AUTOREGRESSIVE FAST MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION FOR JOINT BLIND SOURCE SEPARATION AND DEREVERBERATION

Kouhei Sekiguchi[1,2] *Yoshiaki Bando*[3,1] *Aditya Arie Nugraha*[1] *Mathieu Fontaine*[1] *Kazuyoshi Yoshii*[2,1]

[1]Center for Advanced Intelligence Project (AIP), RIKEN, Japan
[2]Graduate School of Informatics, Kyoto University, Japan
[3]National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper describes a joint blind source separation and dereverberation method that works adaptively and efficiently in a reverberant noisy environment. The modern approach to blind source separation (BSS) is to formulate a probabilistic model of multichannel mixture signals that consists of a *source model* representing the time-frequency structures of source spectrograms and a *spatial model* representing the inter-channel covariance structures of source images. The cutting-edge BSS method in this thread of research is fast multichannel nonnegative matrix factorization (FastMNMF) that consists of a low-rank source model based on nonnegative matrix factorization (NMF) and a full-rank spatial model based on jointly-diagonalizable spatial covariance matrices. Although FastMNMF is computationally efficient and can deal with both directional sources and diffuse noise simultaneously, its performance is severely degraded in a reverberant environment. To solve this problem, we propose autoregressive FastMNMF (AR-FastMNMF) based on a unified probabilistic model that combines FastMNMF with a blind dereverberation method called weighted prediction error (WPE), where all the parameters are optimized jointly such that the likelihood for observed reverberant mixture signals is maximized. Experimental results showed the superiority of AR-FastMNMF over conventional methods that perform blind dereverberation and BSS jointly or sequentially.

Index Terms— Blind source separation, blind dereverberation, multichannel nonnegative matrix factorization, joint diagonalization.

1. INTRODUCTION

Multichannel audio signal processing has been used in a wide variety of applications such as smart speakers, conversational robots, and hearing aid systems [1, 2], where the recorded signals are usually contaminated with utterances of non-target speakers, environmental noise, and reverberations in an unknown environment. To improve the speech intelligibility and the performance of automatic speech recognition (ASR), one may sequentially perform dereverberation and source separation (in the reverse order) for reverberant noisy recorded signals. This approach, however, is sub-optimal because the dereverberation and separation processes have mutually-dependent relationships. This calls for *joint blind* source separation (BSS) and dereverberation, where the acoustic characteristics of an environment are estimated adaptively without using any prior information.

A modern approach to BSS is to formulate a probabilistic generative model of observed mixture signals. A representative example is an underdetermined BSS method called multichannel nonnegative matrix factorization (MNMF) [3–5] that consists of an NMF-based source model representing the low-rank structure of the power spectral densities (PSDs) of sources and a full-rank spatial model representing the covariance structure of source images. MNMF, however, suffers from the high computational cost and the strong sensitivity to parameter initialization because of the high degree of freedom of the full-rank spatial covariance matrices (SCMs) of sources. Although a constrained version of MNMF called independent low-rank matrix analysis (ILRMA) [6] that restricts the SCMs to rank-1 matrices can mitigate this problem, it works only under a determined condition and the rank-1 assumption is often violated in reality. As an intermediate method between MNMF and ILRMA, FastMNMF [7–9] restricts the SCMs to jointly-diagonalizable (JD) yet full-rank matrices. FastMNMF was experimentally shown to be as computationally efficient as ILRMA and works better than MNMF and ILRMA [9].

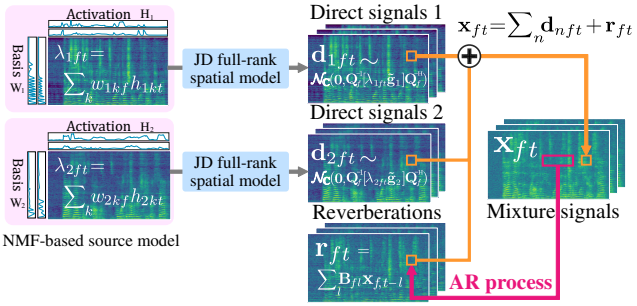

Fig. 1: The generative model of the reverberant mixture signals.

