Minimum-volume regularized ILRMA for blind audio source separation

Jianyu Wang, Shanzheng Guan and Xiao-Lei Zhang
CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, China
E-mail: {alexwang96, gshanzheng}@mail.nwpu.edu.cn, xiaolei.zhang@nwpu.edu.cn

***Abstract*—Multichannel blind audio source separation aims to recover the latent sources from their multichannel mixture without priors. A state-of-the-art blind audio source separation method called independent low-rank matrix analysis (ILRMA) unified independent vector analysis (IVA) and nonnegative matrix factorization (NMF). However, the spectra matrix produced from NMF may not find a compact representation. Also, the matrix may not guarantee that each source is identifiable. To address the problem, here we propose a modified blind audio source separation method that enhances the identifiability of the source model. It combines ILRMA with a new penalty term, named *minimum volume regularization* The proposed method is optimized by standard majorization-minimization framework based multiplication updating rule, which ensures the stability of convergence. Experimental results demonstrate the effectiveness of the proposed method compared with AuxIVA, MNMF and ILRMA.**

I. Introduction

Blind audio source separation separates a mixture of multiple sources into their components without prior information of the recording environments, mixing system, or source locations [1], [2], [3]. A typical approach to blind audio source separation is based on unsupervised learning of a probabilistic model. It can be categorized into single-channel source separation and multichannel source separation. This paper focuses on multichannel source separation. A multichannel source separation method usually consists of a source model representing the time-frequency structure of source images and a spatial model representing their inter-channel covariance structure. A widely used source model is the low-rank model based on nonnegative matrix factorization (NMF) for mitigating the permutation problem. The time-frequency bins of each source in the spatial model are usually assumed to be multivariate complex Gaussian [4].

A representative of multichannel source separation is multichannel nonnegative matrix factorization (MNMF) [5], [6], [7], [8]. 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 at bad local optima since a large number of unconstrained spatial covariance matrices need to be estimated iteratively. To address this problem, Kitamura *et al.* [9], [10] proposed independent low-rank matrix analysis (ILRMA) which makes rank-1 assumption for 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) [11] and NMF respectively, which are optimized iteratively. The aforementioned NMF-based methods, e.g. MNMF, ILRMA [10] and its variants [4] use NMF to decompose a given spectrogram into several spectral bases and temporal activations. Although the spatial properties of the source images constrain the bases of NMF for the uniqueness of the decomposition, it may not guarantee that the spectral content of each source is identifiable. Therefore, a good source model has the potential to improve the source separation performance [10].

To improve the source identifiability of separation algorithms, here we propose a new geometric inference method for MNMF, named *MinVol*. It penalizes the columns of the spectral bases of NMF by volume minimization [12], [13], so that their convex hull has a small volume. Volume minimization factorizes a given data matrix into a basis matrix and a structured coefficient matrix by finding a minimum-volume simplex that encloses all columns of the data matrix [14]. It guarantees the identifiability of the factorized matrices under a so-called *sufficiently scattered condition* [15], [16]. We associate the minimum-volume penalty with the Itakura-Saito (IS) divergence for MNMF. To our knowledge, this is the first time that the minimum-volume penalty is used in MNMF. Also, the minimum-volume constraint implicitly enhances the sparsity of the temporal activations, so that many frequency bands will be located on the facets of the cone of the spectral bases. The proposed *MinVol* method is optimized by a multiplicative update (MU) rule under the standard majorization-minimization framework. Experimental results show that the proposed method outperforms Auxiliary-IVA (AuxIVA) [17], MNMF [6], and ILRMA [10] in speech separation tasks.

II. Methods

A. Problem formulation

Suppose the short-time Fourier transform (STFT) of a multichannel mixture is $\mathbf{x}_{ij} = [x_{ij,1}, \ldots, x_{ij,m}, x_{ij,M}]^T \in \mathbb{C}^M$, 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, and ^{T} denotes the transpose operator. Its source components are denoted as $\mathbf{s}_{ij} = [s_{ij,1}, \ldots, s_{ij,n}, \ldots, s_{ij,N}]^T \in \mathbb{C}^N$, where N is the number of sources and $n = 1, \ldots, N$ is the index of the sources.

