Minimum-Volume Multichannel Nonnegative Matrix Factorization for Blind Audio Source Separation

Jianyu Wang, Shanzheng Guan, *Graduate Student Member, IEEE*, Shupei Liu, and Xiao-Lei Zhang, *Senior Member, IEEE*

***Abstract*—Multichannel blind audio source separation aims to recover the latent sources from their multichannel mixtures without supervised information. One state-of-the-art blind audio source separation method, named independent low-rank matrix analysis (ILRMA), unifies independent vector analysis (IVA) and nonnegative matrix factorization (NMF). However, the spectra matrix produced from NMF may not find a compact spectral basis. It may not guarantee the identifiability of each source as well. To address this problem, here we propose to enhance the identifiability of the source model by a minimum-volume prior distribution. We further regularize a multichannel NMF (MNMF) and ILRMA respectively with the minimum-volume regularizer. The proposed methods maximize the posterior distribution of the separated sources, which ensures the stability of the convergence. Experimental results demonstrate the effectiveness of the proposed methods compared with auxiliary independent vector analysis, MNMF, ILRMA and its extensions. The source code is available at https://github.com/alexwang9654/m-ILRMA.**

***Index Terms*—Blind source separation, multichannel nonnegative matrix factorization, independent low-rank matrix analysis.**

I. Introduction

BLIND source separation (BSS) is a technique of separating source components from a given multichannel mixture without any knowledge about the mixing system or microphone positions. Most BSS methods aim to cluster the time-frequency units of the spectrogram of the mixture into different sources. A promising approach of multichannel BSS to achieve the above goal is to represent the hierarchical generative process of the time-frequency spectrogram of the mixture by a source model and a spatial model, where the source model represents the generative process of source spectrograms, and the spatial model represents the mixing process of the sources.

This paper focuses on nonnegative matrix factorization (NMF) based multichannel BSS [1]–[4]. It usually decomposes the spectrogram of a mixture into several spectral bases and temporal activations. Existing NMF-based BSS methods usually have the following major problems. First, because the NMF decomposition is an NP-hard problem [5], it is difficult to obtain a meaningful representation of the spectral bases. Second, they may not guarantee that the spectral structure of each source is identifiable. For example, a matrix $\mathbf{V}$ may be decomposed by $\mathbf{V} = \mathbf{W}'\mathbf{H}' = (\mathbf{W}\mathbf{Q})(\mathbf{Q}^{-1}\mathbf{H})$ where $\mathbf{Q}$ is an invertible matrix and $\mathbf{W}'$ and $\mathbf{H}'$ are two factors. We see that different invertible matrices $\mathbf{Q}$ may yield different solutions of the factors $\mathbf{W}'$ and $\mathbf{H}'$. In other words, the spectral structure of each source $\mathbf{W}$ may be unidentifiable. Finally, the non-sparse solution of standard NMF may lose some local information of the sources.

Simplex volume minimization [6], which learns an identifiable spectral basis, provides a reliable estimation to the source model of BSS. To our knowledge, it has not been explored in multichannel BSS yet.

A. Contributions

In this paper, we aim to explore the *minimum-volume* (MinVol) prior for multichannel BSS. Specifically, we apply the MinVol prior as a regularizer for multichannel nonnegative matrix factorization (MNMF) [2] and independent low-rank matrix analysis (ILRMA) [3], which are named m-MNMF and m-ILRMA, respectively. Because the object function of MinVol is to minimize $|\mathbf{W}^T\mathbf{W}|$ where $\mathbf{W}$ is the basis matrix of NMF and $|\cdot|$ denotes the determinant operator for a nonsingular matrix, it is formulated as a complicated optimization problem. To overcome this difficulty, we design two auxiliary functions for the object functions of m-MNMF and m-ILRMA respectively, and combine them with the maximum a posteriori (MAP) estimation. Each auxiliary function is solved by iteratively updating the demixing matrix, spectrogram basis, and temporal activations in the function. The proposed methods improve the source model of MNMF and ILRMA, which leads to better empirical performance. The contributions of the MinVol regularizer to MNMF and ILRMA are as follows:

- MinVol improves the identifiability of the separated spectrograms that are produced from the source model, since MinVol has been proven to be able to lead to the identifiability for blind source separation [7].
- MinVol improves the sparseness levels of the factorized spectral basis matrices of the source model, which enhances the learning ability of the source model to capture the local information of sources.

Manuscript received March 31, 2021; revised July 14, 2021 and September 29, 2021; accepted September 29, 2021. Date of publication October 20, 2021; date of current version October 27, 2021. This work was supported by the National Science Foundation of China (NSFC) under Grant 62176211. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zheng-Hua Tan. *(Corresponding author: Xiao-Lei Zhang.)*

The authors are with the School of Marine Science and Technology, and the Center of Intelligent Acoustics and Immersive Communications (CIAIC), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: alexwang96@mail.nwpu.edu.cn; gshanzheng@mail.nwpu.edu.cn; shupei.liu@mail.nwpu.edu.cn; xiaolei.zhang@nwpu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2021.3120603

- MinVol enhances the orthogonality of the factorized spectral basis matrices of the source model, which leads the spectral basis matrices to a rigorous clustering interpretation.

In this paper, we first introduce some related work and preliminaries in the following two subsections, then present the proposed MinVol prior distribution, as well as m-MNMF and m-ILRMA in Section III. Section IV presents the experimental results. Finally, Section VI concludes our findings.

B. Related Work

A multichannel BSS method is composed of a spatial model and a source model. A mixture sound is usually represented as a sum of multiple source signals convolved with the room impulse responses of the corresponding source directions. It is equivalent to an instantaneous mix up in the frequency domain. The mixture is usually separated by the spatial model, where the phase difference between microphones is important for the demixing system. Common algorithms for multichannel BSS are independent component analysis (ICA) [8] and its extensions such as independent vector analysis (IVA) [9]. They make a statistical independence assumption between the sources. However, they do not utilize the spectral structures of the source signals.

Recently, the importance of source models has been fully aware. According to the difference of the source models, modern multichannel BSS methods can be categorized mainly into the NMF-based methods, probability-based methods, and deep neural network (DNN) based methods, which will be introduced in the next three subsections respectively.

1) NMF-Based Models: The original ICA and IVA employ a spherical multivariate Laplace distribution as the source model to ensure higher-order correlations between the frequency bins in each source. However, these source models do not fully utilize the spectral structure of sources. As we know, the spectral structure may significantly help improve the BSS performance if properly incorporated into source models.

To overcome this weakness, NMF [10], [11], which is a nonnegative-parts-based low-rank decomposition of an observed nonnegative data matrix, can be used as a source model. It generates a "clustering-friendly" latent spectrogram basis for each source by introducing a low-rank structure into the source model. Generally, NMF-based BSS models adopt Itakura-Saito divergence to evaluate the reconstruction error between the mixture and the estimated sources. MNMF [1], [2], which is an extension of the NMF methods, estimates the mixing system of convolutive mixtures in a similar way to ICA and IVA, which is used for the clustering of spectrogram bases. It consists of a low-rank source model and a full-rank spatial model. The full-rank spatial model is capable of representing a wide variety of source directivity under an echoic condition.

However, MNMF tends to get stuck in bad local optima, since that a large number of unconstrained spatial covariance matrices are needed to be estimated iteratively. To address this problem, Kitamura *et al.* [3], [12] proposed ILRMA. It makes a rank-1 assumption to the spatial model. It performs well for directional sources in practice. Essentially, the spatial model and source model of ILRMA are independent vector analysis (IVA) [9] and NMF respectively, which are optimized iteratively.

In the original MNMF and ILRMA, the observed signal is assumed to follow a time-variant multivariant complex Gaussian distribution. Recently, the methods *t*-MNMF [13] and *t*-ILRMA [14] use the isotropic complex Cauchy distribution [15] and its generalization—complex student's *t*-distribution [16] respectively to replace the original complex Gaussian distribution. Because the Student's *t*-distribution belongs to the family of the α-stable distribution which is more suitable for modeling complex-valued signals than the complex Gaussian distribution, it is suitable for audio source modeling [14]. Moreover, Kitamura *et al.* [4], [17]–[19] developed a complex generalized Gaussian distribution for ILRMA, which takes *t*-ILRMA and ILRMA as its special cases. To reduce the huge computational cost of the spatial covariance matrices, Sekiguchi *et al.* [20], [21] proposed a fast MNMF, which restricts the covariance matrices to jointly-diagonalizable full-rank matrices in a frequency-wise manner. However, its source separation performance was not improved and the physical meaning of the joint-diagonalization process was unclear [22]. To address this issue, Kamo *et al.* [22] proposed FastMNMF with a new regularization, where the authors declared that the regularization can be applied to ILRMA as well.

2) Probability-Based Models: If the frequency bins of each source are sparsely distributed, the source spectrograms can be assumed to be disjoint with each other in most time-frequency units. Under this assumption, Otsuka *et al.* [23] proposed a Bayesian mixture model, called hierarchical latent Dirichlet allocation (LDA) [24], to classify each time-frequency unit into one source only, and classify each source into a single direction. However, it does not build a source model, which is insufficient in utilizing the spectral structure of sources.

To overcome this weakness, probabilistic models were employed to build priors for the distributions of the parameters of the source model. Itakura *et al.* [25] improve the LDA-based method [23] by combining the low-rank structure of the NMF-based source model. The method iteratively updates the spectrogram basis and temporal activations of the source model, and the variables of the LDA-based spatial model. Itakura *et al.* [26] further introduced an anechoic spatial correlation matrix as a prior distribution of a real spatial correlation matrix for each direction, which avoids the impulse response assumption in previous studies. Recently, Itakura *et al.* [27] proposed a unified Bayesian framework for multichannel BSS and incorporated prior knowledge of the microphone array into BSS. Based on the fundamental categorization of probabilistic models which can be categorized to mixture models and factor models, they proposed four methods for joint modeling the source and spatial models: factor-factor model, mixture-factor model, factor-mixture model, and mixture-mixture model. The above models jointly estimate low-rank sources and spatial covariances on the fly. However, the low-rank assumption does not always hold for speech spectra. To remedy this problem, Sekiguchi *et al.* [28] proposed a semi-supervised method based on an extension of MNMF which consists of a deep generative model called variational auto-encoder (VAE) for speech spectra and a

standard low-rank model for noise spectra. Narisetty *et al.* [29] used Bayesian non-parametric modeling of sources to avoid parameter tuning.

3) DNN-Based Models: To provide a highly accurate estimation to the parameters of the source model, supervised DNN has been introduced into multichannel BSS for the estimation to the source model [28], [30]–[35]. Sekiguchi *et al.* [28], [31] proposed a deep pre-trained generative model of speech spectra and an NMF-based generative model of noise spectra for multichannel speech enhancement. Makishima *et al.* [32] proposed independent deeply learned matrix analysis (IDLMA), which utilizes mutually independent DNN source models for the separation. Kameoka *et al.* [30], [33], [34], [36], [37] proposed a multichannel variational autoencoder (MVAE), which uses a conditional VAE to estimate the power spectrograms of sources. Although the convergence of the optimization of MVAE is guaranteed, its computational complexity is high. Moreover, the accuracy of the source classification of MVAE is unsatisfied. To solve the problems, Li *et al.* [38] employed an auxiliary classifier VAE, which is an information-theoretic extension of the conditional VAE, to learn the generative model of source spectrograms. Togami [39] trained a source model by bidirectional long short-term memory networks with the multichannel Itakura-Saito distance as the training objective. Li *et al.* [35] modeled power spectrograms of sources by a star generative adversarial network (StarGAN). Although more and more DNN models were used in BSS, these models require clean sources for pre-training, which is out of the focus of this paper. Therefore, we will not discuss and compare with the DNN-based models anymore.

II. Problem Formulation

In this section, we formulate the BSS problem. Suppose the short-time Fourier transform (STFT) of a multichannel mixture is $\mathbf{x}_{ij} = [x_{ij1}, \ldots, x_{ijm}, \ldots, x_{ijM}]^T \in \mathbb{C}^M$, where $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $m = 1, \ldots, M$ are the indices of the frequency bins, time frames, and microphones, respectively. The spectrograms of source signals are defined as $\mathbf{s}_{ij} = [s_{ij1}, \ldots, s_{ijn}, \ldots, s_{ijN}]^T \in \mathbb{C}^N$, where N is the number of sources and $n = 1, \ldots, N$ is the index of the nth source, and ^{T} denotes the transpose operator.

