ACCURATE DELAYED SOURCE MODEL
FOR MULTI-FRAME FULL-RANK SPATIAL COVARIANCE ANALYSIS

Shinya Furunaga[1], *Hiroshi Sawada*[2], *Rintaro Ikeshita*[2], *Tomohiro Nakatani*[2], *Shoji Makino*[1]

[1]Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
[2]NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

Multi-frame Full-rank Spatial Covariance Analysis (mfFCA) is a technique for a blind source separation method and can be applied to reverberant underdetermined conditions where the sources outnumber the microphones and the reverberation time is long. This model, however, does not express all direct and delayed source components in multi-frame observation vector. This paper proposes a new model that takes into account accurately the direct and delayed source components, by introducing delay-wise spatial covariance matrices. We have then derived new expectation-maximization and multiplicative update algorithms for the proposed model. Experimental results show that the proposed method performed better than the conventional mfFCA for the task to separate three sources with two microphones.

Index Terms— Blind source separation (BSS), full-rank spatial covariance analysis (FCA), reverberation, expectation-maximization (EM) algorithm, multiplicative update (MU) algorithm

1. INTRODUCTION

Blind Source Separation (BSS) aims to separate N sources from M observed signals without prior information of source signals and the mixing system [1–5]. Independent Component Analysis (ICA) [6–8] is a basic BSS method and many extensions of ICA have been studied [9]. However, they are basically applicable in determined ($N = M$) and overdetermined ($N < M$) conditions. Full-rank Spatial Covariance Analysis (FCA) [10–14], on the other hand, can also be applied to underdetermined conditions ($N > M$).

In a real room environment, it is necessary to perform BSS taking reverberation into account. Applying blind dereverberation methods such as Weighted Prediction Error (WPE) [15] as a preprocessing of BSS is helpful to reduce the adverse effects of reverberation. However, WPE becomes much less effective when it is applied to underdetermined conditions.

Researchers have proposed to incorporate delayed source components in the FCA model [16–23]. Especially, multi-frame FCA (mfFCA) [22, 23] models the source components spanning multiple time frames and takes correlations into account between different time frames. Figure 1 describes the difference between the original FCA, conventional mfFCA, and the proposed modification of mfFCA. In conventional mfFCA (**mfFCAo**), the multi-frame observation vector consists of one part of direct and delayed source components (red arrows in Fig. 1). The other source components (purple arrows) are accounted for by a covariance matrix of the multi-frame observation vector (see Subsection 2.2). Therefore, mfFCA has introduced only an approximate optimization algorithm such as Expectation-Maximization (EM) algorithm.

In this paper, we propose to modify mfFCA on how to model source signals over multi-frames and derive optimization algorithms. We improve the model of the observation signals by correctly including the all direct and delayed source components in the multi-frame observation signals. Owing to that, the parameters can be obtained for each time lag. We then derived the update rules using the EM algorithm and the Multiplicative Update (MU) algorithm, which cannot be derived by the conventional mfFCA. Hereinafter, we call the proposed method as mfFCA with all direct and delayed source components (**mfFCAa**).

2. CONVENTIONAL METHODS

2.1. Full-rank Spatial Covariance Analysis (FCA)

We briefly review FCA [10]. Suppose that N sources are mixed and observed by M microphones. In this paper, we will omit the frequency bin index to simplify notations. We first apply a Short-Time Fourier Transform (STFT) to the time-domain observed signals. The observed signal $\mathbf{x}_t$ at time frame $t \in \{1, \dots, T\}$ is expressed as the superimposition of N source components $\mathbf{c}_{nt}$:

$$\mathbf{x}_t = \textstyle\sum_{n=1}^{N} \mathbf{c}_{nt} \in \mathbb{C}^M. \tag{1}$$

The source components and the observation vector are modeled as

$$p(\mathbf{c}_{nt} \mid \theta) = \mathcal{N}(\mathbf{c}_{nt} \mid \mathbf{0}, \mathsf{C}_{nt}),\ \ \mathsf{C}_{nt} = \mathsf{s}_{nt}\mathsf{A}_n, \tag{2}$$

$$p(\mathbf{x}_t \mid \theta) = \mathcal{N}(\mathbf{x}_t \mid \mathbf{0}, \mathsf{X}_t),\ \ \mathsf{X}_t = \textstyle\sum_{n=1}^{N} \mathsf{C}_{nt} + \beta\mathsf{I} \tag{3}$$

with the set of parameters $\theta = \{\{\mathsf{s}_{nt}\}_{t=1}^{T}, \mathsf{A}_n\}_{n=1}^{N}$, where $\mathcal{N}$ indicates a zero-mean multivariate complex Gaussian distribution, $\mathsf{s}_{nt} \in \mathbb{R}_{\geq 0}$ represents the time-variant power of source n at time frame t, and $\mathsf{A}_n \in \mathbb{C}^{M \times M}$ is the time-invariant spatial covariance matrix of source n. Also, β is the noise power and I is the identity matrix.

