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

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

where $(\cdot)^\mathsf{T}$ 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 $\mathbf{s}(t) = [s_1(t), \cdots, s_K(t)]^\mathsf{T} \in \mathbb{R}^K$ 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_ℓ and $\ell = 0, \cdots, \ell_{\max} - 1$, where $\ell_{\max}$ represents the number of frequencies. In the qth 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

$$\check{\mathbf{x}}(q) \approx \mathbf{A}_\ell \check{\mathbf{s}}_\ell(q), \tag{2}$$

where $\mathbf{A}_\ell = [\mathbf{a}_{1,\ell}, \cdots, \mathbf{a}_{K,\ell}] \in \mathbb{C}^{N \times K}$ denotes frequency-dependent mixing matrix with $\mathbf{a}_{k,\ell} \in \mathbb{C}^N$ being the spatial channel from source k to the sensors at frequency f_ℓ, $\check{\mathbf{x}}_\ell(q) = [\check{x}_{1,\ell}(q), \cdots, \check{x}_{N,\ell}(q)]^\mathsf{T} \in \mathbb{C}^N$ denotes the frequency components of the mixtures at f_ℓ, $\check{\mathbf{s}}_\ell(q) = [\check{s}_{1,\ell}(q), \cdots, \check{s}_{K,\ell}(q)]^\mathsf{T} \in \mathbb{C}^K$ denotes the frequency components of the sources at f_ℓ. Assume that the speech sources are wide-sense stationary signals in short durations, one can chop $\check{s}(q)$ into short frames of length L. Therefore, recalling the independency of sources, the local covariance of sources in the mth frame is given by $\mathbf{D}_\ell[m] = \mathrm{E}\{\check{\mathbf{s}}_\ell(q)\check{\mathbf{s}}_\ell(q)^H\} = \mathrm{Diag}(\mathbf{d}_\ell[m])$, where $q \in [(m-1)L+1, mL]$ and $\mathbf{d}_\ell = [d_{1,\ell}[m], \cdots, d_{K,\ell}[m]]^T$ with $d_{k,\ell}[m] = \mathrm{E}\{|\check{s}_{k,q}(q)|^2\}$ being the power of the kth source in the mth frame. Since speech sources are generally non-stationary in long term, it is reasonable to assume that $\mathbf{D}_\ell[m]$ is static in each frame but varies from frame to frame. Based on the above signal model, the local covariance of $\check{x}_\ell(q)$ in the mth frame can be written as

$$\mathbf{R}_\ell[m] = \mathrm{E}\{\check{\mathbf{x}}_\ell(q)\check{\mathbf{x}}_\ell(q)^H\} = \sum_{k=1}^{K} d_{k,\ell}[m]\mathbf{a}_{k,\ell}\mathbf{a}_{k,\ell}^H, \tag{3}$$

which is estimated as $\mathbf{R}_\ell[m] \approx \frac{1}{L}\sum_{q=(m-1)L+1}^{mL} \check{x}_\ell(q)\check{x}_\ell(q)^H$ in practice. The BSS problem in frequency-domain is thus amounted to estimate $\mathbf{A}_\ell$ at each frequency f_ℓ from $\mathbf{R}_\ell[m]$'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 frame, 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 $\mathbf{R}_\ell[m_{\ell k}] = d_{k,\ell}[m_{\ell k}]\mathbf{a}_{k,\ell}\mathbf{a}_{k,\ell}^H$. Once these locally dominant covariances are available, then $\mathbf{a}_k' s$ can be retrieved by computing the principal eigenvector of the locally dominant $\mathbf{R}_\ell[m_{\ell k}]$. Hence, the estimated mixing matrix $\hat{\mathbf{A}}_\ell = [\hat{\mathbf{a}}_{1,\ell}, ..., \hat{\mathbf{a}}_{K,\ell}]$ can be obtained by $\hat{\mathbf{a}}_{k,\ell} = \mathcal{P}(\mathbf{R}_\ell[\hat{m}_{\ell k}])$, $k = 1, \cdots, K$, where $\mathcal{P}(\mathbf{X})$ denotes the principal eigenvector of $\mathbf{X}$. Thus, the frequency components of sources $\hat{\mathbf{s}}_\ell = [\hat{\mathbf{s}}_{1,\ell}, \cdots, \hat{\mathbf{s}}_{K,\ell}]^T \in \mathbb{C}^{K \times M}$ can be obtain by $\hat{\mathbf{A}}_\ell$, where $\hat{\mathbf{s}}_{k,\ell} = [\hat{s}_{k,\ell}[1], \cdots, \hat{s}_{k,\ell}[M]]^T \in \mathbb{C}^M$. Following the classical work [8], let $\mathbf{y}_\ell[m] = \mathrm{vec}(\mathbf{R}_\ell[m])$ and $\mathbf{z}_\ell[m] = \frac{\mathbf{y}[m]}{\mathrm{Tr}(\mathbf{R}_\ell[m])}$, we can find locally dominant frame of each frequency as

