MULTI-FRAME FULL-RANK SPATIAL COVARIANCE ANALYSIS FOR UNDERDETERMINED BSS IN REVERBERANT ENVIRONMENTS

Hiroshi Sawada, Rintaro Ikeshita, Keisuke Kinoshita, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

Full-rank spatial covariance analysis (FCA) is a blind source separation (BSS) method, and can be applied to underdetermined cases where the sources outnumber the microphones. This paper proposes a new extension of FCA, aiming to improve BSS performance for mixtures in which the length of reverberation exceeds the analysis frame. There has already been proposed a model that considers delayed source components as the exceeded parts. In contrast, our new extension models multiple time frames with multivariate Gaussian distributions of larger dimensionality than the existing FCA models. We derive an expectation-maximization algorithm to optimize the model parameters. Experiments to separate four speech sources with three microphones show that the proposed extension outperforms the existing models, and more specifically, the original FCA by around 2 dB measured with signal-to-distortion ratio.

***Index Terms*—** blind source separation (BSS), full-rank spatial covariance analysis (FCA), expectation-maximization (EM) algorithm, multivariate complex Gaussian distribution

1. INTRODUCTION

Blind source separation (BSS) has been studied for several decades [1–6]. Independent component analysis (ICA) [3, 7, 8] is a well-established BSS method in which the mixing system is assumed to be invertible. Full-rank spatial covariance analysis (FCA) [9–11], on the other hand, models a more flexible mixing system than ICA does. The most crucial difference is that FCA can be applied to **underdetermined** ($N > M$) cases where the number N of sources is larger than the number M of sensors (e.g., microphones in audio cases).

In a real room environment, signals are mixed in a **convolutive** manner with reverberations. Frequency-domain approach [12, 13], where we first apply a short-time Fourier transform (STFT) to the input time-domain signals, is effective for such convolutive mixtures. When the room reverberation time is so long that the analysis window of STFT cannot cover the dominant part of the reverberations, applying blind dereverberation methods such as weighted prediction error (WPE) [14] prior to BSS is helpful to reduce the adverse effects of reverberations. However, WPE is based on the invertible assumption as in the case of ICA, and basically applies to over-determined cases ($N < M$).

In this paper, we extend FCA to cope with long reverberations that span multiple STFT time frames even in an underdetermined case. There have been already proposed to consider delayed source components in the FCA model [15–17]. However, in their proposals, the methods are combined with WPE and have been confirmed only for over-determined cases. And more importantly, the dimensionality of multivariate Gaussian distributions, which are employed as probabilistic models, remains M as the original FCA. This fact might limit the source separation capability when considering delayed source components, because the correlations between different time frames, i.e., the block off-diagonal parts of (15), are ignored. In contrast, our extension of FCA models multiple time frames with multivariate Gaussian distributions of dimensionality $M \times (L+1)$, where L is the number of delayed components. We expect better separation capability by modeling the source propagation to all sensors with all considered time lags.

In the rest of the paper, Section 2 reviews the original FCA and its extension FCAd considering delayed source components. Section 3 proposes our extension multi-frame FCA (mfFCA) that models source components spanning multi frames employing Gaussian distributions of larger dimensionality. Section 4 experimentally shows that mfFCA considerably outperforms FCA and FCAd when the reverberation time was not very short. Section 5 concludes the paper with mentioning some future work.

2. FULL-RANK SPATIAL COVARIANCE ANALYSIS (FCA)

2.1. Model and objective function

Here we define the original FCA model. Suppose that $n = 1, \ldots, N$ sources are mixed and observed at $m = 1, \ldots, M$ sensors. Let the sensor observations at time frame $t \in \{1, \ldots, T\}$ and frequency bin $f \in \{1, \ldots, F\}$ be denoted by an M-dimensional complex vector $\mathbf{x}_{tf} \in \mathbb{C}^M$ with $\mathbf{x}_{tf} = [x_{1tf}, \ldots, x_{Mtf}]^\mathsf{T}$, and assumed to be the sum of N source components $\mathbf{c}_{ntf} \in \mathbb{C}^M$:

$$\mathbf{x}_{tf} = \textstyle\sum_{n=1}^{N} \mathbf{c}_{ntf} \,. \tag{1}$$

Each of the source components follows a zero-mean multivariate complex Gaussian distribution $p(\mathbf{c}_{ntf}) = \mathcal{N}(\mathbf{c}_{ntf} \mid \mathbf{0}, \mathsf{C}_{ntf})$ with covariance matrices

$$\mathsf{C}_{ntf} = \mathsf{s}_{ntf}\mathsf{A}_{nf} \,. \tag{2}$$

By these Sans-serif fonts, the lower and upper cases represent that the parameters are positive scalars and Hermitian positive definite matrices, respectively. A_{nf} encodes the time-invariant spatial property from source n to all M sensors. s_{ntf} represents the time-variant power of source n at time frame t and frequency bin f.

The objective function to maximize is the log-likelihood

$$\textstyle\sum_{f=1}^{F} \sum_{t=1}^{T} \log p(\mathbf{x}_{tf} \mid \theta) \tag{3}$$

with the set of parameters

$$\theta = \{\{\{\mathsf{s}_{ntf}\}_{t=1}^{T}, \mathsf{A}_{nf}\}_{n=1}^{N}\}_{f=1}^{F} \,. \tag{4}$$

According to the sum model (1) where the component vectors $\mathbf{c}_{ntf}$ are assumed to be independent of each other, the observation vector $\mathbf{x}_{tf}$ also follows a zero-mean Gaussian distribution $p(\mathbf{x}_{tf} \mid \theta) = \mathcal{N}(\mathbf{x}_{tf} \mid \mathbf{0}, \mathsf{X}_{tf})$ with a covariance matrix

$$\mathsf{X}_{tf} = \textstyle\sum_{n=1}^{N} \mathsf{C}_{ntf} \,. \tag{5}$$

2.2. EM algorithm

The objective function (3) can be locally maximized [9, 10] by the EM algorithm [18]. In the **E-step**, we calculate the conditional distributions $p(\mathbf{c}_{ntf} \mid \mathbf{x}_{tf}, \theta)$, which are Gaussian distributions with means and covariance matrices

$$\boldsymbol{\mu}_{ntf}^{(c)} = \mathsf{C}_{ntf}\mathsf{X}_{tf}^{-1}\mathbf{x}_{tf}\,, \quad \Sigma_{ntf}^{(c)} = \mathsf{C}_{ntf} - \mathsf{C}_{ntf}\mathsf{X}_{tf}^{-1}\mathsf{C}_{ntf}\,. \tag{6}$$

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

$$\mathsf{A}_{nf} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\mathsf{s}_{ntf}}\tilde{\mathsf{C}}_{ntf}\,, \tag{7}$$

$$\mathsf{s}_{ntf} \leftarrow \frac{1}{M}\mathrm{tr}\left(\mathsf{A}_{nf}^{-1}\tilde{\mathsf{C}}_{ntf}\right)\,, \tag{8}$$

where tr calculates the trace of a matrix, and

$$\tilde{\mathsf{C}}_{ntf} = \mathbb{E}\Big[\mathbf{c}_{ntf}\mathbf{c}_{ntf}^{\mathsf{H}} \mid \mathbf{x}_{tf}, \theta\Big] = \boldsymbol{\mu}_{ntf}^{(c)}\boldsymbol{\mu}_{ntf}^{(c)\mathsf{H}} + \Sigma_{ntf}^{(c)}\,. \tag{9}$$

