsf{I}})\,, \tag{33}$$

$$p(\bar{\mathbf{c}}_{nt} \mid \theta) = \mathcal{N}(\bar{\mathbf{c}}_{nt} \mid \mathbf{0}, \bar{\mathsf{C}}_{nt})\,, \quad \bar{\mathsf{C}}_{nt} = \mathsf{s}_{nt}\bar{\mathsf{A}}_n\,, \tag{34}$$

with an identity matrix $\bar{\mathsf{I}}$ of size $M(L+1) \times M(L+1)$ and the set of parameters to be optimized

$$\theta = \{\{\mathsf{s}_{nt}\}_{t=1}^T, \bar{\mathsf{A}}_n\}_{n=1}^N\,. \tag{35}$$

These three equations are similar to (4) - (6) of FCA, but the differences are in the dimensionality and the introduction of the covariance matrix (32).

The following four equations are also similar to (7) - (10) of FCA. The multi-frame component vectors are assumed to be independent of each other:

$$p(\{\bar{\mathbf{c}}_{nt}\}_{n=1}^N \mid \theta) = \textstyle\prod_{n=1}^N p(\bar{\mathbf{c}}_{nt} \mid \theta)\,. \tag{36}$$

Then, the joint distribution $p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N \mid \theta)$ turns out to be a zero-mean Gaussian distribution with a covariance matrix

$$\begin{bmatrix} \bar{\mathsf{X}}_t & \bar{\mathsf{C}}_{1t} & \cdots & \bar{\mathsf{C}}_{Nt} \\ \bar{\mathsf{C}}_{1t} & \bar{\mathsf{C}}_{1t} & & \bar{\mathsf{0}} \\ \vdots & & \ddots & \\ \bar{\mathsf{C}}_{Nt} & \bar{\mathsf{0}} & & \bar{\mathsf{C}}_{Nt} \end{bmatrix} \tag{37}$$

with

$$\bar{\mathsf{X}}_t = \textstyle\sum_{n=1}^N \bar{\mathsf{C}}_{nt} + \bar{\mathsf{D}}_t + \beta\bar{\mathsf{I}}\,, \tag{38}$$

as Appendix C describes. Therefore, $\bar{\mathbf{x}}_t$ also follows a zero-mean Gaussian distribution

$$p(\bar{\mathbf{x}}_t \mid \theta) = \mathcal{N}(\bar{\mathbf{x}}_t \mid \mathbf{0}, \bar{\mathsf{X}}_t)\,. \tag{39}$$

C. Objective function and EM algorithm

The parameters θ of (35) are optimized according to multi-frame observations $\{\bar{\mathbf{x}}_t\}_{t=1}^{T-l_L}$ by maximizing the objective function, which is the log-likelihood $\ln p(\{\bar{\mathbf{x}}_t\}_{t=1}^{T-l_L} \mid \theta)$ with the assumption (31)

$$\textstyle\sum_{t=1}^{T-l_L} \ln p(\bar{\mathbf{x}}_t \mid \theta)\,. \tag{40}$$

The EM algorithm to maximize the objective function (40) from some initial values of parameters has been derived as follows. In the **E-step**, we calculate the conditional distributions of the multi-frame component vector $\bar{\mathbf{c}}_{nt}$ as

$$p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta) = \mathcal{N}(\bar{\mathbf{c}}_{nt} \mid \bar{\boldsymbol{\mu}}_{nt}, \bar{\mathsf{\Sigma}}_{nt})\,, \tag{41}$$

$$\bar{\boldsymbol{\mu}}_{nt} = \bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}\bar{\mathbf{x}}_t\,,\ \ \bar{\mathsf{\Sigma}}_{nt} = \bar{\mathsf{C}}_{nt} - \bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}\bar{\mathsf{C}}_{nt}\,. \tag{42}$$

In the **M-step**, we optimize the parameters by maximizing the so-called $\mathcal{Q}$ function (71) derived in Appendix D. Since the $\mathcal{Q}$ function is complicated according to the existence of $\bar{\mathsf{D}}_t$ defined in (32), we introduce an approximation that the parameter values used in $\bar{\mathsf{D}}_t$ are fixed at the previous parameter values in θ' when optimizing the parameters. Consequently, the parameters are updated as

$$\mathsf{s}_{nt} \leftarrow \frac{1}{M(L{+}1)} \mathrm{tr}\left[\bar{\mathsf{A}}_n^{-1}\widetilde{\widetilde{\bar{\mathsf{C}}}}_{nt}\right]\,, \tag{43}$$

$$\bar{\mathsf{A}}_n \leftarrow \frac{1}{T}\sum_{t=1}^{T-l_L} \frac{1}{\mathsf{s}_{nt}} \widetilde{\widetilde{\bar{\mathsf{C}}}}_{nt} \tag{44}$$

with

$$\widetilde{\widetilde{\bar{\mathsf{C}}}}_{nt} = \bar{\boldsymbol{\mu}}_{nt}\bar{\boldsymbol{\mu}}_{nt}^* + \bar{\mathsf{\Sigma}}_{nt} \tag{45}$$

derived in (73). The plausibility of the approximated EM algorithm will be demonstrated in Section V. The convergence analysis of the approximated EM algorithm has not been done. This is left as future work. Or, we aim to invent a perfect algorithm without approximation as another future work.

D. Source separation and dereverberation

After the parameters θ are optimized by the EM algorithm, we obtain separated signals $\mathbf{y}_{nt}$ and further dereverberated signals $\mathbf{d}_{nt}$ for sources $n = 1, \ldots, N$ in the following manner. First, we apply the multi-frame multichannel Wiener filter $\bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}$ to multi-frame observation vectors $\bar{\mathbf{x}}_t$ as in (42) to obtain $\bar{\boldsymbol{\mu}}_{nt}$. Note that our definition of the multi-frame multichannel Wiener filter differs from that in [48]. In particular, our multi-frame multichannel Wiener filter is defined as the posterior mean of the multivariate Gaussian distribution developed in Subsection III-B. Then, we introduce a $\sqcap_i$ operator that extracts the $(i+1)$-th single-frame vector from a multi-frame vector (see Fig. 3). To perform source separation, we accumulate all the components originating from source n and observed at time frame t as

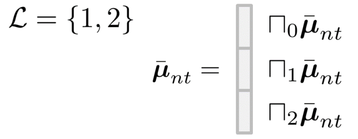

$\mathcal{L} = \{1, 2\}$
$\bar{\boldsymbol{\mu}}_{nt} =$
$\sqcap_0\bar{\boldsymbol{\mu}}_{nt}$
$\sqcap_1\bar{\boldsymbol{\mu}}_{nt}$
$\sqcap_2\bar{\boldsymbol{\mu}}_{nt}$

Fig. 3. $\sqcap_i$ operators applied to a multi-frame vector $\bar{\boldsymbol{\mu}}_{nt}$ of dimension $M(L+1)$. Each grey rectangular represents an M dimensional single-frame vector.

$$\mathbf{y}_{nt} = \sqcap_0\bar{\boldsymbol{\mu}}_{nt} + \textstyle\sum_{i=1}^L \sqcap_i\bar{\boldsymbol{\mu}}_{n(t-l_i)}\,. \tag{46}$$

To perform dereverberation further, we eliminate the delayed components, i.e., the second term of (46) to obtain

$$\mathbf{d}_{nt} = \sqcap_0\bar{\boldsymbol{\mu}}_{nt}\,. \tag{47}$$

E. Model differences to [34]

Having described the detailed model of the proposed multi-frame FCA, let us discuss the differences to a similar multi-frame model proposed in [34]. The most notable difference is in the covariance matrices regarding the purple arrows shown in Fig. 1. In our mfFCA, these purple arrows are modeled by $\bar{\mathsf{D}}_t$ in (32), in which shifted components in both directions by $\nwarrow^i$ and $\searrow^i$, illustrated in Fig. 2, are considered. On the other hand in [34], only the shifted components in either direction are considered. As will be experimentally shown in Subsection V-F, shifts in both directions clearly performed better than shifts in either direction. Therefore, the multi-frame model proposed in this paper is novel and effective for precisely modeling a real-world acoustic situations.

IV. Practical issues

We have intentionally omitted some practical aspects of FCA methods in the previous two sections to make the explanations as simple as possible. In this section, we explain some practical issues that should be considered to separate and dereverberate actual sound mixtures effectively.

A. Full-band processing and permutation alignment

In the previous two sections, we omit frequency dependency f for notational simplicity, and denote m-th microphone observations at time frame t as x_{mt}. For practical separation and dereverberation tasks, we actually have microphone observations x_{mtf} for frequency bins $f = 1, \ldots, F$ as the results of STFT. Thus, we need to perform $f = 1, \ldots, F$ separate FCA executions for the STFT results $\mathbf{x}_{tf}, t = 1, \ldots, T$.

