Autoregressive Moving Average Jointly-Diagonalizable Spatial Covariance Analysis for Joint Source Separation and Dereverberation

Kouhei Sekiguchi, Yoshiaki Bando, Aditya Arie Nugraha, Mathieu Fontaine, Kazuyoshi Yoshii, Tatsuya Kawahara

Autoregressive Moving Average Jointly-Diagonalizable Spatial Covariance Analysis for Joint Source Separation and Dereverberation

Kouhei Sekiguchi, *Member, IEEE*, Yoshiaki Bando, *Member, IEEE*, Aditya Arie Nugraha, *Member, IEEE*, Mathieu Fontaine, *Member, IEEE*, Kazuyoshi Yoshii, *Member, IEEE*, and Tatsuya Kawahara, *Fellow, IEEE*

***Abstract*—This article describes a computationally-efficient statistical approach to joint (semi-)blind source separation and dereverberation for multichannel noisy reverberant mixture signals. A standard approach to source separation is to formulate a generative model of a multichannel mixture spectrogram that consists of source and spatial models representing the time-frequency power spectral densities (PSDs) and spatial covariance matrices (SCMs) of source images, respectively, and find the maximum-likelihood estimates of these parameters. A state-of-the-art blind source separation method in this thread of research is fast multichannel nonnegative matrix factorization (FastMNMF) based on the low-rank PSDs and jointly-diagonalizable full-rank SCMs. To perform mutually-dependent separation and dereverberation jointly, in this paper we integrate both moving average (MA) and autoregressive (AR) models that represent the early reflections and late reverberations of sources, respectively, into the FastMNMF formalism. Using a pretrained deep generative model of speech PSDs as a source model, we realize semi-blind joint speech separation and dereverberation. We derive an iterative optimization algorithm based on iterative projection or iterative source steering for jointly and efficiently updating the AR parameters and the SCMs. Our experimental results showed the superiority of the proposed ARMA extension over its AR- or MA-ablated version in a speech separation and/or dereverberation task.**

***Index Terms*—Multichannel audio signal processing, source separation, dereverberation, joint diagonalization.**

Manuscript received 4 October 2021; revised 31 March 2022; accepted 18 June 2022. Date of publication 13 July 2022; date of current version 28 July 2022. This work was supported in part by JSPS KAKENHI under Grants 19H04137, 20K19833, and 20H01159, and in part by NII CRIS Collaborative Research Program operated by NII CRIS and LINE Corporation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Xie. (*Corresponding author: Kouhei Sekiguchi.*)

Kouhei Sekiguchi and Aditya Arie Nugraha are with the Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: sekiguchi92@gmail.com; adityaarie.nugraha@riken.jp).

Yoshiaki Bando is with the Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan, and also with the National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan (e-mail: y.bando@aist.go.jp).

Mathieu Fontaine is with the Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan, and also with the LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Paris, France (e-mail: mathieu.fontaine@telecom-paris.fr).

Kazuyoshi Yoshii is with the Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan, and also with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: yoshii@i.kyoto-u.ac.jp).

Tatsuya Kawahara is with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: kawahara@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2022.3190734

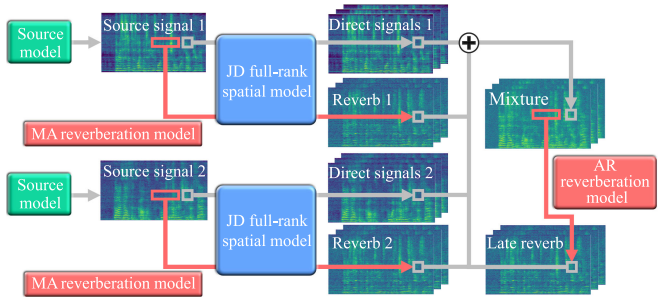


Fig. 1. The probabilistic generative model of a multichannel reverberant mixture spectrogram based on an arbitrary source model, a jointly-diagonalizable full-rank spatial model, and autoregressive (AR) and moving average (MA) reverberation models for joint source separation and dereverberation.

I. Introduction

MULTICHANNEL audio source separation and dereverberation play essential roles for computational auditory scene analysis with smart speakers, conversational robots, and hearing aid systems [1], [2] because real recordings are usually contaminated by utterances of non-target speakers, environmental noise, and reverberation. To improve the accuracy of automatic speech recognition (ASR) for a target speaker, *supervised* methods based on deep neural networks (DNNs) have actively been proposed for speech separation (enhancement) and dereverberation. Such methods, however, often work poorly in a real noisy echoic environment whose acoustic characteristics are not covered by training data. This calls for *unsupervised* methods that use neither prior information about the acoustic environment nor that about the sound sources (*blind* condition) or use only the latter (*semi-blind* condition).

A standard approach to unsupervised blind source separation (BSS) is to perform maximum-likelihood (ML) estimation for a probabilistic generative model of a multichannel mixture spectrogram (complex-valued tensor in the time-frequency-spatial domain) that consists of source and spatial models representing the time-frequency (TF) power spectral densities (PSDs) and spatial covariance matrices (SCMs) of sources, respectively [3]–[13]. Independent component/vector analysis (ICA [3] and IVA [4], [5]) are the most basic BSS methods assuming the SCMs to be *rank-1* matrices. To deal with moderate reverberation, the SCMs can be relaxed to *unconstrained* or *jointly-diagonalizable* (JD) full-rank matrices. Combining these three

types of spatial models with a source model based on nonnegative matrix factorization (NMF) [14], modern BSS methods such as independent low-rank matrix analysis (ILRMA) [6], multichannel NMF (MNMF) [7]–[10], and FastMNMF [11]–[13] are derived, respectively. When the target source is human speech, a DNN-based latent variable model learned from clean speech data [15] can be used as a precise source model for semi-blind speech enhancement and separation [16]–[19].

To jointly perform mutually-dependent BSS and blind source dereverberation (BSD), one can integrate the aforementioned generative model of a *dry* multichannel mixture spectrogram used for BSS with a multivariate autoregressive (AR) generative model of *late reverberation* used for BSD [20]–[23]. The AR reverberation model was originally proposed for a BSD method called weighted prediction error (WPE) that estimates both the PSDs of a target source with *early reflection* and the AR coefficients [24], [25]. To deal with multiple sources, WPE was integrated with ICA (called AR-ICA) [20], which was extended to use an NMF-based source model (called AR-ILRMA) [21] or a DNN-based source model [22]. More recently, WPE was integrated with FastMNMF based on an NMF-based source model and a JD full-rank spatial model to attain the state-of-the-art performance (called AR-FastMNMF) [23].

A practical problem of these joint blind source separation and dereverberation (BSSD) methods [20]–[23] is that a large number of computationally-expensive matrix inversions need to be computed for optimizing the AR coefficients (dereverberation matrices). To mitigate this problem, one can use an efficient algorithm called iterative projection (IP) [26] or iterative source steering (ISS) [27] for optimizing integrated dereverberation and demixing matrices [28], [29]. Note that IP was originally proposed for optimizing demixing matrices in ICA [26] and then used for IVA [30], ILRMA [6], and FastMNMF [12], [13] and that ISS was proposed as an alternative to IP. The joint diagonalizability (including the rank-1 constraint) of source SCMs plays a key role in the applicability of IP and ISS.

Another essential problem of the aforementioned BSSD methods [20]–[23] is that the early reflection is not represented explicitly. The early reflection strongly reflects the acoustic characteristics (e.g., PSDs) of dry sources. In contrast, the late reverberation is mainly affected by the acoustic characteristics of the surrounding environment because source-specific features are obscured in a complicated mixture of sounds corresponding to a large number of echoic propagation paths. This observation motivates us to use an autoregressive moving average (ARMA) model that represents the early reflection and late reverberation with a source-dependent moving average (MA) model and a source-independent AR model, respectively. For BSSD, the ARMA model was integrated with a BSS method called full-rank spatial covariance analysis (FCA) [31] based on a non-structured source model and an unconstrained full-rank spatial model (called ARMA-FCA) [32], [33]. Unfortunately, the IP- or ISS-based optimization algorithm [28], [29] cannot be used for ARMA-FCA because of the non-joint-diagonalizability of the source SCMs, resulting in heavy computational cost.

In this paper, we propose a unified BSSD framework based on a probabilistic generative model that consists of an arbitrary source model, a JD full-rank spatial model, and an ARMA reverberation model (Fig. 1)[1]. To use the efficient optimization algorithms [28], [29], we assume the SCMs of not only direct sounds but also the early reflections represented by the MA model to be jointly diagonalizable at once in each frequency bin, i.e., each of the SCMs is given as a weighted sum of common rank-1 SCMs. Using an NMF-based source model, we instantiate a new versatile BSSD method called ARMA-FastMNMF, which was experimentally found to achieve a state-of-the-art performance of BSSD. For a reverberant noisy environment, we propose a semi-blind method called ARMA-FastMNMF-DP that uses DNN- and NMF-based source models for representing the PSDs of directional speech and those of diffuse noise, respectively. A wide variety of directivity can be dealt with thanks to the full-rankness of the spatial and reverberation models.

The main contribution of this study is to propose the unified framework of computationally-efficient BSSD based on the JD full-rank spatial model and the ARMA reverberation model that encompasses BSS methods such as ICA [3], [26], IVA [4], [5], [30], ILRMA [6], FastMNMF [12], [13], and FastFCA [34] (a JD version of FCA [31]), a BSD method called WPE [24], and BSSD methods such as AR-ICA [20], AR-ILRMA [21], and AR-FastMNMF [23]. We have comprehensively investigated ARMA-FastMNMF with possible combinations of AR and MA parameters using IP or ISS, in comparison with the state-of-the-art existing blind and semi-blind methods.

The rest of this paper is organized as follows. Section II reviews unsupervised methods for separation, dereverberation, and joint separation and dereverberation. Section III explains a unified statistical framework based on source, spatial, and reverberation models for joint separation and dereverberation. Section IV describes the proposed blind and semi-blind methods as instances of the unified framework. Section V reports comparative experiments. Section VI concludes this paper.

II. Related Work

We here review *unsupervised* methods for source separation and/or dereverberation under a blind or semi-blind condition (Table I). The main focus of this paper is on unsupervised learning of some probabilistic model with a maximum-likelihood principle. Supervised methods, which are typically implemented with signal processing techniques informed by DNNs (e.g., beamforming and/or WPE with DNN-based TF mask estimation for speech enhancement [37], [38], joint speech enhancement and dereverberation [39], [40], and joint speech separation and dereverberation [41]), are thus not dealt with in this paper.

A. Separation

We review versatile BSS methods based on rank-1 and unconstrained and jointly-diagonalizable full-rank spatial models. We also review extensions of these methods based on a DNN-based source model for semi-blind speech separation.

1) Rank-1 Spatial Modeling: Assuming an instantaneous mixing system, each TF bin of a source image is typically

[1]The source code is available at https://github.com/sekiguchi92/TASLP2022

TABLE I
COMPARISON OF JOINT SOURCE SEPARATION AND DEREVERBERATION METHODS (SEMI-BLIND METHODS ARE INDICATED BY "*")

Method	Spatial model	Source model		Reverb. model	Condition
AR-ICA [20]	Rank-1	Non-structured		AR	$N = M$
AR-ILRMA [21], [28]		NMF			
AR-MVAE* [22]		DNN (speech)			
AR-OverIVA [35]		Freq. invariant (speech)	Time invariant (noise)		$N_{\text{speech}} + N_{\text{noise}} = M$
AR-OverILRMA [36]		NMF (speech)	Time invariant (noise)		
ARMA-FCA [32], [33]	Unconstrained full-rank	Non-structured		ARMA	Not applicable
ARMA-MNMF		NMF			
ARMA-FastFCA	Jointly-diag. full-rank	Non-structured			Not applicable ($N \leq M$ in practice)
ARMA-FastFIA		Frequency invariant			
ARMA-FastMNMF (cf. [23])		NMF			
ARMA-FastMNMF-TI		NMF (speech)	Time invariant (noise)		
ARMA-FastMNMF-DP*		DNN (speech)	NMF (noise)		

assumed to follow a *degenerate* multivariate complex Gaussian distribution whose covariance matrix is given by the product of a TF-varying PSD and a frequency-dependent rank-1 SCM. ICA [42] is the most basic method that estimates a demixing matrix in each frequency bin such that the separated sources are made independent. To avoid the permutation problem, independent vector analysis (IVA) [4], [5] jointly considers all frequency bins. Assuming the low-rankness of source PSDs, ILRMA [6] introduces an NMF-based source model. These BSS methods, however, are applicable to only a determined condition because as many source images with rank-1 SCMs as microphones should be added up to yield a mixture with a full-rank SCM. In an overdetermined condition, OverIVA [43], [44] and OverILRMA [36] internally recover a determined condition by padding additional sources of no interest.

