BLIND SEPARATION OF CONVOLUTIVE MIXTURES OF SPEECH SOURCES: EXPLOITING LOCAL SPARSITY

Xiao Fu and Wing-Kin Ma

Department of Electronic Engineering, The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
E-mail: xfu@ee.cuhk.edu.hk, wkma@ieee.org

ABSTRACT

This paper presents an efficient method for blind source separation of convolutively mixed speech signals. The method follows the popular frequency-domain approach, wherein researchers are faced with two main problems, namely, per-frequency mixing system estimation, and permutation alignment of source components at all frequencies. We adopt a novel concept, where we utilize local sparsity of speech sources in transformed domain, together with non-stationarity, to address the two problems. Such exploitation leads to a closed-form solution for per-frequency mixing system estimation and a numerically simple method for permutation alignment, both of which are efficient to implement. Simulations show that the proposed method yields comparable source recovery performance to that of a state-of-the-art method, while requires much less computation time.

Index Terms— Blind Source Separation, Convolutive Mixture, Speech Separation, Permutation Ambiguity

1. INTRODUCTION

1.1. Background

We consider blind separation of convolutive mixtures of speech sources. This problem has attracted much interest, motivated by real-world applications such as teleconferencing and mobile telephony. Compared with the instantaneous mixtures, convolutive mixtures are known to be more difficult to separate. One way to tackle this problem is to employ the so-called frequency-domain approach [1, 2]: by transforming signals to the frequency domain, the convolutive mixtures are decoupled into many per-frequency instantaeous mixtures and thus the existing blind source separation (BSS) algorithms for instantaneous mixtures can be directly applied [3–6].

However, the difficulty of the frequency-domain approach lies in implementation efficiency. First, by the frequency-domain approach, a frequency-dependent mixing system needs to be estimated at each frequency. Hence, the computation load of the adopted BSS algorithm is scaled up by the number of frequencies, which could be very large. Second, after the mixing system estimation at each frequency, a permutation alignment stage needs to be considered. This is because that the estimated mixing systems may be permuted differently from one frequency to another, which could result in false alignment of frequency-components of sources and thus compromise the source recovery performance [2]. Permutation alignment usually involves solving highly nonconvex optimization problems [1,7] or clustering thousands of vectors with large size [8], both of which may be time consuming. Thus, computationally cheap per-frequency mixing system estimation algorithms and fast permutation alignment methods are desirable when dealing with practical blind speech separation.

This work is supported by a General Research Fund of Hong Kong Research Grant Council (CUHK415509).

1.2. Review of Prior Works

Prior works adopt techniques such as independent component analysis (ICA) [9], joint diagonalizaton (JD) [2] or three-way tensor decomposition [1, 8] to tackle the per-frequency mixing system estimation problem. These techniques are developed based on the exploitations of the statistical independence between sources and time-varying characteristics of source power profiles. For permutation alignment, an arguably popular class of methods make use of the fact that frequency-components of the same source are correlated in some way, especially in neighboring frequencies. Such methods are essentially based on evaluating the correlations between specified features of source components at adjacent frequencies, and then aligning highly correlated ones together [1, 10]. Using the same insight, various algorithms are introduced: in [9, 11, 12], besides the local alignment between adjacent frequencies, a global adjustment for all frequencies is also incorporated for better robustness; in [8], K-means clustering is employed for efficiency improvement.

1.3. Contributions

In this paper, we focus on developing an efficient method for frequency-domain blind speech separation. The distinguishing feature of this work is that the local sparsity property of speech sources, i.e., the local disjointness of source supports in transform domain, is extensively exploited. Consequently, we come up with efficient algorithms for both mixing system estimation and permutation alignment, which show different flavors compared with existing methods. To be specific, we first propose to employ a recently devised BSS algorithm [13] for per-frequency mixing system estimation. By exploiting the local sparsity, this algorithm admits a closed-form solution to the per-frequency mixing system estimation problem, rather than using iterative optimization algorithms as in existing methods. We then propose a local sparsity-based permutation alignment method, which does not rely on the correlation evaluations of source components and also admits an easy-to-implement structure. Simulations in an artificial room using speech sources demonstrate the efficacy of the proposed methods.