We assume that each source of the mixture is a point source,

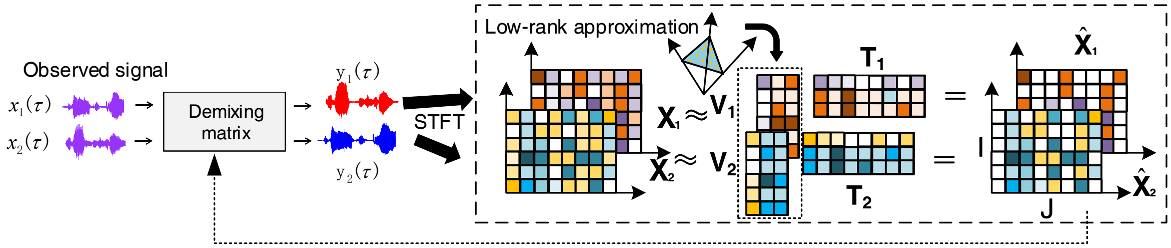


Fig. 1. Principle of the proposed *MinVol* algorithm.

then the mixture and its sources have the following connection:

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

where $\mathbf{A}_i$ is the mixing matrix at the ith frequency bin. If $\mathbf{A}_i$ is invertible and $M = N$, we can find a demixing matrix $(\mathbf{A}_i)^{-1}$ for recovering $\mathbf{s}_{ij}$.

The problem of source separation is to find an estimation of $(\mathbf{A}_i)^{-1}$, denoted as $\mathbf{W}_i = [\mathbf{w}_{i,1}, \ldots, \mathbf{w}_{i,M}]^H$, such that when we apply $\mathbf{W}_i$ to $\mathbf{x}_{ij}$, we obtain the separated signal $\mathbf{y}_{ij}$:

$$\mathbf{y}_{ij} = \mathbf{W}_i\mathbf{x}_{ij} \tag{2}$$

where ^{H} denotes the Hermitian transpose, and $\mathbf{y}_{ij}$ is an estimation of $\mathbf{s}_{ij}$.

Many MNMF methods model the power spectrogram by $\mathbf{X}_{ij} = \mathbf{x}_{ij}\mathbf{x}_{ij}^H$, and use NMF [5], [18], [6] to decompose $\mathbf{X}_{ij}$ by:

$$\mathbf{X}_{ij} \approx \hat{\mathbf{X}}_{ij} = \sum_{k=1}^{K}\left(\sum_{n=1}^{N}\mathbf{R}_{i,n}\right) v_{ik,n}t_{kj,n} \qquad \forall i = 1, \ldots, I,\ \forall j = 1, \ldots, J. \tag{3}$$

where K is the number of basis vectors, $v_{ik,n}$ is the element of a spectral basis matrix $\mathbf{V}_n \in \mathbb{R}_{\geq 0}^{I\times K}$ for the nth source, $t_{kj,n}$ is the element of a temporal activation matrix $\mathbf{T}_n \in \mathbb{R}_{\geq 0}^{K\times J}$ for the nth source, and $\mathbf{R}_{i,n} \in \mathbb{C}^{M\times M}$ is the spatial covariance at the ith frequency band for the nth source. We denote the full representation of $\mathbf{R}_{i,n}$ at all frequency bands for all sources as a tensor $\mathbf{R} \in \mathbb{C}^{I\times N\times M\times M}$, and the full representation of $\mathbf{X}_{i,j}$ at all time-frequency bins as a tensor $\mathbf{X} \in \mathbb{C}^{I\times J\times M\times M}$.

