Blind Separation of Convolutive Speech Mixtures Based on Local Sparsity and K-means

Yuyang Huang, Ping Chu, Bin Liao
College of Electronics and Information Engineering
Shenzhen University, Shenzhen 518060, China
e-mail: binliao@szu.edu.cn

Abstract—In this paper, an accurate and efficient blind source separation method based on local sparsity and K-means (LSK-BSS) is proposed. Specifically, the proposed LSK-BSS approach exploits the local sparsity of speech sources in the transformed domain to obtain closed-form solution for per-frequency mixing system estimation. On this basis, through designing superior initial points of clustering, the well-established K-means algorithm is employed to achieve accurate permutation alignment. Simulations with real reverberant speech sources show that the LSK-BSS approach yields competitive efficiency, robustness and effectiveness, in comparison with the state-of-the-arts methods.

Index Terms—Blind source separation, convolutive speech mixture, K-means, permutation ambiguity.

I. INTRODUCTION

It is known that blind source separation (BSS) aims to separate hidden sources from a mixture without information about mixing system and the characteristics of the sources. BSS has been successfully applied to various areas, such as speech processing, array signal processing, mobile communication, and analysis of astronomical data and satellite images. In this paper, we focus on BSS of convolutive speech mixtures. One of the popular and effective methods tackling this problem is frequency-domain approach [1], which transforms the signals into frequency domain and decouples the convolutive mixtures into a number of per-frequency instantaneous mixtures. By doing so, the instantaneous BSS algorithms can be straightforwardly utilized to perform BSS [2], [3].

Although existing BSS methods for convolutive speech mixtures have shown promising performance [4]–[6], their applicability in practice may be affected by the computational burden and permutation problem. Specifically, the frequency-dependent mixing system should be estimated for each frequency bin, which will consume much time. Moreover, the accuracy as well as computational efficiency of permutation alignment are usually insufficient, due to the fact that this process involves clustering thousands of vectors with large size [4] or solving highly nonconvex optimization problems [6]. Thus, a frequency domain BSS method based on local sparsity (LD-BSS) for over-determined convolutive speech mixtures has been proposed in [7], [8]. This method provides closed-form solution for per-frequency mixing system estimation as well as a numerically simple implementation of permutation alignment. Therefore, it has higher computational efficiency than traditional methods. However, as mentioned in [8], the permutation alignment process of LD-BSS may suffer error accumulation, which will certainly result in serious performance deterioration.

In this paper, we shall present a new frequency domain BSS method based on local sparsity and K-means (LSK-BSS) for over-determined convolutive speech mixtures. Following the concept of LD-BSS, we take advantage of the local sparsity property to achieve an efficient mixing system estimation. However, unlike the permutation alignment process adopted in [8], we propose to use a K-means clustering algorithm with superior initial clustering points. The K-means based permutation alignment process can significantly reduce the implementation complexity as compared to iterative techniques [9]. More importantly, we make use of the local sparsity property to produce superior initial clustering points of K-means (i.e., produce a set of seeds used as starting points) to achieve better clustering results, and hence, to achieve better performance of permutation alignment. The simulations using real speech sources demonstrate the efficiency and the effectiveness of LSK-BSS.

II. PROBLEM FORMULATION

Let us consider an array with $N$ sensors receiving $K$ source signals and assume that $N > K$ (i.e., over-determined mixing system). The received signals by the sensors is expressed in a convolutive mixture model as

$$\mathbf{x}(t) = [x_1(t), \cdots, x_N(t)]^T = \sum_{\tau=0}^{T-1} \mathbf{A}(\tau) \mathbf{s}(t - \tau)$$

(1)

where $(\cdot)^T$ is the transpose operator, $\mathbf{A}(\tau) \in \mathbb{R}^{N \times K}$ denotes the impulse response of mixing system, $T$ denotes the maximal number of delays, and $\mathbf{s}(t) = [s_1(t), \cdots, s_K(t)]^T \in \mathbb{R}^K$ contains the $K$ mutually independent speech sources.