2. PROBLEM STATEMENT

Consider the convolutive mixture model [1, 2], i.e.,

$$\mathbf{x}(t) = \sum_{\tau=0}^{\tau_{\max}-1} \mathbf{A}(\tau)\mathbf{s}(t-\tau), \quad t = 0, 1, 2, \ldots \tag{1}$$

where $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^T \in \mathbb{R}^N$ denotes the received signals by sensors, $\mathbf{s}(t) = [s_1(t), \ldots, s_K(t)]^T \in \mathbb{R}^K$ denotes K

mutually independent speech sources, τ is the index of delay, $\tau_{\max}$ represents the maximal number of delays and $\mathbf{A}(\tau) \in \mathbb{R}^{N\times K}$ represents the impulse response of the mixing system. When $\tau_{\max} = 1$, the signal model in (1) becomes the instantaneous mixture model. We consider overdetermined mixing systems in this paper, i.e., $N > K$. The objective is to recover $\mathbf{s}(t)$ from the mixtures $\mathbf{x}(t)$ without knowing the mixing system. To achieve this goal, we employ the *frequency-domain approach* [1, 8]. Specifically, by applying the short time Fourier transform (STFT) on consecutive time blocks of $\mathbf{x}(t)$, we can obtain an approximately instantaneous model at multiple frequencies f_ℓ, $\ell = 0, \ldots, \ell_{\max} - 1$, where $\ell_{\max}$ denotes the number of frequencies; i.e., we have

$$\check{\mathbf{x}}_\ell(q) \approx \mathbf{A}_\ell \check{\mathbf{s}}_\ell(q), \quad q = 0, 1, 2, \ldots, \tag{2}$$

where $\mathbf{A}_\ell = [\, \mathbf{a}_{1,\ell}, \ldots, \mathbf{a}_{K,\ell}\,] \in \mathbb{C}^{N\times K}$ is a frequency-dependent mixing matrix, $\mathbf{a}_{k,\ell} \in \mathbb{C}^N$ denotes the spatial channel from source k to the sensors at frequency f_ℓ, $\check{\mathbf{x}}_\ell(q) = [\, \check{x}_{1,\ell}(q), \ldots, \check{x}_{N,\ell}(q)\,]^T \in \mathbb{C}^N$ and $\check{\mathbf{s}}_\ell(q) = [\, \check{s}_{1,\ell}(q), \ldots, \check{s}_{K,\ell}(q)\,]^T \in \mathbb{C}^K$ are the frequency-components of the mixture and sources at f_ℓ, obtained by applying STFT on the qth time block of $\mathbf{x}(t)$ and $\mathbf{s}(t)$, respectively. Readers are referred to [1,8] for details of obtaining such transformed signals.

It is known that speech sources can be regarded as wide-sense stationary signals in short durations [8] . Specifically, if we chop $\check{\mathbf{s}}_\ell(q)$ into short frames of length L and define the local covariance of sources in frame m by

$$\mathbf{D}_\ell[m] = \mathrm{E}\{\check{\mathbf{s}}_\ell(q)\check{\mathbf{s}}_\ell(q)^H\}, \quad q \in [\,(m-1)L+1, mL\,],$$

the local covariance $\mathbf{D}_\ell[m]$ can be regarded as being static within frame m. Since speech sources are generally non-stationary in long term, $\mathbf{D}_\ell[m]$ varies from frame to frame. Note that due to the independence assumption of sources, $\mathbf{D}_\ell[m]$ is a diagonal matrix; i.e., we have $\mathbf{D}_\ell[m] = \mathrm{Diag}(\mathbf{d}_\ell[m])$, where $\mathbf{d}_\ell[m] = [d_{1,\ell}[m], \ldots, d_{K,\ell}[m]]^T$ is a power vector of sources at frame m and $d_{k,\ell}[m] = \mathrm{E}\{|\check{s}_{k,\ell}(q)|^2\}$, for $q \in [\,(m-1)L+1, mL\,]$, denotes the power of source k in frame m. We may also define the local covariance of $\check{\mathbf{x}}_\ell(q)$ in frame m by