However, it is essential to relate the F FCA executions so that n-th source in every frequency bin f corresponds to the same source. This is known as the permutation problem as in the case of ICA [49]. There are three major approaches to align the permutation ambiguities of ICA or FCA solutions. The first one is post-processing [50]. The second one is by sharing the source power parameters s_{ntf} among frequency bins [19]–[21]. The third one is by modeling the source power parameters with nonnegative matrix factorization (NMF) [22]–[24], which is especially effective for music separation. In the experiments for speech separation reported in Section V, we combined the first and second approaches. The details are explained in Subsection V-E.

B. Parameter regularization

The parameters (6), (19) and (35) of FCA models tend to be overfit as reported in [51]. More specifically, the spatial matrices A_{nf}, $\mathsf{A}^{(l)}_{nf}$, and $\bar{\mathsf{A}}_{nf}$ tend to be pushed towards rank deficient by maximizing the likelihood. To avoid the overfitting, we regularize the parameters in two ways.

The first one is flooring. Let ϵ be a flooring parameter, e.g., $\epsilon = 10^{-4}$. We apply the following additional updates

$$\mathsf{s}_{ntf} \leftarrow \max(\mathsf{s}_{ntf}, \epsilon) \tag{48}$$

after (14), (23) and (43), and

$$\mathsf{A}_{nf} \leftarrow \mathsf{A}_{nf} + \epsilon\mathsf{I} \tag{49}$$

$$\mathsf{A}^{(l)}_{nf} \leftarrow \mathsf{A}^{(l)}_{nf} + \epsilon\mathsf{I} \tag{50}$$

$$\bar{\mathsf{A}}_{nf} \leftarrow \bar{\mathsf{A}}_{nf} + \epsilon\bar{\mathsf{I}} \tag{51}$$

after (15), (24) and (44), respectively.

The second one is sharing the source power parameters s_{ntf} among adjacent frequency bins. Let the total F frequency bins be partitioned into B blocks $\mathcal{F}_b,\ b = 1, \dots, B$ in a consecutive and disjoint manner, e.g., $F = 12$, $B = 3$, $\mathcal{F}_1 = \{1, 2, 3, 4\}, \mathcal{F}_2 = \{5, 6, 7, 8\}, \mathcal{F}_3 = \{9, 10, 11, 12\}$. We regularize s_{ntf} to be the same in a block $\mathcal{F}_b$, e.g., $b = 1$, $\mathsf{s}_{nt1} = \mathsf{s}_{nt2} = \mathsf{s}_{nt3} = \mathsf{s}_{nt4}$. This can be implemented by averaging the parameters in a block $\mathcal{F}_b$

$$\mathsf{s}_{ntf} \leftarrow \frac{1}{|\mathcal{F}_b|} \sum_{f \in \mathcal{F}_b} \mathsf{s}_{ntf}\ , \quad f \in \mathcal{F}_b \tag{52}$$

after the updates (14), (23) and (43), or after (48) if we employ the first regularization. The parameter sharing might seem to be a very strong constraint, but actually not as we have seen that independent vector analysis (IVA) [19]–[21] work very well by sharing the parameters even among all the frequency bins. Thus, it is effective also for permutation alignment as the second approach explained in the last Subsection.

C. Spatial whitening

Making the vector of sensor observations spatially white [5] is effective for the FCA methods to work robustly. Intuitively speaking, it makes the time-invariant spatial properties A_{nf}, $\mathsf{A}^{(l)}_{nf}$, and $\bar{\mathsf{A}}_{nf}$ significantly different among different sources n. Without it, the spatial properties would be very similar among different sources n especially in low frequencies f where the phase differences between microphones are small. And there remains a risk of inaccurate estimation for the spatial properties. Spatial whitening can be performed by

$$\mathbf{x}_{tf} \leftarrow \mathbf{V}_f\, \mathbf{x}_{tf} \tag{53}$$

with the whitening matrix $\mathbf{V}_f$ calculated by $\mathbf{V}_f = \mathbf{D}^{-1/2}\mathbf{E}^*$ where $\mathbf{D}$ and $\mathbf{E}$ contain the eigenvalues and eigenvectors of $\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{tf}\mathbf{x}_{tf}^* = \mathbf{EDE}^*$, respectively. Once we employ the spatial whitening as a preprocessing, we need a corresponding post-processing to be applied to separated signals (26) and (46) and further dereverberated signals (27) and (47) as

$$\mathbf{y}_{ntf} \leftarrow \mathbf{V}_f^{-1}\, \mathbf{y}_{ntf}\ , \tag{54}$$

$$\mathbf{d}_{ntf} \leftarrow \mathbf{V}_f^{-1}\, \mathbf{d}_{ntf}\ . \tag{55}$$

D. Equally-spaced condition for time lags

We experimentally confirmed the following condition. For the separated signal construction (26) and (46) of the FCA extensions to work correctly, the time lags in a set $\mathcal{L}$ should be equally spaced originating from the 0 lag, whose corresponding terms are $\boldsymbol{\mu}^{(0)}_{nt}$ and $\sqcap_0\bar{\boldsymbol{\mu}}_{nt}$. For example, $\mathcal{L} = \{1, 2\}$, $\mathcal{L} = \{2\}$, and $\mathcal{L} = \{2, 4\}$ work fine, but $\mathcal{L} = \{1, 3\}$ and $\mathcal{L} = \{2, 3\}$ do not.

V. Experiments

A. Conditions and tasks

We performed experiments to separate and dereverberate N speech sources with M microphones. The combinations (M, N) of M and N were (2,3), (3,3), and (3,4), with the second one being a determined case while the others being underdetermined cases. For each (M, N), we have tested 16 combinations of N speech sources to mix. We measured the impulse responses from the sources (loudspeakers) to the microphones under the room conditions shown in Fig. 4. We varied the room reverberation time from 130 ms to 450 ms by attaching or detaching cushion walls. For each condition, i.e., (M, N), N source combinations, and reverberation time, time-domain mixtures at the microphones were constructed by convolving the impulse responses and N 6-second English speech sources and then mixing the convolution results. We believe that the experimental variations described above were sufficient to validate the performance of the FCA methods. This is because we worked on blind tasks where estimated parameter values varied from case to case, and thus overfitting to the whole experimental variations hardly occurred.

In order to evaluate separation and dereverberation performances, we made two types of time-domain source n images img_{mn} and $\mathrm{img}^{(cut)}_{mn}$ at each microphone m. The former was made by convolving source n and the corresponding impulse response. The latter was made by convolving source n and the corresponding impulse response that was cut to 64 ms. We chose this specific duration of 64 ms because we used a 128 ms Hann window for the STFT, as will be explained in Subsection V-B, and not because we intended to define that the transition from early echoes to late reverberation occurs at 64 ms.

Relating to the two types of source images, we set two types of tasks. The first task was blind source separation (BSS)

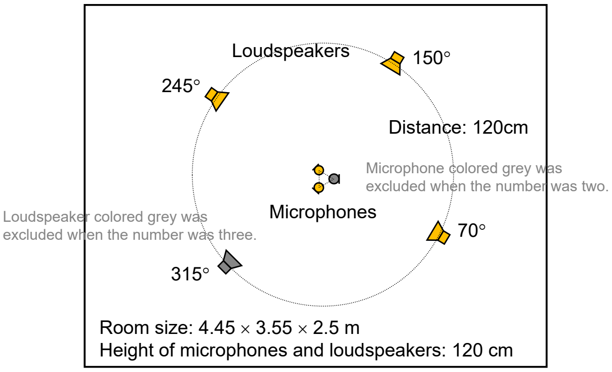

Loudspeakers
150°
245°
Distance: 120cm
Microphone colored grey was excluded when the number was two.
Loudspeaker colored grey was excluded when the number was three.
Microphones
70°
315°
Room size: 4.45 × 3.55 × 2.5 m
Height of microphones and loudspeakers: 120 cm

Fig. 4. Experimental setup

aiming at separated signals $\mathbf{y}_{nt}$ defined in (26) and (46), whose time-domain representation should be close to img_{mn}. The second task was blind source separation and dereverberation (BSS+BD) aiming at separated and dereverberated signals $\mathbf{d}_{nt}$ defined in (27) and (47), whose time-domain representation should be close to $\mathrm{img}^{(cut)}_{mn}$. Note that in the original FCA, $\mathbf{y}_{nt} = \mathbf{d}_{nt}$ by definition. The performances were measured in terms of the Signal to Distortion Ratio (SDR) [52] by setting img_{mn} and $\mathrm{img}^{(cut)}_{mn}$ as reference signals for BSS and BSS+BD, respectively. The notion of SDR is decomposed into the source Image to Spatial distortion Ratio (ISR), Source to Interference Ratio (SIR), and Source to Artifacts Ratio (SAR). For the calculations of SDR, ISR, SIR and SAR, we used the MUSEVAL V4 toolkit [53] with bss_eval_images configuration.