2) Unconstrained Full-Rank Spatial Modeling: Relaxing the idealized rank-1 constraint, each TF bin of a source image is assumed to follow a *non-degenerate* multivariate complex Gaussian distribution with a full-rank SCM. FCA [31] pioneered this approach for dealing with diffuse noise and moderate reverberation. To avoid the permutation problem of FCA with the non-structured source model, MNMF [7]–[10] used the NMF-based source model, where ILRMA is its special case. These BSS methods are computationally demanding and hard to optimize because of the large degree of freedom (DOF).

3) Jointly-Diagonalizable Full-Rank Spatial Modeling: To attain a smaller DOF, the SCMs in each frequency bin are constrained to JD yet full-rank matrices. This strategy was first used for deriving FastFCA [34], [45] from FCA and then used for deriving FastMNMF [11]–[13] from MNMF, where a non-singular matrix called a *diagonalizer* used for jointly diagonalizing the SCMs can be optimized with fixed point iteration (FPI) [11], [34], [45] or IP [12], [13]. Note that IP is guaranteed to converge, whereas FPI is not. In this paper we focus on a well-behaved variant of FastMNMF with IP (called FastMNMF2 in [13]) that shares the direction weights of each source over all frequencies.

4) DNN-Based Source Modeling: Under a semi-blind condition that some sources are known to be human speeches, one can use a DNN-based latent variable model of speech PSDs as a precise source model. Such a speech model, for example, is obtained as the decoder of a variational autoencoder (VAE) [46] pretrained with clean speech signals in an unsupervised manner. Fixing the parameters of the speech model, the latent variables and a spatial model are adaptively estimated in an unsupervised manner at run-time. This approach was originally proposed for single-channel speech enhancement [15], and then used for multichannel speech enhancement and separation based on a full-rank spatial model [16]–[18], a rank-1 spatial model (called MVAE) [17], [19], and a JD spatial model [12].

B. Dereverberation

Linear prediction (LP) and its multichannel extension (MCLP) have effectively been used for BSD, where the reverberation is represented with an AR model in the time domain [47] or in the frequency domain [24], [25]. Assuming that the direct signal is Gaussian white noise, the AR coefficients are estimated in the maximum likelihood sense. This approach, however, tends to yield an over-whitened estimate of the direct signal.

WPE [24], [25] is an extension of MCLP based on a local Gaussian source model. The reverberation consists of the early reflection and late reverberation and the latter is empirically known to be more harmful to the speech intelligibility and ASR performance [48]. To avoid the over-whitening, a delay parameter is introduced for removing only the late reverberation. The PSDs of the target signal with the early reflection and the AR coefficients are updated alternately until convergence. A DNN can be used for estimating the PSDs at once [49].

C. Joint Separation and Dereverberation

AR-ICA [20] is an extension of ICA with the AR reverberation model for BSSD. Various BSS methods have been extended in the same way, resulting in AR-ILRMA [21], AR-MVAE [22], AR-FastMNMF [23], AR-OverIVA [35], and AR-OverILRMA [36]. ARMA-FCA [32] is an extension of FCA with the ARMA reverberation model. Both AR-FastMNMF and ARMA-FCA can deal with diffuse noise thanks to the full-rankness of the SCMs, but ARMA-FCA suffers from the permutation problem because of the frequency-wise independence of the non-structured source model. Under a determined condition, AR-ILRMA is used for initializing ARMA-FCA [33].

III. Unified Framework

This section provides a unified view of multichannel audio source separation and dereverberation based on a combination of source, spatial, and reverberation models.

A. Problem Specification

Let M, N, T, and F be the numbers of microphones, sources, time frames, and frequency bins, respectively. Let $\mathbf{S}_n \triangleq \{s_{nft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T}$ be the single-channel complex spectrogram (short-time Fourier transform (STFT) coefficients) of source n. Let $\mathbf{X} \triangleq \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the multichannel complex spectrogram of a (reverberant) mixture and $\mathbf{X}_n \triangleq \{\mathbf{x}_{nft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be that of source n (called an *image*). We assume the additivity of complex spectra:

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{nft}. \tag{1}$$

Given $\mathbf{X}$ as observed data, we aim to optimize a generative model of $\mathbf{X}$ such that the log-likelihood $\log p(\mathbf{X})$ is maximized and then decompose $\mathbf{X}$ into the source images $\{\mathbf{X}_n\}_{n=1}^{N}$.

B. Source Models

The source model represents a generative process of each source spectrogram $\mathbf{S}_n$. Assuming both the independence of sources and that of time-frequency bins, s_{nft} is assumed to follow a complex Gaussian distribution as follows:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{nft}\right), \tag{2}$$

where λ_{nft} represents the power spectral density (PSD) of source n at frequency f and time t and $\mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$ indicates a univariate complex Gaussian distribution with mean μ and variance σ^2. Let $\mathbf{\Lambda} \triangleq \{\lambda_{nft}\}_{n,f,t=1}^{N,F,T}$.

1) Frequency-Invariant Source Model: As in IVA [4], [5], an effective way of avoiding the permutation problem is to use a frequency-invariant (FI) source model given by

$$\lambda_{nft} = \gamma_{nt}, \tag{3}$$

where γ_{nt} is the frequency-invariant PSD of source n at time t. Let $\boldsymbol{\gamma} \triangleq \{\gamma_{nt}\}_{n,t=1}^{N,T}$.

2) Time-Invariant Source Model: As in OverIVA [43], [44], stationary sources (e.g., background noise) are well represented with a time-invariant (TI) source model given by

$$\lambda_{nft} = \eta_{nf}, \tag{4}$$

where η_{nf} is the PSD of source n at frequency f. Let $\boldsymbol{\eta} \triangleq \{\eta_{nf}\}_{n,f=1}^{N,F}$.

3) NMF-Based Source Model: The NMF-based source model assumes the PSDs $\{\lambda_{nft}\}_{f,t=1}^{F,T}$ of each source n to have low-rank structure as follows:

$$\lambda_{nft} = \sum_{k=1}^{K} w_{nkf} h_{nkt}, \tag{5}$$

where K is the number of bases, $w_{nkf} \geq 0$ is the PSD 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. Let $\mathbf{W} \triangleq \{w_{nkf}\}_{n,k,f=1}^{N,K,F}$ and $\mathbf{H} \triangleq \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$.

4) DNN-Based Source Model: To precisely represent the PSDs of a particular type of source (e.g., speech), one can formulate a DNN-based source model as follows [15]:

$$\lambda_{nft} = \alpha_{nf}\beta_{nt}\left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2\left(\mathbf{z}_{nt}\right)\right]_f \tag{6}$$

where $\alpha_{nf} \geq 0$ and $\beta_{nt} \geq 0$ are scaling factors of source n at frequency f and time t, respectively, $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\cdot)$ is a DNN with parameters $\boldsymbol{\theta}$ that maps a latent variable $\mathbf{z}_{nt} \in \mathbb{R}^D$ to a nonnegative vector $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_{nt}) \in \mathbb{R}_+^F$, and $[\cdot]_f$ indicates the f-th element of a vector. To estimate the parameters $\boldsymbol{\theta}$, a VAE is trained in an unsupervised manner using a large amount of clean speech data. The decoder is used as $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\cdot)$ and the encoder can be used for initializing and/or inferring $\mathbf{z}_{nt}$ at run-time. Let $\boldsymbol{\alpha} \triangleq \{\alpha_{nf}\}_{n,f=1}^{N,F}$, $\boldsymbol{\beta} \triangleq \{\beta_{nt}\}_{n,t=1}^{N,T}$, and $\mathbf{Z} \triangleq \{\mathbf{z}_{nt}\}_{n,t=1}^{N,T}$.

C. Spatial Models

The spatial model represents a generative model of each source image $\mathbf{X}_n \in \mathbb{C}^{F\times T\times M}$ based on the corresponding source spectrogram $\mathbf{S}_n \in \mathbb{C}^{F\times T}$.

1) Rank-1 Spatial Model: Assuming a time-invariant linear system, we have

$$\mathbf{x}_{nft} = \mathbf{a}_{nf} s_{nft}, \tag{7}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the steering vector of source n at frequency f. Using (2) and (7), we have

$$\mathbf{x}_{nft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{nft}\mathbf{G}_{nf}), \tag{8}$$

where $\mathbf{G}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+^M$ is a rank-1 SCM of source n at frequency f and $\mathbb{S}_+^M$ indicates the set of positive semidefinite matrices of size M. Using (1) and (8) and the additive property of the Gaussian distribution, we have

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathbf{G}_{nf}\right). \tag{9}$$

2) Full-Rank Spatial Model: To deal with diffuse noise and modest reverberation, the SCMs $\{\mathbf{G}_{nf}\}_{n,f=1}^{N,F}$ are assumed to be unconstrained full-rank matrices. Since the full-rank spatial model has a considerably larger number of parameters than the rank-1 spatial model, an iterative parameter optimization algorithm is computationally expensive and tends to get stuck in bad local optima.

3) Jointly-Diagonalizable Full-Rank Spatial Model: To mitigate these problems, all the SCMs $\{\mathbf{G}_{nf}\}_{n=1}^{N}$ are restricted to jointly-diagonalizable (JD) matrices as follows [13]:

$$\mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\,\mathbf{Q}_f^{-\mathsf{H}} = \sum_{m=1}^{M} \tilde{g}_{nm}\mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathsf{H}}, \tag{10}$$

where $\mathbf{Q}_f \triangleq [\mathbf{q}_{f1}, \ldots, \mathbf{q}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M\times M}$ is a non-singular matrix called *a diagonalizer*, $\tilde{\mathbf{g}}_n \triangleq [\tilde{g}_{n1}, \ldots, \tilde{g}_{nM}]^{\mathsf{T}} \in \mathbb{R}_+^M$ is a nonnegative vector specific to source n, and $\mathbf{u}_{fm}$ is the m-th column of $\mathbf{U}_f \triangleq \mathbf{Q}^{-1}$. Under a determined condition, when each $\tilde{\mathbf{g}}_n$ is a one-hot vector, the JD spatial model reduces to

the rank-1 spatial model. Because $\{\mathbf{u}_{fm}\}_{m=1}^{M}$ act like steering vectors corresponding to M directions, $\tilde{\mathbf{g}}_n$ can be regarded as the weights of these directions for source n [13]. FastMNMF sharing $\tilde{\mathbf{g}}_n$ over all frequency bins [13] outperforms FastMNMF using frequency-wise weight vectors [11], [12].

D. Reverberation Models

When the reverberation is longer than the window size of STFT, each source image $\mathbf{x}_{nft}$ is represented as the sum of the direct sound $\mathbf{x}^{\mathrm{d}}_{nft}$ corresponding to (7), the early reflection $\mathbf{x}^{\mathrm{e}}_{nft}$, and the late reverberation $\mathbf{x}^{\mathrm{l}}_{nft}$ as follows:

$$\mathbf{x}_{nft} = \mathbf{x}^{\mathrm{d}}_{nft} + \mathbf{x}^{\mathrm{e}}_{nft} + \mathbf{x}^{\mathrm{l}}_{nft}. \tag{11}$$

The observed reverberant mixture $\mathbf{x}_{ft}$ is then given by

$$\mathbf{x}_{ft} = \mathbf{x}^{\mathrm{d}}_{ft} + \mathbf{x}^{\mathrm{e}}_{ft} + \mathbf{x}^{\mathrm{l}}_{ft}, \tag{12}$$

where $\mathbf{x}^{*}_{ft} \triangleq \sum_{n=1}^{N} \mathbf{x}^{*}_{nft}$ $(* \in \{\mathrm{d}, \mathrm{e}, \mathrm{l}\})$.