$$\mathbf{R}_\ell[m] = \mathrm{E}\{\check{\mathbf{x}}_\ell(q)\check{\mathbf{x}}_\ell(q)^H\} = \sum_{k=1}^K d_{k,\ell}[m]\mathbf{a}_{k,\ell}\mathbf{a}_{k,\ell}^H,$$

where $q \in [\,(m-1)L+1, mL\,]$. In practice, $\mathbf{R}_\ell[m]$ can be estimated by local sampling, i.e., $\mathbf{R}_\ell[m] \approx (1/L)\sum_{q=(m-1)L+1}^{mL} \check{\mathbf{x}}_\ell(q)\check{\mathbf{x}}_\ell(q)^H$. By the frequency-domain approach, the BSS problem amounts to estimating $\mathbf{A}_\ell$ at each frequency f_ℓ using $\mathbf{R}_\ell[m]$'s. This is merely a BSS problem for instantaneous mixture, which has been extensively studied [3–6]. However, when the number of frequencies is large, the amount of computation resource for the BSS process will be scaled up. Therefore, it is motivated to consider efficient per-frequency BSS algorithms.

3. PER-FREQUENCY MIXING SYSTEM ESTIMATION

We propose to employ a recently devised BSS algorithm for the instantaneous mixture model [13] at each frequency to estimate $\mathbf{A}_\ell$. This algorithm admits a closed-form solution and thus is very efficient. In this section, we will briefly review this algorithm. The idea is to exploit the local sparsity of sources in time-frequency domain; i.e., we assume that

(A1)(*local dominance*) for each source k and at each frequency f_ℓ, there exists a frame m_{ℓ_k} such that $d_{k,\ell}[m_{\ell_k}] > 0$ and $d_{j,\ell}[m_{\ell_k}] = 0, \forall j \neq k$.

Physically (A1) means that there exist frames at each frequency f_ℓ, in which only one source is active and the others are inactive. In other words, the source supports are locally disjoint at these frames and thus these frames are dominated by one source. This assumption is considered reasonable for sources like speech and audio, which have been demonstrated to exhibit sparsity in transformed domains [14, 15]. By (A1), in those frames locally dominated by source k, the local covariances take a rank-one form; i.e., we have

$$\mathbf{R}_\ell[m_{\ell_k}] = d_{k,\ell}[m_{\ell_k}]\mathbf{a}_{k,\ell}\mathbf{a}_{k,\ell}^H.$$

If these locally dominant covariances are identified, we can estimate $\mathbf{a}_{k,\ell}$ by taking the principal eigenvector of $\mathbf{R}_\ell[m_{\ell_k}]$; i.e., the estimated mixing matrix $\hat{\mathbf{A}}_\ell = [\hat{\mathbf{a}}_{1,\ell}, \ldots, \hat{\mathbf{a}}_{K,\ell}]$ can be obtained by

$$\hat{\mathbf{a}}_{k,\ell} = \boldsymbol{q}_{\max}\left(\mathbf{R}_\ell[\hat{m}_{\ell_k}]\right), \quad k = 1, \ldots, K, \tag{3}$$

where $\boldsymbol{q}_{\max}(\mathbf{X})$ denotes the principal eigenvector of $\mathbf{X}$.

The local dominance-based BSS (LD-BSS) algorithm [13] aims at identifying K locally dominant frames corresponding to different sources. This algorithm can be described as follows. By vectorizing all local covariances, we obtain

$$\begin{aligned}\mathbf{y}_\ell[m] = \mathrm{vec}(\mathbf{R}_\ell[m]) &= (\mathbf{A}_\ell^* \odot \mathbf{A}_\ell)\,\mathbf{d}_\ell[m] \\ &= \sum_{k=1}^K d_{k,\ell}[m]\mathbf{a}_{k,\ell}^* \otimes \mathbf{a}_{k,\ell},\end{aligned}$$

where $\odot$ and $\otimes$ denote the Khatri-Rao product and the Kronecker product, respectively. Based on this structure of $\mathbf{y}_\ell[m]$, it can be shown that