2.3. Considering delayed source components

To accommodate room reverberations, the FCA mixture model (1) can be extended [15–17] as

$$\mathbf{x}_{tf} = \textstyle\sum_{n=1}^{N}\left(\mathbf{c}_{ntf} + \sum_{l\in\mathcal{L}}\mathbf{c}_{ntf}^{(l)}\right) \tag{10}$$

by considering the delayed source components $\mathbf{c}_{ntf}^{(l)}$ for the set $\mathcal{L} = \{l_1, \ldots, l_L\}$ of time lags. We denote this model as FCAd. The component $\mathbf{c}_{ntf}^{(l)}$ is coming from the source n emitted at time frame $t-l$ and observed at time frame t through the time lag l. The component is assumed to follow a zero-mean Gaussian distribution

$$p(\mathbf{c}_{ntf}^{(l)} \mid \mathbf{0}, \mathsf{C}_{ntf}^{(l)})\,, \quad \mathsf{C}_{ntf}^{(l)} = \mathsf{s}_{n(t-l)f}\mathsf{A}_{nf}^{(l)}\,, \tag{11}$$

where $\mathsf{A}_{nf}^{(l)}$ encodes the spatial property of source n affecting to all M sensors with time lag l. According to the extended sum model (10), the observation $\mathbf{x}_{tf}$ follows a zero-mean Gaussian distribution $p(\mathbf{x}_{tf} \mid \theta) = \mathcal{N}(\mathbf{x}_{tf} \mid \mathbf{0}, \mathsf{X}_{tf})$ with a covariance matrix

$$\mathsf{X}_{tf} = \textstyle\sum_{n=1}^{N}\left(\mathsf{C}_{ntf} + \sum_{l\in\mathcal{L}}\mathsf{C}_{ntf}^{(l)}\right)\,, \tag{12}$$

in which C_{ntf} and $\mathsf{C}_{ntf}^{(l)}$ are defined by (2) and (11), respectively.

2.4. Permutation alignment and parameter sharing

The solutions of FCA have permutation ambiguities among frequency bins as in the case of ICA. We need to align the permutations for blindly separating sources in a full-band manner. There are two major approaches. The first one is post-processing [19]. The second one is by sharing the source power parameters s_{ntf} among frequency bins [20–24]. In the experiments reported in Section 4, we employed the post-processing approach because of high separation performance for speech separation.

Nonetheless, sharing the source power parameters s_{ntf} among time frames or frequency bins improves the source separation performance by FCA as reported in [25]. Since the distinction between s_{ntf} and $\mathsf{s}_{n(t-l)f}$ is crucial in identifying time-lagged source components, we do not share s_{ntf} among time frames but among frequency bins. Specifically for the experiments (Section 4), we shared s_{ntf} among adjacent four frequency bins in a disjoint manner.

3. THE PROPOSED EXTENSION

To better model the source component $\mathbf{c}_{ntf}$ with its time-lagged version, we propose a new extension of FCA, multi-frame FCA (mfFCA), that considers the correlation between, e.g., $\mathbf{c}_{ntf}$ and $\mathbf{c}_{n(t+1)f}^{(1)}$, both of which originate from the same source s_{ntf} at the same time frame t but are observed at adjacent two frames. In the rest of this Section, we will omit frequency dependency f in the notations for the sake of simplicity.

3.1. Component vector spanning multiple time frames

Let us consider a long component vector

$$\bar{\mathbf{c}}_{nt} = [\mathbf{c}_{nt}^{\mathsf{T}}, \mathbf{c}_{n(t+l_1)}^{(l_1)\mathsf{T}}, \ldots, \mathbf{c}_{n(t+l_L)}^{(l_L)\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M(L+1)} \tag{13}$$

made by concatenating all the delayed versions originating from the same source s_{nt} with the same time instant. Then we assume that $\bar{\mathbf{c}}_{nt}$ follows a zero-mean Gaussian distribution

$$p(\bar{\mathbf{c}}_{nt}) = \mathcal{N}(\bar{\mathbf{c}}_{nt} \mid \mathbf{0}, \bar{\mathsf{C}}_{nt}) \tag{14}$$

with covariance matrix $\bar{\mathsf{C}}_{nt} = \mathsf{s}_{nt}\bar{\mathsf{A}}_n$, where

$$\bar{\mathsf{A}}_n = \begin{bmatrix} \mathsf{A}_n^{(0)} & \cdots & \mathbf{A}_n^{(0,l_L)} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_n^{(l_L,0)} & \cdots & \mathsf{A}_n^{(l_L)} \end{bmatrix} \tag{15}$$