The early reflection $\mathbf{x}^{\mathrm{e}}_{ft}$ strongly reflects the PSDs of individual sources, whereas the late reverberation $\mathbf{x}^{\mathrm{l}}_{ft}$ is a complicated mixture of sounds arriving through numerous echoic paths. This calls for an ARMA model that represents $\mathbf{x}^{\mathrm{e}}_{nft}$ with a *source-dependent* MA model (FIR filter) with a tap length L_{MA} and $\mathbf{x}^{\mathrm{l}}_{ft}$ with a *source-independent* AR model (IIR filter) with a tap length L_{AR} as follows [32]:

$$\mathbf{x}_{nftl} \triangleq \mathbf{a}_{nfl} s_{nf,t-l}, \tag{13}$$

$$\mathbf{x}^{\mathrm{d}}_{nft} = \mathbf{x}_{nft0}, \tag{14}$$

$$\mathbf{x}^{\mathrm{e}}_{nft} = \sum_{l \in \mathbb{I}_{\mathrm{MA}}} \mathbf{x}_{nftl}, \tag{15}$$

$$\mathbf{x}^{\mathrm{d+e}}_{nft} \triangleq \sum_{l \in \mathbb{I}^{+}_{\mathrm{MA}}} \mathbf{x}_{nftl} = \mathbf{x}^{\mathrm{d}}_{nft} + \mathbf{x}^{\mathrm{e}}_{nft}, \tag{16}$$

$$\mathbf{x}^{\mathrm{d+e}}_{ft} \triangleq \sum_{n=1}^{N} \mathbf{x}^{\mathrm{d+e}}_{nft} = -\sum_{l \in \mathbb{I}^{+}_{\mathrm{AR}}} \mathbf{B}_{fl} \mathbf{x}_{f,t-l} \tag{17}$$

$$\mathbf{x}^{\mathrm{l}}_{ft} = \sum_{l \in \mathbb{I}_{\mathrm{AR}}} \mathbf{B}_{fl} \mathbf{x}_{f,t-l}, \tag{18}$$

where $\mathbb{I}_{\mathrm{MA}} \triangleq [1, L_{\mathrm{MA}}]$, $\mathbb{I}^{+}_{\mathrm{MA}} \triangleq \{0\} \cup \mathbb{I}_{\mathrm{MA}}$, $\mathbb{I}_{\mathrm{AR}} \triangleq [\Delta, \Delta + L_{\mathrm{AR}} - 1]$, $\mathbb{I}^{+}_{\mathrm{AR}} \triangleq \{0\} \cup \mathbb{I}_{\mathrm{AR}}$ are index sets, $\Delta > 0$ is the *delay* of the late reverberation [24], $\mathbf{a}_{nfl} \in \mathbb{C}^{M}$ is the steering vector of source n with delay l at frequency f and, $\mathbf{B}_{fl} \triangleq [\mathbf{b}_{fl1}, \ldots, \mathbf{b}_{flM}]^{\mathsf{T}} \in \mathbb{C}^{M \times M}$ is an AR coefficient. Let $\mathbf{B} \triangleq \{\mathbf{B}_{fl}\}^{F}_{f=1, l \in \mathbb{I}^{+}_{\mathrm{AR}}}$, where $\mathbf{B}_{f0} \triangleq -\mathbf{I}_M$ and $\mathbf{I}_M$ denotes an identity matrix of size M.

When $L_{\mathrm{MA}} = 0$ and $N = 1$, the AR model proposed for WPE [24] is obtained, where $\Delta = 2$ or 3 is used. To improve the expressiveness of the ARMA model, one can allow the MA and AR periods to overlap with a configuration of $L_{\mathrm{MA}} \geq \Delta$.

E. Integration of Source, Spatial, and Reverberation Models

A unified probabilistic model of $\mathbf{X}$ is obtained by arbitrarily integrating the aforementioned source, spatial, and reverberation models. Using (2), (14), and (15), we have

$$\mathbf{x}^{\mathrm{d}}_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{Y}_{nft0} \triangleq \mathbf{Y}^{\mathrm{d}}_{nft}\right), \tag{19}$$

$$\mathbf{x}^{\mathrm{e}}_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{l \in \mathbb{I}_{\mathrm{MA}}} \mathbf{Y}_{nftl} \triangleq \mathbf{Y}^{\mathrm{e}}_{nft}\right), \tag{20}$$

where $\mathbf{Y}_{nftl} \triangleq \lambda_{nf,t-l} \mathbf{G}_{nfl} \in \mathbb{S}^{M}_{+}$ and $\mathbf{G}_{nfl} = \mathbf{a}_{nfl} \mathbf{a}^{\mathsf{H}}_{nfl} \in \mathbb{S}^{M}_{+}$ is a rank-1 SCM of source n at frequency f and delay $l \in \mathbb{I}^{+}_{\mathrm{MA}}$. Let $\mathbf{G} \triangleq \{\mathbf{G}_{nfl}\}^{N,F}_{n,f=1,l \in \mathbb{I}^{+}_{\mathrm{MA}}}$. The rank-1 constraint (Section III-C1) can be removed (Section III-C2) or relaxed to the joint diagonalizability (Section III-C3). The PSDs of each source can be represented with a frequency-invariant (Section III-B1), time-invariant (Section III-B2), NMF-based (Section III-B3), or DNN-based (Section III-B4) source model. Using (12), (18), (19), and (20), the observed reverberant mixture $\mathbf{x}_{ft}$ can be modeled in an AR manner as follows:

$$\mathbf{x}_{ft} \mid \{\mathbf{x}_{f,t-l}\}_{l \in \mathbb{I}_{\mathrm{AR}}} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}^{\mathrm{l}}_{ft}, \sum_{n=1}^{N} \sum_{l \in \mathbb{I}^{+}_{\mathrm{MA}}} \mathbf{Y}_{nftl} \triangleq \mathbf{Y}^{\mathrm{d+e}}_{ft}\right). \tag{21}$$

Given $\mathbf{X}$ as observed data, our goal is to estimate the PSDs $\mathbf{\Lambda}$, the SCMs $\mathbf{G}$, and the AR coefficients $\mathbf{B}$ such that the total likelihood for $\mathbf{X}$ given by aggregating (21) over all TF bins is maximized. Once these parameters are estimated, the late reverberation $\mathbf{x}^{\mathrm{l}}_{ft}$ is given by (18) and the direct sound $\mathbf{x}^{\mathrm{d}}_{nft}$ and early reflection $\mathbf{x}^{\mathrm{e}}_{nft}$ of each source n can be inferred with Wiener filtering as follows:

$$\mathbb{E}\left[\mathbf{x}^{\mathrm{d}}_{nft} \mid \mathbf{X}\right] = \mathbf{Y}^{\mathrm{d}}_{nft} \left(\mathbf{Y}^{\mathrm{d+e}}_{ft}\right)^{-1} \left(\mathbf{x}_{ft} - \mathbf{x}^{\mathrm{l}}_{ft}\right), \tag{22}$$

$$\mathbb{E}\left[\mathbf{x}^{\mathrm{e}}_{nft} \mid \mathbf{X}\right] = \mathbf{Y}^{\mathrm{e}}_{nft} \left(\mathbf{Y}^{\mathrm{d+e}}_{ft}\right)^{-1} \left(\mathbf{x}_{ft} - \mathbf{x}^{\mathrm{l}}_{ft}\right). \tag{23}$$

Various unsupervised (semi-)blind methods can be instantiated from the unified model given by (21). Modern BSS methods such as ILRMA [6], MNMF [9], and FastMNMF [13] without the reverberation model ($L_{\mathrm{MA}} = L_{\mathrm{AR}} = 0$) are given by integrating the NMF-based source model with the rank-1, unconstrained, and JD full-rank spatial models, respectively. Semi-blind extensions for speech enhancement such as ILRMA-DP [17], MNMF-DP [16], [17], and FastMNMF-DP [12] are given by using the NMF- and DNN-based source models for noise and speech sources, respectively. A semi-blind extension of ILRMA for speech separation called MVAE [19] is given by using only the DNN-based source model. As listed in Table I, extensions for joint separation and dereverberation such as AR-ILRMA [21], AR-FastMNMF [23], and AR-MVAE [22] are given by introducing the AR reverberation model ($L_{\mathrm{MA}} = 0$ and $L_{\mathrm{AR}} > 0$). A BSSD method called ARMA-FCA [32] without any assumption on the source PSDs $\mathbf{\Lambda}$ is given by integrating the full-rank spatial model and the ARMA reverberation model, which could be straightforwardly extended to ARMA-MNMF based on the NMF-based source model.

IV. Proposed Method

Based on the unified model given by (21), we propose a state-of-the-art joint BSSD method called ARMA-FastMNMF

(including AR-FastMNMF [23]) using the NMF-based source model, the JD full-rank spatial model, and the ARMA reverberation model (Table I). For stable initialization, we also derive a variant of ARMA-FastMNMF called ARMA-FastFIA using the FI source model.

Under a *semi-blind* condition, we propose ARMA-FastMNMF-DP using the DNN- and NMF-based source models for speech and noise, respectively. For initialization of ARMA-FastMNMF-DP, we propose ARMA-FastMNMF-TI using the NMF-based and TI source models for speech and noise, respectively.

A. Probabilistic Formulation

We integrate the MA model with the JD full-rank spatial model given by (10). Specifically, we assume that all the SCMs $\{\mathbf{G}_{nfl}\}_{n=1,l=0}^{N,L_{\mathrm{MA}}}$ corresponding to not only the direct sounds but also the early reflections are jointly diagonalizable for each frequency bin as follows:

$$\mathbf{G}_{nfl} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_{nl})\mathbf{Q}_f^{-\mathsf{H}} = \sum_{m=1}^{M} \tilde{g}_{nlm}\mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathsf{H}}, \quad (24)$$

where $\mathbf{Q}_f \triangleq [\mathbf{q}_{f1}, \ldots, \mathbf{q}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M\times M}$ is a diagonalizer, $\tilde{\mathbf{g}}_{nl} \triangleq [\tilde{g}_{nl1}, \ldots, \tilde{g}_{nlM}]^{\mathsf{T}} \in \mathbb{R}_+^M$ is a nonnegative vector specific to source n and delay l, and $\mathbf{u}_{fm}$ is the m-th column vector of $\mathbf{U}_f \triangleq \mathbf{Q}^{-1}$. This means that each of $N(L_{\mathrm{MA}}+1)$ SCMs corresponding to N direct paths and NL_{MA} echoic paths are represented by a weighted sum of M rank-1 matrices $\{\mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathsf{H}}\}_{m=1}^{M}$. Let $\tilde{\mathbf{G}} \triangleq \{\tilde{\mathbf{g}}_{nl}\}_{n=1,l\in\mathbb{I}_{\mathrm{MA}}^+}^{N}$ and $\mathbf{Q} \triangleq \{\mathbf{Q}_f\}_{f=1}^{F}$.

For the sake of brevity, we define several symbols. Let $\tilde{\mathbf{x}}_{ft}$ and $\tilde{\mathbf{x}}_{ft}^*$ be the absolute squares of the linear transforms of $\mathbf{x}_{ft}$ and $\mathbf{x}_{ft}^*$ ($* \in \{\mathrm{d}, \mathrm{e}, \mathrm{l}, \mathrm{d{+}e}\}$), respectively, as follows:

$$\tilde{\mathbf{x}}_{ft} \triangleq |\mathbf{Q}_f\mathbf{x}_{ft}|^{\cdot 2}, \quad (25)$$

$$\tilde{\mathbf{x}}_{ft}^* \triangleq |\mathbf{Q}_f\mathbf{x}_{ft}^*|^{\cdot 2}, \quad (26)$$

where $|\mathbf{z}|^{\cdot 2}$ denotes the element-wise absolute square for a vector $\mathbf{z}$. Let $\tilde{\mathbf{y}}_{nftl}$, $\tilde{\mathbf{y}}_{nft}^{\mathrm{d}}$, $\tilde{\mathbf{y}}_{nft}^{\mathrm{e}}$, $\tilde{\mathbf{y}}_{nft}^{\mathrm{d+e}}$, and $\tilde{\mathbf{y}}_{ft}^{\mathrm{d+e}}$ be the predicted PSDs corresponding to (13)–(17) as follows:

$$\tilde{\mathbf{y}}_{nftl} \triangleq \lambda_{nf,t-l}\tilde{\mathbf{g}}_{nl}, \quad (27)$$

$$\tilde{\mathbf{y}}_{nft}^{\mathrm{d}} \triangleq \tilde{\mathbf{y}}_{nft0}, \quad (28)$$

$$\tilde{\mathbf{y}}_{nft}^{\mathrm{e}} \triangleq \sum_{l\in\mathbb{I}_{\mathrm{MA}}} \tilde{\mathbf{y}}_{nftl}, \quad (29)$$

$$\tilde{\mathbf{y}}_{nft}^{\mathrm{d+e}} \triangleq \sum_{l\in\mathbb{I}_{\mathrm{MA}}^+} \tilde{\mathbf{y}}_{nftl} = \tilde{\mathbf{y}}_{nft}^{\mathrm{d}} + \tilde{\mathbf{y}}_{nft}^{\mathrm{e}}, \quad (30)$$

$$\tilde{\mathbf{y}}_{ft}^{\mathrm{d+e}} \triangleq \sum_{n=1}^{N} \tilde{\mathbf{y}}_{nft}^{\mathrm{d+e}}. \quad (31)$$

Substituting (24) into (21), the generative model of an observed reverberant mixture $\mathbf{x}_{ft}$ is given by

$$\mathbf{x}_{ft} \mid \{\mathbf{x}_{f,t-l}\}_{l\in\mathbb{I}_{\mathrm{AR}}} \sim \mathcal{N}_{\mathbb{C}}\Big(\mathbf{x}_{ft}^{\mathrm{l}}, \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{y}}_{ft}^{\mathrm{d+e}})\mathbf{Q}_f^{-\mathsf{H}}\Big). \quad (32)$$