We assume the mixing process in the frequency domain is instantaneous, and each source of the mixture is a point source. Then, the mixture and its sources have the following connection:

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij} \tag{1}$$

where $\mathbf{A}_i = [\mathbf{a}_{i1}, \ldots, \mathbf{a}_{in}, \ldots, \mathbf{a}_{iN}] \in \mathbb{C}^{M \times N}$ is the mixing matrix at the ith frequency bin. Similar to [1], [2], [27], we assume that s_{ijn} follows a zero-mean complex Gaussian distribution as follows, since the distribution is well adapted for Multichannel BSS methods:

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

where λ_{ijn} is a power spectrum density of the source n at time j and frequency i. Substituting (2) into (1), the observation x_{ijm} is found to follow the complex Gaussian distribution as follows:

$$x_{ijn} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{ijn}\mathbf{G}_{in}\right) \tag{3}$$

where $\mathbf{G}_{in} = \mathbf{a}_{in}\mathbf{a}_{in}^H$, and ^{H} denotes the Hermitian transpose.

If the matrix $\mathbf{A}_i$ is not singular, the problem of source separation is to find an estimation of $(\mathbf{A}_i)^{-1}$, denoted as $\mathbf{D}_i = [\mathbf{d}_{i,1}, \ldots, \mathbf{d}_{i,N}]^H$, where $\mathbf{D}_i \in \mathbb{C}^{N \times M}$, such that when we apply $\mathbf{D}_i$ to $\mathbf{x}_{ij}$, we obtain the separated signal:

$$\mathbf{y}_{ij} = \mathbf{D}_i \mathbf{x}_{ij} \tag{4}$$

where $\mathbf{y}_{ij}$ is an estimation of $\mathbf{s}_{ij}$. Here, we emphasize that MNMF is suitable for both the underdetermined situation ($M < N$) and the determined situation ($M = N$), while ILRMA is only suitable for the determined situation.

III. Proposed Methods

In this section, we first propose the MinVol based source model in Section III-A, and then present the MinVol regularized MNMF and ILRMA respectively in Sections III-B and III-C.

A. Minimum-Volume Prior Distribution for Source Models

We propose a minimum-volume prior distribution for the source model of the NMF-based BSS models. Specifically, we formulate the generative process of source power spectrograms $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n, \ldots, \boldsymbol{\lambda}_N] = \{\lambda_{ijn}\}_{i,j,n=1}^{I,J,N}$ as follows: $\boldsymbol{\lambda}$ is generated by a basis spectra $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_n, \ldots, \mathbf{W}_N] = \{w_{nik}\}_{n,i,k=1}^{N,I,K}$ and activations $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_n, \ldots, \mathbf{H}_N] = \{h_{nkj}\}_{n,k,j=1}^{N,K,J}$, where K is the number of the bases of the basis matrix $\mathbf{W}_n$. The power spectrogram of each source is decomposed into basis spectra and temporal activations by low-rank factorization:

$$\lambda_{ijn} = \sum_{k=1}^{K} w_{nik} h_{nkj} \tag{5}$$

Based on the above factorization, we can derive the following conditional probability function:

$$p(\boldsymbol{\lambda}_n | \mathbf{W}_n, \mathbf{H}_n) = \prod_{i=1}^{I}\prod_{j=1}^{J} \delta\left(\lambda_{ijn} - \sum_{k=1}^{K} w_{nik} h_{nkj}\right) \tag{6}$$

where $\delta(\cdot)$ is the Dirac delta function. In many existing decomposition methods, the prior over w_{nik} is constructed as a uniform distribution over the non-negative real numbers,

$$\begin{aligned} p(w_{nik}) &= \lim_{u_w \to \infty} \frac{1}{u_w} \mathbb{I}[0 \le w_{nik} \le u_w] \\ &\propto \mathbb{I}[w_{nik} \ge 0] \end{aligned} \tag{7}$$

where $\mathbb{I}[\cdot]$ denotes an indicator function, which has the value one when its argument is true and zero otherwise. The prior for h_{nkj} is chosen as uniform between zero and one:

$$p(h_{nkj}) = \mathbb{I}[0 \le h_{nkj} \le 1] \tag{8}$$

Under the Bayes' rule, the posterior density of w_{nik} and h_{nkj} is given by:

$$p(\mathbf{W}_n, \mathbf{H}_n|\boldsymbol{\lambda}_n) \propto \frac{1}{Z}\prod_{i=1}^{I}\prod_{j=1}^{J}\delta\left(\lambda_{ijn} - \sum_{k=1}^{K} w_{nik}h_{nkj}\right) \times \mathbb{I}[w_{nik} \geq 0]\mathbb{I}[h_{nkj} \geq 0]\mathbb{I}\left[\sum_{k=1}^{K} h_{nkj} = 1\right] \tag{9}$$

where Z is a normalization constant.

Many algorithms have been developed to find a unique and identifiable factorization for NMF, e.g. [6], [7], [40]–[42]. We are interested in the MinVol criterion among these algorithms. MinVol is motivated by the nice geometrical interpretation of the constraints in [41],[43]. Under the MinVol constraints, all the data points lie in a *convex hull* spanned by the spectrogram basis. For convenience, we here present a probabilistic Bayesian formulation of a prior about *the volume of the data simplex:*

$$p(\mathbf{W}|\gamma) \propto \exp(-\gamma \log|\mathbf{W}^T\mathbf{W} + \eta\mathbf{I}|) \tag{10}$$

where $|\cdot|$ is the determinant operator on a matrix, γ is a parameter that reflects the influence of the prior to the likelihood function, and η is a hyperparameter of the prior distribution (10). We choose MinVol as a prior distribution of the spectrogram basis. It encourages the simplex spanned by the estimated spectrograms to be small, and constrains each element of the spectrogram basis to be non-negative.

The posterior density of the source model based on the volume of the data simplex can be represented as follows:

$$p(\mathbf{W}_n, \mathbf{H}_n|\boldsymbol{\lambda}_n) \propto \frac{1}{Z}\prod_{i=1}^{I}\prod_{j=1}^{J}\delta(\lambda_{ijn} - \sum_{k=1}^{K} w_{nik}h_{nkj}) \times \exp(-\gamma \log|\mathbf{W}_n^T\mathbf{W}_n + \eta\mathbf{I}|) \times \mathbb{I}[w_{nik} \geq 0]\mathbb{I}[h_{nkj} \geq 0]\mathbb{I}\left[\sum_{k=1}^{K} h_{nkj} \geq 0\right] \tag{11}$$

Maximizing the likelihood of (11) is equivalent to the volume-minimization problem of the data simplex.

B. MNMF With Minimum-Volume Regularizer

1) Preliminary: MNMF [2] decomposes the spatial covariance matrix of each source into a weighted sum of direction-dependent matrices for joint source separation and location, which can be formulated as the following maximization problem:

$$\begin{aligned}&\log\left[p(\mathbf{X}|\mathbf{W}_n, \mathbf{H}_n, \mathbf{G})p(\mathbf{W}_n)p(\mathbf{W}_n)\right]\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\log\mathcal{N}_{\mathbb{C}}(\mathbf{x}_{ij}|\mathbf{0}, \hat{\mathbf{X}}_{ij})\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(-\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\right) - \log|\hat{\mathbf{X}}_{ij}|\right) + \mathrm{C}\end{aligned} \tag{12}$$

where $\mathbf{X}_{ij} = \mathbf{x}_{ij}\mathbf{x}_{ij}^H$, $\hat{\mathbf{X}}_{ij} = \sum_n \lambda_{ijn}\mathbf{G}_{ni}$, and 'C' represents a constant.

It is usually solved by a multiplicative update rule which iteratively updates one of the parameters according to the conditional posterior distribution with the other parameters fixed:

$$w_{nik} \leftarrow w_{nik}\sqrt{\frac{\sum_j h_{nkj}\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}\mathbf{G}_{ni}\right)}{\sum_j h_{nkj}\mathrm{tr}\left(\hat{X}_{ij}^{-1}\mathbf{G}_{ni}\right)}} \tag{13}$$

$$h_{nkj} \leftarrow h_{nkj}\sqrt{\frac{\sum_i w_{nik}\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}\mathbf{G}_{ni}\right)}{\sum_i w_{nik}\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{G}_{ni}\right)}} \tag{14}$$

To update $\mathbf{G}_{ni}$, Sawada *et al.* solve an algebraic Riccati equation:

$$\mathbf{G}_{ni}\mathbf{A}\mathbf{G}_{ni} = \mathbf{B} \tag{15}$$

with $\mathbf{A}$ and $\mathbf{B}$ defined as:

$$\mathbf{A} = \sum_j h_{nkj}\hat{\mathbf{X}}_{ij}^{-1}, \quad \mathbf{B} = \mathbf{G}_{ni}{}^{\star}\left(\sum_j h_{nkj}\hat{\mathbf{X}}_{ij}^{-1}\right)\mathbf{G}_{ni}{}^{\star} \tag{16}$$

where $\mathbf{G}^{\star}$ is the old value of the variable $\mathbf{G}$ calculated in the previous step.

2) Objective Function of m-MNMF: To remedy the non-unique identifiable problem of the source model of MNMF, here we propose m-MNMF. Fig. 1 shows a conceptual model of m-MNMF, which is described as follows.

The likelihood function of the unknown variables $\mathbf{W}, \mathbf{H}, \mathbf{G}$ of m-MNMF is formulated as:

$$\begin{aligned}&\log\left[p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{G})p(\mathbf{W}|\gamma)p(\mathbf{H})\right]\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\log\mathcal{N}_{\mathbb{C}}(\mathbf{x}_{ij}|\mathbf{0}, \hat{\mathbf{X}}_{ij}) - \sum_{n=1}^{N}\gamma\log|\mathbf{W}_n^T\mathbf{W}_n + \eta\mathbf{I}|\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(-\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\right) - \log|\hat{\mathbf{X}}_{ij}|\right)\\ &\quad - \sum_{n=1}^{N}\gamma\log|\mathbf{W}_n^T\mathbf{W}_n + \eta\mathbf{I}| + \mathrm{C}\end{aligned} \tag{17}$$

where $\mathbf{X}_{ij} = \mathbf{x}_{ij}\mathbf{x}_{ij}^H$, $\hat{\mathbf{X}}_{ij} = \sum_{n=1}^{N}\lambda_{ijn}\mathbf{G}_{ni}$.

The logarithmic absolute value term in (17) is too difficult to optimize directly. Therefore, we propose to maximize its lower bound instead. To derive a lower bound for (17), we use the following two inequalities [16], [44] to relax the logarithmic determinant term in (17):

First, for a convex function $f(\mathbf{Z}) = -\log|\mathbf{Z}|$ with $\mathbf{Z} \geq \mathbf{0}$ being a positive semi-definite matrix, we have the following lower bound at an arbitrary positive semi-definite matrix $\mathbf{U} \geq \mathbf{0}$:

$$f(\mathbf{Z}) = -\log|\mathbf{Z}| \geq -\log|\mathbf{U}| - \mathrm{tr}(\mathbf{U}^{-1}\mathbf{Z}) + M \tag{18}$$

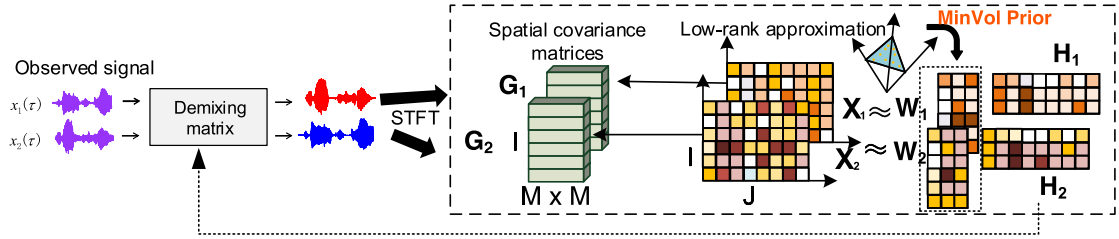


Fig. 1. Principle of the proposed m-MNMF algorithm.

where the equality holds when $\mathbf{U} = \mathbf{Z}$.

