Under-determined Convolutive Blind Source Separation Combining Density-based Clustering and Sparse Reconstruction in Time-Frequency Domain

Junjie Yang Yi Guo, Zuyang Yang, and Shengli Xie [*][†][‡]

April 15, 2020

Abstract

Blind source separation (BSS) in time-frequency (TF) domain is a versatile framework to recover sources from the recorded mixture signals in a reverberant environment. In general, a two-stage strategy is one of the popular BSS frameworks for the underdetermined BSS case (the number of mixtures is less than the number of sources), which is a tough problem due to the mixing matrix is not invertible. In this paper, we propose a new two-stage scheme combining density-based clustering and sparse reconstruction to estimate mixing matrix and sources, respectively. At the first stage, we transform the mixing matrix estimation as an eigenvector clustering problem based on a particular local dominant assumption. The eigenvectors are first exploited from the rank-one structure of local covariance matrices of mixture TF vectors. These eigenvectors are then clustered and adjusted to give estimated mixing matrix by cooperating density-based clustering and weight clustering. At the second stage, we transform the source reconstruction as a ℓ_p norm $(0 < p \leq 1)$ minimization by an iterative Lagrange multiplier method. With a proper initialization, the obtained solution is a global minimum for any p in $(0, 1]$ with convergence guarantee. The proposed approach is demonstrated to be superior to the state-of-the-art baseline methods in various underdetermined experiments.

Keywords: Blind source separation (BSS), reverberation environment, underdetermined case, local dominance, clustering, sparse reconstruction, speech.

1 Introduction

Blind source separation (BSS) aims at recovering multiple unknown sources from mixture signals which are captured by multiple microphones [ćević7501858, 1, 2]. The application of BSS is found in various areas, e.g., speech processing, biomedical signal processing, image processing [4–6]. Convolutive BSS is proven suitable to depict the speech signal mixing mechanism in a reverberant environment, where multiple sources are convolved from multiple delay mixing system model [1, 2]. In time-frequency (TF) domain, the convolutive BSS problem can be modeled as a sequence of linear mixing system [7, 8]. The objective of BSS in TF domain is to design a series of unmixing filters to separate latent sources by exploiting prior knowledge of sources, e.g., nonnegative feature of source spectrum [9] or sparsity of source [10]. In this paper, BSS in TF domain is mainly discussed under the scenario where the mixing system can be underdetermined, i.e., the number of mixtures is less than the number of sources. This underdetermined BSS in TF domain is identified as an open problem [11, 12] because the mixing matrix is not invertible, leading to a difficulty in designing

[*] J. Yang, Z. Yang, and S. Xie are with the School of Automation, Guangdong University of Technology, and Guangdong Key Laboratory of IoT Information Processing, Guangzhou, 510006, China. (E-mail: yangjunjie1985@gmail.com, yangzuyuan@aliyun.com, shlxie@gdut.edu.cn).

[†] Y. Guo is with the Centre for Research in Mathematics, School of Computing, Engineering and Mathematics, Western Sydney University, Parramatta, NSW, 2150, Australia. (E-mail: Y.Guo@westernsydney.edu.au).

[‡] Corresponding author: Zuyuan Yang, (yangzuyuan@aliyun.com).

the unmixing filters.

Some underdetermined BSS methods utilize nonnegative matrix factorization (NMF) to exploit the nonnegativeness of sources [9, 13]. In general, NMF for BSS aims to decompose the spectrum of mixture TF vectors into a product of spectral basis matrix and its coefficient matrix. In the work of [14], it tries to combine maximum-likelihood estimation and NMF based on Itakura-Saito divergence measurement to solve single channel audio source separation problem. The authors in [15] seek to extend the NMF-based framework into a multichannel case by estimating the basis matrix and coefficient matrix with expectation-maximization (EM) and multiplicative update NMF methodology, respectively. Furthermore, other works like [16, 17] provide supervised NMF method to enhance the source separation performance. These methods introduce additional training procedure to learn the basis matrix with a portion of pure sound samples generated by the target source signals. Currently, the implementation of NMF-based methods to the underdetermined BSS case is still ongoing by cooperating with various priori knowledge of sources. For example, [18] takes the sparsity of sources into account on designing the NMF-based algorithm.

Several underdetermined BSS approaches employ TF masking strategy to separate sources from observed mixtures [19]. Yilmaz introduces a condition called approximately Window-disjoint orthogonal (WDO) assumption in [20], which means that source components are disjoint at each TF slot. Based on this assumption, it provides a degenerate unmixing estimation technique (DUET) to design binary masking and extract source components. The method proposed in [21] tries to extend the binary masking into complex-value version by utilizing a supervised learning method. Furthermore, the TF masking technique is combined with the K-means clustering in the work of [22]. This strategy designs a hard TF masking to classify each mixture TF vector into a particular cluster, thus separate the sources without scaling alignment. It is worth noting that the above TF masking methods are mainly designed based on the mentioned WDO assumption. However, this disjoint assumption may not usually hold beyond a certain scale, e.g., the spectrum of mixtures at some TF slots may be mixed by more than one source component, which results in a so-called inter-interference problem [23].

A major class of underdetermined BSS algorithms adapt the sparse component analysis (SCA) with a two-stage framework to unmix latent sources from the mixture signals [24, 38]. In general, the two-stage strategy includes mixing matrix estimation and source reconstruction. At the first stage, the mixing matrix estimation problem is converted to a joint-approximate-diagonalization (JAD) minimization model by exploiting the non-stationary property of source signals from a sequence of local covariance matrices of mixtures [26, 27]. The authors in [28, 29] transform the JAD minimization into a tensor decomposition problem. It is worth noting that these tensor-based methods are only suitable for some particular under-determined cases due to the algebraic uniqueness restriction of tensor decomposition, e.g., the number of microphones must be greater than three. Since the unmixing operation of underdetermined case has no unique solution, at the second stage, the source reconstruction is transformed as building various optimization models to reconstruct most likely version of true sources. For example, the authors in [30] utilizes the volume minimization to exploit the sparsity of source and recover them. The method proposed in [31] transforms the source reconstruction as a minimization problem of Bayes risk with the squared loss measurement. Furthermore, a ℓ_p norm-based optimization model is proposed for the source reconstruction problem based on a Laplacian distribution assumption of sources in [32]. However, it shows that the solution of [32] may stuck in a local minimum when p is less than 0.75. In general, the source reconstruction in the underdetermined case is still a tricky problem.

To alleviate the problems as mentioned above based on the two-stage framework, in this paper, we design a different two-stage strategy to estimate the mixing matrix and reconstruct sources for the underdetermined BSS problem in TF domain. Inspired by the local dominance of sources [33, 34], i.e., each source component is assumed to be locally dominant at least in one successive TF slots, at the first stage, we transform the mixing matrix estimation to an eigenvector clustering problem. In the proposed clustering scheme, the objects are the leading eigenvectors by exploiting the rank-one structure of local covariance matrices. Next, a density-based clustering method [35] is developed to choose particular eigenvectors as the cluster centers. These eigenvectors are required to be satisfying criterion of dense local density and distinctive distance from other eigenvectors with higher local density. To overcome the interference of outliers, these clusters are further adjusted by additional weight clustering scheme. At the second stage, we transform the source reconstruction to a sparse reconstruction minimization model based on a ℓ_p norm $(0 < p \leq 1)$ measurement. The ℓ_p norm based minimization is a flexible framework to exploit sparsity of various source signals. Next, this minimization problem is solved by a developed iterative Lagrange multiplier strategy with a proper initialization procedure. Finally, the constructed sources are converted back into the time-domain.

The main contributions of this paper are as follows:

A1. The proposed clustering strategy reveals the inherent connection of local dominance of sources and mixing matrix estimation in terms of eigenvector clustering. The proposed scheme is improved by cooperating density-based clustering and weight clustering, which is a robust scheme to estimate the mixing matrix with the presence of outliers.

A2. The proposed ℓ_p norm based iterative Lagrange multiplier approach can be viewed as a generalization of [32] for the source reconstruction, whose solution is a global minimum for any p in $(0, 1]$ with convergence guarantee.

A3. The provided new two-stage BSS strategy has been demonstrated to be competitive to the state-of-the-art methods with various synthetic and real-recorded experiments.

The remainder of this paper is organized as follows. First, the system model and assumptions are discussed in Section II. Next, the proposed algorithm including mixing matrix estimation and source reconstruction are provided respectively in Section III. Numerical results are presented in section IV. Finally, conclusions are drawn in Section V. Table I is the list of notations to be used in the rest of this paper.