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. The new set of parameters to be estimated is

$$\theta = \{\{\{\mathsf{s}_{nt}\}_{t=1}^{T}, \bar{\mathsf{A}}_n\}_{n=1}^{N}\}\,. \tag{16}$$

Letting $\mathsf{A}_n^{(0)} = \mathsf{A}_n$, we have already seen the block diagonal submatrices $\mathsf{A}_n^{(0)}, \ldots, \mathsf{A}_n^{(l_L)}$ in (2) and (11). The newly introduced block off-diagonal submatrices $\mathbf{A}_n^{(l,l')}$ satisfy $\left(\mathbf{A}_n^{(l,l')}\right)^{\mathsf{H}} = \mathbf{A}_n^{(l',l)}$. 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 models, e.g., (10) and (11), 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.

3.2. Probabilistic models

In this subsection, we develop new probabilistic models to reflect the introduction of $\bar{\mathbf{c}}_{nt}$. We first consider which parts of observations the component $\bar{\mathbf{c}}_{nt}$ affects. According to the time lags $\mathcal{L} = \{l_1, \ldots, l_L\}$ we care, they are the observations at time frames t and $t+l_1, \ldots, t+l_L$ (see Fig. 1). Thus, we define a long observation vector

$$\bar{\mathbf{x}}_t = [\mathbf{x}_t^{\mathsf{T}}, \mathbf{x}_{t+l_1}^{\mathsf{T}}, \ldots, \mathbf{x}_{t+l_L}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M(L+1)} \tag{17}$$

for which we assume the independence among different time t, i.e.,

$$p(\{\bar{\mathbf{x}}_t\}_{t=1}^{T-l_L} \mid \theta) = \textstyle\prod_{t=1}^{T-l_L} p(\bar{\mathbf{x}}_t \mid \theta)\,. \tag{18}$$

We then consider the joint distribution among $\bar{\mathbf{x}}_t$ and $\{\bar{\mathbf{c}}_{nt}\}_{n=1}^{N}$

$$p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_{n=1}^{N} \mid \theta) = p(\bar{\mathbf{x}}_t \mid \{\bar{\mathbf{c}}_{nt}\}_{n=1}^{N}, \theta)\textstyle\prod_{n=1}^{N} p(\bar{\mathbf{c}}_{nt} \mid \theta) \tag{19}$$

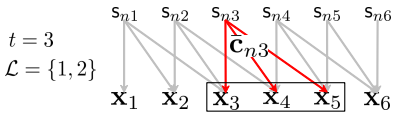

Fig. 1. Component vector spanning multiple time frames

$$\bar{\mathbf{c}}_{nt} = \begin{bmatrix} \mathbf{c}_{nt} \\ \mathbf{c}^{(l_1)}_{n(t+l_1)} \\ \vdots \\ \mathbf{c}^{(l_L)}_{n(t+l_L)} \end{bmatrix} \qquad \bar{\mathsf{C}}_{nt} = \begin{bmatrix} \mathsf{C}_{nt} & & & \\ & \mathsf{C}^{(l_1)}_{n(t+l_1)} & & \\ & & \ddots & \\ & & & \mathsf{C}^{(l_L)}_{n(t+l_L)} \end{bmatrix}$$

$$\sqcap_0 \bar{\mathbf{c}}_{nt} = \mathbf{c}_{nt} \qquad \sqcap_0 \bar{\mathsf{C}}_{nt} = \mathsf{C}_{nt}$$
$$\sqcap_i \bar{\mathbf{c}}_{nt} = \mathbf{c}^{(l_i)}_{n(t+l_i)} \qquad \sqcap_i \bar{\mathsf{C}}_{nt} = \mathsf{C}^{(l_i)}_{n(t+l_i)}$$
$$\sqcap_i \bar{\mathbf{c}}_{n(t-l_i)} = \mathbf{c}^{(l_i)}_{nt} \qquad \sqcap_i \bar{\mathsf{C}}_{n(t-l_i)} = \mathsf{C}^{(l_i)}_{nt}$$