Second, for a concave function $g(\mathbf{Z}) = -\mathrm{tr}(\mathbf{Z}^{-1}\mathbf{A})$ with any matrix $\mathbf{A} \geq \mathbf{0}$, we have the following lower bound:

$$g(\{\mathbf{Z}_l\}_{l=1}^L) = -\mathrm{tr}\left(\left(\sum_{l=1}^L \mathbf{Z}_l\right)^{-1}\mathbf{A}\right) \geq -\sum_{l=1}^L \mathrm{tr}(\mathbf{Z}_l^{-1}\mathbf{\Phi}_l\mathbf{A}\mathbf{\Phi}_l^H) \quad (19)$$

where $\{\mathbf{Z}_l\}_{l=1}^L$ is a set of arbitrary matrices, $\{\mathbf{\Phi}_l\}_{l=1}^L$ is a set of auxiliary matrices that satisfies $\sum_l \mathbf{\Phi}_l = \mathbf{I}$, and the equality holds when $\mathbf{\Phi}_k = \mathbf{Z}_k(\sum_{l'}\mathbf{Z}_{l'})^{-1}$.

Substituting the two inequalities (18) and (19) into (17) derives the following lower bound of (17), denoted as $\mathcal{L}$:

$$\begin{aligned}
&\log\left[p(\mathbf{X}|\mathbf{W},\mathbf{H},\mathbf{G})p(\mathbf{W}|\gamma)p(\mathbf{H})\right] \\
&\geq \sum_{i=1}^I\sum_{j=1}^J\left(-\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}\mathbf{U}_{ij}^{-1}\right) - \log|\mathbf{U}_{ij}| + M\right) \\
&\quad - \sum_{i=1}^I\sum_{j=1}^J\sum_{n=1}^N \mathrm{tr}\left(\hat{\mathbf{X}}_{ijn}^{-1}\mathbf{\Phi}_{ijn}\mathbf{X}_{ij}\mathbf{\Phi}_{ijn}^H\right) \\
&\quad + \gamma\sum_{n=1}^N\left(-\log|\mathbf{V}^{-1}| - \mathrm{tr}(\mathbf{V}\mathbf{W}_n^T\mathbf{W}_n) + K\right) \\
&= -\sum_{i=1}^I\sum_{j=1}^J\sum_{n=1}^N \lambda_{ijn}\mathrm{tr}\left(\mathbf{G}_{ni}\mathbf{U}_{ij}^{-1}\right) - \sum_{i=1}^I\sum_{j=1}^J\log|\mathbf{U}_{ij}| \\
&\quad - \sum_{i=1}^I\sum_{j=1}^J\sum_{n=1}^N \lambda_{ijn}^{-1}\mathrm{tr}\left(\mathbf{G}_{ni}^{-1}\mathbf{\Phi}_{ijn}\mathbf{X}_{ij}\mathbf{\Phi}_{ijn}^H\right) \\
&\quad + \gamma\sum_{n=1}^N\left(-\log|\mathbf{V}^{-1}| - \mathrm{tr}(\mathbf{V}\mathbf{W}_n^T\mathbf{W}_n)\right) + \mathrm{C} = \mathcal{L}
\end{aligned} \quad (20)$$

where $\boldsymbol{\lambda}$ is a function of $\mathbf{H}$ and $\mathbf{W}$ defined in (5), and $\mathbf{U}_{ij}$, $\mathbf{\Phi}_{ijn}$, and $\mathbf{V}$ are auxiliary variables. The above lower bound is a tight one when $\mathbf{U}_{ij}$, $\mathbf{\Phi}_{ijn}$ and $\mathbf{V}$ satisfy:

$$\mathbf{U}_{ij} = \hat{\mathbf{X}}_{ij} \quad (21)$$

$$\mathbf{\Phi}_{ijn} = \hat{\mathbf{X}}_{ijn}\hat{\mathbf{X}}_{ij}^{-1} \quad (22)$$

$$\mathbf{V} = (\mathbf{W}_n^T\mathbf{W}_n + \eta\mathbf{I})^{-1} \quad (23)$$

The objective function of m-MNMF is to maximize $\mathcal{L}$.

3) Optimization of m-MNMF: m-MNMF is optimized by the multiplicative updating (MU) rule [7], which optimizes $\mathbf{H}_n$, $\mathbf{W}_n$, and $\mathbf{G}_{ni}$, alternatively.

Given $\mathbf{W}_n$ and $\mathbf{G}_{ni}$ fixed, $\mathbf{H}_n$ is calculated as follows. Letting the partial derivative of (20) with respect to h_{nkj} equal to zero derives:

$$\sum_{i=1}^I h_{nkj}^{-2}w_{nik}^{-1}\mathrm{tr}\left(\mathbf{G}_{ni}^{-1}\mathbf{\Phi}_{ijn}\mathbf{X}_{ij}\mathbf{\Phi}_{ijn}^H\right) - \sum_{i=1}^I w_{nik}\mathrm{tr}\left(\mathbf{G}_{ni}\mathbf{U}_{ij}^{-1}\right) = 0 \quad (24)$$

and h_{nkj} can be obtained as:

$$h_{nkj} \leftarrow h_{nkj}^{\star}\sqrt{\frac{\sum_{i=1}^I w_{nik}\mathrm{tr}\left(\mathbf{G}_{ni}\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}\right)}{\sum_{i=1}^I w_{nik}\mathrm{tr}\left(\mathbf{G}_{ni}\hat{\mathbf{X}}_{ij}^{-1}\right)}} \quad (25)$$

Lemma 1 ([7]): Let $\mathbf{w}_{ni} \in \mathbb{R}^{1\times K}$ be the ith row of $\mathbf{W}_n$, and $\hat{\mathbf{w}}_{ni} \in \mathbb{R}^{1\times K}$ be an auxiliary variable of $\mathbf{w}_{ni}$. $\mathbf{V}$ can be decomposed as $\mathbf{V} = \mathbf{V}^+ - \mathbf{V}^-$ with $\mathbf{V}^+ = \max(\mathbf{V}, \mathbf{0})$ and $\mathbf{V}^- = \max(-\mathbf{V}, \mathbf{0})$ where $\max(\cdot,\cdot)$ is a component-wise operation that returns the same size matrix with the greater components in each element, and $\mathbf{\Omega}(\mathbf{w}_{ni}^T)$ is the diagonal matrix $\mathbf{\Omega}(\mathbf{w}_{ni}^T) = \mathrm{Diag}(2\frac{[\mathbf{V}^+\mathbf{w}_{ni}^T + \mathbf{V}^-\mathbf{w}_{ni}^T]}{[\mathbf{w}_{ni}^T]})$ where $\frac{[\mathbf{A}]}{[\mathbf{B}]}$ is the component-wise division between $\mathbf{A}$ and $\mathbf{B}$, and $\Delta\hat{\mathbf{w}}_{ni} = \hat{\mathbf{w}}_{ni} - \mathbf{w}_{ni}$. Then

$$\hat{g}(\mathbf{w}_{ni}^T|\hat{\mathbf{w}}_{ni}^T) = g(\hat{\mathbf{w}}_{ni}^T) + \Delta\mathbf{w}_{ni}\nabla g(\hat{\mathbf{w}}_{ni}^T) + \frac{1}{2}\Delta\mathbf{w}_{ni}\mathbf{\Omega}(\hat{\mathbf{w}}_{ni}^T)\Delta\mathbf{w}_{ni}^T \quad (26)$$

is a separable auxiliary function for $g(\mathbf{w}_{ni}^T) = \mathbf{w}_{ni}\mathbf{V}\mathbf{w}_{ni}^T$ at $\hat{\mathbf{w}}_{ni}$.

Given $\mathbf{H}_n$ and $\mathbf{G}_{ni}$ fixed, $\mathbf{W}_n$ is calculated as follows. Because $\mathrm{tr}(\mathbf{V}\mathbf{W}_n^T\mathbf{W}_n)$ of (20) is quadratic and not separable, we optimize a compact lower-bound of (20) with an approximate separable auxiliary function. Specifically, given Lemma 1, we obtain the lower-bound of $\mathcal{L}$ with respect to w_{nik} as:

$$\begin{aligned}
\mathcal{L}_{w_{nik}} &= -\sum_{i=1}^I\sum_{j=1}^J\sum_{n=1}^N\left(\sum_{k=1}^K w_{nik}h_{nkj}\right)\mathrm{tr}(\mathbf{G}_{ni}\mathbf{U}_{ij}^{-1}) \\
&\quad - \sum_{i=1}^I\sum_{j=1}^J\sum_{n=1}^N\left(\sum_{k=1}^K w_{nik}h_{nkj}\right)^{-1} \\
&\quad \mathrm{tr}(\mathbf{G}_{ni}^{-1}\mathbf{\Phi}_{jin}\mathbf{X}_{ij}\mathbf{\Phi}_{ijn}^H)
\end{aligned}$$

$$- \gamma \sum_{n=1}^{N} \sum_{i=1}^{I} \left[\hat{\mathbf{w}}_{ni} \mathbf{V} \hat{\mathbf{w}}_{ni}^{T} + 2\Delta \hat{\mathbf{w}}_{ni} \mathbf{V} \hat{\mathbf{w}}_{ni}^{T} + \Delta \hat{\mathbf{w}}_{ni} \boldsymbol{\Omega} (\hat{\mathbf{w}}_{ni}^{T}) \Delta \hat{\mathbf{w}}_{ni}^{T} \right] \tag{27}$$

We let the partial derivative of $\mathcal{L}_{w_{nik}}$ to zero:

$$\begin{aligned} \frac{\partial \mathcal{L}_{w_{nik}}}{\partial w_{nik}} &= w_{nik}^{-2} h_{nkj}^{-1} \mathrm{tr}\left(\mathbf{G}_{ni}^{-1} \boldsymbol{\Phi}_{ijn} \mathbf{X}_{ij} \boldsymbol{\Phi}_{ijn}^{H} \right) \\ &\quad - h_{nkj} \mathrm{tr}\left(\mathbf{G}_{ni} \mathbf{U}_{ij}^{-1} \right) + 2\gamma [\mathbf{V} \hat{\mathbf{w}}_{ni}^{T}]_k \\ &\quad + 2\gamma \left[\mathrm{Diag}\left(\frac{\mathbf{V}^{+} \hat{\mathbf{w}}_{ni}^{T} + \mathbf{V}^{-} \hat{\mathbf{w}}_{ni}^{T}}{\hat{\mathbf{w}}_{ni}} \right) \right]_k w_{nik} \\ &\quad - 2\gamma \left[\mathrm{Diag}\left(\frac{\mathbf{V}^{+} \hat{\mathbf{w}}_{ni}^{T} + \mathbf{V}^{-} \hat{\mathbf{w}}_{ni}^{T}}{\hat{\mathbf{w}}_{ni}^{T}} \right) \right]_k \hat{w}_{nik} = 0 \end{aligned} \tag{28}$$

and further make the following abbreviations for clarity:

$$a = 2\gamma \left[\mathrm{Diag}\left(\frac{\mathbf{V}^{+} \hat{\mathbf{w}}_{ni}^{T} + \mathbf{V}^{-} \hat{\mathbf{w}}_{ni}^{T}}{\hat{\mathbf{w}}_{ni}^{T}} \right) \right]_k \tag{29}$$

$$\begin{aligned} b &= 2\gamma [\mathbf{V} \hat{\mathbf{w}}_{ni}^{T}]_k - h_{nkj} \mathrm{tr}\left(\mathbf{G}_{ni} \mathbf{U}_{ij}^{-1} \right) \\ &\quad - 2\gamma \left[\mathrm{Diag}\left(\frac{\mathbf{V}^{+} \hat{\mathbf{w}}_{ni}^{T} + \mathbf{V}^{-} \hat{\mathbf{w}}_{ni}^{T}}{\hat{\mathbf{w}}_{ni}^{T}} \right) \right]_k \hat{w}_{nik} \end{aligned} \tag{30}$$

$$d = h_{nkj}^{-1} \mathrm{tr}\left(\mathbf{G}_{ni}^{-1} \boldsymbol{\Phi}_{ijn} \mathbf{X}_{ij} \boldsymbol{\Phi}_{ijn}^{H} \right) \tag{31}$$