Table 1: Conventional Symbols

$\star$	Linear convolutive operator
$\lfloor \cdot \rfloor$	Rounding down operator
diag	Retain only the diagonal elements and make the diagonal elements as a vector
$(\cdot)^*$	Complex conjugation
$(\cdot)^T$	Transpose
$(\cdot)^H$	Hermitian transpose
$(\cdot)^{-1}$	Inverse
$\mathrm{E}(\cdot)$	Expectation operator
$\|\cdot\|_0$	ℓ_0 norm
$\|\cdot\|_p$	ℓ_p norm
$\|\cdot\|_F$	Frobenius norm

2 System Model and Problem Description

2.1 Convolutive Mixing System Model

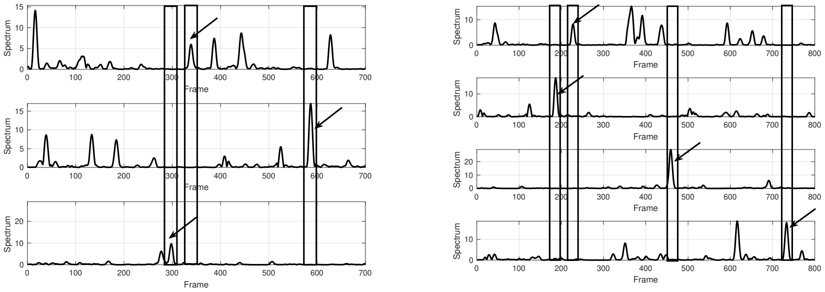
In the following system model, we consider multiple sources are recorded by multiple microphones in the reverberant environment. Let M, N denote the number of microphones and sources, respectively. Let $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$ and $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ denote as mixture signals and source signals, respectively. With the above notations, we consider the convolutive blind source separation (BSS) problem based on a sequence of multiple input multiple output (MIMO) finite impulse response system with order L as

$$\mathbf{x}(t) = \mathbf{H} \star \mathbf{s}(t) = \sum_{\tau=0}^{L-1} \mathbf{H}(\tau)\mathbf{s}(t-\tau), \tag{1}$$

where $\mathbf{H}(\tau) \in \mathbb{R}^{M \times N}$ is the mixing matrix at time lag τ. The elements of $\mathbf{H}(\tau)$ denoted by $h_{i,j}(\tau)$ are the room impulse response coefficients (RIRs) between the ith microphone and the jth source.

Similar to the works of [36, 37], the formulation of (1) can be approximately transformed to a multiplicative narrowband model in the time-frequency (TF) domain by performing a F-length short-time Fourier transform (STFT) to the mixture signals $\mathbf{x}(t)$, such as

$$\mathbf{x}_{f,d} = \mathbf{H}_f \mathbf{s}_{f,d} + \mathbf{e}_{f,d}, \tag{2}$$



(a)

(b)

Figure 1: Illustration of local dominance assumption in TF domain. The sub-blocks pointed by arrows exhibit the local dominance of speech sources. (a) $N = 3$, (b) $N = 4$.

where $\mathbf{x}_{f,d} = [x_{f1,d}, \ldots, x_{fM,d}]^T$, $\mathbf{s}_{f,d} = [s_{f1,d}, \ldots, s_{fN,d}]^T$, $\mathbf{e}_{f,d} = [e_{f1,d}, \ldots, e_{fM,d}]^T$ are the complex-valued vectors of mixture, source and error resulted by the narrowband approximation at TF slot (f, d), respectively. The length of $\mathbf{x}_{f,d}$ at frequency bin f is denoted as D. In the system model of (2), $\mathbf{H}_f = [\mathbf{h}_{f1}, \ldots, \mathbf{h}_{fN}]$ is a $M \times N$ complex valued mixing matrix at the fth frequency bin. Each column of $\mathbf{H}_f$, e.g., $\mathbf{h}_{fi}$, is called as steering vector representing each direction of mixing matrix, $i = 1, \ldots, N$. It is worth noting that the system model of (2) holds when the window length of F satisfies $F \geq L/2 + 1$.

2.2 Local Covariance Matrix

The second-order statistics of $\mathbf{x}_{f,d}$, i.e., local covariance matrix, is introduced to exploit the non-stationary property of latent sources. We divide the mixture TF vectors into Q non-overlapping blocks, such that each sub-block contains $P = \lfloor D/Q \rfloor$ successive vectors. In this case, we define the qth local covariance matrix of mixture TF vectors by $\mathbf{R}^{\mathbf{x}}_{f,q} \triangleq \mathrm{E}(\mathbf{x}_{f,d}\mathbf{x}^H_{f,d})$, $d = (q-1)P + 1, \ldots, qP$. The local covariance matrix of $\mathbf{R}^{\mathbf{x}}_{f,q}$ can be further expanded as

$$\mathbf{R}^{\mathbf{x}}_{f,q} = \mathbf{H}_f \mathbf{R}^{\mathbf{s}}_{f,q} \mathbf{H}^H_f, \tag{3}$$

where $\mathbf{R}^{\mathbf{s}}_{f,q} \triangleq \mathrm{E}(\mathbf{s}_{f,d}\mathbf{s}^H_{f,d})$. Suppose that the source TF vectors at each sub-block are wide-sense quasi-stationary with zero-mean and uncorrelated from each other, the covariance of $\mathbf{R}^{\mathbf{s}}_{f,q}$ in (3) can be written in a diagonal matrix formulation,

$$\mathbf{R}^{\mathbf{s}}_{f,q} = \begin{bmatrix} \sigma^2_{f1,q} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma^2_{fN,q} \end{bmatrix},$$

where $\sigma^2_{fi,q} \triangleq \mathrm{E}(s_{fi,d}s^*_{fi,d}), d = (q-1)P + 1, \ldots, qP, i = 1, \ldots, N$. In practical implementation, the local covariance matrix of $\mathbf{R}^{\mathbf{x}}_{f,q}$ can be approximately calculated by

$$\hat{\mathbf{R}}^{\mathbf{x}}_{f,q} = \frac{1}{P} \sum_{d=q(P-1)+1}^{qP} \mathbf{x}_{f,d}\mathbf{x}^H_{f,d}. \tag{4}$$

2.3 Assumptions

We introduce the following assumptions to the system model of (2):

A1) The number of microphones is less than the number of sources, i.e., $M < N$.

A2) For each source i at any frequency bin f, there exists at least a sub-block indexed by q_i, such that $\sigma^2_{fi,q_i} > 0$ and $\sigma^2_{fj,q_i} = 0$, $\forall\ j \neq i$, where $\forall q_i \in \{1, 2, \ldots, Q\}$, $i = 1, \ldots, N$.

Assumption A1) considers the under-determined scenario, which is a tough problem in convolutive BSS.

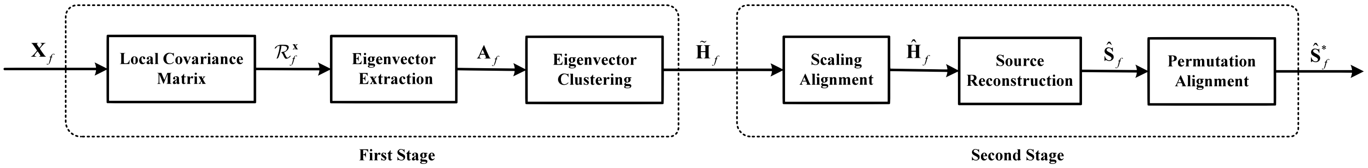


Figure 2: Block diagram of the proposed two-stage framework.

In this case, the mixing matrix is not invertible and result in difficulty of designing unmixing filters. A2) is the so-called local dominance assumption introduced in [33, 34, 38], which stems from the observation of some signals, e.g., speeches, exhibiting sparsity feature in that each of them is locally dominant in the time domain or TF domain. As an example shown in Fig. 1, the local dominance of speech signals can be easily observed in TF domain.

It is worth noting that A2) has relaxed the assumption of WDO [20] to a great extent, which refers that the source components are required to be disjoint at each TF slot, i.e., $s_{fi,d} \times s_{fj,d} = 0, i, j = 1, ..., N, i \neq j$. Alternatively, A2) infers that the source components are only required to be disjoint at several sub-blocks. In this paper, A2) is a key assumption to ensure that the steering vector of the mixing matrix is detectable as long as those particular indices of TF sub-block, i.e., q_1,$\ldots$,q_N, are correctly identified.

2.4 Objectives

In this paper, the under-determined BSS problem in TF domain will be discussed in a new two-stage framework based on the system model (2) and provided assumptions. First, we develop an eigenvector clustering approach to estimate the mixing matrix. Second, we present a sparsity-based reconstruction approach to estimate latent sources based on the estimated mixing matrix.