B. Maximum Likelihood Estimation

We aim to estimate the parameters $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\Lambda}, \mathbf{Q}, \tilde{\mathbf{G}}, \mathbf{B}\}$ such that the log-likelihood $\log p(\mathbf{X})$ obtained by aggregating (32) over all TF bins is maximized. Using (25)–(31), $\log p(\mathbf{X})$ can be briefly written as follows:

$$\log p(\mathbf{X}) = \sum_{f,t,m=1}^{F,T,M}\left(\log\tilde{y}_{ftm}^{\mathrm{d+e}} - \frac{\tilde{x}_{ftm}^{\mathrm{d+e}}}{\tilde{y}_{ftm}^{\mathrm{d+e}}}\right) + T\sum_{f=1}^{F}\log\left|\mathbf{Q}_f\mathbf{Q}_f^{\mathsf{H}}\right|, \quad (33)$$

where $\tilde{x}_{ftm}^{\mathrm{d+e}}$ and $\tilde{y}_{ftm}^{\mathrm{d+e}}$ are given by

$$\tilde{x}_{ftm}^{\mathrm{d+e}} = \left|\mathbf{q}_{fm}^{\mathsf{H}}\left(\mathbf{x}_{ft} - \sum_{l\in\mathbb{I}_{\mathrm{AR}}}\mathbf{B}_{fl}\mathbf{x}_{f,t-l}\right)\right|^2, \quad (34)$$

$$\tilde{y}_{ftm}^{\mathrm{d+e}} = \sum_{n=1}^{N}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\lambda_{nf,t-l}\tilde{g}_{nlm}. \quad (35)$$

1) Updating $\boldsymbol{\Lambda}$: Using the current estimates of $\mathbf{Q}, \tilde{\mathbf{G}}$, and $\mathbf{B}$, we update $\boldsymbol{\Lambda}$. Since $\boldsymbol{\Lambda}$ is involved only in the first and second terms of (33), the maximization of (33) with respect to $\boldsymbol{\Lambda}$ is equivalent to the minimization of the Itakura-Saito (IS) divergence between $\{\tilde{x}_{ftm}^{\mathrm{d+e}}\}_{f,t,m=1}^{F,T,M}$ and $\{\tilde{y}_{ftm}^{\mathrm{d+e}}\}_{f,t,m=1}^{F,T,M}$.

a) Frequency-Invariant Source Model: In the same way as the NMF-based source model (explained later), we use a convergence-guaranteed minorization-maximization (MM) algorithm for deriving multiplicative update (MU) rules of $\boldsymbol{\gamma}$:

$$\gamma_{nt} \leftarrow \gamma_{nt}\sqrt{\frac{\sum_{f,m=1}^{F,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{\tilde{g}_{nlm}\tilde{x}_{f,t+l,m}^{\mathrm{d+e}}}{\left(\tilde{y}_{f,t+l,m}^{\mathrm{d+e}}\right)^2}}{\sum_{f,m=1}^{F,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{\tilde{g}_{nlm}}{\tilde{y}_{f,t+l,m}^{\mathrm{d+e}}}}}. \quad (36)$$

b) Time-Invariant Source Model: We can fix $\alpha_{nf} = 1$ to avoid the scale ambiguity between α_{nf} and $\mathbf{G}_{nfl}$.

c) NMF-Based Source Model: As in NMF based on the IS divergence [14], we use an MM algorithm for deriving MU rules of $\mathbf{W}$ and $\mathbf{H}$:

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m=1}^{T,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{h_{nk,t-l}\tilde{g}_{nlm}\tilde{x}_{ftm}^{\mathrm{d+e}}}{\left(\tilde{y}_{ftm}^{\mathrm{d+e}}\right)^2}}{\sum_{t,m=1}^{T,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{h_{nk,t-l}\tilde{g}_{nlm}}{\tilde{y}_{ftm}^{\mathrm{d+e}}}}}, \quad (37)$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m=1}^{F,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{w_{nkf}\tilde{g}_{nlm}\tilde{x}_{f,t+l,m}^{\mathrm{d+e}}}{\left(\tilde{y}_{f,t+l,m}^{\mathrm{d+e}}\right)^2}}{\sum_{f,m=1}^{F,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{w_{nkf}\tilde{g}_{nlm}}{\tilde{y}_{f,t+l,m}^{\mathrm{d+e}}}}}, \quad (38)$$

d) DNN-Based Source Model: We use an MM algorithm for deriving MU rules of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$\alpha_{nf} \leftarrow \alpha_{nf}\sqrt{\frac{\sum_{t,m=1}^{T,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{\beta_{n,t-l}\left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_{n,t-l})\right]_f\tilde{g}_{nlm}\tilde{x}_{ftm}^{\mathrm{d+e}}}{\left(\tilde{y}_{ftm}^{\mathrm{d+e}}\right)^2}}{\sum_{t,m=1}^{T,M}\sum_{l\in\mathbb{I}_{\mathrm{MA}}^+}\frac{\beta_{n,t-l}\left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_{n,t-l})\right]_f\tilde{g}_{nlm}}{\tilde{y}_{ftm}^{\mathrm{d+e}}}}}, \quad (39)$$

$$\beta_{nt} \leftarrow \beta_{nt} \sqrt{\frac{\sum_{f,m=1}^{F,M} \sum_{l \in \mathbb{I}_{\mathrm{MA}}^{+}} \frac{\alpha_{nf} \left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_{nt})\right]_{f} \tilde{g}_{nlm} \tilde{x}_{f,t+l,m}^{\mathrm{d+e}}}{\left(\tilde{y}_{f,t+l,m}^{\mathrm{d+e}}\right)^{2}}}{\sum_{f,m=1}^{F,M} \sum_{l \in \mathbb{I}_{\mathrm{MA}}^{+}} \frac{\alpha_{nf} \left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_{nt})\right]_{f} \tilde{g}_{nlm}}{\tilde{y}_{f,t+l,m}^{\mathrm{d+e}}}}}, \tag{40}$$

Since the latent variables $\mathbf{Z}$ are hard to optimize such that (33) is maximized, we use a stochastic gradient descent method based on backpropagation as in [12].

2) Updating $\tilde{\mathbf{G}}$: We also use an MM algorithm for deriving MU rules of $\mathbf{G}$:

$$\tilde{g}_{nlm} \leftarrow \tilde{g}_{nlm} \sqrt{\frac{\sum_{f,t=1}^{F,T} \frac{\lambda_{nf,t-l} \tilde{x}_{ftm}^{\mathrm{d+e}}}{\left(\tilde{y}_{ftm}^{\mathrm{d+e}}\right)^{2}}}{\sum_{f,t=1}^{F,T} \frac{\lambda_{nf,t-l}}{\tilde{y}_{ftm}^{\mathrm{d+e}}}}}. \tag{41}$$

3) Updating $\mathbf{Q}$ and $\mathbf{B}$: Instead of alternately updating $\mathbf{Q}$ and $\mathbf{B}$ as proposed in [23], we jointly update $\mathbf{Q}$ and $\mathbf{B}$ with IP or ISS for significantly better time-space complexity. IP and ISS were originally used for jointly updating demixing and dereverberation matrices (corresponding to $\mathbf{Q}$ and $\mathbf{B}$) in AR-ILRMA based on the rank-1 spatial model [28], [29].

We here aim to estimate an integrated demixing and dereverberation matrix $\mathbf{P}_f \triangleq [\mathbf{p}_{f1}, \ldots, \mathbf{p}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M \times M(L_{\mathrm{AR}}+1)}$ determined by $\mathbf{Q}$ and $\mathbf{B}$ as follows:

$$\mathbf{P}_f \triangleq [\mathbf{Q}_f, -\mathbf{Q}_f \mathbf{B}_{f,\Delta}, \ldots, -\mathbf{Q}_f \mathbf{B}_{f,\Delta+L_{\mathrm{AR}}-1}]. \tag{42}$$

Then, $\tilde{x}_{ftm}^{\mathrm{d+e}}$ involved in (33) can be written as follows:

$$\tilde{x}_{ftm}^{\mathrm{d+e}} = \left|\mathbf{p}_{fm}^{\mathsf{H}} \bar{\mathbf{x}}_{ft}\right|^{2} = \left|\mathbf{e}_{m}^{\mathsf{T}} \mathbf{P}_f \bar{\mathbf{x}}_{ft}\right|^{2}, \tag{43}$$

where $\mathbf{e}_m \in \{0,1\}^M$ is a one-hot vector whose m-th element is 1 and $\bar{\mathbf{x}}_{ft} \triangleq [\mathbf{x}_{ft}^{\mathsf{T}}, \mathbf{x}_{f,t-\Delta}^{\mathsf{T}}, \ldots, \mathbf{x}_{f,t-\Delta-L_{\mathrm{AR}}+1}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M(L_{\mathrm{AR}}+1)}$ is an aggregated observed vector.

a) Iterative Projection: Since (33) with (43) has the similar form as the likelihood function of IVA [30], as proposed in [28], $\{\mathbf{p}_{fm}\}_{f,m=1}^{F,M}$ can be updated one by one as follows:

$$\mathbf{c}_{fm} \triangleq \left(\left(\mathbf{Q}_f^{-1} \mathbf{e}_m\right)^{\mathsf{T}}, \mathbf{0}_{ML_{\mathrm{AR}}}^{\mathsf{T}}\right)^{\mathsf{T}}, \tag{44}$$

$$\boldsymbol{\Phi}_{fm} \triangleq \sum_{t=1}^{T} \frac{\bar{\mathbf{x}}_{ft} \bar{\mathbf{x}}_{ft}^{\mathsf{H}}}{\tilde{y}_{ftm}}, \tag{45}$$

$$\mathbf{p}_{fm} \leftarrow \boldsymbol{\Phi}_{fm}^{-1} \mathbf{c}_{fm} \left(\mathbf{c}_{fm}^{\mathsf{H}} \boldsymbol{\Phi}_{fm}^{-1} \mathbf{c}_{fm}\right)^{-\frac{1}{2}}, \tag{46}$$

where $\mathbf{0}_J$ denotes a zero vector of size J.

b) Iterative Source Steering: As an efficient alternative to IP, we can use either of ISS variants called ISS1 and ISS2, as proposed in [29]. Let $\bar{\mathbf{P}}_f \triangleq [\bar{\mathbf{p}}_{f1}, \ldots, \bar{\mathbf{p}}_{f,M(L_{\mathrm{AR}}+1)}]^{\mathsf{H}} \in \mathbb{C}^{M(L_{\mathrm{AR}}+1) \times M(L_{\mathrm{AR}}+1)}$ be a square matrix given by

$$\bar{\mathbf{P}}_f \triangleq \begin{pmatrix} \mathbf{P}_f \\ \mathbf{0}_{ML_{\mathrm{AR}},M} \quad \mathbf{I}_{ML_{\mathrm{AR}}} \end{pmatrix}, \tag{47}$$

where $\mathbf{0}_{I,J}$ denotes a zero matrix of size $I \times J$. Then, $\bar{\mathbf{P}}_f$ can be updated for each $m \in [1, M]$ as follows:

$$\bar{\mathbf{P}}_f \leftarrow \bar{\mathbf{P}}_f - \begin{pmatrix} \mathbf{c}_{fm} \\ \mathbf{0}_{ML_{\mathrm{AR}}} \end{pmatrix} \bar{\mathbf{p}}_{fm}^{\mathsf{H}}, \tag{48}$$

where $\mathbf{c}_{fm} \triangleq [c_{fm1}, \ldots, c_{fmM}] \in \mathbb{C}^M$ and $c_{fmm'}$ is calculated so that the log-likelihood is maximized as follows:

$$c_{fmm'} = \begin{cases} \frac{\bar{\mathbf{p}}_{fm'}^{\mathsf{H}} \boldsymbol{\Phi}_{fm'} \bar{\mathbf{p}}_{fm}}{\bar{\mathbf{p}}_{fm}^{\mathsf{H}} \boldsymbol{\Phi}_{fm'} \bar{\mathbf{p}}_{fm}} & (m \neq m'), \\ 1 - (\bar{\mathbf{p}}_{fm'}^{\mathsf{H}} \boldsymbol{\Phi}_{fm'} \bar{\mathbf{p}}_{fm'})^{-\frac{1}{2}} & (m = m'). \end{cases} \tag{49}$$

It is proven that updating the whole $\bar{\mathbf{P}}_f$ with (48) is equivalent to updating only the m-th column vector of $\bar{\mathbf{P}}_f^{-1}$ [27].

For each $m \in [M+1, M(L_{\mathrm{AR}}+1)]$, in ISS1, (48) is used to update the remaining M column vectors of $\bar{\mathbf{P}}_f^{-1}$. In ISS2, they are updated at once as follows:

$$\bar{\mathbf{P}}_f \leftarrow \bar{\mathbf{P}}_f - \begin{pmatrix} \mathbf{0}_{M \times M} \ \bar{\mathbf{C}}_f \\ \mathbf{0}_{ML_{\mathrm{AR}} \times M(L_{\mathrm{AR}}+1)} \end{pmatrix}, \tag{50}$$

where $\bar{\mathbf{C}}_f \triangleq [\bar{\mathbf{c}}_{f1}, \ldots, \bar{\mathbf{c}}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M \times ML_{\mathrm{AR}}}$ and

$$\bar{\mathbf{c}}_{fm}^{\mathsf{H}} = \left(\sum_{t=1}^{T} \frac{\tilde{x}_{ftm}^{\mathrm{d+e}} \check{\mathbf{x}}_{ft}^{\mathsf{H}}}{\tilde{y}_{ftm}}\right) \left(\sum_{t=1}^{T} \frac{\check{\mathbf{x}}_{ft} \check{\mathbf{x}}_{ft}^{\mathsf{H}}}{\tilde{y}_{ftm}}\right)^{-1}, \tag{51}$$

where $\check{\mathbf{x}}_{ft} \triangleq [\mathbf{x}_{f,t-\Delta}^{\mathsf{T}}, \ldots, \mathbf{x}_{f,t-\Delta-L_{\mathrm{AR}}+1}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML_{\mathrm{AR}}}$ is an aggregated observed vector.