Fig. 2. $\sqcap_i$ operator applied to a vector (left) and a matrix (right)

where we assume the independence of $\bar{\mathbf{c}}_{nt}$. We further assume the independence of subvectors in (17) when conditioned on $\{\bar{\mathbf{c}}_{nt}\}_{n=1}^N$

$$p(\bar{\mathbf{x}}_t \mid \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N, \theta) = \textstyle\prod_{i=0}^L p(\mathbf{x}_{t+l_i} \mid \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N, \theta)\,. \tag{20}$$

Note that the subvectors are still correlated as will be shown in (26). From here we let $l_0 = 0$, i.e., $\mathbf{x}_{t+l_0} = \mathbf{x}_t$ in the above, for notational convenience. Each subvector then follows a Gaussian distribution

$$p(\mathbf{x}_{t+l_i} \mid \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N, \theta) = \mathcal{N}(\mathbf{x}_{t+l_i} \mid \boldsymbol{\mu}^{(x)}_{t+l_i}, \boldsymbol{\Sigma}^{(x)}_{t+l_i})\,, \tag{21}$$

$$\boldsymbol{\mu}^{(x)}_{t+l_i} = \textstyle\sum_{n=1}^N \sqcap_i \bar{\mathbf{c}}_{nt}\,, \tag{22}$$

$$\boldsymbol{\Sigma}^{(x)}_{t+l_i} = \textstyle\sum_{n=1}^N \sum_{j=0,\ldots,i-1,i+1,\ldots,L} \sqcap_j \bar{\mathsf{C}}_{n(t+l_i-l_j)}\,. \tag{23}$$

Here we define $\sqcap_i$ operator that extracts the $(i{+}1)$-th subvector when applied to a vector and extracts the $(i{+}1)$-th submatrix on the block diagonal when applied to a matrix (see Fig. 2). Please observe that the mean vector (22) is defined by the component vectors (red lines in Fig. 1), whereas the covariance matrix (23) is defined by the parameters θ (grey lines in Fig. 1).

Although we omit the detailed derivation, we have obtained a specific form of the joint distribution $p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N \mid \theta)$ according to the above assumptions, and it turns out to be a zero-mean Gaussian distribution with a covariance matrix of the form:

$$\begin{bmatrix} \bar{\mathsf{X}}_t & \bar{\mathsf{C}}_{1t} & \bar{\mathsf{C}}_{2t} & \cdots & \bar{\mathsf{C}}_{Nt} \\ \bar{\mathsf{C}}_{1t} & \bar{\mathsf{C}}_{1t} & 0 & \cdots & 0 \\ \bar{\mathsf{C}}_{2t} & 0 & \bar{\mathsf{C}}_{2t} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{\mathsf{C}}_{Nt} & 0 & 0 & \cdots & \bar{\mathsf{C}}_{Nt} \end{bmatrix} \tag{24}$$

with

$$\bar{\mathsf{X}}_t = \begin{bmatrix} \mathsf{X}_t & & \\ & \ddots & \\ & & \mathsf{X}_{t+l_L} \end{bmatrix} + \textstyle\sum_{n=1}^N \mathrm{BoffDiag}\bar{\mathsf{C}}_{nt}\,, \tag{25}$$

where X_t has already been defined in (12) without omitting frequency f dependency, and BoffDiag extracts the block off-diagonal submatrices and zeros out the $L+1$ block diagonal submatrices of

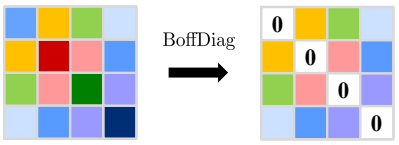

Fig. 3. BoffDiag operator applied to a matrix consisting of submatrices. Each colored box represents a submatrix of size $M \times M$.

size $M \times M$ (see Fig. 3). Once the joint distribution is derived, it is easy to derive the marginal (see below) and conditional (see Sec. 3.3) distributions, both of which are also Gaussian distributions [26].