Then, (28) can be rewritten as:

$$a w_{nik}^{3} + b w_{nik}^{2} + d = 0 \tag{32}$$

We employ the *cubic roots* procedure [45] to solve problem (32). In order to ensure the nonnegativity of w_{nik}, we simply use the Descarte's rules of sign in the root of (32) as in [7]:

$$w_{nik} \leftarrow \max(0, w_{nik}) \tag{33}$$

Given $\mathbf{H}_n$ and $\mathbf{W}_n$ fixed, the spatial model $\mathbf{G}_{ni}$ is calculated as follows. We let the partial derivative of $\mathcal{L}$ with respect to $\mathbf{G}_{ni}$ equal to zero:

$$\sum_{j=1}^{J} \lambda_{ijn}^{-1} \mathbf{G}_{ni}^{-1} \boldsymbol{\Phi}_{ijn} \mathbf{X}_{ij} \boldsymbol{\Phi}_{ijn}^{H} \mathbf{G}_{ni}^{-1} - \sum_{j=1}^{J} \lambda_{ijn} \mathbf{U}_{ij}^{-1} = \mathbf{0} \tag{34}$$

where $\mathbf{0}$ is an all-zero matrix of size $M \times M$. Substituting $\mathbf{U}_{ij}$ and $\boldsymbol{\Phi}_{ijn}$ into (34) derives:

$$\mathbf{G}_{ni}^{\star} \mathbf{A}_{\mathbf{G}} \mathbf{G}_{ni}^{\star} = \mathbf{G}_{ni} \mathbf{B}_{\mathbf{G}} \mathbf{G}_{ni} \tag{35}$$

where $\mathbf{G}_{ni}^{\star}$ is value of $\mathbf{G}_{ni}$ at the previous step, $\mathbf{A}_{\mathbf{G}}$ and $\mathbf{B}_{\mathbf{G}}$ are short for

$$\mathbf{A}_{\mathbf{G}} = \sum_{j=1}^{J} \lambda_{ijn} \hat{\mathbf{X}}_{ij}^{-1} \mathbf{X}_{ij} \hat{\mathbf{X}}_{ij}^{-1} \tag{36}$$

Algorithm 1: m-MNMF.

Input : Mixture $\mathbf{x}_{ij}$, number of sources N, *MaxIteration*, hyperparameter $\eta \geq 0$.
Output: Separated signal $\mathbf{y}_{ij}$.

Initialize: $\mathbf{W}_n$, $\mathbf{H}_n$, $\mathbf{G}_{ni}$;
for *iteration = 1 to MaxIteration* **do**
  for *n=1 to N* **do**
    for *i=1 to I* **do**
      for *k=1 to K* **do**
        Update h_{nkj} by (25);
        Update w_{nik} by solving (32) and (33);
      end
      for *j=1 to J* **do**
        Compute $\boldsymbol{\Phi}_{ijn} = \hat{\mathbf{X}}_{ijn} \hat{\mathbf{X}}_{ij}^{-1}$, $\hat{\mathbf{X}}_{ij} = \sum_{n=1}^{N} \hat{\mathbf{X}}_{ijn}$;
      end
    end
    Compute $\mathbf{V}_n = (\mathbf{W}_n^{T} \mathbf{W}_n + \delta \mathbf{I})^{-1}$;
  end
  for *n=1 to N* **do**
    for *i = 1 to I* **do**
      Update spatial covariance matrix $\mathbf{G}_{ni}$ by (36), (37), (38);
      for *j=1 to J* **do**
        Compute $\boldsymbol{\Phi}_{ijn} = \hat{\mathbf{X}}_{ijn} \hat{\mathbf{X}}_{ij}^{-1}$, $\hat{\mathbf{X}}_{ij} = \sum_{n=1}^{N} \hat{\mathbf{X}}_{ijn}$;
      end
    end
  end
end
Compute $\mathbf{y}_{ij}$ by multichannel Wiener filter.

$$\mathbf{B}_{\mathbf{G}} = \sum_{j=1}^{J} \lambda_{ijn} \hat{\mathbf{X}}_{ij}^{-1} \tag{37}$$

(35) has a closed-form updating rule for $\mathbf{G}_{ni}$:

$$\mathbf{G}_{ni} \leftarrow \mathbf{G}_{ni}^{\star} (\mathbf{G}_{ni}^{\star} \mathbf{A}_{\mathbf{G}} \mathbf{G}_{ni}^{\star}) \sharp (\mathbf{B}_{\mathbf{G}})^{-1} \tag{38}$$

where $\mathbf{A} \sharp \mathbf{B}$ is the geometric mean of two positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{A} \sharp \mathbf{B} = \mathbf{A}^{\frac{1}{2}} \left(\mathbf{A}^{-\frac{1}{2}} \mathbf{B} \mathbf{A}^{-\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}} = \mathbf{A} (\mathbf{A}^{-1} \mathbf{B})^{\frac{1}{2}} \tag{39}$$

The above algorithm is summarized in Algorithm 1.

C. ILRMA With Minimum-Volume Regularizer

1) Preliminary: ILRMA utilizes the assumption of the invertibility of mixing matrix $\mathbf{A}_i$ to transform the spatial optimization of MNMF into the estimation problem of the demixing matrix $\mathbf{D}_i$. We note that ILRMA cannot be applied to the underdetermined BSS problem because the mixing matrix $\mathbf{A}_i$ must be invertible. ILRMA employs a flexible source model to estimate the demixing matrix $\mathbf{D}_i$ in a stable manner as in AuxIVA [46]. When $\mathbf{G}_{ni}$ is a rank-1 matrix given by $\mathbf{G}_{ni} = \mathbf{a}_{in}^{H} \mathbf{a}_{in}$, $\hat{\mathbf{X}}_{ij}$ can

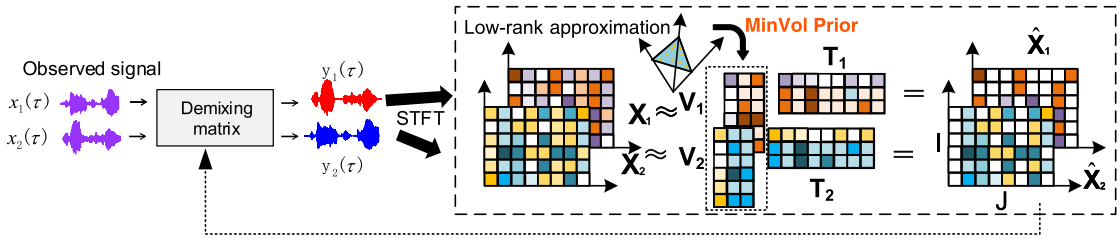


Fig. 2. Principle of the proposed m-ILRMA algorithm.

be calculated by:

$$\begin{aligned}\hat{\mathbf{X}}_{ij} &= \sum_{n=1}^{N} \lambda_{ijn}\mathbf{a}_{in}^H\mathbf{a}_{in} \\ &= \mathbf{A}_i\mathbf{\Lambda}_{ij}\mathbf{A}_i^H \\ &= \mathbf{D}_i^{-1}\mathbf{\Lambda}_{ij}\mathbf{D}_i^{-H} \end{aligned} \tag{40}$$

where $\mathbf{\Lambda}_{ij} = \mathrm{Diag}(\lambda_{ij1}, \ldots, \lambda_{ijN})$ is a diagonal matrix.

By substituting (40) into the cost function of MNMF (12), we obtain:

$$\begin{aligned}&\log\left[p(\mathbf{X}|\mathbf{W},\mathbf{H},\mathbf{G})p(\mathbf{W})p(\mathbf{H})\right] \\ &= -\sum_{i=1}^{I}\sum_{j=1}^{J} \mathrm{tr}\left(\mathbf{y}_{ij}^H\mathbf{D}_i^{-H}\left(\mathbf{D}_i^H\mathbf{\Lambda}_{ij}^{-1}\mathbf{D}_i\right)\mathbf{D}_i^{-1}\mathbf{y}_{ij}\right) \\ &\quad + J\sum_{i=1}^{I}\log|\mathbf{D}_i\mathbf{D}_i^H| - \sum_{i=1}^{I}\sum_{j=1}^{J}\log|\mathbf{\Lambda}_{ij}| + \mathrm{C}\end{aligned} \tag{41}$$

The demixing matrix $\mathbf{D}_i$ of the spatial model in ILRMA is updated based on the rules of AuxIVA which can be represented as follows:

$$\begin{aligned}\mathbf{G}_{ni} &= \frac{1}{J}\sum_j \frac{1}{\lambda_{ijn}}\mathbf{x}_{ij}\mathbf{x}_{ij}^h \\ \mathbf{d}_{in} &\leftarrow (\mathbf{D}_i\mathbf{G}_{ni})^{-1}\mathbf{e}_m \\ \mathbf{d}_{in} &\leftarrow \mathbf{d}_{in}(\mathbf{d}_{in}^h\mathbf{G}_{ni}\mathbf{d}_{in})^{-\frac{1}{2}}\end{aligned} \tag{42}$$

where $\mathbf{d}_{in}$ is a row vector of $\mathbf{D}_i$, and $\mathbf{e}_m$ denotes the nth column vector of an $M \times M$-dimensional identity matrix.

The parameters of the source model $\mathbf{W_n}$ and $\mathbf{T}_n$ are updated by MU:

$$w_{nik} \leftarrow w_{nik}\sqrt{\frac{\sum_j |y_{ijn}|^2 h_{nkj}\left(\sum_k w_{nik}h_{nkj}\right)^{-2}}{\sum_j h_{nkj}\left(\sum_k w_{nik}h_{nkj}\right)^{-1}}} \tag{43}$$

$$h_{nkj} \leftarrow h_{nkj}\sqrt{\frac{\sum_i |y_{ijn}|^2 w_{nik}\left(\sum_k w_{nik}h_{nkj}\right)^{-2}}{\sum_i w_{nik}\left(\sum_k w_{nik}h_{nkj}\right)^{-1}}} \tag{44}$$

2) Objective Function of m-ILRMA: To remedy the non-unique identifiable problem of the source model of ILRMA, here we propose m-ILRMA. Fig. 2 shows a conceptual model of m-ILRMA. Specifically, substituting (40) into (17) derives the objective of m-ILRMA:

$$\begin{aligned}&\log\left[p(\mathbf{X}|\mathbf{W},\mathbf{H},\mathbf{G})p(\mathbf{W}|\gamma)p(\mathbf{H})\right] \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(-\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\right) - \log|\hat{\mathbf{X}}_{ij}|\right) \\ &\quad - \sum_{n=1}^{N}\gamma\log|\mathbf{W}_n^T\mathbf{W}_n + \eta\mathbf{I}| + \mathrm{C} \\ &= -\sum_{i=1}^{I}\sum_{j=1}^{J}\mathrm{tr}\left(\mathbf{y}_{ij}^H\mathbf{D}_i^{-H}\left(\mathbf{D}_i^H\mathbf{\Lambda}_{ij}^{-1}\mathbf{D}_i\right)\mathbf{D}_i^{-1}\mathbf{y}_{ij}\right) \\ &\quad + J\sum_{i=1}^{I}\log|\mathbf{D}_i\mathbf{D}_i^H| - \sum_{n=1}^{N}\gamma\log|\mathbf{W}_n^T\mathbf{W}_n + \delta\mathbf{I}| \\ &\quad - \sum_{i=1}^{I}\sum_{j=1}^{J}\log|\mathbf{\Lambda}_{ij}| + \mathrm{C} \\ &= -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}\left(\frac{|y_{ijn}|^2}{\lambda_{ijn}} + \log\lambda_{ijn}\right) + \mathrm{C} \\ &\quad + J\sum_{i=1}^{I}\log|\mathbf{D}_i\mathbf{D}_i^H| - \sum_{n=1}^{N}\gamma\log|\mathbf{W}_n^T\mathbf{W}_n + \delta\mathbf{I}|\end{aligned} \tag{45}$$

Because the full rank spatial model in MNMF has more parameters than the rank-1 spatial model in ILRMA, ILRMA is less sensitive to its parameter initialization, so as to the advantage of m-LIRMA over m-MNMF. As will be shown in the experiments, m-ILRMA achieves better separation performance than m-MNMF.