4) Normalization: In each iteration, we adjust the scales of $\boldsymbol{\Lambda}$, $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\mathbf{B}$ without affecting (33) such that

$$\sum_{f=1}^{F} w_{nkf} = 1, \tag{52}$$

$$\sum_{f=1}^{F} \alpha_{nf} = 1, \tag{53}$$

$$\mathrm{tr}\left(\mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}}\right) = M, \tag{54}$$

$$\sum_{l \in \mathbb{I}_{\mathrm{MA}}^{+}} \sum_{m=1}^{M} \tilde{g}_{nlm} = 1. \tag{55}$$

C. Multichannel Wiener Filtering

Using (22) and (23), the late reverberation $\mathbf{x}_{ft}^{\mathrm{l}}$ is given by (18) and the direct sound $\mathbf{x}_{nft}^{\mathrm{d}}$ and early reflection $\mathbf{x}_{nft}^{\mathrm{e}}$ of each source n can be inferred as follows:

$$\mathbb{E}[\mathbf{x}_{nft}^{\mathrm{d}} \mid \mathbf{X}] = \mathbf{Q}_f^{-1} \mathrm{Diag}\left(\frac{\tilde{\mathbf{y}}_{nft}^{\mathrm{d}}}{\tilde{\mathbf{y}}_{ft}^{\mathrm{d+e}}}\right) \mathbf{Q}_f \left(\mathbf{x}_{ft} - \mathbf{x}_{ft}^{\mathrm{l}}\right), \tag{56}$$

$$\mathbb{E}[\mathbf{x}_{nft}^{\mathrm{e}} \mid \mathbf{X}] = \mathbf{Q}_f^{-1} \mathrm{Diag}\left(\frac{\tilde{\mathbf{y}}_{nft}^{\mathrm{e}}}{\tilde{\mathbf{y}}_{ft}^{\mathrm{d+e}}}\right) \mathbf{Q}_f \left(\mathbf{x}_{ft} - \mathbf{x}_{ft}^{\mathrm{l}}\right), \tag{57}$$

where $\frac{\mathbf{a}}{\mathbf{b}}$ denotes the element-wise division for vectors $\mathbf{a}$ and $\mathbf{b}$. For better intelligibility, $\sum_{l=0}^{L'} \mathbf{x}_{nftl}$ (the sum of $\mathbf{x}_{nft}^{\mathrm{d}}$ and an initial part of $\mathbf{x}_{nft}^{\mathrm{e}}$ up to L' frames) obtained by replacing $\tilde{\mathbf{y}}_{nft}^{\mathrm{d}}$ with $\sum_{l=0}^{L'} \tilde{\mathbf{y}}_{nftl}$ in (56) for some number $L' \leq L_{\mathrm{MA}}$ can be used instead of $\mathbf{x}_{nft}^{\mathrm{d}}$ as the final result.

D. Stabilized Optimization

We explain parameter initialization and updating techniques for ARMA-FastMNMF(-DP) for better performance. In this

paper we deal with speech enhancement/separation and dereverberation under a practically-important (over)determined situation with $N = N_{\text{speech}} + N_{\text{noise}} \leq M$, where N_{speech} and N_{noise} are the numbers of speech and noise sources, respectively. In particular, the spatial parameters $\mathbf{Q}$ and $\tilde{\mathbf{G}}$ should be initialized appropriately. Let $\tilde{\mathbf{G}}_l \triangleq [\tilde{\mathbf{g}}_{1^{\sim}l}, \ldots, \tilde{\mathbf{g}}_{Nl}]^{\mathsf{T}} \in \mathbb{R}_+^{N\times M}$.

1) Progressive Update: For ARMA-FastMNMF, we can use a modified version of a progressive update technique proposed for FastMNMF [13]. First, the NMF parameters $\mathbf{W}$ and $\mathbf{H}$ are initialized randomly and the spatial parameters $\mathbf{Q}$ and $\tilde{\mathbf{G}}$ and the reverberation parameters $\mathbf{B}$ are initialized as follows:

$$\mathbf{Q}_f \leftarrow \mathbf{I}_M, \tag{58}$$

$$\tilde{\mathbf{G}}_0 \leftarrow \begin{pmatrix} 1 & \epsilon & \ldots & \epsilon & 1 & \epsilon & \ldots \\ \epsilon & 1 & \ldots & \epsilon & \epsilon & 1 & \ldots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \epsilon & \epsilon & \ldots & 1 & \epsilon & \epsilon & \ldots \end{pmatrix}, \tag{59}$$

$$\tilde{\mathbf{G}}_l \leftarrow \epsilon \mathbf{1}_{N,M} \ (l \in \mathbb{I}_{\text{MA}}), \tag{60}$$

$$\mathbf{B}_{fl} \leftarrow \mathbf{0}_{M,M} \ (l \in \mathbb{I}_{\text{AR}}), \tag{61}$$

where ϵ is set to some small number ($\epsilon = 10^{-2}$ in this paper). Then, $\mathbf{Q}$, $\tilde{\mathbf{G}}_0$, and $\mathbf{B}$ are updated 50 times with AR-FastFIA (ARMA-FastFIA with $L_{\text{MA}} = 0$) as an initialization-insensitive method with a smaller DOF, where the source parameters $\boldsymbol{\gamma}$ specific to AR-FastFIA are randomly initialized and updated as well. The circular initializer given by (59) makes the SCMs given by (24) close to rank-1 or low-rank matrices and is expected to mitigate the initialization sensitivity as in IVA or (AR-)ILRMA based on the rank-1 spatial model [13]. Finally, $\mathbf{B}$ is initialized again with (61) because $\mathbf{B}$ estimated by AR-FastFIA is not accurate due to the severely limited expressive capability of the FI source model.

For ARMA-FastMNMF-DP, the NMF parameters $\mathbf{W}$ and $\mathbf{H}$ are initialized randomly and the scaling parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are initialized with all-one vectors. $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\mathbf{B}$ are initialized with (58)–(61) and updated 50 times with AR-FastMNMF-TI, a *blind* method with a smaller DOF, using the NMF-based model with $K = 2$ for N_{speech} sources and the TI source model for N_{noise} sources, where the NMF parameters specific to AR-FastMNMF-TI (different from those of ARMA-FastMNMF-DP) are initialized randomly and updated as well. Finally, $\mathbf{B}$ is initialized again with (61), and the latent variables $\mathbf{Z}$ are initially inferred by the encoder of a VAE given $\{\mathbf{x}_{nft}^{\text{d}}\}_{n=1}^{N_{\text{speech}}}$ estimated by AR-FastMNMF-TI.

2) Rank-Constrained Initialization: For ARMA-FastMNMF, we can use an extended version of a rank-constrained technique proposed for vanilla FastMNMF [13] with $L_{\text{AR}} = L_{\text{MA}} = 0$. The ranks of the SCMs $\{\mathbf{G}_{nfl}\}_{f=1}^F$ can be fixed by initializing a specified number of elements of $\tilde{\mathbf{g}}_{nl}$ to zero for each source n and delay l. Once $\tilde{g}_{nlm}$ is set to zero, it remains zero due to the multiplicative nature of (41). As proposed in [13], for example, if source n is directional speech with strong directivity, $\tilde{\mathbf{g}}_{n0}$ can be set to a one-hot vector such that $\{\mathbf{G}_{nf0}\}_{f=1}^F$ are rank-1 matrices as in ILRMA [6]. If source n is diffuse noise with weak directivity, in contrast, all elements of $\tilde{\mathbf{g}}_{n0}$ should be initialized to non-zero values such that $\{\mathbf{G}_{nf0}\}_{f=1}^F$ are full-rank matrices as in MNMF [9].

When the MA model is used ($L_{\text{MA}} > 0$), the rank constraint (RC) for $\{\mathbf{G}_{nfl}\}_{f=1,l\in\mathbb{I}_{\text{MA}}}^F$ plays another important role to avoid partially modeling the direct sound as the early reflection. To keep $\mathbf{G}_{nfl}$ ($l \in \mathbb{I}_{\text{MA}}$) far from $\mathbf{G}_{nf0}$, one can use (59) and (60) and then $\tilde{g}_{nln} \leftarrow 0$ ($l \in \mathbb{I}_{\text{MA}}$) such that $\{\mathbf{G}_{nfl}\}_{f=1,l\in\mathbb{I}_{\text{MA}}}^F$ become rank-$(M-1)$ matrices.

V. Evaluation

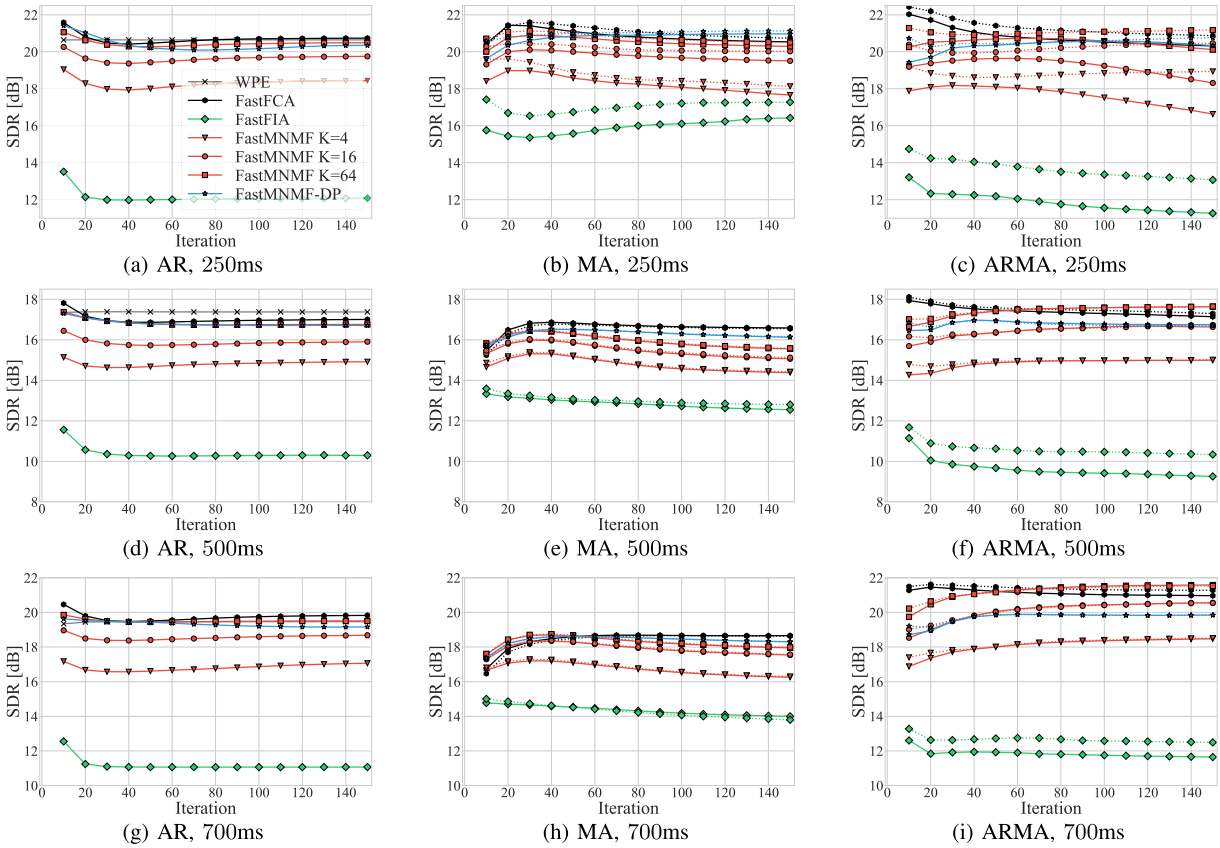
This section reports comparative experiments conducted for evaluating ARMA-FastMNMF(-DP). First, we compare various combinations of source and reverberation models in a speech dereverberation task. Second, we investigate the impacts of the hyperparameters L_{MA}, L_{AR}, and Δ, and the choice of IP or ISS in a speech separation and dereverberation task. Finally, we compare the proposed methods with the state-of-the-art methods under noisy reverberant conditions.