For blind dereverberation, a monaural reverberant signal is often represented by an autoregressive (AR) model that can deal with longer reverberation thanks to the long-term nature of the infinite impulse response. Using a vector extension of the AR model, weighted prediction error (WPE) [10, 11] has been proposed for blind dereverberation of multichannel audio signals, where the PSDs of a dry signal and a dereverberation filter are estimated jointly. To perform BSS in a reverberant environment, WPE was recently integrated with ILRMA (called AR-ILRMA in this paper) [12].

In this paper, we propose a joint blind source separation and dereverberation method called AR-FastMNMF that integrates FastMNMF with WPE in a statistically-principled manner (Fig. 1). Specifically, we formulate a probabilistic model of reverberant mixture signals that consists of 1) the NMF-based low-rank source model representing dry monaural source spectrograms, 2) the jointly-diagonalizable full-rank spatial model representing multichannel direct signals, and 3) the AR model adding reverberation to the sum of the direct signals through a feedback loop. Note that the proposed AR extension can be applied to any BSS methods based on the jointly-diagonalizable

This work is partially supported by JSPS KAKENHI Nos. 19H04137, 20H01159, 20K21813, and 20K19833.

spatial model and an arbitrary source model (*e.g.*, DNN-based speech model [13, 14] as in [8]). Thanks to the full-rank nature of the spatial model, AR-FastMNMF can deal with diffuse noise under an overdetermined or determined condition. We experimentally confirm that AR-FastMNMF outperform AR-ILRMA and a two-step method that uses WPE and FastMNMF sequentially in terms of separation performance and speech intelligibility in noisy reverberant environments.

2. RELATED WORK

For joint speech dereverberation and enhancement, one can sequentially use WPE and beamforming based on deep neural networks (DNNs) [15], where the time-frequency (TF) masks of dry speech are estimated with a DNN for calculating dereverberation filters [16, 17] and those of speech and noise are then estimated with another DNN for calculating demixing filters [18, 19]. While these DNNs are concatenated and jointly optimized in the training phase such that the ASR performance for the dereverberated enhanced speech is maximized, such a supervised approach increases the sensitivity to the environment. In the test phase, WPE and DNN-based beamforming can be used alternately and iteratively [20]. Extending this approach to multiple speech separation under a condition that the TF-masks of each source are given, a joint separation, dereverberation, and denoising method was proposed [21]. Although DNN-based mask estimation is computationally efficient, robust mask estimation from noisy reverberant mixture signals is still an open problem because the acoustic characteristics of a real environment may significantly differ from those covered by the training data.

For joint *blind* source separation and dereverberation, a BSS method called full-rank covariance analysis (FCA) [22] based on the full-rank spatial model was integrated with an autoregressive moving average (ARMA) model representing the reverberant process [23] (called ARMA-FCA in this paper). Although ARMA-FCA can deal with diffuse noise thanks to the full-rank spatial model, it needs to solve the permutation problem in a post-processing step because of the frequency-wise nature of source component estimation. To avoid the permutation problem under a determined condition, autoregressive ILRMA (AR-ILRMA) [12] that combines ILRMA [6] based on the rank-1 spatial model with WPE [10, 11] was proposed. In [24], the permutation problem of ARMA-FCA was alleviated by utilizing the parameters estimated by AR-ILRMA. The computational cost of ARMA-FCA, however, is larger than those of AR-ILRMA and AR-FastMNMF because of the unconstrained full-rank SCMs.

3. PROPOSED METHOD

This section explains the joint blind source separation and dereverberation method integrating the jointly-diagonalizable spatial model with an autoregressive (AR) model.

3.1. Model formulation

Assuming that a mixture of N sources are recorded by M microphones, let $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the short-time Fourier transform (STFT) coefficients of the observed multichannel mixture signals, where F and T represent the number of frequency bins and that of time frames, respectively. We formulate the observed reverberant mixture $\mathbf{x}_{ft}$ using an AR model, where the reverberation is represented by the convolutive mixture of the observed spectra of the previous frames as follows:

$$\mathbf{x}_{ft} = \mathbf{d}_{ft} + \mathbf{r}_{ft} = \sum_{n=1}^{N} \mathbf{d}_{nft} + \sum_{l=\Delta}^{\Delta+L-1} \mathbf{B}_{fl}\mathbf{x}_{f,t-l}, \qquad (1)$$

where $\mathbf{d}_{nft} \in \mathbb{C}^M$ is the direct signal of source n at frequency f and time t, $\mathbf{r}_{ft} \in \mathbb{C}^M$ is the reverberation represented by the AR model, $\mathbf{B}_{fl} = [\mathbf{b}_{fl1}, \ldots, \mathbf{b}_{flM}]^{\mathsf{T}} \in \mathbb{C}^{M\times M}$ is a set of AR coefficients, and $L\,(\geq 1)$ and $\Delta\,(\geq 1)$ represent the tap length and delay, respectively. The delay parameter is used for preserving the correlations inherent in the clean speech signals, and $\Delta = 3$ was used in our experiments.