3 Proposed Algorithm

3.1 Overview

We provide a flowchart to show the significant steps of the proposed two-stage strategy in Fig.2. At the first stage, the mixture TF vectors $\mathbf{X}_f \triangleq [\mathbf{x}_{f,1}, \ldots, \mathbf{x}_{f,D}]$ are first used to obtain a sequence of local covariance matrices, e.g., $\mathcal{R}_f^{\mathbf{x}} \triangleq [\mathbf{R}_{f,1}^{\mathbf{x}}, \ldots, \mathbf{R}_{f,Q}^{\mathbf{x}}]$. Then, the eigenvector with the largest eigenvalue of each local covariance matrix is extracted and all of these eigenvectors are collected as $\mathbf{A}_f$. Third, the columns of $\mathbf{A}_f$ are clustered to give the estimated mixing matrix $\tilde{\mathbf{H}}_f$. At the second stage, the estimated mixing matrix $\tilde{\mathbf{H}}_f$ is first rescaled to give $\hat{\mathbf{H}}_f$. Then, the mixture vectors $\mathbf{X}_f$ are column-wisely unmixed along with the estimated $\hat{\mathbf{H}}_f$ to give $\hat{\mathbf{S}}_f$. Third, additional permutation alignment is performed to give reconstructed sources $\hat{\mathbf{S}}_f^*$. Detailed description of the steps mentioned above will be given in the following Sections. The purpose of this paper is to estimate the source signals under a two-stage framework combining density-based clustering and sparsity-based reconstruction. As the two-stage strategy are discussed frequency wisely, the superscript f of provided symbols, e.g., $\mathbf{x}_{f,d}$, $\mathbf{H}_f$, $\mathbf{s}_{f,d}$, are omitted in following discussion for simplicity.

3.2 Mixing Matrix Estimation

In this part, we show that the mixing matrix estimation can be cast as an eigenvector clustering problem. This clustering scheme includes three steps: 1) extract leading eigenvectors from local covariance matrices; 2) obtain clusters by a density-based clustering method; 3) identify the mixing matrix via an additional adjustment of clusters.

3.2.1 Eigenvector Extraction

To begin, the local covariance matrix of (3) is expanded as follows,

$$\mathbf{R}_q^{\mathbf{x}} = \sum_{i=1}^{N} \sigma_{i,q}^2 \mathbf{h}_i \mathbf{h}_i^H. \tag{5}$$

Based on the assumption of A2), there exists at least one sub-block indexed by q_i, where the corresponding local covariance of $\mathbf{R}_{q_i}^{\mathbf{x}}$ can be approximately expanded as follows,

$$\mathbf{R}_{q_i}^{\mathbf{x}} \approx \sigma_{i,q_i}^2 \mathbf{h}_i \mathbf{h}_i^H. \tag{6}$$

In (6), it indicates that the local covariance matrix is roughly a rank-one structure if the local dominance condition holds.

To exploit approximate rank-one structure of the local covariance matrix, we try to employ the eigenvector extraction approaches similar to the works of [39–41]. Specifically, we perform eigenvalue decomposition (EVD) to the local covariance matrix of $\mathbf{R}_q^{\mathbf{x}}$, such that

$$\mathbf{R}_q^{\mathbf{x}} = \mathbf{U}_q \mathbf{\Sigma}_q \mathbf{U}_q^H, \tag{7}$$

where $\mathbf{U}_q$ and $\mathbf{\Sigma}_q$ are the eigenvector matrix and eigenvalue matrix, respectively. The extracted vector denoted by $\mathbf{a}_q$ is the first eigenvector in $\mathbf{U}_q$ corresponding to the largest eigenvalue of $\mathbf{\Sigma}_q$. Without loss of generality, the first entry of eigenvector $\mathbf{a}_q$ is restricted to be a positive. The eigenvector extraction is performed subblock wisely to give an eigenvector matrix defined by $\mathbf{A} \triangleq [\mathbf{a}_1, \ldots, \mathbf{a}_Q]$. Based on local dominance assumption of A2), the particular eigenvectors dominated by only one source component, i.e., $\mathbf{a}_{q_1}, \ldots, \mathbf{a}_{q_N}$, are crucial to the mixing matrix estimation. In the following, we focus on how to estimate the steering vectors by the proposed eigenvector clustering strategy.

3.2.2 Density-based Clustering

In this part, we will exploit the directions of steering vectors based on the extracted eigenvectors. Here, we compute a similarity matrix from eigenvectors $\mathbf{A}$, such as,

$$\mathbf{V} \triangleq \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1Q} \\ \vdots & \vdots & & \vdots \\ v_{Q1} & v_{Q2} & \cdots & v_{QQ} \end{bmatrix}, \tag{8}$$

where $v_{qk} = \parallel \mathbf{a}_q - (\mathbf{a}_q^H \mathbf{a}_k)\mathbf{a}_k \parallel_F^2$, $q, k = 1, .., Q$. Utilizing the similarity matrix of $\mathbf{V}$, we can visualize the distribution of eigenvectors by projecting the high-dimensional eigenvectors into a two-dimensional space by maintaining the similarity of any pair of eigenvectors. As an example shown in Fig.3 (a), we observe that the eigenvectors are mostly concentrated around various perfect steering vectors. The distribution of eigenvectors of Fig.3 (a) has two significant characteristics: 1) there are N local regions with high density; 2) the local density regions are far from each other. Based on these observations, it is reasonable to employ the density-based clustering strategy [35] to identify the steering vectors based on the similarity matrix of $\mathbf{V}$.

Two factors are taken into account in the eigenvector clustering, i.e., local density ρ_q and minimum distance δ_q from the eigenvectors of higher density. First, the local density ρ_q is defined by using a sum of Gaussian kernel functions,

$$\rho_q \triangleq \sum_{k \neq q} e^{-\frac{v_{qk}^2}{\tau_c^2}}, \tag{9}$$

where τ_c is a cutoff distance used to define a region for each data point. Usually, parameter τ_c is often empirically chosen to ensure around 6% to 8% of the total number of points in local region. Second, the minimum distance between point q and any other points with a higher density is defined as

$$\delta_q = \min_{k:\rho_k > \rho_q} (v_{qk}). \tag{10}$$

It is worth noting that the point with global maximum in the density, indexed as q^*, whose minimum distance δ_{q^*} is defined as follows,

$$\delta_{q^*} = \max_{q,k=1,\ldots,Q} (v_{qk}), \;\; if \;\; \rho_{q^*} = \max_{q=1,\ldots,Q} (\rho_q). \tag{11}$$

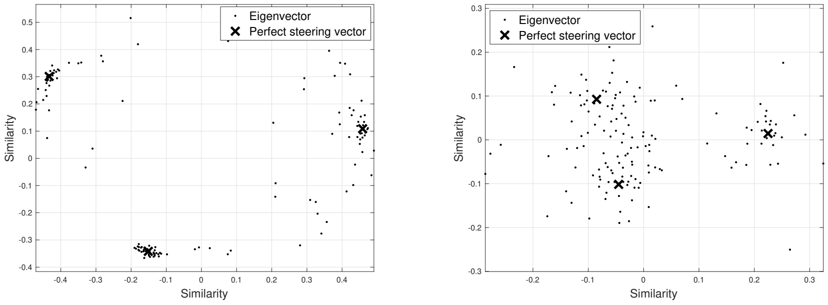
Third, the two factors are multiplied together to obtain a score as follows,

$$\gamma_q = \rho_q \times \delta_q. \tag{12}$$

The scores applying (12) are performed for all of the sub-blocks to give $\{\gamma_q\}_{q=1}^Q$. The obtained scores are further ranked in a descending order. In this way, the eigenvectors with the first highest N scores are extracted as the clusters, which can be denoted by $\mathbf{C} \triangleq [\mathbf{c}_1, \ldots, \mathbf{c}_N]$.

3.2.3 Adjustment of weight clustering

It is worth noting that in practical implementation, the eigenvectors may not be well distributed at each frequency bin as shown in Fig. 3 (b), e.g., there may be



(a)

(b)

Figure 3: Scatter plot of eigenvectors in $Case\ (M, N) = (2, 3)$. The eigenvectors are compressed into a two dimensional space by maintaining the similarity of any pair of eigenvectors. This procedure can be performed by Matlab function of 'mdscale'. (a) case of few outliers ($f = 80$); (b) case of heavy outliers ($f = 244$).

lots of outliers at some frequency bins. In this circumstance, it would be difficult to cluster eigenvectors by only using above density-based method. To alleviate this problem, we further adjust the estimated clusters by a weight clustering scheme proposed by our previous work of [42]. This scheme aims to suppress the effects of outliers during clustering by introducing weight penalty, thus it is able to identify the clusters robustly. In this case, the density-based method provides an initialization of the weight clustering scheme.

The procedures of weighted eigenvector clustering can be concluded in the following three steps. First, we weight the eigenvector by a kernel function as follows,

$$\mathbf{b}_{qk} \triangleq e^{-w_{qk}^2/\tau_0^2}\mathbf{a}_q, k = 1, \dots, N, \tag{13}$$

where $w_{qk} = \| \ \mathbf{a}_q - (\mathbf{a}_q^H\mathbf{c}_k)\mathbf{c}_k \ \|_F^2$ and τ_0 is a preset threshold, e.g., $\tau_0 = 0.05$. Second, we construct a weighted covariance matrix as follows,

$$\mathbf{R}_k^{\mathbf{b}} = \sum_{q=1}^{Q} \mathbf{b}_{qk}\mathbf{b}_{qk}^H. \tag{14}$$

Third, we still perform EVD to the wighted covariance matrix of $\mathbf{R}_k^{\mathbf{b}}$, such that

$$\mathbf{R}_k^{\mathbf{b}} = \mathbf{U}_{q_k}\mathbf{\Sigma}_{q_k}\mathbf{U}_{q_k}^H. \tag{15}$$

The eigenvector corresponds to the largest eigenvalue from (15) is extracted as an updated version of cluster $\mathbf{c}_k$, $k = 1, \dots, N$. More details of above procedures can be found in [42]. The implementation of mixing matrix estimation is concluded in Algorithm 1.