B. Methods

We examined the original FCA (Subsection II-A), the conventional FCAd (Subsection II-C), and the proposed mfFCA (Section III). The following settings were common in all the methods. The sampling frequency was 8 kHz. The STFT window width and shift were 1024 and 256 samples, i.e., 128 ms and 32 ms, respectively. Consequently, the numbers of time frames and frequency bins were $T = 201$ and $F = 513$, respectively. As shown in [54], the finer the shift amount, the better the separation performance. We chose this quarter-shift scheme to balance the performance and the computational complexity.

After applying spatial whitening (Subsection IV-C), the FCA parameters (6), (19) and (35) were initialized by the procedure shown in [55]. The initialization procedures of FCAd and mfFCA were exactly the same as that of FCA, as will be explained in the first paragraph of Subsection V-E. The noise power parameter was $\beta = 10^{-3}$. The flooring parameter was $\epsilon = 10^{-4}$. The size of blocks for parameter sharing was four $|\mathcal{F}_b| = 4$. The number of iterations for the EM algorithms was 500.

Regarding FCAd and mfFCA, we prepared three sets of time lags $\mathcal{L} = \{2\}$, $\mathcal{L} = \{2, 4\}$ and $\mathcal{L} = \{2, 4, 6\}$. According to the discussion of Subsection IV-D, a set $\mathcal{L} = \{1, 2, 3, 4\}$ of time lags would be fine to replace $\mathcal{L} = \{2, 4\}$, for example. However, we employed these three sets of time lags with only even numbers. The reason was again to balance the performance and the computational complexity. We chose the quarter-shift scheme with a Hann window for the STFT as described above. In this setting, adjacent STFT frames were 75 % overlapped, and skipping one frame still had 50 % overlap. Imagining the shape of a Hann window, time lags corresponding to 50 % overlap were enough to cover the past signals of interest for modeling reverberations.

C. Results of FCA methods

Figure 5 shows the BSS and BSS+BD results of the FCA methods under various reverberation times measured in SDR, ISR, SIR, and SAR. We examined 7 methods as listed in the right top corner of Fig. 5. And only for the SDR as a representative measure, Fig. 6 shows the differences from the baseline, which was FCA, to highlight how much improvements were achieved by the FCA extensions.

From these two figures, we observe the followings. The improvements by the extensions FCAd and mfFCA from the baseline were apparent except the lowest reverberant 130 ms cases. Comparing these two extensions, the proposed mfFCA clearly outperformed the conventional FCAd. From Fig. 5, we observe that mfFCA achieved clearly higher SIRs than FCAd in both BSS and BSS+BD tasks, and clearly higher ISRs in BSS tasks. Regarding the three sets of time lags, Fig. 6 clearly shows that mfFCA maximized the performance by employing an adequate set of time lags according to a reverberation time, e.g., $\mathcal{L} = \{2\}$ for 200 ms and $\mathcal{L} = \{2, 4, 6\}$ for 450 ms, whereas the results of FCAd were not affected much by the differences of the time lag sets. We believe that the introduction of the block off-diagonal parts, e.g., $\mathbf{A}_n^{(0,1)}$ in Fig. 1, contributes to better model reverberant situations. Separated sound examples produced by the original FCA and the proposed mfFCA with $\mathcal{L} = \{2, 4, 6\}$ for a case of 450 ms reverberation time can be heard at [56].

Comparing the improvements by the proposed mfFCA between the BSS and BSS+BD tasks, we observed that the improvements for the BSS tasks was more prominent than the improvements for the BSS+BD tasks. Therefore, we tried to improve the results of the BSS+BD tasks further by WPE as will be reported next.

D. WPE preprocessing for the BSS+BD tasks

We then report the results of the BSS+BD tasks when we employed WPE [31]–[33] as a preprocessing of FCA and mfFCA. WPE is a well-established BD method for over-determined mixtures. However, even for underdetermined mixtures, we expect that WPE removes reverberations to a certain degree despite that the number M of microphones is insufficient for the number N of sources.

Figure 7 shows the results. The horizontal axis in each plot corresponds to the set of time lags for WPE. We examined three sets of time lags $\mathcal{L} = \{2, 3, 4, 5, 6\}$, $\mathcal{L} = \{2, 3, 4, 5, 6, 7, 8\}$ and $\mathcal{L} = \{4, 5, 6, 7, 8\}$. Note that for WPE we did not have to care about the equally-spaced condition described in Subsection IV-D. The empty set $\mathcal{L} = \{\}$ indicates that the preprocessing by WPE was not employed. To see how

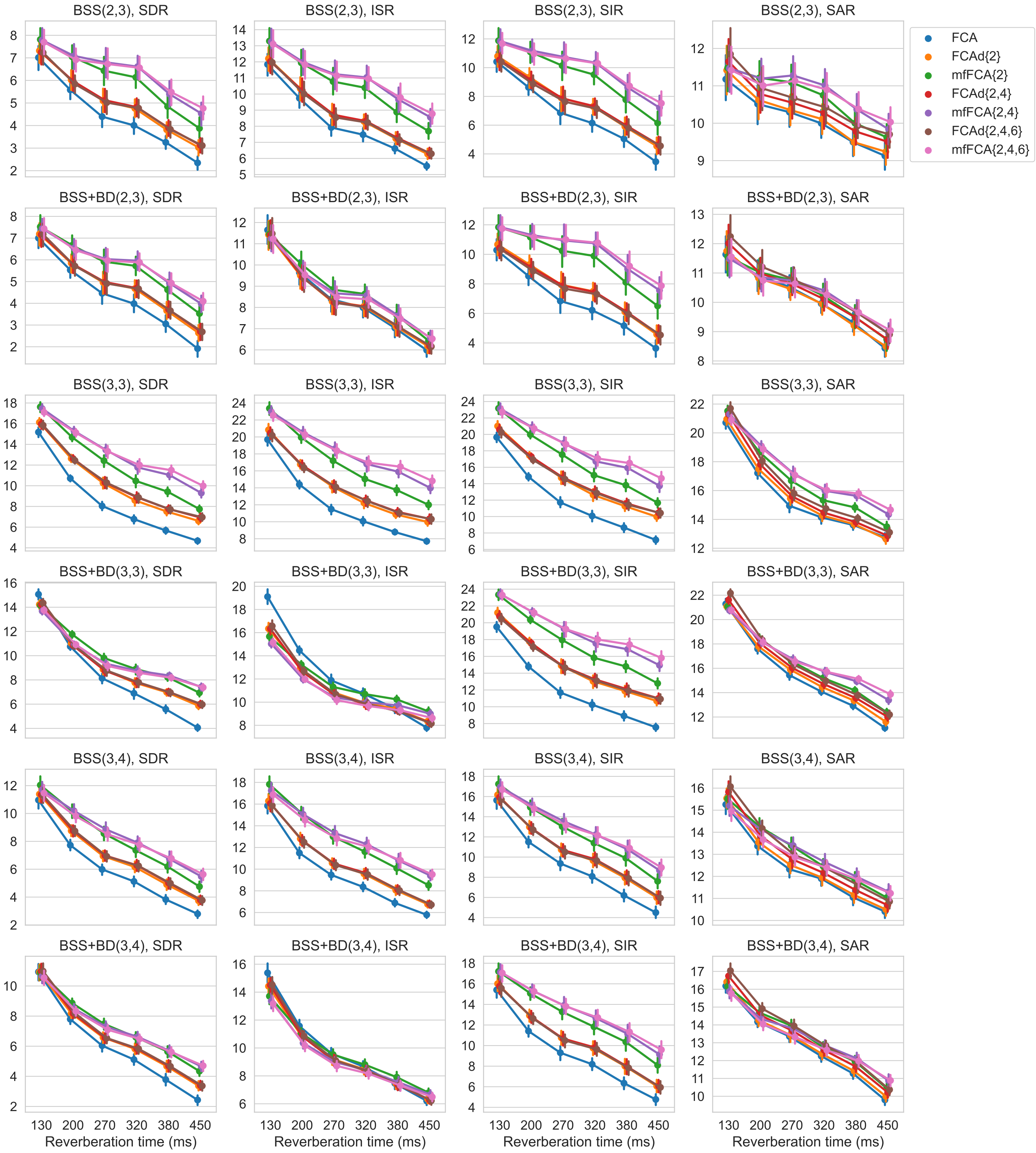


Fig. 5. BSS (odd rows) and BSS+BD (even rows) performances measured in SDR, ISR, SIR, and SAR (shown in each column). Every two rows from the top correspond to (M, N) combinations (2,3), (3,3), and (3,4). The horizontal axis in each plot corresponds to reverberation times. The markers ● represent the sample means (the point estimates) and the vertical lines represent their 95% confidence interval.