$$\hat{m}_{\ell k} = \begin{cases} \arg \max\limits_{m_\ell = 1, \cdots, M} \|\mathbf{z}[m_\ell]\|_2, & \text{if } k = 1 \\ \arg \max\limits_{m_\ell = 1, \cdots, M} \left\|\mathbf{P}^\perp_{\hat{\mathbf{H}}_{1:k-1,\ell}} \mathbf{z}[m_\ell]\right\|_2, & \text{if } k \geq 2 \end{cases} \tag{4}$$

where $\hat{\mathbf{h}}_{\ell i} = \mathbf{z}[\hat{m}_{\ell i}]$, $\hat{\mathbf{H}}_{1:k-1,\ell} = [\hat{\mathbf{h}}_{\ell 1}, \cdots, \hat{\mathbf{h}}_{\ell k-1}]$, and $\mathbf{P}^\perp_{\mathbf{X}}$ 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 $d_{k,\ell}[m] > 0$ and $d_{i,\ell}[m] = 0$ for $i \neq k$, we have $d_{k,\ell\pm v}[m] > 0$ and $d_{i,\ell\pm v}[m] = 0$, where v is a positive integer and small enough, $d_{k,\ell}[m] > 0$ and $d_{k,\ell\pm v}[m] > 0$ 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 $\forall\ \hat{m}_{\ell k} \in F_\ell = \{\hat{m}_{\ell 1}, \cdots, \hat{m}_{\ell K}\}$, we have

$$\frac{\mathbf{d}_\ell[\hat{m}_{\ell k}]}{||\mathbf{d}_\ell[\hat{m}_{\ell k}]||_1} = \frac{\mathbf{d}_{\ell-1}[\hat{m}_{\ell k}]}{||\mathbf{d}_{\ell-1}[\hat{m}_{\ell k}]||_1}, \tag{5}$$

where $\hat{m}_{lk}$ is the index of the kth identified locally dominant frame. The above identity shows that the normalized source power vectors are identical unit vectors at frequencies ℓ 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 positive integer and 2^n is smaller than the number of frequency bins. For the ith segment, we obtain lower limit and upper limit of frequency segment as ℓ_{li} and ℓ_{ui} respectively, $i \in \{1, 2, \cdots, 2^n\}$. For neighboring frequency $\ell_i - 1, \ell_i \in \{\ell_{li}, \ell_{li}+1, \cdots, \ell_{ui}\}$, we can obtain $\hat{\mathbf{d}}_{\ell_i - 1}[\hat{m}_{\ell k}]$ as

$$\hat{\mathbf{d}}_\ell[m] = (\hat{\mathbf{A}}_\ell^* \odot \hat{\mathbf{A}}_\ell)^\dagger \mathbf{y}_\ell[m], \tag{6}$$

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

$$\mathbf{P}_\ell = \left[\frac{\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell 1}]}{||\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell 1}]||_1}, \cdots, \frac{\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell K}]}{||\hat{\mathbf{d}}_{\ell-1}[\hat{m}_{\ell K}]||_1}\right]. \tag{7}$$

It should be noticed that the permutation matrix $\mathbf{P}_{\ell_i}$ may imprecise in practice, because of the modeling error. Therefore, we need map $\mathbf{P}_{\ell_i}$ to permutation matrix by Hungarian algorithm [10]. According to $\mathbf{P}_{\ell_i}$, the permutation ambiguity of $\mathbf{s}_{\ell_i}$ can be removed by $\hat{\mathbf{s}}_{\ell_i} = \mathbf{P}_{\ell_i}\mathbf{s}_{\ell_i}$, i.e., $\mathbf{s}_{\ell_i}$ is aligned to $\mathbf{s}_{\ell_i - 1}$. By this way, all source frequency components in segment i can be aligned and are expressed as $\hat{\mathbf{s}}_i = [\hat{\mathbf{s}}_{\ell_{li}}, \cdots, \hat{\mathbf{s}}_{\ell_{ui}}]$, which is a three-dimensional tensor. Let $\mathbf{T}_i$ be the average of $\hat{\mathbf{s}}_i$ on ℓ, 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 $\mathbf{T}_i$ to obtain final K initial clustering points. Specifically, in each iteration, we first pair all $\mathbf{T}_i$ randomly, then apply Hungarian algorithm to each pair. After the two $\mathbf{T}_i$ of each pair have been matched, we compute their mean as a new $\mathbf{T}_i$. The number of $\mathbf{T}_i$ is halved after each iteration. In the nth iteration, we can obtain the ultimate $\mathbf{T}_i$ 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, thy 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 $k = 1, \cdots, K$. Define the matrix $\hat{\mathbf{\Gamma}}_\ell \in K \times N_\ell$ which collets the K profiles $\hat{\gamma}_{k,\ell}$, where N_ℓ 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 $\hat{\mathbf{\Gamma}} \in \ell_{max}K \times N_\ell$ be the concatenation of the matrices $\hat{\mathbf{\Gamma}}_\ell, \ell = 1, \cdots, L$.