To recover $\mathbf{s}(t)$ from the mixtures $\mathbf{x}(t)$ without knowing the mixing system, the received signal are first transformed into frequency-domain by applying short time Fourier transform (STFT) on consecutive time blocks of $\mathbf{x}(t)$ [6], [8]. Thus, we can obtain an approximately instantaneous model at multiple frequencies $f_\ell$ and $\ell = 0, \cdots, \ell_{\text{max}} - 1$, where $\ell_{\text{max}}$ represents the number of frequencies. In the $q$th time block, the

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under grants 2018A030310576 and 2020A1515010410, in part by the National Natural Science Foundation of China under grant 61771316, and in part by the Guangdong Special Support Program.
frequency-domain mixture model is written as

\[ \mathbf{x}(q) \approx \mathbf{A}_L \hat{\mathbf{s}}_L(q), \quad (2) \]

where \( \mathbf{A}_L = [a_{1,L}, \ldots, a_{K,L}] \in \mathbb{C}^{N \times K} \) denotes the spatial mixing matrix with \( a_{k,L} \in \mathbb{C}^N \) being the spatial channel from source \( k \) to the sensors at frequency \( f_L \). \( \hat{x}_L(q) = [\hat{x}_{1,L}(q), \ldots, \hat{x}_{N,L}(q)]^T \in \mathbb{C}^N \) denotes the frequency components of the mixtures at \( f_L \), and \( \hat{s}_L(q) = [\hat{s}_{1,L}(q), \ldots, \hat{s}_{K,L}(q)]^T \in \mathbb{C}^K \) denotes the frequency components of the sources at \( f_L \).

Assume that the speech sources are wide-sense stationary \( \mathbb{C}^m \)-mixing \( C \). The means clustering algorithm is applied to perform permutation alignment. In detail, the LKS-BSS approach consists of \( \mathbf{A}_L \)'s.

III. THE LSK-BSS APPROACH

In this section, we shall introduce the LSK-BSS approach in detail. Specifically, the LKS-BSS approach consists of three steps. In the first step, we estimate the per-frequency mixing systems using the local sparsity property [7]. Next, the initial clustering points are determined. Finally, the K-means clustering algorithm is applied to perform permutation alignment.

A. Per-frequency Mixing System Estimation

Under the assumption of local sparsity (local dominance) [7] (i.e., for each source \( k \), there exists a time index, indexed by \( m_k \), such that \( d_k[m_k] > 0 \) and \( d_j[m_k] = 0 \) for all \( j \neq k \)), the local covariances of those frames locally dominated by source \( k \) can be expressed as \( \mathbf{R}_L[m_k] = \sum_{k=1}^{K} d_k[m_k] \mathbf{a}_k \mathbf{a}_k^H \). Once these locally dominant covariances are available, then \( \mathbf{a}_k \)'s can be retrieved by computing the principal eigenvector of the locally dominant \( \mathbf{R}_L[m_k] \). Hence, the estimated mixing matrix \( \hat{\mathbf{A}}_L = [\mathbf{a}_{1,L}, \ldots, \mathbf{a}_{K,L}] \) can be obtained by \( \hat{\mathbf{a}}_k = \mathcal{P}(\mathbf{R}_L[m_k]), \quad k = 1, \ldots, K \), where \( \mathcal{P}(\mathbf{X}) \) denotes the principal eigenvector of \( \mathbf{X} \). The frequency components of source \( \hat{s}_k = [\hat{s}_{1,k}(q), \ldots, \hat{s}_{K,k}(q)]^T \in \mathbb{C}^K \) can be obtained by \( \hat{\mathbf{A}}_L \), where \( \hat{\mathbf{a}}_k = [\hat{s}_{1,k}(q), \ldots, \hat{s}_{K,k}(q)]^T \in \mathbb{C}^K \). Following the classical work [8], let \( \mathbf{y}_L[m] = \text{vec}(\mathbf{R}_L[m]) \) and \( \mathbf{z}_L[m] = y_L[m] \frac{\hat{\mathbf{a}}_k}{\|\hat{\mathbf{a}}_k\|_2} \), we can find locally dominant frame of each frequency as