B. Minimum-volume multichannel source separation

Because there exists several valid solutions of $\mathbf{V}_n$ in (3), the decomposition of the source model of MNMF is not unique. To improve the identifiability of ILRMA (see Section II-D for the definition of the identifiability), we propose the minimum-volume based MNMF (*MinVol*). The principle of *MinVol* is shown in Fig. 1. Its objective function is:

$$\min_{\mathbf{R}_{i,n},\mathbf{V}_n,\mathbf{T}_n} \mathcal{L} = \min_{\mathbf{R}_{i,n},\mathbf{V}_n,\mathbf{T}_n} \sum_n \lambda\mathrm{vol}(\mathbf{V}_n) + \sum_{i,j}\ell(\mathbf{X}_{ij}, \hat{\mathbf{X}}_{ij}) \quad \text{subject to} \quad \mathbf{V}_n^T\mathbf{1} = \mathbf{1}, \quad \mathbf{T}_n \geq 0 \tag{4}$$

where $\mathbf{1}$ is an all-one vector and

$$\mathrm{vol}(\mathbf{V}_n) = \log|\det(\mathbf{V}_n^T\mathbf{V}_n + \delta\mathbf{I}_K)| \tag{5}$$

is the minimum-volume regularization with δ as a small positive constant that ensures $\mathrm{vol}(\mathbf{V}_n)$ is bounded from below, unlike the quantity $\log|\det(\mathbf{V}_n^T\mathbf{V}_n)|$. $\mathbf{I}_K$ is the identity matrix with dimensions K, and $\ell(\mathbf{X}_{ij}, \hat{\mathbf{X}}_{ij})$ is the loss of the approximation. λ is the regularization coefficient.

The reason for using the minimum-volume is that minimizing the volume of $\mathbf{V}_n$ makes the columns of $\mathbf{V}_n$ to be as close as possible to each other within the unit simplex. For different assumptions of data distribution, the loss ℓ should be chosen differently. Because we assume that the data is multiplicative Gamma distribution in this paper, we choose the IS divergence as the loss. The IS divergence is the only one in the β divergence family that has the scale-invariant property. It implies that the distribution of the time-frequency bins with low power is as important as that with high power during the divergence computation [19].

C. Optimization algorithm

The objective function $\mathcal{L}$ based on the IS divergence is formulated as:

$$\mathcal{L} = \sum_{i,j}\left[\mathrm{tr}\left(\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}\right) + \log\det\hat{\mathbf{X}}_{ij}\right] + \sum_n \lambda\mathrm{vol}(\mathbf{V}_n) \tag{6}$$

According to ILRMA [10], the spatial covariance $\mathbf{R}_{i,n}$ can be modeled by the rank-1 assumption. With the assumption, (6) can be formulated as:

$$\begin{aligned}\mathcal{L} = \sum_{i,j}\Bigg[&\sum_n \frac{|y_{ij,n}|^2}{\sum_k v_{ik,n}t_{kj,n}} + \sum_n \log\sum_k v_{ik,n}t_{kj,n} \\ &- 2\log|\det\mathbf{W}_i|\Bigg] + \sum_n \lambda\mathrm{vol}(\mathbf{V}_n)\end{aligned} \tag{7}$$

where the term $\sum_{i,j} -2\log|\det\mathbf{W}_i|$ is called the spatial model, and the sum of all other terms are called the source model. The spatial and source models of the objective are optimized iteratively.

For each single iteration, to optimize the spatial model, an IVA-based auxiliary function [17] is used, which results in the following solution:

$$\begin{aligned}\mathbf{G}_{i,n} &= \frac{1}{J}\sum_j \frac{1}{d_{ij,n}}\mathbf{x}_{ij}\mathbf{x}_{ij}^h \\ \mathbf{w}_{i,n} &\leftarrow (\mathbf{W}_i\mathbf{G}_{i,n})^{-1}\mathbf{e}_m \\ \mathbf{w}_{i,n} &\leftarrow \mathbf{w}_{i,n}(\mathbf{w}_{i,n}^h\mathbf{G}_{i,n}\mathbf{w}_{i,n})^{-\frac{1}{2}}\end{aligned} \tag{8}$$

where $\mathbf{e}_m$ denotes the nth column vector of the $M \times M$ identity matrix, $\mathbf{D}_n \in \mathbb{C}^{I\times J\times N}$ is the estimated spectrogram of the nth source, and $d_{ij,n} = \sum_k v_{ik,n}t_{kj,n}$ is the element of $\mathbf{D}_n$.