In the FCA model, the optimization of parameters θ has been proposed based on an EM [10] and MU [24] algorithms.

2.2. Multi-frame FCA with one part of direct and delayed source components (mfFCAo)

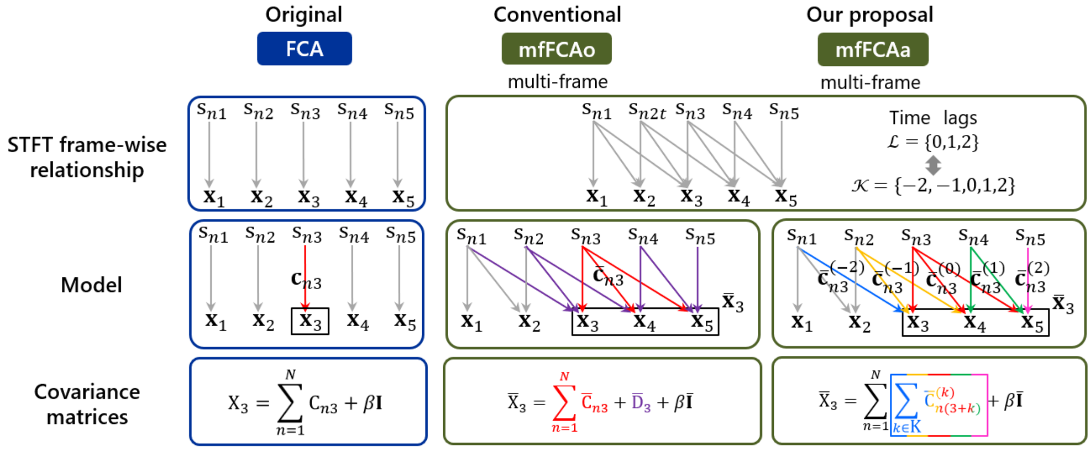
To better model reverberation than FCA does, mfFCAo [22, 23] treats a multi-frame vector for the observed signal

$$\bar{\mathbf{x}}_t = [\mathbf{x}_t^{\mathsf{T}}, \mathbf{x}_{(t+l_1)}^{\mathsf{T}}, \dots, \mathbf{x}_{(t+l_L)}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M(L+1)}, \tag{4}$$

where $\mathcal{L} = \{0, l_1, \dots, l_L\}$ is the set of time lags and L is the number of time lags considered in mfFCAo. Analogously, a multi-frame source component is defined as

$$\bar{\mathbf{c}}_{nt} = [\mathbf{c}_{nt}^{(0)\mathsf{T}}, \mathbf{c}_{n(t+l_1)}^{(l_1)\mathsf{T}}, \dots, \mathbf{c}_{n(t+l_L)}^{(l_L)\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M(L+1)}. \tag{5}$$

See Fig. 1 for an illustration of $\bar{\mathbf{c}}_{nt}$ with $\mathcal{L} = \{0, 1, 2\}$.


Original
FCA
Conventional
mfFCAo
multi-frame
Our proposal
mfFCAa
multi-frame
STFT frame-wise relationship
Model
Covariance matrices
Time lags
$\mathcal{L} = \{0,1,2\}$
$\mathcal{K} = \{-2,-1,0,1,2\}$
s_{n1} s_{n2} s_{n3} s_{n4} s_{n5}
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$
$\mathbf{c}_{n3}$
$\bar{\mathbf{c}}_{n3}$
$\bar{\mathbf{x}}_3$
$\bar{\mathbf{c}}_{n3}^{(-2)}$ $\bar{\mathbf{c}}_{n3}^{(-1)}$ $\bar{\mathbf{c}}_{n3}^{(0)}$ $\bar{\mathbf{c}}_{n3}^{(1)}$ $\bar{\mathbf{c}}_{n3}^{(2)}$
$\mathsf{X}_3 = \sum_{n=1}^{N} \mathsf{C}_{n3} + \beta\mathbf{I}$
$\bar{\mathsf{X}}_3 = \sum_{n=1}^{N} \bar{\mathsf{C}}_{n3} + \bar{\mathsf{D}}_3 + \beta\bar{\mathbf{I}}$
$\bar{\mathsf{X}}_3 = \sum_{n=1}^{N} \sum_{k\in\mathcal{K}} \bar{\mathsf{C}}_{n(3+k)}^{(k)} + \beta\bar{\mathbf{I}}$

s_{nt}: time-variant power, $\mathbf{x}_t$: observation vector, $\mathbf{c}_{nt}$: source component vector, $\mathsf{X}_t, \mathsf{C}_{nt}$: covariance matrix

Fig. 1: Illustrations of FCA, mfFCA (denoted as mfFCAo), and the proposed mfFCAa.