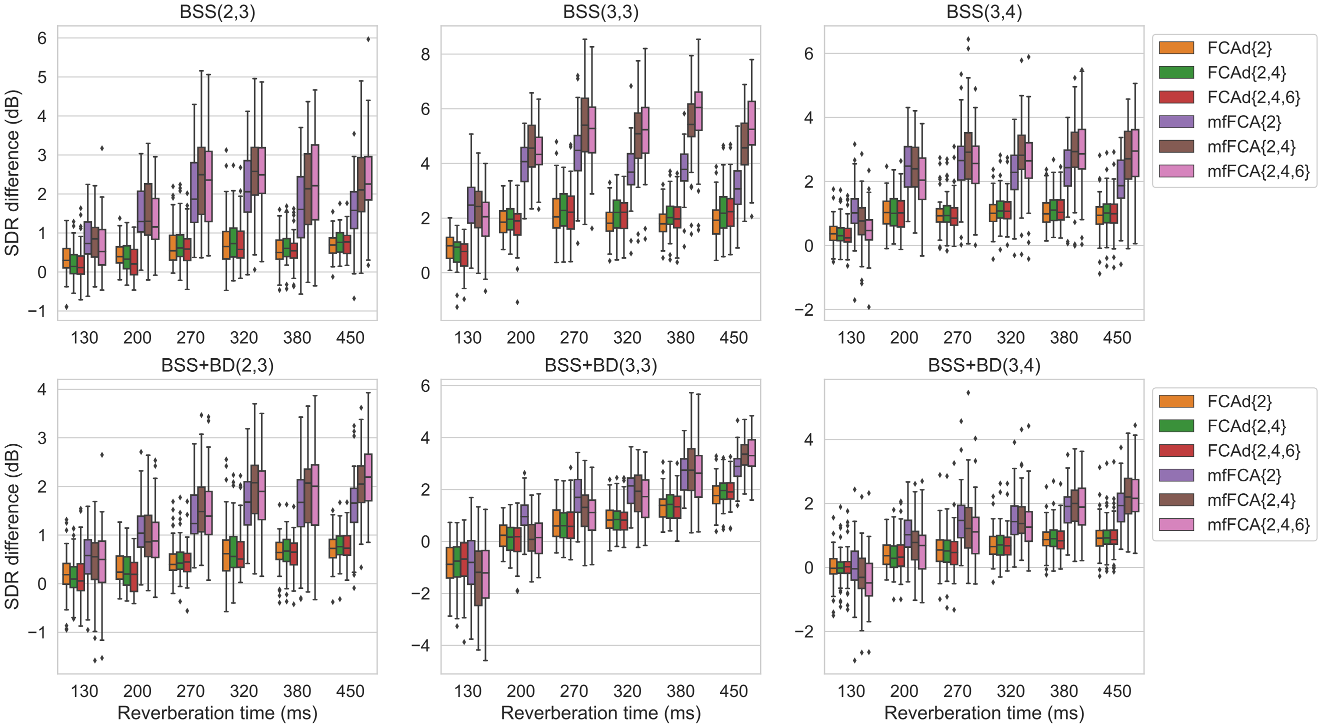
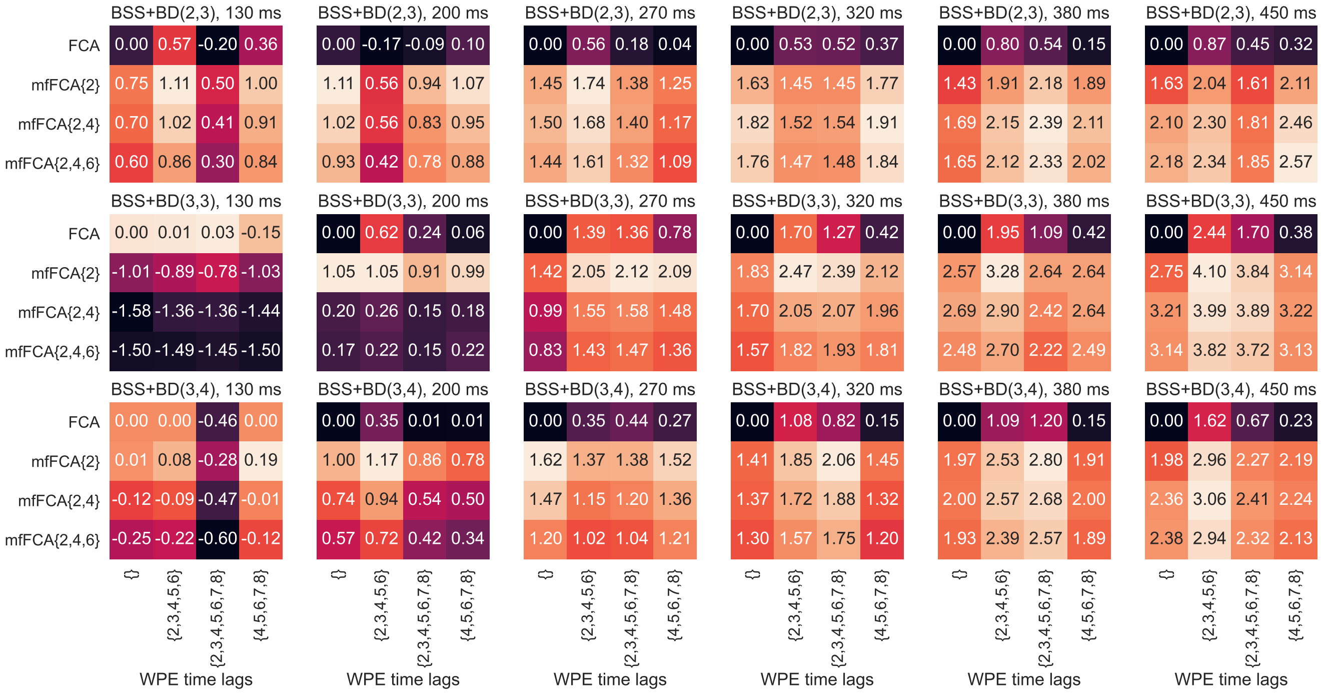


Fig. 6. (Continued from Fig. 5) SDR differences from the baseline. Each box plot shows the distribution of 48 and 64 differences (16 combinations of N sources) for $N = 3$ and $N = 4$ cases, respectively.



BSS+BD(2,3), 130 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.57	-0.20	0.36
mfFCA{2}	0.75	1.11	0.50	1.00
mfFCA{2,4}	0.70	1.02	0.41	0.91
mfFCA{2,4,6}	0.60	0.86	0.30	0.84

BSS+BD(2,3), 200 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	-0.17	-0.09	0.10
mfFCA{2}	1.11	0.56	0.94	1.07
mfFCA{2,4}	1.02	0.56	0.83	0.95
mfFCA{2,4,6}	0.93	0.42	0.78	0.88

BSS+BD(2,3), 270 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.56	0.18	0.04
mfFCA{2}	1.45	1.74	1.38	1.25
mfFCA{2,4}	1.50	1.68	1.40	1.17
mfFCA{2,4,6}	1.44	1.61	1.32	1.09

BSS+BD(2,3), 320 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.53	0.52	0.37
mfFCA{2}	1.63	1.45	1.45	1.77
mfFCA{2,4}	1.82	1.52	1.54	1.91
mfFCA{2,4,6}	1.76	1.47	1.48	1.84

BSS+BD(2,3), 380 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.80	0.54	0.15
mfFCA{2}	1.43	1.91	2.18	1.89
mfFCA{2,4}	1.69	2.15	2.39	2.11
mfFCA{2,4,6}	1.65	2.12	2.33	2.02

BSS+BD(2,3), 450 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.87	0.45	0.32
mfFCA{2}	1.63	2.04	1.61	2.11
mfFCA{2,4}	2.10	2.30	1.81	2.46
mfFCA{2,4,6}	2.18	2.34	1.85	2.57

BSS+BD(3,3), 130 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.01	0.03	-0.15
mfFCA{2}	-1.01	-0.89	-0.78	-1.03
mfFCA{2,4}	-1.58	-1.36	-1.36	-1.44
mfFCA{2,4,6}	-1.50	-1.49	-1.45	-1.50