A. Configurations

1) Test Data: Multichannel *reverberant* speech signals used for evaluation were synthesized by convolving single-channel dry speech signals with room impulse responses (RIRs) taken from the development and evaluation subsets of the REVERB Challenge dataset [50]. The RIRs were measured with an eight-channel circular array with a diameter of 0.2 [m] under six conditions where the reverberation time RT_{60} was 250, 500, or 700 [ms] and the distance between the source and array was 0.5 (near) or 2.0 [m] (far). To compensate for the time gaps between the source and microphones, multichannel *dry* speech signals were also synthesized as ground-truth data by using non-echoic RIRs obtained by masking the original RIRs except at the peak and its previous and next two samples. Through all experiments, audio signals were sampled at 16 kHz and processed by STFT with a shifting interval of 256 samples and a Hann window of 1024 samples ($F = 513$).

2) Training Data: The DNN-based source model was obtained as the decoder of a VAE trained from clean dry speech signals taken from the training subset of the REVERB Challenge dataset (14.8 [h] in total). The configuration of the VAE was the same as that of [19] except for the dimensions of the observed and latent spaces ($F = 513$ and $D = 16$ in our work and $F = 2048$ and $D = 256$ in [19]) and the speaker conditioning [19] was not used. There was no overlap between the speakers of the training data and those of the test data. The volume of each utterance was perturbed with a uniformly distributed random number between 0.5 and 1.5.

3) Optimization: The number of iterations was set to 150 for achieving convergence (100 and 150 iterations for $M = 3$ and $M = 8$ were sufficient in our preliminary evaluation, respectively). $\boldsymbol{\Lambda}$ (the parameters of the source model), $\tilde{\mathbf{G}}$, and $\mathbf{P}$ (i.e., $\mathbf{Q}$ and $\mathbf{B}$) were updated in this order. For the NMF-based source model, $\mathbf{W}$ and $\mathbf{H}$ were updated in this order. For the DNN-based source model, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{Z}$ were updated in this order, and $\mathbf{Z}$ was updated five times per iteration by using the Adam optimizer [51] with a learning rate of 0.01.



(a) AR, 250ms (b) MA, 250ms (c) ARMA, 250ms

(d) AR, 500ms (e) MA, 500ms (f) ARMA, 500ms

(g) AR, 700ms (h) MA, 700ms (i) ARMA, 700ms

Fig. 2. The evolutions of average SDRs. The dotted lines indicate the rank-constrained versions.

4) Evaluation Measures: We used the signal-to-distortion ratio (SDR) [52], [53], the perceptual evaluation of speech quality (PESQ) [54], the frequency-weighted segmental SNR (FWSegSNR) [55], and the cepstrum distance (CD) [55] for evaluating the source separation and dereverberation performance. Larger SDR, PESQ, and FWSegSNR and smaller CD mean better performance.

B. Comparison of Source and Reverberation Models

Using reverberant speech signals, we investigated the combinations of the NMF-based, DNN-based, FI, and non-structured source models and the AR, MA, and ARMA reverberation models in a speech dereverberation task.

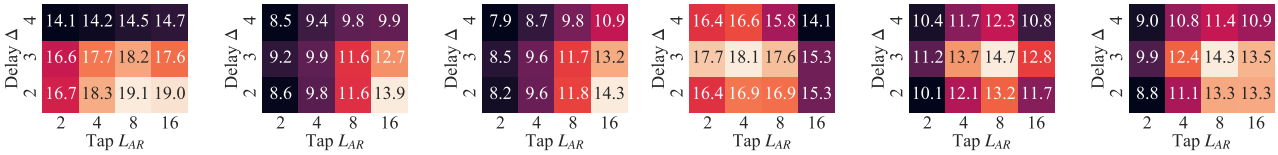
1) Experimental Conditions: Twenty reverberant speech signals were randomly selected from the development subset of the REVERB Challenge dataset [50] under each of the six conditions (120 signals in total). Each signal included only a single utterance without noise contamination. The hyperparameters were set as $N = 1$ ($N_{\text{speech}} = 1$ and $N_{\text{noise}} = 0$), $M = 8$, $L_{\text{MA}} = 8$, $L_{\text{AR}} = 4$, and $\Delta = 3$. We compared ARMA-FastMNMF(-DP) with $K \in \{4, 16, 64\}$ with its ablated versions, AR- and MA-FastMNMF(-DP), where FastMNMF-DP used only the DNN-based source model for a single speech source. We also tested AR-, MA-, and ARMA-FastFCA based on the non-structured source model and AR-, MA-, and ARMA-FastFIA based on the FI source model. As a reasonable baseline, we tested WPE [24], which can be interpreted as a special case of AR-FastFCA with the SCMs equal to identity matrices.

Given the PSDs $\{|x_{ftm}|^2\}_{f,t=1}^{F,T}$ of the observed reverberant speech, which were considered to be close to the PSDs $\{|x^{\text{d}}_{ftm}|^2\}_{f,t=1}^{F,T}$ of the dry speech, $\mathbf{W}$ and $\mathbf{H}$ of the NMF-based source model were initially estimated with NMF [14], $\mathbf{Z}$ of the DNN-based source model was initially inferred with the encoder of the VAE, $\boldsymbol{\lambda}$ of the non-structured model was initialized as $\lambda_{1\text{˜}ft} = \frac{1}{M}\sum_m |x_{ftm}|^2$, and $\boldsymbol{\gamma}$ of the FI source mode was initialized as $\gamma_{1t} = \frac{1}{FM}\sum_{fm} |x_{ftm}|^2$.

Regarding the spatial and reverberation models, $\mathbf{Q}$ and $\mathbf{B}$ were initialized with (58) and (61), respectively. $\tilde{\mathbf{G}}_0 \in \mathbb{R}^{1\times 8}$ was initialized with $[1, \epsilon, \ldots, \epsilon]$ and $\tilde{\mathbf{G}}_l$ ($l \in \mathbb{I}_{\text{MA}}$) was initialized with $[1, \ldots, 1]$ or with the RC technique $[0, 1, \ldots, 1]$. Under the simple experimental condition with $N = 1$, the progressive update technique (Section IV-D1) was not used. $\mathbf{B}$ was optimized with IP.

2) Experimental Results: Fig. 2 shows the evolutions of the average SDRs obtained by the AR, MA, and ARMA methods under $\text{RT}_{60} \in \{250, 500, 700\ [\text{ms}]\}$. The dotted lines indicate the MA and ARMA methods initialized by the RC technique.

As for the AR methods, AR-FastMNMF with $K = 64$, AR-FastMNMF-DP, and AR-FastFCA achieved almost the same performance as WPE, and AR-FastFIA performed worst in all cases. AR-FastMNMF with larger K performed better. When $N = 1$, a source model capable of precisely representing the speech PSDs $\boldsymbol{\lambda}$ is crucial for precisely estimating the AR



(a) $M = 3$, 250ms (b) $M = 3$, 500ms (c) $M = 3$, 700ms (d) $M = 8$, 250ms (e) $M = 8$, 500ms (f) $M = 8$, 700ms

Fig. 3. The average SDRs of AR-FastMNMF with ISS2.

coefficients $\mathbf{B}$. The NMF-based source model with large K and the DNN-based source model have sufficiently large DOFs, whereas the FI source model does not, resulting in the significant performance difference in speech dereverberation. We found that the performance as well as the average power of the dereverberated speech slightly decreased during the initial several tens of iterations. This indicates that the AR model excessively removed the reverberation, i.e., the direct sound was partially regarded as the reverberation.

As for the MA methods, a source model with a larger DOF was crucial for better performance as in the AR methods. When RT_{60} was 250 [ms], the MA methods were comparable with their AR counterparts. When RT_{60} was 500 or 700 [ms], in contrast, the MA methods slightly underperformed their AR counterparts. The ARMA methods outperformed their AR and MA counterparts by a larger margin under a longer RT_{60}. When RT_{60} was 250 [ms], the performance of the methods without the RC technique decreased as the number of iterations increased, especially when the source model had small degree of freedom. The roughly estimated PSDs lead to the overestimation of the reverberations. The RC technique alleviated this problem. When RT_{60} was 500 or 700 [ms], the RC technique yielded no significant performance difference. For ARMA-FastFCA, its performance and the average power of the dereverberated speech decreased continuously, probably because of the too high DOF of the source model. These results indicate the importance of the source model for dereverberation.

C. Investigation of Hyperparameters and Optimizers

Using reverberant speech mixtures, we investigate the performance of ARMA-FastMNMF according to the hyperparameters L_{MA}, L_{AR}, and Δ, and the optimizer (IP, ISS1, or ISS2) in a speech separation and dereverberation task.

1) Experimental Conditions: Twenty reverberant speech mixtures were made from the REVERB Challenge dataset [50] under each of the six conditions (120 signals in total). Each mixture was obtained by superimposing a reverberant speech signal randomly taken from the development subset and another one from the evaluation subset.

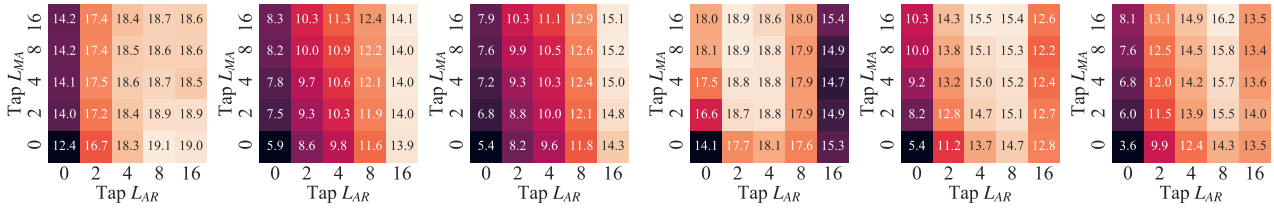
We tested ARMA-FastMNMF with $N = 2$, $M \in \{3, 8\}$, and $K = 16$. We investigated the impact of the delay parameter $\Delta \in \{2, 3, 4\}$ using $L_{\mathrm{MA}} = 0$ and $L_{\mathrm{AR}} \in \{2, 4, 8, 16\}$. We then investigated the combination of $L_{\mathrm{MA}} \in \{0, 2, 4, 8, 16\}$ and $L_{\mathrm{AR}} \in \{0, 2, 4, 8, 16\}$ using $\Delta = 2$ for $M = 3$ or $\Delta = 3$ for $M = 8$. ARMA-FastMNMF reduces to vanilla FastMNMF [13] when $L_{\mathrm{MA}} = 0$ and $L_{\mathrm{AR}} = 0$, MA-FastMNMF when $L_{\mathrm{MA}} \neq 0$ and $L_{\mathrm{AR}} = 0$, and AR-FastMNMF [23] when $L_{\mathrm{MA}} = 0$ and $L_{\mathrm{AR}} \neq 0$. The parameters were initialized with the progressive update technique (Section IV-D1). The integrated spatial and AR coefficients $\mathbf{P}$ were updated with IP, ISS1, or ISS2. In addition to the SDR, we measured the elapsed time per iteration for processing a mixture signal of 9.2 [s] (average length) on NVIDIA GeForce Titan RTX. We used the dependent one-sided t-test for statistical significance assessment.

TABLE II
THE p VALUES OBTAINED BY THE DEPENDENT ONE-SIDED t-TESTS FOR THE SDRS OF 40 SAMPLES OBTAINED BY IP, ISS1, AND ISS2 WITH THE BEST CONFIGURATIONS

	$M = 3$			$M = 8$		
RT_{60}	250	500	700	250	500	700
H_0: IP = ISS1, H_1: IP > ISS1	0.058	0.020	0.000	0.042	0.140	0.158
H_0: ISS2=ISS1, H_1: ISS2>ISS1	0.000	0.020	0.000	0.008	0.001	0.000
H_0: IP = ISS2, H_1: IP > ISS2	0.630	0.225	0.038	0.260	0.364	0.573

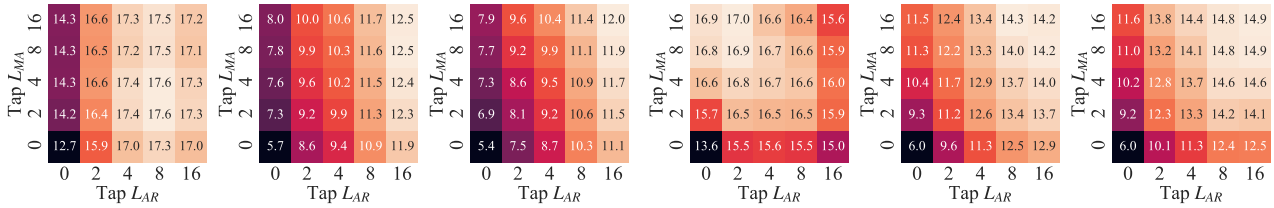
2) Experimental Results: Fig. 3 shows the average SDRs obtained by AR-FastMNMF with ISS2 with respect to $M \in \{3, 8\}$ and $\mathrm{RT}_{60} \in \{250, 500, 700\ [\mathrm{ms}]\}$. The best configuration of the delay parameter was $\Delta = 2$ for $M = 3$ or $\Delta = 3$ for $M = 8$, regardless of RT_{60}. When $\Delta = 2$ for $M = 8$, the direct sound was partially represented by the AR model, resulting in the excessive dereverberation. We thus decided to use the same configuration for ARMA-FastMNMF.