\[
\hat{m}_{\ell k} = \begin{cases} 
\arg \max_{m_k=1, \ldots, M} \| \mathbf{z}_L[m_k] \|_2, & \text{if } k = 1 \\
\arg \max_{m_k=1, \ldots, M} \left\| \mathbf{P}_k^{H} \hat{\mathbf{R}}_{1:k-1, \ell} \mathbf{z}_L[m_k] \right\|_2, & \text{if } k \geq 2
\end{cases}
\quad (4)
\]

where \( \hat{\mathbf{R}}_{1:k-1, \ell} = [\hat{\mathbf{R}}_{\ell_1}, \ldots, \hat{\mathbf{R}}_{\ell_{k-1}}] \), and \( \mathbf{P}_k \) is the orthogonal complement project of \( \mathbf{X} \). It is seen that one can achieve per-frequency mixing system estimation efficiently by the local sparsity.

B. Initial Clustering Points Determination

The local dominance frames estimated in previous subsection have the property that the same frame is usually dominated by the same source at neighboring frequencies. In other words, if \( d_{k,L}[m] > 0 \) and \( d_i,L[m] = 0 \) for \( i \neq k \), we have \( d_{k,L}[m] > 0 \) and \( d_{i,L}[m] = 0 \), where \( v \) is a positive integer and small enough, \( d_{k,L}[m] > 0 \) and \( d_{i,L}[m] > 0 \) are dominated by the same source [7]. Furthermore, if there is no permutation ambiguity in the estimated frequency component, i.e., estimated per-frequency mixing system has been aligned, then \( \forall \hat{m}_{\ell k} \in F_{\ell} = [\hat{m}_{\ell_1}, \ldots, \hat{m}_{\ell K}] \), we have

\[
\frac{d_{\ell K}^T[m_{\ell k}]}{||d_{\ell K}^T[m_{\ell k}]||_1} = \frac{d_{\ell K}^T[m_{\ell L}]}{||d_{\ell K}^T[m_{\ell L}]||_1},
\quad (5)
\]

where \( m_{\ell k} \) is the index of the \( k \)-th identified locally dominant frame. The above identity shows that the normalized source power vectors are identical unit vectors at frequencies \( \ell \) and \( \ell - 1 \).

By exploiting the property of local dominance, in the sequel we devise a scheme to obtain superior initial clustering points by modifying the method in [8]. More precisely, we first divide the entire frequency band into \( 2^n \) equal segments, where \( n \) is a positive integer and \( 2^n \) is smaller than the number of frequency bins. For the \( i \)-th segment, we obtain lower limit and upper limit of frequency segment as \( \ell_{li} \) and \( \ell_{ui} \) respectively, \( i \in \{1, 2, \ldots, 2^n\} \). For neighboring frequency \( \ell_{li} - 1, \ell_{li} \in \{\ell_{li}, \ell_{li} + 1, \ldots, \ell_{ui}\} \), we can obtain \( \hat{d}_{\ell_{li}-1} \) as

\[
\hat{d}_{\ell_{li}-1} = \left( \hat{A}_L^* \odot \hat{A}_L \right)^T \mathbf{y}_{\ell}[m],
\quad (6)
\]

where \( \mathbf{y}_{\ell}[m] = \text{vec}(\mathbf{R}_{\ell}[m]) = \left( \hat{A}_L^* \odot \hat{A}_L \right) \mathbf{d}_{\ell}[m] \) can be derived from (3), and \( \odot \) denote the Khatri-Rao product. Moreover, we can obtain permutation matrix \( \mathbf{P}_\ell \) as

\[
\mathbf{P}_\ell = \begin{bmatrix} \hat{d}_{\ell_{li}-1, \hat{m}_{\ell 1}} & \cdots & \hat{d}_{\ell_{li}-1, \hat{m}_{\ell K}} \end{bmatrix}
\quad (7)
\]