BSS+BD(3,3), 200 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.62	0.24	0.06
mfFCA{2}	1.05	1.05	0.91	0.99
mfFCA{2,4}	0.20	0.26	0.15	0.18
mfFCA{2,4,6}	0.17	0.22	0.15	0.22

BSS+BD(3,3), 270 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	1.39	1.36	0.78
mfFCA{2}	1.42	2.05	2.12	2.09
mfFCA{2,4}	0.99	1.55	1.58	1.48
mfFCA{2,4,6}	0.83	1.43	1.47	1.36

BSS+BD(3,3), 320 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	1.70	1.27	0.42
mfFCA{2}	1.83	2.47	2.39	2.12
mfFCA{2,4}	1.70	2.05	2.07	1.96
mfFCA{2,4,6}	1.57	1.82	1.93	1.81

BSS+BD(3,3), 380 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	1.95	1.09	0.42
mfFCA{2}	2.57	3.28	2.64	2.64
mfFCA{2,4}	2.69	2.90	2.42	2.64
mfFCA{2,4,6}	2.48	2.70	2.22	2.49

BSS+BD(3,3), 450 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	2.44	1.70	0.38
mfFCA{2}	2.75	4.10	3.84	3.14
mfFCA{2,4}	3.21	3.99	3.89	3.22
mfFCA{2,4,6}	3.14	3.82	3.72	3.13

BSS+BD(3,4), 130 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.00	-0.46	0.00
mfFCA{2}	0.01	0.08	-0.28	0.19
mfFCA{2,4}	-0.12	-0.09	-0.47	-0.01
mfFCA{2,4,6}	-0.25	-0.22	-0.60	-0.12

BSS+BD(3,4), 200 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.35	0.01	0.01
mfFCA{2}	1.00	1.17	0.86	0.78
mfFCA{2,4}	0.74	0.94	0.54	0.50
mfFCA{2,4,6}	0.57	0.72	0.42	0.34

BSS+BD(3,4), 270 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	0.35	0.44	0.27
mfFCA{2}	1.62	1.37	1.38	1.52
mfFCA{2,4}	1.47	1.15	1.20	1.36
mfFCA{2,4,6}	1.20	1.02	1.04	1.21

BSS+BD(3,4), 320 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	1.08	0.82	0.15
mfFCA{2}	1.41	1.85	2.06	1.45
mfFCA{2,4}	1.37	1.72	1.88	1.32
mfFCA{2,4,6}	1.30	1.57	1.75	1.20

BSS+BD(3,4), 380 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	1.09	1.20	0.15
mfFCA{2}	1.97	2.53	2.80	1.91
mfFCA{2,4}	2.00	2.57	2.68	2.00
mfFCA{2,4,6}	1.93	2.39	2.57	1.89

BSS+BD(3,4), 450 ms

	{}	{2,3,4,5,6}	{2,3,4,5,6,7,8}	{4,5,6,7,8}
FCA	0.00	1.62	0.67	0.23
mfFCA{2}	1.98	2.96	2.27	2.19
mfFCA{2,4}	2.36	3.06	2.41	2.24
mfFCA{2,4,6}	2.38	2.94	2.32	2.13

Fig. 7. BSS+BD performances when WPE was employed as preprocessing. The numbers in the colors represent the averaged SDR differences from the baseline, i.e., FCA without WPE. The three rows correspond to (M, N) combinations (2,3), (3,3), and (3,4). The columns correspond to reverberation times.

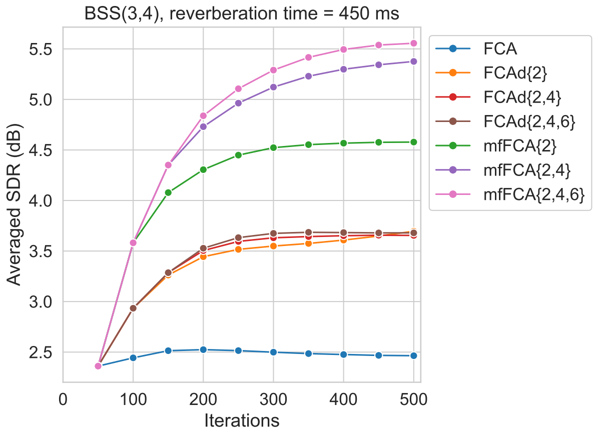


Fig. 8. Typical convergence behaviors for a task BSS(3,4) under 450 ms reverberation time.

TABLE I
COMPUTATIONAL TIME (IN SECONDS) OF THE FCA METHODS WITH 500 ITERATIONS FOR THE EM ALGORITHMS.

	FCA	mfFCA		
		$\mathcal{L} = \{2\}$	$\mathcal{L} = \{2,4\}$	$\mathcal{L} = \{2,4,6\}$
BSS(2,3)	87.2	90.9	110.6	119.1
BSS(3,3)	89.3	113.3	140.7	174.2
BSS(3,4)	104.6	130.5	173.0	200.0

TABLE II
THE DIRECTIONS OF THE DIAGONAL SHIFT OPERATORS AFFECTED THE PERFORMANCE MEASURED IN AVERAGED SDRS (IN DB). THE REVERBERATION TIME WAS 450 MS.

	BSS(3,4)			BSS+BD(3,4)		
time lags	both	$\nwarrow^i$	$\searrow^i$	both	$\nwarrow^i$	$\searrow^i$
$\mathcal{L} = \{2\}$	4.79	2.31	3.21	4.36	1.02	2.66
$\mathcal{L} = \{2,4\}$	5.53	0.41	3.13	4.68	0.29	2.65
$\mathcal{L} = \{2,4,6\}$	5.69	-0.95	3.07	4.73	0.17	2.63

BSS+BD performances were improved by extending FCA and by employing WPE, we set FCA without WPE as the baseline. The numbers in Fig. 7 represent the averaged SDR differences from the baseline.

From the results, we observe the followings. For situations with moderate and higher reverberations (from 270 ms to 450 ms), the introduction of WPE preprocessing generally improved the BSS+BD performance further with the FCA extension mfFCA. The best set of WPE time lags depended on the situation, i.e., (M, N) combination, reverberation time, and the set of time lags of mfFCA.

E. Convergence behavior

Figure 8 shows typical convergence behaviors as examples. The vertical axes show the average of $N = 4$ SDRs for a source combination. The horizontal axes show the iteration numbers of the EM algorithms for the FCA methods. The first 50 iterations were common to the baseline FCA. Then, FCAd and mfFCA inherited the FCA parameters, and augmented the time lag set from an empty set to the specified set $\mathcal{L}$ by adding one time lag at the beginning of every 50-iterations until the completion of $\mathcal{L}$.

We employed a hybrid approach for permutation alignment (Subsection IV-A), which was further related to parameter regularization (Subsection IV-B). In the first 50 iterations, we did not employ the second regularization of parameter sharing, and employed only the first regularization of flooring. After the first 50 iterations, we aligned the permutation ambiguities of FCA solutions by post-processing [50]. The alignment results were expected to be roughly correct globally, but there might be a few misaligned frequency bins. Then, starting from the 51-th iteration, we employed the second regularization of parameter sharing, which could be regarded as the second approach of permutation alignment. We expected the above-mentioned misaligned frequency bins to be aligned locally within each of the blocks $\mathcal{F}_b$, $, b = 1, \ldots, B$ by sharing the source power parameters s_{ntf} $, f \in \mathcal{F}_b$.

From Fig. 8, we observe that the effectiveness of employing suitable set of time lags are clear as mfFCA{2} saturated around 350 iterations due to the insufficient time lags for the 450 ms reverberation time.

Table I shows the computational times that took for FCA and mfFCA to run the EM algorithms with 500 iterations. All the methods were coded with Python using CuPy [57] and run on an AMD EPYC 7313 processor together with NVIDIA RTX A6000 as a GPU. Although we accelerated the algorithm computation by the GPU and CuPy similar to [58], the matrix inverse calculations in (13), (14), (42) and (43) still took long times. We also observed that the newly introduced $\nwarrow^i$ and $\searrow^i$ operators took long times in mfFCA as M and $L = |\mathcal{L}|$ increased. However, we consider the computation times were worth paying for the performance improvements.

F. On the directions of the diagonal shift operators

Table II shows that the directions of the diagonal shift operators, introduced for (32), affected the performance. For the sake of limited space, only the cases of BSS(3,4) and BSS+BD(3,4) with the 450 ms reverberation time are shown. However, also for the other $(M, N) = (2, 3), (3, 3)$ combinations and the reverberation times from 130 ms to 380 ms, we confirmed that the tendencies were almost the same. As discussed in Subsection III-E, our proposed mfFCA shifted the covariance matrices in both directions, and therefore performed clearly better than the other methods that only shifted in either direction. Especially the method only using $\nwarrow^i$ could not successfully separate the mixtures. This can be explained as this method ignored the $\searrow^i$ shifted components which were the dominant direct components as illustrated by the red boxes in Fig. 2.