We assume $\mathbf{d}_{nft}$ follows a circularly symmetric complex Gaussian distribution as follows:

$$\mathbf{d}_{nft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \lambda_{nft}\mathbf{G}_{nf}), \qquad (2)$$

where λ_{nft} represents the power spectral density (PSD) of source n. $\mathbf{G}_{nf} \in \mathbb{S}_+^M$ is the full-rank spatial covariance matrix (SCM) and $\mathbb{S}_+^M$ indicates the set of positive semidefinite matrices of size M. Using the low-rank source model based on NMF, the PSDs $\{\lambda_{nft}\}_{f,t=1}^{F,T}$ of each source n are assumed to have low-rank structure as follows:

$$\lambda_{nft} = \sum_{k=1}^{K} w_{nkf}h_{nkt}, \qquad (3)$$

where K is the number of bases, $w_{nkf} \geq 0$ is the magnitude of basis k of source n at frequency f, and $h_{nkt} \geq 0$ is the activation of basis k of source n at time t.

To reduce the degree of freedom as in FastMNMF [9], we restrict the SCMs of all sources to jointly-diagonalizable (JD) full-rank matrices using a set of diagonalizers $\mathbf{Q} \triangleq \{\mathbf{Q}_f \triangleq \mathbf{U}_f^{-1}\}_{f=1}^{F}$ and a set of nonnegative vectors $\tilde{\mathbf{G}} \triangleq \{\mathbf{g}_n\}_{n=1}^{N}$ as follows:

$$\forall n,\ \ \mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\mathbf{g}_n)\mathbf{Q}_f^{-\mathsf{H}} = \sum_{m=1}^{M} g_{nm}\mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathsf{H}}, \qquad (4)$$

where $\mathbf{g}_n \triangleq [g_{n1}, \ldots, g_{nM}]^{\mathsf{T}} \in \mathbb{R}_+^M$ is a frequency-independent nonnegative vector of source n, $\mathbf{Q}_f \triangleq [\mathbf{q}_{f1}, \ldots, \mathbf{q}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M\times M}$ and $\mathbf{U}_f \triangleq [\mathbf{u}_{f1}, \ldots, \mathbf{u}_{fM}] \in \mathbb{C}^{M\times M}$ are non-singular matrices at frequency f. Here, $\mathbf{g}_n$ is shared over all frequency bins for jointly diagonalizing the SCMs $\{\mathbf{G}_{nf}\}_{f=1}^{F}$ of source n. Because each rank-1 matrix $\mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathsf{H}}$ at frequency f corresponds to a particular source direction and $\mathbf{g}_n$ is considered to indicate the weights of M directions, $\mathbf{g}_n$ should be shared over all frequency bins [9].

Using Eqs. (2), (3), and (4) and the reproductive property of the Gaussian distribution, we say

$$\mathbf{d}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, \mathbf{Q}_f^{-1}\left(\sum_{n=1}^{N}\lambda_{nft}\mathrm{Diag}(\mathbf{g}_n)\right)\mathbf{Q}_f^{-\mathsf{H}}\right) \qquad (5)$$

$$\triangleq \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, \mathbf{Q}_f^{-1}\mathrm{Diag}(\mathbf{y}_{ft})\,\mathbf{Q}_f^{-\mathsf{H}}\right) \triangleq \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, \mathbf{Y}_{ft}\right), \qquad (6)$$

where $\mathbf{y}_{ft} \triangleq [y_{ft1}, \ldots, y_{ftM}]^{\mathsf{T}}$ and $y_{ftm} \triangleq \sum_{n=1}^{N}\lambda_{nft}g_{nm}$. Since the SCMs are full-rank matrices unlike AR-ILRMA based on rank-1 SCMs, $\mathbf{d}_{nft}$ can represent diffuse noise. From Eqs. (1) and (5), we have