It should be noticed that the permutation matrix \( \mathbf{P}_\ell \) may imprecise in practice, because of the modeling error. Therefore, we need map \( \mathbf{P}_\ell \) to permutation matrix by Hungarian algorithm [10]. According to \( \mathbf{P}_\ell \), the permutation ambiguity of \( s_{\ell} \) can be removed by \( \hat{s}_{\ell} = \mathbf{P}_\ell s_{\ell} \), i.e., \( s_{\ell} \) is aligned to \( s_{\ell_{li}-1} \). By this way, all source frequency components in segment \( i \) can be aligned and are expressed as \( \hat{s}_i = [\hat{s}_{\ell_{li}}, \ldots, \hat{s}_{\ell_{ui}}] \), which is a three-dimensional tensor. Let \( \mathbf{T}_i \) be the average of \( \hat{s}_i \) on \( \ell \), we can obtain a group initial clustering points and each row of \( \mathbf{T}_i \) denotes an initial point. Finally, we can obtain \( 2^n \) \( \mathbf{T}_i \)’s. Owing to the fact that the permutation alignment is achieved with \( 2^n \) small segments rather than entire frequency band, the problem of error accumulation can be overcome effectively. Since there are permutations ambiguity in different \( \mathbf{T}_i \), we perform the pairing by \( n \) times
and apply the Hungarian algorithm to $2^n$ $T_i$ to obtain final $K$ initial clustering points. Specifically, in each iteration, we first pair all $T_i$ randomly, then apply Hungarian algorithm to each pair. After the two $T_i$ of each pair have been matched, we compute their mean as a new $T_i$. The number of $T_i$ is halved after each iteration. In the $n$th iteration, we can obtain the ultimate $T_i$ as final $K$ initial clustering points.

It is worth mentioning that if the local dominance frames in each segment are obvious enough, the obtained final initial points should belong to different sources and are close to the centroid of the frequency components of different sources. These properties can improve the cluster accuracy of K-means and speed up the convergence of K-means. In other words, the K-means is able to achieve the permutation alignment of frequency components precisely and efficiently.

C. Permutation Alignment Based on K-means

Before achieve permutation alignment, we use minimum distortion principle to deal with the scaling ambiguity [11].

The rationale of applying K-means to permutation alignment is that source profiles come from the same source, but at different frequencies, they are still more similar than those from other sources [12]. The process of achieving permutation alignment by K-means is summarized as follows.

Step 1: The source profile $\hat{\gamma}_{k,\ell}$ is calculated, where $k = 1, \ldots, K$. Define the matrix $\hat{\Gamma}_\ell \in K \times N_\ell$ which collets the $K$ profiles $\hat{\gamma}_{k,\ell}$, where $N_\ell$ is the length of the profile, the profiles $\hat{\gamma}_{k,\ell}(q)$ are computed for overlapping frames over the whole signal in practice. Let $\hat{\Gamma} \in \ell_{\max}K \times N_\ell$ be the concatenation of the matrices $\hat{\Gamma}_\ell, \ell = 1, \ldots, L$.

Step 2: Get the superior initial clustering points by the scheme discussed in the above subsection.

Step 3: With the obtained initial clustering points, the K-means algorithm is applied to $\hat{\Gamma}_\ell$. This process produces a frequency independent centroid matrix $C \in [c_1, \ldots, c_K]^T \in \mathbb{C}^{K \times N_\ell}$, which makes the sum of the within-cluster-distances of all clusters is minimized (within-cluster-distance is the sum of point-to-cluster-centroid distances).

Step 4: Find the $K \times K$ permutation matrix $\Pi_\ell$ for each frequency bin by

$$\min_{\Pi_\ell} ||\hat{C} - \hat{\Gamma}_\ell\Pi_\ell||^2_2, \quad \ell = 1, \ldots, L, \quad (8)$$

Step 5: Achieve permutation alignment by $\Pi_\ell$.

Simulation results below show that the superior initial clustering points can effectively improve the clustering accuracy of K-means.