$$\begin{aligned}\frac{\|\mathbf{y}_\ell[m]\|_2}{\mathrm{Tr}(\mathbf{R}_\ell[m])} &= \frac{\|\sum_{k=1}^K d_{k,\ell}[m]\mathbf{a}_{k,\ell}^* \otimes \mathbf{a}_{k,\ell}\|_2}{\sum_{k=1}^K d_{k,\ell}[m]\|\mathbf{a}_{k,\ell}\|_2^2} \\ &\le \frac{\sum_{k=1}^K d_{k,\ell}[m]\|\mathbf{a}_{k,\ell}^* \otimes \mathbf{a}_{k,\ell}\|_2}{\sum_{k=1}^K d_{k,\ell}[m]\|\mathbf{a}_{k,\ell}\|_2^2} = 1,\end{aligned} \tag{4}$$

where the inequality is resulted from the triangle inequality of 2-norm and the non-negativity of $d_{k,\ell}[m]$; the last equality is obtained by the fact $\|\mathbf{a}_{k,\ell}^* \otimes \mathbf{a}_{k,\ell}\|_2 = \|\mathbf{a}_{k,\ell}\mathbf{a}_{k,\ell}^H\|_F^2 = \|\mathbf{a}_{k,\ell}\|_2^2$. Note that the inequality in (4) holds if and only if $\mathbf{d}_\ell[m]$ is a unit vector [13], i.e., frame m is locally dominated by a source at frequency f_ℓ. According to this observation, we can do the following. Initially, we identify the first locally dominant frame at frequency f_ℓ by

$$\hat{m}_{\ell_1} = \arg\max_{m=1,\ldots,M} \frac{\|\mathbf{y}_\ell[m]\|_2}{\mathrm{Tr}(\mathbf{R}_\ell[m])}. \tag{5}$$

Then, a successive approach can be employed to identify the remaining locally dominant frames at frequency f_ℓ. To present the process, suppose that we have found $k-1$ locally dominant frames, indexed by $\hat{m}_{\ell_1}, \ldots, \hat{m}_{\ell_{k-1}}$. Also, denote $\mathbf{Y}_{k-1,\ell} = [\, \mathbf{y}_\ell[\hat{m}_{\ell_1}], \ldots, \mathbf{y}_\ell[\hat{m}_{\ell_{k-1}}]\,]$ and let $\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^\dagger\mathbf{X}^T$ be the orthogonal complement projector of $\mathbf{X}$. Given these notations and by similar derivations as in Eq. (4), it can be easily shown that

$$\hat{m}_{\ell_k} = \arg\max_{m=1,\ldots,M} \frac{\left\|\mathbf{P}_{\mathbf{Y}_{k-1,\ell}}^{\perp}\mathbf{y}_\ell[m]\right\|_2}{\mathrm{Tr}(\mathbf{R}_\ell[m])}, \quad k > 1, \tag{6}$$

corresponds to a frame locally dominated by a new source (not the sources dominating frames indexed by $\hat{m}_{\ell_1}, \ldots, \hat{m}_{\ell_{k-1}}$).

In summary, the LD-BSS algorithm consists of Eq. (5)-(6), which take closed-form expressions and involve only basic matrix operations, such as 2-norm computation and multiplication. Hence, this process can be carried out quite efficiently at each frequency. This is a desirable feature for frequency-domain approach-based speech separation, especially when $\ell_{\max}$ is large.

4. PERMUTATION ALIGNMENT

One may notice that, although by Eq. (5)-(6) we can identify indices of the locally dominant frames $\hat{m}_{\ell_1}, \ldots, \hat{m}_{\ell_K}$ at frequency f_ℓ, there is no knowledge revealed about which source dominates frame $\hat{m}_{\ell_k}$, according to the process. The scaling factor of $\mathbf{a}_{k,\ell}$ is also unknown since the obtained $\hat{\mathbf{a}}_{k,\ell}$ always admits unit 2-norm. In other words, as in other BSS methods, the permutation and scaling ambiguities exist when applying LD-BSS; i.e, we have

$$\hat{\mathbf{A}}_\ell = \mathbf{A}_\ell \mathbf{P}_\ell \mathbf{\Lambda}_\ell,$$

where $\mathbf{P}_\ell$ is a permutation matrix and $\mathbf{\Lambda}_\ell$ is a full rank diagonal matrix. It is necessary to compensate the effects brought by these two ambiguities before doing source recovery. In practice, one may employ the *minimum distortion principle* (MDP) method to deal with the scaling ambiguity. See [8, 11] and references therein. In the sequel, we will focus on solving the permutation ambiguity problem.