VI. CONCLUSION

We have newly proposed multi-frame FCA mfFCA in which source components spanning multiple STFT time frames are modeled with covariance matrix (30) of larger dimensionality than the existing FCA models. Then, we have developed the corresponding probabilistic model and derived an EM algorithm to optimize the model parameters. The model and algorithm derivations are fully described in Section III and Appendices. Our experimental tasks were BSS and BSS+BD for determined and underdetermined cases, the latter of which was the first experimental confirmation of underdetermined separation and dereverberation in a fully blind manner. Experimental results show that the proposed mfFCA considerably improves the separation performance for reverberant BSS tasks. For BSS+BD tasks, preprocessing by WPE contributed to improve the performance of mfFCA further even in underdetermined cases. Therefore, future work includes the development of a technique that integrates WPE and mfFCA as WPE and FCAd are integrated [42]–[44]. The execution times were far from the signal lengths (6 seconds) even we have accelerated the algorithm computation by a GPU. In this sense, we would like to reduce the computational complexity as another future work. The joint diagonalization approach [15] would contribute to both the integration and the computational reduction.

Appendix A
The joint distribution (FCA)

In this appendix, we derive the specific form of the joint distribution $p(\mathbf{x}_t, \{\mathbf{c}_{nt}\}_{n=1}^N \mid \theta)$. According to the product rule of probability and the independence assumption (7), we have

$$p(\mathbf{x}_t, \{\mathbf{c}_{nt}\}_{n=1}^N \mid \theta) = p(\mathbf{x}_t \mid \{\mathbf{c}_{nt}\}_n, \theta) \prod\nolimits_n p(\mathbf{c}_{nt} \mid \theta) \, . \tag{56}$$

The log likelihood of (56) can be expressed as

$$\ln p(\mathbf{x}_t \mid \{\mathbf{c}_{nt}\}_n, \theta) + \sum\nolimits_n \ln p(\mathbf{c}_{nt} \mid \theta) \overset{\text{const}}{=} \\ -\beta^{-1}(\mathbf{x}_t - \sum_n \mathbf{c}_{nt})^*(\mathbf{x}_t - \sum_n \mathbf{c}_{nt}) - \sum_n \mathbf{c}_{nt}^* \mathsf{C}_{nt}^{-1} \mathbf{c}_{nt} \tag{57}$$

where $\overset{\text{const}}{=}$ denotes equality up to constants that do not include variables $\mathbf{x}_t$ and $\mathbf{c}_{nt}$. By simple mathematical manipulations to (57), we confirm that the joint distribution (56) is a zero-mean Gaussian distribution with the precision matrix

$$\beta^{-1} \begin{bmatrix} \mathsf{I} & -\mathsf{I} & \cdots & -\mathsf{I} \\ -\mathsf{I} & \mathsf{I} & \cdots & \mathsf{I} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathsf{I} & \mathsf{I} & \cdots & \mathsf{I} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \mathsf{C}_{1t}^{-1} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \mathsf{C}_{Nt}^{-1} \end{bmatrix} \tag{58}$$

whose inverse is the covariance matrix. To calculate the inverse of (58), we employ a well-known matrix identity

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{S}^{-1} \end{bmatrix} \tag{59}$$

with the Schur complement $\mathbf{S} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$. Letting $\mathbf{A} = \beta^{-1}\mathsf{I}$, $\mathbf{B} = \beta^{-1}\begin{bmatrix} -\mathsf{I} & \cdots & -\mathsf{I} \end{bmatrix}$ and so on, we have the Schur complement in this case

$$\mathbf{S} = \begin{bmatrix} \mathsf{C}_{1t}^{-1} & & 0 \\ & \ddots & \\ 0 & & \mathsf{C}_{Nt}^{-1} \end{bmatrix} , \tag{60}$$

which is a block diagonal matrix, and the inverse of (58) turns out to be (8).

Appendix B
$\mathcal{Q}$ function (FCA)

In this appendix, we derive the form of $\mathcal{Q}$ function for optimizing the FCA parameters. The log likelihood of the joint distribution $p(\mathbf{x}_t, \{\mathbf{c}_{nt}\}_n \mid \theta)$, which can be decomposed as (56), is expressed as

$$\ln p(\mathbf{x}_t \mid \{\mathbf{c}_{nt}\}_n, \theta) + \sum\nolimits_n \ln p(\mathbf{c}_{nt} \mid \theta) \overset{\text{const}}{=} \\ -\sum\nolimits_n \ln \det \mathsf{C}_{nt} - \sum\nolimits_n \mathbf{c}_{nt}^* \mathsf{C}_{nt}^{-1} \mathbf{c}_{nt} \tag{61}$$

where $\overset{\text{const}}{=}$ denotes equality up to constants that do not include parameters in θ. Then, the $\mathcal{Q}$ function is defined by taking the expectation using the posterior distribution (12) with the set θ' of previous parameters

$$\mathcal{Q}(\theta, \theta') = \sum_{t=1}^{T} \mathbb{E}_{\{p(\mathbf{c}_{nt} | \mathbf{x}_t, \theta')\}_{n=1}^N} \ln p(\mathbf{x}_t, \{\mathbf{c}_{nt}\}_n \mid \theta) \overset{\text{const}}{=} \\ -\sum_t \sum_n \left\{ \ln \det (\mathsf{s}_{nt} \mathsf{A}_n) + \operatorname{tr} \left[(\mathsf{s}_{nt} \mathsf{A}_n)^{-1} \widetilde{\mathsf{C}}_{nt} \right] \right\} \tag{62}$$

with

$$\widetilde{\mathsf{C}}_{nt} = \mathbb{E}_{p(\mathbf{c}_{nt} | \mathbf{x}_t, \theta')} \left[\mathbf{c}_{nt} \mathbf{c}_{nt}^* \right] = \boldsymbol{\mu}_{nt} \boldsymbol{\mu}_{nt}^* + \Sigma_{nt} \, . \tag{63}$$

The partial derivative of $\mathcal{Q}(\theta, \theta')$ with respect to s_{nt} and A_n are given as

$$\frac{\partial \mathcal{Q}(\theta, \theta')}{\partial \mathsf{s}_{nt}} = -M \mathsf{s}_{nt}^{-1} + \mathsf{s}_{nt}^{-2} \operatorname{tr} \left[\mathsf{A}_n^{-1} \widetilde{\mathsf{C}}_{nt} \right] , \tag{64}$$

$$\frac{\partial \mathcal{Q}(\theta, \theta')}{\partial \mathsf{A}_n} = -T \mathsf{A}_n^{-1} + \sum_{t=1}^{T} \mathsf{s}_{nt}^{-1} \mathsf{A}_n^{-1} \widetilde{\mathsf{C}}_{nt} \mathsf{A}_n^{-1} , \tag{65}$$

respectively. Setting these zero gives the updates (14) and (15).

Appendix C
The joint distribution (mfFCA)

In this appendix, we derive the specific form of the joint distribution $p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N \mid \theta)$ by a similar manner with Appendix A. According to the product rule of probability and the independence assumption (36), we have

$$p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N \mid \theta) = p(\bar{\mathbf{x}}_t \mid \{\bar{\mathbf{c}}_{nt}\}_n, \theta) \prod\nolimits_n p(\bar{\mathbf{c}}_{nt} \mid \theta) \, . \tag{66}$$

The log likelihood of (66) can be expressed as

$$\ln p(\bar{\mathbf{x}}_t \mid \{\bar{\mathbf{c}}_{nt}\}_n, \theta) + \sum\nolimits_n \ln p(\bar{\mathbf{c}}_{nt} \mid \theta) \overset{\text{const}}{=} \\ -(\bar{\mathbf{x}}_t - \sum_n \bar{\mathbf{c}}_{nt})^* \bar{\mathsf{E}}_t^{-1} (\bar{\mathbf{x}}_t - \sum_n \bar{\mathbf{c}}_{nt}) - \sum_n \bar{\mathbf{c}}_{nt}^* \bar{\mathsf{C}}_{nt}^{-1} \bar{\mathbf{c}}_{nt} \tag{67}$$

with $\bar{\mathsf{E}}_t = \bar{\mathsf{D}}_t + \beta \bar{\mathsf{I}}$, where $\overset{\text{const}}{=}$ denotes equality up to constants that do not include variables $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{c}}_{nt}$. The definitions of $\bar{\mathsf{C}}_{nt}$ and $\bar{\mathsf{D}}_t$ are given in (30) and (32), respectively. By simple mathematical manipulations to (67), we confirm that the joint distribution (66) is a zero-mean Gaussian distribution with the precision matrix

$$\begin{bmatrix} \bar{\mathsf{E}}_t^{-1} & -\bar{\mathsf{E}}_t^{-1} & \cdots & -\bar{\mathsf{E}}_t^{-1} \\ -\bar{\mathsf{E}}_t^{-1} & \bar{\mathsf{E}}_t^{-1} + \bar{\mathsf{C}}_{1t}^{-1} & \cdots & \bar{\mathsf{E}}_t^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\bar{\mathsf{E}}_t^{-1} & \bar{\mathsf{E}}_t^{-1} & \cdots & \bar{\mathsf{E}}_t^{-1} + \bar{\mathsf{C}}_{Nt}^{-1} \end{bmatrix} . \tag{68}$$