3.3 Source Reconstruction

3.3.1 The ℓ_p Norm-based Minimization Model

In the under-determined case, the source reconstruction is impossible via directly inversing the estimated mixing matrix to the system of (2) since it is a fat matrix. Alternatively, we employ a sparsity-based method to reconstruct the sources as follows. To begin, each source component is assumed to satisfy the following complex-valued super-Gaussian distribution [32], such as

$$P(\mid s_{i,d} \mid) = p\frac{\gamma^{1/p}}{\Gamma(1/p)}e^{-|s_{i,d}|^p}, \tag{16}$$

where $0 < p \leq 1$ and $\gamma > 0$ control shape and variance of the probability function, respectively; $\Gamma(\cdot)$ is the gamma function. The objective of source reconstruction is to find the sparsest term of $\mathbf{s}_d$ based on the linear mixing system of (2). For this purpose, a maximum posterior likelihood of $\mathbf{s}_d$ is given by

$$\begin{aligned} &\max_{\mathbf{s}_d} \prod_{i=1}^{N} P(\mid s_{i,d} \mid) \\ &s.t.\ \mathbf{x}_d = \hat{\mathbf{H}}\mathbf{s}_d, \end{aligned} \tag{17}$$

where $\hat{\mathbf{H}}$ is the estimated mixing matrix. The optimization model of (17) is equivalent to

$$\begin{aligned} &\min_{\mathbf{s}_d} \|\mathbf{s}_d\|_p^p \\ &s.t.\ \mathbf{x}_d = \hat{\mathbf{H}}\mathbf{s}_d, \end{aligned} \tag{18}$$

where $\|\mathbf{s}_d\|_p^p \triangleq \sum_{i=1}^{N} \mid s_{i,d} \mid^p$. In the work of [32], the optimization of (18) is solved by transforming it into a subspace minimization problem. However, the global minimal solution based on the method of [32] may not exist for all values of p, i.e., it can only guarantee the global convergence when $p \leq 0.75$. To avoid this problem, we propose the following strategy to solve (18) based on the Lagrange multiplier method. In the next part, we will show that the proposed method can achieve a global solution with convergence guarantee for any p in $(0, 1]$.

Algorithm 1 Implementation of Mixing Matrix Estimation

```
Input: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_D]$.
Whitening $\mathbf{X}$ applying (25).
**Stage 1**: Eigenvector extraction
**for** $q = 1$ to $Q$ **do**
    Calculate $\hat{\mathbf{R}}_q^{\mathbf{x}}$ applying (4).
    Calculate $\mathbf{a}_q$ applying (7).
**end for**
Construct eigenvector matrix as $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_Q]$.
**Stage 2**: Eigenvector clustering
Calculate similarity matrix $\mathbf{V}$ applying (8).
**for** $q = 1$ to $Q$ **do**
    Calculate $\rho_q$ applying (9).
    Calculate $\delta_q$ applying (10).
    Calculate $\gamma_q$ applying (12).
**end for**
Calculate $\delta_{q^*}$ applying (11).
Obtain score sequence of $\Upsilon = [\gamma_1, \ldots, \gamma_Q]$.
Reorder the eigenvector matrix with the same permutation of a descending alignment to the score sequence of $\Upsilon$.
Truncate the first $N$ reordered eigenvectors to give the estimated clusters $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$.
**Stage 3**: Mixing matrix estimation
**for** $k = 1$ to $N$ **do**
    **repeat**
        **for** $q = 1$ to $Q$ **do**
            Using $\mathbf{a}_q$ and $\mathbf{c}_k$, calculate weighted eigenvector $\mathbf{b}_{qk}$ applying (13).
            Calculate $\mathbf{R}_{qk}^{\mathbf{b}}$ applying (14).
            Using $\mathbf{R}_{qk}^{\mathbf{b}}$, calculate $\tilde{\mathbf{h}}_k$ applying (15).
        **end for**
    **until** some stopping criterion is satisfied.
**end for**
Output estimated mixing matrix: $\tilde{\mathbf{H}} \triangleq [\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_N]$.
De-whitening and re-scaling the estimated mixing matrix to give $\hat{\mathbf{H}}$ by applying (26), (28), respectively.
```

3.3.2 Proposed Lagrange Multiplier Method

To begin with, the model of (18) is reformulated to an unconstrained optimization problem as follows:

$$\min_{\mathbf{s}_d, \boldsymbol{\alpha}} \mathcal{F}(\mathbf{s}_d, \boldsymbol{\alpha}) \triangleq \|\mathbf{s}_d\|_p^p + \boldsymbol{\alpha}^H(\mathbf{x}_d - \hat{\mathbf{H}}\mathbf{s}_d), \tag{19}$$

where $\boldsymbol{\alpha} \in \mathbb{C}^M$ is denoted as Lagrange multiplier. Applying Lagrange multiplier method, the optimal solution of (19) can be deduced as follows (See appendix A),

$$\mathbf{s}_d = \mathbf{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H(\hat{\mathbf{H}}\mathbf{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H)^{-1}\mathbf{x}_d, \tag{20}$$

where

$$\mathbf{\Psi}^{-1}(\mathbf{s}_d) \triangleq \begin{bmatrix} |s_{1,d}|^{2-p} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & |s_{N,d}|^{2-p} \end{bmatrix}.$$

Since (20) is an implicit function, we apply iterative scheme to obtain the solution of $\mathbf{s}_d$ as shown in (21). The iterative scheme can be measured by $\|\hat{\mathbf{s}}_d^{(iter)}\|_p^p - \|\hat{\mathbf{s}}_d^{(iter+1)}\|_p^p$ and terminate when it is below a preset threshold, e.g., 10^{-2}. The convergence of proposed Lagrange multiplier method has been discussed in [43], and included here for completeness.

$$\hat{\mathbf{s}}_d^{(iter+1)} = \begin{cases} \boldsymbol{\Psi}^{-1}(\hat{\mathbf{s}}_d^{(iter)})\hat{\mathbf{H}}^H(\hat{\mathbf{H}}\boldsymbol{\Psi}^{-1}(\hat{\mathbf{s}}_d^{(iter)})\hat{\mathbf{H}}^H)^{-1}\mathbf{x}_d, & if\ \ \|\hat{\mathbf{s}}_d^{(iter)}\|_0 \geq M, \\ \boldsymbol{\Psi}^{-1}(\hat{\mathbf{s}}_d^{(iter)})\hat{\mathbf{H}}^H(\hat{\mathbf{H}}(\boldsymbol{\Psi}(\hat{\mathbf{s}}_d^{(iter)}) + \epsilon\mathbf{I})^{-1}\hat{\mathbf{H}}^H)^{-1}\mathbf{x}_d, & elseif\ \ \|\hat{\mathbf{s}}_d^{(iter)}\|_0 < M. \end{cases} \tag{21}$$

Theorem 1. *For $0 < p \leq 1$, let $\{\hat{\mathbf{s}}_d^{(iter)}\}_{iter=0}^{+\infty}$ denote as the iterative sequences, assuming 1) $\hat{\mathbf{s}}_d^{(0)} \neq \mathbf{0}$ and 2) $\hat{\mathbf{H}}$ is a column-wise linearly independent mixing matrix, then the sequence $\{\hat{\mathbf{s}}_d^{(iter)}\}_{iter=0}^{+\infty}$ obtained by* (21) *is convergent.*

3.3.3 Initialization Issue

It is worth mentioning that since (20) is a non-convex optimization model, the iterative solution of (21) may stuck in a local minimum with an inappropriate initialization of $\hat{\mathbf{s}}_d^{(0)}$. Inspired by the work of [31], we provide the following scheme to avoid the local minimum problem. It is worth noting that a local minimum solution of (21) can be obtained when a portion of $N - M$ components of $\mathbf{s}_d$ are inactive. Let C_N^M be the number of M combinations of set $\{1, \ldots, N\}$ and denote $\mathbf{y}_{j,d}$ as the jth local minimum, $j = 1, \ldots, C_N^M$, satisfying

$$\mathbf{x}_d = \hat{\mathbf{H}}\mathbf{y}_{j,d}. \tag{22}$$

The local minimum of $\mathbf{y}_{j,d}$ in (22) can be estimated by

$$\mathbf{y}_{j,d} \triangleq \begin{bmatrix} \hat{\mathbf{H}}_j^{-1}\mathbf{x}_d \\ \mathbf{0} \end{bmatrix} \begin{matrix} \}M\ nonzero\ indices, \\ \}N - M\ zero\ indices, \end{matrix} \tag{23}$$

where $\hat{\mathbf{H}}_j$ is the jth sub-matrix of $\hat{\mathbf{H}}$ whose columns correspond to the non-zero indices of $\mathbf{y}_{j,d}$. Based on the above definitions, the initialization of sources is selected as a summation of C_N^M local minimums with weight penalty, such as

$$\hat{\mathbf{s}}_d^{(0)} = \sum_{j=1}^{C_N^M} \omega_j \mathbf{y}_{j,d}, \tag{24}$$

where ω_j is the jth weight parameter. In the work of [31], the weight ω_j is exploited from a Bayes-risk probabilistic model, which results in a sophisticate procedure of finding optimal weights. Here, we only adapt the Monte Carlo strategy to randomly generate the weight but restricting that $\sum_{j=1}^{C_N^M} \omega_j = 1$, where $\omega_j \in (0, 1)$.