Our idea is, again, to make use of the local dominance frames, by the observation that same frame is usually dominated by the same source at neighboring frequencies. Specifically, if $d_{k,\ell}[m] > 0$ and $d_{j,\ell}[m] = 0$ for $j \neq k$, it is likely that $d_{k,\ell\pm1}[m] > 0$ and $d_{j,\ell\pm1}[m] = 0$ for $j \neq k$, especially when the frequency grids are dense. This means that the local power dominance relationship between sources exhibits some similarity at frequencies close by. In this paper, instead of evaluating the correlations of source features across frequencies as existing algorithms do, we exploit the dominance similarity in local frames to come up with a simple method with low implementation complexity.

Recall that at frequency f_ℓ, we have obtained an index set of locally dominant frames by Eq. (5)-(6), denoted by $\mathcal{D}_\ell = \{\hat{m}_{\ell_1}, \ldots, \hat{m}_{\ell_K}\}$, where $\hat{m}_{\ell_k}$ is the index of the kth identified locally dominant frame, corresponding to the kth column in $\hat{\mathbf{A}}_\ell$. The key insight is that if the same source dominates $\hat{m}_{\ell_k}$ at f_ℓ and the neighboring frequency $f_{\ell-1}$, we must have

$$\frac{\mathbf{d}_\ell[\hat{m}_{\ell_k}]}{\|\mathbf{d}_\ell[\hat{m}_{\ell_k}]\|_1} = \frac{\mathbf{d}_{\ell-1}[\hat{m}_{\ell_k}]}{\|\mathbf{d}_{\ell-1}[\hat{m}_{\ell_k}]\|_1}, \quad \forall \hat{m}_{\ell_k} \in \mathcal{D}_\ell.$$

In other words, the normalized source power vectors in frame m_{ℓ_k} are identical unit vectors at frequency f_ℓ and $f_{\ell-1}$. Therefore, the permutation alignment problem amounts to permuting the estimated source power vector in frame m_{ℓ_k} at f_ℓ, such that it can be aligned to the source power vector at $f_{\ell-1}$; i.e., by letting the estimated source power vector be

$$\hat{\mathbf{d}}_\ell[m] = (\hat{\mathbf{A}}_\ell^* \odot \hat{\mathbf{A}}_\ell)^\dagger \mathbf{y}_\ell[m], \tag{7}$$

we want to find a permutation matrix $\hat{\mathbf{P}}_\ell$, such that

$$\hat{\mathbf{P}}_\ell^T \frac{\hat{\mathbf{d}}_\ell[\hat{m}_{\ell_k}]}{\|\hat{\mathbf{d}}_\ell[\hat{m}_{\ell_k}]\|_1} = \frac{\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell_k}]}{\|\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell_k}]\|_1}, \quad \forall \hat{m}_{\ell_k} \in \mathcal{D}_\ell. \tag{8}$$

Given the same source dominating $\hat{m}_{\ell_k}$ at frequencies f_ℓ and $f_{\ell-1}$, it can be easily verified that

$$\hat{\mathbf{P}}_\ell = \left[\frac{\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell_1}]}{\|\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell_1}]\|_1}, \ldots, \frac{\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell_K}]}{\|\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell_K}]\|_1} \right]^T, \tag{9}$$

can satisfy the identity (8). Note that if a source dominates frame $\hat{m}_{\ell_k}$ at f_ℓ and $f_{\ell-1}$ simultaneously, each column of such $\hat{\mathbf{P}}_\ell$ only admits one non-zero element and thus is a permutation matrix. Eq. (9) means that to remove permutation ambiguity at frequency f_ℓ, one just needs to identify the indices of locally dominant frames at frequency f_ℓ and compute the source power vectors in these frames at frequency $f_{\ell-1}$. As permuting the source power vector by $\hat{\mathbf{P}}_\ell^T \hat{\mathbf{d}}_\ell[m]$ is equivalent to permuting the columns of the mixing matrix by $\hat{\mathbf{P}}_\ell$, we can remove the permutation ambiguity at frequency f_ℓ by using $\hat{\mathbf{A}}_\ell \hat{\mathbf{P}}_\ell$ as the permutation compensated mixing matrix.