In mfFCAo, it is assumed that $\bar{\mathbf{c}}_{nt}$ follows the zero mean Gaussian distribution with covariance matrix $\bar{\mathsf{C}}_{nt}$:

$$p(\bar{\mathbf{c}}_{nt} \mid \theta) = \mathcal{N}(\bar{\mathbf{c}}_{nt} \mid \mathbf{0}, \bar{\mathsf{C}}_{nt}), \tag{6}$$

where $\bar{\mathsf{C}}_{nt} \in \mathbb{C}^{M(L+1)\times M(L+1)}$ is modeled as $\bar{\mathsf{C}}_{nt} = \mathsf{s}_{nt}\bar{\mathsf{A}}_n$ with $\mathsf{s}_{nt} \in \mathbb{R}_{\geq 0}$ and

$$\bar{\mathsf{A}}_n = \begin{bmatrix} \mathsf{A}_n^{(0)} & \dots & \mathbf{A}_n^{(0,l_L)} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_n^{(l_L,0)} & \dots & \mathsf{A}_n^{(l_L)} \end{bmatrix}. \tag{7}$$

It is also assumed that the multi-frame observed signal $\bar{\mathbf{x}}_t$ follows

$$p(\bar{\mathbf{x}}_t \mid \theta) = \mathcal{N}(\bar{\mathbf{x}}_t \mid \mathbf{0}, \bar{\mathsf{X}}_t). \tag{8}$$

To establish the relation between the covariance matrices $\bar{\mathsf{X}}_t$ and $\bar{\mathsf{C}}_{nt}$, let us look at an example in Fig. 1. When $\mathcal{L} = \{0, 1, 2\}$, $\bar{\mathbf{x}}_3$ is affected by the red arrows $\bar{\mathbf{c}}_{n3}$ and also affected by direct and delayed components corresponding to the purple arrows. This relationship is modeled in mfFCAo as

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

$$\bar{\mathsf{D}}_t = \textstyle\sum_{n=1}^{N} \sum_{i=1}^{L} \left(\nwarrow^{i} \bar{\mathsf{C}}_{n(t-l_i)} + \searrow^{i} \bar{\mathsf{C}}_{n(t+l_i)}\right), \tag{10}$$

where β is the noise power, $\bar{\mathsf{I}}$ is the identity matrix, and $\nwarrow^{i}$and $\searrow^{i}$are shift operators that diagonally shift the submatrices of size $M \times M$ (see [23] for the definition of this shift operator). The set of parameters of mfFCAo is $\theta = \{\{\mathsf{s}_{nt}\}_{t=1}^{T}, \bar{\mathsf{A}}_n\}_{n=1}^{N}$. In the mfFCAo paper, the authors proposed an EM algorithm for estimating the parameters. The update rule can be found in [22, 23].

2.3. Drawback of conventional mfFCAo

The drawback of conventional mfFCAo is that $\bar{\mathbf{x}}_t$ does not contain all direct and delayed source components. For the sake of simplicity, let's consider the situation in Fig. 1. $\bar{\mathbf{x}}_3$ contains the direct and delayed source components from $\mathsf{s}_{n1}, \mathsf{s}_{n2}, \mathsf{s}_{n3}, \mathsf{s}_{n4}$, and s_{n5}. However, in the conventional mfFCAo model, the direct and delayed source components of purple arrows in Fig. 1 are modelled by $\bar{\mathsf{D}}_t$ in (9, 10), not in $\bar{\mathbf{x}}_3$. According to (10), $\bar{\mathsf{D}}_t$ is modeled by components with a shifted $\bar{\mathsf{C}}_{nt}$ and contaminated with components of $\bar{\mathsf{C}}_{nt}$. Therefore, mfFCAo only uses an approximate optimization algorithms such as the EM algorithm.

3. PROPOSED METHOD

We propose to modify mfFCAo by considering all source component vectors affecting the multi-frame observation vector $\bar{\mathbf{x}}_t$ and modeling them with an appropriate Gaussian distribution. This modification allows us to develop an EM and MU algorithms for mfFCAa, unlike mfFCAo. We expect the proposed mfFCAa to improve separation performance and to allow faster convergence of parameter optimization.