3) Optimization of m-ILRMA: m-ILRMA is optimized by the multiplicative updating (MU) rule, which optimizes $\mathbf{H}_n$, $\mathbf{W}_n$, and $\mathbf{G}_{ni}$, alternatively.

Given $\mathbf{H}_n$ and $\mathbf{G}_{ni}$ fixed, the objective (45) with respect to $\mathbf{W}_n$ is a difficult maximization problem. In order to solve this problem, we propose to maximize a lower-bound of (45) by an auxiliary function proposed in [47]. Specifically, we first define a Q function by Lemma 2.

Lemma 2: Let $\hat{y}_{ijn} = \sum_k \hat{w}_{nik} h_{nkj}$ and $\hat{s}_{ijn} \geq 0, \hat{w}_{nik} \geq 0$. Then, the following Q function:

$$Q(\mathbf{w}_{ni}^T|\hat{\mathbf{w}}_{ni}^T) = \left[\sum_k \frac{\hat{w}_{nik} h_{nkj}}{\hat{y}_{ijn}} \breve{\rho}\left(y_{ijn}|\hat{y}_{ijn}\frac{w_{nik}}{\hat{w}_{nik}}\right)\right] + \bar{\rho}(\hat{y}_{ijn}) + \left[\hat{\rho}'(y_{ijn}|\hat{y}_{ijn})\sum_k (w_{nik} - \hat{w}_{nik}) h_{nkj} + \hat{\rho}(y_{ijn}|\hat{y}_{ijn})\right] \quad (46)$$

is an auxiliary function to $Q(\mathbf{w}_{ni\cdot})$ at $\hat{w}_{nik}$ [47], where $\breve{\rho}$ is a convex function with respect to $\hat{y}_{ijn}\frac{w_{nik}}{\hat{w}_{nik}}$, $\hat{\rho}$ is a concave function with respect to $\hat{y}_{ijn}$, and $\bar{\rho}$ is a constant function with respect to y_{ijn}. $\hat{\rho}'$ is the differential of $\hat{\rho}(y_{ijn}|\hat{y}_{ijn})$ at $\hat{y}_{ijn}$. In the case of Itakura-Saito (IS) divergence, we have $\breve{\rho}(x|y) = xy^{-1}$, $\hat{\rho}(x|y) = \log y$, $\bar{\rho}(x) = x(\log x - 1)$, $\hat{\rho}'(x|y) = y^{-1}$.

Similar to m-MNMF, we use (18) to construct a low-bound of the likelihood function with respect to $\mathbf{W}_n$:

$$\mathcal{L}_{w_{nik}} = \sum_{n=1}^{N}\sum_{i=1}^{I}\left[Q(\mathbf{w}_{ni}^T|\hat{\mathbf{w}}_{ni}^T) + \gamma\left[\log|\det(\mathbf{V}_n^{-1})| + \mathrm{tr}(\mathbf{V}_n\mathbf{W}_n^T\mathbf{W}_n) - K\right]\right] \quad (47)$$

where $\mathbf{V}_n = (\mathbf{Z}^T\mathbf{Z} + \eta\mathbf{I})^{-1}$ with $\eta \geq 0$, $\mathbf{Z} \in \mathbb{R}^{I\times K}$ is an arbitrary positive definite matrix. We can set $\mathbf{Z} = \mathbf{W}_n$ in the experiments, since $\mathbf{W}_n$ is a positive definite matrix. Finally, the right side of (18) is an auxiliary function for $\log|\mathbf{W}_n^T\mathbf{W}_n|$. Because it is quadratic and inseparable, we use an approximation to represent the right side of (18). Specifically, let $\mathbf{V}_n = \mathbf{V}_n^+ - \mathbf{V}_n^-$ with $\mathbf{V}_n^+ = \max(\mathbf{V}_n, \mathbf{0})$ and $\mathbf{V}_n^- = \max(-\mathbf{V}_n, \mathbf{0})$, Then, the right side of (47) can be written as:

$$\mathcal{L}_{w_{nik}} = -\sum_{n=1}^{N}\sum_{i=1}^{I}\left[Q(\mathbf{w}_{ni}^T|\hat{\mathbf{w}}_{ni}^T) + \gamma\left[\hat{\mathbf{w}}_{ni}\mathbf{V}_n\hat{\mathbf{w}}_{ni}^T + 2\Delta\hat{\mathbf{w}}_{ni}\mathbf{V}_n\hat{\mathbf{w}}_{ni}^T + \Delta\hat{\mathbf{w}}_{ni}\boldsymbol{\Omega}(\hat{\mathbf{w}}_{ni}^T)\Delta\hat{\mathbf{w}}_{ni}^T\right]\right] \quad (48)$$

with $\boldsymbol{\Omega}(\mathbf{w}_{ni}^T) = \mathrm{Diag}(2\frac{[\mathbf{V}^+\mathbf{w}_{ni}^T + \mathbf{V}^-\mathbf{w}_{ni}^T]}{[\mathbf{w}_{ni}^T]})$, where the operator "$\frac{[\mathbf{A}]}{[\mathbf{B}]}$" is the component-wise division between $\mathbf{A}$ and $\mathbf{B}$.

We let the partial derivative of $\mathcal{L}_{w_{nik}}$ equal to zero and derive the MU update rule of the factor w_{nik} as follows:

$$\frac{\partial\mathcal{L}_{w_{nik}}}{\partial w_{nik}} = \left(\sum_j \frac{h_{nkj}}{\hat{y}_{ijn}} - \sum_j h_{nkj}\frac{\hat{w}_{nik}^2 y_{ijn}}{w_{nik}^2\hat{y}_{ijn}^2} + 2\gamma[\mathbf{V}_n\hat{\mathbf{w}}_{ni}^T]_k + 2\gamma\left[\mathrm{Diag}\left(\frac{\mathbf{V}^+\hat{\mathbf{w}}_{ni}^T + \mathbf{V}^-\hat{\mathbf{w}}_{ni}^T}{\hat{\mathbf{w}}_{ni}^T}\right)\right]_k w_{nik} - 2\gamma\left[\mathrm{Diag}\left(\frac{\mathbf{V}^+\hat{\mathbf{w}}_{ni}^T + \mathbf{V}^-\hat{\mathbf{w}}_{ni}^T}{\hat{\mathbf{w}}_{ni}^T}\right)\right]_k \hat{w}_{nik} \quad (49)$$

To make the above objective function easier, we let

$$a = 2\gamma\left[\mathrm{Diag}\left(\frac{\mathbf{V}^+\hat{\mathbf{w}}_{ni}^T + \mathbf{V}^-\hat{\mathbf{w}}_{ni}^T}{\hat{\mathbf{w}}_{ni}^T}\right)\right]_k \quad (50)$$

$$b = -\sum_j \frac{h_{nkj}}{\hat{y}_{ijn}} + 2\gamma[\mathbf{V}\hat{\mathbf{w}}_{ni}^T]_k - 2\gamma\left[\mathrm{Diag}\left(\frac{\mathbf{V}^+\hat{\mathbf{w}}_{ni}^T + \mathbf{V}^-\hat{\mathbf{w}}_{ni}^T}{\hat{\mathbf{w}}_{ni}^T}\right)\right]_k \hat{w}_{nik} \quad (51)$$

$$d = \sum_j h_{nkj}\frac{\hat{w}_{nik}^2 y_{ijn}}{\hat{y}_{ijn}^2} \quad (52)$$

Setting the derivative to zero is equivalent to finding the roots of the following degree-three polynomial:

$$aw_{nik}^3 + bw_{nik}^2 + d = 0 \quad (53)$$

$$w_{nik} \leftarrow \max(0, w_{nik}) \quad (54)$$

Similar to (32), we use the cubic roots procedure [45] to solve the above polynomial problem.

Similar to m-MNMF, given $\mathbf{W}_n$ and $\mathbf{G}_{ni}$ fixed, the closed-form MU rules for $\mathbf{H}_n$ is:

$$a^h = \sum_{i=1}^{I} w_{nik}|y_{ijn}|^2\lambda_{ijn}^{-2} \quad (55)$$

$$b^h = \sum_{i=1}^{I} w_{nik}\lambda_{ijn}^{-1} \quad (56)$$

$$h_{nkj} \leftarrow h_{nkj}^{\star}\sqrt{\frac{a_h}{b_h}} \quad (57)$$

Given $\mathbf{W}_n$ and $\mathbf{H}_n$ fixed, an IVA-based auxiliary function [46] is used to optimize the spatial model $\mathbf{G}_{ni}$, which results in the following solution:

$$\mathbf{G}_{ni} = \frac{1}{J}\sum_j \frac{1}{\lambda_{ijn}}\mathbf{x}_{ij}\mathbf{x}_{ij}^h$$
$$\mathbf{d}_{in} \leftarrow (\mathbf{D}_i\mathbf{G}_{ni})^{-1}\mathbf{e}_m$$
$$\mathbf{d}_{in} \leftarrow \mathbf{d}_{in}(\mathbf{d}_{in}^h\mathbf{G}_{ni}\mathbf{d}_{in})^{-\frac{1}{2}} \quad (58)$$

The above algorithm is summarized in Algorithm 2.

D. On the Hyper-Parameter Selection and Estimation

The objectives (17) and (45) have two hyper-parameters η and γ.

The hyper-parameter η in the objectives (17) and (45) is a small positive constant that prevents the term $\log|\mathbf{W}_n^T\mathbf{W}_n + \eta\mathbf{I}|$ from $-\infty$. It should not be chosen too small, otherwise $\mathbf{W}_n^T\mathbf{W}_n + \eta\mathbf{I}$ might be badly conditioned which results in the optimization problems hard to solve.

The regularization coefficient γ strongly affects the model performance. Here we have to select an appropriate value for γ. First, the variables $\hat{\mathbf{X}}_{ij}$ and $\mathbf{W}_n$ are initialized with the

Algorithm 2: m-ILRMA.

Input : Mixture $\mathbf{x}_{ij}$, number of sources N, *MaxIteration*, hyperparameter $\eta \geq 0$.
Output: Separated signal $\mathbf{y}_{ij}$.

Initialize: $\mathbf{W}_n$, $\mathbf{H}_n$, $\mathbf{G}_{ni}$;
for *iteration = 1 to MaxIteration* **do**
  for *n=1 to N* **do**
    for *i=1 to I* **do**
      for *k=1 to K* **do**
        Update w_{nik} by solving (53) and (54);
        Update h_{nkj} by (57);
      end
    end
    for *i=1 to I* **do**
      Update $\mathbf{d}_{in}$ by (58);
    end
  end
end
$y_{ij,n} \leftarrow \mathbf{d}_{in}^{h}\mathbf{x}_{ij}$

successive nonnegative projection algorithm [44], then γ is chosen by:

$$\gamma = \gamma^{\star}\frac{\sum_{i,j}\left[\mathrm{tr}\left(\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}\right) + \log|\hat{\mathbf{X}}_{ij}|\right]}{\log|\mathbf{W}_n^T\mathbf{W}_n + \delta\mathbf{I}|} \tag{59}$$

where we recommend to select $\gamma^{\star}$ from a range of $[10^{-3}, 1]$.

IV. Experiments

In this section, we compare the proposed m-MNMF and m-ILRMA methods with 5 representative multichannel BSS methods in both simulated environments and real-world environments.

A. Experimental Settings

1) Simulated Databases: We used SISEC2011 [48] as the first experimental dataset. SISEC2011 consists of three subsets named *dev1*, *dev2* and *dev3*. For the speech separation problem, the clean speech was used to construct an underdetermined BSS task. After mixing up, 192 stereo mixture signals with female and male speech were generated, where the microphone spacing is 1 m or 5 cm and the reverberation time is 130 ms or 250 ms. For the music separation problem, we used non-percussive music sources and the music sources including drums in the *dev1* and *dev2* datasets, which has 12 music mixtures in total, where the experiments setting is consistent with that of the speech separation problem.

Then, we used SISEC2018 [49] as the second experimental dataset. Specifically, we used the clean speech in the *asynchronous recordings of speech mixtures* of SISEC2018 [49] as the speech source. After mixing up, SISEC2018 includes 72 mixture signals (*dev* and *dev2* datasets) with female and male speech, where the microphone spacing is 2.15 cm or 7.65 cm and the reverberation time is 150 ms or 300 ms.