The pseudo code of the permutation alignment process is presented in Algorithm 1. It can be seen that this permutation alignment procedure is quite simple to implement. The mapping process in line 4 is added because in practice, owing to modeling errors, the calculated $\hat{\mathbf{P}}_\ell$ may not be a permutation matrix precisely. Therefore, the closest (in Euclidean distance sense) permutation matrix to it may serve as a good surrogate. Besides this approximation, we are also aware of the possibility that the sequential alignment scheme might suffer error accumulation [1]. Nevertheless, as an easily implemented method, the proposed procedure exhibits quite satisfactory performance, as will be demonstrated in the next section.

Algorithm 1: Local Sparsity-based Permutation Alignment

input : $\hat{\mathbf{A}}_\ell, \mathbf{y}[m], \mathcal{D}_\ell, \forall \ell, m$;
1 **for** $\ell = 1, \ldots, \ell_{\max-1}$ **do**
2 obtain $\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell_k}], \forall \hat{m}_{\ell_k} \in \mathcal{D}_\ell$, by Eq. (7);
3 construct $\hat{\mathbf{P}}_\ell$ by Eq. (9);
4 map $\hat{\mathbf{P}}_\ell$ to a permutation matrix $\tilde{\mathbf{P}}_\ell$ via Hungarian algorithm [1].
5 align the permutation by $\hat{\mathbf{A}}_\ell := \hat{\mathbf{A}}_\ell \tilde{\mathbf{P}}_\ell$;
6 **end**
output: $\{\hat{\mathbf{A}}_\ell\}_{\ell=0}^{\ell_{\max}-1}$.

5. SIMULATION

In this section, we use simulations to demonstrate the performance of the proposed local sparsity-based speech separation methods. To build up convolutive mixtures similar to those obtained in real world, we follow the image method in [17] to simulate a room with sound reflective walls. This artificial room is set to have a size of 5m×3.5m×3m. One typical channel response (with reverberation time $T60 = 120$ms) under such scenario is plotted in Fig. 1. We set $(K, N) = (4, 6)$ in our simulations. Sources are set in positions $(1, 0.8, 1.6)$, $(1, 1.6, 1.6)$, $(1, 2.4, 1.6)$ and $(1, 3.2, 1.6)$; sensors are in $(4, 0.5, 1.6)$, $(4, 1, 1.6)$, $(4, 1.5, 1.6)$, $(4, 2, 1.6)$, $(4, 2.5, 1.6)$ and $(4, 3, 1.6)$. We create a noisy environment by adding Gaussian distributed noise to the received signal $\mathbf{x}(t)$.The source data base provided by [8] is employed, which contains 8 cleanly recorded speeches. Each source in this data base is 16KHz sampled and we truncate them into 10 seconds. In each independent trial in simulations, we randomly choose K sources from this data base and the results are obtained by averaging 50 trials.

We implement the local sparsity-based speech separation by the following steps: 1) transform $\mathbf{x}(t)$ into $\ell_{\max} = 2048$ frequencies by STFT; 2) estimate $\mathbf{R}_\ell[m]$ at each frequency; 3) estimate the mixing system by LD-BSS [i.e., Eq. (5)-(6)], apply scaling and permutation alignment methods (MDP and Algorithm 1) to obtain $\hat{\mathbf{A}}_\ell$, and let $\mathbf{W}_\ell = \hat{\mathbf{A}}_\ell^\dagger$ at each frequency; 4) apply inverse discrete Fourier transform on $\{\mathbf{W}_\ell\}_{\ell=0}^{\ell_{\max}-1}$ to obtain the deconvolution matrix filter $\mathbf{W}(t)$. Note that in practice, we prewhiten $\mathbf{R}_\ell[m]$'s at each frequency before step 3). This is because that prewhitening has been

[1] Hungarian algorithm is a fast solver for mapping (or matching) problems. For details and MATLAB implementation, see [16].