$$p(\mathbf{x}_{ft}|\boldsymbol{\Theta}, \{\mathbf{x}_{f,t-l}\}_{l\in\mathbb{L}}) = \mathcal{N}_{\mathbb{C}}\left(\sum_{l\in\mathbb{L}}\mathbf{B}_{fl}\mathbf{x}_{f,t-l}, \mathbf{Y}_{ft}\right), \qquad (7)$$

where $\mathbb{L} \triangleq \{\Delta, \ldots, \Delta+L-1\}$, $\boldsymbol{\Theta} \triangleq \{\mathbf{W}, \mathbf{H}, \mathbf{Q}, \tilde{\mathbf{G}}, \mathbf{B}\}$, $\mathbf{W} \triangleq \{w_{nkf}\}_{n,k,f=1}^{N,K,F}$, $\mathbf{H} \triangleq \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$, and $\mathbf{B} \triangleq \{\mathbf{B}_{fl}\}_{f,l=1}^{F,L}$. If L is set to zero, AR-FastMNMF reduces to FastMNMF.

To estimate the direct signal $\mathbf{d}_{nft}$, we use a multichannel Wiener filter as follows:

$$\mathbb{E}[\mathbf{d}_{nft}|\mathbf{x}_{ft}] = \mathbf{Y}_{nft}\mathbf{Y}_{ft}^{-1}\left(\mathbf{x}_{ft} - \sum_{l\in\mathbb{L}}\mathbf{B}_{fl}\mathbf{x}_{f,t-l}\right) \qquad (8)$$

where $\mathbf{Y}_{nft} \triangleq \lambda_{nft}\mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\lambda_{nft}\mathbf{g}_n)\mathbf{Q}_f^{-\mathsf{H}}$.

3.2. Parameter estimation

The parameters $\boldsymbol{\Theta}$ are estimated such that the log-likelihood function $\log p(\mathbf{X}|\boldsymbol{\Theta})$ is maximized. From Eq. (7), $\log p(\mathbf{X}|\boldsymbol{\Theta}) = \sum_{f,t} \log p(\mathbf{x}_{ft}|\boldsymbol{\Theta}, \{\mathbf{x}_{f,t-l}\}_{l\in\mathbb{L}})$ is given by

$$\log p(\mathbf{X}|\boldsymbol{\Theta}) = -\sum_{f,t,m=1}^{F,T,M} \left(\frac{\tilde{d}_{ftm}}{y_{ftm}} + \log y_{ftm} \right) + T \sum_{f=1}^{F} \log |\mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}}|, \tag{9}$$

where $\tilde{d}_{ftm} \triangleq |\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{d}_{ft}|^2$ and $\mathbf{d}_{ft} = \mathbf{x}_{ft} - \sum_{l\in\mathbb{L}} \mathbf{B}_{fl} \mathbf{x}_{f,t-l}$.

3.2.1. Separation: Updating $\mathbf{W}$*,* $\mathbf{H}$*,* $\tilde{\mathbf{G}}$*, and* $\mathbf{Q}$

On condition that $\mathbf{B}$ is given, since Eq. (9) has the same form as the log-likelihood of FastMNMF [9], $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ can be updated in almost the same way as FastMNMF. $\mathbf{Q}_f$ is updated with iterative projection (IP) [8, 25] as follows:

$$\mathbf{V}_{fm} \triangleq \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbf{d}_{ft} \mathbf{d}_{ft}^{\mathsf{H}}}{y_{ftm}}, \tag{10}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f \mathbf{V}_{fm})^{-1} \mathbf{e}_m, \tag{11}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{V}_{fm} \mathbf{q}_{fm})^{-\frac{1}{2}} \mathbf{q}_{fm}, \tag{12}$$

where $\mathbf{e}_m$ is a one-hot vector whose m-th element is 1. The multiplicative update (MU) rules for $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ are given by

$$w_{nkf} \leftarrow w_{nkf} \sqrt{\frac{\sum_{t,m=1}^{T,M} h_{nkt} g_{nm} \tilde{d}_{ftm} y_{ftm}^{-2}}{\sum_{t,m=1}^{T,M} h_{nkt} g_{nm} y_{ftm}^{-1}}}, \tag{13}$$