Substituting the solution (8) into (7) derives the following optimization objective of the nth source model:

$$\mathcal{L}_{nS} = \sum_{i,j} \Big(\frac{d_{ij,n}}{\sum_k v_{ik,n} t_{kj,n}} + \log \sum_k v_{ik,n} t_{kj,n} \Big) + \lambda \text{vol}(\mathbf{V}_n) \tag{9}$$

where each source model is optimized independently as follows.

Because the first term of (9) is a difficult optimization problem, we propose to optimize a new auxiliary function instead of the difficult problem. The design of the auxiliary function follows that in [20]:

Lemma 1 ([20])**.** *Let $\tilde{d}_{ij,n} = \sum_k \tilde{v}_{ik,n} t_{kj,n}$ and $\tilde{d}_{ij,n} \geq 0$, $\tilde{v}_{ik,n} \geq 0$. Then, the function:*

$$Q(\mathbf{v}_{i\cdot,n}|\tilde{\mathbf{v}}_{i\cdot,n}) = \left[\sum_k \frac{\tilde{v}_{ik,n} t_{kj,n}}{\tilde{d}_{ij,n}} \check{\rho}(d_{ij,n}|\tilde{d}_{ij,n} \frac{v_{ik,n}}{\tilde{v}_{ik,n}}) \right] + \bar{\rho}(\tilde{d}_{ij,n}) \\ + \left[\hat{\rho}'(d_{ij,n}|\tilde{d}_{ij,n}) \sum_k (v_{ik,n} - \tilde{v}_{ik,n}) t_{kj,n} + \hat{\rho}(d_{ij,n}|\tilde{d}_{ij,n}) \right] \tag{10}$$

is an auxiliary function for $Q(\mathbf{v}_{i\cdot,n})$ at $\tilde{v}_{ik,n}$. where $\check{\rho}$ is a convex function with respect to $\tilde{d}_{ij,n} \frac{v_{ik,n}}{\tilde{v}_{ik,n}}$, $\hat{\rho}$ is a concave function with respect to $\tilde{d}_{ij,n}$, and $\bar{\rho}$ is the constant of $d_{ij,n}$. $\hat{\rho}'$ is the differential of $\hat{\rho}(d_{ij,n}|\tilde{d}_{ij,n})$ at $\tilde{d}_{ij,n}$. Due to the IS divergence, we have $\check{\rho}(x|y) = xy^{-1}$, $\hat{\rho}(x|y) = \log y$, $\bar{\rho}(x) = x(\log x - 1)$, $\hat{\rho}'(x|y) = y^{-1}$.

Because the second term of (9), i.e. the minimum-volume regularization, is also a difficult optimization problem, we use its first-order Taylor expansion as an approximation which constructs an upper bound of the expansion:

$$\text{vol}(\mathbf{V}_n) \leq \log|\det(\mathbf{U}^{-1})| + \text{tr}(\mathbf{U}\mathbf{V}_n^T\mathbf{V}_n) - K \tag{11}$$

where $\mathbf{U} = (\mathbf{Z}^T\mathbf{Z} + \delta\mathbf{I})^{-1}$ with $\delta \geq 0$, $\mathbf{Z} \in \mathbb{R}^{I \times K}$ is an arbitrary positive definite matrix. We can set $\mathbf{Z} = \mathbf{V}_n$ in the experiments, since $\mathbf{V}_n$ is a positive definite matrix. Finally, the right side of (11) is an auxiliary function for $\text{vol}(\mathbf{V}_n)$. However, it is quadratic and inseparable, which makes the problem hard to optimize over the nonnegative orthant. We use an approximation to represent the right side of (11). The non-constant part can be written as $l(\mathbf{V}_n) = \mathbf{V}_n\mathbf{U}\mathbf{V}_n^T$. Let $\mathbf{U} = \mathbf{U}^+ - \mathbf{U}^-$ with $\mathbf{U}^+ = \max(\mathbf{U}, 0)$ and $\mathbf{U}^- = \max(-\mathbf{U}, 0)$, Then, the right side of (11) can be written as:

$$l(\mathbf{V}_n, \tilde{\mathbf{V}}_n) = \frac{1}{2} \Delta\mathbf{V}_n^T \text{Diag}\left(2\frac{[\mathbf{U}^+\tilde{\mathbf{V}}_n + \mathbf{U}^-\tilde{\mathbf{V}}_n]}{[\tilde{\mathbf{V}}_n]} \right) \Delta\mathbf{V}_n \\ + \Delta\mathbf{V}_n^T \nabla l(\tilde{\mathbf{V}}_n) + l(\tilde{\mathbf{V}}_n) \tag{12}$$

where $\frac{[x]}{[y]}$ is the component division between x and y, $\text{Diag}(\cdot)$ is the diagonal matrix, and $\Delta\mathbf{V}_n = \mathbf{V}_n - \tilde{\mathbf{V}}_n$.

At last, we replace the first term of (9) by (10) and the second term of (9) by (12), which results in the following auxiliary function at $\tilde{\mathbf{V}}_n$:

$$F(\mathbf{V}_n|\tilde{\mathbf{V}}_n) = \lambda \left(\log|\det(\mathbf{U}^{-1})| + \text{tr}(\mathbf{U}\mathbf{V}_n^T\mathbf{V}_n) \right) \\ \sum_i Q(\mathbf{v}_{i\cdot,n}|\tilde{\mathbf{v}}_{i\cdot,n}) + \text{const} \tag{13}$$

where const is a constant for $\mathbf{V}_n$. Similarly with (13), we obtain:

$$F(\mathbf{T}_n|\tilde{\mathbf{T}}_n) = \sum_j Q(\mathbf{t}_{\cdot j,n}|\tilde{\mathbf{t}}_{\cdot j,n}) + \text{const} \tag{14}$$

as an auxiliary function at $\tilde{\mathbf{T}}_n$ for $\mathbf{T}_n$

Setting the derivative of the auxiliary function $F(\mathbf{V}_n|\tilde{\mathbf{V}}_n)$ to zero:

$$\nabla_{v_{ik,n}} F(v_{ik,n}|\tilde{v}_{ik,n}) = \Big(\sum_j \frac{t_{kj,n}}{\tilde{d}_{ij,n}} - \sum_j t_{kj,n} \frac{\tilde{v}_{ik,n}^2 d_{ij,n}}{v_{ik,n}^2 \tilde{d}_{ij,n}^2} \\ + 2\lambda[\tilde{v}_{ik,n}\mathbf{U}]_k + 2\lambda[\text{Diag}(\frac{\tilde{v}_{ik,n}\mathbf{U}^+ + \tilde{v}_{ik,n}\mathbf{U}^-}{\tilde{v}_{ik,n}})]_k (v_{ik,n} - \tilde{v}_{ik,n}) \Big) = 0 \tag{15}$$

and solve (15) by Vieta's theorem [21] derives the updating function of $\mathbf{V}_n$. Similarly, setting the derivative of (14) to zero derives the updating function of $\mathbf{T}_n$:

$$\mathbf{T}_n \leftarrow \tilde{\mathbf{T}}_n \odot \sqrt{\frac{|\mathbf{Y}_n|^{\cdot 2}\mathbf{V}_n^T(\mathbf{V}_n\tilde{\mathbf{T}}_n)^{\cdot -2}}{\mathbf{V}_n^T(\mathbf{V}_n\tilde{\mathbf{T}}_n)^{\cdot -1}}} \tag{16}$$

where $\mathbf{V}_n$ is the solution of (15). (9) is solved.

The regularization coefficient λ affects the model performance. Here we update λ automatically. First, the variables $\hat{\mathbf{X}}_{ij}$ and $\mathbf{V}_n$ are initialized with the successive nonnegative projection algorithm [22], then λ is updated by:

$$\lambda \leftarrow \hat{\lambda} \frac{\sum_{i,j} \left(\frac{d_{ij,n}}{\sum_k v_{ik,n} t_{kj,n}} + \log \sum_k v_{ik,n} t_{kj,n} \right)}{\log|\det(\mathbf{V}_n^T\mathbf{V}_n + \delta\mathbf{I})|} \tag{17}$$

where $\hat{\lambda}$ is the value of λ at the previous iteration, and recommended to be chosen between 10^{-3} and 1 at the first iteration.