To calculate the inverse of (68), we employ (59) again. Letting $\mathbf{A} = \bar{\mathsf{E}}_t^{-1}$, $\mathbf{B} = \left[\begin{array}{ccc} -\bar{\mathsf{E}}_t^{-1} & \cdots & -\bar{\mathsf{E}}_t^{-1} \end{array} \right]$ and so on, we have the Schur complement in this case

$$\mathbf{S} = \begin{bmatrix} \bar{\mathsf{C}}_{1t}^{-1} & & \mathsf{0} \\ & \ddots & \\ \mathsf{0} & & \bar{\mathsf{C}}_{Nt}^{-1} \end{bmatrix} \tag{69}$$

and the inverse of (68) turns out to be (37).

Appendix D
$\mathcal{Q}$ function (mfFCA)

In this appendix, we derive the form of $\mathcal{Q}$ function for optimizing the mfFCA parameters by a similar manner with Appendix B. The log likelihood of the joint distribution $p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_n \mid \theta)$, which can be decomposed as (66), is expressed as

$$\begin{aligned} \ln p(\bar{\mathbf{x}}_t \mid \{\bar{\mathbf{c}}_{nt}\}_n, \theta) + \textstyle\sum_n \ln p(\bar{\mathbf{c}}_{nt} \mid \theta) \overset{\text{const}}{=} \\ -\ln\det\left(\bar{\mathsf{D}}_t + \beta\bar{\mathsf{I}}\right) - \bar{\mathbf{r}}_t^* \left(\bar{\mathsf{D}}_t + \beta\bar{\mathsf{I}}\right)^{-1} \bar{\mathbf{r}}_t \\ -\textstyle\sum_n \ln\det\bar{\mathsf{C}}_{nt} - \textstyle\sum_n \bar{\mathbf{c}}_{nt}^* \bar{\mathsf{C}}_{nt}^{-1} \bar{\mathbf{c}}_{nt} \end{aligned} \tag{70}$$

with $\bar{\mathbf{r}}_t = \bar{\mathbf{x}}_t - \sum_n \bar{\mathbf{c}}_n$. Here $\overset{\text{const}}{=}$ denotes equality up to constants that do not include parameters in θ. Then, the $\mathcal{Q}$ function is defined by taking the expectation using the posterior distribution (41) with the set θ' of previous parameters

$$\begin{aligned} \mathcal{Q}(\theta, \theta') = \sum_{t=1}^{T-l_L} \mathbb{E}_{\{p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta')\}_{n=1}^N} \ln p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_n \mid \theta) \overset{\text{const}}{=} \\ -\sum_t \left\{ \ln\det\left(\bar{\mathsf{D}}_t + \beta\bar{\mathsf{I}}\right) + \operatorname{tr}\left[\left(\bar{\mathsf{D}}_t + \beta\bar{\mathsf{I}}\right)^{-1} \tilde{\tilde{\bar{\mathsf{R}}}}_t \right] \right\} \\ -\sum_t \sum_n \left\{ \ln\det\left(\mathsf{s}_{nt}\bar{\mathsf{A}}_n\right) + \operatorname{tr}\left[\left(\mathsf{s}_{nt}\bar{\mathsf{A}}_n\right)^{-1} \tilde{\tilde{\bar{\mathsf{C}}}}_{nt} \right] \right\} \end{aligned} \tag{71}$$

with

$$\tilde{\tilde{\bar{\mathsf{R}}}}_t = \mathbb{E}_{p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta')} \left[\bar{\mathbf{r}}_t \bar{\mathbf{r}}_t^* \right], \tag{72}$$

$$\tilde{\tilde{\bar{\mathsf{C}}}}_{nt} = \mathbb{E}_{p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta')} \left[\bar{\mathbf{c}}_{nt} \bar{\mathbf{c}}_{nt}^* \right] = \bar{\boldsymbol{\mu}}_{nt} \bar{\boldsymbol{\mu}}_{nt}^* + \bar{\mathsf{\Sigma}}_{nt}. \tag{73}$$

This $\mathcal{Q}$ function of mfFCA is complicated compared to that (62) of FCA, and the exact maximization by θ is not simple as (62). To deal with this difficulty, we make an approximation that the parameter values used in $\bar{\mathsf{D}}_t$ are fixed at the previous parameter values in θ'. Then, the partial derivative of $\mathcal{Q}(\theta, \theta')$ with respect to s_{nt} and $\bar{\mathsf{A}}_n$ are given as

$$\frac{\partial \mathcal{Q}(\theta, \theta')}{\partial \mathsf{s}_{nt}} = -M(L+1)\mathsf{s}_{nt}^{-1} + \mathsf{s}_{nt}^{-2} \operatorname{tr}\left[\bar{\mathsf{A}}_n^{-1} \tilde{\tilde{\bar{\mathsf{C}}}}_{nt} \right], \tag{74}$$

$$\frac{\partial \mathcal{Q}(\theta, \theta')}{\partial \bar{\mathsf{A}}_n} = -T\bar{\mathsf{A}}_n^{-1} + \sum_{t=1}^{T} \mathsf{s}_{nt}^{-1} \bar{\mathsf{A}}_n^{-1} \tilde{\tilde{\bar{\mathsf{C}}}}_{nt} \bar{\mathsf{A}}_n^{-1}, \tag{75}$$

respectively. Setting these zero gives the updates (43) and (44).

References

[1] H. Sawada, R. Ikeshita, K. Kinoshita, and T. Nakatani, “Multi-frame full-rank spatial covariance analysis for underdetermined BSS in reverberant environments,” in *Proc. ICASSP*, 2022, pp. 496–500.

[2] ——, “Evaluating the dereverberation-separation capability of multi-frame full-rank spatial covariance analysis,” in *Proc. International Congress on Acoustics*, 2022.

[3] C. Jutten and J. Herault, “Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture,” *Signal processing*, vol. 24, no. 1, pp. 1–10, 1991.

[4] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. John Wiley & Sons, 2000.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[6] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. John Wiley & Sons, 2002.

[7] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Springer, 2007.

[8] M. Pedersen, J. Larsen, U. Kjems, and L. Parra, “Convolutive blind source separation methods,” in *Springer handbook of speech processing*. Springer, 2008, pp. 1065–1094.

[9] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[10] P. Comon, “Independent component analysis, a new concept?” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[11] A. Bell and T. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[12] N. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[13] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, “Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation,” in *Proc. ISSPA 2010*, May 2010, pp. 1–4.

[14] A. Ozerov, E. Vincent, and F. Bimbot, “A general flexible framework for the handling of prior information in audio source separation,” *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1118–1133, 2011.

[15] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, “A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel wiener filter,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1950–1965, 2021.

[16] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time fourier transform domain,” *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[17] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[18] L. Schobben and W. Sommen, “A frequency domain blind signal separation method based on decorrelation,” *IEEE Trans. Signal Processing*, vol. 50, no. 8, pp. 1855–1865, Aug. 2002.

[19] A. Hiroe, “Solution of permutation problem in frequency domain ICA using multivariate probability density functions,” in *Proc. ICA 2006 (LNCS 3889)*. Springer, Mar. 2006, pp. 601–608.

[20] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

[21] N. Ono, “Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions,” in *Proc. APSIPA ASC*, Dec. 2012, pp. 1–4.

[22] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, May 2013.

[23] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[24] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, “A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF,” *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[25] R. Talmon, I. Cohen, and S. Gannot, “Relative transfer function identification using convolutive transfer function approximation,” *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 546–555, 2009.

[26] X. Li, L. Girin, R. Horaud, and S. Gannot, “Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[27] X. Li, L. Girin, S. Gannot, and R. Horaud, “Multichannel speech separation and enhancement using the convolutive transfer function,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 645–659, 2019.