$$h_{nkt} \leftarrow h_{nkt} \sqrt{\frac{\sum_{f,m=1}^{F,M} w_{nkf} g_{nm} \tilde{d}_{ftm} y_{ftm}^{-2}}{\sum_{f,m=1}^{F,M} w_{nkf} g_{nm} y_{ftm}^{-1}}}, \tag{14}$$

$$g_{nm} \leftarrow g_{nm} \sqrt{\frac{\sum_{f,t,k=1}^{F,T,K} w_{nkf} h_{nkt} \tilde{d}_{ftm} y_{ftm}^{-2}}{\sum_{f,t,k=1}^{F,T,K} w_{nkf} h_{nkt} y_{ftm}^{-1}}}. \tag{15}$$

To avoid the scale ambiguity, we adjust the scales of $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\mathbf{W}$ in this order in each iteration as follows:

$$\mu_f \triangleq \frac{1}{M} \mathrm{tr}(\mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}}), \qquad \begin{cases} \mathbf{Q}_f \leftarrow \mu_f^{-\frac{1}{2}} \mathbf{Q}_f, \\ w_{nkf} \leftarrow \mu_f^{-1} w_{nkf}, \end{cases} \tag{16}$$

$$\phi_n \triangleq \sum_{m=1}^{M} g_{nm}, \qquad \begin{cases} g_{nm} \leftarrow \phi_n^{-1} g_{nm}, \\ w_{nkf} \leftarrow \phi_n w_{nkf}, \end{cases} \tag{17}$$

$$\nu_{nk} \triangleq \sum_{f=1}^{F} w_{nkf}, \qquad \begin{cases} w_{nkf} \leftarrow \nu_{nk}^{-1} w_{nkf}, \\ h_{nkt} \leftarrow \nu_{nk} h_{nkt}. \end{cases} \tag{18}$$

3.2.2. Dereverberation: Updating $\mathbf{B}$

$\mathbf{B}_{fl}$ depends on only the first term of Eq. (9), and $\mathbf{r}_{ft}$ is rewritten as

$$\mathbf{r}_{ft} = \sum_{l\in\mathbb{L}} \mathbf{B}_{fl} \mathbf{x}_{f,t-l} = \hat{\mathbf{X}}_{ft} \hat{\mathbf{b}}_f. \tag{19}$$

$\hat{\mathbf{b}}_f$ and $\hat{\mathbf{X}}_{ft}$ are given as follows:

$$\hat{\mathbf{b}}_f \triangleq [\mathbf{b}_{f:1}^{\mathsf{T}}, \ldots, \mathbf{b}_{f:M}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M^2 L}, \tag{20}$$

$$\mathbf{b}_{f:m} \triangleq [\mathbf{b}_{f,\Delta,m}^{\mathsf{T}}, \ldots, \mathbf{b}_{f,\Delta+L-1,m}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML}, \tag{21}$$

$$\hat{\mathbf{X}}_{ft} \triangleq \mathbf{I}_M \otimes \bar{\mathbf{x}}_{ft}^{\mathsf{T}} \in \mathbb{C}^{M \times M^2 L}, \tag{22}$$

$$\bar{\mathbf{x}}_{ft} \triangleq [\mathbf{x}_{f,t-\Delta}^{\mathsf{T}}, \ldots, \mathbf{x}_{f,t-(\Delta+L-1)}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML}, \tag{23}$$

where $\mathbf{b}_{flm}$ is the m-th row vector of $\mathbf{B}_{fl}$. Substituting Eq. (19) into Eq. (9) and letting the partial derivative of Eq. (9) with respect to $\hat{\mathbf{b}}_f$ equal to zero, the update rule for $\hat{\mathbf{b}}_f$ is given by

$$\hat{\mathbf{b}}_f = \left(\sum_{t=1}^{T} \hat{\mathbf{X}}_{ft}^{\mathsf{H}} \mathbf{Y}_{ft}^{-1} \hat{\mathbf{X}}_{ft} \right)^{-1} \left(\sum_{t=1}^{T} \hat{\mathbf{X}}_{ft}^{\mathsf{H}} \mathbf{Y}_{ft}^{-1} \mathbf{x}_{ft} \right). \tag{24}$$