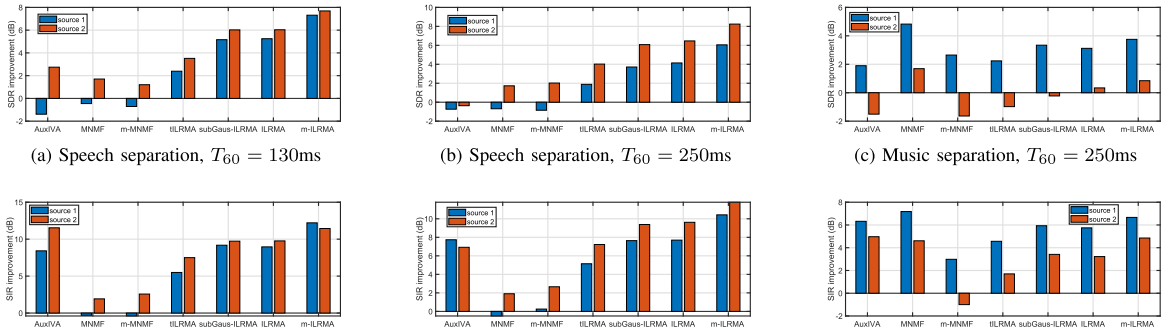
In our experiments, we followed the environment of the SISEC challenge [48] to construct a determined multichannel speech separation task with the number of channels $M = 2$ and number of speakers per mixture $N = 2$.

Besides the above two test corpora, we also followed the environment of the SISEC challenge [48] to construct a determined multichannel speech separation task with $M = N = 2$, where we used the Wall Street Journal (WSJ0) corpus [50] as the clean speech source. We evaluated the comparison methods on all gender combinations. We generated two test conditions for this test corpus, denoted as condition 1 and condition 2. In both conditions, the room size was set to $6 \times 6 \times 3$ m; the two speakers were positioned 2 m from the center of the two microphones. The differences between the two conditions are that (i) the distance between the two microphones are 5.66 cm and 2.83 cm respectively, and (ii) the incident angles of the two speakers follow [4, Figs. 9 a and 9b]. The image source model [51] was used to generate the room impulse response with the reverberation time T_{60} selected from $[130, 150, 200, 250, 300, 350, 400, 450, 500]$ ms. For each condition, we generated 200 mixtures for each gender combination at each T_{60}, which amounts to 7200 mixtures. The sampling rate was set to 16 kHz. We named the simulated data without reverberation as *WSJ0-anechoic*, and the simulated data with reverberation as *WSJ0-reverb*.

2) Semi-Real Database: As shown in Fig. 5, we also conducted an experiment in a real world environment. Specifically, we used a circular array of 48 equiangular-placed loudspeakers with a radius of 1 m and a height of 1 m to produce a desired sound field. Then, we transcribed SISEC2011 as shown in Fig. 5. A linear array of 8 microphones, indexed as mic1 to mic8 from left to right, was placed at the center of the circular loudspeaker array. The target speech was located at the 45° and 120° of the linear array respectively. We named this semi-real database as *semi-real-SISEC2011*.

3) Comparison Algorithms: The hyperparameters of m-MNMF and m-ILRMA in all experiments were set as follows: $K = 10$, $\eta = 0.5$, and the number of iterations was set to 100. The frame-length and frame-shift were set to 1024 and 512, respectively. We compared m-MNMF and m-ILRMA with AuxIVA and four NMF-based multichannel BSS methods which are described as follows:

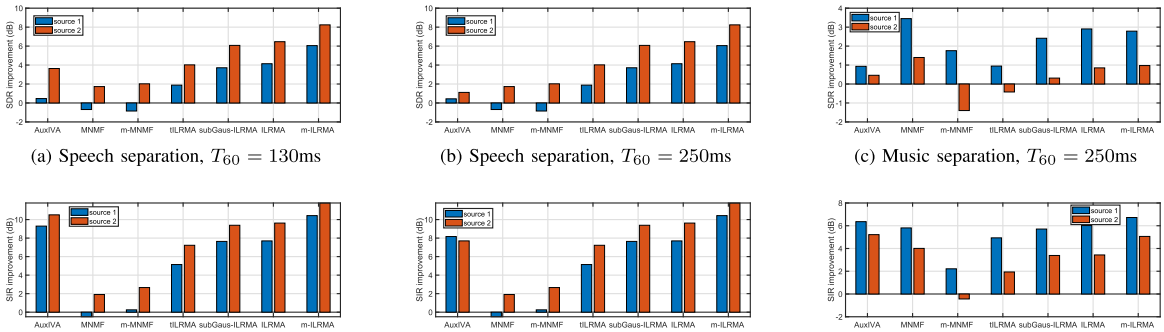
- Auxiliary-function-based Independent Vector Analysis (AuxIVA) [46]: It introduces an auxiliary function for IVA, which is solved by a stable and fast update rule.
- Multichannel Nonnegative Matrix Factorization (MNMF) [2]: It is modeled by the spatial covariance of a zero-mean multivariate Gaussian distribution. It can be considered as a natural extension of NMF, since the Hermitian positive semi-definite is utilized as a multichannel counterpart of nonnegativity. The number of basis vectors K was set to 10 as default.
- Independent Low-Rank Matrix Analysis (ILRMA) [3]: It is a unification of IVA and NMF, which assumes both the statistical independence between sources and a low-rank time-frequency structure for each source. The demixing systems of ILRMA are estimated without encountering the permutation problem. The iteration was set to 100. The number of basis vectors K was set to 10.



(a) Speech separation, $T_{60} = 130$ms (b) Speech separation, $T_{60} = 250$ms (c) Music separation, $T_{60} = 250$ms

(d) Speech separation, $T_{60} = 130$ms (e) Speech separation, $T_{60} = 250$ms (f) Music separation, $T_{60} = 250$ms

Fig. 3. Performance of the comparison methods on SISEC2011 when the distance between the sources and the microphones is 1 m.



(a) Speech separation, $T_{60} = 130$ms (b) Speech separation, $T_{60} = 250$ms (c) Music separation, $T_{60} = 250$ms

(d) Speech separation, $T_{60} = 130$ms (e) Speech separation, $T_{60} = 250$ms (f) Music separation, $T_{60} = 250$ms

Fig. 4. Performance of the comparison methods on SISEC2011, where the distance between source and microphone is 5 cm.

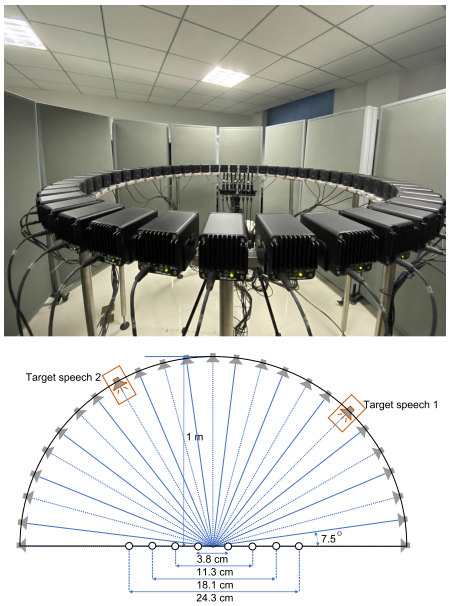


Fig. 5. Recording conditions of impulse responses.

- T-distribution for Independent Low-Rank Matrix Analysis (tILRMA) [14]: It generalizes the source generative model of ILRMA from the complex Gaussian distribution to a complex Student's t-distribution, which is expected to further improve the performance as well as the stability of the parameter initialization. In our experiment, we set the hyperparameters $\nu = 1000$ and $\rho = 10$ respectively.
- Sub-Gaussian Independent Low-Rank Matrix Analysis (subGaus-ILRMA) [18]: The generalization of subGaus-ILRMA is similar to t-ILRMA. SubGaus-ILRMA differs from t-ILRMA in that the distribution of its source generative model is a generalized Gaussian distribution. We set the hyperparameters $\beta = 1.99$ and $\rho = 0.5$ respectively.

4) Evaluation Metrics: We used the source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) [52] to evaluate the *quality* of the separated speech, which are defined as follow:

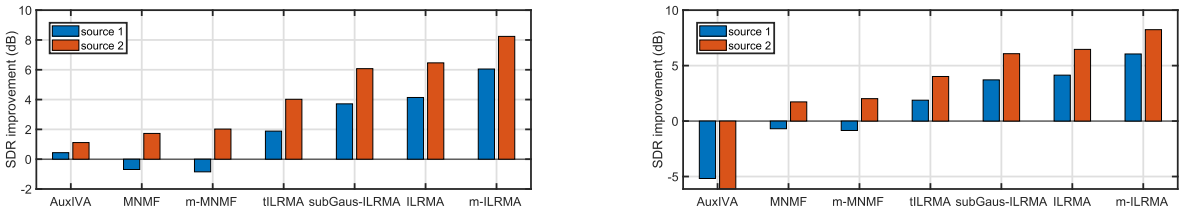
$$\mathrm{SDR} := 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}}\|^2} \tag{60}$$

$$\mathrm{SIR} := 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}}\|^2} \tag{61}$$

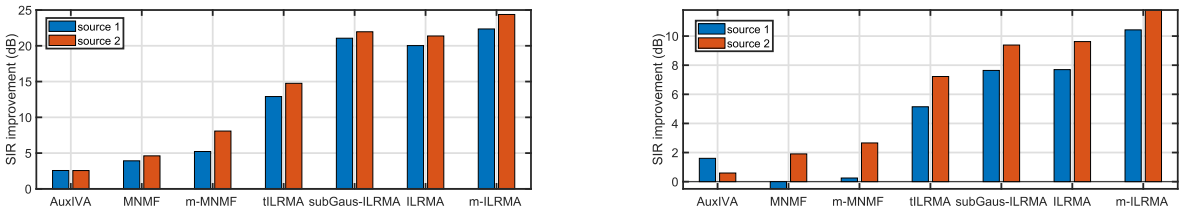
where s_{target} is a version of the wanted source modified by an allowed distortion, and e_{interf}, e_{noise}, and e_{artif} are respectively the interferences, noise, and artifacts error terms.

B. Main Results

1) Results on SISEC2011: The comparison results on SISEC2011 are summarized in Figs. 3 and 4. Specifically, Figs. 3(a), 4(a) and Figs. 3(b), 4(b) show the SDR scores of the comparison methods on the speech separation problem with the reverberation time of 130 ms and 250 ms respectively.

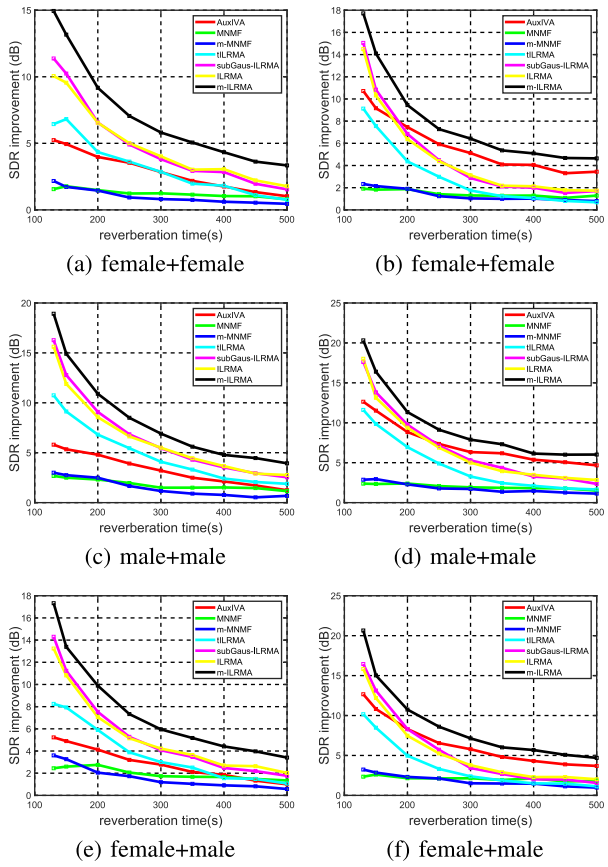


(a) Speech separation, $T_{60} = 150$ms (b) Speech separation, $T_{60} = 300$ms



(c) Speech separation, $T_{60} = 150$ms (d) Speech separation, $T_{60} = 300$ms

Fig. 6. Performance of the comparison methods on SISEC2018.