The detail of source reconstruction is concluded in Algorithm 2. As an example shown in Fig.4, we illustrate the performance of source reconstruction for $Case\ (M, N) = (2, 3)$. In Fig.4 (a), it is observed that the iterative error of $\|\hat{\mathbf{s}}_d^{(iter)}\|_p^p - \|\hat{\mathbf{s}}_d^{(iter+1)}\|_p^p$ decreases along with the iterative step for $d = 1, \ldots, D$. The proposed method can be convergent with several iterations, e.g., less than 25 steps. In Fig.4 (b), it shows a comparison between the spectrum of original sources and the reconstructed version by the proposed method. We see that the sources are precisely reconstructed with minor distortions.

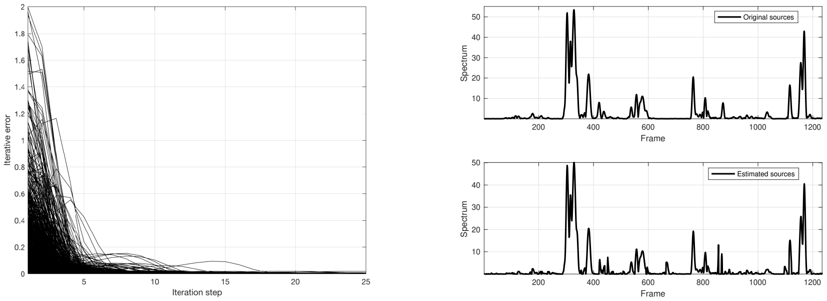
Algorithm 2 Implementation of proposed ℓ_p Norm Minimization based Source Reconstruction

```
Input: Mixtures $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_D]$, estimated mixing matrix $\hat{\mathbf{H}}$.
for $d = 1$ to $D$ do
    Initialize $\hat{\mathbf{s}}_d^{(0)}$ applying (24).
    repeat
        Update $\hat{\mathbf{s}}_d^{(iter)}$ applying (21).
        $iter = iter + 1$.
    until some stopping criterion is satisfied, e.g., the iterative error of $(\|\hat{\mathbf{s}}_d^{(iter)}\|_p^p - \|\hat{\mathbf{s}}_d^{(iter+1)}\|_p^p)$ is less than a given threshold, e.g., $10^{-2}$.
    Output: $\hat{\mathbf{s}}_d^*$.
end for
```

3.4 Pre- and Post- Processing Issue

At the stage of mixing matrix estimation, each $\mathbf{x}_d$ is whitened as the pre-processing step such that

$$\mathbf{x}_d^{\mathrm{W}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{U}_{\mathbf{x}}^H\mathbf{x}_d, \tag{25}$$



(a)

(b)

Figure 4: Example of speech source reconstruction for $Case\ (M,N)=(2,3)$ when $f=52$. The reverberant time is set as 200ms. (a) Iterative errors with iterative steps. The curve represents the reconstruction error of $\hat{\mathbf{s}}_d$, $d=1,\dots,D$; (b) spectrum comparison between original sources and reconstructed sources.

where $\mathbf{U}_\mathbf{x}$ and $\mathbf{\Sigma}_\mathbf{x}$ are the eigenvector matrix and eigenvalue matrix of $\mathrm{E}(\mathbf{x}_d\mathbf{x}_d^H)$, $d=1,..,D$. At the post-processing step, the estimated mixing matrix are dewhitened by

$$\hat{\mathbf{H}}=\mathbf{U}_\mathbf{x}\mathbf{\Sigma}_\mathbf{x}^{\frac{1}{2}}\tilde{\mathbf{H}}. \tag{26}$$

In general, the inherent connection between the estimated mixing matrix and the true one is given by

$$\hat{\mathbf{H}}=\mathbf{H}\mathbf{\Lambda}\mathbf{\Pi}, \tag{27}$$

where $\mathbf{\Lambda}$ and $\mathbf{\Pi}$ are a diagonal matrix with arbitrary scaling and a permutation matrix with arbitrary order, respectively. It shows that the estimated mixing matrix is restricted by ambiguities of scaling and permutation. Such ambiguities are the inherent problem of BSS methods, a more detail description of these ambiguities can be found in [1,2,30].

To solve the ambiguity of scaling, similar as the work of [31], we rescale the estimated mixing matrix by restricting the first row of all 1's, i.e.,

$$\hat{\mathbf{H}}=\begin{bmatrix}1 & \dots & 1\\ \hat{h}_{2,1}/\hat{h}_{1,1} & \dots & \hat{h}_{2,N}/\hat{h}_{1,N}\\ \vdots & \ddots & \vdots\\ \hat{h}_{M,1}/\hat{h}_{1,1} & \dots & \hat{h}_{M,1}/\hat{h}_{1,N}\end{bmatrix}. \tag{28}$$

To solve the ambiguity of permutation, similar as the work of [44], we align the order of reconstructed sources by clustering the adjacent source TF vectors based on their correlation in terms of power ratio. It is skipped here as the focus of this paper is the two-stage scheme in under-determined case rather than permutation alignment.

4 Simulation results

In this section, we will introduce data sets, evaluation criterion and algorithm settings, we then apply proposed two-stage scheme to process these data to evaluate the performance of mixing matrix estimation and source reconstruction, respectively.

4.1 Datasets

In the following experiments, various scenarios are considered based on three public benchmark audio data sets (see Dataset A, B and C). In the provided data sets, the speech signals are recorded by various female or male speakers with sampling rate $F_s=16$ kHz. The data sets also provide synthetic RIRs in a virtual room or authentic RIRs recorded in a real room environment. Using the provided source signals and RIRs function, we can generate various convolutive mixture signals by varying reverberant time of RT_{60}, which is crucial to reflect the room reverberation by measuring the transmission time of signal decay to 60 dB. Four under-determined cases were introduced in the following experiments, such as

$Case$ $(M, N) = (2, 3)$, $(2, 4)$, $(3, 4)$ and $(4, 5)$, respectively.

- **Dataset A** is a data collection artificially generated by convolving speech data with the authentic RIRs from real room, which is recorded with the aid of an acoustic impulse response measuring software named as "Sample Champion" [45]. In this Dataset, three sources and two microphones are located in a room size of 4.9m × 2.8m × 2.65m, whose positions are shown in Fig.5 (a). The reverberation time RT_{60} from source to microphone is 127ms, whose RIRs are shown in Fig.5 (b). Source $s_1(t)$ and $s_2(t)$ are collinear, i.e., they have the same directions to the center of microphones.

- **Dataset B** is a data collection created by a group of speeches and the artificial RIRs function [46], which can simulate various reverberant scenarios with arbitrary settings, e.g., the physical sizes of room, the locations of sources and microphones. The reverberation time RT_{60} is varied from 100ms to 500ms with duration time of 20ms. The synthetic room size is 5m × 5m × 2.3m. The coordinate of sources is (2m, 1m, 1.6m), (2m, 1.4m, 1.6m), (2m, 1.8m, 1.6m), (2m, 2.2m, 1.6m), and (2m, 2.6m, 1.6m), respectively. The coordinate of microphones is (3m, 1m, 1.6m), (3m, 1.5m, 1.6m), (3m, 2m, 1.6m), and (3m, 2.5m, 1.6m), respectively.

- **Dataset C** is a data collection provided by the Signal Separation Evaluation Campaign (SiSEC 2011) [47]. The first development data set includes 28 different sets of synthetic and live recorded mixtures with varying speech types, reverberation time and microphone spacing. It is worth noting that the synthetic recorded mixtures is generated by the Roomsi toolbox for a rectangular room and omni-directional microphone arrays [48]. The room reverberation time RT_{60} is 130ms and 250ms in a room size of 4.45m ×3.55m × 2.5m. Here, we use 8 sets of mixtures of female and male speech sources obtained from an array of microphones. The spacing distance of microphones is 5cm or 1m and the distances between sources and microphone pair center is 1.2m.