The marginal distribution is obtained as

$$p(\bar{\mathbf{x}}_t \mid \theta) = \mathcal{N}(\bar{\mathbf{x}}_t \mid \mathbf{0}, \bar{\mathsf{X}}_t) \tag{26}$$

with (25) as the covariance matrix. The critical distinction between the proposed mfFCA and the existing FCAd is whether the off-diagonal submatrices $\mathbf{A}^{(l,l')}_n$ in (15) are introduced or not. If all the off-diagonal submatrices are set to zero, (25) becomes a block diagonal matrix with the RHS's second term becoming zero, and mfFCA reduces to FCAd. Thus, we understand that the second term is unique to the new model mfFCA.

3.3. Objective function and EM algorithm

According to the discussion so far, the objective function of our extension mfFCA is defined as the log-likelihood sum

$$\textstyle\sum_{t=1}^{T-l_L} \log p(\bar{\mathbf{x}}_t \mid \theta)\,, \tag{27}$$

by employing the long observation vector $\bar{\mathbf{x}}_t$ and assuming their independence (18). The EM algorithm to maximize the objective function (27) has been derived as follows. In the **E-step**, we calculate the conditional distributions of the long component vector $\bar{\mathbf{c}}_{nt}$ as

$$p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta) = \mathcal{N}(\bar{\mathbf{c}}_{nt} \mid \boldsymbol{\mu}^{(\bar{c})}_{nt}, \boldsymbol{\Sigma}^{(\bar{c})}_{nt})\,, \tag{28}$$

$$\boldsymbol{\mu}^{(\bar{c})}_{nt} = \bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}\bar{\mathbf{x}}_t\,, \quad \boldsymbol{\Sigma}^{(\bar{c})}_{nt} = \bar{\mathsf{C}}_{nt} - \bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}\bar{\mathsf{C}}_{nt}\,. \tag{29}$$

The part $\bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}$ for the mean $\boldsymbol{\mu}^{(\bar{c})}_{nt}$ calculation is called multi-frame multichannel Wiener filter in [27]. In the **M-step**, we want to optimize the parameters in θ by maximizing the so-called $\mathcal{Q}$ function defined with the set θ' of previous parameters

$$\mathcal{Q}(\theta, \theta') = \textstyle\sum_t \mathbb{E}_{\{p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta')\}_{n=1}^N} \log p(\bar{\mathbf{x}}_t, \{\bar{\mathbf{c}}_{nt}\}_{n=1}^N \mid \theta) \tag{30}$$

$$= -\textstyle\sum_{t=1}^T \sum_{i=0}^L \left\{ \log\det \boldsymbol{\Sigma}^{(x)}_{t+l_i} + \mathbb{E}_{\{p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta')\}_{n=1}^N} \mathcal{S}_{ti} \right\} \tag{31}$$

$$-\textstyle\sum_{t=1}^T \sum_{n=1}^N \left\{ \log\det\left(\mathsf{s}_{nt}\bar{\mathsf{A}}_n\right) + \mathrm{tr}\left[\left(\mathsf{s}_{nt}\bar{\mathsf{A}}_n\right)^{-1} \tilde{\mathsf{C}}_{nt} \right] \right\} \tag{32}$$

with $\mathcal{S}_{ti} = \mathrm{tr}\left[\left(\boldsymbol{\Sigma}^{(x)}_{t+l_i}\right)^{-1} \left(\mathbf{x}_{t+l_i} - \boldsymbol{\mu}^{(x)}_{t+l_i}\right) \left(\mathbf{x}_{t+l_i} - \boldsymbol{\mu}^{(x)}_{t+l_i}\right)^{\mathsf{H}} \right]$,