Although the first term accumulates T matrices of size $M^2 L \times M^2 L$, the memory usage and the computational cost can be reduced by rewriting the update rule as in [21] as follows:

$$\boldsymbol{\psi}_f \triangleq \sum_{m=1}^{M} \mathbf{q}_{fm} \otimes \left(\sum_{t=1}^{T} \frac{\mathbf{x}_{ft}^{\mathsf{H}} \mathbf{q}_{fm}}{y_{ftm}} \bar{\mathbf{x}}_{ft} \right)^{*}, \tag{25}$$

$$\boldsymbol{\Phi}_f \triangleq \sum_{m=1}^{M} (\mathbf{q}_{fm} \mathbf{q}_{fm}^{\mathsf{H}}) \otimes \left(\sum_{t=1}^{T} \frac{\bar{\mathbf{x}}_{ft} \bar{\mathbf{x}}_{ft}^{\mathsf{H}}}{y_{ftm}} \right)^{\mathsf{T}}, \tag{26}$$

$$\hat{\mathbf{b}}_f = \boldsymbol{\Phi}_f^{-1} \boldsymbol{\psi}_f, \tag{27}$$

where $(\cdot)^*$ indicates the complex conjugate. This rewrite is also applicable to AR-ILRMA, where y_{ftm} and $\mathbf{q}_{fm}^{\mathsf{H}}$ are replaced with λ_{nft} and the m-th row vector of the demixing matrix, but not applicable to ARMA-FCA because of the unconstrained full-rank SCMs.

4. EVALUATION

This section reports comparative experiments for evaluating AR-FastMNMF. We compare our method with the state-of-the-art unsupervised joint source separation and dereverberation methods using reverberant mixture signals of two speeches and noise signals.

4.1. Experimental conditions

We prepared a dataset of noisy reverberant mixture signals using the simulation data of REVERB Challenge dataset [26]. Each mixture signal consisted of diffuse noise recorded in real environments and two reverberant speech signals synthesized by convolving dry speech signals with real impulse responses from the development and evaluation subsets of REVERB Challenge dataset. The signal-to-noise ratio (SNR) between mixture of direct speech signals and noise was set to 0 or 10 dB. The impulse responses were recorded in three rooms with the reverberation times RT_{60} of 250 ms, 500 ms, and 700 ms. The distances between sound sources and microphones were set to 0.5 m (near) and 2.0 m (far). We thus tested six conditions in total, where 20 signals were used for each condition. Audio signals were sampled at 16 kHz and processed by STFT with a Hann window of 1024 points ($F = 513$) and a shifting interval of 256 points.

For comparison, we tested the sequential use of WPE [10, 11] and ILRMA [6], that of WPE and FastMNMF [9], AR-ILRMA [12], AR-FastMNMF (proposed), and two-step ARMA-FCA [27]. All methods were configured with $N = M = 8$ and the delay was set to $\Delta = 3$. The tap length was set to $L \in [0, 2, 4, 8]$, where AR-FastMNMF and AR-ILRMA with $L = 0$ reduce to vanilla FastMNMF and ILRMA, respectively. In ARMA-FCA, the tap length of the MA model was set to 4. The number of bases was set to $K \in [2, 4, 16]$. We showed the scores with the best K for each method in each SNR. In the sequential methods, WPE was updated 10 times, and FastMNMF or ILRMA was updated 150 times. $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\mathbf{B}$ of AR-FastMNMF were initialized to those estimated by AR-FastMNMF with $K = 2$ and 50 iterations. Similarly, AR-ILRMA was initialized using AR-ILRMA with $K = 2$. AR-FastMNMF and AR-ILRMA were then updated 100 times. In two-step ARMA-FCA, AR-ILRMA was updated 100 times, and then ARMA-FCA was updated 50 times as in [27].

Table 1: SDRs [dB] obtained by AR-FastMNMF and the conventional methods with best K.

Method	Observed	WPE + ILRMA			WPE + FastMNMF			AR-ILRMA				AR-FastMNMF				Two-step ARMA-FCA		
Tap length L	-	2	4	8	2	4	8	0	2	4	8	0	2	4	8	2	4	8
SNR=0, far	-4.2	7.3	7.6	7.5	8.9	9.4	9.4	5.2	7.7	7.9	6.5	6.6	9.1	**9.5**	9.2	8.9	8.8	7.9
SNR=0, near	-4.2	7.2	7.1	7.1	9.8	9.8	9.6	6.1	7.6	7.7	7.0	8.8	9.9	**10.1**	9.7	8.9	8.8	8.4
SNR=10, far	-2.5	10.9	11.6	11.7	11.0	11.7	11.8	7.8	10.8	11.6	10.6	8.1	11.1	12.0	**12.1**	11.1	10.9	9.5
SNR=10, near	-1.2	11.7	11.7	11.5	12.6	12.6	12.5	10.3	12.1	12.1	11.2	11.2	12.8	**13.1**	12.9	11.9	11.8	10.7

Table 2: PESQs obtained by AR-FastMNMF and the conventional methods with best K.