IV. SIMULATION

In this subsection, we demonstrate the performance of proposed LSK-BSS. The size of artificial room is $5m \times 3.5m \times 3m$. We set the number of sources to be 4 and the number of sensors to be 6. The sensor positions are $(4, 0.5, 1.6), (4, 1.1, 1.6), (4, 1.5, 1.6), (4, 2, 1.6), (4, 2.5, 1.6)$ and $(4, 3, 1.6)$, while the source positions are $(1, 0.8, 1.6), (1, 1.6, 1.6), (1, 2.4, 1.6)$ and $(1, 3.2, 1.6)$. In the data base, each source have been truncated into 10 seconds and sampled at 16 KHz. The sources are randomly chosen and built to convolutive mixtures in each independent trial. The final experimental results are averaged over 50 trials.


The settings of LSK-BSS are as follow: the number of frequency bins is 2048; the percentage of overlap between two consecutive fast fourier transformation (FFT) frames is 0.5; the window for FFT computation is hanning; the number of consecutive overlapping FFT frames that are used to compute the sample mean estimate of covariance matrices is 7; $n$ is 3; log profiles is used. The number of frequency bins of benchmarks are set to 2048 and other settings of benchmarks are keep default.

We test the algorithms under different reverberation times (7’60). Fig. 1 shows the BSS performance under SIR criterion [15], which compares the LSK-BSS method and benchmarks and T60 varies form 80ms to 180ms. The average input SIR is $-4.93$ dB. From Fig. 1 it can be seen that: 1) LSK-BSS provides much better SIR performance than other algorithms; 2) the SIR of LSK-BSS is higher than PARAFAC-kmeans, PARAFAC-kmeans++, LD-BSS, LD-kmeans, LD-kmeans++ by around 1.26-4.59dB, 1.11-5.37dB, 2.84-3.99dB, 1.86-6.42dB, 1.73-4.25dB, respectively; 3) the SIR of LSK-BSS is better than LD-BSS, LD-kmeans and LD-kmeans++, even the used methods for per-frequency mixing system estimation are identical; and 4) the SIR of LSK-BSS is better than other K-means methods including PARAFAC-kmeans, PARAFAC-kmeans++, LD-kmeans and LD-kmeans++. The main reason is that LSK-BSS utilize the property of local sparsity to find superior initial clustering points for K-means, which improve the clustering accuracy and robustness of K-means significant effectively.

Table I shows the worst three SIR performance of partial algorithms on various T60. From this table, we can observe that: 1) the worst SIR performance of LSK-BSS is better than other methods, which means LSKM-BSS has excellent effectiveness in each trial; 2) K-means and K-means++ always fall into local optimum and cause effect of permutation alignment deterioration, since they sensitive to initial points, LSK-BSS can ensure high accurate and stability of permutation alignment by the superior initial points.

Fig. 2 shows the corresponding average runtimes of the
TABLE I
WORST 1, WORST 2, AND WORST 3 PERFORMANCES OF VARIOUS ALGORITHMS ON VARIOUS T60

<table>
<thead>
<tr>
<th>Algorithm/T60</th>
<th>PARAFAC-kmean++</th>
<th>LD-BSS</th>
<th>LD-kmeans</th>
<th>LD-kmeans++</th>
<th>LSK-BSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>worst1</td>
<td>worst2</td>
<td>worst3</td>
<td>worst1</td>
<td>worst2</td>
</tr>
</tbody>
</table>

Fig. 1. SIR performance of various algorithms under different T60.

algorithms. We can observe that LD-BSS, LD-kmeans, LD-kmeans++ and LSK-BSS have almost the same performance of runtime, while LD-kmeans and LD-kmeans++ perform slightly better than LD-BSS and LSK-BSS. The PARAFAC-based methods have much longer run time, owing to the fact that PARAFAC-based methods use three-way tensor decomposition to tackle the per-frequency mixing system estimation problem, while LD-based approaches utilize local sparsity of sources and have closed-form solutions for per-frequency mixing system estimation.

V. CONCLUSION

In this paper, we presented a over-determine BSS approach named as LSK-BSS based on local sparsity and K-means for convolutive speech mixtures in frequency domain. LSK-BSS combines the property of local sparsity with K-means. More exactly, it exploits the property of local sparsity to find the superior initial clustering points for K-means. These points can help K-means to achieve permutation alignment precisely. Simulation results show that LSK-BSS provides better separation performance and robustness with much higher efficiency than existing approaches.

REFERENCES