(a) female+female (b) female+female

(c) male+male (d) male+male

(e) female+male (f) female+male

Fig. 7. SDR improvement of the comparison methods on WSJ0-reverb. (a), (c), (e) are the results in condition 1. (b), (d), (f) are the results in condition 2.

Figs. 3(d), 3(e) and Figs. 4(d), 4(e) show the corresponding SIR scores of the comparison methods. From the figures, we see that the performance of the proposed m-ILRMA is significantly better than the other methods. For example, it achieves an SDR improvement of about 2 dB higher than the best baselines, i.e. ILRMA and subGaus-ILRMA, in both of the test environments.

TABLE I
AVERAGE SDR IMPROVEMENT (DB) OF THE COMPARISON METHODS OVER DIFFERENT REVERBERATION TIME ON WSJ0-REVERB

Methods	Condition 1			Condition 2		
	f+f	m+m	f+m	f+f	m+m	f+m
AuxIVA [46]	2.98	3.40	2.95	5.92	7.55	7.60
MNMF [2]	1.25	1.84	1.97	1.47	2.00	2.11
t-ILRMA [14]	3.30	5.10	3.95	3.29	4.95	3.92
subGaus-ILRMA [18]	5.13	7.08	5.81	5.27	7.40	6.14
ILRMA [3]	5.03	6.89	5.72	5.17	7.31	6.00
m-MNMF	1.05	1.55	1.69	1.37	1.87	1.90
m-ILRMA	7.39	8.77	7.87	8.31	10.06	9.29

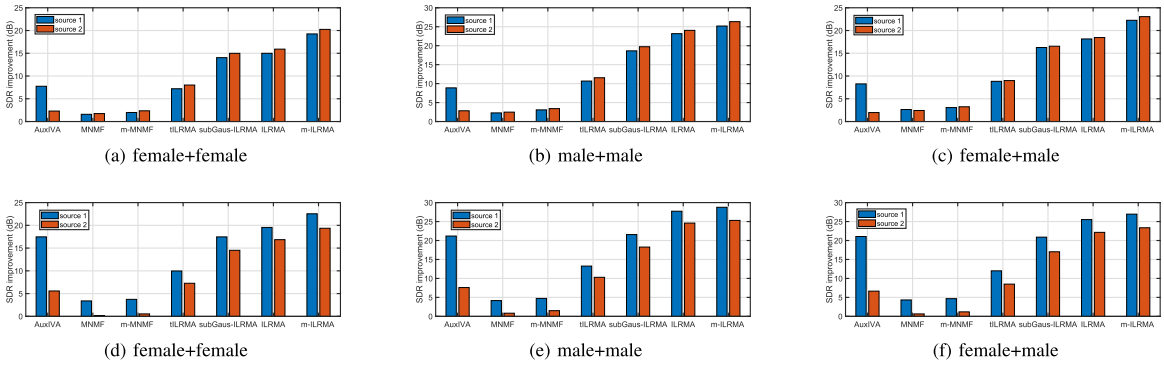
TABLE II
AVERAGE SIR IMPROVEMENT (DB) OF THE COMPARISON METHODS OVER DIFFERENT REVERBERATION TIME ON WSJ0-REVERB

Methods	Condition 1			Condition 2		
	f+f	m+m	f+m	f+f	m+m	f+m
AuxIVA [46]	10.09	11.86	10.20	12.19	14.58	13.69
MNMF [2]	1.58	2.34	2.57	1.87	2.59	2.76
t-ILRMA [14]	6.02	8.35	6.91	5.80	8.00	6.69
subGaus-ILRMA [18]	8.10	10.55	8.96	7.96	10.75	9.06
ILRMA [3]	7.83	10.11	8.67	7.65	10.31	8.68
m-MNMF	1.60	2.33	2.60	2.17	2.85	2.96
m-ILRMA	10.80	12.63	11.47	11.69	14.06	12.98

Figs. 3(c), 3(f) and Figs. 4(c), 4(f) show the comparison result on the music separation problem. From the figures, we see that m-ILRMA achieves better performance than the other methods except MNMF.

2) Results on SISEC2018: Fig. 6 shows the comparison results on speech separation in terms of the average SDR and SIR improvement. From the figure, we see that m-MNMF outperforms MNMF, and m-ILRMA outperforms ILRMA, which demonstrate the effectiveness of the proposed MinVol prior for the multichannel BSS.

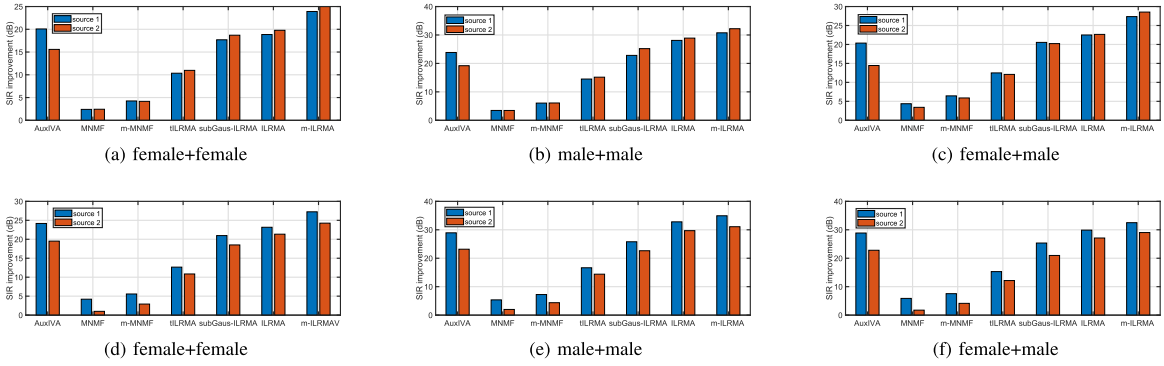
3) Results on semi-Real-SISEC2011: Tables III and IV show the separation performance of the comparison methods on the real-world recording environment of semi-real-SISEC2011.



(a) female+female (b) male+male (c) female+male

(d) female+female (e) male+male (f) female+male

Fig. 8. SDR improvement of the comparison methods on the WSJ0-anechoic corpus. (a), (b), (c) are the results in condition 1. (d), (e), (f) are the results in condition 2.



(a) female+female (b) male+male (c) female+male

(d) female+female (e) male+male (f) female+male

Fig. 9. SIR improvement of the comparison methods on the WSJ0-anechoic corpus. (a), (b), (c) are the results in condition 1. (d), (e), (f) are the results in condition 2.

TABLE III
SDR IMPROVEMENT (DB) OF THE COMPARISON METHODS ON SEMI-REAL SISEC2011

Methods	mic4 / mic5	mic3 / mic6	mic2 / mic7	mic1 / mic8
AuxIVA [46]	2.59 / 2.41	0.45 / 3.04	-0.01 / 2.79	1.84 / 1.25
MNMF [2]	-0.12 / 1.72	-1.34 / 1.17	-2.06 / 1.51	-0.22 / -0.46
t-ILRMA [14]	2.81 / 3.46	0.38 / 2.97	0.10 / 3.69	1.76 / 1.60
subG-ILRMA [18]	3.43 / 4.63	0.75 / 3.42	0.73 / 4.58	1.90 / 1.86
ILRMA [3]	4.20 / 5.16	1.34 / 3.94	0.87 / 4.62	2.94 / 2.62
m-MNMF	-0.99 / 0.91	-2.16 / 0.33	-2.75 / 0.91	-0.81 / -1.04
m-ILRMA	5.98 / 6.90	1.50 / 4.53	2.57 / 7.04	3.87 / 4.35

TABLE IV
SIR IMPROVEMENT (DB) OF THE COMPARISON METHODS ON SEMI-REAL SISEC2011

Methods	mic4 / mic5	mic3 / mic6	mic2 / mic7	mic1 / mic8
AuxIVA [46]	10.81 / 9.93	8.49 / 8.74	8.65 / 9.41	11.09 / 7.94
MNMF [2]	0.60 / 2.65	-0.15 / 3.43	-0.74 / 3.86	1.17 / 1.90
t-ILRMA [14]	7.16 / 6.47	4.26 / 5.90	4.18 / 6.70	6.11 / 4.65
subG-ILRMA [18]	8.38 / 8.12	5.09 / 6.74	5.41 / 8.26	6.65 / 5.24
ILRMA [3]	8.64 / 8.05	5.40 / 6.86	5.61 / 7.79	7.62 / 5.55
m-MNMF	-0.23 / 1.42	-0.92 / 2.11	-1.38 / 2.76	0.64 / 0.84
m-ILRMA	12.47 / 10.71	5.78 / 8.06	8.53 / 10.93	9.40 / 8.25

From Table III, we see that the SDR improvement of m-ILRMA is 2 dB higher than that of ILRMA on average in all four situations. From Table IV, we see that the SIR improvement of m-ILRMA is competitive with the best comparison method.

V. RESULTS ON WSJ0-ANECHOIC AND WSJ0-REVERB

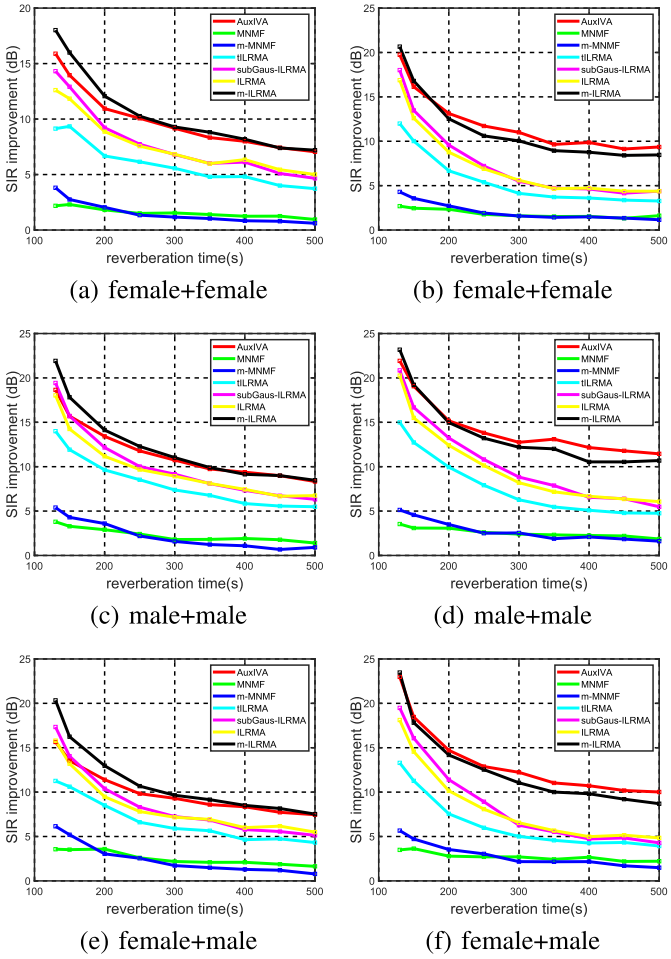
A. Results on WSJ0-Anechoic

Figs. 8 and 9 show respectively the average SDR and SIR improvement of the comparison methods over the mixed speech in the simulated anechoic environment of WSJ0-anechoic. From the figures, we see that the performance of the proposed m-ILRMA is significantly better than that of the other methods. For example, m-ILRMA achieves an SDR improvement of about 3 dB higher than the best reference method, i.e. ILRMA.

B. Results on WSJ0-Reverb

Figs. 7 and 10 show the SDR and SIR improvement respectively over the mixed speech in the simulated reverberant environment of WSJ0-reverb. From the figures, we see that the curves of the SDR improvement produced by m-ILRMA are always higher than those produced from the comparison methods. The minimum improvement of m-ILRMA over the comparison methods is 2 dB.

To clearly show the general improvement of m-ILRMA over the referenced methods, we average the SDR improvement with respect to different gender combinations and T_{60} for each condition. The average results are listed in Tables I and II, respectively. From the tables, we see that the average SDR improvement brought by the proposed m-ILRMA is 2 dB higher than ILRMA in condition 1, and 3 dB higher than the latter in condition 2.