Method	Observed	WPE + ILRMA			WPE + FastMNMF			AR-ILRMA				AR-FastMNMF				Two-step ARMA-FCA		
Tap length L	-	2	4	8	2	4	8	0	2	4	8	0	2	4	8	2	4	8
SNR=0, far	1.13	1.26	1.28	1.28	1.36	1.38	1.38	1.24	1.28	1.30	1.24	1.30	1.37	**1.40**	1.36	1.35	1.34	1.32
SNR=0, near	1.10	1.35	1.34	1.33	1.54	1.55	1.51	1.33	1.39	1.40	1.31	1.47	1.55	**1.57**	1.50	1.45	1.47	1.41
SNR=10, far	1.15	1.47	1.52	1.52	1.54	1.59	1.59	1.32	1.49	1.54	1.47	1.38	1.55	**1.61**	1.60	1.55	1.56	1.47
SNR=10, near	1.10	1.78	1.80	1.75	1.97	1.99	1.97	1.65	1.85	1.87	1.74	1.82	2.00	**2.05**	**2.05**	1.83	1.83	1.70

We used the signal-to-distortion ratio (SDR) [28, 29] and the perceptual evaluation of speech quality (PESQ) [30] for evaluating the source estimation performance and the speech intelligibility, respectively. In both measures, we used dry speech signals without reverberation as reference signals.

4.2. Experimental results

Table 1 shows the average SDRs in the three different rooms with the SNRs of 0 dB and 10 dB under the far and near conditions. The sequential methods worked best when $K = 4$ and two-step ARMA-FCA worked best when $K = 2$. AR-FastMNMF worked best when $K = 4$ and $K = 16$ in the SNRs of 0 dB and 10 dB, respectively, and the same applied to AR-ILRMA. In all conditions, AR-FastMNMF outperformed the other methods. In particular, AR-FastMNMF performed significantly better than FastMNMF in the far condition and performed significantly better than AR-ILRMA when the SNR was 0 dB. Although longer reverberation violates the rank-1 spatial model of ILRMA, AR-ILRMA worked well even under a reverberant condition where standard ILRMA failed, because the reverberation, which increases the rank of the speech SCMs, was successfully removed by the AR model. However, since AR-ILRMA cannot deal with diffuse noise as shown in Fig. 2, the performance gap between AR-ILRMA and AR-FastMNMF became larger in noisier environments. One promising extension of AR-FastMNMF is to restrict only the SCMs of directional sources to rank-1 matrices and keeping the SCMs of diffuse noise to full-rank matrices as in [9].

Although two-step ARMA-FCA can deal with diffuse noise, it underperformed AR-FastMNMF. In two-step ARMA-FCA, the SCM of each source n is represented by the weighted sum of M rank-1 matrices unique to source n estimated by AR-ILRMA, and only the weights were optimized. In AR-FastMNMF, the SCMs of N sources are represented by the weighted sums of M common rank-1 matrices, where both the weights and matrices were optimized.

Table 2 shows the average PESQs in each condition. In all the situations, AR-FastMNMF outperformed the other methods. Although in terms of the SDR, AR-FastMNMF outperformed WPE+FastMNMF, the speech intelligibilities of these methods were almost the same in terms of the PESQ. One promising approach to further improving the performance is to integrate a moving average (MA) model as in ARMA-FCA [23], because in the MA model, the reverberations of each source can be formulated individually, while in the AR model, the reverberations of all sources are formulated as a whole.