$Step$ 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 $\hat{\mathbf{\Gamma}}_\ell$. This process produces a frequency independent centroid matrix $\hat{\mathbf{C}} = [\hat{\mathbf{c}}_1, \cdots, \hat{\mathbf{c}}_K]^T \in \mathbb{C}^{K \times N_\ell}$, 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 $\mathbf{\Pi}_\ell$ for each frequency bin by

$$\min_{\mathbf{\Pi}_\ell} ||\hat{\mathbf{C}} - \hat{\mathbf{\Gamma}}_\ell \mathbf{\Pi}_\ell||_F^2, \quad \ell = 1, \cdots, L, \tag{8}$$

$Step$ 5: Achieve permutation alignment by $\mathbf{\Pi}_\ell$.

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×3.5m×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, 1.6), (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.

For comparison, the following methods are compared: 1) The BSS package [6] based on the parallel factor analysis via simultaneous diagonalization (PARAFAC-SD)-based $\mathbf{A}_\ell$ estimation [13] and the K-means clustering-based permutation alignment (denoted by "PARAFAC-Kmeans"), 2) The sparsity based convolutive BSS algorithm based on the local sparsity based $\mathbf{A}_\ell$ estimation and permutations alignment (denoted by "LD-BSS") [7], 3) The latest local sparsity based convolutive BSS algorithm proposed in [8] based on the local sparsity based $\mathbf{A}_\ell$ estimation and the K-means clustering-based permutation alignment (denoted by "LD-Kmeans"), and 4) The extended algorithms of PARAFAC-Kmeans and LD-Kmeans which use K-means++ algorithm [14] to replace K-means algorithm (denoted by "PARAFAC-Kmeans++" and "LD-Kmeans++" respectively).

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 ($T60$). 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 LSMK-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
WORST1, WORST2, AND WORST3 PERFORMANCES OF VARIOUS ALGORITHMS ON VARIOUS T60

Algorithm/T60	PARAFAC-kmean++			LD-BSS			LD-kmeans			LD-kmeans++			LSK-BSS		
	worst1	worst2	worst3	worst1	worst2	worst3	worst1	worst2	worst3	worst1	worst2	worst3	worst1	worst2	worst3
80	4.5076	6.5200	6.7492	9.5526	13.0917	13.5115	1.9271	3.6326	5.5743	4.1033	7.2091	7.6422	**16.3626**	**17.6915**	**17.7605**
90	8.9401	9.2025	9.2093	10.4091	12.3816	13.4618	0.7186	4.4518	7.2126	4.0343	5.8426	6.3817	**16.5722**	**16.8935**	**17.7045**
100	1.2024	5.7270	8.4511	6.9061	6.9685	8.9011	5.2332	5.4323	6.8862	5.6526	6.4111	7.3955	**15.5188**	**15.6357**	**16.6997**
110	4.9466	6.2989	6.5425	4.9644	6.3190	7.0948	3.9844	6.5847	8.6641	1.5710	2.9238	4.7039	**12.6546**	**14.1102**	**14.3179**
120	1.9130	2.9333	5.7400	7.9406	8.6189	8.6870	3.4944	4.9234	5.2621	4.5134	5.8578	6.6322	**12.9955**	**13.4967**	**13.8708**
130	0.9690	5.6698	6.2748	3.2561	5.1018	5.2921	2.8018	4.4553	4.5748	5.0761	5.8923	6.3572	**11.2799**	**12.8049**	**13.3241**
140	2.0592	4.0159	6.7402	0.9828	4.2776	6.1965	3.1690	4.6588	4.7195	4.2128	4.7955	4.8106	**12.0356**	**12.2244**	**13.3274**
150	1.0364	5.4412	5.5879	1.9682	4.6807	4.9105	5.1969	5.4198	5.6762	4.1349	6.1176	6.4479	**7.3111**	**7.5246**	**11.6582**
160	4.3016	5.5885	6.1702	0.8885	1.8230	4.3842	2.5257	3.7935	4.9926	3.8835	4.9199	4.9847	**11.1875**	**11.2573**	**11.5353**
170	2.7060	4.4953	4.9008	2.5924	4.0706	4.5891	5.8808	5.9505	6.2181	0.2856	3.1108	5.7081	**6.0441**	**6.6463**	**11.3312**
180	4.1700	4.1919	4.6409	4.5799	5.5668	6.3261	0.9078	3.7822	4.5209	3.7474	4.1966	4.8949	**8.7019**	**9.9226**	**10.4288**