3.1. Model

We introduce a set of source component vectors, denoted as $\bar{\mathbf{c}}_{nt}^{(k)}$ ($k \in \mathcal{K}$), such that it holds that

$$\bar{\mathbf{x}}_t = \textstyle\sum_{n=1}^{N} \sum_{k\in\mathcal{K}} \bar{\mathbf{c}}_{nt}^{(k)} \in \mathbb{C}^{M(K+1)}, \tag{11}$$

where $\mathcal{K} = \{-k_K, \dots, -k_1, 0, k_1, \dots, k_K\}$ is a set of time lags and K is the number of time lags considered in mfFCAa. To explain the definition of $\bar{\mathbf{c}}_{nt}^{(k)}$, let us look at an example in Fig. 1 which describes the case where $K = 2$, $\mathcal{K} = \{-2, -1, 0, 1, 2\}$, and $t = 3$. In this case, $\bar{\mathbf{x}}_t \in \mathbb{C}^{M(K+1)}$ and $\bar{\mathbf{c}}_{nt}^{(k)}$ ($k \in \mathcal{K}$) in (11) are defined as

$$\bar{\mathbf{x}}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_{(t+1)} \\ \mathbf{x}_{(t+2)} \end{bmatrix}, \bar{\mathbf{c}}_{nt}^{(-2)} = \begin{bmatrix} \mathbf{c}_{nt}^{(2)} \\ 0 \\ 0 \end{bmatrix}, \bar{\mathbf{c}}_{nt}^{(-1)} = \begin{bmatrix} \mathbf{c}_{nt}^{(1)} \\ \mathbf{c}_{n(t+1)}^{(2)} \\ 0 \end{bmatrix},$$

$$\bar{\mathbf{c}}_{nt}^{(0)} = \begin{bmatrix} \mathbf{c}_{nt}^{(0)} \\ \mathbf{c}_{n(t+1)}^{(1)} \\ \mathbf{c}_{n(t+2)}^{(2)} \end{bmatrix}, \bar{\mathbf{c}}_{nt}^{(1)} = \begin{bmatrix} 0 \\ \mathbf{c}_{n(t+1)}^{(0)} \\ \mathbf{c}_{n(t+2)}^{(1)} \end{bmatrix}, \bar{\mathbf{c}}_{nt}^{(2)} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{c}_{n(t+2)}^{(0)} \end{bmatrix}. \tag{12}$$

Each of $\bar{\mathbf{c}}_{nt}^{(-2)}$, $\bar{\mathbf{c}}_{nt}^{(-1)}$, $\bar{\mathbf{c}}_{nt}^{(0)}$, $\bar{\mathbf{c}}_{nt}^{(1)}$, and $\bar{\mathbf{c}}_{nt}^{(2)}$ corresponds to a colored arrow in Fig. 1. Here, for example, the elements of $\bar{\mathbf{c}}_{n3}^{(1)}$ (green arrows in Fig. 1) consist of the source components from s_{n4} into $\mathbf{x}_4$ or $\mathbf{x}_5$, namely $\mathbf{c}_{n4}^{(0)}$ or $\mathbf{c}_{n5}^{(1)}$. We then assume that $\bar{\mathbf{c}}_{nt}^{(k)}$ follows a zero-mean Gaussian distribution

$$p(\bar{\mathbf{c}}_{nt}^{(k)} \mid \theta) = \mathcal{N}(\bar{\mathbf{c}}_{nt}^{(k)} \mid \mathbf{0}, \bar{\mathsf{C}}_{nt}^{(k)}) \tag{13}$$

with covariance matrix $\bar{\mathsf{C}}_{nt}^{(k)} = \mathsf{s}_{n(t+k)}\bar{\mathsf{A}}_n^{(k)}$. Here $\mathsf{s}_{nt} \in \mathbb{R}_{\geq 0}$ is the time-variant power of source n at time frame t, and $\bar{\mathsf{A}}_n^{(k)} \in$

$\mathbb{C}^{M(K+1)\times M(K+1)}$ is a covariance matrix which encodes the time-invariant spatial property from source n to M microphones and for all considered time lags. Along with the definitions of the source component vectors in (12), the covariance matrices $\bar{\mathsf{A}}_n^{(k)}$ ($k \in \mathcal{K}$) should have the following structure:

$$\bar{\mathsf{A}}_n^{(-2)} = \begin{bmatrix} \mathsf{A}_n^{(2),-2} & 0 & 0 \\ 0 & \epsilon\mathsf{I} & 0 \\ 0 & 0 & \epsilon\mathsf{I} \end{bmatrix}, \bar{\mathsf{A}}_n^{(-1)} = \begin{bmatrix} \mathsf{A}_n^{(1),-1} & \mathbf{A}_n^{(1,2),-1} & 0 \\ \mathbf{A}_n^{(2,1),-1} & \mathsf{A}_n^{(2),-1} & 0 \\ 0 & 0 & \epsilon\mathsf{I} \end{bmatrix},$$