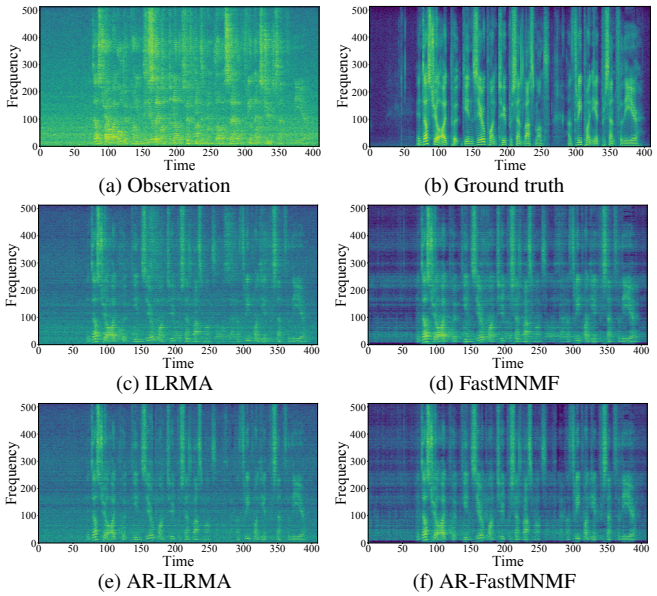

Fig. 2: Spectrogram examples in the far condition with the SNR of 0 dB and RT_{60} of 700 ms.

5. CONCLUSION

This paper presented a joint blind source separation and dereverberation method called AR-FastMNMF that integrates the source and spatial models of FastMNMF with the AR-based reverberation model. AR-FastMNMF is based on the maximum likelihood estimation of a unified probabilistic model of observed reverberant mixture signals, where the reverberation is modeled by an AR process and the time-frequency low-rank structures and channel covariance structures of direct signals are represented by NMF and jointly-diagonalizable full-rank SCMs, respectively. Thanks to the full-rank spatial model capable of dealing with diffuse noise, AR-FastMNMF outperformed AR-ILRMA by a large margin in a highly noisy environment.

For joint blind source separation, dereverberation, and denoising, we plan to use both the AR and MA models as in ARMA-FCA [24] and restrict the SCMs of only directional sources to rank-1 matrices as in rank-constrained FastMNMF [9]. Another interesting direction would be to integrate richer source models that can deal with the time-frequency covariance structures [31, 32] as in [33, 34]

6. REFERENCES

[1] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Punduk, K. Chin, K. C. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. Mcdermott, R. Rose, and M. Shannon, "Acoustic modeling for Google Home," in *Interspeech*, 2017, pp. 399–403.

[2] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, M. Souden, and Speech, "Speech processing for digital home assistants," *IEEE SP Mag.*, vol. 36, no. 6, pp. 111–124, 2019.

[3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[4] 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 *ISSPA*, 2010, pp. 1–4.

[5] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[6] 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 TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[7] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP*, 2019, pp. 371–375.

[8] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *EUSIPCO*, 2019.

[9] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM TASLP*, vol. 28, pp. 2610–2625, 2020.

[10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE TASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[11] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE TASLP*, vol. 20, no. 10, pp. 2707–2720, 2012.

[12] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *ICASSP*, 2018, pp. 31–35.

[13] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *ICASSP*, 2018, pp. 716–720.

[14] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2197–2212, 2019.

[15] A. S. Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," *arXiv preprint arXiv:1904.09049*, 2019.

[16] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Interspeech*, 2017, pp. 384–388.

[17] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," in *IWAENC*, 2018, pp. 466–470.

[18] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200.

[19] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Interspeech*, 2016, pp. 1981–1985.

[20] L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation," *Interspeech*, 2018.

[21] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM TASLP*, vol. 28, pp. 2267–2282, 2020.

[22] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

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

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

[25] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *WASPAA*, 2011, pp. 189–192.

[26] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Sign. Process*, vol. 2016, no. 7, pp. 1–19, 2016.

[27] M. Togami, "Joint training of deep neural networks for multi-channel dereverberation and speech source separation," in *ICASSP*, 2020, pp. 3032–3036.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A transparent implementation of common MIR metrics," in *ISMIR*, 2014, pp. 367–372.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

[31] K. Yoshii, "Correlated tensor factorization for audio source separation," in *ICASSP*, 2018, pp. 731–735.

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

[33] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "A unifying framework for blind source separation based on a joint diagonalizability constraint," in *EUSIPCO*, 2019, pp. 1–5.

[34] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "Independent low-rank matrix analysis with decorrelation learning," in *WASPAA*, 2019, pp. 288–292.