4.2 Estimation Evaluation Criteria

First, the estimated mixing matrix is evaluated by the mean square errors (MSEs) [30] as follows:

$$\mathrm{MSEs} = \min_{\pi_i \in \Pi} \frac{1}{N} \sum_{i=1}^{N} \| \frac{\mathbf{h}_i}{\| \mathbf{h}_i \|_2} - \frac{\hat{\mathbf{h}}_{\pi_i}}{\| \hat{\mathbf{h}}_{\pi_i} \|_2} \|_F^2, \quad (29)$$

where Π is the set of all permutations of $\{1, .., N\}$, $\mathbf{h}_i$ and $\hat{\mathbf{h}}_i$ are the original steering vector and the estimated version, respectively. The overall performance of mixing matrix estimation is obtained by averaging the calculated MSEs for all frequency bins.

Second, the estimated sources is evaluated by the criterion of [49], which is to calculate signal distortion ratio (SDR) between the target source signal and a series of decomposed terms of source signal including distortion, noises or errors. In general, the source signal can be decomposed into a sum of several components, such as

$$\hat{s}_i(t) = s_i^{target}(t) + e_i^{interf}(t) + e_i^{noise}(t) + e_i^{artif}(t), \quad (30)$$

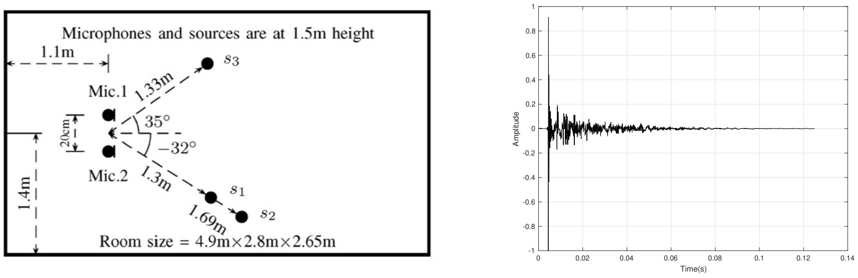
where $s_i^{target}(t)$, $e_i^{interf}(t)$, $e_i^{noise}(t)$, $e_i^{artif}(t)$ are the target source with allowed distortion, interferences, noises and artifacts error terms, respectively. Based on above decomposition, SDR at the ith source component is defined by

$$\mathrm{SDR}_i = 10log_{10}\times \frac{\|s_i^{target}(t)\|_F^2}{\|e_i^{interf}(t) + e_i^{artif}(t) + e_i^{noise}(t)\|_F^2}. \quad (31)$$

The average SDR is calculated by $\mathrm{SDR} = (\sum_{i=1}^{N} \mathrm{SDR}_i)/N$, which reflects the overall accuracy of source reconstruction performance.

4.3 Algorithm Settings

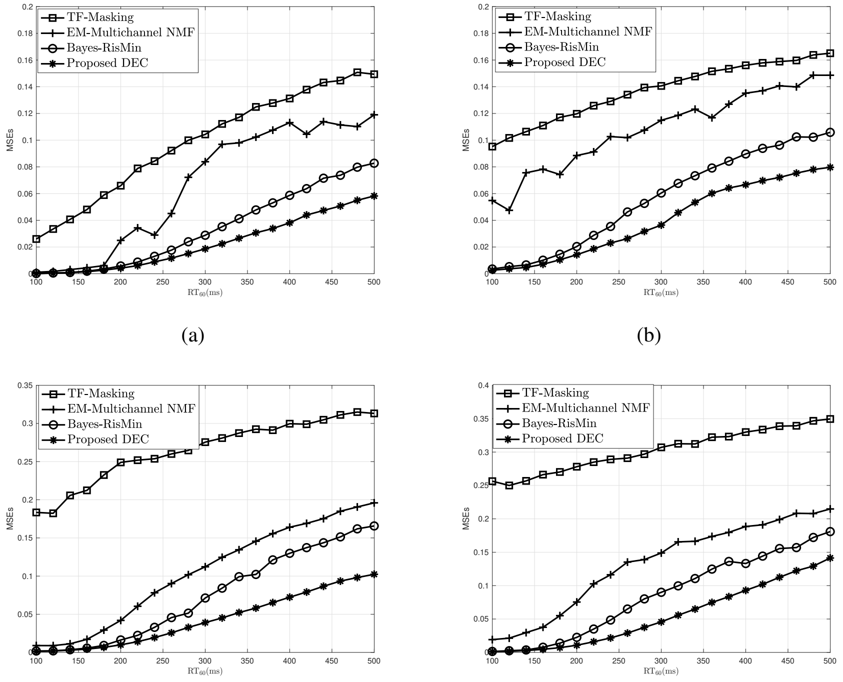
In the following, the proposed algorithm was tested with various experiments based on provided Datasets. All the experiments were carried out by a MacBook Air laptop with Intel Core i5, CPU 1.8 GHz under the system of macOS 10.13.6, and the programs were coded by Matlab R2018b. In the proposed algorithm, the window was selected as Hanning function. The window length F of



(a)

(b)

Figure 5: Settings of Dataset A. (a) source-microphone configurations, (b) measured authentic RIRs from first source to first microphone.



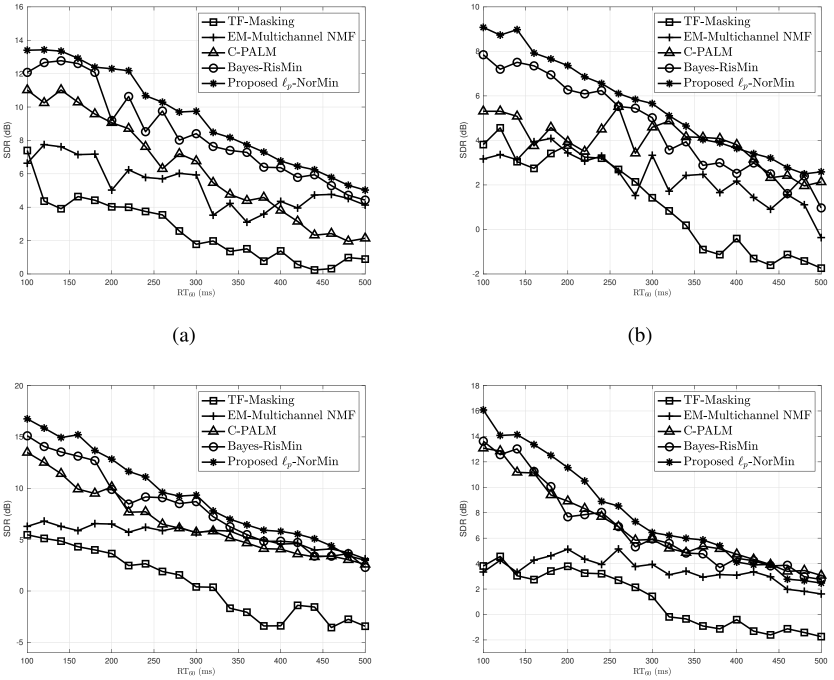
(a)

(b)

(c)

(d)

Figure 6: Channel estimation performance based on Dataset B: MSEs versus RT_{60} (Sec.). (a) $Case\ (M, N) = (2, 3)$, (b) $Case\ (M, N) = (2, 4)$, (c) $Case\ (M, N) = (3, 4)$, (d) $Case\ (M, N) = (4, 5)$.



(a)

(b)

(c)

(d)

Figure 7: Source reconstruction performance of Dataset B: SDRs versus RT_{60} (Sec.). (a) $Case\ (M, N) = (2, 3)$, (b) $Case\ (M, N) = (2, 4)$, (c) $Case\ (M, N) = (3, 4)$, (d) $Case\ (M, N) = (4, 5)$.

STFT is chosen to exploit local stationarity of speech signals while maintaining computational efficiency. It is pointed out in the work of [36] that large window length can decrease narrowband approximation error, thus we set the window length and the STFT frame shift as 2048 and 128 samples (8ms) in the following experiments, respectively. The speech signal duration was truncated as 10 seconds and the duration of each sub-block was set as 62.5ms, i.e., the recordings were partitioned into 160 segments.

For the convenience of following discussion, the proposed approaches on mixing matrix estimation and source reconstruction are labeled as 'Density-based Eigenvector Clustering (DEC)', 'ℓ_p-Norm-based Minimization (ℓ_p-NorMin) ', respectively. To demonstrate validity of provided methods, several state-of-the-art algorithms are compared in the following experiments. In the work of [22], the methods on mixing matrix estimation and source reconstruction are labeled as 'Fuzzy-C Means clustering (FCM)', 'TF-Masking', respectively; In the works of [15, 50], and [31], the methods are labeled as 'EM-based-Multichannel Nonnegative Matrix Factorization (EM-Multichannel NMF)', 'Convolutive- Proximal Alternating Linearized Minimization (C-PALM)' and 'Bayes Risk-based-Minimization (Bayes-RisMin)', respectively. It is worth noting that C-PALM is developed based on a convolutive narrowband system model, which means that the system in TF domain is still considered as a convolutive mixing model. In this case, the mixing matrix are estimated as a series of delayed versions. For this reason, we would not compare the performance of mixing matrix estimation by utilizing C-PALM in our following experiments.