Table 2: The SIRs and running times, under various $T60$s. $(K, N) = (4, 6)$; SNR= 25dB.

Method \ $T60$ (ms)		100	110	120	130	140	150	160
LD-BSS	SIR (dB)	**20.95**	**20.55**	**18.95**	16.59	15.15	13.32	10.51
&	T_m (Sec.)	3.37	3.31	3.30	**3.24**	**3.21**	**3.14**	**3.27**
Algo. 1	T_p(Sec.)	**0.38**	**0.38**	**0.40**	**0.40**	**0.41**	**0.38**	**0.41**
LD-BSS	SIR (dB)	18.12	19.02	16.73	**17.41**	15.19	**14.51**	**13.45**
&	T_m (Sec.)	**3.16**	**3.12**	**3.17**	3.31	3.25	3.20	3.17
K-means	T_p (Sec.)	3.82	3.94	4.12	3.94	4.08	4.18	3.94
PARAFAC-SD	SIR (dB)	17.52	17.91	17.56	17.21	**15.27**	14.31	13.45
&	T_m (Sec.)	14.12	14.48	14.56	15.02	15.12	15.13	15.09
K-means	T_p (Sec.)	3.80	3.91	3.96	4.05	4.01	3.92	3.92

Table 1: The SIRs and running times, under various SNRs; $(K, N) = (4, 6)$; $T60 = 120$ms.

Method \ SNR		0	10	20	30	40
LD-BSS	SIR (dB)	6.36	10.93	**15.80**	**20.18**	**19.94**
&	T_m (Sec.)	**2.77**	**2.76**	**2.76**	**2.75**	**2.76**
Algo. 1	T_p (Sec.)	**0.90**	**0.71**	**0.37**	**0.32**	**0.32**
PARAFAC-SD	SIR (dB)	**8.56**	**12.55**	15.44	18.66	18.95
&	T_m(Sec.)	17.33	16.29	13.74	12.08	11.35
K-means	T_p (Sec.)	5.52	4.20	3.34	3.28	3.21

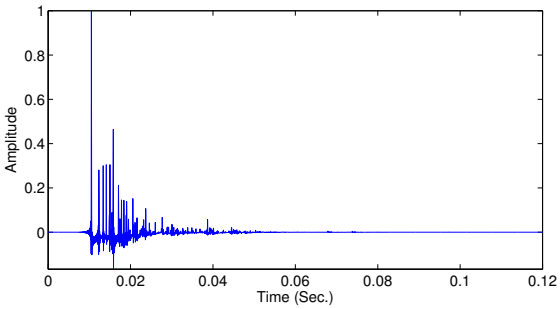

Fig. 1: The channel response between position $(1, 0.8, 1.6)$ and $(4, 0.5, 1.6)$ in the artificial room.

found helpful in improving the robustness of LD-BSS to modeling errors, such as source correlations. For details of modeling error reduction and the prewhitening technique, readers are referred to [13] and references therein.

The recovered signal-to-interference ratio (SIR) averaged across sources is adopted as the performance measure. The SIR of recovered source i is defined as [8],

$$\mathrm{SIR}_i = 10 \log_{10} \frac{\sum_t \hat{s}_{ii}^2(t)}{\sum_t \sum_{k \neq i} \hat{s}_{ik}^2(t)},$$

where $\hat{s}_{ik}(t) = \sum_{n=1}^{N} \mathbf{W}_{in}(t) \star \tilde{x}_{nk}(t)$, $\tilde{x}_{nk}(t)$ is the recorded signal at sensor n when only source k is active and $\star$ denotes the linear convolution operator.