$$\bar{\mathsf{A}}_n^{(0)} = \begin{bmatrix} \mathsf{A}_n^{(0),0} & \mathbf{A}_n^{(0,1),0} & \mathbf{A}_n^{(0,2),0} \\ \mathbf{A}_n^{(1,0),0} & \mathsf{A}_n^{(1),0} & \mathbf{A}_n^{(1,2),0} \\ \mathbf{A}_n^{(2,0),0} & \mathbf{A}_n^{(2,1),0} & \mathsf{A}_n^{(2),0} \end{bmatrix}, \tag{14}$$

$$\bar{\mathsf{A}}_n^{(1)} = \begin{bmatrix} \epsilon\mathsf{I} & 0 & 0 \\ 0 & \mathsf{A}_n^{(0),1} & \mathbf{A}_n^{(0,1),1} \\ 0 & \mathbf{A}_n^{(1,0),1} & \mathsf{A}_n^{(1),1} \end{bmatrix}, \bar{\mathsf{A}}_n^{(2)} = \begin{bmatrix} \epsilon\mathsf{I} & 0 & 0 \\ 0 & \epsilon\mathsf{I} & 0 \\ 0 & 0 & \mathsf{A}_n^{(0),2} \end{bmatrix},$$

where $\mathsf{A}_n^{(k),*}$ and $\mathbf{A}_n^{(k,k'),*}$ are block submatrices that should be equal for arbitrary $*$, e.g., $\mathsf{A}_n^{(0),0} = \mathsf{A}_n^{(0),1} = \mathsf{A}_n^{(0),2}$, $\mathbf{A}_n^{(0,1),0} = \mathbf{A}_n^{(0,1),1}$. The constant ϵ is extremely low value, e.g., as $\epsilon = 10^{-6}$. The conventional mfFCAo optimize only $\bar{\mathsf{A}}_n^{(0)}$. On the other hand, our proposal mfFCAa can optimize $\bar{\mathsf{A}}_n^{(k)}$ for each time lag k in $\mathcal{K}$.

The parameters of mfFCAa to be optimized are

$$\theta = \left\{ \{\mathsf{s}_{nt}\}_{t=1}^{T}, \{\bar{\mathsf{A}}_n^{(k)}\}_{k\in\mathcal{K}} \right\}_{n=1}^{N}. \tag{15}$$

For model tractability, we assume that the multi-frame source components $\bar{\mathbf{c}}_{nt}^{(k)}$ are mutually independent:

$$p(\{\bar{\mathbf{c}}_{nt}^{(k)}\}_{n,t,k} \mid \theta) = \prod_{n=1}^{N} \prod_{t=1}^{T-k_K} \prod_{k\in\mathcal{K}} p(\bar{\mathbf{c}}_{nt}^{(k)} \mid \theta). \tag{16}$$

Then with the additive model (11), $\bar{\mathbf{x}}_t$ follows

$$p(\bar{\mathbf{x}}_t \mid \theta) = \mathcal{N}(\bar{\mathbf{x}}_t \mid \mathbf{0}, \bar{\mathsf{X}}_t) \tag{17}$$

with the covariance matrix

$$\bar{\mathsf{X}}_t = \textstyle\sum_{n=1}^{N} \sum_{k\in\mathcal{K}} \bar{\mathsf{C}}_{nt}^{(k)} + \beta\bar{\mathsf{I}}, \tag{18}$$

where β is the noise power, The parameters θ can be estimated by maximizing the log-likehood of the multi-frame observation vector:

$$\ln p(\{\bar{\mathbf{x}}_t\}_{t=1}^{T-k_K} \mid \theta) = \textstyle\sum_{t=1}^{T-k_K} \ln p(\bar{\mathbf{x}}_t \mid \theta). \tag{19}$$

3.2. EM algorithm

We derive an EM algorithm to maximize the log-likelihood (19) for estimating the parameters θ in (15). In the E-step, in a similar manner to the EM algorithm for FCA, we calculate the conditional distribution of the source component vector $\bar{\mathbf{c}}_{nt}^{(k)}$ as

$$p(\{\bar{\mathbf{c}}_{nt}^{(k)}\}_{n,k\in\mathcal{K}} \mid \bar{\mathbf{x}}_t, \theta) = \mathcal{N}(\bar{\mathbf{c}}_{nt}^{(k)} \mid \bar{\boldsymbol{\mu}}_{nt}^{(k)}, \bar{\boldsymbol{\Sigma}}_{nt}^{(k)}) \tag{20}$$

with

$$\bar{\boldsymbol{\mu}}_{nt}^{(k)} = \bar{\mathsf{C}}_{nt}^{(k)} \bar{\mathsf{X}}_t^{-1} \bar{\mathbf{x}}_t\,, \ \ \bar{\boldsymbol{\Sigma}}_{nt}^{(k)} = \bar{\mathsf{C}}_{nt}^{(k)} - \bar{\mathsf{C}}_{nt}^{(k)} \bar{\mathsf{X}}_t^{-1} \bar{\mathsf{C}}_{nt}^{(k)}. \tag{21}$$