4.4 Experiment Results

4.4.1 Selection of parameter p in ℓ_p-NorMin

Considering the parameter p of proposed ℓ_p-NorMin method plays an important role in source reconstruction performance, we provide a series of tests with various p to evaluate the impact of SDRs based on provided Dataset A, B and C. Table 2 presents the SDRs by setting various parameter of p ranging from 0.1 to 1 for $Case\ (M, N) = (2, 3)$. Based on provided results, we can observe that SDR slightly grows with increasing of p and reaches the peak when $p = 0.8$. In the following experiments, the parameter p is set as 0.8 for better performance. We can conclude that the ℓ_p-NorMin method is a flexible framework by adjusting the parameter of p to exploit sparsity of various types of sources.

4.4.2 Results on Dataset A

In this case, the mixtures is artificially convoluted by the speech signals and authentic RIRs in a real room. It is worth noting that sources $s_1(t)$ and $s_2(t)$ are located in the same direction to the center of microphones in this experiment. It follows that the corresponding steering vectors of mixing matrix, i.e., $\mathbf{h}_1$ and $\mathbf{h}_2$, are similar to each other, which increases the estimation difficulty. The results of mixing matrix estimation applying provided algorithms are as follows. The MSEs of proposed DEC is 0.012 while the MSEs of TF-Masking methodEM-Multichannel NMF and Bayes-RisMin are 0.097, 0.068 and 0.019, respectively. The results of source reconstruction applying provided algorithms are provided in Table 3. In general, the proposed ℓ_p-NorMin algorithm achieves a better averaged SDRs than other four algorithms by approximately 3.33 dB, 4.49 dB, 1.89 dB and 0.98 dB, respectively. Furthermore, we see that SDR_2 is relatively lower than SDR_1 and SDR_3 in all of algorithms due to the collinear interference of source $s_1(t)$ and $s_2(t)$. In general, the proposed method has improved the SDR performance of each source component, especially for the SDR of source $s_2(t)$.

4.4.3 Results on Dataset B

In this experiment, the proposed algorithms were tested with various parameter settings on Dataset B. In the aspect of mixing matrix estimation, the proposed DEC method is compared with the methods of FCM, EM-Multichannel NMF and Bayes-RisMin, whose results are illustrated in Fig.6. The mixing matrix performance of proposed DEC is better than the Bayes-RisMin for all of cases, especially when the reverberant time RT_{60} is over 200ms. Moreover, the proposed method outperforms FCM and EM-Multichannel NMF in all under-determined cases. In addition, The EM-Multichannel NMF yields better MSEs performance than FCM.

In the aspect of source reconstruction, the impact on SDRs performance with reverberant time were tested

Table 2: SDRs Evaluation of Various Parameter p in proposed ℓ_p-NorMin

L_p norm (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Dataset A	5.28	5.52	5.66	5.76	5.82	5.81	5.80	**5.85**	5.76	5.36
Dataset B	11.89	12.04	12.59	13.15	13.36	13.35	13.67	**13.76**	13.64	13.44
Dataset C	6.89	7.39	7.71	7.92	7.98	7.90	8.22	**8.24**	8.05	7.50

Table 3: SDRs Evaluation of Dataset A

	s_1	s_2	s_3	Average
TF-Masking	2.36	1.81	3.31	2.49
EM-Multichannel NMF	1.69	0.59	1.71	1.33
C-PALM	4.54	2.81	4.45	3.93
Bayes-RisMin	5.12	4.65	4.77	4.84
Proposed ℓ_p-NorMin	**6.60**	**4.77**	**6.08**	**5.82**

Table 4: SDRs Evaluation of Dataset C (Microphone Distance: 1m)

Dataset	'dev1' (synthetic)				'dev1' (live record)			
$Case\ (M,N)$	$(2,3)$				$(2,3)$			
RT$_{60}$	130ms		250ms		130ms		250ms	
Source type	Female	Male	Female	Male	Female	Male	Female	Male
TF-Masking	3.84	1.24	0.64	0.25	3.69	0.87	1.81	0.56
EM-Multichannel NMF	2.95	1.32	1.36	1.18	4.54	1.13	2.91	1.50
C-PALM	7.44	5.14	4.35	**3.91**	7.47	5.72	4.24	3.88
Bayes-RisMin	7.73	5.49	4.69	2.20	7.22	5.76	3.59	3.49
Proposed ℓ_p-NorMin	**8.76**	**5.54**	**4.88**	3.71	**8.11**	**5.96**	**5.18**	**3.90**

Table 5: SDRs Evaluation of Dataset C (Microphone Distance: 5cm)

Dataset	'dev1' (synthetic)				'dev1' (live record)			
$Case\ (M,N)$	$(2,3)$				$(2,3)$			
RT$_{60}$	130ms		250ms		130ms		250ms	
Source type	Female	Male	Female	Male	Female	Male	Female	Male
TF-Masking	1.16	0.34	1.02	0.75	1.82	1.32	-0.11	0.09
EM-Multichannel NMF	4.74	2.21	1.58	0.27	3.19	3.81	1.70	1.27
C-PALM	5.94	4.59	4.83	**3.61**	5.85	4.55	3.42	**3.39**
Bayes-RisMin	7.34	**5.89**	4.85	3.12	6.39	5.86	3.37	2.06
Proposed ℓ_p-NorMin	**8.24**	5.52	**5.09**	3.05	**6.42**	**6.25**	**3.64**	2.53

based on Dataset B. Fig.7 illustrates the average SDR curves with various RT_{60}. The SDRs of all proposed algorithms descend when RT_{60} is increasing from 100ms to 500ms. It is worth noting that the provided ℓ_p NorMin method is slightly better than the results of TF-Masking, EM-Multichannel NMF, C-PALM and Bayes-RisMin, respectively. Moreover, TF-Masking and EM-Multichannel NMF exhibit slightly better performance when RT_{60} is less than 160ms. Based on above results, we utilize Bayes-RisMin for benchmarking as it demonstrates superior performance to other provided algorithms in provided underdetermined cases.

4.4.4 Results on Dataset C

In this experiment, the algorithms were tested based on Dataset C where synthetic data and real recorded data are both taken into account. Table 4 and Table 5 present the average SDRs by setting the microphone distance at 5cm and 1m for $Case\ (M,N) = (2,3)$, respectively. The proposed ℓ_p-NorMin is slightly better than the Bayes-RisMin in most of cases including the synthetic and real recorded data. C-PALM achieves a better SDR result in several highly reverberant scenario such as the case of male speech when RT_{60} is $250ms$. Moreover, the proposed ℓ_p-NorMin, C-PALM and Bayes-RisMin achieve stable SDRs performance while EM-Multichannel NMF and TF-Masking do not work well in these experiments. The SDR results of proposed ℓ_p-NorMin is still approximately 3 dB higher than EM-Multichannel NMF and TF-Masking of all cases. Similar results also can be found in Table 5. Overall, the provided experiments show that the SDRs of proposed two-stage algorithm is competitive than C-PALM and Bayes-RisMin in various tests, and it outperforms the other two algorithms, i.e., EM-Multichannel NMF and TF-Masking.

5 Conclusion

A new two-stage strategy for solving the underdetermined convolutive BSS problem has been proposed in this paper. At the first stage, we transform the mixing matrix estimation to an eigenvector clustering problem. First, the eigenvectors were extracted by exploiting the rank-one structure of the local covariance matrices of mixture signals; second, these eigenvectors were clustered by a density-based clustering method to give clusters; third, a wight clustering scheme was applied to adjust the clusters to give the estimated mixing matrix. At the second stage, we transform the source reconstruction to a sparse minimization model based on the ℓ_p norm $(0 < p \leq 1)$, whose solution has solved by an iterative Lagrange multiplier method with a proper initialization. The experiment results have demonstrated the effectiveness of the proposed two-stage algorithm compared to the state-of-the-art methods in various under-determined BSS cases. As the future work, we will extend the study of underdetermined BSS in TF domain from a linear narrowband system to a convolutive narrowband systems, which is more suitable to depict the source separation problem in a highly reverberant environment [50]. The two-stage scheme based on the convolutive narrowband system will be an interesting problem to investigate.

6 Acknowledgment

This work was supported in part by China Postdoctoral Science Foundation under grants 2018M643022. This work was also supported in part by the National Natural Science Foundation of China under grants 61722304, 61333013, 61603099. This work was also supported in part by Guangdong Natural Science Funds under grants 2014A030306037, 2018A0303130080. This work was also supported in part by Pearl River S&T Nova Program of Guangzhou under grants 201806010176.