D. Theoretical analysis

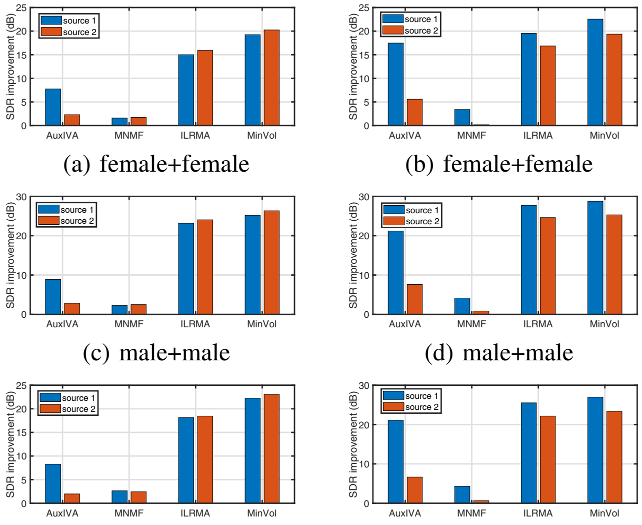
Similar to [23], we prove the identifiability of $\mathbf{V}_n$ in *MinVol*, which supports the superiority of the proposed *MinVol*-ILRMA over ILRMA theoretically.

Theorem 1. *Let $(\mathbf{V}_n^\star, \mathbf{T}_n^\star)$ be an optimal solution of* (4)*. If the ground truth $\mathbf{V}_n^\natural$ and $\mathbf{T}_n^\natural$ satisfies the scattered condition [23] and rank$(\mathbf{D}_n) = K$. Then $\mathbf{V}_n^\star = \mathbf{V}_n^\natural\mathbf{B}$ and $\mathbf{T}_n^\star = \mathbf{B}^T\mathbf{T}_n^\natural$, where $\mathbf{B}$ is a permutation matrix.*

Proof 1. *The method can be repeated here*

$$\min_{\mathbf{V}_n, \mathbf{T}_n} \log|\det \mathbf{V}_n^T\mathbf{V}_n| \\ \textit{s.t.} \quad \mathbf{D}_n = \mathbf{V}_n\mathbf{T}_n, \quad \mathbf{V}_n^T\mathbf{1} = \mathbf{1}, \quad \mathbf{T}_n \geq 0 \tag{18}$$

Denote the optimal solution of (18) *as $\mathbf{V}_n'$ and $\mathbf{T}_n'$. There exists a permutation matrix $\mathbf{B}$ such that $\mathbf{V}_n' = \mathbf{V}_n^\star\mathbf{B}$,$\mathbf{T}_n' = \mathbf{B}^T\mathbf{T}_n^\star$. Because rank$(\mathbf{D}_n) = K$, there exists a non-singular*



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

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

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

Fig. 2. Average SDR improvement of the comparison methods over mixed speech in anechoic environments. (a), (c), (e) are the results in condition 1. (b), (d), (f) are the results in condition 2.

matrix $\mathbf{C}$ *such that* $\mathbf{V}'_n = \mathbf{V}^\star_n \mathbf{C}$, $\mathbf{T}'_n = \mathbf{C}^{-1}\mathbf{T}^\star_n$. *Because we assume* $\mathbf{V}'_n$ *and* $\mathbf{T}'_n$ *are the optimal solution, we have*

$$|\det \mathbf{V}'^T_n \mathbf{V}'_n| = |\det \mathbf{C}^T \det \mathbf{T}^{T\star}_n \det \mathbf{T}^\star_n \det \mathbf{C}| \le |\det \mathbf{V}^{T\star}_n \mathbf{V}^\star_n| \tag{19}$$