The part $\bar{\mathsf{C}}_{nt}^{(k)}\bar{\mathsf{X}}_t^{-1}$ in the mean $\bar{\boldsymbol{\mu}}_{nt}^{(k)}$ calculation in (21) is called multi-frame multichannel Wiener filter in [25]. In the M-step, we update the parameters as

$$\mathsf{s}_{nt} \leftarrow \frac{1}{|\mathcal{K}|M(K+1)} \sum_{k\in\mathcal{K}} \left\{ \mathrm{tr}\left[(\bar{\mathsf{A}}_n^{(k)})^{-1} \tilde{\bar{\mathsf{C}}}_{nt}^{(k)} \right] \right\}, \tag{22}$$

$$\bar{\mathsf{A}}_n^{(k)} \leftarrow \frac{1}{T-k_K} \sum_{t=1}^{T-k_K} (\mathsf{s}_{n(t+k)})^{-1} \tilde{\bar{\mathsf{C}}}_{nt}^{(k)}, \tag{23}$$

where tr calculates the trace of a matrix, $|\mathcal{K}|$ means the number of elements in set $\mathcal{K}$, and

$$\tilde{\bar{\mathsf{C}}}_{nt}^{(k)} = \bar{\boldsymbol{\mu}}_{nt}^{(k)} (\bar{\boldsymbol{\mu}}_{nt}^{(k)})^* + \bar{\boldsymbol{\Sigma}}_{nt}^{(k)}. \tag{24}$$

After updating $\bar{\mathsf{A}}_n^{(k)}$ with (23), we post-process $\bar{\mathsf{A}}_n^{(k)}$ to satisfy the structure of (14). The post-processing includes averaging the should-be-equal block submatrices, e.g., $\mathsf{A}_n^{(0),*} \leftarrow (\mathsf{A}_n^{(0),0} + \mathsf{A}_n^{(0),1} + \mathsf{A}_n^{(0),2})/3$ and replacing the corresponding submatrices with $\mathsf{0}$ or $\epsilon\mathsf{I}$.

3.3. MU algorithm

We drive an MU algorithm to minimize the negative log-likelihood for the parameters θ in (15). In a similar manner to the MU algorithm for FCA [14,24], a surrogate function for the negative log-likelihood is obtained as:

$$\begin{aligned} &D^{+}(\mathsf{s}_{nt}, \bar{\mathsf{A}}_n^{(k)}, \boldsymbol{R}_{n(t+k)}^{(k)}, \boldsymbol{U}_t) \\ &= \sum_{t=1}^{T-k_K} \Big[\sum_{n=1}^{N} \sum_{k\in\mathcal{K}} \Big(\mathrm{tr}\{ \boldsymbol{R}_{n(t+k)}^{(k)} \bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^{\mathsf{H}} (\bar{\mathsf{C}}_{nt}^{(k)})^{-1} \boldsymbol{R}_{n(t+k)}^{(k)}{}^{\mathsf{H}} \} \Big) \\ &\qquad + \mathrm{tr}(\bar{\mathsf{X}}_t \boldsymbol{U}_t{}^{-1}) \Big], \end{aligned} \tag{25}$$

where $\boldsymbol{R}_{n(t+k)}^{(k)}$ and $\boldsymbol{U}_t$ are auxiliary variables given by

$$\boldsymbol{R}_{n(t+k)}^{(k)} = \mathsf{s}_{n(t+k)} \bar{\mathsf{A}}_n^{(k)} \bar{\mathsf{X}}_t^{-1}, \qquad \boldsymbol{U}_t = \bar{\mathsf{X}}_t. \tag{26}$$

Then, we optimize the parameters θ based on the minimization of the surrogate function (25). The derived update formulas are

$$\mathsf{s}_{nt} \leftarrow \mathsf{s}_{nt} \sqrt{ \frac{ \sum_{k\in\mathcal{K}} \mathrm{tr}\left\{ \bar{\mathsf{X}}_{(t-k)}^{-1} \bar{\mathbf{x}}_{(t-k)} \bar{\mathbf{x}}_{(t-k)}^{\mathsf{H}} \bar{\mathsf{X}}_{(t-k)}^{-1} \bar{\mathsf{A}}_n^{(k)} \right\} }{ \sum_{k\in\mathcal{K}} \mathrm{tr}\left\{ \bar{\mathsf{A}}_n^{(k)} \bar{\mathsf{X}}_{(t-k)}^{-1} \right\} } }, \tag{27}$$

$$\begin{aligned} \bar{\mathsf{A}}_n^{(k)} &\leftarrow \left[\sum_{t=1}^{T-k_K} \mathsf{s}_{n(t+k)} \bar{\mathsf{X}}_t^{-1} \right]^{-1} \\ &\# \left[\bar{\mathsf{A}}_n^{(k)} \sum_{t=1}^{T-k_K} \left\{ \mathsf{s}_{n(t+k)} \bar{\mathsf{X}}_t^{-1} \bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^{\mathsf{H}} \bar{\mathsf{X}}_t^{-1} \right\} \bar{\mathsf{A}}_n^{(k)} \right]. \end{aligned} \tag{28}$$