Figs. 4, 5, and 6 show the average SDRs obtained by ARMA-FastMNMF with IP, ISS1, and ISS2 with respect to $M \in \{3, 8\}$ and $\mathrm{RT}_{60} \in \{250, 500, 700\ [\mathrm{ms}]\}$. Note that ISS1 is equivalent to ISS2 when $L_{\mathrm{AR}} = 0$. MA-FastMNMF always outperformed FastMNMF in any condition (see the most-left column of each table). When $M = 8$, ARMA-FastMNMF outperformed AR-FastMNMF by a large margin. For IP, when $M = 8$ and $\mathrm{RT}_{60} = 250$ [ms], MA-FastMNMF with $L_{\mathrm{MA}} = 8$ outperformed vanilla FastMNMF ($p < 0.001$), and ARMA-FastMNMF with $L_{\mathrm{MA}} = 8$ and $L_{\mathrm{AR}} = 2$ outperformed AR-FastMNMF with $L_{\mathrm{AR}} = 4$ ($p = 0.022$). This indicates the effectiveness of using both the source-independent AR and source-dependent MA models for precise reverberation modeling. When $M = 3$, in contrast, the performance gain of ARMA-FastMNMF from AR-FastMNMF was smaller. Since each of the $N(L_{\mathrm{MA}} + 1)$ SCMs of the direct sounds and early reflections was represented as a weighted sum of common M rank-1 SCMs in (24), the effectiveness of the MA model was limited by the severely restricted representation capability. Comparing IP, ISS1, and ISS2 used for AR- and ARMA-FastMNMF, where the best configuration of L_{AR} and L_{MA} was used for each method, the SDRs and likelihoods of IP and ISS2 were higher than those of ISS1. Table II shows the p values in comparison of IP, ISS1, and ISS2. This indicates that



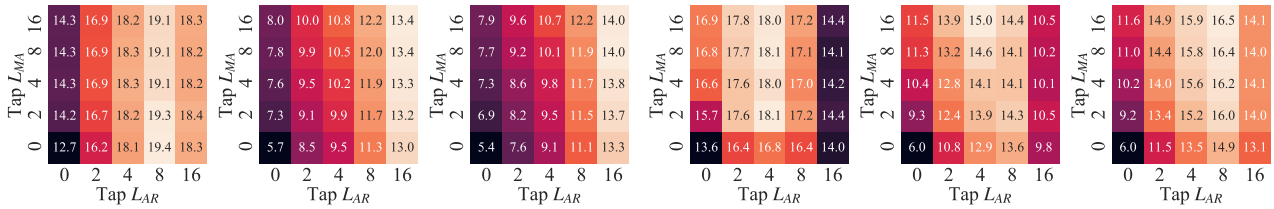
(a) $M = 3$, 250ms (b) $M = 3$, 500ms (c) $M = 3$, 700ms (d) $M = 8$, 250ms (e) $M = 8$, 500ms (f) $M = 8$, 700ms

Fig. 4. The average SDRs of ARMA-FastMNMF with IP.



(a) $M = 3$, 250ms (b) $M = 3$, 500ms (c) $M = 3$, 700ms (d) $M = 8$, 250ms (e) $M = 8$, 500ms (f) $M = 8$, 700ms

Fig. 5. The average SDRs of ARMA-FastMNMF with ISS1.



(a) $M = 3$, 250ms (b) $M = 3$, 500ms (c) $M = 3$, 700ms (d) $M = 8$, 250ms (e) $M = 8$, 500ms (f) $M = 8$, 700ms

Fig. 6. The average SDRs of ARMA-FastMNMF with ISS2.

AR- and ARMA-FastMNMF with ISS1 are more likely to get stuck at bad local optima. This was because only M elements of $\bar{\mathbf{P}}_f$ are updated with (48) for each m ($> M$) in ISS1, while $M(L_{\mathrm{AR}}+1)$ and ML_{AR} elements are updated at once with IP and ISS2, respectively. We also found that there were strong correlations between the SDRs of ISS1 and ISS2, resulting in the small p values.

Table III shows the elapsed times of ARMA-FastMNMF with respect to the optimizer. ISS1 was significantly faster than the others, especially for $M = 8$ and a larger L_{AR}, because matrices of size $M(L_{\mathrm{AR}}+1)$ and ML_{AR} should be inverted in IP and ISS2, respectively. Considering the performances and computational costs, we decided to use $L_{\mathrm{MA}} = 8$ and $L_{\mathrm{AR}} = 4$, and IP or ISS2 in the following experiment.

TABLE III
THE ELAPSED TIME [ms] PER ITERATION OF ARMA-FASTMNMF FOR PROCESSING A 9.2 SECONDS SIGNAL ON GPU

(a) $M = 3$, IP

$L_{\mathrm{MA}} \backslash L_{\mathrm{AR}}$	0	2	4	8	16
16	52	55	57	64	120
8	39	41	43	50	110
4	32	34	36	43	100
2	28	31	33	39	96
0	25	27	29	36	91

(b) $M = 8$, IP

$L_{\mathrm{MA}} \backslash L_{\mathrm{AR}}$	0	2	4	8	16
16	160	180	320	590	1600
8	110	140	270	550	1500
4	94	120	250	520	1500
2	83	100	240	510	1500
0	73	94	230	500	1500

(c) $M = 3$, ISS1

$L_{\mathrm{MA}} \backslash L_{\mathrm{AR}}$	0	2	4	8	16
16	36	41	45	54	75
8	23	27	31	40	61
4	16	21	25	33	54
2	13	17	21	30	51
0	9.3	14	18	27	48

(d) $M = 8$, ISS1

$L_{\mathrm{MA}} \backslash L_{\mathrm{AR}}$	0	2	4	8	16
16	110	130	160	210	330
8	71	94	120	170	290
4	51	74	98	150	270
2	41	63	88	140	260
0	30	53	78	130	250

(e) $M = 3$, ISS2

$L_{\mathrm{MA}} \backslash L_{\mathrm{AR}}$	0	2	4	8	16
16	36	65	67	73	130
8	23	51	53	59	110
4	16	44	46	52	110
2	13	40	42	48	100
0	9.3	36	38	44	97

(f) $M = 8$, ISS2

$L_{\mathrm{MA}} \backslash L_{\mathrm{AR}}$	0	2	4	8	16
16	110	190	210	390	1200
8	71	150	170	350	1100
4	51	130	150	330	1100
2	41	120	140	320	1100
0	30	100	130	300	1100

D. Comparison in Noisy Environments

Using noisy reverberant speech mixtures, we compared ARMA-FastMNMF(-DP) with the state-of-the-art methods in a speech separation and dereverberation task.

1) Experimental Conditions: 100 noisy reverberant speech mixtures were made from the REVERB Challenge dataset [50] under each of the six conditions (600 signals in total). Each

TABLE IV
THE SEPARATION AND DEREVERBERATION PERFORMANCES OF THE CONVENTIONAL AND PROPOSED METHODS FOR $M = 3$

Method	SDR (Observation = −4.1 dB)							PESQ (Observation = 1.1)							FWSegSNR (Observation = −0.2 dB)							CD (Observation = 7.6)						
	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS
IVA	1.2	-	**3.0**	2.9	-	-	-	1.28	-	1.35	**1.39**	-	-	-	0.74	-	**0.82**	0.77	-	-	-	6.95	-	6.81	**6.80**	-	-	-
ILRMA K=4	1.1	-	3.0	**3.3**	-	-	-	1.28	-	1.33	**1.37**	-	-	-	0.76	-	0.85	**0.90**	-	-	-	6.99	-	6.85	**6.81**	-	-	-
ILRMA K=16	0.9	-	2.8	**3.2**	-	-	-	1.28	-	1.32	**1.37**	-	-	-	0.77	-	0.85	**0.92**	-	-	-	7.01	-	6.87	**6.83**	-	-	-
ILRMA K=64	0.8	-	2.7	**3.1**	-	-	-	1.27	-	1.34	**1.36**	-	-	-	0.76	-	0.84	**0.92**	-	-	-	7.01	-	6.88	**6.84**	-	-	-
ILRMA-DP	1.0	-	2.8	**3.2**	-	-	-	1.27	-	1.37	**1.38**	-	-	-	0.80	-	0.88	**0.94**	-	-	-	7.04	-	6.91	**6.87**	-	-	-
FastFIA	1.3	1.8	3.1	2.9	3.5	3.5	**3.5**	1.27	1.29	1.32	**1.36**	1.32	1.34	1.34	0.69	0.71	0.76	0.70	**0.79**	0.74	0.75	6.97	6.96	6.85	6.84	6.83	6.81	**6.80**
FastMNMF K=4	3.6	5.0	5.1	5.5	6.0	**6.5**	6.4	1.37	1.37	1.39	1.42	1.39	**1.43**	1.43	0.90	1.33	1.07	1.14	1.40	1.56	**1.57**	6.81	6.84	6.67	6.60	6.70	**6.59**	6.60
FastMNMF K=16	2.5	4.0	4.1	4.6	5.1	**5.5**	5.5	1.32	1.36	1.39	**1.43**	1.39	1.42	1.42	0.79	1.10	0.86	0.91	1.13	1.20	**1.22**	6.91	6.74	6.80	6.76	6.67	6.63	**6.62**
FastMNMF K=64	1.4	2.8	3.2	3.7	4.0	**4.5**	4.5	1.29	1.34	1.36	1.39	1.38	**1.40**	1.40	0.73	0.79	0.78	0.84	0.81	**0.91**	0.88	6.99	6.87	6.88	6.84	6.81	**6.77**	6.78
FastMNMF-TI	2.7	4.0	4.2	4.6	5.0	4.7	**5.4**	1.32	1.35	1.36	1.40	1.37	1.40	**1.43**	0.91	1.14	0.92	0.95	1.11	0.99	**1.18**	6.98	6.89	6.88	6.86	6.83	6.92	**6.79**
FastMNMF-DP	4.3	5.3	5.9	6.5	6.4	6.1	**6.8**	1.35	1.38	1.42	**1.46**	1.43	1.43	1.46	1.76	1.94	1.97	2.18	2.02	2.00	**2.20**	7.20	7.26	7.04	**6.94**	7.12	7.15	7.01

TABLE V
THE SEPARATION AND DEREVERBERATION PERFORMANCES OF THE CONVENTIONAL AND PROPOSED METHODS FOR $M = 8$

Method	SDR (Observation = −4.1 dB)							PESQ (Observation = 1.1)							FWSegSNR (Observation = −0.2 dB)							CD (Observation = 7.6)						
	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS	- ISS	MA ISS	WPE ISS	AR ISS	WPE +MA ISS	AR MA IP	AR MA ISS
IVA	6.3	-	**8.1**	7.7	-	-	-	1.65	-	**1.71**	1.66	-	-	-	1.86	-	**1.96**	1.78	-	-	-	6.26	-	**6.13**	6.21	-	-	-
ILRMA K=4	6.6	-	8.4	**9.1**	-	-	-	1.63	-	1.72	**1.73**	-	-	-	2.01	-	2.11	**2.24**	-	-	-	6.25	-	6.12	**6.05**	-	-	-
ILRMA K=16	6.1	-	8.2	**9.0**	-	-	-	1.62	-	1.72	**1.73**	-	-	-	2.01	-	2.12	**2.30**	-	-	-	6.28	-	6.14	**6.05**	-	-	-
ILRMA K=64	6.0	-	8.0	**8.9**	-	-	-	1.59	-	1.70	**1.73**	-	-	-	1.99	-	2.11	**2.30**	-	-	-	6.30	-	6.15	**6.07**	-	-	-
ILRMA-DP	5.7	-	7.7	**8.6**	-	-	-	1.63	-	1.70	**1.72**	-	-	-	2.03	-	2.15	**2.38**	-	-	-	6.40	-	6.27	**6.19**	-	-	-
FastFIA	6.1	7.0	7.9	7.8	**8.7**	8.5	8.4	1.60	1.62	1.66	1.63	1.69	**1.74**	1.71	1.69	1.82	1.81	1.71	**1.90**	1.87	1.81	6.33	6.26	6.21	6.22	6.15	**6.14**	6.17
FastMNMF K=4	9.1	9.9	10.9	11.2	11.3	**11.6**	11.5	1.69	1.73	1.78	1.79	1.82	1.85	**1.85**	2.35	2.32	2.53	2.64	2.46	**2.69**	2.63	6.04	6.39	5.90	**5.87**	6.18	6.02	6.08
FastMNMF K=16	8.3	9.5	10.2	10.3	10.9	**11.2**	11.1	1.68	1.73	1.75	1.77	1.79	**1.88**	1.82	2.11	2.52	2.25	2.30	2.53	**2.66**	2.65	6.09	6.06	5.96	5.99	5.96	**5.88**	5.91
FastMNMF K=64	7.3	8.6	9.2	9.4	10.0	**10.2**	10.1	1.65	1.67	1.71	1.72	1.74	**1.77**	1.75	1.88	2.23	1.96	2.12	2.14	**2.30**	2.25	6.26	6.07	6.13	6.12	6.04	**5.99**	6.03
FastMNMF-TI	6.1	8.3	9.1	9.8	9.9	10.3	**10.4**	1.59	1.77	1.64	1.67	1.77	1.80	**1.82**	2.12	2.56	2.30	2.49	2.58	2.69	**2.74**	6.47	6.24	6.19	6.13	6.10	6.06	**6.04**
FastMNMF-DP	7.6	9.6	10.5	11.0	10.9	11.3	**11.4**	1.70	1.73	1.79	1.80	1.80	**1.84**	1.83	2.43	2.83	2.70	2.98	2.80	3.06	**3.08**	6.75	6.76	6.58	**6.49**	6.65	6.51	6.50

mixture was obtained by superimposing a reverberant speech signal randomly taken from the development subset, another one from the evaluation subset, and a real diffuse noise signal (mainly caused by air conditioners) from the development or evaluation subset. The SNR of the clean speech mixture was set to 0 dB.