On the other hand, because $\mathbf{T}'_n$ *is an optimal solution of* (19), *we have:*

$$\mathbf{C}^{-1}\mathbf{T}^\star_n \ge 0, \mathbf{T}^{T\star}_n (\mathbf{C}^{-1})^T \mathbf{1} = \mathbf{1} \tag{20}$$

We assume that $\mathbf{T}^\star_n$ *is sufficiently scattered, therefore* $\|\mathbf{C}(:,k)^{-1}\|_2 \le \mathbf{1}^T\mathbf{C}(:,k)^{-1}$. *Then, due to the Hadamard inequality, we have:*

$$|\det \mathbf{C}^{-1}| \le \prod_{k=1}^{K} \|\mathbf{C}(:,k)^{-1}\|_2 \le \prod_{k=1}^{K} \mathbf{1}^T\mathbf{C}(:,k)^{-1} = 1 \tag{21}$$

Combining (20) *and* (21) *derives that* $|\det \mathbf{C}| = 1$. *The above conclusions imply that the columns of* $\mathbf{C}$ *can only be selected from the columns of the identity matrix. So* $\mathbf{C}$ *should be a non-singular and permutation matrix.*

III. Experiments

Experimental settings: We followed the environment of the SISEC challenge [24] to construct a determined multichannel speech separation task with $M = N = 2$. We used the Wall Street Journal (WSJ0) corpus [25] as the speech source. We evaluated the comparison methods on all gender combinations.

We generated two test conditions, 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 microphone spacing is 5.66 cm and 2.83 cm respectively, and (ii) the incident angles of the two speakers follow [4, Figs. 9a and 9b]. The image source model [26] was used to generate the room impulse responses with the reverberation time T_{60} selected from $[130, 150, 200, 250, 300, 350, 400, 450, 500]$ ms. For each gender combination and each T_{60} in each condition, we generated 200 mixtures for evaluation. The sampling rate was set to 16 kHz.

TABLE I
The average SDR improvement (dB).

	Condition 1			Condition 2		
	f+f	m+m	f+m	f+f	m+m	f+m
AuxIVA [17]	2.98	3.40	2.95	5.92	7.55	7.60
MNMF [6]	1.25	1.84	1.97	1.47	2.00	2.11
ILRMA [10]	5.03	6.89	5.72	5.17	7.31	6.00
MinVol	7.39	8.77	7.87	8.31	10.06	9.29

The parameter δ of *MinVol* in (5) was set to 0.5. Note that MinVol is insensitive to the selection of δ, since it is only used to prevent (5) from infinity. We compared *MinVol* with AuxIVA [17], MNMF [6], and ILRMA [10]. For each comparison method, we set the frame length and frame shift of STFT to 64 ms and 32 ms respectively. Hamming window was also applied to each frame. The number of basis vectors were set to 10 in MNMF, ILRMA and *MinVol* by default. The evaluation metric is signal-to-distortion ration (SDR) [27].