Here, $\#$ denotes the geometric mean [26, 27] of two positive semidefinite matrices and is defined as

$$\mathbf{A}\#\mathbf{B} = \mathbf{A}^{\frac{1}{2}} \left(\mathbf{A}^{-\frac{1}{2}} \mathbf{B} \mathbf{A}^{-\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}}. \tag{29}$$

As in the EM algorithm, after updating $\bar{\mathsf{A}}_n^{(k)}$ with (28), we post-process $\bar{\mathsf{A}}_n^{(k)}$ to meet the structure of (14).

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We conducted experiments to evaluate the separation performance of the conventional mfFCAo and the proposed mfFCAa with EM or MU algorithm. In this experiment, we tested the case where $N = 3$ speech sources were mixed and observed with $M = 2$ omni-directional microphones. We measured impulse responses from the sources to the microphones under the room conditions shown in Fig. 2. The room reverberation time (RT) was set to 270 and 450 ms.

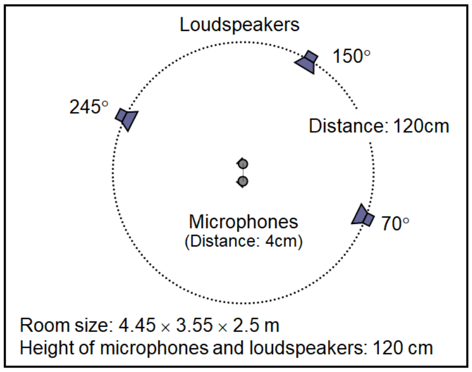

Loudspeakers
150°
245°
Distance: 120cm
Microphones
(Distance: 4cm)
70°
Room size: 4.45 × 3.55 × 2.5 m
Height of microphones and loudspeakers: 120 cm

Fig. 2: Experimental setup

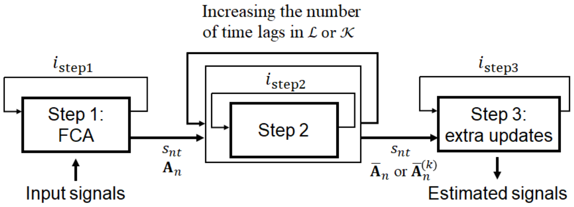


Fig. 3: Optimization scheduling with three steps. i is number of iterations in each step. In the experiments, we set them to $i_{\text{step1}} = 40$, $i_{\text{step2}} = 40$, and $i_{\text{step3}} = 240$.

The source images were generated by convolving 6 seconds English speech source signals with the impulse responses. We generated 8 mixtures by adding these source images. The sampling frequency was 8 kHz. The STFT window size and shift were 1024 and 256 samples (128 and 32 ms), respectively. Separation performance was evaluated using Signal-to-Distortion Ratios (SDR) [28].

We specified the sets of time lags, $\mathcal{L}$ and $\mathcal{K}$, according to the RT. When the RT was 270 ms, they were set to $\mathcal{L} = \{0, 2, 4\}$ and $\mathcal{K} = \{-4, -2, 0, 2, 4\}$. When the RT was 450 ms, they were set to $\mathcal{L} = \{0, 2, 4, 6\}$ and $\mathcal{K} = \{-6, -4, -2, 0, 2, 4, 6\}$.

To effectively optimize both mfFCAo and mfFCAa, we introduced an optimization scheduling technique to gradually increase the number of the time lags during the iterative optimization, as shown in Fig. 3. To begin with, we optimized the parameters, s_{nt} and A_n, using the original FCA (Subsection 2.1). We then switched to mfFCAo or mfFCAa and optimized the parameters by gradually increasing the number of time lags in $\mathcal{L}$ or $\mathcal{K}$ one by one from the initial set $\mathcal{L} = \{0, l_1\}$ in the case of mfFCAo and $\mathcal{K} = \{-k_1, 0, k_1\}$ in the case of mfFCAa until the set reached the specified set. After that, we further updated the parameters in an extra step. We used mfFCAo with EM or mfFCAa with EM/MU in the second and third steps. We tested four combinations: EM+EM, EM+MU, mfFCAo+EM, and mfFCAo+MU, where "A+B" means that method "A" was used in Step 2 and method "B" was used in Step 3. Here EM/MU denotes mfFCAa with EM/MU, respectively.