$$\tilde{\mathsf{C}}_{nt} = \mathbb{E}\left[\bar{\mathbf{c}}_{nt}\bar{\mathbf{c}}^{\mathsf{H}}_{nt} \mid \bar{\mathbf{x}}_t, \theta' \right] = \boldsymbol{\mu}^{(\bar{c})}_{nt}\boldsymbol{\mu}^{(\bar{c})\mathsf{H}}_{nt} + \boldsymbol{\Sigma}^{(\bar{c})}_{nt}\,. \tag{33}$$

However, we have found out that the exact maximization of $\mathcal{Q}$ by θ is difficult. To deal with the difficulty, we make an approximation that the parameter values $\sqcap_j \bar{\mathsf{C}}_{n(t+l_i-l_j)}$ used in (23) and appearing in the line (31) are fixed as the previous parameter values in θ', and as a consequence we optimize the parameters in θ as

$$\bar{\mathsf{A}}_n \leftarrow \frac{1}{T}\sum_{t=1}^T \frac{1}{\mathsf{s}_{nt}} \tilde{\mathsf{C}}_{nt}\,, \quad \mathsf{s}_{nt} \leftarrow \frac{1}{M(L{+}1)} \mathrm{tr}\left(\bar{\mathsf{A}}_n^{-1}\tilde{\mathsf{C}}_{nt}\right)\,. \tag{34}$$

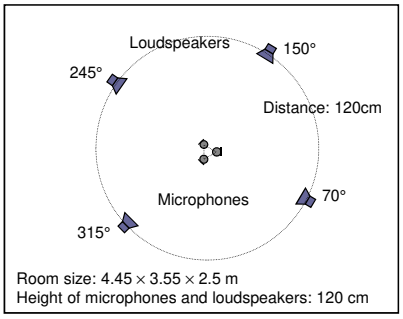

Fig. 4. Experimental setup

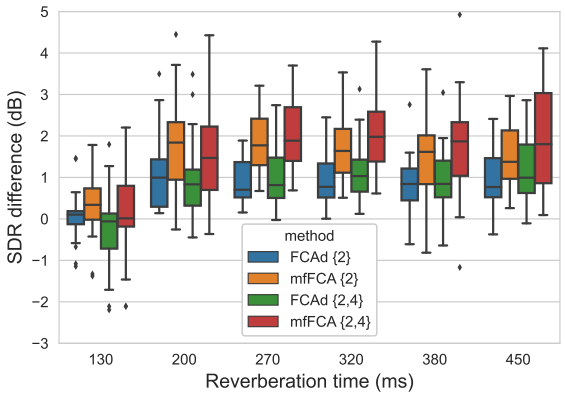

Fig. 5. SDR differences to the baseline FCA under various reverberation times. Each box plot shows the distribution of 32 differences (8 mixture combinations of $N = 4$ sources).

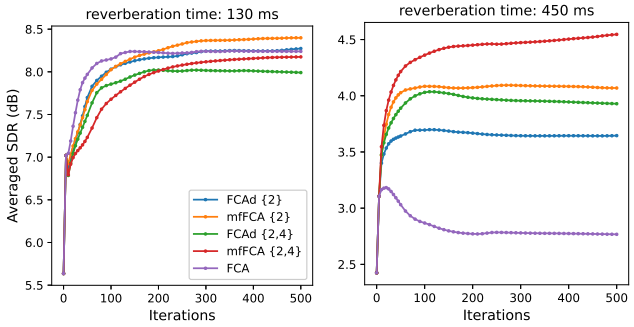

Fig. 6. Convergence behavior for a combination of 4 sources.