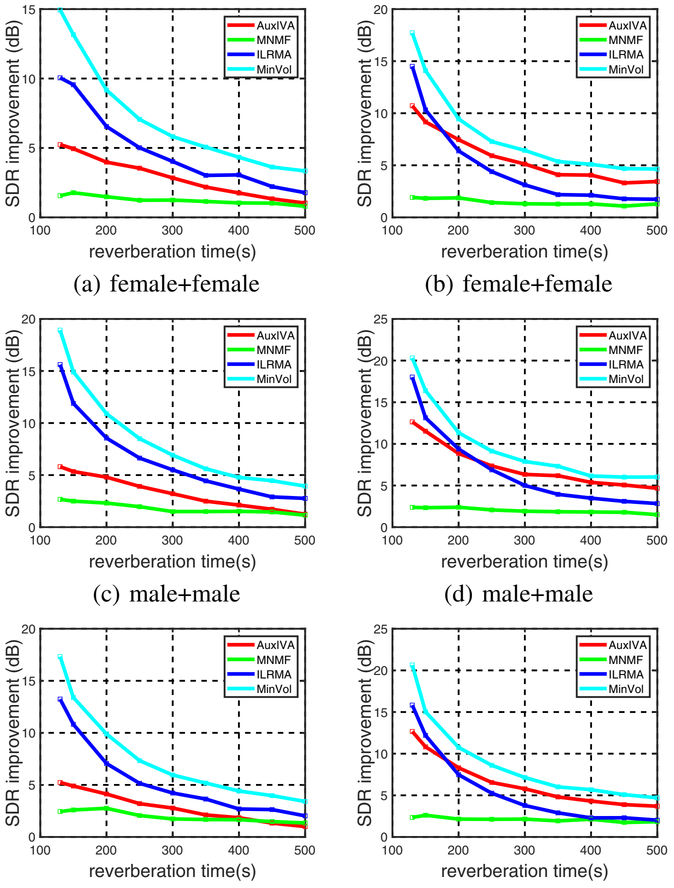
Results We first conducted an experiment in anechoic environments. Fig. 2 shows the average SDR improvement of the comparison methods over the mixed speech. From the figure, we see that the performance of the proposed *MinVol* is significantly better than that of MNMF. Compared to AuxIVA and ILRMA, *MinVol* achieves an SDR improvement of about 3 dB on average.

Then, we studied the performance of the comparison methods in reverberant environments. Fig. 3 shows the SDR improvement over the mixed speech with respect to T_{60}. From the figure, we see that the curves of the SDR improvement produced by *MinVol* are always higher than those produced from the comparison methods.

To clearly show the general improvement of *MinVol* 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 Table I. From the table, we see that the average SDR improvement brought by the proposed *MinVol* is 2 dB higher than ILRMA in condition 1, and 3 dB higher in condition 2.

IV. Conclusion

This paper proposes *MinVol* source separation method. It constrains ILRMA with the volume minimization to improve the identifiability of the source model estimation of ILRMA. It further unifies the IVA-based blind spatial optimization and the minimum-volume constrained MNMF. It is optimized by the alternating fast projected gradient algorithm. We have also proved the identifiability of the volume minimum regularizer. Experimental results show that the proposed algorithm outperforms three representative blind audio source separation methods.



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

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

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

Fig. 3. The curves of the SDR improvement of the comparison methods in reverberant environments. (a), (c), (e) are in condition 1. (b), (d), (f) are in condition 2.

References

[1] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 112–114, 1997.

[2] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[3] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Blind speech extraction based on rank-constrained spatial covariance matrix estimation with multivariate generalized gaussian distribution," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1948–1963, 2020.

[4] S. Mogami, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, and N. Ono, "Independent low-rank matrix analysis based on time-variant sub-gaussian source model for determined blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 503–518, 2020.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

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

[7] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 654–669, 2015.

[8] 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 Transactions on Audio, Speech, and Language Processing*, 2020.

[9] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *2015 IEEE ICASSP*. IEEE, 2015, pp. 276–280.

[10] ——, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[11] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2006.

[12] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 328–332, 2018.

[13] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2306–2320, 2015.

[14] X. Fu, K. Huang, B. Yang, W. Ma, and N. D. Sidiropoulos, "Robust volume minimization-based matrix factorization for remote sensing and document clustering," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6254–6268, 2016.

[15] V. Leplat, N. Gillis, and A. M. S. Ang, "Blind audio source separation with minimum-volume beta-divergence nmf," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3400–3410, 2020.

[16] 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, 2019.

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

[18] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *10th International Conference on ISSPA*. IEEE, 2010, pp. 1–4.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[20] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[21] J. S. Tanton, *Encyclopedia of mathematics*. Infobase Publishing, 2005.

[22] V. Leplat, A. M. Ang, and N. Gillis, "Minimum-volume rank-deficient nonnegative matrix factorizations," in *2019 IEEE ICASSP*. IEEE, 2019, pp. 3402–3406.

[23] X. Fu, K. Huang, N. D. Sidiropoulos, Q. Shi, and M. Hong, "Anchor-free correlated topic modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 5, pp. 1056–1071, 2019.

[24] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (sisec2011):-audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 414–422.

[25] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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