4.2. Results

Figure 4 shows the convergence behavior for a certain mixture when changing the combinations of mfFCAo and mfFCAa with EM/MU in Steps 2 and 3 in the optimization scheduling. The proposed methods show higher separation performance than mfFCAo. Additionally, EM+MU and mfFCAo+MU converged more quickly than other methods, demonstrating the advantage of the MU algorithm.

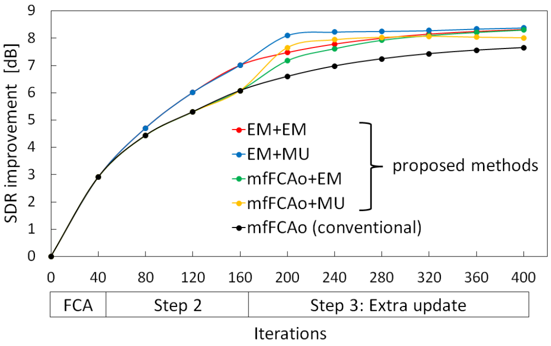


Fig. 4: Convergence behavior of conventional mfFCAo and proposed mfFCAa with EM/MU when RT was 450 ms. Optimization scheduling described in Fig. 3 was used for proposed methods.

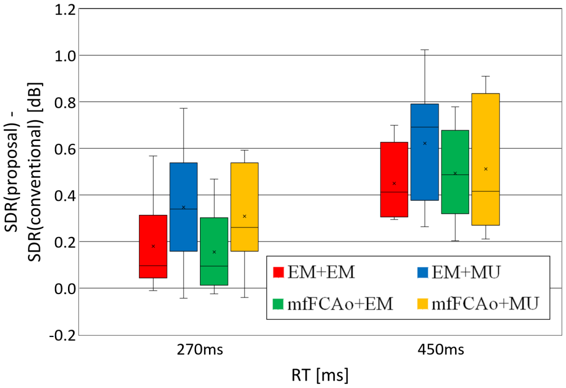


Fig. 5: Scatter plot for comparing SDR [dB] between conventional mfFCAo and proposed four methods: EM+EM, EM+MU, mfFCAo+EM, and mfFCAo+MU.

Figure 5 compares the SDR performance between the conventional mfFCAo and the four methods using the proposed mfFCAa with EM/MU. The horizontal axis denotes the SDR of conventional mfFCAo, and the vertical axis denotes the SDR obtained by the proposed methods. Each point corresponds to the separation result for one mixture signal. Overall, the points are above the diagonal dashed line, implying that the proposed mfFCAa improves mfFCAo. In particular, the average SDR obtained by EM+MU were 6.1 dB when RT was 450 ms and 7.7 dB when RT was 270 ms, which were 0.35 dB and 0.62 dB higher than the conventional mfFCAo, respectively.

5. CONCLUSION

We modified the conventional mfFCA to handle the direct and delayed source components more accurately. We also derived the EM and MU algorithms for the new formulation of mfFCA. The experimental results showed that by using an optimization scheduling, the proposed mfFCA can improve the separation performance of the original mfFCA.

6. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 23K28113.

7. REFERENCES

[1] 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.

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

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

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

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

[6] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

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

[8] 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.

[9] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. 24, pp. 695–709, 2005.

[10] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Underdetermined 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, Sept. 2010.

[11] S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for underdetermined reverberant audio source separation," in *Proc. ISSPA 2010*, May 2010, pp. 1–4.

[12] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2011.

[13] N. Ito and T. Nakatani, "Fastfca-AS: Joint diagonalization based acceleration of full-rank spatial covariance analysis for separating any number of sources," in *Proc. IWAENC*, 2018, pp. 151–158.

[14] 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 Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 1950–1965, 2021.

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

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

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

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

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

[20] K. Sekiguchi, Y. Bando, A.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.

[21] K. Sekiguchi, Y. Bando, A.A. Nugraha, M. Fontaine, and K. Yoshii, "Joint blind source separation and dereverberation based on ARMA-FastMNMF," in *Proc. the Acoustical Society of Japan*, Mar. 2021, pp. 129–132, (in Japanese).

[22] 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.

[23] H. Sawada, R. Ikeshita, K. Kinoshita, and T. Nakatani, "Multi-frame full-rank spatial covariance analysis for underdetermined blind source separation and dereverberation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 31, pp. 3589–3602, 2023.

[24] 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, 2013.

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

[26] T. Ando, C.-K. Li, and R. Mathias, "Geometric means," *Linear Algebra and its Applications*, vol. 385, no. 1, pp. 305–334, 2002.

[27] K. Yoshii, K. Kitamura, Y. Bando, and E. Nakamura, "Independent low-rank tensor analysis for audio source separation," in *Proc. EUSIPCO*, 2018, pp. 1671–1675.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.