(a) female+female (b) female+female

(c) male+male (d) male+male

(e) female+male (f) female+male

Fig. 10. SIR improvement of the comparison methods on WSJ0-reverb. (a), (c), (e) are the results in condition 1. (b), (d), (f) are the results in condition 2.

The average SIR improvement of m-ILRMA is comparable to AuxIVA, and outperforms the other methods.

C. Discussion

In this section, we demonstrate the effectiveness of the MinVol prior on the sparsity, orthogonality, and uniqueness of the spectra matrix by comparing ILRMA with m-ILRMA. Before analysis, we first define the sparsity, orthogonality, and uniqueness of a matrix as follows:

Define 1: Sparseness measurement [53]: The sparseness of a matrix is built on the relationship between the L_1 norm and the L_2 norm:

$$\zeta(\mathbf{w}_k) = \frac{\sqrt{n} - (\sum_i |w_{ik}|)/\sqrt{\sum_i w_{ik}^2}}{\sqrt{n} - 1} \tag{62}$$

$$\hat{\zeta}(\mathbf{W}) = \frac{1}{K}\sum_k \zeta(\mathbf{w}_k) \tag{63}$$

where $\mathbf{w}_k = [w_{1k}, \ldots, w_{ik}, \ldots, w_{Ik}]^T$ is the kth column of the matrix $\mathbf{W}$, $\zeta(\mathbf{w}_k)$ calculates the sparseness of the vector $\mathbf{w}_k$, and $\hat{\zeta}(\mathbf{W}_n)$ defines the sparseness of the matrix $\mathbf{W}_n$.

The higher the sparsity score is, the stronger the part-based representation ability of the matrix $\mathbf{W}$ is.

Define 2: Orthogonality measurement [54], [55]: Two nonnegative vectors are orthogonal if and only if they do not have the same non-zero elements, and we measure the orthogonality of a matrix by:

$$\text{Orthogonality}(\mathbf{W}) = \|\mathbf{W}^T\mathbf{W} - \mathbf{I}\| \tag{64}$$

where $\mathbf{I}$ is an identity matrix.

The lower the orthogonality score is, the stronger the orthogonality between the basis vectors of the matrix $\mathbf{W}$ is.

Define 3: Uniqueness measurement [56]: Assume that $\mathbf{T}$ can be approximated by $\mathbf{T} \approx \mathbf{WH}$. In the ideal case, we have $\mathbf{T} = \mathbf{W}'\mathbf{H}$. When the equality holds, we have $\mathbf{W}' = \mathbf{TH}^{-1}$. However, the equality relation can hardly be achieved in practice. Therefore, the closer the two different solutions $\mathbf{W}$ and $\mathbf{W}'$ are to degeneracy, the better the unique solution of the source model $\mathbf{T}$ is. Here, we use the squared Frobenius norm to measure their difference:

$$\text{Dif}(\mathbf{W}, \mathbf{W}') = \|\mathbf{W} - \mathbf{W}'\|_F^2 \tag{65}$$

The lower the uniqueness score is, the stronger the identifiability of the matrix $\mathbf{W}$ is.

To analyze the sparsity, orthogonality and uniqueness of the spectra matrix generated by ILRMA and m-ILRMA, we averaged the results of 50 spectra matrices in terms of the three measurements. Before calculating (65), we first measured the cosine similarity of the basis vectors in the basis matrix $\mathbf{W}$ to alleviate their permutation problem, then changed the order of the basis vectors, and finally normalized the basis vectors to prevent their scale ambiguity. The results are that (i) the sparsity scores of ILRMA and m-ILRMA are 0.65 and 0.68 respectively, (ii) the orthogonality scores are 0.99 and 0.74 respectively, and (iii) the uniqueness scores are 68.53 and 2.73, respectively. The results show that the spectral matrix of m-ILRMA has stronger representation ability than that of ILRMA, which proves the effectiveness of the MinVol prior for the multichannel BSS. Fig. 11 shows the SDR improvement of the speech signals separated from 2 channel mixtures with different initialization parameters. From the figure, we further see the advantages of m-ILRMA. We also observe that the proposed methods are insensitive to the initialization parameters.

VI. Conclusion

In this paper, we have proposed a MinVol prior for the source model of multichannel BSS methods. To our knowledge, this is the first MinVol prior regularized multichannel BSS model. The novelty of the MinVol prior lies in the following aspect. First, we propose a novel MinVol prior distribution for the source model which improves the identifiability, sparseness, and orthogonality of the separated spectrograms produced from the source model. It performs as a regularization of the source model in the objective functions of the multichannel BSS. To evaluate its effectiveness, we implement two multichannel MinVol-based BSS algorithms, denoted as m-MNMF and

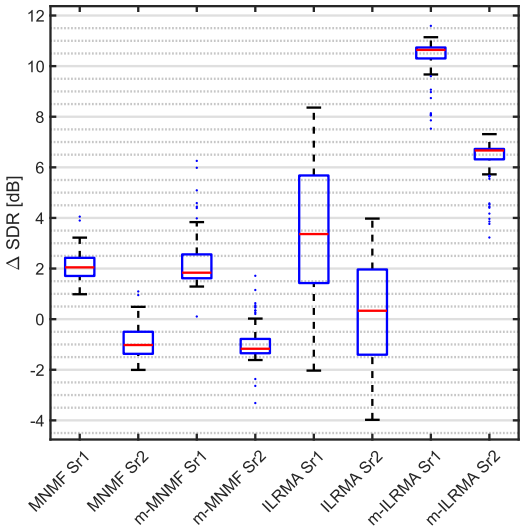


Fig. 11. SDR improvement of the separated speech signals over the original 2-channel mixtures. The central lines indicate the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The experimental settings are described in Section IV-A3, where the number of iterations was set to 100.

m-ILRMA. The optimization of the two proposed methods is intractable since that the objective functions contain logarithmic determinant terms. To overcome this problem, we relax the logarithmic determinant terms with their tightened lower bounds. Finally, we apply multiplicative update rules to solve the optimization problems. We have conducted an extensive experimental comparison with five representative comparison methods on four simulated datasets and a real dataset, which are SISEC2011, SISEC2018, WSJ0-anechoic, WSJ0-reverb, and semi-real-SISEC2011, respectively. Experimental results show that the proposed m-ILRMA outperforms the comparison methods significantly in terms of SDR and SIR. Although m-MNMF does not reach the top performance, it performs better than its counterpart MNMF. Moreover, we analyzed the identifiability, sparseness, and orthogonality of the spectral matrix produced by ILRMA and m-ILRMA. The results show that the spectral matrix of m-ILRMA has stronger representation ability than that of ILRMA, which proves the effectiveness of the MinVol prior for the multichannel BSS.

REFERENCES

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

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

[3] 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.

[4] S. Mogami *et al.*, "Independent low-rank matrix analysis based on time-variant sub-Gaussian source model for determined blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 12, pp. 503–518, Dec. 2019.

[5] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM J. Optim.*, vol. 20, no. 3, pp. 1364–1377, 2010.

[6] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Robust volume minimization-based matrix factorization via alternating optimization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 2534–2538.

[7] V. Leplat, N. Gillis, and A. M. S. Ang, "Blind audio source separation with minimum-volume beta-divergence NMF," *IEEE Trans. Signal Process.*, vol. 68, no. 5, pp. 3400–3410, May 2020.

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

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

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*. Berlin, Germany: Springer, 2018, pp. 125–155.

[13] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.

[14] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex student's t-distribution for blind audio source separation," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[15] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.

[16] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 51–55.

[17] D. Kitamura *et al.*, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, pp. 1–25, 2018.

[18] R. Ikeshita and Y. Kawaguchi, "Independent low-rank matrix analysis based on multivariate complex exponential power distribution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 741–745.

[19] K. Kamo *et al.*, "Joint-diagonalizability-constrained multichannel nonnegative matrix factorization based on multivariate complex sub-Gaussian distribution," *Signal Process.*, vol. 188, Aug. 2020, Art. no. 108183.

[20] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[21] 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, no. 8, pp. 2610–2625, Aug. 2020.

[22] K. Kamo *et al.*, "Regularized fast multichannel nonnegative matrix factorization with ILRMA-based prior distribution of joint-diagonalization process," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 606–610.

[23] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 493–504, Feb. 2014.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[25] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, and K. Yoshii, "A unified Bayesian model of time-frequency clustering and low-rank approximation for multi-channel source separation," in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 2280–2284.

[26] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Bayesian multichannel nonnegative matrix factorization for audio source separation and localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 551–555.

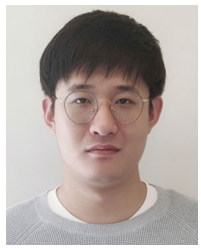
[27] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Bayesian multichannel audio source separation based on integrated source and spatial models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 831–846, Apr. 2018.

[28] 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.
[29] C. Narisetty, T. Komatsu, and R. Kondo, "Bayesian non-parametric multi-source modelling based determined blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 111–115.
[30] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," 2018, *arXiv:1808.00892*.
[31] 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.
[32] N. Makishima *et al.*, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, Oct. 2019.
[33] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.
[34] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, 2019.
[35] L. Li, H. Kameoka, and S. Makino, "Determined audio source separation with multichannel star generative adversarial network," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process.*, 2020, pp. 1–6.
[36] 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.
[37] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Underdetermined source separation based on generalized multichannel variational autoencoder," *IEEE Access*, vol. 7, pp. 168104–168115, 2019.
[38] L. Li, H. Kameoka, and S. Makino, "Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 546–550.
[39] M. Togami, "Multi-channel Itakura Saito distance minimization with deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 536–540.
[40] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 211–224, Jan. 2014.
[41] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 328–332, Mar. 2018.
[42] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, Mar. 2019.
[43] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May 2015.
[44] V. Leplat, A. M. Ang, and N. Gillis, "Minimum-volume rank-deficient nonnegative matrix factorizations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 3402–3406.
[45] E. Rechtschaffen, "92.35 real roots of cubics: Explicit formula for quasi-solutions," *Math. Gazette*, vol. 92, no. 524, pp. 268–276, 2008.
[46] 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.
[47] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, 2011.
[48] S. Araki *et al.*, "The 2011 signal separation evaluation campaign (SISEC2011):-audio source separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Berlin, Germany, 2012, pp. 414–422.
[49] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Berlin, Germany, 2018, pp. 293–305.
[50] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," *Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.
[51] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[52] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[53] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, no. 9, pp. 1457–1469, 2004.
[54] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. IEEE Int. Joint Conf. Neural Netw. IEEE World Congr. Comput. Intell.*, 2008, pp. 1828–1832.
[55] Z. Yuan, Z. Yang, and E. Oja, "Projective nonnegative matrix factorization: Sparseness, orthogonality, and clustering," *Neural Process. Lett.*, pp. 11–13, 2009.
[56] F. J. Theis, K. Stadlthanner, and T. Tanaka, "First results on uniqueness of sparse non-negative matrix factorization," in *Proc. 13th Eur. Signal Process. Conf.*, 2005, pp. 1–4.

Jianyu Wang received the B.E. degree in information countermeasure technology from Northwestern Polytechnical University, Xi'an, China, in 2018, where he is currently working toward the Ph.D. degree in information and communication engineering. In 2018, he joined the Center for Intelligent Acoustics and Immersive Communications.

His research interests include machine learning, audio and speech processing, source separation, and latent variable model.

Shanzheng Guan (Graduate Student Member, IEEE) received the bachelor's degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2015, where he is currently working toward the Ph.D. degree in information and communication engineering. His research interests include deep learning, audio and speech processing, speech enhancement, and array signal processing.



Shupei Liu received the bachelor's degree in 2019 in detection guidance and control techniques from Northwestern Polytechnical University, Xi'an, China, where he is currently working toward the master's degree in signal and information processing. His research interests include sound source localization and speech enhancement.

Xiao-Lei Zhang (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Full Professor with the Center for Intelligent Acoustics and Immersive Communications, and the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China. He was a Postdoctoral Researcher with Perception and Neurodynamics Laboratory, The Ohio State University, Columbus, OH, USA. His research interests include speech processing, machine learning, statistical signal processing, and artificial intelligence. He is a Member of IEEE SPS and ISCA.