7 Appendix

7.1 Derivation of Proposed ℓ_p Norm Minimization Method

According to the optimization model of (19), the gradient of objective function respects to $\mathbf{s}_d$ and $\boldsymbol{\alpha}$ are derived as follows, respectively,

$$\begin{cases} \frac{\partial \mathcal{F}(\boldsymbol{\alpha},\mathbf{s}_d)}{\partial \mathbf{s}_d} = \frac{\partial \mathcal{J}(\mathbf{s}_d)}{\partial \mathbf{s}_d} + \hat{\mathbf{H}}^H \boldsymbol{\alpha}, \\ \frac{\partial \mathcal{F}(\boldsymbol{\alpha},\mathbf{s}_d)}{\partial \boldsymbol{\alpha}} = \mathbf{x}_d - \hat{\mathbf{H}}\mathbf{s}_d = \mathbf{0}, \end{cases} \tag{32}$$

where $\frac{\partial \mathcal{J}(\mathbf{s}_d)}{\partial \mathbf{s}_d} = p\boldsymbol{\Psi}(\mathbf{s}_d)\mathbf{s}_d$. Note that the optimal solution can be obtained if $\frac{\partial \mathcal{F}(\boldsymbol{\alpha},\mathbf{s}_d)}{\partial \mathbf{s}_d} = \mathbf{0}$, thus we have

$$\hat{\mathbf{H}}^H \boldsymbol{\alpha} = -p\boldsymbol{\Psi}(\mathbf{s}_d)\mathbf{s}_d. \tag{33}$$

Multiply $\boldsymbol{\Psi}^{-1}(\mathbf{s}_d)$ on both sides of (33), such as

$$\boldsymbol{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H \boldsymbol{\alpha} = -p\mathbf{s}_d. \tag{34}$$

Next, multiply $\hat{\mathbf{H}}$ on both sides of (34), such as

$$\hat{\mathbf{H}}\boldsymbol{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H\boldsymbol{\alpha} = -p\mathbf{x}_d. \tag{35}$$

Multiply $(\hat{\mathbf{H}}\boldsymbol{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H)^{-1}$ on both sides of (35), such as

$$\boldsymbol{\alpha} = -p(\hat{\mathbf{H}}\boldsymbol{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H)^{-1}\mathbf{x}_d. \tag{36}$$

Using the results of (34) and (36), we finally have

$$\mathbf{s}_d = \boldsymbol{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H(\hat{\mathbf{H}}\boldsymbol{\Psi}^{-1}(\mathbf{s}_d)\hat{\mathbf{H}}^H)^{-1}\mathbf{x}_d. \tag{37}$$

It is worth noting that in some circumstance the active source component is less than M, i.e., $\|\mathbf{s}_d\|_0 < M$. In this case, we can substitute $(\boldsymbol{\Psi}(\mathbf{s}_d) + \epsilon\mathbf{I})^{-1}$ instead of $\boldsymbol{\Psi}^{-1}(\mathbf{s}_d)$ to avoid the ill-conditioned matrix inversion problem in (37).

References

[1] A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing. New York: Wiley, 2003.

[2] S. Makino,T. W. Lee, and H. Sawada, Blind Speech Separation. Berlin: Springer-Verlag, 2007.

[3] M. Stanaćević, S. Li, and G. Cauwenberghs, “Micropower mixed-signal VLSI independent component analysis for gradient flow acoustic source separation,”*IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 7, pp. 972-981, Jul. 2016.

[4] Z. Yang, G. Zhou, S. Xie, S. Ding, J. Yang, and J. Zhang, “Blind spectral unmixing based on sparse nonnegative matrix factorization,”*IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1112-1125, 2011.

[5] K. Rahbar, J. Reilly, and J. H. Manton, “Blind identification of MIMO-FIR systems driven by quasistationary sources using second order statistics: A frequency domain approach,”*IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 406-417, Feb. 2004.

[6] G. Zhou, Q. Zhao, Y. Zhang, T. Adal, S. Xie, and A. Cichocki, “Linked component analysis from matrices to high-order tensors: Applications to biomedical data,”*Proceedings of the IEEE*, vol. 104, no. 2, pp. 310-331, Feb. 2016.

[7] T. Kim, “Real-time independent vector analysis for convolutive blind source separation,”*IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1431-1438, Jul. 2010.

[8] S. X. Ding, J. Huang, D. M. Wei, and A. Cichocki, “A near real-time approach for convolutive blind source separation,”*IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 53, no. 1, pp. 114-128, Jan. 2006.

[9] B. Gao, W. L. Woo, and S. S. Dlay, “Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and Itakura-Saito nonnegative matrix two-dimensional factorizations,”*IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 3, pp. 662-675, Mar. 2013.

[10] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,”*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066-1074, Mar. 2007.

[11] P. Comon, “Blind identification and source separation in 2 × 3 underdetermined mixtures,”*IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 11-22, Jan. 2004.

[12] P. Comon and C. Jutten, Handbook of Blind Source Separation-Independent Component Analysis and Applications. Academic Press, 2010.

[13] Z. Yang, Y. Xiang, K. Xie, and Y. Lai, “Adaptive method for nonsmooth nonnegative matrix factorization,”*IEEE Trans. Neural Networks Learning Syst.*, vol. 28, no. 4, pp. 948-960, 2017.

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

[15] 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.

[16] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1-12, Jan. 2007.

[17] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140-2151, 2013.

[18] F. Feng and M. Kowalski,"Sparsity and low-rank amplitude based blind source separation,"in *2017 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 571-575.

[19] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483-492, Mar. 2016.

[20] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking,"*IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830-1847, Jul. 2004.

[21] D. S. Williamson and D. Wang,"Time-frequency masking in the complex domain for speech dereverberation and denoising,"*IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492-1501, Jul. 2017.

[22] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time-frequency masking,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 101-116, Jan. 2010.

[23] S. Rickard, The DUET Blind Source Separation Algorithm. Springer Netherlands, 2007, pp. 217-241.

[24] Z. He, S. Xie, S. Ding, and A. Cichocki, "Convolutive blind source separation in the frequency domain based on sparse representation,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1551-1563, Sep. 2007.

[25] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture,"*IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 121-133, Jan. 2010.

[26] K. Rahbar and J. P. Reilly,"A frequency domain method for blind source separation of convolutive audio mixtures,"*IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 832-844, Sep. 2005.

[27] D. Nion and N. D. Sidiropoulos,"Adaptive algorithms to track the PARAFAC decomposition of a third-order tensor,"*IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2299-2310, Jun. 2009.

[28] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1193-1207, Aug. 2010.

[29] P. Tichavský, and Z. Koldovský,"Weight adjusted tensor method for blind separation of underdetermined mixtures of nonstationary sources,"*IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1037-1047, Mar. 2011.

[30] 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.

[31] J. Cho and C. Yoo,"Underdetermined convolutive BSS: Bayes risk minimization based on a mixture of super-Gaussian posterior approximation,"*IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 828-839, May. 2015.

[32] E. Vincent,"Complex nonconvex Lp norm minimization for underdetermined source separation,"in *Int. Conf. Ind. Compon. Anal. and Signal Separat.*, 2007, pp. 430-437.

[33] T. Chan, W. Ma, C. Chi, and Y. Wang,"A convex analysis framework for blind separation of nonnegative sources,"*IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5120-5134, Oct. 2008.

[34] C. Lin, C. Chi, L. Chen, D. J. Miller, and Y. Wang,"Detection of sources in non-negative blind source separation by minimum description length criterion,"*IEEE Trans. Neural Networks Learning Syst.*, vol. 29, no. 9, pp. 4022-4037, Sep. 2018.

[35] A. Rodriguez and L. Alessandro, "Clustering by fast search and find of density peaks,"*Science*, vol. 344, no. 6191, pp. 1492-1496, 2014.

[36] W. Kellermann and H. Buchner, "Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain,"in *The Thrity-Seventh IEEE Asilomar Conf. Signals, Syst. Comput.*, 2003, vol. 2, pp. 1278-1282.

[37] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830-1840, Sep. 2010.

[38] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture,"*IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 121-133, Jan. 2010.

[39] S. Markovich-Golan and S. Gannot,"Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method, "in *2015 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2015, pp. 544-548.

[40] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR, "*IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780-793, Apr. 2017.

[41] G. Zhou, A. Cichocki, Y. Zhang, and D. P. Mandic, "Group component analysis for multiblock data: Common and individual feature extraction,"*IEEE Trans. Neural Networks Learning Syst.*, vol. 27, no. 11, pp. 2426-2439, Nov. 2016.

[42] J. Yang and H. Liu,"Blind identification of the underdetermined mixing matrix based on K-weighted hyperline clustering,"*Neurocomput.*, vol. 149, no. PB, pp. 483-489, Feb. 2015.

[43] K. Xie, Z. He, and A. Cichocki, "Convergence analysis of the FOCUSS algorithm,"*IEEE Trans. Neural Netw. Learning Syst.*, vol. 3, no. 26, pp. 601-613, Mar. 2015.

[44] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516-527, Mar. 2011.

[45] http://www.purebits.com, [Online].

[46] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics,"*J. Acoust. Soc. Amer.*, vol. 65, no. 4, Apr. 1979.

[47] http://sisec2011.wiki.irisa.fr/tiki-index.html, [Online].

[48] Campbell, http://media.paisley.ac.uk/campbell/Roomsim/, [Online].

[49] 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.

[50] F. Feng and M. Kowalski, "Underdetermined reverberant blind source separation: Sparse approaches for multiplicative and convolutive narrowband approximation, "*IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 442-456, Feb. 2019.