We tested ARMA-FastMNMF(-DP/TI) and ARMA-FastFIA with $M \in \{3, 8\}$, $L_{\mathrm{MA}} \in \{0, 8\}$, and $L_{\mathrm{AR}} \in \{0, 4\}$. The delay parameter was set as $\Delta = 2$ for $M = 3$ and $\Delta = 3$ for $M = 8$ and the number of bases was set as $K \in \{4, 16, 64\}$ for ARMA-FastMNMF and $K = 16$ for ARMA-FastMNMF-DP/TI. ISS2 was used for updating the integrated coefficients $\mathbf{P}$. IP was also tested only when $L_{\mathrm{MA}} = 8$ and $L_{\mathrm{AR}} = 4$. For comparison, we tested AR-IVA and AR-ILRMA(-DP) with the same configuration except for L_{MA}. In addition to these *joint* separation and dereverberation methods, we tested *sequential* methods that perform dereverberation with WPE and separation (and further dereverberation) with the MA methods ($L_{\mathrm{MA}} \in \{0, 8\}$ and $L_{\mathrm{AR}} = 0$) in this order. Since the rank-1 spatial model is applicable to only a determined condition with $N = M$ and an unknown number of noise sources were included, we set $N = 3$ ($N_{\mathrm{speech}} = 2$ and $N_{\mathrm{noise}} = 1$) for $M = 3$ and $N = 8$ ($N_{\mathrm{speech}} = 2$ and $N_{\mathrm{noise}} = 6$) for $M = 8$, where two sources were selected out of N sources such that the performance was maximized in terms of each measure.

2) Experimental Results: We first validate the effectiveness of the combination of the AR and MA reverberation models. Tables IV and V show the SDRs, PESQs, FWSegSNRs, and CDs averaged over all conditions. In most cases, ARMA-FastMNMF outperformed AR-FastMNMF in terms of all measures. However, ARMA-FastMNMF attained only a marginal gain (0.4 dB when $K = 4$) over AR-FastMNMF for $M = 8$, in which $N = 8$ sources were estimated. These extra sources were exploited by AR-FastMNMF to represent the early reflection and the residual late reverberation that was not represented with the AR model. In ARMA-FastMNMF, these reflection and reverberation were represented by the MA model. Therefore, the dereverberation performance of ARMA-FastMNMF and AR-FastMNMF were not so different. Nonetheless, ARMA-FastMNMF is still considered to be advantageous in estimating the actual number of speech sources from the separated signals under a noisy reverberant condition thanks to the little leakage of speech components to noise components. Note that ARMA-FastMNMF has a clear performance advantage under a noise-free condition (Section V-C).

We then compare the source models. ARMA-FastMNMF with $K = 4$ and ARMA-FastMNMF-DP performed best in terms of the SDR. The NMF-based source model with a larger K worked worse for speech separation and dereverberation under a noisy condition because it fit not only the PSDs of clean isolated

speech but also those of any noisy sound mixtures. Note that it worked better for speech dereverberation under a noise-free condition (Section V-B). In contrast, the DNN-based source model, which also had rich expression capability, was trained to represent only the PSDs of clean isolated speech. In fact, however, the performance improvement was small because the latent variables $\mathbf{Z}$ of the test data were hard to update gradually towards the global optimum with the gradient descent algorithm (backpropagation) through unseen areas that were not covered by the training data. To make the latent space widely covered by the training data, adversarially learned inference (ALI) [56] would help.

We finally validate the effectiveness of the joint separation and dereverberation approach. As for the NMF-based source model, the joint method, AR(MA)-FastMNMF, attained a marginal gain over its sequential counterpart, WPE+(MA-)FastMNMF, in terms of the SDR. One reason is that although the PSDs of each direct sound were estimated individually, only the sum of those PSDs over all the sources made an effect on the estimation of the AR coefficients $\mathbf{B}$ with (46) or (48). As for the FI source model, in contrast, the joint method, AR(MA)-FastFIA, underperformed its sequential counterpart, WPE+(MA-)FastFIA because the expression capability of the source model was not enough to precisely estimate $\mathbf{B}$, as discussed in Section V-B.

VI. Conclusion

This paper presented a computationally-efficient joint source separation and dereverberation framework based on a unified probabilistic model consisting of the non-structured, FI, TI, NMF-based, and/or DNN-based source model(s), the JD full-rank spatial model, and the source-independent AR and/or source-dependent MA reverberation model(s). We derived three optimization methods called IP, ISS1, and ISS2 for jointly updating the diagonalizers used for separation and the AR coefficients used for dereverberation.

In our comparative experiments, we comprehensively validated MA-, AR-, and ARMA-FastMNMF(-DP) with IP, ISS1, and ISS2 in terms of the computational cost and the separation and dereverberation performance. We revealed the mutual benefit of the AR and MA models used for representing the late reverberation and the early reflection, respectively. We found that a source model with a higher expression capability is crucial in speech dereverberation, whereas such a rich source model is not necessarily effective for joint source separation and dereverberation. We also showed the superiority of the proposed joint method, ARMA-FastMNMF(-DP), over the sequential counterpart and the conventional (semi-)blind methods based on the rank-1 spatial model.

In the future, we will extend both the spatial and reverberation models to deal with time-varying acoustic environments (e.g., moving sources and microphones). One promising way is to use normalizing flows (NFs) for representing the diagonalizers and AR coefficient matrices in a time-varying manner as proposed for a determined BSS method called NF-IVA [57] with time-varying demixing matrices.

References

[1] B. Li *et al.*, "Acoustic modeling for google home," in *Proc. Interspeech*, 2017, pp. 399–403.

[2] R. Haeb-Umbach *et al.*, "Speech processing for digital home assistants," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124, Nov. 2019.

[3] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[4] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Int. Conf. Independent Compon. Anal. Blind Signal Separation*, 2006, pp. 165–172.

[5] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. Int. Conf. Independent Compon. Anal. Blind Signal Separation*, 2006, pp. 601–608.

[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 Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

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

[8] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. Int. Symp. Signal Process. Appl.*, 2010, pp. 1–4.

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

[10] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[11] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 371–375.

[12] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[13] 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 Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2610–2625, 2020.

[14] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[15] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 716–720.

[16] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 1233–1239.

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

[18] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 101–105.

[19] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neuro Comput.*, vol. 31, no. 9, pp. 1–24, 2019.

[20] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[21] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 31–35.

[22] S. Inoue, H. Kameoka, L. Li, S. Seki, and S. Makino, "Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 96–100.

[23] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 511–515.
[24] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.
[25] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.
[26] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian source," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2010, pp. 165–172.
[27] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 236–240.
[28] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "A unifying framework for blind source separation based on a joint diagonalizability constraint," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.
[29] T. Nakashima, R. Scheibler, M. Togami, and N. Ono, "Joint dereverberation and separation with iterative source steering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 216–220.
[30] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.
[31] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
[32] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1369–1380, Jul. 2013.
[33] M. Togami, "Multi-channel speech source separation and dereverberation with sequential integration of determined and underdetermined models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 231–235.
[34] N. Ito, S. Araki, and T. Nakatani, "FastFCA: A joint diagonalization based fast algorithm for audio source separation using a full-rank spatial covariance model," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 1667–1671.
[35] R. Ikeshita and T. Nakatani, "Independent vector extraction for fast joint blind source separation and dereverberation," *IEEE Signal Process. Lett.*, vol. 28, pp. 972–976, 2021.
[36] M. Togami and R. Scheibler, "Over-determined speech source separation and dereverberation," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 705–710.
[37] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 196–200.
[38] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-Channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.
[39] L. Drude *et al.*, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *Proc. Interspeech*, 2018, pp. 3043–3047.
[40] 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," 2019, *arXiv:1904.09049*.
[41] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2267–2282, 2020.
[42] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ, USA: Wiley, 2004.
[43] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 185–189.
[44] R. Ikeshita, T. Nakatani, and S. Araki, "Overdetermined independent vector analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 591–595.
[45] 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. Int. Workshop Acoust. Signal Enhance.*, 2018, pp. 151–155.
[46] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.
[47] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust. Signal Enhance.*, 2003, pp. 99–102.
[48] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.
[49] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Proc. Interspeech*, 2017, pp. 384–388.
[50] K. Kinoshita *et al.*, "The REVERB challenge : A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.
[51] D. P. Kingma and J. Lei Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[52] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[53] C. Raffel *et al.*, "MIR_EVAL: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 367–372.
[54] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.
[55] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[56] V. Dumoulin *et al.*, "Adversarially learned inference," in *Proc. Int. Conf. Learn. Representations*, 2017.
[57] A. A. Nugraha, K. Sekiguchi, M. Fontaine, Y. Bando, and K. Yoshii, "Flow-based independent vector analysis for blind source separation," *IEEE Signal Process. Lett.*, vol. 27, pp. 2173–2177, 2020.

Kouhei Sekiguchi (Member, IEEE) received the B.E. degree, in 2015, and the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2017 and 2021, respectively. He is currently a Postdoctoral Researcher with the Sound Scene Understanding Team, Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include microphone array signal processing and machine learning. He is also a Member of IPSJ.

Yoshiaki Bando (Member, IEEE) received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2015 and 2018, respectively. He is currently a Senior Researcher with the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. He is also a Visiting Researcher with Advanced Intelligence Project, RIKEN, Tokyo, Japan. His research interests include microphone array signal processing, deep Bayesian learning, and robot audition. He is a Member of RSJ and IPSJ.

Aditya Arie Nugraha (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Institut Teknologi Bandung, Bandung, Indonesia, in 2008 and 2011, respectively, the M.E. degree in computer science and engineering from the Toyohashi University of Technology, Toyohashi, Japan, in 2013, and the Ph.D. degree in informatics from the Université de Lorraine, Nancy, France, and Inria Nancy–Grand-Est, France, in 2017. He is currently a Research Scientist of the Sound Scene Understanding Team with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include audio-visual signal processing and machine learning.

Mathieu Fontaine (Member, IEEE) received the M.S. degree in applied & fundamentals mathematics from the Université de Poitiers, Poitiers, France, in 2015, and the Ph.D. degree in informatics from the Université de Lorraine and Inria Nancy Grand-Est, France, in 2018. He was a Postdoctoral Researcher with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. He is currently an Assistant Professor with LTCI, Télécom Paris, Palaiseau, France. He is also a Visiting Researcher with the Advanced Intelligence Project (AIP), RIKEN. His research interests include machine listening topics, such as audio source separation, sound event detection, and speaker diarization using microphone array.

Kazuyoshi Yoshii (Member, IEEE) received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2005 and 2008, respectively. He is currently an Associate Professor with the Graduate School of Informatics, Kyoto University, and concurrently the Leader of the Sound Scene Understanding Team, Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include music informatics, audio signal processing, and statistical machine learning.

Tatsuya Kawahara (Fellow, IEEE) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively. From 1995 to 1996, he was a Visiting Researcher with Bell Laboratories, Murray Hill, NJ, USA. He is currently a Professor and the Dean of the School of Informatics, Kyoto University. He was also an Invited Researcher with ATR and NICT. He has authored or coauthored more than 450 academic papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several projects including speech recognition software Julius, automatic transcription system deployed in the Japanese Parliament (Diet), and autonomous android ERICA.

He received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a Member of IEEE SPS Speech Technical Committee. He was the General Chair of IEEE ASRU 2007. He was also the Tutorial Chair of INTERSPEECH 2010, Local Arrangement Chair of ICASSP 2012, and General Chair of APSIPA ASC 2020. He was the Editorial Board Member of the *Elsevier Journal of Computer Speech and Language* and IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. From 2018 to 2021, he was the Editor-in-Chief of *APSIPA Transactions on Signal and Information Processing*. He is a Board Member of APSIPA and ISCA.