[28] F. Feng and M. Kowalski, “Underdetermined reverberant blind source separation: Sparse approaches for multiplicative and convolutive narrowband approximation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 442–456, 2018.

[29] T. Wang, F. Yang, and J. Yang, “Convolutive transfer function-based multichannel nonnegative matrix factorization for overdetermined blind source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 802–815, 2022.

[30] P. Naylor and N. Gaubitch, *Speech dereverberation*. Springer, 2010.

[31] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech dereverberation based on variance-normalized delayed linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[32] T. Yoshioka and T. Nakatani, “Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[33] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, “Multichannel linear prediction-based speech dereverberation with sparse priors,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.

[34] M. Togami, “Multi-channel time-varying covariance matrix model for late reverberation reduction,” *arXiv preprint arXiv:1910.08710*, 2019.

[35] H. Buchner, R. Aichner, and W. Kellermann, “TRINICON: A versatile framework for multichannel blind signal processing,” in *Proc. ICASSP*, vol. 3. IEEE, 2004, pp. iii–889.

[36] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno, “Blind separation and dereverberation of speech mixtures by joint optimization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69–84, 2011.

[37] R. Ikeshita and T. Nakatani, “Independent vector extraction for fast joint blind source separation and dereverberation,” *IEEE Signal Processing Letters*, vol. 28, pp. 972–976, 2021.

[38] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, “Relaxed disjointness based clustering for joint blind source separation and dereverberation,” in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 268–272.

[39] M. Togami and Y. Kawaguchi, “Noise robust speech dereverberation with kalman smoother,” in *Proc. ICASSP*. IEEE, 2013, pp. 7447–7451.

[40] N. Ito, S. Araki, and T. Nakatani, “Probabilistic integration of diffuse noise suppression and dereverberation,” in *Proc. ICASSP*. IEEE, 2014, pp. 5167–5171.

[41] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, “Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1369–1380, 2013.

[42] M. Togami, “Multi-channel speech source separation and dereverberation with sequential integration of determined and underdetermined models,” in *Proc. ICASSP*, 2020, pp. 231–235.

[43] K. Sekiguchi, Y. Bando, A. Nugraha, M. Fontaine, and K. Yoshii, “Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation,” in *Proc. ICASSP*, 2021, pp. 511–515.

[44] K. Sekiguchi, Y. Bando, A. Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara, “Autoregressive moving average jointly-diagonalizable spatial covariance analysis for joint source separation and dereverberation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2368–2382, 2022.

[45] T. Adali, P. J. Schreier, and L. L. Scharf, “Complex-valued signal processing: The proper way to deal with impropriety,” *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5101–5125, 2011.

[46] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[47] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[48] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, D. Raj, S. Watanabe, Z. Chen, and J. Hershey, “Sequential multi-frame neural beamforming for speech separation and enhancement,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 905–911.

[49] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[50] H. Sawada, S. Araki, and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[51] H. Sawada, R. Ikeshita, and T. Nakatani, “Experimental analysis of EM and MU algorithms for optimizing full-rank spatial covariance model,” in *Proc. EUSIPCO 2020*, Jan. 2021, pp. 885–889.

[52] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Duong, “The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges,” *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, Aug. 2012.

[53] “Museval,” https://github.com/sigsep/sigsep-mus-eval.

[54] S. Araki, S. Makino, H. Sawada, and R. Mukai, “Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask,” in *Proc. ICASSP 2005*, vol. 3, Mar. 2005, pp. iii–81.

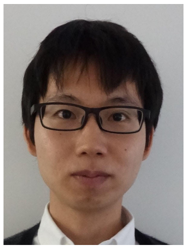
[55] H. Sawada, R. Ikeshita, N. Ito, and T. Nakatani, “Computational acceleration and smart initialization of full-rank spatial covariance analysis,” in *Proc. EUSIPCO*, 2019, pp. 1–5.

[56] [Online]. Available: http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/mffca/

[57] R. Okuta, Y. Unno, D. Nishino, S. Hido, and C. Loomis, “CuPy: A NumPy-compatible library for NVIDIA GPU calculations,” in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017. [Online]. Available: http://learningsys.org/nips17/assets/papers/paper_16.pdf

[58] D. Raj, D. Povey, and S. Khudanpur, “GPU-accelerated guided source separation for meeting transcription,” *arXiv preprint arXiv:2212.05271*, 2022.

Hiroshi Sawada (M’02-SM’04-F’18) received the B.E., M.E. and Ph.D. degrees in information science from Kyoto University, in 1991, 1993 and 2001, respectively. He joined NTT Corporation in 1993. He is now a senior distinguished researcher at the NTT Communication Science Laboratories. His research interests include statistical signal processing, audio source separation, array signal processing, latent variable models, second-order optimization, and computer architecture. He served as an Associate Editor for the IEEE Trans. Audio, Speech and Language Processing from 2006 to 2009, and as an Associate Editor of the IEEE Open Journal of Signal Processing from 2019 to 2022. He was a Member and an Associate Member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE SP Society from 2006 to 2018. He received the Best Paper Award of the IEEE Circuit and System Society in 2000, and the Best Paper Award of the IEEE Signal Processing Society in 2014. He is a 2022 Distinguished Lecturer of the IEEE Signal Processing Society. He is an IEEE Fellow, an IEICE Fellow, and a member of the ASJ.



Rintaro Ikeshita (Member, IEEE) received the B.E. and M.S. degrees from the University of Tokyo, Tokyo, Japan, in 2013 and 2015, respectively. He is currently a Researcher with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. From 2015 to 2018, he was a Researcher with Research & Development Group, Hitachi, Ltd., Tokyo, Japan.

Keisuke Kinoshita is a research scientist at Google. Before joining Google, he was a distinguished research scientist at NTT Communication Science Laboratories (from 2003 to 2022), where he did most of the work on the project described in this manuscript. He received the M. Eng. degree and Ph.D degree from Sophia University in Tokyo in 2003 and 2010, respectively. In this research career, he has been engaged in fundamental research on various types of speech, audio, and music signal processing, including 1ch/multi-channel speech enhancement (blind dereverberation, source separation, noise reduction), speaker diarization, robust speech recognition, and distributed microphone array processing, and developed several innovative commercial software. He is an author or a co-author of more than 20 journal papers, 5 book chapters, more than 100 papers presented at peer-reviewed international conferences, and an inventor or a co-inventor of more than 20 Japanese patents and 5 international patents. He serves as an associate editor of IEEE Transactions on Audio, Speech and Language Processings (TASLP) since 2021, and a member of IEEE Audio and Acoustic Signal Processing Technical Committee (AASP-TC) since 2019, and served as the Chief Coordinator of the REVERB challenge (2014), an editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (from 2013 to 2017), a guest editor of EURASIP journal on advances in signal processing (2015). He was honored to receive the 2006 IEICE Paper Award, the 2010 ASJ Outstanding Technical Development Prize, the 2011 ASJ Awaya Prize, the 2012 Japan Audio Society Award, 2015 IEEE-ASRU Best Paper Award Honorable Mention, and 2017 Maejima Hisoka Award. He is a member of IEEE, ASJ ,and IEICE.

Tomohiro Nakatani (Fellow, IEEE) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 1989, 1991, and 2002, respectively. He is a Senior Distinguished Researcher with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. Since joining NTT Corporation, as a Researcher in 1991, he has been investigating audio signal processing technologies for intelligent human-machine interfaces, including dereverberation, denoising, source separation, and robust ASR. He was a Visiting Scholar with the Georgia Institute of Technology, Atlanta, GA, USA, for a year in 2005, and Visiting Assistant Professor with the Department of Media Science, Nagoya University, Nagoya, Japan, from 2008 to 2017. From 2008 to 2010 he was an Associate Editor for the IEEE TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. From 2009 to 2014, he was a Member of the IEEE Signal Processing Society (SPS), Audio and Acoustics Technical Committee, and from 2016 to 2021, Member of the IEEE SPS, Speech and Language Processing Technical Committee. From 2011 to 2012, he was the Chair of the IEEE Kansai Section Technical Program Committee and from 2019 to 2020, Chair of the IEEE SPS Kansai Chapter. He was the Technical Program Co-Chair of the IEEE WASPAA-2007, Co-Chair of the 2014 REVERB Challenge Workshop, and General Co-Chair of the IEEE ASRU-2017. He is a Fellow of IEICE and Member of ASJ. He was the recipient of the 2005 IEICE Best Paper Award, 2009 ASJ Technical Development Award, 2012 Japan Audio Society Award, 2015 IEEE ASRU Best Paper Award Honorable Mention, 2017 Maejima Hisoka Award, and 2018 IWAENC Best Paper Award.