Table 1 shows the output SIRs and running times of the local sparsity-based BSS package (including the LD-BSS for $\mathbf{A}_\ell$ estimation and the proposed permutation alignment method, denoted by "LD-BSS & Algo. 1") under various SNRs, when the reverberation time is fixed to be 120ms and the average input SIR (i.e., the SIR without separation process) is -5.03dB. We benchmark the proposed approach by a state-of-the-art BSS package [8], which includes the *parallel factor analysis via simultaneous diagonalization (PARAFAC-SD)*-based $\mathbf{A}_\ell$ estimation [5] and the K-means clustering-based permutation alignment (denoted by "PARAFAC-SD & K-means"). By clustering, centroids of source features at all frequencies are determined so that sources at each frequency can be aligned to their most correlated centroids. In this simulation, the "dominance measure" feature in [9] is employed for clustering. It can be seen that the proposed local sparsity-based BSS package yields comparable SIRs to the benchmarked method in this scenario. In particular, when SNR$<$ 20dB, the benchmarked package slightly outperforms the proposed package, while when SNR$\geq$ 20dB, the SIRs of these two methods are quite on a par. In Table 1, it can also be seen that the advantage of the proposed methods lies in the computational time. From the mixing system estimation time (T_m) and the permutation alignment time (T_p) listed in Table 1, it can be seen that the LD-BSS and proposed permutation alignment require much less time compared to that of the benchmarked methods. Specifically, in this scenario, the LD-BSS can be around 5 times faster than the PARAFAC-SD method and the local sparsity-based permutation alignment algorithm is around 10 times faster than the K-means clustering-based algorithm.

We are also interested in testing the algorithms under different reverberation times ($T60$). Generally, higher $T60$ values means more critical environments for source separation. The results are shown in Table 2. It can be seen that when $T60 \leq 130$ms, the proposed method performs best in terms of both SIR and running time. When $T60$ increases, the benchmarked package outputs slightly better SIRs. One way to get a trade-off between implementation efficiency and robustness to reverberation is to combine the LD-BSS and the clustering-based permutation alignment. In Table 2, it can be seen that such combination (denoted by "LD-BSS & K-means") can yield better SIRs than the benchmarked method when $T60 \geq 150$ms, with much less running time in total.

6. CONCLUSION

In conclusion, we have proposed an efficient method for local sparsity-based blind speech separation, using the frequency-domain approach. Specifically, we first proposed to apply the LD-BSS algorithm for mixing system estimation at each frequency. Then, a simple permutation alignment method with low implementation complexity was presented. Such simplicity stems from the exploitation of local disjointness of the source supports. Simulations in an artificial room have illustrated that the proposed methods are not only capable of recovering speech sources from convolutive mixtures with good SIRs, but also are computationally much more efficient than a state-of-the-art method.

7. REFERENCES

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

[2] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Speech Audio Process.*, vol. 8, no. 3, pp. 320 –327, May 2000.

[3] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[4] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Muller, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," *Journal of Machine Learning Research*, pp. 777–800, 2004.

[5] L. De Lathauwer and J. Castaing, "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1096 –1105, Mar. 2008.

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

[7] K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Blind speech separation using PARAFAC analysis and integer least squares," in *Proc. ICASSP 2006*, May 2006, vol. 5.

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

[9] H. Sawada, S. Araki, and S. Makino, "MLSP 2007 data analysis competition: Frequency-domain blind source separation for convolutive mixtures of speech/audio signals," in *2007 IEEE Workshop on Machine Learning for Signal Process.*, Aug. 2007, pp. 45 –50.

[10] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. ICA 1999*, 1999, vol. 99, pp. 365–370.

[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Speech Audio Process.*, vol. 12, no. 5, pp. 530 – 538, Sep. 2004.

[12] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. ISCAS 2007*, 2007, pp. 3247–3250.

[13] X. Fu and W.-K. Ma, "A simple closed-form solution for overdetermined blind separation of locally sparse quasi-stationary sources," in *Proc. ICASSP 2012*, Mar. 2012, pp. 2409–2412.

[14] A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, "Underdetermined blind separation of non-disjoint sources in the time-frequency domain," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 897 –907, Mar. 2007.

[15] 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.

[16] P. Tichavsky and Z. Koldovsky, "Optimal pairing of signal components separated by blind techniques," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 119 – 122, Feb. 2004.

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