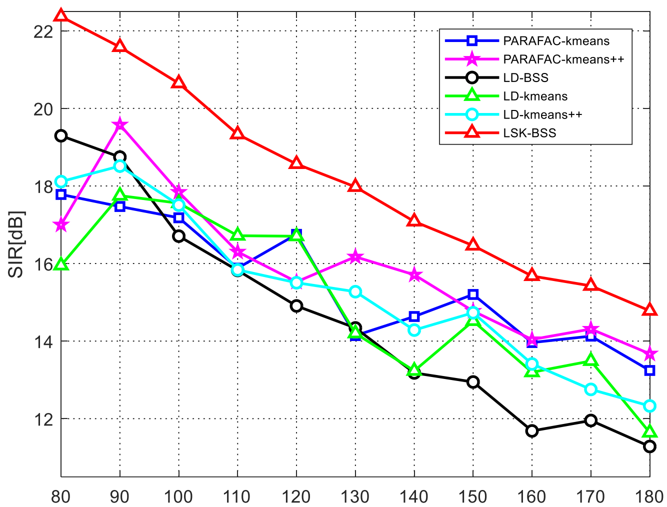

Fig. 1. SIR performance of various algorithms under different $T60$.

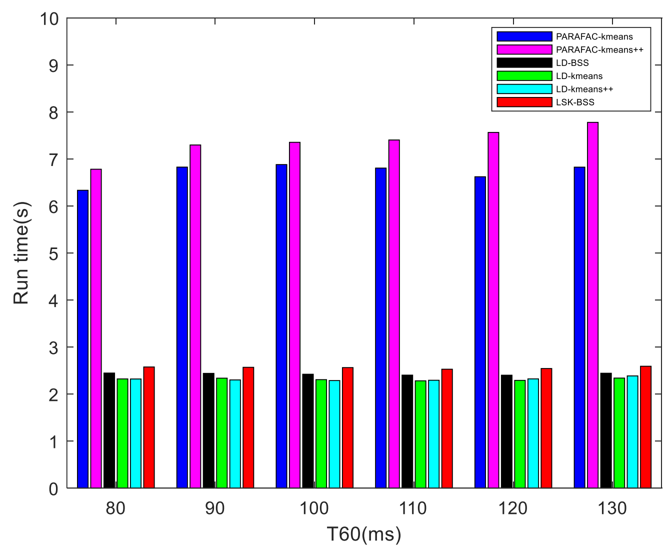

Fig. 2. Run time of various algorithms under different $T60$.

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

[1] K. Rahbar and J. P. Reilly, "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 832–844, 2005.
[2] M. Behr and A. Munk, "Identifiability for blind source separation of multiple finite alphabet linear mixtures," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5506–5517, 2017.
[3] Z. Liang, C. Xun, X. Ji, and Z. Wang, "Underdetermined joint blind source separation of multiple datasets," *IEEE Access*, vol. 5, pp. 7474–7487, 2017.
[4] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.
[5] L. J. Zhang, J. Yang, K. W. Lu, and Y. D. Sun, "Underdetermined blind source separation based on time-frequency method using cyclostationary characteristic," *Acta Armamentarii*, vol. 36, no. 4, pp. 703–709, 2015.

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

[7] X. Fu and W. Ma, "Blind separation of convolutive mixtures of speech sources: Exploiting local sparsity," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4315–4319, May 2013.

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

[9] 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 Trans Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

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

[11] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of the 41st SICE Annual Conference. SICE 2002*, vol. 4, pp. 2138–2143, Aug 2002.

[12] D. . Pham, C. Serviere, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings*, vol. 2, pp. 73–76, July 2003.

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

[14] D. ARTHUR, "k-means++ : the advantages of careful seeding," in *Eighteenth Acm-siam Symposium on Discrete Algorithms*, vol. 9, pp. 1027–1035, January 2007.

[15] C. Servire and D. T. Pham, "Permutation correction in frequency-domain in blind separation of speech mixtures," *Eurasip Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–16, 2006.