Once the parameters are optimized by the EM algorithm, we obtain separated signal $\mathbf{y}_{nt}$ for source n firstly by applying the multi-frame multichannel Wiener filter $\bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}$ to the long observation vector $\bar{\mathbf{x}}_t$ as in (29) to obtain $\boldsymbol{\mu}_{nt}^{(\bar{c})}$, and then by accumulating all the components originating from source n and observed at time frame t as

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

4. EXPERIMENTS

We performed experiments to separate $N = 4$ speech sources with $M = 3$ microphones. We measured the impulse responses from the sources (loudspeakers) to the microphones under the room conditions shown in Fig. 4. The room reverberation time was varied from 130 ms to 450 ms. For each reverberation time, 8 mixtures at the microphones were constructed by convolving the impulse responses and 6-second English speech sources. The sampling frequency was 8 kHz. The STFT window width and shift were 128 ms and 32 ms, respectively. Consequently, the numbers of time frames and frequency bins were $T = 201$ and $F = 513$, respectively. The separation performances were measured in signal-to-distortion ratios (SDRs) [28] by setting the source images with reverberations at the microphones as reference signals.

We examined two FCA extensions, i.e., FCAd (Subsection 2.3, conventional) and mfFCA (Section 3, our proposal), with two sets $\mathcal{L} = \{2\}$ and $\mathcal{L} = \{2, 4\}$ of time lags, which we considered effective for the quarter shift of STFT windows. Figure 5 shows the SDR differences to the baseline FCA under various reverberation times. To get an idea on how easy or hard these BSS tasks were, i.e., the absolute SDR values, please refer to Fig. 6 where two cases are shown. We observe that mfFCA outperformed the baseline by around 2 dB with an appropriate set of time lags accommodating to the reverberation time, except for a low reverberant 130 ms case. Although FCAd also outperformed the baseline, the improvements were clearly inferior to the proposed extension mfFCA.

Figure 6 shows convergence behavior as examples from low and high reverberant cases. At the 0-th iteration, the FCA parameters were initialized by the procedure shown in [29]. Then, after 5 iterations of FCA parameter updates, FCAd and mfFCA inherited the FCA parameters and iterated the updates of their own EM algorithms. In the low reverberant case, the baseline FCA converged the fastest and performed the best at 100 iterations. However, mfFCA with $\{2\}$ time lag gradually improved the separation and performed the best at 500 iterations. In the high reverberant case, mfFCA with $\{2, 4\}$ time lags performed the best by monotonically improving the performance as the iteration went on.

The algorithms were coded with Python using CuPy [30] and run on an Intel Core i7-8700K (3.70GHz) processor together with GeForce GTX 1080 Ti as a GPU. Thanks to the acceleration by the GPU and CuPy, FCA, FCAd $\{2\}$, mfFCA $\{2\}$, FCAd $\{2, 4\}$, mfFCA $\{2, 4\}$ took 94.3, 123.6, 151.4, 146.9, 260.6 seconds, respectively, for the 500 iterations.

5. CONCLUSION

We have proposed a new FCA model in which source components spanning multiple time frames are probabilistically modeled with covariance matrix (15) of larger dimensionality than the original FCA. We have developed new probabilistic models and then derived an EM algorithm to optimize the model parameters. Experimental results show that the proposed extension considerably improves the separation performance for underdetermined reverberant convolutive mixtures. While the new FCA model and algorithm can be employed as a dereverberation method such as WPE [14], we have not yet evaluated its dereverberation capability. Therefore, future work includes the assessment of dereverberation performance. Even we have accelerated the algorithm computation by a GPU, the execution times were far from the signal lengths (6 seconds). In this sense, we would like to reduce the computational complexity as another future work, for example, by the joint diagonalization approach [11].

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

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

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

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

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

[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] N.Q.K. 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, Sept. 2010.

[10] 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 under-determined reverberant audio source separation," in *Proc. ISSPA 2010*, May 2010, pp. 1–4.

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[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, Sept. 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] 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.

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

[27] 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 *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 905–911.

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

[29] 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.

[30] 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.